



ALAGAPPA UNIVERSITY

[Accredited with 'A+' Grade by NAAC (CGPA:3.64) in the Third Cycle
and Graded as Category-I University by MHRD-UGC]

KARAIKUDI – 630 003

DIRECTORATE OF DISTANCE EDUCATION



INFORMATION PROCESSING AND RETRIEVAL

Master of Library & Information Science 323 11



INFORMATION PROCESSING AND RETRIEVAL

I - Semester

Master of Library & Information Science



ALAGAPPA UNIVERSITY

[Accredited with 'A+' Grade by NAAC (CGPA:3.64) in the Third Cycle
and Graded as Category-I University by MHRD-UGC]

(A State University Established by the Government of Tamil Nadu)

KARAIKUDI – 630 003



Directorate of Distance Education

Master of Library & Information Science

I - Semester

323 11

INFORMATION PROCESSING AND RETRIEVAL

Author:

Gayatri Kalbag, Freelance Author

"The copyright shall be vested with Alagappa University"

All rights reserved. No part of this publication which is material protected by this copyright notice may be reproduced or transmitted or utilized or stored in any form or by any means now known or hereinafter invented, electronic, digital or mechanical, including photocopying, scanning, recording or by any information storage or retrieval system, without prior written permission from the Alagappa University, Karaikudi, Tamil Nadu.

Information contained in this book has been published by VIKAS® Publishing House Pvt. Ltd. and has been obtained by its Authors from sources believed to be reliable and are correct to the best of their knowledge. However, the Alagappa University, Publisher and its Authors shall in no event be liable for any errors, omissions or damages arising out of use of this information and specifically disclaim any implied warranties or merchantability or fitness for any particular use.



VIKAS® is the registered trademark of Vikas® Publishing House Pvt. Ltd.

VIKAS® PUBLISHING HOUSE PVT. LTD.

E-28, Sector-8, Noida - 201301 (UP)

Phone: 0120-4078900 • Fax: 0120-4078999

Regd. Office: 7361, Ravindra Mansion, Ram Nagar, New Delhi 110 055

• Website: www.vikaspublishing.com • Email: helpline@vikaspublishing.com

Work Order No. AU/DDE/DE1-238/Preparation and Printing of Course Materials/2018 Dated 30.08.2018 Copies - 500

SYLLABI-BOOK MAPPING TABLE

Information Processing and Retrieval

Syllabi	Mapping in Book
BLOCK I: CLASSIFICATION SCHEMES Unit I : Concepts of Information transfer – Universe of subjects Unit II : Structure & development – Impact on the schemes for classification - CC, DDC, UDC, & LC	Unit 1: Concepts of Information Transfer (Pages 1-12); Unit 2: Structure and Development (Pages 13-38)
BLOCK II: INDEXING TECHNIQUES Unit III : Indexing Languages – Vocabulary Control – Thesaurus Unit IV : Design of indexing languages, general theory of subject indexing languages. Unit V : Indexing Systems & Techniques – Pre coordinate indexing – PRECIS, POPSI, Chain indexing – Relational indexing, Unit VI : Post Coordinate Indexing Systems, Uniterm Indexing, Citation Indexing, KWIC and KWOC, Evaluative Studies – Crane field. I.	Unit 3: Indexing Languages, Vocabulary Control and Thesaurus (Pages 39-50); Unit 4: Design of Indexing Languages (Pages 51-64); Unit 5: Indexing Systems and Techniques – Precoordinate Indexing (Pages 65-74); Unit 6: Indexing Systems and Techniques – Post Coordinate Indexing (Pages 75-86)
BLOCK III: BIBLIOGRAPHIC STANDARDS AND FORMATS Unit VII : Bibliographic Standards – ISBD, (G), AACR 2R, ISBN, ISDN, ISSN, ISO 2709 Unit VIII : Bibliographic Formats - Bibliographic Standards : MARC, CCF, UNIMARC, MARC21, MARC XML, Dublin Core Z39.5.	Unit 7: Bibliographic Standards (Pages 87-100); Unit 8: Bibliographic Formats (Pages 101-112)
BLOCK IV: INFORMATION RETRIEVAL SYSTEM Unit IX : Information Retrieval System – Structure, Functions and Components Unit X : Search strategy – Criteria for evaluation – Recall, Precision – Relevance and failure analysis.	Unit 9: Information Retrieval System: Basics (Pages 113-120); Unit 10: Overview of Search Strategy (Pages 121-138)
BLOCK V: WEB TECHNOLOGY Unit XI : Automatic Indexing, Web Ontology Unit XII : Sequential file, structure of a sequential file, inverted file, structure of an index file, matching criteria,	Unit 11: Boolean Logic (Pages 139-150); Unit 12: Recent Trends in IRS (Pages 151-170);

Unit XIII : Boolean logic, limitations of Boolean logic, processing query expression: rules for operations

Unit XIV : Recent Trends in IRS - Internet information retrieval
- Web-based information retrieval

Unit 13: Automatic Indexing and
Web Ontology

(Pages 171-200);

Unit 14: Sequential File Access and
Structure of Index
(Pages 201-218)

CONTENTS

INTRODUCTION

BLOCK I: CLASSIFICATION SCHEMES

UNIT 1 CONCEPTS OF INFORMATION TRANSFER 1-12

- 1.0 Introduction
- 1.1 Objectives
- 1.2 Concept of Information Transfer and Universe of Subjects
- 1.3 Answers to Check Your Progress Questions
- 1.4 Summary
- 1.5 Key Words
- 1.6 Self-Assessment Questions and Exercises
- 1.7 Further Readings

UNIT 2 STRUCTURE AND DEVELOPMENT 13-38

- 2.0 Introduction
- 2.1 Objectives
- 2.2 Structure of Library Classification
- 2.3 Colon Classification
- 2.4 Dewey Decimal Classification
- 2.5 Universal Decimal Classification (USC)
- 2.6 Library of Congress Classification
- 2.7 Answers to Check Your Progress Questions
- 2.8 Summary
- 2.9 Key Words
- 2.10 Self Assessment Questions and Exercises
- 2.11 Further Readings

BLOCK II: INDEXING TECHNIQUES

UNIT 3 INDEXING LANGUAGES, VOCABULARY CONTROL AND THESAURUS 39-50

- 3.0 Introduction
- 3.1 Objectives
- 3.2 Indexing Languages
- 3.3 Vocabulary Control
- 3.4 Thesaurus
- 3.5 Answers to Check Your Progress
- 3.6 Summary
- 3.7 Key Words
- 3.8 Self-Assessment Questions and Exercises
- 3.9 Further Readings

UNIT 4 DESIGN OF INDEXING LANGUAGES 51-64

- 4.0 Introduction
- 4.1 Objectives
- 4.2 Design of Indexing Languages
- 4.3 General Theory of Subject Indexing and Thesaurus
- 4.4 Answers to Check Your Progress Questions
- 4.5 Summary
- 4.6 Key Words
- 4.7 Self Assessment Questions and Exercises
- 4.8 Further Readings

UNIT 5 INDEXING SYSTEMS AND TECHNIQUES – PRECOORDINATE INDEXING 65-74

- 5.0 Introduction
- 5.1 Objectives
- 5.2 Techniques of Indexing Systems
 - 5.2.1 Precis
 - 5.2.2 POPSI
 - 5.2.3 Chain Indexing
 - 5.2.4 Relational Indexing
 - 5.2.5 Other Indexing Processes
- 5.3 Answers to Check Your Progress Questions
- 5.4 Summary
- 5.5 Key Words
- 5.6 Self-Assessment Questions and Exercises
- 5.7 Further Readings

UNIT 6 INDEXING SYSTEMS AND TECHNIQUES – POST COORDINATE INDEXING 75-86

- 6.0 Introduction
- 6.1 Objectives
- 6.2 Post Coordinate Indexing System
 - 6.2.1 Uni-Term Indexing
- 6.3 Non-Conventional Indexing
 - 6.3.1 Citation Indexing
 - 6.3.2 Kwic and Kwoc
- 6.4 Evaluation Studies of Indexing Systems
 - 6.4.1 Crane Field-I
- 6.5 Answers to Check Your Progress Questions
- 6.6 Summary
- 6.7 Key Words
- 6.8 Self-Assessment Questions and Exercises
- 6.9 Further Readings

BLOCK III: BIBLIOGRAPHIC STANDARDS AND FORMATS

UNIT 7 BIBLIOGRAPHIC STANDARDS 87-100

- 7.0 Introduction
- 7.1 Objectives
- 7.2 International Standard Bibliographic Description (ISBD) Standards
- 7.3 AACR
- 7.4 ISBN Standards
- 7.5 ISDN
- 7.6 ISSN
- 7.7 ISO 2709
- 7.8 Answers to Check Your Progress Question
- 7.9 Summary
- 7.10 Key Words
- 7.11 Self Assessment Questions and Exercises
- 7.12 Further Readings

UNIT 8 BIBLIOGRAPHIC FORMATS 101-112

- 8.0 Introduction
- 8.1 Objectives
- 8.2 Kinds of Bibliographic Records
- 8.3 Bibliographic Standards
 - 8.3.1 MARC
 - 8.3.2 UNIMARC
 - 8.3.3 CCF
 - 8.3.4 MARC21
 - 8.3.5 MARCXML
 - 8.3.6 Dublin Core Z39.5
- 8.4 Answers to Check Your Progress Questions
- 8.5 Summary
- 8.6 Key Words
- 8.7 Self Assessment Questions and Exercises
- 8.8 Further Readings

BLOCK IV: INFORMATION RETRIEVAL SYSTEM

UNIT 9 INFORMATION RETRIEVAL SYSTEM: BASICS 113-120

- 9.0 Introduction
- 9.1 Objectives
- 9.2 Structure, Functions and Components
- 9.3 Answers to 'Check Your Progress' Questions
- 9.4 Summary
- 9.5 Key Words
- 9.6 Self Assessment Questions and Exercises
- 9.7 Further Readings

UNIT 10 OVERVIEW OF SEARCH STRATEGY

121-138

- 10.0 Introduction
- 10.1 Objectives
- 10.2 Search Strategy
- 10.3 Criteria of Evaluation: Recall and Precision
- 10.4 Relevance and Failure Analysis
- 10.5 Answers to Check Your Progress Questions
- 10.6 Summary
- 10.7 Key Words
- 10.8 Self Assessment Questions and Exercises
- 10.9 Further Readings

BLOCK V: WEB TECHNOLOGY

UNIT 11 BOOLEAN LOGIC

139-150

- 11.0 Introduction
- 11.1 Objectives
- 11.2 Boolean Logic: An Overview
 - 11.2.1 Limitations of Boolean Logic
- 11.3 Processing the Boolean Query or Search Operations: rules for operations
- 11.4 Answers to Check Your Progress Questions
- 11.5 Summary
- 11.6 Key Words
- 11.7 Self Assessment Questions and Exercises
- 11.8 Further Readings

UNIT 12 RECENT TRENDS IN IRS

151-170

- 12.0 Introduction
- 12.1 Objectives
- 12.2 An Overview of Recent Trends in IRS
- 12.3 Internet Information Retrieval and Web based Information Retrieval Trends
- 12.4 Answers to Check Your Progress Questions
- 12.5 Summary
- 12.6 Key Words
- 12.7 Self Assessment Questions and Exercises
- 12.8 Further Readings

UNIT 13 AUTOMATIC INDEXING AND WEB ONTOLOGY

171-200

- 13.0 Introduction
- 13.1 Objectives
- 13.2 Automatic Indexing
- 13.3 Web Ontology Language (OWL) Source Wikipedia
- 13.4 Answers to Check Your Progress Questions
- 13.5 Summary
- 13.6 Key Words

13.7 Self-Assessment Questions and Exercises

13.8 Further Readings

UNIT 14 SEQUENTIAL FILE ACCESS AND STRUCTURE OF INDEX

201-218

14.0 Introduction

14.1 Objectives

14.2 Structure of Sequential File Access

14.3 Inverted File and its structure

14.4 Matching Criteria for Index Files

14.5 Answers to Check Your Progress Questions

14.6 Summary

14.7 Key Words

14.8 Self Assessment Questions and Exercises

14.9 Further Readings

INTRODUCTION

NOTES

Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing. Automated information retrieval systems are used to reduce information overload. The concept of information retrieval presupposes that there are some documents or records containing information that have been organized in an order suitable for easy retrieval. An information retrieval system is designed to retrieve the documents or information required by the user community. It should make the right information available to the right user. Thus, an information retrieval system aims at collecting and organizing information in one or more subject areas in order to provide it to the user as soon as it is asked for.

This book, *Information Processing and Retrieval*, is written with the distance learning student in mind. It is presented in a user-friendly format using a clear, lucid language. Each unit contains an Introduction and a list of Objectives to prepare the student for what to expect in the text. At the end of each unit are a Summary and a list of Key Words, to aid in recollection of concepts learnt. All units contain Self-Assessment Questions and Exercises, and strategically placed Check Your Progress questions so the student can keep track of what has been discussed.

BLOCK - I

CLASSIFICATION SCHEMES

*Concepts of
Information Transfer*

**UNIT 1 CONCEPTS OF
INFORMATION TRANSFER**

NOTES

Structure

- 1.0 Introduction
- 1.1 Objectives
- 1.2 Concept of Information Transfer and Universe of Subjects
- 1.3 Answers to Check Your Progress Questions
- 1.4 Summary
- 1.5 Key Words
- 1.6 Self-Assessment Questions and Exercises
- 1.7 Further Readings

1.0 INTRODUCTION

The concept of subject, as interpreted by multiple researchers in the discipline of Library and Information Science, is generally preceded by the concept of document. It will therefore be prudent to remember that within the area of library and information science, the concept of subject (or subjects to be precise) will have to be studied within the concept of document. The reason for this must be made clear.

While studying the concept of subjects in the domain of library and information science, it becomes absolutely necessary that a single subject, or multiple subjects at that, would need a proper document to hold all the information that is comprised within that specific subject, or within those individual subjects. This means the document is as important or significant as individual subjects that are included and studied within library and information science.

In an attempt to make the study of this easier, researchers have been over the years been dividing this section into three different areas, each of which have been explored as separate entities in this unit.

The term information transfer is also significant here. But before we venture into that let us understand first why library and information science is important to educational institutions and academia in general. All educational institutions, whether at the school or college levels, need to provide their students with a wide range of referral material to provide them with information beyond mere text books. At the higher educational levels, especially, students need to refer to additional information on a wide range of topics and subjects that may be used in their research projects,

NOTES

classroom activities, and generally information that goes beyond the printed textbook prescribed at each grade level. As students begin to proceed towards higher education studies, it will become necessary for them to engage in research projects, carry out dissertations, use their new information to write detailed theses papers, and from there to earn their doctorate degrees.

It is important to understand that different researchers have interpreted both the concept of information and information transfer, as well as the concept of subject differently. There are some interpretations that place as much importance on the concept of subject in the domain of library and information science, as they would place on the concept of word in the domain of linguistics. In this unit, we will discuss the concepts of information transfer and universe of subjects.

1.1 OBJECTIVES

After going through this unit, you will be able to:

- Discuss the concept of information transfer
- Describe universe of subjects in information processing and retrieval

1.2 CONCEPT OF INFORMATION TRANSFER AND UNIVERSE OF SUBJECTS

The concept of subject may be interpreted either as a macro-concept or a microconcept. Within the realm or domain of library and information science, however, subject may be perceived as a macro-concept because of the vastness of its content. It is important to understand too that the concept of subject is possible to be defined independent of the concept of document when it is described as a science in itself. This is actually when it becomes ready to be defined with the concept of linguistics.

Information transfer simply refers to the process of transfer of data from one source to another using a communication medium. It includes several processes under its subject including creation, collection, storage, dissemination and retrieval of information. These processes are taken care of by the library professionals. We will learn about the input of these people engaged in library science helps in the transfer of knowledge.

The universe of subject on the other hand refers to the categorization of vast amount of knowledge in different forms present in the libraries. The universe of subjects are created as different segments based on shared characteristic, their use to the readers, the types of readers and the how these subjects are connected to each other.

Universe of Subject

There are different kinds of knowledge, be it personal or social. Social knowledge which is publicly available is what is relevant to library science since libraries contain sources of information for the perusal of the public. The different kinds of knowledge can be referred to as discipline.

Ranganathan has defined the term 'subject' as, 'An organized or systematized body of ideas, whose extension and intension are likely to fall coherently within the field of interest and comfortably within the intellectual competence and the field of inevitable specialisation of a normal person.' Subject can then be called as the logical progression of knowledge.

Since the kinds of knowledge are vast, it is important that they be classified into different categories as per the field of interest. Subjects define the categories of knowledge contained in different documents.

Library Science and retrieval systems then essentially study the universe of readers and universe of subjects since the categorization and classification of various information into neat and recognizable categories for easy retrieval is the primary aim of information processing and retrieval systems.

The concept of information transfer can be understood through the functions involved and performed in Library Science.

It is interesting, therefore, to understand how each of these concepts are inter-related, and at times, even inter-dependent.

Let us now examine Library Science in some detail.

Library Science:

Library Science is the multi-disciplinary as well as the inter-disciplinary study of the collection, compilation, segregation and management of information. To refine this definition a little more clearly, the science of library science is the science of processing a vast body of information, organizing it according to subject and topic, using it for as well as providing it for dissertation at various levels and platforms, and ensuring that academicians, educators, research scholars, and/perhaps just subject or topic enthusiasts, are able to access any of the information being so managed an organized.

It is interesting to observe the progress and development of library and information in an effort to keep pace with the development of technology. Perhaps to begin with people had to be satisfied with oral or verbal traditions of information management and distribution. As technology developed and printers came into existence, libraries began to collect and organize printed resource material. People who wished to access this information would obviously have referred to the printed sheets that had been organized and arranged according to subject as well as topic.

With the advent of more technological evolution and further inventions, this entire process must have been made easier, quicker and more professional. It may

NOTES

NOTES

also be remembered that information and knowledge of any type was traditionally accessible only to the upper castes in India, and traditionally all information of any significance used to be organized and monitored only by the upper castes. Perhaps it is only with the development of technology and the advent of such electronic devices as the printer, the printing press, the typewriter, the electric and then the electronic typewriter, and finally the computer, that people from other castes have been able to access information at will.

The library as a structure or symbol of information has been in existence since the times of Alexander, while the actual concept of library science appears to have been fairly recent. In fact, it was perhaps not until early in the 20th century that the study of the management of the library as an applied science got to be included in the curriculum at the bachelor's program in India. This corroborates the earlier information that all information traditionally used to be handled and managed by the upper castes in India.

According to research material available on this subject, the subject of library and information science has existed from the time scientific information has been accessible. Perhaps it can therefore be surmised that the concept of library and information as being applicable to other subjects as well, has evolved only in recent years.

However, there are other research scholars who believe and argue that the subject of library and information science is so vast and in-depth in itself that it should be treated as a separate science in itself. They believe that the fact that it is multidisciplinary and provides services to all the subjects within the academic and non-academic spectra, for it to be treated and identified only as an applied science. The subject in fact appears to be requiring more rigorous and in-depth examination of its basic concept as well as the entire methodology that governs its management.

There are three main groups or areas of study that students can explore within this single subject. These include information science practitioners, information system designers and information scientists.

Information Science Practitioners:

Information science is a field primarily concerned with the analysis, collection, classification, manipulation, storage, retrieval, movement, dissemination, and protection of information. Practitioners within and outside the field study application and usage of knowledge in organizations along with the interaction between people, organizations, and any existing information systems with the aim of creating, replacing, improving, or understanding information systems. Historically, information science is associated with computer science, psychology, and technology. However, information science also incorporates aspects of diverse fields such as archival science, cognitive science, commerce, law, linguistics, museology, management, mathematics, philosophy, public policy, and social sciences.

So basically, practitioners of information science learn the science and its application within an institution or organization, where it is possible for them to learn the art of information science in practice or while practicing it.

Within the scope or purview of information science, the most common issues or problems that practitioners may encounter will relate to the problems of stakeholders.

When working at libraries, people may often approach information science practitioners with problems relating to their interactions with the computers assigned to them. Practitioners will need to be thorough in their knowledge of computer technology as well as their knowledge of information retrieval and dissemination on a multitude of subjects and topics. technologies to help solve those problems and issues.

Information science practitioners will also need to keep themselves abreast of current affairs, including developments across a wide range of subjects. They will need to have sufficient knowledge to enable them to retrieve information and transfer it to the individual computers being used by stakeholders.

In some instances, information science may also be referred to as information systems. However, both information science as well as information systems are usually considered and studied as two different fields.

The philosophy of information allows students and practitioners to study the issues and problems that crop up at the intersection of computer systems, retrieval of the required information, conceptual handling of these issues, the dynamics of information science and the maximization of the use of information to the stakeholders concerned.

The role of information practitioners within a library would therefore be the collection, classification, categorization, dissemination, retrieval, storage, distribution or transfer of that information to the concerned stakeholder, protection of that information, and analysis of that information when needed.

Information Systems Analyst:

An information systems analyst is a trained professional who designs, analyses and implements information. An information systems analyst is responsible for analysing and assessing the suitability of specific information in the context of user needs of all their stakeholders. Within the purview of their job profile, information systems analysts are also responsible for liaising with vendors, users, programmers, information technology professionals, software engineers, etc., in an overall effort to derive the desired outcomes. Information systems analysts are trained to use information technology enriched with analytical skills and design techniques to solve business problems. In simple terms, information systems analysts may be described as people who are able to identify the intrinsic changes required within organizations, design specific systems that will implement those changes and train

NOTES

NOTES

the personnel and professionals working within that organization to help them use those systems effectively and successfully.

However, it is important to understand that although information systems analysts are generally conversant with multiple computer languages, computer hardware options, operating systems, as well as computer operating languages, they do not as a rule involve themselves in or interfere in either the hardware or software development within organization they work with.

On the other hand, systems analysts may take over the responsibility of implementing systems timelines, designing staff training programs, creating cost analyses reports, and such like. Usually, systems analysts are assigned or allotted individual systems, and are expected to work in collaboration with business analysts.

A business analyst is usually in charge of identifying the business requirement, designing required changes in the business format but does not delve too deep into the software or systems programming nuances of the organization. A business analyst is responsible, in other words, for identifying the business need of an organization, then identifying the appropriate solution that will address or meet the immediate business need. However, for all matters relating to systems analyses, a business analyst will depend upon the services and cooperation of a systems analyst. A systems analyst, meanwhile, will usually identify and design coding and scripting.

This does not mean that a systems analyst should never assay the role of a business analyst. There are plenty of systems analysts who are dedicated to their professions and are able to take on the dual role of a business analyst. This way they are able to blur the thin line between the two roles of a systems analyst and a business analyst.

A trained, professional systems analyst may carry out the following responsibilities:

- Plan the systems flow from ground level up
- Identify, understand and plan for organizational, operational and human impacts and outcomes of planned and newly implemented systems
- Ensure that the planned technological and systems changes are properly and effectively implemented with the existing skill sets and processes
- Interact with internal users as well as customers to be able to understand and identify the intricacies of their business requirements that they will then be expected to record in documents
- Use the said documents in collaboration with the business analyst to recognize and address the business requirements
- Write technical requirements from a critical point of view
- Interact with the organization's software professional or architect to be able to understand and identify the software limitations

- Collaborate with the software programmers to help them during the organizational systems development by providing them with previous case studies, BPMN diagrams, flowcharts as well as UML diagrams
- Record and document the requirements or contribute to user manuals
- Design the component and provide the relevant information to the developers during the process of development within the organization

Traditionally, larger business organizations have been implementing the systems development life cycle during larger information technology projects. The systems development life cycle may be described as being a structured framework that enables the development of information systems. In effect, the systems development life cycle comprises of:

- The systems investigation
- The systems analysis
- The systems designing
- Programming
- Testing
- Implementation

Operation and maintenance

The systems analysis phase begins once the required approvals have been received from all the concerned participants. Here the term, systems analysis refers to the examination of the business problem within the organization that the concerned organization plans to resolve with the implementation of an appropriate information systems development.

It may be said that the entire purpose of conducting the systems analysis is primarily to examine the existing information system in order to gather required information that will help the systems analyst design and implement either an enhanced information system or an entirely new information system. This stage is also referred to as the system deliverable. Perhaps the most critical and sensitive aspect of information systems analysis is identifying the specific systems requirement and developing the solution that will meet that specific requirement. This set of system requirements is known as user requirements, obviously because the system users have identified them. When the systems analyst has been able to gather a considerable number of user requirements, they proceed to the next stage, which is referred to as the systems design stage.

Computer systems analysis is a profession that falls within the purview of computer information technology. Perhaps this is why a computer systems analyst is also referred to as, or may be referred to primarily referred to as, an information systems analyst.

NOTES

NOTES

Computer systems analysts endeavour to solve problems related to computer systems and the working of computers. As a matter of fact, many computer analysts establish their individual business platforms where they develop solutions in both the hardware as well as software domains. They also develop new software solutions and applications that will help enhance and enrich computer productivity. Some systems analysts may work as systems developers or as systems analysts, while a vast majority of them may assay the role of information systems specialists.

Information Scientist:

The role of information scientists is distinct from that of librarians. In order to be able to work as or be hired as information scientists, people must have gained at least a bachelor's degree in the relevant subject. It is also possible to be hired as a computer information scientist if one has relevant working experience and knowledge as a computer information scientist. However, the traditional use of the term information scientist has decreased considerably over the years, resulting in the fact that these professionals are now being referred to as information officers, or in other cases as information professionals.

The term information scientist or information professional has traditionally been used, and is still being used, to refer to people who are conducting research in the field of information science.

Researcher Brian Vickey has mentioned that the Institute of Information Scientists had been set up in London in the year 1848, and that it had published a list of requirements.

Vickey mentions that the institute of information scientists merged with the Library Association in the year 2002, and then went on to form the Chartered Institute of Library and Information Professionals (CILIP).

Library and Information Scientist:

Library and information scientists are professionals who conduct research in their respective subject area. These library and information scientists may also be academics who take part in the writing of scholarly journals or essays in the field of library and information science.

It may be said that library and information scientists are not confined to one particular or specific or exclusive substrata of library and information science.

Library and information scientists may not work within the confines of any one specific form of library.

Library and information scientists may belong to every sector that is even remotely related to the gathering, retrieval and transfer of information of any type, on any subject and at any level.

Librarians:

Librarians are people who are hired at libraries. They work on professional levels to provide users access to information on a wide range of subjects. There may also be times when librarians are expected to provide users access to social and technical programming services. Apart from this, librarians may provide users instruction or training on information theory. In countries such as the United States of America and Canada, librarians are expected to possess graduate degrees in library science from recognized library schools or institutes.

These qualifications may include the Master of Library Science / Information Systems (MLSIS), the Master of Library Science (MLS); or the Master of Science in Library Science (MSLS).

The term librarian has been derived from the Latin word *liber* which means book. So, traditionally, librarians were expected to assay the role of gatherers of books or collectors of books. In that capacity, librarians were expected to gather and categorize books of all genres, retrieve them and provide necessary information to users, usually scholars and academicians when and where needed. This traditional role has been used to evolve continually in order to keep pace with the evolving changes in social and technological sciences.

In modern times, however, librarians are expected to provide users with information in a wide range of formats. These may include printed materials such as books, magazines, newspapers, and in some cases manuscripts; graphic resources such as maps, photographs, film clips, audio visual clips, audio clips and recordings; web based resource material such as digitized content, online access to web pages and private websites, etc.

In addition to all this, librarians may also provide users with such services as training in information literacy and science, computer system analysis, information science, etc.

In conclusion, it may be noted that people with the requisite qualifications and skill sets in any of the sub strata of the wide umbrella of library and information science will be able to work in large or small libraries public or private libraries or spaces with large or small collections of books and other forms of information, within business organizations to work with business analysts in order to provide solutions to systems requirements, or purely in academics.

NOTES

Check Your Progress

1. What does the philosophy of information allow students and practitioners to study?
2. List the activities which comprises the role of information practitioners.
3. State the entire purpose of conducting the systems analysis.

NOTES

1.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The philosophy of information allows students and practitioners to study the issues and problems that crop up at the intersection of computer systems, retrieval of the required information, conceptual handling of these issues, the dynamics of information science and the maximization of the use of information to the stakeholders concerned.
2. The role of information practitioners within a library includes the collection, classification, categorization, dissemination, retrieval, storage, distribution or transfer of that information to the concerned stakeholder, protection of that information, and analysis of that information when needed.
3. The entire purpose of conducting the systems analysis is primarily to examine the existing information system in order to gather required information that will help the systems analyst design and implement either an enhanced information system or an entirely new information system.

1.4 SUMMARY

- The concept of subject may be interpreted either as a macro-concept or a microconcept. Within the realm or domain of library and information science, however, subject may be perceived as a macro-concept because of the vastness of its content.
- Library Science is the multi-disciplinary as well as the inter-disciplinary study of the collection, compilation, segregation and management of information.
- There are three main groups or areas of study that students can explore within this single subject. These include information science practitioners, information system designers and information scientists.
- The role of information practitioners within a library includes the collection, classification, categorization, dissemination, retrieval, storage, distribution or transfer of that information to the concerned stakeholder, protection of that information, and analysis of that information when needed.
- An information systems analyst is a trained professional who designs, analyses and implements information. An information systems analyst is responsible for analysing and assessing the suitability of specific information in the context of user needs of all their stakeholders.
- In effect, the systems development life cycle comprises of:
 - o The systems investigation
 - o The systems analysis
 - o The systems designing

- o Programming
- o Testing
- o Implementation
- The systems analysis phase begins once the required approvals have been received from all the concerned participants. Here the term, systems analysis refers to the examination of the business problem within the organization that the concerned organization plans to resolve with the implementation of an appropriate information systems development.
- It may be said that the entire purpose of conducting the systems analysis is primarily to examine the existing information system in order to gather required information that will help the systems analyst design and implement either an enhanced information system or an entirely new information system.
- Computer systems analysis is a profession that falls within the purview of computer information technology. Perhaps this is why a computer systems analyst is also referred to as, or may be referred to primarily referred to as, an information systems analyst.
- The role of information scientists is distinct from that of librarians. In order to be able to work as or be hired as information scientists, people must have gained at least a bachelor's degree in the relevant subject. Library and information scientists are professionals who conduct research in their respective subject area. These library and information scientists may also be academics who take part in the writing of scholarly journals or essays in the field of library and information science.

NOTES

1.5 KEY WORDS

- **Library Science:** It is the multi-disciplinary as well as the inter-disciplinary study of the collection, compilation, segregation and management of information.
- **Information science:** It is a field primarily concerned with the analysis, collection, classification, manipulation, storage, retrieval, movement, dissemination, and protection of information.
- **Information systems analyst:** It refers to a trained professional who designs, analyses and implements information.

1.6 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Give a refined definition of Library Science.
2. What are the responsibilities carried out by an information analyst?

3. List the constituents of the systems development life cycle.
4. What are user requirements in operations and maintenance?
5. What is the role of a librarian in modern times?

NOTES

Long-Answer Questions

1. Discuss the subject of library science.
2. Explain the information analysts who work in the field of information science.
3. How are information scientists and library science related?
4. Describe the concept of operations and maintenance in Library science.

1.7 FURTHER READINGS

Alberico, R. and Micco M. 1990. *Expert systems for reference and Information retrieval*. West Port: Meckler.

Atchison, J. & Gilchrist, A. 1972. *Thesaurus construction: a practical manual*. London: Aslib.

Chowdhry, G.G. 2003. *Introduction to modern Information retrieval*. 2nd Ed. London: Facet Publishing.

Cleaveland, D. B. 2001. *Introduction to Indexing and abstracting*. 3rd Ed. Englewood, Colo: Libraries Unlimited.

Ghosh, S.B. and Biswas, S.C. 1998. *Subject Indexing systems: Concepts, methods and techniques*. Rev. ed. Calcutta: IASLIC.

Pandey, S.K. Ed. 2000. *Library Information retrieval*. New Delhi: Anmol.

UNIT 2 STRUCTURE AND DEVELOPMENT

NOTES

Structure

- 2.0 Introduction
- 2.1 Objectives
- 2.2 Structure of Library Classification
- 2.3 Colon Classification
- 2.4 Dewey Decimal Classification
- 2.5 Universal Decimal Classification (USC)
- 2.6 Library of Congress Classification
- 2.7 Answers to Check Your Progress Questions
- 2.8 Summary
- 2.9 Key Words
- 2.10 Self Assessment Questions and Exercises
- 2.11 Further Readings

2.0 INTRODUCTION

The history of modern library classification may be traced back to the year 1876, when the Dewey Decimal Classification, was first published in the United States of America. From that year on, different librarians across the world have been giving their own individual interpretations of the basic classification system, depending on the angle which they may have adopted to approach the task of classification.

For instance, C A Cutter, in the year 1879, published the Expansive Classification, in which he attempted to define a different type of notation by using a different method of approach to the task of classification. Further, a British librarian named James Duff Brown used yet another approach in the year 1906 for classify knowledge according to subjects, with the publication of his paper Subject Classification.

Later on the UDC, or the Universal Decimal Classification had been recognized as being significantly important, however, and Indian librarian named S R Ranganathan, in the year 1933, published his research paper titled Colon Classification, causing a watershed of sorts in the history of library classification. The Ranganathan adopted an entirely different approach to the classification from those that had been adopted up until then. Ranganathan referred to his scheme of classification as a faceted scheme of classification. In sharp contrast, the other schemes of library classification discussed in this unit would come to be referred to as the enumerative classification system. The classification schemes that would

come to be included in this category of classification would include the Dewey Decimal Classification and all other schemes that were to be patterned along similar lines of classification.

NOTES

In this unit, we will learn the different types of library classifications.

2.1 OBJECTIVES

After going through this unit, you will be able to:

- Discuss the different structures of library classification systems
- Describe the Colon Classification
- Explain the Dewey Decimal Classification
- Examine the important points of Universal Decimal Classification
- Discuss the classes of Library of Congress Classification

2.2 STRUCTURE OF LIBRARY CLASSIFICATION

According to Ranganathan, as he mentions in his *Prolegomena to Library Classification*, there are about six existing schemes or systems of library classification, besides some other special systems of library classification. These systems include the following:

- The rigidly faceted classification scheme, or the library classification with the pre-determined facets
- The almost faceted scheme of classification
- The freely faceted classification, which is an analytica-systemic system of classification which is prompted by possibilities and principles
- The enumerative classification
- The almost enumerative classification

Of these, the rigidly faceted classification and the freely faceted classification schemes have been further categorized under the main group titled the fully faceted classification.

Fully Faceted Classification

A faceted classification is made up of timelines of basic types; special isolates and common isolates primarily. Apart from this, there are a few devices that are intended for sharpening isolates that already exist, and/or developing new isolates. These timelines are brief, not compound or complicated but only simple and straightforward subjects are enumerated. Type numbers for such subjects cannot be found readymade. They are required to be systemized each and every time as per the

specified rules (grammar) mandated by the concerned schemes. As a result, the type numbers of those systemized subjects are polythetic or used only once with their structure comprised of facets is transparent. During the evolution of these classification systems, these faceted systems are consequently fairly recent besides being better equipped to bear the impending knowledge revolution.

The classifications are of two kinds: Rigidly-Faceted Classification and Freely Faceted Classification.

Rigidly-Faceted Classification

This is the initial stage in the development process of faceted classifications. As may be implied from the name, during a rigidly'-faceted classification, the facets along with their order of citation are fixed while their facet formula is pre-ordained. No facet can be left out or ignored. The first three editions (Le, of 1931 1939 and 1950 respectively) from the Colon Classification (CC) have been identified as rigidly-faceted because they developed an individual facet formula for each basic type. Due to a cluttering of facets in the type number, it had been difficult to identify the category to which category a specific facet belonged. This issue cropped up as there was only one connecting digit colon. In case an intermediate facet was missing, there was a need to insert the connecting digit also for missing facets

Freely Faceted Classification

This is said to be the final stage in the evolution process of library classification. As the name suggests, a freely faceted classification is founded on principles and postulates. There is no rigid or pre-determined facet founded on complex or compound subjects together with basic subjects.

Because this kind of categorization is founded on analysis and synthesis, each subject is able to determine its own facet formula. The facet formula is transparent and open. This allows it to carry on the analysis and synthesis of facets. This allows the sequence of facets to be guided by postulates and principles.

A synonym for this kind of classification is Analytical-Synthetic Classification. Edition 4 (1952) to Edition 6 (460 of the Colon Classification are examples of a freely faceted classification. Yet another example of this type of species could be said to be the Bibliographic Classification Edition 2 (BC-2) (1977-) modified by J. Mills. The immunity of this type of species is practically infinite. The class numbers are-co-extensive, as well as short-lived.

But some people may perceive these as almost-freely faceted classification schemes in which the use of varying indicator digits for diverse kinds of facets. When the concept of Rounds and Levels is removed, the strict rigidity in the numeric and sequence of facets that could take place in a compound subject. In spite of this, a few of the rigidity linked facts may exist that pertain to the levels of facet within a round.

NOTES

NOTES

However, with support from Sector Notation, the strictness in the number of levels of the facets and their sequence in a round lurking as many as CC Ed. 6 has been taken out of CC Ed. 7, as it realizes that facets belong to compound subjects and not to one primary subject. This precisely why re-determination of the facets for all compound subjects may possibly go with any basic or primary subject is ruled out. Obviously this is the reason why it has been defined as fully Freely Faceted Scheme of Classification. Ranganathan's Colon Classification, is therefore a very good example of a Freely Faceted Analytico-Synthetic Classification guided by postulates and principles.

Let us now explore the four main schemes of classification, namely Colon Classification, Dewey Decimal Classification, Universal Decimal Classification and Library Classification in some detail.

Check Your Progress

1. Where did Ranganathan first mentioned his six existing schemes of library classification?
2. Give examples of rigidly faceted classifications.

2.3 COLON CLASSIFICATION

Colon Classification is a system or method of library classification that was developed by an Indian librarian named S R Ranganathan in the year 1933. Although it was first published in the year 1933, it has had six more editions published since then. This was perhaps the watershed of library classification, because it had been the first time that someone had attempted to use the analytico systemic approach for the purpose of library classification. Due to the analytico systemic approach, this system of library classification has also been referred to by Ranganathan himself as the first faceted library classification ever attempted.

Perhaps because S R Ranganathan was an Indian, the Colon Classification system of library classification is primarily used in libraries across India. The term colon classification has obviously been derived from to the use of colons to separate class, or category, numbers facets.

However, perhaps inspired by Ranganathan and his colon classification scheme, many other library classification schemes have also been known to adopt various other punctuation marks to identify distinctive facets However, it is important to understand that the use of the other punctuation marks by the other libraries is completely unrelated to the original system of colon classification, and the other punctuation marks are used for various other functions and not actually classification.

Within the domain of colon classification, the term facet is used to define the

personality which is the most commonly used subject, time, space, matter, and energy.

It is important to remember that these are facets that are most commonly used to define every item in any library at any given time, which is perhaps why the colon classification system is so easy to adopt. This also explains the reason libraries across India choose to adopt the colon classification scheme.

For instance, during the first half of the 20th century, a great deal of research was being conducted in India in the area of tuberculosis because the medical profession had then been working hard to find a cure for this life endangering disease. So obviously, medical research in the field of tuberculosis was then and continues to be a very common subject of study and reference. Here is the colon number of this resource material.

"research in the cure of tuberculosis of lungs by x-ray conducted in India in 1950" would be categorized as:

Medicine,Lungs;Tuberculosis:Treatment;X-ray:Research.India'1950

This is summarized in a specific call number:

L,45;421:6;253:f.44'N5

Organization: The system of colon classification uses about 42 main classes, that are then combined with various letters of the alphabet and numbers or numerals, in a manner that resembles the Library of Congress Classification to sort or derive a publication.

Facets: The colon classification uses five primary facets or categories that are then further sorted or categorized before publication. Collectively, these facet or categories are known as PMEST. As mentioned earlier, the colon classification system is perhaps the only library classification or categorization system that is known to use facets for such classification. PMEST represents Personality, Matter, Energy, Space and Time.

Classes: S R Ranganathan has developed the following classes or categories as listed below. These are then further divided into subtopics using the PMEST system:

- z Generalia
 - 1 Universe of Knowledge
 - 2 Library Science
 - 3 Book science
 - 4 Journalism
- A Natural science
- B Mathematics
 - B2 Algebra

NOTES

NOTES

- C Physics
- D Engineering
- E Chemistry
- F Technology
- G Biology
- H Geology
- HX Mining
 - I Botany
 - J Agriculture
- J1 Horticulture
- J2 Feed
- J3 Food
- J4 Stimulant
- J5 Oil
- J6 Drug
- J7 Fabric
- J8 Dye
- K Zoology
- KZ Animal Husbandry
 - L Medicine
- LZ3 Pharmacology
- LZ5 Pharmacopoeia
- M Useful arts
 - M7 Textiles *[material]:[work]*
 - Ä Spiritual experience and mysticism *[religion],[entity]:[problem]*
- N Fine arts
- ND Sculpture
- NN Engraving
- NQ Painting
- NR Music
- O Literature
- P Linguistics
- Q Religion

R Philosophy
 S Psychology
 T Education
 U Geography
 V History
 W Political science
 X Economics
 Y Sociology
 YZ Social Work
 Z Law

NOTES

The Colon Classification system of library classification was developed by S R Ranganathan between the years 1922 and 1928 in his research project at the Madras University, and so was first adopted at the Madras University Library after it was first published in the year 1933. It was also adopted at the Madras Library which was founded by S R Ranganathan himself in the year 1928. The Colon Classification has been subsequently published many times. The last time it got published was after his death in the year 1987. This edition published in the year 1987 is known to be the seventh publication or edition of the Colon Classification. The fact that S R Ranganathan had been a mathematician is perhaps reflected in his method of classification. The Colon Classification is said to have been greatly impacted by mathematics, and critics have discerned similarities between mathematical equations and the method of classification developed by S R Ranganathan.

2.4 DEWEY DECIMAL CLASSIFICATION

The Dewey Decimal Classification was first developed by Melvil Dewey in the United States of America in the year 1876. When it had first been published, it had less than one thousand classes, and had been compiled in a short four page pamphlet. Since the year 1876, the Dewey Decimal Classification has been adapted and published many times over, the latest edition being reported in the year 2011. Currently, the Dewey Decimal Classification has been compiled into as many as four volumes. The Dewey Decimal Classification is also known as the Dewey Decimal System. The DDC may be described also as a proprietary library classification system.

The Dewey Decimal Classification System was developed by Melvil Dewey in the year 1876 and published in the same year. It should be noted that it was developed as a knowledge classification system, which requires it obviously to be revised constantly in order to keep pace with the growth of knowledge. The Dewey

NOTES

Decimal Classification System is fully owned by the Online Computer Library Center Inc, which has so far published 23 editions, as mentioned earlier the last edited version being published in the year 2011. Obviously therefore all the four volumes in the complete set are owned by the Online Computer Library Center Inc.

Libraries across the world can have paid access to the Dewey Decimal Classification System. This is perhaps the reason or rather one of the reasons why it is so widely and popularly used across the world. Currently, libraries in across 140 countries are using the Dewey Decimal Classification System to organize, categorize and manage their collections.

As a matter of fact, the Dewey Decimal Classification numbers are currently featured in the national bibliographies of more than 60 countries across the world. Libraries across the world access and apply the Dewey Decimal Classification numbers almost on a daily basis, through a range of access modes including the OCLC Online Union Catalog. This catalogue is also referred to as the WorldCat. The Dewey Decimal Classification numbers are used for many other purposes, including as browsing identification to search for online resources or websites and research material.

The Dewey Decimal Classification has been translated into more than thirty languages. Translations of the latest edition published in the year 2011 have either been completed, or are underway in such languages as Vietnamese, German, French, Arabic, Spanish and Swedish.

Development: Perhaps one of the greatest strengths of the Dewey Decimal Classification system is that the system is developed, updated and managed in a national bibliographic agency, namely the Library of Congress. The Dewey Decimal Classification system editorial wing is situated within the Decimal Classification Division of the Library of Congress. This is the space within which decimal classification specialists allot literally tens of thousands of numbers to new collections or arrivals of books that have been catalogued by the Library of Congress over the past one year.

How does this benefit the editing of the classification numbers? The fact that the editorial office is situated within the CIP and the Dewey section enables the editorial team to detect or identify new trends in the literature that they believe need to be included or incorporated into the Dewey classification list, the Dewey editorial team sets to work. This is how it goes about it:

1. The Dewey editorial team first prepares a Proposed Schedule of Revisions and Expansions
2. The editorial team then forwards their list of proposals to the next higher agency, the Decimal Classification Editorial Policy Committee or the EPC for further action

3. The Decimal Classification Editorial Policy Committee – the EPC – now reviews the proposals put forth by the Swqwy editorial team and recommends an appropriate course of action

The Decimal Classification Editorial Policy Committee: The EPC or the Decimal Classification Editorial Policy Committee is an international board comprising of ten members from across the world. The primary function of the Editorial Policy Committee is to advise both the Online Computer Library Center Inc, or the OCLC, and the Dewey Decimal Classification editorial team on all matters pertaining to proposed or intended change, expansion as well as overall, general maintenance of the Online Computer Library Center Inc or the OCLC.

The Dewey Decimal Classification Editorial Policy Committee is required to represent the interests of all their users from across the world. In keeping with this requirement, the members of the Decimal Classification Editorial Policy Committee come from library schools, academic schools, national libraries, special libraries, central libraries, as well as privately owned libraries.

Editions: The Dewey Decimal Classification numbers are always published in both the print as well as the web or online versions. The online or web versions are also accessible to all their users. The Dewey Decimal Classification numbers are published in both the full as well as the abridged versions. While the print versions may take time to be reprinted for obvious reasons, the online or the electronic editions re updated as frequently as possible. So this means the online or the web versions have indices and titles in addition to the latest print version The online or electronic version also has mapped vocabulary, as is described in the passage below:

Notations and Structure: It will be obvious from the passage quoted below that the Dewey Decimal Classification system is professionally managed and had first been developed on a set of solid principles. The primary intention that initiated and prompted the first version or edition of the Dewey Decimal Classification system appears to have been solely for the purpose of encouraging people, especially students, academicians and professionals, to build their knowledge on a vase spectrum of subjects and topics. Towards that end, the Dewey Decimal Classification system is intended to provide a structured and logical tool that comprises notations that every person across the world can understand. This means the notations are devised or assigned in Arabic numerals that are easily identifiable and recognizable by every person across the world. The purpose is to provide information or general knowledge enrichers in approximately 20,000 titles or maybe less.

Having said this, the primary intention of the Dewey Decimal Classification appears to have been to reach out to as many people or users across the world as possible. This is obvious from the fact that although the main editions are in Arabic

NOTES

NOTES

numerals, all editions are also routinely translated into approximately thirty international languages.

The DDC has been developed around strong principles that help develop it into an ideal model of a general knowledge organization tool: in other words, meaningful notation in universally recognized Arabic numerals, well-defined categories, fully formed hierarchies, as well as an enriched network of relationships between the topics. In the Dewey Decimal Classification, the basic or primary classes are organized by the domains or fields of study.

At the broadest level, the Dewey Decimal Classification has been divided into ten main types, all of which combined cover the entire spectrum of knowledge. Each main type has been again divided into ten sections, and each section into ten sections (but all the numbers for these sections and sub sections have not been used).

Arabic numerals are used to represent each class in the DDC. A decimal point follows the third digit in a class number, after which division by ten continues to the specific degree of classification needed.

A subject may appear in more than one discipline. For example, "clothing" has aspects that fall under several disciplines. The psychological influence of clothing belongs in 155.95 as part of the discipline of psychology; customs associated with clothing belong in 391 as part of the discipline of customs; and clothing in the sense of fashion design belongs in 746.92 as part of the discipline of the arts.

Heirarchy: The important or main subjects appear at the top while subtopics appear below them. The main or important subjects are given complete numerical notations while their subtopics are usually given numerical values or notations that are smaller in value.

The passages quoted below describe the hierarchy of the notations, and go on to describe how the information has been categorized and divided into the four volumes that form the latest edition which was published in the year 2011.

Hierarchy in the Dewey Decimal Classification has been defined with the help of a structure and notation. Structural hierarchy indicates that all the topics (apart from the ten primary types) have been incorporated in all the broader topics that appear before them. Any note pertaining to the nature of a type is valid for all the subordinate types, and will include the logically subordinate topics that have been classified at coordinate numbers.

Notational hierarchy has been defined by the length of notation. Numbers at any specific level are normally subordinate to a type the notation of which has been even a single digit shorter; coordinate with a type whose notation has the same number of significant digits; and superordinate to a type that has numbers that are longer even by a single digit. The underlined digits in the following example are indicative of this notational hierarchy:

- 600 Technology
- 630 Agriculture and related technologies
- 636 Animal husbandry
- 636.7 Dogs
- 636.8 Cats

NOTES

"Dogs" and "Cats" are more specific than (i.e., are subordinate to) "Animal husbandry"; they are equally specific as (i.e., are coordinate with) each other; and "Animal husbandry" is less specific than (i.e., is superordinate to) "Dogs" and "Cats."

Sometimes, other devices must be used to express the hierarchy when it is not possible or desirable to do so through the notation. Special headings, notes, and entries indicate relationships among topics that violate notational hierarchy.

Number Building: Perhaps the most important part of the entire Dewey Decimal Classification system is the method that the Online Computer Library Center Inc and the classification specialists follow in assigning numerals to each entity. This is obviously a painstaking and laborious task requiring tremendous team work and concentration. This is how the Online Computer Library Center Inc website describes the entire process:

Just a small percentage of the DDC numbers are listed in the schedules, thus making it essential to develop or create numbers that are not listed in the actual schedules. This means numbers that have been so developed or constructed enable greater depth in the analysis of content. There are usually four sources of notation to construct numbers. These include (A) Table 1 Standard Subdivisions; (B) Tables 2–6; (C) other parts of the schedules; and (D) add tables in the schedules.

Number building is usually started only when there are instructions in the schedules (except for the addition of standard subdivisions, which could occur at any place unless there is an instruction to the contrary). Construction of numbers begins with a base number (always stated in the instruction note) to which another number is added.

Check Your Progress

3. Name the first faceted library classification as per Ranganathan.
4. From where has the term colon classification been derived?
5. Who developed the Dewey Decimal Classification?

2.5 UNIVERSAL DECIMAL CLASSIFICATION (USC)

The Universal Decimal Classification system or the UC is analytical, systematic and faceted system of classification of knowledge across a wide range of subjects

NOTES

and its organization, storage and retrieval within those large collaborative and collective spaces.

The UDC Consortium is a non-profit association of publishers from across the world. The UDC Consortium is an international body, and perhaps for that reason has its headquarters in the Hague, the Netherlands.

Every other library classification systems across the world may have been conceptualized and initiated at their respective national levels, perhaps as bibliographical and library classification systems before progressing to the international platform, however, the Universal Decimal Classification system had been conceived and initiated at the International platform at the turn of the 20th century. From that moment on, the Universal Decimal Classification system has been organized and maintained as an international body that allows publishers from across the world to ensure that their printed and published works are included in the Universal Decimal Classification list and catalog. The Universal Decimal Classification catalog has been translated into numerous international languages from the year of its inception and initiation. Currently, the Universal Decimal Classification catalog is being translated into more than forty international languages.

The UDC Summary is available in over fifty international languages.

The Universal Decimal Classification catalog pertains to all areas of human knowledge and achievement. Being primarily based on the acquisition, organization and management of human knowledge across the world. The Universal Decimal Classification catalog is consistently and constantly reviewed and updated in an effort to keep pace with and match new global developments in every aspect and area of global knowledge building and development.

It is important to note that the Universal Decimal Classification had been primarily conceived as a knowledge indexing system that could be used in libraries across the world. However, perhaps due to its far-reaching strengths, the Universal Decimal Classification system is now used not just for indexing but also for management and retrieval of items listed in their catalogs. Libraries across the world today use the Universal Decimal Classification system for indexing their knowledge content, for managing and organizing their shelf content, or in most cases both of these.

Once assigned, the Universal Decimal Classification codes or notations can describe or define any type of subject or item, at any desire or required level of depth or detail.

The Universal Decimal Classification of indexing and cataloging can be adapted and used to catalog any thing, including text documents, audio-video clips, audio clips, videos, film clips, illustrations, maps, photographs of relics relating to museum curating, actual relics or antiques from museums and archeological

sites, art work both contemporary as well as from historical collections, and so much more.

The Universal Decimal Classification system had been designed, conceived and developed by a pair of Belgian bibliographers named Paul Ottet and Henri La Fontaine towards the end of the 19th century. It was in the year 1895 that Paul Ottet and Henri La Fontaine conceived and designed the first Universal Bibliographical Repertory, or in Belgian the Repertoire Bibliographique Universel, also known as the RBU. The Universal Bibliographical Repertory or the UBR was conceptualized and developed to become a catalog of all the ‘comprehensive and classified index’ of all.

It was around this time that a young American zoologist named Herbert Haviland Field was in Zurich, attempting to set up his own bibliographic business which he intended to name Concilium Bibliographium. Interestingly, it was Herbert Haviland Field who met Paul Ottet and Henri La Fontaine and suggested that they conceive a card indexing system for their new venture. Paul Ottet, it appears, had heard of Melvil Dewey and his Dewey Decimal Classification system. Paul Ottet wrote to Melvil Dewey to seek his permission to translate the Dewey Decimal Classification index into the French language and use that to list their own entities. However, somewhere along the way this plan appears to have fall through, because it must be remembered that the Dewey Decimal Classification system uses an entirely numerical indexing format.

Paul Ottet, Henri La Fontaine and Herbert Haviland Field decided to make remarkable innovations in the original indexing format devised by Melvil Dewey. It was decided that the Universal Bibliographical Repertory would include syllables, or letters of the alphabet in combination with the numerals, allowing the new system to be always operational and never be allowed to exhaust itself.

It was during their experiments and research that plenty of connections and inter-links and inter-relations were recognized and identified between different subjects. So symbols were appropriately created and included to represent all of these.

The first printed list of indexes named the Manuel du Repertoire Bibliographique Universel, published at the turn of the century in the year 1905, Paul Ottet and Henri La Fontaine had included plenty of revolutionary and innovative features. These included in the context of knowledge classification, the following features:

1. tables of generally applicable (aspect-free) concepts—called common auxiliary tables;
2. a series of special auxiliary tables with specific but re-usable attributes in a specific field of knowledge;
3. an expressive notational system with connecting symbols

NOTES

4. syntax rules to enable coordination of subjects and finally,
5. the creation of a documentation language proper.

NOTES

From the year 1905 which was the year it had been published for the first time, the Universal Bibliographic Repertory has appeared to have expanded and grown in leaps and bounds. In the years before the outbreak of the World War II in fact, the Universal Bibliographic Repertory had recorded as many as 11 million entries. The entire catalog and its complete content can still be viewed at the Mundaunum in Mons in Belgium as it has been first organized and indexed. In the year 2013, the Universal Decimal Classification had been recommended for the UNESCO Memory of the World Register.

The Application of the Universal Decimal Classification index: The Universal Decimal Classification index is currently used across more than 150 countries, in more than 130,000 libraries of all types, including special libraries, public libraries, academic institutional libraries, public schools and colleges. The Universal Decimal Classification index is also used in the indexing of bibliographies at the national level in many of these countries as well. Some of the larger databases that have been using the Universal Decimal Classification index over the years include the following:

1. *NEBIS (The Network of Libraries and Information Centers in Switzerland)* – 2.6 million records
2. COBIB.SI (Slovenian National Union Catalogue) – 3.5 million records
3. Hungarian National Union Catalogue (MOKKA) – 2.9 million records
4. VINITI RAS database (All-Russian Scientific and Technical Information Institute of
5. Russian Academy of Science) with 28 million records
6. Meteorological & Geostrophysical Abstracts (MGA) with 600 journal titles
7. PORBASE (Portuguese National Bibliography) with 1.5 million records

In the years preceding the inventions and growth of the electronic technology, it may be remembered that the Universal Decimal Classification indexing system had traditionally been used to index inventions and discoveries in the different areas and fields of science. This had probably begun with the indexing of zoology through the partnership with Herbert Haviland Field. Compilations of scientific research material demonstrating decades of research in various areas of scientific innovations and experiments bear witness to the use of the Universal Decimal Classification indexing codes in countries across the world. Some journals at the international level that bear the Universal Decimal Classification codes have been listed here as examples.

- UDC code 663.12:57.06 in the article "Yeast Systematics: from Phenotype to Genotype" in the journal *Food Technology and Biotechnology* (ISSN 1330-9862)^[20]
- UDC code 37.037:796.56, provided in the article "The game method as means of interface of technical-tactical and psychological preparation in sports orienteering" in the Russian journal "Pedagogico-psychological and medico-biological problems of the physical culture and sport" (ISSN 2070-4798).^[21]
- UDC code 663.12:57.06 in the article "Yeast Systematics: from Phenotype to Genotype" in the journal *Food Technology and Biotechnology* (ISSN 1330-9862)^[20]
- UDC code 37.037:796.56, provided in the article "The game method as means of interface of technical-tactical and psychological preparation in sports orienteering" in the Russian journal "Pedagogico-psychological and medico-biological problems of the physical culture and sport" (ISSN 2070-4798).^[21]
- UDC code 621.715:621.924:539.3 in the article Residual Stress in Shot-Peened Sheets of AlMg4.5Mn Alloy – in the journal *Materials and technology* (ISSN 1580-2949).^[22]

The Universal Decimal Classification indexing system appears to be flexible, and seems to have been designed for use with multiple types of machinery. For instance the Universal Decimal Classification index can be used with modern electronic devices such as websites and webpages, while it has also been used successfully with the earlier mechanical sorting and classifying machinery that used to be in existence in libraries across the world before the invention of the computer and the electronic or electric typewriters.

The structure of Universal Decimal Classification:

Let us now examine some basic points that make up the structure of Universal Decimal Classification.

Notation:

A notation is a code identification that is be used to represent a subject and from thereon its various subtopics in their proper hierarchy. It is important to understand that in bibliographies as well as in libraries where databases need to be maintained to facilitate such basic and important tasks as storing, sorting and filing of the information pertaining to relevant subjects, subtopics or any other relevant information, it is important to allot or assign code numbers or index numbers to identify each item that is being stored or sorted. Assigning or allotting such code or index numbers within the large scheme of things or within the larger database makes it easier for librarians or users to identify specific subjects, subtopics or

NOTES

NOTES

object and be able to retrieve such required information from a vast database without confusion and without wasting much time. Assigning or allotting such index or code numbers to each main class of information, and from thereon to each category or sub-topic ensures that users, within each user space especially in academia or any other profession, are in a position to categorize each class of subject matter.

This makes it easier for school or college students, for instance, to identify and retrieve information on specific plants or flowers within the larger stream of science such as botany. Students at the school or college level who are studying zoology will be able to identify and retrieve information pertaining to specific animal species within a few seconds.

This implies that each main class of information may be assigned or allotted a main number or code or index notation which may have different symbols or syllables added to them to denote specific sub-categories or sub-species or sub-topics. In order to make it easier for users of the Universal Decimal Classification indexing scheme, a specific punctuation mark is usually used after every third digit or syllable so users find it ready to read such index or code numbers. Obviously, if only numerals or syllables of the alphabet were to be used without such punctuation marks, it would be very difficult for users to read such index or code numbers without some level of confusion or strain on the eyes.

Let us look at some examples of such notations that are given below so that it becomes clearer to understand why such punctuation marks are so essential.

Caption (Class description)

539.120	Theoretical problems of elementary particles physics. Theories and models of fundamental interactions
539.120.2	Symmetries of quantum physics
539.120.22	Conservation laws
539.120.222	Translations. Rotations
539.120.224	Reflection in time and space
539.120.226	Space-time symmetries
539.120.23	Internal symmetries
539.120.3	Currents
539.120.4	Unified field theories
539.120.5	Strings

In UDC the notation has two features that make the scheme easier to browse and work with:

- hierarchically expressed – the specificity of the class always matches or is proportionate to the length of the notation. This means when the final digit is removed it causes a broader class code.
- combined, the sequence of digits is interrupted by a precise type of punctuation sign which indicates that the expression is a combination of classes rather than a simple class e.g. the colon in 34:32 indicates that there are two distinct notational elements: 34 Law. Jurisprudence and 32 Politics; the closing and opening parentheses and double quotes in the following code 913(574.22)"19"(084.3) indicate four separate notational elements: 913 Regional geography, (574.22) North Kazakhstan (Soltüstik Qazaqstan); "19" 20th century and (084.3) Maps (document form)

NOTES

Universal Decimal Classification indexing and coding system allows users to use an infinite patterns of combinations and permutations between attributes and relationships between subjects that need to be expressed or used. We are giving various examples in this unit to help students understand these things more easily. Universal Decimal Classification coding system allows users to use combinations from different subject tables to help them represent or express various aspects of document content and form. For instance, *94(410)"19"(075) History (main subject) of United Kingdom (place) in 20th century (time), a textbook (document form). Or: 37:2 Relationship between Education and Religion.*

It is also possible to parse complicated Universal Decimal Classifications into constituent elements. Universal Decimal Classification system is also known as a disciplinary classification system because it can be used to cover the entire universe of knowledge. In other words, one can describe this type of classification as the aspect of perspective. Effectively this means that concepts can be subsumed and placed under the subject field in which that are being studied. This allows the same concept to reappear in different subject tables under which they are currently being studied. Usually this specific feature can be implemented in Universal Decimal Classification by re-using the same concept in different combinations with the one main subject.

A language code in common auxiliaries of language is used to obtain numbers intended for ethnic grouping, individual languages in linguistics and individual literatures. Or maybe, a code from the auxiliaries of place, such as (410) United Kingdom, uniquely representing the concept of United Kingdom can be used to express 911(410) Regional geography of United Kingdom and 94(410) History of United Kingdom.

Organization of classes in Inversal Decimal Classification system:

- It may be noted that Class 4 had earlier been assigned to the subject of linguistics. During the 1960s, the subject of linguistics had been moved

NOTES

to Class 8 in order to make space available for rapidly developing and expanding knowledge and discoveries, especially in the areas of natural sciences and technology. So here we go.

- Common auxiliary tables (such as specific auxiliary signs). Such tables comprise of facets of concepts that represent, general repetitive characteristics that are applicable across a range of subjects spanning the main tables, including such as place, language of the text and physical form of the document, which could take place in practically any subject. UDC numbers from these tables, known as common auxiliaries are just added at the end of the number for the subject taken from the main tables. There are over 15,000 of common auxiliaries in UDC.
- The main tables or main schedules comprising of the different domains as well as branches of knowledge, categorized into nine main classes, numbered between 0 and 9 (with class 4 being vacant). At the beginning of each class there are also series of special auxiliaries, which express aspects that are recurrent within this specific class. Main tables in UDC contain more than 60,000 subdivisions.

Main classes

- *0 Science and Knowledge. Organization. Computer Science. Information Science. Documentation. Librarianship. Institutions. Publications*
- *1 Philosophy. Psychology*
- *2 Religion. Theology*
- *3 Social Sciences*
- *4 vacant*
- *5 Mathematics. Natural Sciences*
- *6 Applied Sciences. Medicine, Technology*
- *7 The Arts. Entertainment. Sport*
- *8 Linguistics. Literature*
- *9 Geography. History*

Some interesting aspects of how coding is implemented in the Universal Decimal Classification system may be gleaned from the nature of subject. For instance, Applied Science, Medicine and Technology are separate subjects by themselves, right? But within the medical profession, it is routinely possible to find instances where all of these three subjects can be found in the same space. For instance, Natural Science is used in the field of research and pharmaceuticals or pharmacology. This would naturally lead one to the field of medicine, or hospitals and health care. Technology is always being innovated and implemented in health

care, that is hospitals and medical research itself. Especially in areas such as X-rays, neuro surgery, neuro physics, cardiology and almost every other area of health care. So Universal Decimal Classification allows the combination of codes and indexes from all three spaces, that is natural sciences, medicine as well as technology to create a code for resulting information tables.

Let's look at some explanations of these main classes.

Class 6 may be described as occupying the largest space within the Universal Decimal Classification schedules. It provides more than 44000 sub-divisions. This is because every industry or business usually will have one main auxiliary table of concepts or content, under which there will be separate sections assigned to operations, processes, materials used and the end products or services. Universal Decimal Classification allows separate codes to be used for each of these sub divisions.

Class 8 in the Universal Decimal Classification systems has been assigned for the main auxiliary table of Languages. Under this main auxiliary table, or within it, one can identify linguistics, group linguistics, individual linguistics, ethnic linguistics, medieval linguistics, medieval languages, pre medieval linguistics, pre medieval languages, modern languages, modern linguistics, modern group languages, modern individual linguistics – the list appears to be almost infinite.

2.6 LIBRARY OF CONGRESS CLASSIFICATION

The Library of Congress Classification is a system of classification that had been developed by the Library of Congress. It is a system of classification that is primarily set by most academic and research libraries, not just within the United States of America, but also in many other countries across the world.

The Library of Congress Classification or the LCC is separate from the LCCN which is the Library of Congress Control Numbers which are coding numbers assigned by the Library of Congress to their books, their authors, and the urls of their online entries. It is important for audiences and users to understand the difference between the two entities.

The Library of Congress or the LCC has often been criticized a great deal because it lacks a solid theoretical foundation. The entire classification system appears to have been prompted, and shall we say founded, by the Library of Congress to meet its own internal requirements. Many of its classifications appear to have been prompted as an immediate answer to an immediate coding requirement. So these classifications are not something that can relate to or meet the needs of libraries anywhere else, because they are not based on entomological classifications. This limitation or drawback sets it starkly apart from the other classification systems that have been explored in this unit, specifically the Colon

NOTES

Classification system, the Dewey Decimal Classification coding system and the Universal Decimal Classification coding or indexing system.

NOTES

In spite of the fact that the Library of Congress Classification system does attempt to divide main subjects into sub categories, it remains essentially enumerative by nature and perhaps by design. This means that the Library of Congress Classification system appears to provide a numerical or coded guided to the collections of books that are contained within the one library located within its own space, rather than provide a larger or more universal coding system that can be adopted or replicated by any library across the world.

As a matter of fact, in the year 2007, a report published in The Wall Street Journal reported that in most of the countries the newspaper routinely surveyed, most of the public libraries, as well as the smaller academic libraries, the oldest global classification system, the Dewey Decimal Classification coding system, was being implemented to classify and code their collections of books and other resource material.

Some national level libraries in the United States of America appear to be still using the original letters, such as W, WS-WZ that had been introduced by the Library of Congress. However, significantly the Library of Congress Classification system itself no longer uses these letters in its internal classifications. In that context, it may be said that most of the original classifications that had been introduced by the Library of Congress are now redundant and extinct, but sadly it would seem are still being used by other national level libraries within the United States of America, something that no one appears to be bothered about either within the Library of Congress itself or within the Federal government.

It may be assumed that the original classifications that had been introduced by the Library of Congress to categorize its collections of books and their authors have never since then been updated or revised at any point in time or to meet advances in knowledge or developments in technology. One is left to wonder in fact whether the collections of books in their library have ever been replaced, edited or exchanged.

We are quoting some of the classes that were first introduced by the Library of Congress Classification system at its inception.

Class D – World History and History of Europe, Asia, Africa, Australia, New Zealand, etc.

- Subclass D – History (General)
- Subclass DA – Great Britain
- Subclass DAW – Central Europe
- Subclass DB – Austria – Liechtenstein – Hungary – Czechoslovakia
- Subclass DC – France – Andorra – Monaco

- Subclass DD – Germany
- Subclass DE – Greco-Roman World
- Subclass DF – Greece
- Subclass DG – Italy – Malta
- Subclass DH – Low Countries – Benelux Countries
- Subclass DJ – Netherlands (Holland)
- Subclass DJK – Eastern Europe (General)
- Subclass DK – Russia. Soviet Union. Former Soviet Republics – Poland
- Subclass DL – Northern Europe. Scandinavia
- Subclass DP – Spain – Portugal
- Subclass DQ – Switzerland
- Subclass DR – Balkan Peninsula
- Subclass DS – Asia
- Subclass DT – Africa
- Subclass DU – Oceania (South Seas)
- Subclass DX – Romanies

Class E – History of the Americas

- Class E does not have any subclasses.

Class F – Local History of the Americas

- Class F does not have any subclasses, however Canadian Universities and the Canadian National Library use FC for Canadian History, a subclass that the LC has not officially adopted, but which it has agreed not to use for anything else^{[7][8]}

Class G – Geography, Anthropology, Recreation

- Subclass G – Geography (General). Atlases. Maps
- Subclass GA – Mathematical geography. Cartography
- Subclass GB – Physical geography
- Subclass GC – Oceanography
- Subclass GE – Environmental Sciences
- Subclass GF – Human ecology. Anthropogeography
- Subclass GN – Anthropology
- Subclass GR – Folklore
- Subclass GT – Manners and customs (General)
- Subclass GV – Recreation. Leisure

NOTES

NOTES

Class J – Political Science

- Subclass J – General legislative and executive papers
- Subclass JA – Political science (General)
- Subclass JC – Political theory
- Subclass JF – Political institutions and public administration
- Subclass JJ – Political institutions and public administration (North America)
- Subclass JK – Political institutions and public administration (United States)
- Subclass JL – Political institutions and public administration (Canada, Latin America, etc.)
- Subclass JN – Political institutions and public administration (Europe)
- Subclass JQ – Political institutions and public administration (Asia, Africa, Australia, Pacific Area, etc.)
- Subclass JS – Local government. Municipal government
- Subclass JV – Colonies and colonization. Emigration and immigration. International migration
- Subclass JX – International law, see JZ and KZ (obsolete)
- Subclass JZ – International relations

Class L – Education

- Subclass L – Education (General)
- Subclass LA – History of education
- Subclass LB – Theory and practice of education
- Subclass LC – Special aspects of education
- Subclass LD – Individual institutions – United States
- Subclass LE – Individual institutions – America (except United States)
- Subclass LF – Individual institutions – Europe
- Subclass LG – Individual institutions – Asia, Africa, Indian Ocean islands, Australia, New Zealand, Pacific islands
- Subclass LH – College and school magazines and papers
- Subclass LJ – Student fraternities and societies, United States
- Subclass LT – Textbooks

A closer examination of these classes would give audiences the impression that none of these classes had been implemented based on any relevance to the main subject of the books in the collection. The main classes appear to have been randomly assigned, to enable the librarians use some sort of categorization. It is

difficult to fathom how no further updation has ever taken place, and it would also appear that technology is not used to maintain these records, which would make it simpler to update.

Interestingly the Dewey Decimal Classification system has its editorial office within the Library of Congress office, and the Dewey Decimal Classification codes are updated and republished at regular intervals, besides being translated into more than thirty international languages. This is under the jurisdiction of the OCSC Inc, as has been explained under the Dewey Decimal Classification system earlier in this unit. One is forced to wonder how the OCC Inc has not been updating the Library of Congress Classification system at any point in time.

NOTES

Check Your Progress

6. Which library classification was developed first at international level and then incorporated at the national?
7. How is the Library of Congress Classification system different from other library classifications?

2.7 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Ranganathan first mentioned the six existing systems of library classification in his book *Prolegomena to Library Classification*.
2. The examples of rigidly faceted classification include the first three editions (Le, of 1931 1939 and 1950 respectively) from the Colon Classification (CC) because they developed an individual facet formula for each basic type.
3. Colon classification is identified as the first faceted library classification.
4. The term colon classification has obviously been derived from the use of colons to separate class, or category, numbers facets.
5. The Dewey Decimal Classification was first developed by Melvil Dewey in the United States of America in the year 1876.
6. The Universal Decimal Classification was developed first at international level and then incorporated at the national.
7. The Library of Congress Classification system is different from other library classifications because the system appears to provide a numerical or coded guided to the collections of books that are contained within the one library located within its own space, rather than provide a larger or more universal coding system that can be adopted or replicated by any library across the world.

2.8 SUMMARY

NOTES

- According to Ranganathan, as he mentions in his *Prolegomena to Library Classification*, there are about six existing schemes or systems of library classification, besides some other special systems of library classification. These systems include the following:
 - o The rigidly faceted classification scheme, or the library classification with the pre-determined facets
 - o The almost faceted scheme of classification
 - o The freely faceted classification, which is an analytical-systemic system of classification which is prompted by possibilities and principles
 - o The enumerative classification
 - o The almost enumerative classification
- The rigidly faceted classification and the freely faceted classification schemes have been further categorized under the main group titled the fully faceted classification.
- Colon Classification is a system or method of library classification that was developed by an Indian librarian named S R Ranganathan in the year 1933. Although it was first published in the year 1933, it has had six more editions published since then. This was perhaps the watershed of library classification, because it had been the first time that someone had attempted to use the analytical-systemic approach for the purpose of library classification.
- The system of colon classification uses about 42 main classes, that are then combined with various letters of the alphabet and numbers or numerals, in a manner that resembles the Library of Congress Classification to sort or derive a publication.
- The Dewey Decimal Classification was first developed by Melvil Dewey in the United States of America in the year 1876. When it had first been published, it had less than one thousand classes, and had been compiled in a short four page pamphlet.
- The DDC was developed as a knowledge classification system, which requires it obviously to be revised constantly in order to keep pace with the growth of knowledge. The Dewey Decimal Classification System is fully owned by the Online Computer Library Center Inc, which has so far published 23 editions, as mentioned earlier the last edited version being published in the year 2011. Obviously therefore all the four volumes in the complete set are owned by the Online Computer Library Center Inc.

- The Universal Decimal Classification system or the UC is analytical, systematic and faceted system of classification of knowledge across a wide range of subjects and its organization, storage and retrieval within those large collaborative and collective spaces.
- The Universal Decimal Classification system had been designed, conceived and developed by a pair of Belgian bibliographers named Paul Ottet and Henri La Fontaine towards the end of the 19th century.
- The Universal Decimal Classification indexing system appears to be flexible, and seems to have been designed for use with multiple types of machinery.
- The Library of Congress Classification is a system of classification that had been developed by the Library of Congress. It is a system of classification that is primarily set by most academic and research libraries, not just within the United States of America, but also in many other countries across the world.
- The Library of Congress or the LCC has often been criticized a great deal because it lacks a solid theoretical foundation. The entire classification system appears to have been prompted, and shall we say founded, by the Library of Congress to meet its own internal requirements.

NOTES

2.9 KEY WORDS

- Faceted classification: It is made up of timelines of basic types; special isolates and common isolates primarily.
- **Facet:** It is a generic term used to denote any component
- **Notation:** It is a code identification that is be used to represent a subject and from thereon its various subtopics in their proper hierarchy.

2.10 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. List the six existing schemes or systems of library classification as mentioned by Ranganathan.
2. Explain the sub-types of rigidly faceted classification.
3. Write a short note on the development of the Colon Classification.
4. What is a UDC Summary?
5. List some of the larger databases that have been using the Universal Decimal Classification index over the years.
6. What are the general criticism against the Library of Congress Classification?

Long-Answer Questions

NOTES

1. Describe the Colon Classification.
2. Explain the notation and structure of the Dewey Decimal Classification.
3. Compare the history and development of Universal Decimal Classification in relation to the Dewey Decimal Classification.
4. Examine the structure of Universal Decimal Classification.
5. Discuss the major classes of Library of Congress Classification.

2.11 FURTHER READINGS

- Alberico, R. and Micco M. 1990. *Expert systems for reference and Information retrieval*. West Port: Meckler.
- Atchison, J. & Gilchrist, A. 1972. *Thesaurus construction: a practical manual*. London: Aslib.
- Chowdhry, G.G. 2003. *Introduction to modern Information retrieval*. 2nd Ed. London: Facet Publishing.
- Cleaveland, D. B. 2001. *Introduction to Indexing and abstracting*. 3rd Ed. Englewood, Colo: Libraries Unlimited.
- Ghosh, S.B. and Biswas, S.C. 1998. *Subject Indexing systems: Concepts, methods and techniques*. Rev. ed. Calcutta: IASLIC.
- Pandey, S.K. Ed. 2000. *Library Information retrieval*. New Delhi: Anmol.

BLOCK - II
INDEXING TECHNIQUES

*Indexing Languages,
Vocabulary Control and
Thesaurus*

**UNIT 3 INDEXING LANGUAGES,
VOCABULARY CONTROL
AND THESAURUS**

NOTES

Structure

- 3.0 Introduction
- 3.1 Objectives
- 3.2 Indexing Languages
- 3.3 Vocabulary Control
- 3.4 Thesaurus
- 3.5 Answers to Check Your Progress
- 3.6 Summary
- 3.7 Key Words
- 3.8 Self-Assessment Questions and Exercises
- 3.9 Further Readings

3.0 INTRODUCTION

Even before the students begin to understand the concept of designing, creating and using the appropriate indexing languages, the students need to understand the actual concept of what indexing is all about. They need to understand that indexing is a two dimensional concept. This means the basic term indexing can be used in two ways – the broad sense as well as the narrow sense.

1. First, looking at it from a broad perspective, the term indexing may be taken as a generic or a general term, which pertains to the entire processes of creating and using verbal representations that are able to take place in both the physical as well as the subject descriptions in a wide range of professional spaces.
2. Second, looking at it from a narrower perspective and in a narrow sense, the term indexing will refer to a major area of practice, within which specific types of indexes and oftentimes abstracts are able to be created. Cataloging is yet another area of practice within which specific types of cataloging records and very often classifications are able to be created. It should be understood however that while verbal representation is able to take place in both indexing as well as cataloging, the terminology or the descriptions may differ to some extent.

In this unit, we will discuss the concepts of indexing languages, vocabulary control and thesaurus.

NOTES

3.1 OBJECTIVES

After going through this unit, you will be able to:

- Discuss the indexing languages
- Explain vocabulary control
- Describe the concept of thesaurus

3.2 INDEXING LANGUAGES

What is meant by the term indexing languages?

Well, if we wish to define the term indexing languages, the first thing to remember is that the term indexing has to be looked at in the broader sense, that is, in a general sense:

- Indexing language is the process that helps create and provide access to objects of information; and
- It may be remembered that this process of creating and providing access to objects of information could either be manual or by human hand or through computer technology.

Next, we need to understand what the term index means.

An index may be defined in different ways, such as

- An organized list of pointers or access points to concepts and items in collection, or document
- May be called index, catalog, or information system
- May be print or electronic

In the present context, the term 'Document' can be used to refer to either textual or non-textual object of information. In the present context, it has been taken to define target documents, not search queries.

Many types of indexes exist, such as . . .

- Back-of-the-book index
- Periodical index for specific periodical(s)
- Index to the literature of a discipline
- Catalog of materials in collection or library

- Citation index of connections among documents based on authors' citations
- Directory of persons, businesses, organizations, etc.
- Inverted file/index of data values drawn from main file of computer database for the purpose of facilitating matches with search terms for retrieval

NOTES

Next, what is meant by an index entry?

Has this indicator or pointer been included in any index?

Similarly, an index term may be defined as:

- Being any word/phrase/number used for physical or subject representation in an information system
- Being any word/phrase/number used for searching and retrieving representations
- Being able to describe any attribute of a document

Therefore, the above definitions will help us define indexing language in the following different ways:

- As terms or vocabularies used to represent document container or content
- Serving as access points for searching
- Varying from one index or system to the next
- Possibly being extracted or derived from document text: natural language
- Possibly being assigned from authority control list: controlled vocabulary

Let us now look at how indexing languages are actually conceptualized and created.

All indexing languages originate as natural language, or the language found in documents. Natural language does not refer to writing style, but to the fact that the language is not under authority control.

Language under authority control is called controlled vocabulary. There is nothing special about the words in controlled vocabulary except the fact that they are standardized for use in certain systems.

This diagram illustrates the processes involved in translating natural language (NL) terms into controlled vocabulary (CV) terms for entry in database records. The diagram helps explain why . . .

1. Natural indexing languages are also called derived-term approaches
2. Controlled indexing languages are also called assigned-term approaches

NOTES

Abstracting and indexing processes

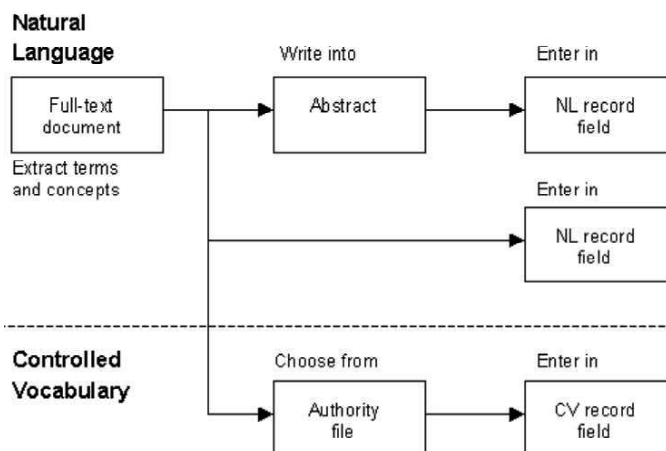


Fig 3.1 Processes Involved in Translating Natural Language Terms into Controlled Vocabulary

It may be observed that these processes include the steps of subject analysis. To review, subject analysis requires you to (1) become familiar with document content; (2) extract significant concepts and terms; (3) translate extracted terms into the language—often controlled—of the system; and (4) formalize the terms (format them, etc.) according to input rules.

It should be remembered, though, that the analysis process is used for physical description well as subject description. Further, individual fields often have their own indexing languages within the same system. For example, name authority control is often used for physical descriptions such as in the names of authors.

How does indexing relate to searching?

The processes of indexing and searching are complementary to each other in the following ways:

- Indexers and catalogers require information from database record fields
- Searchers enter these terms in the fields of database search interfaces

Depending on how accurately the terms used for searches match those terms in the records, the systems retrieve and return the data from the records.

Here's how the processes differs for natural language and controlled vocabulary:

Natural language	Controlled vocabulary
Terms are . . .	based on existing vocabulary of documents (which may be inconsistent) based on standardized vocabulary intended to describe concepts consistently
Indexers/catalogers . . .	extract terms from documents and enter them (or their own terms) in various subject fields

	extract terms from documents, choose appropriate authorized terms from controlled vocabulary list, and enter terms in designated controlled vocabulary field
Searchers . . .	may enter any search terms that are likely to occur in natural language must enter search terms that are in controlled vocabulary

NOTES

What is subject indexing?

Library Information Science aims most of its attention on languages used for subject definition because these are the most complicated. Basically, subject indexing can be said to be the process of creating an index for the purpose of representing and providing access to intellectual content.

Any term that is used to define or describe the content within a document is known as a subject term.

- A subject term can also be used to describe any term that may be used to search for a specific document based primarily on its content.
- A subject term may also be described as a compact synonym or surrogate for a specific subject representation.

For subject terms under authority control (or vocabulary control), a subject authority file or list . . .

- May be described as a list of terms that are permitted to be used in describing or representing specific subjects.
- May be said to standardize one of two synonyms that are used to assign or represent specific topics.
- May be used to determine the preferred term when multiple terms are used to define or describe a single topic.
- May be used to provide cross references for terms that are on par with, hierarchical or alternate in position or relationships.

Cataloging and indexing professionals have created different subject authority control structures:

- Subject headings lists are used by catalogers in cases where subject terms have been used as subject headings.
- A thesaurus is used by indexers when subject terms are known as descriptors.

Indexing languages can be studied as the following types:

- **Controlled indexing language:** This refers to the indexing language in which only approved terms are allowed to be used to describe the document.
- **Natural language indexing language:** This is a slightly broader language in which the description of the document can be done using any of the terms present in the document.

NOTES

- **Free indexing language:** As the name suggests, this type of indexing language brings into use any term within or outside the document for its description.

Another concept of much importance while indexing a document is selecting the level of indexing exhaustivity. This is nothing but the level of detail with which the document is described, and this is the discretion of the indexer. Minor details of the documents are not described in the low indexing exhaustivity whereas the higher level signifies more indexed terms.

In today's times, the searching mechanism and trends have changed and there is a higher use of free text search. This demands that the natural language with the highest possible indexing ideally indexing every text be done. Of course, the research of whether free text search or expert-driven well-chosen vocabularies is being done to check which is more efficient.

3.3 VOCABULARY CONTROL

To ensure that free text searching is of maximum use to the users, it is important to make sure that the accuracy is improved. And a great help in this case is 'controlled vocabularies'. It does so simply by eliminating from the retrieval list any item which is irrelevant to the search query. These irrelevant items are also known as 'false positives' and these occur merely because of the ambiguity of the natural language itself. Let us look at an example, to understand the concept a little better. Say for instance, one uses the term 'football' in the search query. Now, football is known as different sports in different parts of the world. Like in many parts this is referred to as an association football, but in many countries the term soccer is used for the same sport. Additionally, the term football is also representative of rugby football, Canadian football, Australian rules football, Gaelic football and much more, each of which have a different meaning even though they use the same word. To ensure that the confusions are kept at bay and only those documents which the user is searching is displayed is what controlled vocabulary seeks to establish. This is done by identifying, grouping and tagging the relevant items to the search query.

In fact, it can be easily said that the performance of the information retrieval system is greatly affected by the controlled vocabulary system, especially in comparison to the free text searching since when measuring the performance based on accuracy it can be noted that the number of actually relevant documents show up more in controlled vocabulary than in free text searching.

It can also be the recall value of the information system under the controlled vocabulary system increases in some cases since the general requirement of using synonyms in the search query as a part of the natural language is eliminated with the use of authorized terms. But there are also chances that the authorized terms in fact do not assist at all in bringing the relevant documents. Such problems arise mostly in situations when the search question itself has tangential references which

might have prompted the indexer to use a distinct term for, where the searcher doesn't recognize as separate from his query in the first place. This requires an additional understanding on part of the user of the controlled language in the sense that it matches with that of the indexer.

Another problem that might come up with controlled vocabulary is that with the concept of low indexing exhaustivity. For instance, if there is an article which has football as a second priority, it will not be tagged as 'football' by the indexer. But the article might in fact be looking for the same, but will not be able to find out due to the restriction of controlled vocabulary. The article will not go untagged under the free text search. But on the flipside, high exhaustivity is very high in the free text search and much lower precision. However, the it will have a higher recall value for the searcher if he/she uses every possible related combination to solve the problem of synonyms.

A very important threat for controlled vocabulary is the fact that it has the negative trait in the field of quickly changing technology. To avoid this, it is imperative that the authorized terms are updated regularly. Further, a crucial factor to note here is that compared to the free text, controlled vocabulary is much less specific given that the indexer's discretion might wrongly interpret the author's actual meaning. This is not a problem in free text search given that the author's words itself are used in the indexing.

A very important disadvantage of controlled vocabulary is that it is a comparatively costly affair given that it requires the services of human experts or automated systems which would index the entries one by one. It also requires the user to be acquainted with the controlled vocabulary system. Homographs solve the problem of the control of synonyms.

Faceted classification as well as other numerous methodologies are used for the creation of controlled vocabulary which facilitates multiple description of the data or documents.

As we have learnt before, the method for subject indexing, subject headings, taxonomies, and other knowledge organization systems is controlled vocabularies. This system requires

Controlled vocabulary schemes require the use of predefined, authorized terms that have been preselected by the designers of the schemes, in contrast to natural language vocabularies, which have no such restriction.

In library and information science, controlled vocabulary is a carefully selected list of words and phrases, which are used to tag units of information (document or work) so that they may be more easily retrieved by a search. Controlled vocabularies solve the problems of homographs, synonyms and polysemes by a bijection between concepts and authorized terms. In short, controlled vocabularies reduce ambiguity inherent in normal human languages where the same concept can be given different names and ensure consistency.

NOTES

NOTES

For example, in the Library of Congress Subject Heading (a subject heading system that uses a controlled vocabulary), authorized terms—subject headings in this case—have to be chosen to handle choices between variant spellings of the same word (American versus British), choice among scientific and popular terms (*cockroach* versus *Periplaneta hesaurus*), and choices between synonyms (*automobile* versus *car*), among other difficult issues.

Choices of authorized terms are based on the principles of *user warrant* (what terms users are likely to use), *literary warrant* (what terms are generally used in the literature and documents), and *structural warrant* (terms chosen by considering the structure, scope of the controlled vocabulary).

Controlled vocabularies also typically handle the problem of homographs, with qualifiers. For example, the term *pool* has to be qualified to refer to either *swimming pool* or the *game pool* to ensure that each authorized term or heading refers to only one concept.

There are two main kinds of controlled vocabulary tools used in libraries: subject headings and thesauri. While the differences between the two are diminishing, there are still some minor differences.

For example, the Library of Congress Subject Heading itself did not have much syndetic structure until 1943, and it was not until 1985 when it began to adopt the thesauri type term "Broader term" and "Narrow term".

The terms are chosen and organized by trained professionals (including librarians and information scientists) who possess expertise in the subject area. Controlled vocabulary terms can accurately describe what a given document is actually about, even if the terms themselves do not occur within the document's text. Well known subject heading systems include the Library of Congress system, MeSH, and Sears. Well known thesauri include the Art and Architecture Thesaurus and the ERIC Thesaurus.

Choosing authorized terms to be used is a tricky business, besides the areas already considered above, the designer has to consider the specificity of the term chosen, whether to use direct entry, inter consistency and stability of the language. Lastly the amount of pre-co-ordinate (in which case the degree of enumeration versus synthesis becomes an issue) and post co-ordinate in the system is another important issue.

Controlled vocabulary elements (terms/phrases) employed as tags, to aid in the content identification process of documents, or other information system entities (e.g. DBMS, Web Services) qualifies as metadata.

3.4 THESAURUS

In general usage, a thesaurus is a reference work that lists words grouped together according to similarity of meaning (containing synonyms and sometimes antonyms),

in contrast to a dictionary, which provides definitions for words, and generally lists them in alphabetical order. The main purpose of such reference works for users "to find the word, or words, by which [an] idea may be most fitly and aptly expressed" – to quote Peter Mark Roget, architect of the best known thesaurus in the English language.

Although including synonyms, a thesaurus should not be taken as a complete list of all the synonyms for a particular word. The entries are also designed for drawing distinctions between similar words and assisting in choosing exactly the right word. Unlike a dictionary, a thesaurus entry does not give the definition of words.

In library science and information science, thesauri have been widely used to specify domain models. Recently, thesauri have been implemented with Simple Knowledge Organization System (SKOS).

The word "thesaurus" is derived from 16th-century New Latin, in turn from Latin *hesaurus*, which is the Latinisation of the Greek ἐϋθάδῆυδ (thçsauros), "treasure, treasury, storehouse". The word *thçsauros* is of uncertain etymology. Douglas Harper derives it from the root of the Greek verb ὀέεῖ Ἰίáέ *tithenai*, "to put, to place."^[2] Robert Beekes rejected an Indo-European derivation and suggested a Pre-Greek suffix *-ar^wo-^l

From the 16th to the 19th centuries, the term "thesaurus" was applied to any dictionary or encyclopedia, as in the *Thesaurus linguae latinae* (1532), and the *Thesaurus linguae graecae* (1572). The meaning "collection of words arranged according to sense" is first attested in 1852 in Roget's title and *thesaurer* is attested in Middle English for "treasurer".

History of the Thesaurus

In ancient times, Philo of Byblos authored the first text that could now be called a thesaurus. In Sanskrit, the Amarakosha is a thesaurus in verse form, written in the 4th century.

The first modern thesaurus was Roget's Thesaurus, first compiled in 1805 by Peter Mark Roget, and published in 1852. Since its publication it has never been out of print and is still a widely used work across the English-speaking world. Entries in Roget's Thesaurus are listed conceptually rather than alphabetically. Roget described his thesaurus in the foreword to the first edition:

It is now nearly fifty years since I first projected a system of verbal classification similar to that on which the present work is founded. Conceiving that such a compilation might help to supply my own deficiencies, I had, in the year 1805, completed a classed catalogue of words on a small scale, but on the same principle, and nearly in the same form, as the Thesaurus now published.

NOTES

NOTES

Check Your Progress

1. List the steps involved in subject analysis.
2. Define level of indexing exhaustivity.
3. What is the basis for choices of authorized terms in controlled vocabulary?
4. List the two main kinds of controlled vocabulary tools used.

3.5 ANSWERS TO CHECK YOUR PROGRESS

1. The steps involved in subject analysis are to: (1) become familiar with document content; (2) extract significant concepts and terms; (3) translate extracted terms into the language—often controlled—of the system; and (4) formalize the terms (format them, etc.) according to input rules.
2. The level of indexing exhaustivity is nothing but the level of detail with which the document is described, and this is the discretion of the indexer.
3. Choices of authorized terms are based on the principles of *user warrant* (what terms users are likely to use), *literary warrant* (what terms are generally used in the literature and documents), and *structural warrant* (terms chosen by considering the structure, scope of the controlled vocabulary).
4. There are two main kinds of controlled vocabulary tools used in libraries: subject headings and thesauri.

3.6 SUMMARY

- Even before the students begin to understand the concept of designing, creating and using the appropriate indexing languages, the students need to understand the actual concept of what indexing is all about. They need to understand that indexing is a two dimensional concept.
- Indexing language is the process that helps create and provide access to objects of information.
- An index is an organized list of pointers or access points to concepts and items in collection, or document
- All indexing languages originate as natural language, or the language found in documents. Natural language does not refer to writing style, but to the fact that the language is not under authority control. Language under authority control is called controlled vocabulary.

- The following are the steps of subject analysis: to (1) become familiar with document content; (2) extract significant concepts and terms; (3) translate extracted terms into the language—often controlled—of the system; and (4) formalize the terms (format them, etc.) according to input rules.
- Library Information Science aims most of its attention on languages used for subject definition because these are the most complicated. Basically, subject indexing can be said to be the process of creating an index for the purpose of representing and providing access to intellectual content.
- Any term that is used to define or describe the content within a document is known as a subject term.
- Indexing languages can be studied as the following types: Controlled, natural and free indexing.
- The performance of the information retrieval system is greatly affected by the controlled vocabulary system, especially in comparison to the free text searching since when measuring the performance based on accuracy it can be noted that the number of actually relevant documents show up more in controlled vocabulary than in free text searching.
- In general usage, a thesaurus is a reference work that lists words grouped together according to similarity of meaning (containing synonyms and sometimes antonyms), in contrast to a dictionary, which provides definitions for words, and generally lists them in alphabetical order. The main purpose of such reference works for users "to find the word, or words, by which [an] idea may be most fitly and aptly expressed" – to quote Peter Mark Roget, architect of the best known thesaurus in the English language.
- In library science and information science, thesauri have been widely used to specify domain models. Recently, thesauri have been implemented with Simple Knowledge Organization System (SKOS).

NOTES

3.7 KEY WORDS

- **Indexing language:** It refers to the process that helps create and provide access to objects of information.
- **Subject indexing:** It refers to the process of creating an index for the purpose of representing and providing access to intellectual content.
- **Controlled vocabulary:** It is a carefully selected list of words and phrases, which are used to tag units of information (document or work) so that they may be more easily retrieved by a search.
- **Thesaurus:** It is a reference work that lists words grouped together according to similarity of meaning (containing synonyms and sometimes antonyms), in contrast to a dictionary, which provides definitions for words, and generally lists them in alphabetical order.

NOTES

3.8 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What are the different types of indexes which exist today?
2. How are indexes defined?
3. How does indexing relate to searching?
4. What are the different types of indexing languages?

Long-Answer Questions

1. Describe in detail the concept of controlled vocabulary.
2. Explain the concept of thesaurus in library science.

3.9 FURTHER READINGS

- Alberico, R. and Micco M. 1990. *Expert systems for reference and Information retrieval*. West Port: Meckler.
- Atchison, J. & Gilchrist, A. 1972. *Thesaurus construction: a practical manual*. London: Aslib.
- Chowdhry, G.G. 2003. *Introduction to modern Information retrieval*. 2nd Ed. London: Facet Publishing.
- Cleaveland, D. B. 2001. *Introduction to Indexing and abstracting*. 3rd Ed. Englewood, Colo: Libraries Unlimited.
- Ghosh, S.B. and Biswas, S.C. 1998. *Subject Indexing systems: Concepts, methods and techniques*. Rev. ed. Calcutta: IASLIC.
- Pandey, S.K. Ed. 2000. *Library Information retrieval*. New Delhi: Anmol.

UNIT 4 DESIGN OF INDEXING LANGUAGES

NOTES

Structure

- 4.0 Introduction
- 4.1 Objectives
- 4.2 Design of Indexing Languages
- 4.3 General Theory of Subject Indexing and Thesaurus
- 4.4 Answers to Check Your Progress Questions
- 4.5 Summary
- 4.6 Key Words
- 4.7 Self Assessment Questions and Exercises
- 4.8 Further Readings

4.0 INTRODUCTION

This unit introduces students to the concept of designing and developing indexing languages. In order to be able to understand the concept of how to design and develop indexing languages, students will, to begin with, need to understand why they would need to design indexing languages at all.

Today we live in times when everything is being computerized, especially in spaces where a great deal of information and data is required to be collected, sorted and organized, stored and retrieved without wasting much time, and with the minimum of effort and time being wasted.

This is the precise reason computers are so essential in spaces such as libraries, where there is so much information that needs to be collected about books, innumerable subjects and topics, categorizing books according to authors, subjects or topics. Once the books have been sorted and categorized according to authors, subjects and topics, they need to be stored appropriately so that they are easy to retrieve when people wish to refer to them or borrow them.

In current times, libraries are established to serve specific needs of people who use them. There are public libraries, privately owned libraries, libraries that respond to special or specific purposes and academic libraries.

In order to allow users to refer to the appropriate books or access the information they need within the shortest span of time, libraries will be required to have all the information computerized. Libraries also would be required to collect and store personal details of the people who borrow those books or access the stored information. All this collected information needs to be stored in computers, to help libraries maintain records, manage the collection of books and reference books they possess and help an increasing number of people access the books and information they possess.

NOTES

In the earlier units, we have discussed the indexing languages, and why libraries require indexing languages at all to function professionally. We have explained the concept of indexing languages in detail. In this unit, we will explain how to design those indexing languages and not just the indexing languages, we will also explain how to design the thesaurus, which is another method to store information in large collections and sizes.

4.1 OBJECTIVES

After going through this unit, you will be able to:

- Describe the concept of designing of indexing languages
- Discuss the general theory of subject indexing and thesaurus

4.2 DESIGN OF INDEXING LANGUAGES

To begin this section of how to design and why to design indexing languages, we will first discuss and explain the concept of indexing in relation to libraries, where documents or books are required to be tracked and accessed and then retrieved at high speed. Indexing is the first step to identifying and accessing the documents and books required by any users or by anyone who is looking for information or data on any specific subject or topic.

Indexing: In simple terms, indexing is the method of speeding up and making efficient the process of searching for relevant documents. The importance of indexing can be realized by the fact that if this process is absent then the search engine would go through the entire library of documents every time a search is made and this would require a considerably long time and energy. For instance, the search query using indexing can query close to 10,000 documents in milliseconds, while without it, the search itself would take a lot of time. One can say the storage space of an index and its update will each a chunk of space and time but this is neutralized with the speed with which the retrieval is done once it is in use.

Design Factors involved in Indexing: Let us now have a look at some of the important factors that play a significant part in the search engine architecture.

1. **Storage techniques:** This involved making decisions related to the how the information that is being developed is stored, including whether it needs to be filtered or compressed.
2. **Size of the index:** This is related to the aforementioned point and involves consideration and planning related to the volume of computer storage which will easily store the index data.
3. **Merge factors:** This pertains to the questions about how the addition of subject features or words and in fact the manner of input of data

related to the index during text corpus transversal. It is important that the indexer is clear whether the system is adding new data or updating the early ones. There should also be decisions as regards to whether there is a possibility of a number of indexers working on different things at different times. The index merging shares traits with SQL merge commands and related merge algorithms.

4. **Maintenance:** The frequency and the manner in which the index is maintained and checked for faults is very important too.
5. **Lookup speed:** It is a very crucial area of focus to decide how quickly the word is identified in the data structure. This is also important in the context of how easily it can be deleted or updated.
6. **Fault tolerance:** It is important to make decisions related to making the service the most reliable. This will include focus on areas such as bad hardware partitioning, hash-based or composite partitioning as well as replication.
7. **Structure of the index:** This will reflect the overall design decisions made which will affect the manner in which the indexing is actually undertaken and stored and shown in the search engine. The various types of indices include:
 - **Suffix tree:** This is used to facilitate linear time lookup. This means that like a tree the suffixes of words are stored along side the main word. Further, extendable hashing also becomes possible due to this. But on the negative side, it requires a considerably large space to store. This problem can be dealt with, with the help of a suffix array which requires lesser memory supports data compression.
 - **Inverted index:** In the form of a binary tree or a hash table, it basically acts like a storage for list of occurrences of different atomic search criterion.
 - **Citation index:** This is a subject which comes under bibliometrics. It stores the in text citations and hyperlinks which will help in citation analysis.
 - **Ngram index:** It assists in storing the sequence of length of data which will help in the process of text mining or simply retrieval of other types of data.
 - **Document term matrix:** This is technique which utilizes the frequency of occurrence of words in a document in the form of a two-dimensional sparse matrix. This is helpful in latent semantic analysis.

NOTES

NOTES

Barriers in Natural Language Processing:

- **Language ambiguity:** While searching for documents, several times the indexing procedure might include associated the search of the search query terms with its root language origin. This is called including an additional information of the search terms like associating its lexical category. These are also complicated as the syntax of different languages are different.
- **Word boundary ambiguity:** Although the process of indexing or tagging words and documents might not seem like a complicated procedure for the native speakers of the language. But this becomes problematic when the multilingual document and tagging is to be considered. To have a more universal appeal considerations of the languages with similar syntax and boundary markers is of utmost importance for it to have relevance to the searchers.
- **Fault in storage:** There are several problems in question here like the natural language lacking perfection, the files not following a standard protocol, the mistakes in the encoding of the binary characters, which might make the index performance not perfect.
- **Varied file formats:** The characters present in the document are very important for proper indexing to take place. But if the file formats are not recognized and correctly handled, it will make indexing inaccurate. Therefore, it is imperative that the search engine supports as many file formats as possible.

Tokenization

The computer system works in a different form than that of natural language. The system does not recognize the structure of the natural language but only the sequencing of the bytes. This process of programming the computer to make it understand what constitutes a word and which character is different is known as tokenization. For instance, a computer would not understand that the character of 'space' means that two different words are being separated here. The program which assists in this process is known as lexer, parser or tokenizer. Some special programs used here are Lex or YACC.

The parser in the process of tokenization includes several activities like identifying the pattern of elements representing words or elements. This includes the recognition of punctuations, numeric characters, other non-printing controls, URLs, email addresses, etc. Apart from such identifications, the characteristics of the elements are also noted like recognizing lexical category, language or encoding, line numbers, token's case, length, sentence positions, etc.

Language recognition

As the name suggests this process involves identifying and tagging the language of the document. This is especially helpful in cases where the search engine is

multilingual. This process is also known as language analysis, language tagging, language classification, language identification. The method of automated language recognition is still under a lot of research and refinement. This may involve the use of tools like language recognition chart.

Format analysis

Format analysis is a very important part of indexing especially when the search engine supports documents of multiple formats. This includes proper recognition of the different information present within the documents which are different from the text material itself. For instance, information related to font size, style or new line with reference to HTML tags must be understood by the system. This requires proper identification of different between the markups and the content itself otherwise the document will not properly be understood and will not display in the search results. Further format analysis, is important because the handling of the formatting content embedded within documents which reflects the manner in which the document is rendered on a computer screen or interpreted by a software program. Another name given to this analysis is structural analysis, analysis, format parsing, tag stripping, format stripping, text normalization, text cleaning and text preparation. Of course the document identification is not an easy process which is complicated by factors like the information actually available about the formats, some of which could have proprietary restrictions. The following are some of the common, well-documented file formats that are compatible with most of the search engines support:

- HTML
- ASCII text files (a text document without specific computer readable formatting)
- Adobe's Portable Document Format (PDF)
- PostScript (PS)
- LaTeX
- UseNet netnews server formats
- XML and derivatives like RSS
- SGML
- Multimedia meta data formats like ID3
- Microsoft Word
- Microsoft Excel
- Microsoft PowerPoint
- IBM Lotus Notes

It is very important to note here that inspection of compressed files are also supported by many search engines. The decompression of these files in the indexing process may result in the creation of more than one files and it is crucial that the

NOTES

NOTES

indexer indexes each of these documents separately. The following the Commonly supported compressed file formats include:

- *ZIP - Zip archive file*
- *RAR - Roshal ARchive file*
- *CAB - Microsoft Windows Cabinet File*
- *Gzip - File compressed with gzip*
- *BZIP - File compressed using bzip2*
- *Tape ARchive (TAR), Unix archive file, not (itself) compressed*
- *TAR.Z, TAR.GZ or TAR.BZ2 - Unix archive files compressed with Compress, GZIP or BZIP2*

It is also important for the indexer to keep in the mind the fact the ‘bad information’ is hidden or avoided from the main text so as to improve its quality. Content can manipulate the formatting information to include additional content. Examples of abusing document formatting for spamdexing:

- This comprises of different characters or words in section which is not displayed to the reader in front of the computer screen but to the indexer only. This is done through formatting with CSS or Javascript which will hide the likes of the hidden ‘div tag’ in the HTML.
- Sometimes the foreground colours of the words are matched with the background colour but visible to the indexer.

Section recognition

Sometimes the search engines include the identification of only sections of the document before its tokenization. This is to say that sometime the documents include such sections which are not the main focus of the document but provide non-related or additional material. For example, this article displays a side menu with links to other web pages. Some file formats, like HTML or PDF, allow for content to be displayed in columns. So, it might occur at times, that some of these extra content is stored in a sequence while preparing the raw markup content. In the indexing procedure, words or documents that occur in different areas of the document are indexed by default in a sequence. So if the search engine indexes this document as a normal content then it will lower the quality of the document in the search given the mix of different irrelevant words.

Check Your Progress

1. Name some of the parser programs used in indexing.
2. What are some of the other names for language recognition in indexing?

4.3 GENERAL THEORY OF SUBJECT INDEXING AND THESAURUS

NOTES

Students of Library and Information Science are the people who would want most frequently to understand what is meant by the term subject indexing or subject analysis. So to begin with let us try to understand the concept of subject indexing and why subject indexing would be required at all in libraries and the task of information retrieval.

Libraries are the spaces where information is usually collected, sorted and organized, stored and retrieved on a large scale, whether manually or through computers. Before computers became commonly used across all spheres of human life in modern times, libraries and to be precise librarians did all of these tasks, that is collecting information about books and documents, authors and poets, people who wished to access such information on a regular basis for specific purposes, sorting and organizing such information on the basis of subject content or topic, or sometimes even on the basis of authors or poets, categorizing such information, storing all that information appropriately in ledgers and journals, and retrieving all the collected and categorized information and data when required on a daily basis, manually.

The invention of computers made the task of librarians and most other industries and business for that matter, much simpler, faster and easier. Large spaces such as libraries, where information and data is collected, sorted and organized and categorized according to subject and topic, and retrieved at high speed, especially, the need for indexing, is obvious. This means librarians would be required to understand the concept of indexing, how to index the information collected on the basis of subject and topic, and sometimes on the basis of author or poet, storing such information and data and retrieval for the use of the students, researchers or other professionals who require such information. So in order to make all this work simpler and easier, it is obvious that librarians would be required to understand and learn how to design and create the indexes they wish to use within their libraries on a professional level.

This section explores the concept of subject indexing, and explains the theory of subject indexing in some detail.

As the name suggests, in the process of subject indexing, the description of the document containing information about the subject is included in the tagging in the form of index term or symbols. Three different stages of indexes are created in the process: terms in a document such as a book; objects in a collection such as a library; and documents (such as books and articles) within a field of knowledge.

Bibliographic indexes are created in subject indexing to help retrieve documents on the basis of subject. Zentralblatt MATH, Chemical Abstracts and

NOTES

PubMed are some of the examples of academic indexing services. The index terms were mostly assigned by experts but author keywords are also common.

The analysis of the subject of the document is the first step of subject indexing. The second step involves taking the texts either right from the document or from the controlled vocabulary to assign to the document. Then the terms are sequentially arranged in the index. The number of terms and the types of terms to be included in the index is the discretion of the indexers. Together this gives a depth of indexing.

As a first step of the indexing, the indexer must first understand the subject matter of the document. In manual indexing procedure, the indexer would consider the subject matter in terms of answer to a set of questions such as "Does the document deal with a specific product, condition or phenomenon?". Of course, the analysis is influenced by two subjective factors: the knowledge and experience of the indexer. It then becomes possible that different indexers may analyse the content differently and so come up with different index terms. This will reflect on the performance of the index.

The advantages of automatic subject analysis against manual analysis

Automatic indexing are based on pre-determined processes of analysis frequencies of patterns of words and then comparing the results with other documents to be able to assign them to subject categories. This does not need any understanding of the content being indexed. So this in turn leads to additional equitable indexing has an adverse effect on the actual meaning that is interpreted. A computer program may perhaps not discern the meaning of statements and so not assign some appropriate terms or perhaps assign inappropriately.

On the other hand, human indexers may perhaps focus their attention on specific aspects of the document such as the title, abstract, summary and conclusions, because an analysis of the complete content in depth may be expensive and time consuming. An automated indexing system helps remove the time constraint and enables the complete content to be analyzed, besides providing the additional option to be directed to specific areas of the content.

Choice of term

It is during the second stage of indexing that the subject analysis gets to be translated into a list of index terms. This could involve retrieving from the content or assigning relevant words from a controlled dictionary or vocabulary. Having the facility to carry out a complete text search, many people are now able to depend on their individual expertise in carrying out information searches so that complete content searches have now become extremely easy.

Subject indexing has its experts, that includes professionals such as indexers, catalogers, as well as librarians, are essential for information organization and retrieval. These experts are able to understand the meaning and context of controlled vocabularies besides being able to locate information that cannot be otherwise located by full length content searches.

The expenses incurred by such expert analysis to create subject indexing cannot be compared with a set of full-text, fully searchable materials. Compared to the expense of hardware, software and labor of new web applications that enable every user to analyse content or information, use has now become increasingly common especially on the Internet.

In spite of the information revolution in recent times, one area where indexing has obviously remained status quo, or unchanged, is the book indexing field.

Extracted or Extraction indexing involves using words directly from the content. It uses the original vocabulary and also becomes adaptable to automated methods in which word frequencies are determined while those with a frequency higher than a pre-fixed limit have been used as index terms. A stop-list indicating common words (such as "the", "and") may be referred to and such stop words are usually not used as index terms.

There are great chances of the loss of meaning of the terms present in the document in the process of automated extraction indexing as the single words are indexed here instead of phrases. Even though it is possible to index phrases but mostly it becomes more difficult if key concepts are inconsistently worded in phrases. Automated extraction indexing also suffers from the disadvantage that, even with use of a stop-list to remove common words, some frequent words may not be useful for allowing the identification of difference between documents. For instance, the term glucose is likely to occur commonly and more likely in any document related to diabetes. Therefore, use of this term would likely return most or all the documents in the database. Post-co-ordinated indexing could help to minimize this problem since the phrases will be combined at the time of searching but this process adds the onus of this on the searcher instead of the indexer. Further, there are chances that the word may not occur as frequently but it has a really high significance. For example, the word drug might not be used as liberally but the gravity of the topic makes any reference to the word very crucial as a subject. To tackle this problem, one thing can be done and that is using a comparison of the frequently mentioned word in the document to its frequency position in the overall database. This will help to a certain extent to include rarer words than the common ones. Therefore, a term that occurs more often in a document than might be expected based on the rest of the database could then be used as an index term, and terms that occur equally frequently throughout will be excluded.

A different problem that might occur with automatic extraction is that the search engine is not able to recognize when a subject is being discussed in the document due to the lack of the indexable word in the document.

Indexing of Assignment

An assignment indexing is always a feasible alternative, in which index terms are used directly from a controlled vocabulary. This method provides the benefit of controlling the use of synonyms since the preferred term is indexed and synonyms

NOTES

NOTES

or related terms lead individual users to the preferred term. This shows that such users are able to locate articles irrespective of the specific term that have been used by the author and enables the user to work without the need to know and identify the possible synonyms.

This method also prevents any confusion that may result from homographs by including a qualifying term. A third benefit is that this method enables linking the related terms irrespective of whether they are connected by hierarchy or association, for instance, an index entry that indicates an oral medication could also list alternate oral medications as related terms on the same level of hierarchy but could also connect the user to broader terms such as treatment.

Assignment indexing is usually used in manual indexing in an effort to improve inter-indexer consistency because varying indexers could have a controlled set of terms from which to choose. Controlled vocabulary does not entirely mitigate inconsistencies because two indexers could yet interpret the same subject differently

Presentation of the designed Index

The last stage of indexing is where all the entries are presented or put forward in a systematic sequence. This stage could involve connecting entries. Within a pre-fixed index the user decides the sequence in which terms are connected in an entry depending on how the user is able to formulate their search. Within a post-coordinated index, however, the entries are presented individually while the user is able to connect the entries through searches that are normally carried out by computer software. Post-coordination usually causes a loss of accuracy in comparison to pre-coordination.

Let us now explore some of the common terms related to the concept of indexing.

Depth of indexing

Indexers need to take decisions on such issues as which of the entries have to be incorporated and the number of entries an index that have to be included. The intensity of the indexing defines the accuracy of the indexing process with regard to its clarity and specificity

Specificity

The specificity has to define how exactly the index terms match the topics they relate to. An index is described as being specific when the indexer makes use of alternate descriptors to the central idea contained in the document and echoes the ideas contained in the sub-topics. Specificity can be said to increase in proportion to its use which means the higher the number of terms that are being included, the narrower those terms will be.

Next, Exclusivity

An exhaustive index may be described as one which includes all the possible index terms. This means that a higher level of exclusivity results in a higher level of retrieval, or the greater the possibility of all the related articles being retrieved. But this can take place at the cost of accuracy. This indicates that the user will be able to recall a greater number of related articles or documents that focus specifically on the topic in much greater detail.

In a manual system a higher level of usage brings with it a greater loss because more man hours are usually needed. On the other hand, the time needed on a computerized system would be much less. At the other end of the spectrum, while using a selective index only the most significant and valid are points are retrieved. This shows that retrieval is minimized while using selective index because an indexer does not usually need to include extra terms. So a highly relevant document could be overlooked. This indicates that indexers need to endeavor to strike a balance and consider what the article would be used for. They could moreover need to decide the implications of time and expense.

Finally, we come to the actual Theory of Indexing

The theory of indexing at the farthest level could be linked to various theories of knowledge. Rationalist theories of indexing (for instance Ranganathan's theory) indicate that topics are developed rationally from a basic set of categories. The fundamental technique of topic analysis is thus "analytico-synthetic", in order to segregate a set of fundamental categories such as analysis and then to develop the topic or theme of any given article by combining those categories as per the mandated rules.

Empiricist theories of indexing are based on choosing alternate articles that are based on their properties, specifically by applying numerical statistical methods. Historicist and hermeneutical theories of indexing show that the topic of a given article is related to a given discussion or field, then the indexing has to show the requirement of a specific discussion or area. As per hermeneutics an article is always written and translated or understood from specific boundaries. The same rule pertains to systems of knowledge organization as well as interpretations of all such systems.

The theory of indexing is, as indicated by Rowley & Farrow is intended to evaluate papers contributed to knowledge and index it accordingly. Or, in the words of Hjørland (1992, 1997) to index its informative capacity.

Check Your Progress

3. What are the stages of indexes created in the indexing procedure?
4. Give some examples of academic indexing services.

NOTES

NOTES

4.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Some special parser programs used in indexing are Lex or YACC.
2. Language recognition process is also known as language analysis, language tagging, language classification, language identification.
3. Three different stages of indexes are created in the process: terms in a document such as a book; objects in a collection such as a library; and documents (such as books and articles) within a field of knowledge.
4. Zentralblatt MATH, Chemical Abstracts and PubMed are some of the examples of academic indexing services.

4.5 SUMMARY

- Indexing is the first step to identifying and accessing the documents and books required by any users or by anyone who is looking for information or data on any specific subject or topic.
- The importance of indexing can be realized by the fact that if this process is absent then the search engine would go through the entire library of documents every time a search is made and this would require a considerably long time and energy.
- Some of the important factors that play a significant part in the search engine architecture include: storage techniques, size of the index, merge factors, maintenance, structure, fault tolerance, etc.
- Barriers in natural language indexing include: word boundary ambiguity, faulty storage, varied formats, etc.
- The computer system works in a different form than that of natural language. The system does not recognize the structure of the natural language but only the sequencing of the bytes. This process of programming the computer to make it understand what constitutes a word and which character is different is known as tokenization.
- Language recognition is the process involves identifying and tagging the language of the document. This is especially helpful in cases where the search engine is multilingual.
- Format analysis is a very important part of indexing especially when the search engine supports documents of multiple formats. This includes proper recognition of the different information present within the documents which are different from the text material itself.
- Sometimes the search engines include the identification of only sections of the document before its tokenization. This is to say that sometime the

documents include such sections which are not the main focus of the document but provide non-related or additional material.

- The invention of computers made the task of librarians and most other industries and business for that matter, much simpler, faster and easier. Large spaces such as libraries, where information and data is collected, sorted and organized and categorized according to subject and topic, and retrieved at high speed, especially, the need for indexing, is obvious.
- As a first step of the indexing, the indexer must first understand the subject matter of the document.
- It is during the second stage of indexing that the subject analysis gets to be translated into a list of index terms.
- Extracted or Extraction indexing involves using words directly from the content. It uses the original vocabulary and also becomes adaptable to automated methods in which word frequencies are determined while those with a frequency higher than a pre-fixed limit have been used as index terms.
- The last stage of indexing is where all the entries are presented or put forward in a systematic sequence. This stage could involve connecting entries.

NOTES

4.6 KEY WORDS

- **Indexing:** It is the method of speeding up and making efficient the process of searching for relevant documents. The importance of indexing can be realized by the fact that if this process is absent then the search engine would go through the entire library of documents every time a search is made and this would require a considerably long time and energy.
- **Format analysis:** It is a very important part of indexing especially when the search engine supports documents of multiple formats. This includes proper recognition of the different information present within the documents which are different from the text material itself.
- **Specificity:** In indexing, it defines how exactly the index terms match the topics they relate to.

4.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What are some of the barriers in the natural language processing?
2. List some of the commonly supported compressed file formats.
3. What is section recognition in indexing?

4. Write a short note on the advantages of automatic subject analysis against manual analysis.
5. Briefly explain the presentation of the designed index.

NOTES

Long-Answer Questions

1. Explain some of the important factors that play a significant part in the search engine architecture.
2. Describe the concept of format analysis in indexing.
3. Discuss in detail the various stages of indexing.

4.8 FURTHER READING

- Alberico, R. and Micco M. 1990. *Expert systems for reference and Information retrieval*. West Port: Meckler.
- Atchison, J. & Gilchrist, A. 1972. *Thesaurus construction: a practical manual*. London: Aslib.
- Chowdhry, G.G. 2003. *Introduction to modern Information retrieval*. 2nd Ed. London: Facet Publishing.
- Cleaveland, D. B. 2001. *Introduction to Indexing and abstracting*. 3rd Ed. Englewood, Colo: Libraries Unlimited.
- Ghosh, S.B. and Biswas, S.C. 1998. *Subject Indexing systems: Concepts, methods and techniques*. Rev. ed. Calcutta: IASLIC.
- Pandey, S.K. Ed. 2000. *Library Information retrieval*. New Delhi: Anmol.

UNIT 5 INDEXING SYSTEMS AND TECHNIQUES – PRECOORDINATE INDEXING

NOTES

Structure

- 5.0 Introduction
- 5.1 Objectives
- 5.2 Techniques of Indexing Systems
 - 5.2.1 Precis
 - 5.2.2 POPSI
 - 5.2.3 Chain Indexing
 - 5.2.4 Relational Indexing
 - 5.2.5 Other Indexing Processes
- 5.3 Answers to Check Your Progress Questions
- 5.4 Summary
- 5.5 Key Words
- 5.6 Self-Assessment Questions and Exercises
- 5.7 Further Readings

5.0 INTRODUCTION

To begin with, the processes of indexing of any given document generally takes place in two stages. These include the following:

1. Establishing the concepts that are contained in any given document, and then
2. Translating those concepts into the components of the given indexing language.

Elaborating on these two stages a bit more, establishing the concepts of any given document would usually require first understanding the content of the entire document/ This again is in two parts, the text and the images. In order to understand the text of the given document would require a person to examine the content in detail, especially the most important points such as the title, the abstract, the introduction, the opening phrases or the key words of each of the chapters and their opening paragraphs. The images also need to be examined in detail. The images would include all the illustrations, the diagrams, the photographs, and the tables and charts. The conclusions also need to be examined and understood in detail.

Once the entire document has been thus examined in detail, the next step would be to identify the concepts of the document that actually need to be indexed.

NOTES

This means the concepts that get to be so selected need to be looked at from the perspective of the purpose for which the indexed data will be used. After this, finally, the concepts that have been so identified and elected will need to be translated into the components of the given indexing language.

Let us now examine, explore and understand the meaning and relevance of the two terms exhaustivity and specificity in indexing.

The gamut of all the different topics discussed in a document is measured through the concept of exhaustivity in indexing. To put it in more simpler terms, it represents the total number of index terms that are used in the indexing of a document. This follows that the greater the exhaustivity of the document the higher the recall value of the document in the indexing system. Ofcourse, the level of exhaustivity is matter of policy decisions.

Whether the subject in the indexing system captures the thought or idea behind the document is measures through the concept of specificity. This quality or measure is rather an intrinsic trait of the document. And this subsequently means that the more high the level of specificity of the document, the greater its precision.

C. A. Cutter was perhaps the first person to have introduced a generic set of rules for topic based indexing in his Rules for Dictionary Catalog that was published in the year 1876. Cutter defined rules intended for both specific as well as compound topic headings. Cutter considered topics as being specific and types as broad. But, in reality, it was these broad types that Cutter entered his specific topics. Moreover, as per Cutter, the order in which the components were entered as terms in a compound subject heading has to be the one which is definitely the most significant. But Cutter was not able to give any guideline that would show users how to decide which of the terms were to be considered the more significant. The degree or level of significance differs from user to user as well as from indexer to indexer.

J. O. Kaiser began precisely at this point where Cutter had been unable to give the guideline concerning the issue of significance. Kaiser, in his Systematic Indexing, which had been published in the year 1911, suggested that the compound topic has to be analysed by deciding the relative importance of the various parts terms of a compound topic through classificatory approach of categorization of terms. He divided the component terms into two basic categories: concrete and process. Kaiser gave the indication that concrete is more important than process and so, he laid a rule that a process has to follow concrete.

J. E. L. Farradane has identified the nine relationships in his Relational Indexing. He proposed nine relationships and their respective relational operators. We will discuss them further in the unit.

5.1 OBJECTIVES

After going through this unit, you will be able to:

- Compare and contrast the different techniques of indexing
- Explain POPSI, web indexing and NLP indexing.

NOTES

5.2 TECHNIQUES OF INDEXING SYSTEMS

In this section, we will have a look at the different techniques of the indexing systems.

5.2.1 Precis

It stands for Preserved Context Index System. It was developed in 1974 by Derek Austin. It is a pre-coordinate indexing system. It is based on two basic principles: principle of one-to-one relationship and principle of context dependency. Let us now look at and discuss how the syntactic and semantic relationships are handled in Precis. It follows a two level process: syntactical and semantical.

Syntactical relationships in PRECIS are usually dealt with by means of a set of logical rules, role operators as well as codes. These rules determine the order of terms that appear in the input string by the indexer and their utilization to create index entries by the computing system. Role operators work as directions to the computer. Semantic relationships in PRECIS are governed by an online thesaurus that works as a source of as well as references point for *the index*. A thesaurus is accessed in tandem with the readiness of input string.

It may be present in standard or inverted format or predicate transformation.

5.2.2 POPSI

It refers to the Postulate Based Permuted Subject Indexing. It was developed by Dr G. Bhattacharya in 1984. It is a pre-coordinate indexing system. Let us now discuss how POPSI-specific was first developed.

The fundamental belief that leads to the creation of POPSI-Specific is that topic indexing is usually a specific purpose-based activity. But, there has usually been a custom of depending on a designer of topic indexing, language, and such total reliance has usually been found to be insufficient to meet the specific needs of topic indexing at the local level. Variance in needs could require variances in syntax of the topic proposition. It has been stipulated that the flexibility has to be the rule of syntax, not the rigidity. Based on this belief, POPSI could be said to strive to discover what is logically fundamental, known as POPSI-Basic, and therefore readily amenable to the systematic use to derive purpose-based specific versions, known as POPSI-Specific.

NOTES

Pre-coordinate indexing involves coordination of component terms within a compound topic by the indexer at the time of indexing in an expectation of users' approach. On the other hand, in post-coordinate indexing, component terms in a compound topic are maintained separately and not coordinated by the indexer, and the user does the coordination of terms in accordance with his requirements at the time of searching. In the pre-coordinate indexing system, the order of the rigidity of the importance connected with the syntactical rules may not meet the approaches of all the people using the index file. However, in the post-coordinate indexing, the searcher has wide options for the free use of the terms during the process of searching in an effort to achieve as much logical operations as may be needed.

Some specific devices can help eliminate false drops in post-coordinate indexing. These include the following.

The following devices are used to eliminate false drops in post-coordinate indexing:

- (a) Roles;
- (b) Weighting;
- (c) Links; and
- (d) Use of bound terms

There are multiple versions of keyword indexing. These include the following.

There are multiple versions of key word indexing. Some of the more important ones may include the following:

- (a) KEYTALPHA (Key Term Alphabetical) Index;
- (b) KWIC (Keyword –In-Context) Index;
- (c) KWOC (Keyword Out-of-Context) Index;
- (d) KWIT (Keyword-In-Title) Index;
- (e) KWWC (Keyword-with-Context) Index;
- (f) SWIFT (Selected Words In Full Titles) Index;
- (g) DKWTC (Double KWIC) Index;
- (h) KLIC (Key-Letter-In-Context) Index;
- (i) WADEX (Word and Author Index); and
- (j) KWAC (Keyword Augmented-in-Context) Index

We will discuss some of these in succeeding unit.

The following are the different methods adopted in measuring the word significance in computerized indexing:

- (a) Use of noun phrase;
- (b) Relative frequency weighting;

- (c) Weighting by location;
- (d) Use of thesaurus;
- (e) Use of association factor; and
- (f) Maximum-depth indexing.

The NLP based topic indexing systems are not very easy to understand. However it could be still possible to follow their meaning with the help of various types of knowledge of a language. These could include the following.

Different levels of knowledge used in natural language understanding for the NLP-based subject indexing system falls into the following groups:

- (a) Morphological knowledge;
- (b) Lexical knowledge;
- (c) Syntactic knowledge;
- (d) Semantic knowledge;
- (e) Pragmatic knowledge; and
- (f) World knowledge.

There are also certain non-conventional indexing techniques. Indexing of scientific citations could be segregated into different sections that usually include as the following.

There are three parts in Science Citation Index: c) Permuterm Subject Index.

- (a) Citation Index;
- (b) Source Index; and

Web indexing

The term web indexing could be said to be multi-dimensional, and could be described in the following words.

The term 'Web indexing' relates to many things, such as to (a) search engine indexing of the Web, developing a range of metadata, (c) developing the online or web to resemble a model of a back-of-book index. On the whole, web indexing is arranged alphabetically, thus allowing a systematic access to information, as well as entries of contain index which have subsections as well as cross-references. Looking at this in a more general way, Web indexing would indicate that provide access at various points or levels for digital information resources that then become available or accessible across the world through what is commonly referred to as world wide web, or www browsing software. The term web indexing could also include the following aspects or issues:

- (a) Uploading or upgrading the so-called 'traditional' indices (as well as the articles that they refer) on to the internet so as to provide a broader audience that give access to them.

NOTES

NOTES

- (b) ‘Micro’ indexing of an individual Web page, so as to be able to give users an option of hyperlinked access points to the resources on that particular page.
- (c) ‘Midi’ indexing of numerous pages, broadly or basically comprised in an individual Web site while yet falling within the onus of an individual Webmaster.
- (d) ‘Web-wide’ indexing, that allows users a more centralized access to a more scattered resources, that could be clubbed below a single heading (for instance, every web page that pertains directly with ‘breast cancer’).
- (e) ‘Macro’ systems that have been developed to make access to a greater number of Web pages falling under many different headings (e.g. every web pages that pertain directly to a specific medical topic easier and simpler).
- (f) The integration of comments and annotations that are intended to help users get connected to the web sites or pages that they are looking for.

A search engine usually has numerous or multiple sections, while each of these sections will have a specific purpose or function. These would include the following:

- (a) **Search and engine mechanism:** In this system, the user poses a query to the index and the results displayed are in order of rank relevancy.
- (b) **Spider:** These refers to the computer programs which include collecting different elements related to identifying and reading Web pages including collecting data from the websites through the links mentioned in the documents.
- (c) **Index:** This refers to the indexed items which are from Web pages and form a searchable database.

Let us look at some of the important, fundamental and specific tech tools that could be used in the designing or creation of semantic web sites or pages:

- (a) Ontology;
- (b) eXtensible Markup Language (XML);
- (c) Uniform Resource Identifier (URI);
- (d) Agent software;
- (e) Resource Description Framework (RDF)

5.2.3 Chain Indexing

Chain indexing involves the process of evaluating subdivision of subject entries to arrive at an alphabetical list of subject entries which are to be indexed in the order

of their specificity. It is a pre-coordinate indexing technique. It utilizes a preferred classification scheme to extract class numbers of the documents concerned for deriving subject index entries.

The important steps involved in the chain indexing procedure include ascertaining the particular subject of the document, assigning the expressive name of the subject, considering the kernel terms, analysed terms, transformed and standard terms of the document, etc.

5.2.4 Relational Indexing

The problem with having multiple objects in a single space is the difficulty in identifying three dimensional objects. This is where relational indexing comes into the picture. Relational indexing is something that can be used to identify objects that exist within a singular space that contains multiple objects of different features and contours.

When considering a database of relational objects or models, it is possible to recognize or identify those models or objects in the sub sects or subcategories whose descriptions or definitions are most akin to, or match, those descriptions or definitions of the unknown situation or scene.

In the relational indexing of ‘System of Relational Analysis’ as developed by J. E. L. Farradane in 1950, the indexing system includes relational operators which connect the separate isolates. The relational operators in fact show how the different isolates are related to each other. These operators are represented through a slash followed by the specific relational sign. This relational indexing is crucial to helps to form relations between terms. He specifically proposed nine relational operators including:

- (a) Association / ;
- (b) Concurrence / 0
- (c) Reaction / -
- (d) Self-activity / *
- (e) Equivalence / =
- (f) Dimensional / +
- (g) Causation / :
- (h) Distinctness /)
- (i) Appurtenance / (

5.2.5 Other Indexing Processes

It is perfectly possible to draw a clear demarcation between pre-coordinate and post-coordinate indexing systems. This is how it would be done:

NOTES

NOTES

Check Your Progress

1. What governs the semantic relationships in PRECIS?
2. List the devices which are used to eliminate false drops in post-coordinate indexing.

**5.3 ANSWERS TO CHECK YOUR PROGRESS
QUESTIONS**

1. Semantic relationships in PRECIS are governed by an online thesaurus that works as a source of as well as references point for *the index*.
2. The following devices are used to eliminate false drops in post-coordinate indexing:
 - (a) Roles;
 - (b) Weighting;
 - (c) Links; and
 - (d) Use of bound terms

5.4 SUMMARY

- The processes of indexing of any given document generally takes place in two stages. These include the following:
 - o Establishing the concepts that are contained in any given document, and then
 - o Translating those concepts into the components of the given indexing language.
- C. A. Cutter was perhaps the first person to have introduced a generic set of rules for topic based indexing in his Rules for Dictionary Catalog that was published in the year 1876. Cutter defined rules intended for both specific as well as compound topic headings. Cutter considered topics as being specific and types as broad.
- J. O. Kaiser began precisely at this point where Cutter had been unable to give the guideline concerning the issue of significance. Kaiser, in his Systematic Indexing, which had been published in the year 1911, suggested that the compound topic has to be analysed by deciding the relative importance of the various parts terms of a compound topic through classificatory approach of categorization of terms.

- J. E. L. Farradane has identified the nine relationships in his Relational Indexing. These nine relationships and their respective relational operators are
 - (a) Association / ;
 - (b) Concurrence / 0
 - (c) Reaction / -
 - (d) Self-activity / *
 - (e) Equivalence / =
 - (f) Dimensional / +
 - (g) Causation / :
 - (h) Distinctness /)
 - (i) Appurtenance / (
- Syntactical relationships in PRECIS are usually dealt with by means of a set of logical rules, role operators as well as codes. These rules determine the order of terms that appear in the input string by the indexer and their utilization to create index entries by the computing system.
- The fundamental belief that leads to the creation of POPSI-Specific is that topic indexing is usually a specific purpose-based activity. But, there has usually been a custom of depending on a designer of topic indexing, language, and such total reliance has usually been found to be insufficient to meet the specific needs of topic indexing at the local level
- Chain indexing involves the process of evaluating subdivision of subject entries to arrive at an alphabetical list of subject entries which are to be indexed in the order of their specificity. It utilizes a preferred classification scheme to extract class numbers of the documents concerned for deriving subject index entries.
- Relational indexing is something that can be used to identify three dimensional objects that exist within a singular space that contains multiple objects of different sizes and contours.

NOTES

5.5 KEY WORDS

- **Semantic relationships:** It refers to the associations that there exist between the meanings of words (semantic relationships at word level), between the meanings of phrases, or between the meanings of sentences.
- **Syntactic relationships:** In linguistics, it refers to functional relationships between constituents in a clause.

- **Spider:** This refers to the computer programs which include collecting different elements related to identifying and reading Web pages including collecting data from the websites through the links mentioned in the documents.

NOTES

5.6 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What is PRECIS?
2. Differentiate between precoordinated and postcoordinate indexing.
3. Explain chain indexing.
4. What is relational indexing? Specify the nine relational operators proposed by J. E. L. Farradane.

Long-Answer Questions

1. Compare and contrast the different techniques of indexing.
2. Explain in detail, the POPSI, web indexing and NLP indexing.

5.7 FURTHER READINGS

- Alberico, R. and Micco M. 1990. *Expert systems for reference and Information retrieval*. West Port: Meckler.
- Atchison, J. & Gilchrist, A. 1972. *Thesaurus construction: a practical manual*. London: Aslib.
- Chowdhry, G.G. 2003. *Introduction to modern Information retrieval*. 2nd Ed. London: Facet Publishing.
- Cleaveland, D. B. 2001. *Introduction to Indexing and abstracting*. 3rd Ed. Englewood, Colo: Libraries Unlimited.
- Ghosh, S.B. and Biswas, S.C. 1998. *Subject Indexing systems: Concepts, methods and techniques*. Rev. ed. Calcutta: IASLIC.
- Pandey, S.K. Ed. 2000. *Library Information retrieval*. New Delhi: Anmol.

UNIT 6 INDEXING SYSTEMS AND TECHNIQUES – POST COORDINATE INDEXING

*Indexing Systems and
Techniques – Post
Coordinate Indexing*

NOTES

Structure

- 6.0 Introduction
- 6.1 Objectives
- 6.2 Post Coordinate Indexing System
 - 6.2.1 Uni-Term Indexing
- 6.3 Non-Conventional Indexing
 - 6.3.1 Citation Indexing
 - 6.3.2 Kwic and Kwoc
- 6.4 Evaluation Studies of Indexing Systems
 - 6.4.1 Crane Field-I
- 6.5 Answers to Check Your Progress Questions
- 6.6 Summary
- 6.7 Key Words
- 6.8 Self-Assessment Questions and Exercises
- 6.9 Further Readings

6.0 INTRODUCTION

This unit helps students of library and information science understand the concept of post coordinate indexing systems. This concept had actually been introduced briefly in our previous unit and is being explored further and in detail in this present unit.

6.1 OBJECTIVES

After going through this unit, you will be able to:

- Discuss the post coordinate indexing system
- Describe uni-term indexing system
- Explain the citation indexing
- Recall the KWIC and KWOC systems
- Evaluate the indexing system and its need

6.2 POST COORDINATE INDEXING SYSTEM

All the pre-coordinate indexing systems that have been discussed in the previous units have been of uni-dimensional and to a great extent based on the level of

NOTES

importance. This level of importance is in turn based on the syntactical rules that govern the given pre-coordinate indexing system. Non flexibility of the order of importance need not necessarily match the needs of all potential users who have a specific system of coordination, in spite of permutations. This means that the requirement for a multi-dimensional model that allows access to topics across any domain was the considered perspective of both users as well as indexers. The alternative was obviously to develop an index, in which the concepts of different topics would be maintained separately and not coordinated by the indexer. Coordination of basic themes or ideas may be handled precisely when users begin to search. It is important to understand that the entire process of coordination of ideas is present in both the pre-as well as post-coordinate indexing systems, although it is carried out in two separate stages. The indexing model that indicates the ability to freely coordinate various terms at the time the search is carried out, so leading to an infinite number of points of access to an article belong either to the post-coordinate or coordinate indexing system.

In a broader context, there are two main kinds of post-coordinate indexing systems, namely the Term Entry model and Item Entry model. In the Term Entry models, indexers carry out entries for an article below each of its appropriate terms which are then re-aligned in alphabetical order. A card catalogue that makes use of unit cards belongs to this type. In this model, the number of index entries carried out for a single article would largely depend on the actual number of terms that represented the central theme of the content in that particular article. These terms would be entered on the item. Uni-term Indexing and Optical Coincidence Card and usually considered to be examples of Term Entry model. On the other hand, the Item Entry model uses the contrasting approach that uses only one singular index entry for each article, that makes use of a physical form that allows access based on the entry from across a range of relevant terms. In this case, items get to be entered on the term. Edge-Notched Card or Punched Card model would be the most common example of item entry model.

The most salient characteristics of the post-coordinate indexing model would include (a) No single sequence of terms the sequence of citations is actually needed. All the terms are given equal importance; (b) ideas that make up a composite topic are maintained separately uncoordinated by the indexer and the coordination of ideas or themes is carried out by the individual user at the search level; (c) The user usually has a wide range of options to the free use of the types at the actual moment of searching. This means that the user is given the full freedom to coordinating the index terms within specific combinations or sequence that may be needed; (d) It is possible to maintain relationships at multi-dimensional levels

The post-coordinate indexing model had been introduced by Mortimer Taube in the year 1953 in an effort to arrange the research papers that had been received by the US Armed Services Technical Information Agency. This indexing model was to become extremely popular precisely because it was so simple to use.

Taube's model had been developed on "uniterm cards", which had been a card-based technique. The rules of this technique have since then been adapted and evolved much more since their inception into the system of computer-based bibliographic information retrieval. When Taube invented the technique, it had been developed on unit term. However, shortly after first inventing it, he renamed the idea of unit term to unit concept.

Most people who use search engines have a similar way or manner of using them. That is to say most people who are looking for specific information start out with a list of terms that they believe will help them find the information they are looking for. When they look at a search engine, they use the terms they have listed out in specific pre-determined or pre decided combination of terms. The search engines then use those combinations of terms in the proper way and try to make sense of what the users may actually have meant when using the combination of terms.

So when the search engine lists out links of information that seems most appropriate to the information the users are looking for, what happens is that the links of information that appear closest to the apparent meaning that may be derived from the combination of terms that had been entered into the search engine will appear at the top of the list, and this list will continue down in receding or diminishing relevance to the combination of terms that had been entered into the search engine.

This means that people who are looking for specific information on any subject or topic must use the combination of terms appropriately when they enter it in the search engine so that the list of links to the information they are looking for will appear in the correct order beginning with the most relevant links to the information closest to the information they are looking for appearing at the top, and those that follow diminishing in the correct order of diminishing importance or relevance.

This indicates that users must learn to use search engines correctly when searching for specific information on specific subjects or topics. This is because they will need to save both time and effort in their search operations.

6.2.1 Uni-Term Indexing

Students of library and information science, therefore, must learn to use uni term indexing so that they are able to access and identify the relevant information correctly, with the most appropriate and relevant information appearing at the top of the web page when they use their search engines. In order to do this, the students of library and information science must first understand the concept of the term uni term indexing systems, and learn to use uni term indexing in the correct manner. We have given the relevant information here to help students understand the term uni term, and understand and interpret the concept of uni term indexing in a manner that will most benefit them.

NOTES

NOTES

The term uni term indexing indicates using a single or singular term in the search engine that matches the information the users are actually looking for, without having to combine two or more terms being typed into the search engines. The single term or uni term, can be a single word, a single phrase or a single sentence pertaining to the information being sought.

Uniterm is actually an amalgam of a two words, namely unit and Term. The word unit denotes the use of a singular theme or idea, while the word term denotes the use of a singular word. In that context, therefore, the term uni-term is intended to represent the use of a singular worded theme or concept.

This system is built on the theory that every heading of the article could be condensed into a single or individual word that holds the ability to identify completely with the entire meaning or theme of that particular article or document. This is perhaps the reason no indexer has provided any pre-conceived or existing list of terms that would represent or depict any topic. On the contrary, every word that the user uses as a term could be identified as an independent point of access within the indexing system.

reduced for indexing purpose to a number of basic ideas capable of being represented by uniterms. Thus, in this system ready-made subj This means that instead of having to use a complete phrase or string of words in the language they find appropriate, users are able to use a single word of their choice to represent their choice of search term. This obviously removes the possibility of problems that normally arise due to errors in sequencing of citations

Let us now take a look at the benefits of the uni-term indexing technique

1. No classification system is required to organise the documents indexed by uniterm. These are arranged by their accession numbers.
2. As all the documents are arranged according to their accession number, retrieving the requisite documents is simple and easy;
3. The work at the indexing stage is simple as the coordination of subject components is done at the output stage. There is no problem of citation order at the input stage;
4. Since the system is simple and uses natural language, the users can understand it without difficulty and can make use of it easily.

Check Your Progress

1. What are Term Entry modes?
2. Who introduced the post-coordinate indexing model?
3. What is the uni-term indexing based on?

6.3 NON-CONVENTIONAL INDEXING

In this section, we will have a look at some different indexing techniques.

6.3.1 Citation Indexing

Yet another technique or system of indexing that can be used in order to access the desired information on any given subject or topic is the citation indexing system. In order to understand how to use the citation indexing systems, the users or in this present context the students of library and information science, will first need to understand the concept of citation indices in the context of library and information science. A citation index provides citations of documents that have been published on the given subject or topics most relevant to the needs of the users, and the citations least relevant or appropriate to the subject or topic that is desired or required by the users or students.

Let's us take a brief glimpse at how and where this entire concept of citation indexing actually began.

"A form of citation index is first found in 12th-century Hebrew religious literature. Legal citation indexes are found in the 18th century and were made popular by citators such as Shepard's Citations (1873). In 1960, Eugene Garfield's Institute for Scientific Information (ISI) introduced the first citation index for papers published in academic journals, first the Science Citation Index (SCI), and later the Social Sciences Citation Index (SSCI) and the Arts and Humanities Citation Index (AHCI). The first automated citation indexing was done by CiteSeer in 1997. Other sources for such data include Google Scholar and Elsevier's Scopus."

Perhaps one of the earliest documented index of citations had been index of biblical citations that had been documented or recorded in rabbinic literature, the *Mafteah ha-Derashot*, that had been credited to Maimonides and perhaps that can be traced to the 12th century. It has been arranged alphabetically through biblical phrase. Later biblical citation indices can be found in the sequence of the relevant verses. These citation indices were used both for generic as well as for legal study. The Talmudic citation index titled *En Mishpat* (1714) also indicated a symbol that helped determine whether a Talmudic decision had been ignored or over-ruled, just as in the instance of the 19th-century Shepard's Citations. Unlike modern academic citation indices, just references to one work, the Holy Bible, were found to have been actually indexed.

The term key word has been used frequently in the previous units, and will continue to be used frequently in the units that follow in this book. This is a term that relates to the most important word in a given document on a specific subject or topic, a word that is used usually in indexing systems or techniques, or in many cases to draw the attention of the readers, users of the information or the audience, to the one most important point in the document or information being presented.

NOTES

NOTES

6.3.2 Kwic and Kwoc

Taking this explanation still further, we now discuss the term KWIC. The term KWIC is a short form or an acronym for the term Key Word in Context. We will discuss here how and where this term first began to be used, and the context in which it began to be so used.

The term KWIC was first coined by Hans Peter Luhn. The system was based on a concept called keyword in titles which was first proposed for Manchester libraries in 1864 by Andrea Crestadoro.

A KWIC index is formed by sorting and aligning the words within an article title to allow each word (except the stop words) in titles to be searchable alphabetically in the index. It was a useful indexing method for technical manuals before computerized full text search became common.

For instance, a search enquiry that includes most of or all the words that appear in the heading phrase or sentence of this document ("KWIC is an acronym for Key Word In Context, the most common style used for concordance lines") as well as the Wikipedia slogan in English ("the free encyclopedia"), that may be searched as an access point to specific web sites, could retrieve a KWIC index as follows. A KWIC index would normally uses a wide field that would enable the retrieve the largest number of in context or relevant information.

"A KWIC index is a special case of a permuted index. This term refers to the fact that it indexes all cyclic permutations of the headings. Books composed of many short sections with their own descriptive headings, most notably collections of manual pages, often ended with a permuted index section, allowing the reader to easily find a section by any word from its heading. This practice, also known as KWOC ("Key Word Out of Context"), is no longer common."

KWOC

The KWOC index is in contrast to the KWIC index. The KWOC index is a short form for or an acronym for the term Key Word Out of Context being used as an index in a search engine or in a document of book. We will discuss these in detail in further units.

6.4 EVALUATION STUDIES OF INDEXING SYSTEMS

Indexing systems are usually both effective as well as efficient. While discussing the evaluation of the indexing systems, we would need to evaluate as well as compare both the efficiency as well as the effectiveness of the indexing systems. Effectiveness of an indexing system refers to the levels to which a given indexing system attains or reaches its mandated objectives. Efficiency of the given indexing system, on the other hand, would refer to how economically the indexing system attains its mandated

objectives. This means any indexing system needs to be both economically efficient and effective in order to be useful to its users.

To be able to determine the effectiveness of the given indexing system, you would actually need to determine how much a specific system is able to retrieve relevant and appropriate documents while withholding or rejecting inappropriate or irrelevant documents or information. On the other hand, determining the efficiency of the indexing system refers to calculating the cost incurred per second, the user time, response time of the search engine, etc.

An indexing or information retrieval system is usually evaluated for its effectiveness at various levels. These include the following:

- (a) System effectiveness;
- (b) Cost effectiveness; and
- (c) Cost-benefit evaluation

It is important to understand that different criteria may usually be used to evaluate any given indexing or information retrieval system. These may include the following criteria:

- (a) Recall;
- (b) Precision;
- (c) Response time;
- (d) User effort;
- (e) Form of presentation; and
- (f) Collection coverage.

Recall and precision ratios are normally used to help determine the performance of the given indexing or information retrieval system. This is how it happens:

The term Recall in the present context would refer to the property of the computing system to retrieve articles or documents most relevant to the term that has been entered into the search engine. The number of relevant articles or documents that are retrieved in proportion to the number of searches made is known as the Recall Ratio.

This recall ratio could therefore enable users to determine the exact level or degree of the rate of retrieval or recall of relevant articles or documents in any given computing system. The formula that may be used to determine of Recall ratio is:

Relevant retrieval/Total relevant*100

On the other hand, precision ratio is the ratio between the relevant retrieved

NOTES

NOTES

and total retrieved documents. It measures how precisely an indexing system functions. The formula for the calculation of the precision ratio is:

$$\text{Relevant retrieval/Total retrieval} * 100$$

An index of a single field of search is normally designed and developed through different stages that would include the following:

1. In the Navigation Pane, right-click the name of the table that you want to create the index in, and then click Design View on the shortcut menu.
2. Click the Field Name for the field that you want to index.
3. Under Field Properties, click the General tab.
4. In the Indexed property, click Yes (Duplicates OK) if you want to allow duplicates, or Yes (No Duplicates) to create a unique index.
5. To save your changes, click Save on the Quick Access Toolbar, or press CTRL+S.

6.4.1 Crane Field-I

Starting from 1953, the Cranfield tests consisted of a series of experiments run over more than a decade, continuing on into the late 1960s. ASLIB (Association for Information Management) was given the grant by National Science Foundation to carry out the investigation to be made into the comparative efficiency of four indexing systems. This work was to be undertaken at the College of Aeronautics, Cranfield, England, under the direction of the Librarian, Mr. Cyril Cleverdon. Within this period, there were two main stages. The first of these ran from 1957 to 1961 and is commonly called Cranfield 1; and the second, Cranfield 2, ran from 1963 to 1966.

Cranfield experiments were different from what constitutes the knowledge of information retrieval systems today. It is crucial to point out here that the experiments were not computerized. Instead, they were carried out manually -- human-made indexes and human-executed retrieval over them.

The Cranfield 1 as defined in the report was "an investigation into the comparative efficiency of indexing systems" and not indexing languages. In the Cranfield tests the focus was on the crucial step how the index was constructed for the information retrieval system.

Cranfield 1 found that, there was little difference in effectiveness between different indexing languages; if anything the simpler the indexing method, the better the retrieval. This unexpected result inspired in Cranfield 2 a closer examination of precisely what components of an indexing language boosted retrieval effectiveness. The recall ration and the precision ratio are said to have been the result of the development of modern evaluation indexing systems. Further, modern systems

like the SMART and TREC systems follow the Cranefield paradigm for their evaluations.

Check Your Progress

4. Mention the earliest document index of citations.
5. What is the KWIC index a special case of?
6. List the levels on which the effectiveness of an indexing system is usually evaluated.

NOTES

6.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. In the Term Entry models, indexers carry out entries for an article below each of its appropriate terms which are then re-aligned in alphabetical order.
2. The post-coordinate indexing model had been introduced by Mortimer Taube in the year 1953 in an effort to arrange the research papers that had been received by the US Armed Services Technical Information Agency.
3. The uni-term index is built on the theory that every heading of the article could be condensed into a single or individual word that holds the ability to identify completely with the entire meaning or theme of that particular article or document.
4. Perhaps one of the earliest documented index of citations had been index of biblical citations that had been documented or recorded in rabbinic literature, the Mafteah ha-Derashot, that had been credited to Maimonides and perhaps that can be traced to the 12th century.
5. A KWIC index is a special case of a permuted index. This term refers to the fact that it indexes all cyclic permutations of the headings.
6. An indexing or information retrieval system is usually evaluated for its effectiveness at various levels. These include the following
 - (a) System effectiveness;
 - (b) Cost effectiveness; and
 - (c) Cost-benefit evaluation

6.6 SUMMARY

- All the pre-coordinate indexing systems that have been discussed in the previous units have been of uni-dimensional and to a great extent based on the level of importance. This level of importance is in turn based on the

NOTES

- syntactical rules that govern the given pre-coordinate indexing system.
- The indexing model that indicates the ability to freely coordinate various terms at the time the search is carried out, so leading to an infinite number of points of access to an article belong either to the post-coordinate or coordinate indexing system.
 - In a broader context, there are two main kinds of post-coordinate indexing systems, namely the Term Entry model and Item Entry model.
 - Uniterm is actually an amalgam of a two words, namely unit and Term. The word unit denotes the use of a singular theme or idea, while the word term denotes the use of a singular word. In that context, therefore, the term uniterm is intended to represent the use of a singular worded theme or concept.
 - Uni term index system is built on the theory that every heading of the article could be condensed into a single or individual word that holds the ability to identify completely with the entire meaning or theme of that particular article or document.
 - Yet another technique or system of indexing that can be used in order to access the desired information on any given subject or topic is the citation indexing system. In order to understand how to use the citation indexing systems, the users or in this present context the students of library and information science, will first need to understand the concept of citation indices in the context of library and information science.
 - A citation index provides citations of documents that have been published on the given subject or topics most relevant to the needs of the users, and the citations least relevant or appropriate to the subject or topic that is desired or required by the users or students.
 - The term KWIC is a short form or an acronym for the term Key Word in Context. The term KWIC was first coined by Hans Peter Luhn. The system was based on a concept called keyword in titles which was first proposed for Manchester libraries in 1864 by Andrea Crestadoro.
 - A KWIC index is formed by sorting and aligning the words within an article title to allow each word (except the stop words) in titles to be searchable alphabetically in the index.
 - The KWOC index is in contrast to the KWIC index. The KWOC index is a short form for or an acronym for the term Key Word Out of Context being used as an index in a search engine or in a document of book.
 - To be able to determine the effectiveness of the given indexing system, you would actually need to determine how much a specific system is able to retrieve relevant and appropriate documents while withholding or rejecting inappropriate or irrelevant documents or information. On the other hand,

determining the efficiency of the indexing system refers to calculating the cost incurred per second, the user time, response time of the search engine, etc.

6.7 KEY WORDS

- **Post-coordinate index:** It refers to an indexing system in which the concepts of different topics would be maintained separately and not coordinated by the indexer. Coordination of basic themes or ideas may be handled precisely when users begin to search.
- **Uni-term indexing:** It indicates using a single or singular term in the search engine that matches the information the users are actually looking for, without having to combine two or more terms being typed into the search engines.
- **Citation index:** It provides citations of documents that have been published on the given subject or topics most relevant to the needs of the users, and the citations least relevant or appropriate to the subject or topic that is desired or required by the users or students.
- **KWIC index:** It is an index formed by sorting and aligning the words within an article title to allow each word (except the stop words) in titles to be searchable alphabetically in the index.

6.8 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What are the salient characteristics of the post-coordinate index?
2. List the benefits of the uni-term indexing technique.
3. Give a brief glimpse of how and where the concept of citation indexing began.
4. Explain KWIC and KWOC index.
5. What is recall ratio?

Long-Answer Questions

1. Explain the concept of post-coordinate indexing.
2. Discuss the major elements of the citation index.
3. How are indexing systems evaluated?

NOTES

NOTES

6.9 FURTHER READINGS

- Alberico, R. and Micco M. 1990. *Expert systems for reference and Information retrieval*. West Port: Meckler.
- Atchison, J. & Gilchrist, A. 1972. *Thesaurus construction: a practical manual*. London: Aslib.
- Chowdhry, G.G. 2003. *Introduction to modern Information retrieval*. 2nd Ed. London: Facet Publishing.
- Cleaveland, D. B. 2001. *Introduction to Indexing and abstracting*. 3rd Ed. Englewood, Colo: Libraries Unlimited.
- Ghosh, S.B. and Biswas, S.C. 1998. *Subject Indexing systems: Concepts, methods and techniques*. Rev. ed. Calcutta: IASLIC.
- Pandey, S.K. Ed. 2000. *Library Information retrieval*. New Delhi: Anmol.

BLOCK - III

BIBLIOGRAPHIC STANDARDS AND FORMATS

NOTES

UNIT 7 BIBLIOGRAPHIC STANDARDS

Structure

- 7.0 Introduction
- 7.1 Objectives
- 7.2 International Standard Bibliographic Description (ISBD) Standards
- 7.3 AACR
- 7.4 ISBN Standards
- 7.5 ISDN
- 7.6 ISSN
- 7.7 ISO 2709
- 7.8 Answers to Check Your Progress Question
- 7.9 Summary
- 7.10 Key Words
- 7.11 Self Assessment Questions and Exercises
- 7.12 Further Readings

7.0 INTRODUCTION

People working at any work place and providing clients any services always need to adhere to specific stands that will ensure that their services and products are of the highest quality, and that their clients receive services of the highest quality and order. In the context of libraries, as well, there are numerous standards that libraries and librarians are required to adhere to in order to provide the best services to their clients. Among all these standards, however, the most important standard would be the bibliographic standard.

Considering that this book is on library science, we have been trying to help students of library and information science understand their chosen subject in greater depth, we are devoting this unit to exploring bibliographic standards in depth.

Bibliography is related to the task of collecting data about new arrivals of books, periodicals and other objects, sorting all of these new arrivals, sorting them and categorizing them all based on specific criteria, storing them appropriately in order to make access by both the librarians as well as their clients using their services and products, easy and quick.

NOTES

Bibliographic data is also about collecting details about the people using those services and when required sharing all that information with other libraries. Obviously, the magnitude or dimension of all this data is too huge to be maintained and managed manually. This means all of these details need to be saved and stored for easy retrieval on computerized systems, or in other words, on machine related systems. Obviously, bibliographic standards would pertain to or refer to the quality of the computerization of these details, the indexing systems, the speed and ease of retrieval of those details and so on.

There are many texts and reference resources that the students of library and information science can refer to in their study of this subject. Beacher Wiggins discusses bibliographic standards in general and then focuses on the Cooperative Online Serials Program—CONSER—a cooperative program for bibliographic control that is now more than a decade old. But in this unit, we will have a look at the major bibliographic standards.

7.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain the ISBD bibliographic standard
- Discuss the development of ISBN standards
- Describe AACR and ISDN standards
- Recognize ISSN and ISO 2709 standards

7.2 INTERNATIONAL STANDARD BIBLIOGRAPHIC DESCRIPTION (ISBD) STANDARDS

It is significant to note and understand that one of the primary reasons or objectives of developing the International Standard Bibliographic Description standard was to design a common format that could help libraries and librarians exchange bibliographic details using a common format that could be understood, accessed and used by libraries and librarians across the world.

The International Standard Bibliographic Description (ISBD) would refer to a set of principles or rules that have been provided or mandated by the International Federation of Library Associations and Institutions (IFLA) intended to help develop a bibliographic description in a standard and one that is especially conducive to being used in library catalogs or bibliographies determined form that is clearly legible to and understood by human beings.

A basic condensed edited version of the ISBD was published in the year 2007 and the unabridged edition has since then been published in the year 2011, upgrading earlier individual versions of the ISBDs for monographs, older monographic publications, cartographic materials, serials as well as other ongoing resources, electronic resources, non-book resources, as also printed music. IFLA's ISBD Review Group is responsible for maintaining the ISBD.

This is intended to support IFLA's Universal Bibliographic Control program.

The ISBD outlines nine clear fields of description. Each of these fields or zones, except for the one marked 7, is comprised of multiple aspects or elements with definitive and structured demarcations. Aspects and zones that do not pertain to any specific resource have been left out of or deleted from the definition. Standard punctuation marks (colons, semicolons, slashes, dashes, commas, and periods) have been made use of in order to identify and divide the aspects from the zones. The sequence of aspects and standardized punctuation marks obviously make it easier for users to interpret bibliographic data when they are not able to understand the language that has been used in the description.

In an effort to help students understand what an International Standard Bibliographic Description would actually look like, we are giving here an example:

A manual that is intended to guide authors of research papers, theses, as well as dissertations, titled *Chicago style for students and researchers*, by Kate L. Turabian and revised by Wayne C. Booth, Gregory G. Colomb, Joseph M. Williams, and University of Chicago Press editorial staff. In its 7th edition, Chicago, at the University of Chicago Press, 2007. xviii, 466 p. : ill. ; 23 cm. (Chicago guides to writing, editing, and publishing), includes bibliographical references (p. 409-435) and index, ISBN 978-0-226-82336-2 (cloth : alk. paper) : USD35.00. — ISBN 978-0-226-82337-9 (pbk. : alk. paper) : USD17.00

NOTES

7.3 AACR

AACR is an acronym or short form for the Anglo American Cataloging Rules.

Anglo-American Cataloguing Rules (AACR) were international Library Cataloging Standards that were first published in the year 1967 that had been edited by C. Sumner Spalding, with a second edition (AACR2) that was edited by Michael Gorman and Paul W. Winkler being published in the year 1978, with later revisions (AACR2R) appearing in the years 1988 and 1998; with all further editions ending in the year 2005.

Published collaboratively by the American Library Association, the Canadian Library Association, as well as the UK Chartered Institute of Library and Information

NOTES

Professionals, the standards had been developed for the establishment of library cataloging and other such bibliographic tools. The standards are comprised of not just the physical definitions of library resources, but also possible choice of names and titles of those resources that may be used as access points.

AACR2 was issued in several print versions, including a concise edition, as well as an online version. Various translations were also available. Principles of AACR included cataloging based on the item 'in hand' rather than inferring information from external sources and the concept of the 'chief source of information' which is preferred where conflicts exist.

These rules and standards are required to be constantly reviewed and updated in order to keep pace with and match the requirements of the changing times. In this context the Anglo American Cataloging Rules have also been constantly reviewed and updated, with revised editions constantly being published.

In spite of the claim laid to be 'Anglo-American', the first edition of AACR had been published in the year 1967 in clearly definite North American and British texts. The second version that had been brought out in the year 1978 managed to consolidate the two sets of rules (following the British spelling besides bringing them in alignment with the International Standard Bibliographic Description). Libraries that want to deviate from the earlier North American content had been compelled to implement 'desuperimposition', a greater edition in the style of titles for corporate bodies.

While the 2002 revision had incorporated plenty of improvements to AACR's treatment of non-book materials, the excessive use of 21st century styles in a networked environment as well as the increased number of electronic publishing demonstrated the need for considerable revision in the cataloging code. However, plans for further editions to these rules were put on hold in the year 2005.

The global cataloging community redirected its attention to the task of developing an entirely revamped standard that was expected to follow AACR. Taking references from earlier research carried out by the International Federation of Library Associations and Institutions Functional Requirements for Bibliographic Records, the new outline was developed so as to be more flexible as well as more in alignment with the growing digital environment: Resource Description and Access (RDA) had been published in June of the year 2010. The Library of Congress, National Library of Medicine, National Agricultural Library, and many other libraries at the national level of other English-speaking countries opted for a formal manuscript of RDA, leading to the publication of a research paper on this study in June of the year 2011.

7.4 ISBN STANDARDS

ISBN is a short form for or an acronym for the term International Standard Book Number. The International Standard Book Number which allows a book or its author to be identified anywhere across the world and not just to be identified but also for the book title to be retrieved and provided or accessed. The ISBN was designed and conceived perhaps to answer the growing need of authors and books being published across the world, and becoming famous internationally, which meant any person anywhere across the world is now able to access and retrieve a book written and published in any part of the world, just by quoting the ISBN number. All publishers are required to quote a special ISBN number on all their publications to ensure that the book being so published will be identified and accessed any time, at any place in any country across the world.

The International Standard Book Number also helps libraries and book sellers identify a book desired or required by any of their clients or users.

Understanding the International Standard Book Number will help students of library and information science understand the concept behind it and why they need to use it, and how it will make things easier for them.

The International Standard Book Number (ISBN) is a special number that is given to every book that gets published anywhere in the world, and is intended to help both publishers, writers as well their audiences across the world identify a specific book, simply by tracking this ISBN code.

Going by this principle, a new or separate ISBN would need to be assigned to every new edition and variation (except in the case of reprints) of a book. For instance, an e-book, a paperback and a hard cover edition of the same book will both be assigned different ISBN. The ISBN is 13 digits long if it has been assigned on or after 1 January 2007, and 10 digits long if it had been assigned before 2007. The technique of assigning an ISBN is nation-based and differs from country to country, depending most times on the size and status of the publishing industry within a specific country.

The first such ISBN enumeration of identification had been initiated in the year 1967 based on the 9-digit Standard Book Numbering (SBN) that had been developed in the year 1966. The 10-digit ISBN style was later developed by the International Organization for Standardization (ISO) which had been published in the year 1970 as international standard ISO 2108 (It is possible to convert the SBN code to a ten digit ISBN by prefixing it with a zero).

Sometimes, it is quite likely that certain books could get published privately, so not being assigned or showing any ISBN code. In such circumstances, the International ISBN authorizing entity could take an initiative and assign an ISBN code to such books on its own volition.

NOTES

Yet another such identifying organization, the International Standard Serial Number (ISSN), identifies the publication of periodicals such as magazines; and the International Standard Music Number (ISMN) covers for musical recordings.

NOTES

Historical Background

The Standard Book Numbering (SBN) code is a 9-digit commercial book identifying method that had been invented by Gordon Foster, Emeritus Professor of Statistics at the Trinity College, Dublin, allocated to the booksellers and stationers WHSmith and others in the year 1965. The ISBN enumeration of identification was later initiated in the year in 1967 in the United Kingdom by David Whitaker (considered to be the "Father of the ISBN) followed in the year 1968 in the United States by Emery Koltay (who was later to be made director of the U.S. ISBN organization R.R. Bowker).

The 10-digit ISBN coding style was later to be designed by the International Organization for Standardization (ISO) that had then been published in the year 1970 with the title international standard ISO 2108. The United Kingdom however had continued to use the 9-digit SBN code until the year 1974. ISO has assigned the International ISBN Agency the responsibility of registering all ISBN codes across the world and the ISBN Standard had thus been designed under the aegis of ISO Technical Committee 46/Subcommittee 9 TC 46/SC 9. The online ISO entity only dates back to the year 1978.

An SBN could be changed to an ISBN version just by prefixing it with the number 0. For instance, the second edition of Mr. J. G. Reeder Returns, that had been brought out by Hodder in the year 1965, has "SBN 340 01381 8" in which 340 identifies the publisher, 01381 their serial number, while 8 would be the check digit. This could be changed to ISBN 0-340-01381-8; in which check digit does not need to be re-assigned.

Beginning 1 January 2007, ISBN codes have all comprised of 13 digits, a style that is suited to "Bookland" European Article Number EAN-13s.

Check Your Progress

1. State the primary reason or objectives of developing the ISBN.
2. Who publishes the AACR?

7.5 ISDN

The ISDN is the short form or acronym for Integrated Services Digital Network and refers to combined service provision of such features as video, data, voice

and other internet services through the traditional wiring system that used to be utilized for the traditional telephones.

Integrated Services Digital Network (ISDN) is basically a collection of communication standards that are intended for simultaneous online transfer of audio, video, data, as well as any other network services through the traditional wiring circuits of the public telephone wiring network system. It had been described for the first time in the year 1988 in the CCITT red book. Before the advent of the ISDN, the telephone networking system had been considered to be a means to relay voice, having the ability to provide some special services intended for data. The main characteristic of the ISDN is its ability to combine audio and data through the same wire, having some additional features that had not earlier been available in the traditional telephone network system. The ISDN standards describe many types of access interfaces, that include Basic Rate Interface (BRI), Primary Rate Interface (PRI), Narrowband ISDN (N-ISDN), and Broadband ISDN (B-ISDN).

ISDN could perhaps be described as a circuit-switched telephone network system, that has the additional ability to provide access to what is known as packet switched networks, intended to enable online transfer audio and data through normal telephone copper lines, in all probability leading to enhanced quality of audio over that provided by the earlier version of the telephone. It provides circuit-switched connections (both for voice as well as data), and packet-switched connections (only intended for data), in increments of 64 kilobit/s. In certain countries, ISDN has been found to provide a greater market application for purposes of internet services in which ISDN generally gives a maximum of 128 kbit/s bandwidth in both outgoing as well as incoming directions. Channel bonding is able to achieve a higher data rate; generally the ISDN B-channels of three or four BRIs (six to eight 64 kbit/s channels) are bonded.

ISDN is generally used as the network, data-link and physical layers in the context of the OSI model. In most cases, ISDN would usually be confined to usage to Q.931 and related protocols, that are a collected of signaling protocols setting up and breaking circuit-switched connections, as well as for enhanced call services for the user. These had been initiated in the year 1986.

In the context of a voice conference, ISDN enables simultaneous voice, video, and text communication between individual desktop video conferencing systems and group (room) video conferencing systems.

Let us consider the elements of ISDN or Integrated Services Digital Network. The term 'Integrated Services' refers to the ability of this network to provide a minimum two services simultaneously over or through one single line or transmission wire, that is voice, video, fax or data. This can be done by attaching multiple

NOTES

devices such as a telephone, fax machine, computer etc to the single line or transmission wire. This obviously indicates that one ISDN or integrated services digital network, line or transmission wire can meet or answer the entire range of communication requirements of most people.

NOTES

7.6 ISSN

The ISSN is an acronym or short form for the term International Standard Serial Number.

An International Standard Serial Number (ISSN) is an eight-digit serial number that is generally used to specifically identify serialized publications. Generally speaking, the ISSN is particularly helpful in telling the difference between such serialized publications that may have the same title. ISSN codes are used while ordering, cataloging, interlibrary borrowing or lending, and other such instances that may be related to serialized literature.

The ISSN system was first drafted as an International Organization for Standardization (ISO) international standard in 1971 and published as ISO 3297 in 1975. ISO subcommittee TC 46/SC 9 is responsible for maintaining the standard.

When a serialized publication having the same content has another such publication is published in more than one medium of publication, a different ISSN is allotted to each of those publications. For instance, plenty of magazines or journals are published both in the print as well as the electronic format. The ISSN coding system refers to these types as print ISSN (p-ISSN) and electronic ISSN (e-ISSN), respectively. In contrast, as described in the ISO 3297:2007, every publication in the ISSN system is also given a linking ISSN (ISSN-L), usually the same as the ISSN has given to the magazine or journal in its first published format, which links together all ISSN codes given to the magazine or journal in every medium.

Let us now consider how the code is formatted, of which we have given some examples below to help students of library and information science understand the concept of the term international standard serial number, and how it actually looks. It will also help them understand its functions a lot better.

The eight digit ISSN code is normally divided or separated into two parts of four digits each by a hyphen. In this form of an integer number, it could be represented by the first seven digits while the last code digit, which may be anything between 0 and 9 or an X, is known as a check digit. As a rule, the general form of the ISSN code (also known as the "ISSN structure" or "ISSN syntax") can be expressed as follows:

NNNN-NNNC

where N is in the set $\{0,1,2,\dots,9\}$, a digit character, and C is in $\{0,1,2,\dots,9,X\}$;

or by a Perl Compatible Regular Expressions (PCRE) regular expression:

$$\text{^\d{4}-\d{3}[\dxX]\$}$$

The ISSN of the journal *Hearing Research*, for example, is 0378-5955, where the final 5 is the check digit, that is $C=5$. To calculate the check digit, the following algorithm may be used:

Calculate the sum of the first seven digits of the ISSN multiplied by its position in the number, counting from the right—that is, 8, 7, 6, 5, 4, 3, and 2, respectively:

$$0 \cdot 8 + 3 \cdot 7 + 7 \cdot 6 + 8 \cdot 5 + 5 \cdot 4 + 9 \cdot 3 + 5 \cdot 2 \quad \{\displaystyle 0 \cdot 8 + 3 \cdot 7 + 7 \cdot 6 + 8 \cdot 5 + 5 \cdot 4 + 9 \cdot 3 + 5 \cdot 2\}$$

$$= 0 + 21 + 42 + 40 + 20 + 27 + 10 \quad \{\displaystyle = 0 + 21 + 42 + 40 + 20 + 27 + 10\}$$

$$= 160 \quad \{\displaystyle = 160\}$$

The modulus 11 of this sum is then calculated; divide the sum by 11 and determine the remainder:

$$160 \div 11 = 14 \text{ remainder } 6 = 14 + \frac{6}{11} \quad \{\displaystyle \frac{160}{11} = 14 \text{ remainder } 6 = 14 + \frac{6}{11}\}$$

If there is no remainder the check digit is 0, otherwise the remainder value is subtracted from 11 to give the check digit:

$$11 - 6 = 5 \quad \{\displaystyle 11 - 6 = 5\}$$

5 is the check digit, C .

For calculations, an upper case X in the check digit position indicates a check digit of 10 (like a Roman ten).

To confirm the check digit, calculate the sum of all eight digits of the ISSN multiplied by its position in the number, counting from the right (if the check digit is X , then add 10 to the sum). The modulus 11 of the sum must be 0.

There is an online ISSN checker that can validate an ISSN, based on the above algorithm.

NOTES

7.7 ISO 2709

NOTES

ISO 2709 is an ISO standard for bibliographic descriptions, titled Information and documentation Format for information exchange.

It is maintained by the Technical Committee for Information and Documentation (TC 9846).

We have given here a brief history of how the ISO 2708 was first conceived and developed.

A style or design intended for the exchange of bibliographic Information was developed in the 1960s under the aegis of Henriette Avram at the Library of Congress in order to encode the information printed on library cards. It was first created as ANSI/NISO Standard Z39.2, one of the first known or recorded standards for information technology, and was named the Information Interchange Format. The 1981 version of the standard was titled Documentation—Format for bibliographic information interchange on magnetic tape. The more recent edition of that standard is ANSI/NISO Z39.2-1994 (ISSN 1041-5653). The ISO standard followed the Z39.2. As of December 2008 while the current standard is known as the ISO 2709:2008.

The ISO 2709 has three basic sections, which we have explained briefly here. These sections constitute the basic structure of the ISO 2709.

- Record label—the first 24 characters of the record. This is the only portion of the record that is fixed in length. The record label includes the record length and the base address of the data contained in the record. It also has data elements that indicate how many characters are used for indicators and subfield identifiers. (See Variable fields, below)
- Directory—the directory provides the entry positions to the fields in the record, along with the field tags. A directory entry has four parts and cannot exceed twelve characters in length:
 - Field tag (3 characters)
 - Length of the field (4 characters)
 - Starting character position of the field (5 characters)
 - (Optional) Implementation-defined part
- Datafields (Variable fields)—a string containing all field and subfield data in the record
- Record separator—a single character (IS₃ of ISO 646)

Check Your Progress

3. What do the ISDN standards describe?
4. How and when was the ISSN first drafted and published?

NOTES

7.8 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. One of the primary reasons or objectives of developing the International Standard Bibliographic Description standard was to design a common format that could help libraries and librarians exchange bibliographic details using a common format that could be understood, accessed and used by libraries and librarians across the world.
2. The AACR is published collaboratively by the American Library Association, the Canadian Library Association, as well as the UK Chartered Institute of Library and Information Professionals. The standards had been developed for the establishment of library cataloging and other such bibliographic tools.
3. The ISDN standards describe many types of access interfaces, that include Basic Rate Interface (BRI), Primary Rate Interface (PRI), Narrowband ISDN (N-ISDN), and Broadband ISDN (B-ISDN).
4. The ISSN system was first drafted as an International Organization for Standardization (ISO) international standard in 1971 and published as ISO 3297 in 1975.

7.9 SUMMARY

- It is significant to note and understand that one of the primary reasons or objectives of developing the International Standard Bibliographic Description standard was to design a common format that could help libraries and librarians exchange bibliographic details using a common format that could be understood, accessed and used by libraries and librarians across the world.
- The International Standard Bibliographic Description (ISBD) would refer to a set of principles or rules that have been provided or mandated by the International Federation of Library Associations and Institutions (IFLA) intended to help develop a bibliographic description in a standard and [re- that is especially conducive to being used in library catalogs or bibliographic determined form that is clearly legible to and understood by human beings,

NOTES

- AACR is an acronym or short form for the Anglo American Cataloging Rules. Anglo-American Cataloguing Rules (AACR) were international Library Cataloging Standards that were first published in the year 1967 that had been edited by C. Sumner Spalding, with a second edition (AACR2) that was edited by Michael Gorman and Paul W. Winkler being published in the year 1978, with later revisions (AACR2R) appearing in the years 1988 and 1998; with all further editions ending in the year 2005.
- AACR is published collaboratively by the American Library Association, the Canadian Library Association, as well as the UK Chartered Institute of Library and Information Professionals, the standards had been developed for the establishment of library cataloging and other such bibliographic tools. The standards are comprised of not just the physical definitions of library resources, but also possible choice of names and titles of those resources that may be used as access points.
- ISBN is a short form for or an acronym for the term International Standard Book Number. The International Standard Book Number which allows a book or its author to be identified anywhere across the world and not just to be identified but also for the book title to be retrieved and provided or accessed.
- The ISBN was designed and conceived perhaps to answer the growing need of authors and books being published across the world, and becoming famous internationally, which meant any person anywhere across the world is now able to access and retrieve a book written and published in any part of the world, just by quoting the ISBN number.
- The ISDN is the short form or acronym for Integrated Services Digital Network and refers to combined service provision of such features as video, data, voice and other internet services through the traditional wiring system that used to be utilized for the traditional telephones.
- The main characteristic of the ISDN is its ability to combine audio and data through the same wire, having some additional features that had not earlier been available in the traditional telephone network system. The ISDN standards describe many types of access interfaces, that include Basic Rate Interface (BRI), Primary Rate Interface (PRI), Narrowband ISDN (N-ISDN), and Broadband ISDN (B-ISDN).
- An International Standard Serial Number (ISSN) is an eight-digit serial number that is generally used to specifically identify serialized publications. Generally speaking, the ISSN is particularly helpful in telling the difference between such serialized publications that may have the same title. ISSN codes are used while ordering, cataloging, interlibrary borrowing or lending, and other such instances that may be related to serialized literature.

- When a serialized publication having the same content has another such publication is published in more than one medium of publication, a different ISSN is allotted to each of those publications.
- ISO 2709 is an ISO standard for bibliographic descriptions, titled *Information and documentation Format for information exchange*.

NOTES

7.10 KEY WORDS

- **The International Standard Bibliographic Description (ISBD):** It refers to a set of principles or rules that have been provided or mandated by the International Federation of Library Associations and Institutions (IFLA).
- **The International Standard Book Number (ISBN):** It is a special number that is given to every book that gets published anywhere in the world, and is intended to help both publishers, writers as well their audiences across the world identify a specific book, simply by tracking this ISBN code.
- **Integrated Services Digital Network (ISDN):** It is basically a collection of communication standards that are intended for simultaneous online transfer of audio, video, data, as well as any other network services through the traditional wiring circuits of the public telephone wiring network system
- **International Standard Serial Number (ISSN):** It is an eight-digit serial number that is generally used to specifically identify serialized publications.
- **ISO 2709:** It is an ISO standard for bibliographic descriptions, titled *Information and documentation Format for information exchange*.

7.11 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. How is the eight-digit ISSN code divided?
2. Why is it said that the ISDN could perhaps be described as a circuit-switched telephone network system?
3. What are the three sections in which the ISO 2709 is divided?

Long-Answer Questions

1. Explain the ISBD bibliographic standard with an example.
2. Discuss the ISBN standards and its development.

7.12 FURTHER READINGS

NOTES

- Alberico, R. and Micco M. 1990. *Expert systems for reference and Information retrieval*. West Port: Meckler.
- Atchison, J. & Gilchrist, A. 1972. *Thesaurus construction: a practical manual*. London: Aslib.
- Chowdhry, G.G. 2003. *Introduction to modern Information retrieval*. 2nd Ed. London: Facet Publishing.
- Cleaveland, D. B. 2001. *Introduction to Indexing and abstracting*. 3rd Ed. Englewood, Colo: Libraries Unlimited.
- Ghosh, S.B. and Biswas, S.C. 1998. *Subject Indexing systems: Concepts, methods and techniques*. Rev. ed. Calcutta: IASLIC.
- Pandey, S.K. Ed. 2000. *Library Information retrieval*. New Delhi: Anmol.

UNIT 8 BIBLIOGRAPHIC FORMATS

Structure

- 8.0 Introduction
- 8.1 Objectives
- 8.2 Kinds of Bibliographic Records
- 8.3 Bibliographic Standards
 - 8.3.1 MARC
 - 8.3.2 UNIMARC
 - 8.3.3 CCF
 - 8.3.4 MARC21
 - 8.3.5 MARC XML
 - 8.3.6 Dublin Core Z39.5
- 8.4 Answers to Check Your Progress Questions
- 8.5 Summary
- 8.6 Key Words
- 8.7 Self Assessment Questions and Exercises
- 8.8 Further Readings

NOTES

8.0 INTRODUCTION

The bibliographic format was originally designed to convey manuscripts as well as printed text material, computer files, maps, audio clips, video clips, computer files, and visual material. So, in that sense, a bibliographic format must ideally contain, titles of books or documents, names of authors, subjects, notes, dates of publication as well as information pertaining to the physical descriptions of each of the items or objects. In this unit, you will learn about the different Bibliographic formats and standards.

8.1 OBJECTIVES

After going through this unit, you will be able to:

- Discuss the kinds of bibliographic record
- Describe the different bibliographic standards like MARC, UNIMARC, CCF, etc.

8.2 KINDS OF BIBLIOGRAPHIC RECORDS

In that sense, a typical bibliographic format should include data elements for the following types of material or resources:

- Books (BK) - used for printed, electronic, manuscript, and microform textual material that is monographic in nature.

NOTES

- Computer files (CF) – These are used for online systems or services, computer software, numeric data, computer-oriented multimedia. Other classes of electronic resources are coded for their most significant aspect. Material may be monographic or serial in nature.
- Music (MU) - used for printed, electronic, manuscript, and microform music, as well as musical sound recordings, and non-musical sound recordings. Material may be monographic or serial in nature.
- Continuing resources (CR) - used for printed, electronic, manuscript, and microform textual material that is issued in parts with a recurring pattern of publication (e.g., periodicals, newspapers, yearbooks). (NOTE: Prior to 2002, Continuing resources (CR) were referred to as Serials (SE)).
- Visual materials (VM) - used for projected media, non-projected media, two-dimensional graphics, three-dimensional artifacts or naturally occurring objects, and kits. Material may be monographic or serial in nature.
- Maps (MP) - used for all types of printed, electronic, manuscript, and microform cartographic materials, including atlases, sheet maps, and globes. Material may be monographic or serial in nature.
- Mixed materials (MX) - used primarily for archival and manuscript collections of a mixture of forms of material. Material may be monographic or serial in nature. (NOTE: Prior to 1994, Mixed materials (MX) were referred to as Archival and manuscript YYmaterial (AM))

It is possible to set apart or distinguish bibliographic records from other types of records by using specific codes. These codes help identify the following types of bibliographic records:

1. Language or linguistic material
2. Non-musical sound recordings
3. Manuscript material in different languages
4. Musical sound recordings
5. Computer files
6. Projected media
7. Cartographic material or maps
8. Two dimensional non-projected graphic material
9. Manuscript cartographic material
10. Three dimensional artefacts or natural objects
11. Musical notations
12. Kits

13. Manuscripted music

14. Combinations of material

Fill Character

The reason why the fill character began to be used in records was to contribute to the development of a national database besides also being reliant on the national level needs that had been defined for each data aspect. The existence of a fill character in any given bibliographic record is usually meant to show that the style of the record indicates a code that has to be used but that the person who developed the record probably decided not to make an effort to provide it.

NOTES

8.3 BIBLIOGRAPHIC STANDARDS

Let's discuss the varied bibliographic standards in this section.

8.3.1 MARC

A MARC record usually engages three basic aspects, specifically the record structure, the content allocation, and the data that is contained in the record.

The Network Development and MARC Standards Office is a center that is intended to encourage library and information network standards and planning. This center is located within the Library of Congress. This center had been set up in the year 1976 in an effort to provide motivation for networking activities at the Library of Congress. The office had later been expanded in the year 1984 in order to include MARC standards' needs. Thus, Currently the faculty is engaged in multiple areas of network development such as:

Standards, that are normally considered fundamental to the efficient, long-lasting exchange with other systems that could include those for Bibliographic Framework Initiative (BIBFRAME) and Machine-Readable Cataloging (MARC), as well as SRU and Z39.50 information retrieval protocols.

Planning, that makes use of the design and development of detailed samples and specifications and experimenting their use with other organizations as also within internal Library of Congress units.

Coordination and testing the implementation that uses the standards development and planning right up to the final stage through the final implementation of operational networking systems.

8.3.2 UNIMARC

The UNIMARC Bibliographic format was first created and proposed by IFLA in 1977, with the title UNIMARC: Universal MARC format. A second version was

NOTES

published in 1980 followed, in 1983, by the UNIMARC Handbook. Finally, in 1987, the UNIMARC Manual was published. The second edition, entitled UNIMARC Manual: bibliographic format was published in 1994 in a loose-leaf binder to facilitate updates. Five updates were issued in 1996, 1998, 2000, 2002 and 2005. The third and current edition was published in 2008, in book format.

UNIMARC was originally designed to be a switching format to enable the wider exchange of bibliographic data. UNIMARC has been developed by a number of countries to become a production format. It has also been used by UNESCO for its library products, mainly to help developing countries move to automated library management systems and standard data formats. UNIMARC currently consists of a set of four formats:

- Bibliographic
- Authorities
- Classification
- Holdings

The feature that differentiates UNIMARC other specifications is for its consistency and rigor. Coherent numerical blocks are used grouping the labels that identify the type of data contained in a UNIMARC field.

8.3.3 CCF

CCF refers to the Common Communication Format which was first developed in 1984 by the UN Ad Hoc group to facilitate the exchange of bibliographic information between different libraries and organizations.

Generally, different organizations keep different formats for input of data, long term storage, retrieval and display of information. While exchanging information, there must be agreed upon a common format so that compatibility issues do not arise.

The CCF is based on certain principles:

- Standards conform of ISO 2709
- Core record consists the mandatory data elements describing the bibliographic information
- Additional optional data may also be present in the standard format
- A common and standard techniques is used to connect different levels, links and relationship of bibliographic data

Each CCF record contains a record label, directory, data field and record separator.

8.3.4 MARC21

Description of items in the library catalogues are maintained by different digital formats like the MARC (Machine-Readable Cataloging) standards. This system was developed in the 1960s by Henriette Avram who was an American computer scientist working with the Library of Congress. The purpose of developing this digital format was to help different computers and libraries create records that could be understood and read by them all. This format was given a national standard by the year 1971 in the US for the sharing of bibliographic data. And these become an international standard in the next two years, but there are different versions of the same. The most famous version of the MARC format used worldwide is MARC 21. The synchronization of the US and Canadian formats along with the use of UNIMARC in 1999 was the event which marked the creation of the MARC 21 format. The scope of the MARC 21 family of standards is no more restricted to bibliographic records and is now more widened in that it now comprises of formats for authority records, holdings records, classification schedules, and community information.

While the above passage was a more collective description, we may describe the two terms individually also. Let us see how this achieved.

Field designation

Every field within a MARC record is intended to give specific information regarding the given item that the record is defining, such as the author, title, publisher, date, language, media type, etc. Because it had been developed first at a time when the ability to computerize was low, and space fairly limited, MARC makes use of a simple three-digit numeric code (from 001-999) to identify each field in the record. MARC describes field 100 as the initial or basic author of a work, field 245 as the title and field 260 as the publisher, for instance.

Those fields marked higher than 008 have been further divided into sub sections with the help of a single letter or number allocation. The number or digits 260, for example, is further divided into subfield "a" for the place of publication, "b" for the name of the publisher, and "c" for the date of publication.

A MARC record foundation could perhaps be defined in the following words:

MARC records are usually saved and communicated in the form of binary files, usually having multiple MARC records connected together into a single file. MARC makes use of the ISO 2709 standard to help describe the status of each individual record. This includes a marker that is used to show where each record starts and ends, as well as a set of characters at the beginning of each record that provide a directory for locating the fields and subfields within the record.

NOTES

NOTES

In the year 2002, the Library of Congress created the MARCXML schema as an alternative record base, that was intended to enable MARC records to be represented in XML; These fields have basically remained the same, but those fields are now described in the record in XML mark up. Libraries usually show or display their records as MARCXML through a web service, many times following the SRU or OAI-PMH standards.

The MARC 21 styles are standards intended to be used to represent and then communicate the bibliographic as well as other related information in a computerized and online format

MARC 21 was first created in an effort to reshape the original MARC record in a form that would be suitable for the 21st century and to ensure that it would be more accessible to the global computing community. MARC 21 has formats therefore that would be compatible to the following five kinds of data: Bibliographic Format, Authority Format, Holdings Format, Community Format, and Classification Data Format.

At present MARC 21 has been implemented very effectively by numerous famous libraries across the world, such as The British Library, the European Institutions and the major library institutions in the United States, and Canada.

MARC 21 could be said to be the creation of or the result of the amalgamation of the United States and Canadian MARC formats (USMARC and CAN/MARC). MARC21 has been founded on the NISO/ANSI standard Z39.2, that enables users of various software products to interact and connect with one another and thereby exchange data.

MARC 21 facilitates the use of two sets of characters, either MARC-8 or Unicode registered as UTF-8. MARC-8 is built around the ISO 2022 and enables the use of Hebrew, Cyrillic, Arabic, Greek, and East Asian scripts. MARC 21 in UTF-8 format could therefore be said to enable all the languages that are globally supported by Unicode.

The structure of MARC records is an installation of national as well as international standards, such as Information Interchange Format (ANSI Z39.2) and Format for Information Exchange (ISO 2709).

Content allocation, the relevant codes and principles set up to help recognize clearly and then further classify the data aspects contained in a record and finally to support the use of the recorded data, is described in the MARC 21 formats.

The information contained in most data elements has been described by standards beyond the formats, such as Anglo-American Cataloguing Rules, Library of Congress Subject Headings, National Library of Medicine Classification. The information contained in other data elements, has also been described in the MARC 21 formats.

A MARC 21 format could be described as a set of codes and content allocators that are intended to help encoding computerized or digital records. Formats are shaped and described for five kinds of data: bibliographic, holdings, authority, classification, and community information.

MARC 21 Format for Bibliographic Data is comprised of format definitions for encoding data elements that are necessary for the purpose of decoding, retrieval, and controlling different forms of bibliographic resources. The MARC 21 Format that is intended for Bibliographic Data is a combined format designed to help identify and describe the various forms of bibliographic resources. MARC 21 descriptions are intended for books, magazines, computer files, maps, music, visual materials, and mixed resources. With the entire integration of the earlier discrete bibliographic formats, continuous description as well as usage are maintained for various types of material.

MARC 21 Format for Holdings Data consists of specifications for format intended for encoding data elements related to holdings and location data used for all types of resources.

MARC 21 Format for Authority Data contains format specifications for encoding data elements that identify or control the content and content designation of those portions of a bibliographic record that may be subject to authority control.

MARC 21 Format for Classification Data contains format specifications for encoding data elements related to classification numbers and the captions associated with them. Classification records are used for the maintenance and development of classification schemes.

MARC 21 Format for Community Information provides format specifications for records containing information about events, programs, services, etc. so that this information can be integrated into the same public access catalogs as data in other record types.

The MARC 21 formats are managed by the Library of Congress in collaboration with a wide range of user communities.

Through management and editing, allocation of information is added to the existing content allocation is then deleted or made redundant or from formats. Content allocation is termed obsolete when it is discovered to be no longer relevant to when the data aspect concerned is not required any more. An obsolete content allocation however could sometimes continue to appear in records that had been developed before the date it was made obsolete. Obsolete content designators are not used in new records. A deleted content designator is one that had been reserved in MARC 21 but had not been defined or one that had been defined but it is known with near certainty that it had not been used.

NOTES

NOTES

The formats could therefore consist of exceptions to the principles because of early format creation decisions. While most of the exceptions could have been made obsolete, some of them remain because of the need to manage upward compatibility of the formats in current times.

8.3.5 MARC XML

The MARCXML is an XML schema based on the common MARC21 standards. The Library of Congress was the one which created and adopted it and others as to facilitate the networked access and sharing of bibliographic information between different libraries and users. This standard is used by different systems to parse as an aggregation format as it is in software packages such as MetaLib, though that package merges it into a wider DTD specification.

The MARCXML primary design goals included:

- The presence of validation tools
- Following a simple to use schema
- Potential to be extended and flexible for changes
- Create minimal losses and reversible conversions from MARC
- Use XML conversions to update the MARC records and data conversions

The MARCXML system uses the MARC records to form well designed XMLs and back. It does so by including only the basic structure of the MARC records in the XML format. A small DTD is present but the converter doesn't necessitate it in order to rebuild the record. So what essentially happens is that the XML file becomes available for editing or reviewing in a browser or a word processor which is then recompiled into a structurally sound MARC record. This is done through the help of a small yet potent XML parser program built in the converter itself.

8.3.6 Dublin Core Z39.5

This is a small set of vocabulary terms that can be used to describe digital resources (video, images, web pages, etc.), as well as physical resources such as books or CDs, and objects like artworks comprised in a **Dublin Core Schema**. The full set of Dublin Core metadata terms can be found on the Dublin Core Metadata Initiative (DCMI) website.

The Dublin Core Metadata Initiative (DCMI) supports shared innovation in metadata design and best practices across a broad range of purposes and business models.

DCMI does this by:

- o Managing long term curation and development of DCMI specifications and metadata terms namespaces;
- o Managing ongoing discussion of current DCMI-wide work themes;
- o Setting up and managing international and regional events;
- o Curation and open availability of meeting assets including proceedings, project reports and meeting minutes;
- o Creation and delivery of training resources in metadata best practices including tutorials, webinars and workshops; and
- o Coordinating the global community of DCMI volunteers.

The original set of 15 classic metadata terms, known as the Dublin Core Metadata Element Set (DCMES), is endorsed in the following standards documents:

- IETF RFC 5013
- ISO Standard 15836-1:2017
- NISO Standard Z39.85

Dublin Core metadata may be used for multiple purposes, from simple resource description to combining metadata vocabularies of different metadata standards, to providing interoperability for metadata vocabularies in the linked data cloud and Semantic Web implementations.

Check Your Progress

1. What are the three basic aspects engaged in a MARC record?
2. When was the CCF first developed?
3. Who developed the MARC standards?

8.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. A MARC record usually engages three basic aspects, specifically the record structure, the content allocation, and the data that is contained in the record.
2. CCF which refers to the Common Communication Format was first developed in 1984 by the UN Ad Hoc group to facilitate the exchange of bibliographic information between different libraries and organizations.
3. MARC (Machine-Readable Cataloging) standards system was developed in the 1960s by Henriette Avram.

NOTES

8.5 SUMMARY

NOTES

- The bibliographic format was originally designed to convey manuscripts as well as printed text material, computer files, maps, audio clips, video clips, computer files, and visual material.
- It is possible to set apart or distinguish bibliographic records from other types of records by using specific codes: Books (BK), Music (MU), Computer files (CF), Visual materials (VM), Mixed materials (MX), Maps (MP), etc.
- These codes help identify the following types of bibliographic records:
 - o Language or linguistic material
 - o Non-musical sound recordings
 - o Manuscript material in different languages
 - o Musical sound recordings
 - o Computer files
 - o Projected media
 - o Cartographic material or maps
 - o Two dimensional non-projected graphic material
 - o Manuscript cartographic material
 - o Three dimensional artefacts or naural objects
 - o Musical notations
 - o Kits
 - o Manuscripted music
 - o Combinations of material
- The reason why the fill character began to be used in records was to contribute to the development of a national database besides also being reliant on the national level needs that had been defined for each data aspect. The existence of a fill character in any given bibliographic record is usually meant to show that the style of the record indicates a code that has to be used but that the person who developed the record probably decided not to make an effort to provide it.
- A MARC record usually engages three basic aspects, specifically the record structure, the content allocation, and the data that is contained in the record.
- CCF refers to the Common Communication Format which was first developed in 1984 by the UN Ad Hoc group to facilitate the exchange of bibliographic information between different libraries and organizations.

- MARC records are usually saved and communicated in the form of binary files, usually having multiple MARC records connected together into a single file. MARC makes use of the ISO 2709 standard to help describe the status of each individual record.
- MARC 21 was first created in an effort to reshape the original MARC record in a form that would be suitable for the 21st century and to ensure that it would be more accessible to the global computing community. MARC 21 has formats therefore that would be compatible to the following five kinds of data: Bibliographic Format, Authority Format, Holdings Format, Community Format, and Classification Data Format.
- The is a small set of vocabulary terms that can be used to describe digital resources (video, images, web pages, etc.), as well as physical resources such as books or CDs, and objects like artworks is comprised in a Dublin Core Schema.

NOTES

8.6 KEY WORDS

- **Bibliographic format:** It contains titles of books or documents, names of authors, subjects, notes, dates of publication as well as information pertaining to the physical descriptions of each of the items or objects.
- **Fill character:** It is a character transmitted solely for the purpose of consuming time. It does this by filling a timeslot on a data transmission line which would otherwise be forced to be idle (empty). In this way, fill characters provide a simple way of timing required idle times.

8.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What are the different types of bibliographic records?
2. What does the existence of fill character in a bibliographic record show?
3. What are the principles on which CCF is based?
4. Why is the Dublin Core Schema used?

Long-Answer Questions

1. What are the different bibliographic codes used to specify different bibliographic records?
2. Explain the MARC21 bibliographic record.
3. How is MARC XML different from MARC standard?

8.8 FURTHER READINGS

NOTES

Alberico, R. and Micco M. 1990. *Expert systems for reference and Information retrieval*. West Port: Meckler.

Atchison, J. & Gilchrist, A. 1972. *Thesaurus construction: a practical manual*. London: Aslib.

Chowdhry, G.G. 2003. *Introduction to modern Information retrieval*. 2nd Ed. London: Facet Publishing.

Cleaveland, D. B. 2001. *Introduction to Indexing and abstracting*. 3rd Ed. Englewood, Colo: Libraries Unlimited.

Ghosh, S.B. and Biswas, S.C. 1998. *Subject Indexing systems: Concepts, methods and techniques*. Rev. ed. Calcutta: IASLIC.

Pandey, S.K. Ed. 2000. *Library Information retrieval*. New Delhi: Anmol.

BLOCK - IV

INFORMATION RETRIEVAL SYSTEM

*Information Retrieval
System: Basics*

**UNIT 9 INFORMATION RETRIEVAL
SYSTEM: BASICS**

NOTES

Structure

- 9.0 Introduction
- 9.1 Objectives
- 9.2 Structure, Functions and Components
- 9.3 Answers to 'Check Your Progress' Questions
- 9.4 Summary
- 9.5 Key Words
- 9.6 Self Assessment Questions and Exercises
- 9.7 Further Readings

9.0 INTRODUCTION

In earlier units we have described the science of information retrieval, its significance and function in a massive environment such as libraries, where massive bodies of data are required to be collected, sorted and categorized, stored and most importantly, retrieve in a split second without any changes or modifications.

In this unit, you will explore this concept in greater detail and depth so that you can increase the understanding of what the information retrieval concept is and how the information retrieval system functions and can be used for the maximum effectiveness in environments such as libraries across the world.

9.1 OBJECTIVES

After going through this unit, you will be able to:

- Discuss the meaning of information retrieval system
- Describe the structure and components of the information retrieval system

9.2 STRUCTURE, FUNCTIONS AND COMPONENTS

Information retrieval (IR) is referred to as the act of obtaining information from computing systems. In simple terms, it can be defined as the manner in which

NOTES

online resources in respond to a requirement for specific information from across a wide range of information resources. Complete content or articles or other content-based indexing can form the foundation for people's queries. Information retrieval is the science of searching for relevant information in a single article, or searching for relevant information independently, and also searching for metadata that describes data, and for databases of texts, images or audio clips.

Automated information retrieval systems are normally used to minimize what has been called an overload of information or excessive information. An IR system is a software that allows users access to books, journals and other articles, stores them and manages the articles. Web search engines are the most visible IR applications.

Students and other users of the information retrieval systems would be greatly benefited by a general overview of an information retrieval system in order to enhance and enrich their understanding of the concept and function. So, here we go:

An object is known as an entity that is represented by relevant information in a collection of content or databases. User searches are matched to the database information. However, in contrast to traditional SQL searches for a database, in information retrieval the information retrieved may or may not match the search, so results are traditionally ranked. This ranking of results is a key difference between information retrieval searching method and database searching.

Information retrieval is based on what the data objects may be used for examples images audio, mind maps or videos. Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates or metadata.

Most IR systems develop a numeric evaluation of how appropriately each object in the database matches the search, and rank the objects according to this value. The top ranking objects are then displayed to the user. The process may then be edited if the user wishes to narrow down the search.

In the next few paragraphs, we will explore the history of the information retrieval systems in order to enable students of library and information science understand the concept of information retrieval systems better, and use it in their work places to greater effectiveness and more successfully and efficiently.

There is ... a machine called the Univac ... whereby letters and figures are coded as a pattern of magnetic spots on a long steel tape. By this means the text of a document, preceded by its subject code symbol, can be recorded ... the machine ... automatically selects and types out those references which have been coded in any desired way at a rate of 120 words a minute " — *J. E. Holmstrom, 1948*

The History

The concept of using computers in order to search for relevant passages of information was popularized in the article *As We May Think* written by Vannevar Bush in the year 1945. It may appear that Bush had been inspired by patents for a ‘statistical machine’ that had been filed by Emanuel Goldberg in the 1920s and ‘30s that looked for documents that had been stored on film. The initial definition of a computer looking for information was described by Holmstrom in the year 1948, giving details of an early description of the Univac computer. Automated information retrieval systems were introduced in the 1950s: one of them apparently had even appeared in *Desk Set*, the 1957 romantic comedy.

In the 1960s, the first large information retrieval research group was formed by Gerard Salton at Cornell. By the 1970s many other retrieval strategies had been shown to perform successfully on small text content such as the Cranfield collection (plenty of thousand documents). Large-scale retrieval systems, such as the Lockheed Dialog system, came into use sometime in the early 1970s.

In the year 1992, the US Department of Defense in collaboration with the National Institute of Standards and Technology (NIST), sponsored the Text Retrieval Conference (TREC) as part of the TIPSTER text program. The primary goal of this program was to delve deeper into the information retrieval community by providing the infrastructure that would be required for assessment of retrieval methodologies on a very large text collection. This catalyzed re text search on methods that scale to huge corpora. The introduction of web search engines has boosted the need for very large scale retrieval systems even further.

A suitable representation is an essential pre-requirement of the process of retrieving the required articles using the IR strategies A certain sample can be utilized towards this end. Two dimensions are followed on which IR strategy is based: the mathematical basis and the properties of the model.

Let us have a look at both these types, in detail.

Mathematical Basis

- *Algebraic models*: These models are characterized by their use of the vectors, tuples and matrices to describe documents and queries. The similarity query and document vector is represented through the scalar value. The following are the popular models of this category:
 - o Vector space model
 - o (Enhanced) Topic-based Vector Space Model
 - o Latent semantic indexing a.k.a. latent semantic analysis

NOTES

NOTES

- o Generalized vector space model
- o Extended Boolean model
- *Set-theoretic* models: As the name suggests word and phrases represent documents here. Set-theoretic operations on those sets are derived. Some common of these categories are:
 - o Fuzzy retrieval
 - o Standard Boolean model
 - o Extended Boolean model
- *Probabilistic models*: In this category, the method of document retrieval is viewed as a probabilistic inference. Hence, probabilities are used to define the manner in which search queries are used to locate documents. Bay's theorem is one of the probabilistic thoerems often used in these models. Others included are:
 - o Uncertain inference
 - o Probabilistic relevance model on which is based the okapi (BM25) relevance function
 - o Language models
 - o Divergence-from-randomness model
 - o Binary Independence Model
 - o Latent Dirichlet allocation
- *Feature-based retrieval*: The vectors of values of the feature functions are used to for document retrieval in this model. Further, these are employed recognize the best combination to arrive at a single relevance score, typically by learning to rank methods. These functions are random functions of document and query, and can easily incorporate almost any other retrieval model as just another feature.

Properties of the Model

- *Models with transcendent term interdependencies* permit a representation of interdependencies between terms, but they isplaye the conjectures about the manner in which the interdependency between two terms is described. An external source for the degree of interdependency between two terms is often relied upon in this model. (For example, a human or sophisticated algorithms.)
- *Models without term-interdependencies* uses the philosophy of treating independently the varied terms/words. This fact is usually isplayed in vector

space models by the orthogonality assumption of term vectors or in probabilistic models by an independency assumption for term variables.

- *Models with immanent term interdependencies* permit a representation of interdependencies between terms. However the degree of the interdependency between two terms is defined by the model itself. It is usually directly or indirectly derived (e.g. by dimensional reduction) from the co-occurrence of those terms in the whole set of documents.

NOTES

Performance and correctness measures

The method by which the valuation of an information retrieval system is made of how understanding how appropriately a system matches the information requirements of its users is called the performance assessment. This assessment considers a collection of articles that are to be searched and a search query. There are different methods by which assessment is done and this includes the Traditional assessment metrics, developed for Boolean retrieval or top-k retrieval, which includes precision and recall. All measures assume the foundational notion of relevancy, that is to say that every article is known to be either relevant or non-relevant to a specific search. In practice, however, searches may be incorrectly termed and there could be varying shades of relevancy.

Check Your Progress

1. What forms the foundation of people's queries?
2. State the most visible IR applications.
3. Who popularized the concept of using the computers in order to search for relevant passages of information?

9.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Complete content or articles or other content-based indexing can form the foundation for people's queries.
2. Web search engines are the most visible IR applications.
3. The concept of using computers in order to search for relevant passages of information was popularized in the article *As We May Think* written by Vannevar Bush in the year 1945.

NOTES

9.4 SUMMARY

- Information retrieval (IR) is referred to as the act of obtaining information from computing systems. In simple terms, it can be defined as the manner in which online resources respond to a requirement for specific information from across a wide range of information resources.
- Information retrieval is the science of searching for relevant information in a single article, or searching for relevant information independently, and also searching for metadata that describes data, and for databases of texts, images or audio clips.
- An IR system is a software that allows users access to books, journals and other articles, stores them and manages the articles. Web search engines are the most visible IR applications.
- Most IR systems develop a numeric evaluation of how appropriately each object in the database matches the search, and rank the objects according to this value. The top ranking objects are then displayed to the user. The process may then be edited if the user wishes to narrow down the search.¹
- The concept of using computers in order to search for relevant passages of information was popularized in the article *As We May Think* written by Vannevar Bush in the year 1945. It may appear that Bush had been inspired by patents for a 'statistical machine' that had been filed by Emanuel Goldberg in the 1920s and '30s that looked for documents that had been stored on film. The initial definition of a computer looking for information was described by Holmstrom in the year 1948, giving details of an early description of the Univac computer. Automated information retrieval systems were introduced in the 1950s: one of them apparently had even appeared in *Desk Set*, the 1957 romantic comedy.
- In the 1960s, the first large information retrieval research group was formed by Gerard Salton at Cornell. By the 1970s many other retrieval strategies had been shown to perform successfully on small text content such as the Cranfield collection (plenty of thousand documents). Large-scale retrieval systems, such as the Lockheed Dialog system, came into use sometime in the early 1970s.
- A suitable representation is an essential pre-requirement of the process of retrieving the required articles using the IR strategies. A certain sample can be utilized towards this end. Two dimensions are followed on which IR strategy is based: the mathematical basis and the properties of the model.

- The method by which the valuation of an information retrieval system is made of how understanding how appropriately a system matches the information requirements of its users is called the performance assessment.

9.5 KEY WORDS

- **Information retrieval (IR):** It is referred to as the manner in which online resources in respond to a requirement for specific information from across a wide range of information resources
- **Object:** It is known as an entity that is represented by relevant information in a collection of content or databases
- **Performance assessment:** It refers to the method by which the valuation of an information retrieval system is made of how understanding how appropriately a system matches the information requirements of its users

9.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. How do IR systems use numeric evaluation?
2. Give some examples of the set-theoretic models.
3. Write a short note on performance and correctness measures.

Long-Answer Questions

1. Trace the history of the information retrieval systems.
2. Explain the two dimensions are followed on which IR strategy is based.

9.7 FURTHER READINGS

- Alberico, R. and Micco M. 1990. *Expert systems for reference and Information retrieval*. West Port: Meckler.
- Atchison, J. & Gilchrist, A. 1972. *Thesaurus construction: a practical manual*. London: Aslib.
- Chowdhry, G.G. 2003. *Introduction to modern Information retrieval*. 2nd Ed. London: Facet Publishing.
- Cleaveland, D. B. 2001. *Introduction to Indexing and abstracting*. 3rd Ed. Englewood, Colo: Libraries Unlimited.

NOTES

Ghosh, S.B. and Biswas, S.C. 1998. Subject Indexing systems: Concepts, methods and techniques. Rev. ed. Calcutta: IASLIC.

Pandey, S.K. Ed.2000 .*Library Information retrieval*. New Delhi: Anmol.

NOTES

UNIT 10 OVERVIEW OF SEARCH STRATEGY

NOTES

Structure

- 10.0 Introduction
- 10.1 Objectives
- 10.2 Search Strategy
- 10.3 Criteria of Evaluation: Recall and Precision
- 10.4 Relevance and Failure Analysis
- 10.5 Answers to Check Your Progress Questions
- 10.6 Summary
- 10.7 Key Words
- 10.8 Self Assessment Questions and Exercises
- 10.9 Further Readings

10.0 INTRODUCTION

Information retrieval systems are extremely crucial to the library system given that the cataloguing of numerous books depends on the system that is adopted by the library system. There are very many search strategies which can be used by the libraries based on which the Informal retrieval system is based. It can be from very general model to a highly efficient one. There have been varied theories based on which the information retrieval system can be evaluated. The primary idea is to analyse the recall and precision with which the retrieval system works based on the search query. But another factor is crucial to the understanding of the information retrieval system. This is the failure analysis which helps in not only detecting the failures that are occurring but also failure modes that might occur in the system. In this unit, we will discuss the concept of search strategy, the different criteria of evaluation and the idea of failure analysis and evaluation.

10.1 OBJECTIVES

After going through this unit, you will be able to:

- Discuss the concept of search strategy
- Describe the criteria of evaluation including recall and precision
- Explain the idea of failure analysis and evaluation

10.2 SEARCH STRATEGY

Language model is the simplest form of automatic information retrieval. Users enter a text or keywords that are used to search the inverted indexes of the

NOTES

document keywords. This approach retrieves documents based on the presence or absence of exact single word strings as specified by the logical representation of the query. Because of this type of searching approach, the user might miss many relevant documents because it does not capture the complete or deep meaning of the user's query. Linguistic and knowledge-based approaches have also been developed to address this problem by performing a morphological, syntactic and semantic analysis to retrieve documents more effectively. In analysis, roots and affixes are analysed to determine the part of speech (noun, verb, adjective, etc.) of the words. Next complete phrases have to be parsed using some form of syntactic analysis. Finally, the linguistic methods have to resolve word ambiguities and generate relevant synonyms based on the semantic relationships between words. The development of a sophisticated linguistic retrieval system is difficult and it requires complex knowledge based of semantic information and retrieval. Hence, these systems often require techniques that are commonly referred to as artificial intelligence or expert systems techniques.

A language model for an Information Retrieval (IR) is defined using an ordinary fuzzy language approach. The ordinary fuzzy language model is presented, and is used for modelling the ambiguity and subjectivity that appear in the information retrieval system. The user's query and Information Retrieval responses are modelled linguistically using the concept of fuzzy or unclear linguistic variables. The system accepts Boolean queries whose terms can be weighted simultaneously by means of language values according to three possible semantics.

- A symmetrical threshold semantic
- A quantitative semantic
- An important semantic

A symmetrical threshold semantic identifies a new threshold semantic used to express qualitative restrictions on the documents retrieved for a given term. It is monotone increasing in index, term, weight for the threshold values that are on the right of the mid-value and decreasing for the threshold values that are on the left of the mid-value.

A quantitative semantic is a new semantic proposal introduced to express quantitative restrictions on the documents retrieved for a term, i.e., restrictions on the number of documents that must be retrieved containing that term.

An important semantic is the usual semantic of relative importance that has an effect when the term is in a Boolean expression.

A bottom-up evaluation mechanism of queries is presented that coherently integrates the use of the three semantics and satisfies the separate property. The advantage of this Information Retrieval System with respect to others is that users can express through language with different restrictions on the desired documents, incorporating more flexibility or elasticity in the user-Information Retrieval System.

For many years, people have realized the importance of retrieving and searching information. With the arrival of computers, it became possible to store large amounts of information, and finding useful information from such collections became our necessity. The field of Information Retrieval (IR) was born in the 1950s out of this necessity. Over the last forty years, the field has matured considerably.

Search Strategy

Building an effective Information Retrieval system requires proper knowledge of information retrieval techniques and tools used for it. Many strategies are used for providing better services to the user. Two important factors to be considered while providing information retrieval system are as follows:

- The indexing knowledge is also required for providing better strategies of IR. The paper deals with importance, strategies of IR and combination of strategies for better IR. Online Library catalogues, Indexing tools and Web portals are discussed in the paper.
- Ranking and Impact of retrieved documents are also to be considered in the IR system.

Keywords also play an important role in information retrieval and strategies. In the age of information explosion, the user is sometimes confused with overload of information. The problem of information overload may be addressed with the help of efficient Information Retrieval techniques. In most of the libraries, the online catalogues are available to users, but still users do not know about the refined strategies of Information Retrieval.

Information Retrieval enables the user with more options of information retrieval. Various researches have been conducted and have provided the base for efficient online library catalogue searching.

Web search portals are another widely used example of online information retrieval. These kinds of strategies of IR are not treated as permanent strategies but temporary in nature. Due to the lack of unified information retrieval strategies, it directly affects the training of users by library professionals and designing of Information Retrieval System (IRS) by them.

There are different kinds of strategies, ranging from general to efficient strategies. For building an information retrieval system, various design system ranging from weighing system to morphological normalization required. Nowadays, combination of strategies is widely used. The combination of various strategies can lead to significant improvement in retrieval effectiveness. The researchers are focusing more on the relative impact of retrieving more relevant documents with improved rankings.

NOTES

NOTES

There are two things that are most prevalent. One strategy is based on relatedness of authors. This strategy includes tendency of authors to cite other authors that do related works and could, therefore, be of high relevance for persons who are finding related information. These situations are mainly emergent situations where behaviour of information authors matters. In this type of strategy, the user is informed about the various documents used by the authors for his literature. These documents are used by the author for related literature on the topic. This can be materialised by using footnotes given in books while searching in the library cataloguing system. This can also be used by footnotes given in book which may be completely retrieved by abstracting and Indexing system. Another related form of strategy is finding regularity in storing collection or information sources. This strategy deals with the regularity of indexed collection, where various simple keywords are stored in a successive manner. It mainly narrows down the search. This is known as successive fraction strategy. Another form of strategy includes enhancement of process of retrieval. It is difficult that the search process will organise the information in various categories as per the user's requirement. Therefore, the user should be aware for organizing the results found from searching. A careful review of result should be done.

Therefore, this strategy advises the user to organize the result of a search in named hierarchy to enable easy access in the future. Other strategies are also used by users, where other ways of enhancing retrieval process are suggested depending on the breadth and depth of the large searchers, keeping the context of topic intact. Observed improvement is entirely due to better ranking of relevant documents. Combinations of runs are generally made by an index using different morphological normalization using the same retrieval and weighting schemes.

Utility of Strategies: These strategies can be utilised in 3 types of Information Retrieval systems:

1. Online Library Catalogue
2. Abstracting and Indexing System
3. Web Search Portals

Very few IR strategies delivers relevant document, whereas the combination of strategies gives tremendous results. Conceptual IRS use domain knowledge and domain indices to solve the query of the user. In a search where no direct match of query is found, the relationship with others is used to enhance the Recall. The query might be found in this, or to narrow down the search, the document may be best suited to the query, which is known as precision. It may be said that pieces of information in the document are somehow important in implicit rules. So, the first strategy may increase Recall by having more relevant documents and the other strategy may increase the precision by better ranking of documents. The

successful combination depends on the different strategies retrieving relevant documents.

The Information Retrieval Process mainly depends on the strategies used for retrieval. The combination of strategies is more successful and more widely used.

This approach improves the retrieval process in a noise environment.

NOTES

10.3 CRITERIA OF EVALUATION: RECALL AND PRECISION

Search operations in an ISAR system can be performed manually or in an automated environment. A library shelf-list or card catalogue is a good example of a manual search system. OPAC or Digital Libraries are examples of automated systems. According to experts like Claverdon, an ISAR system should be evaluated on the following criteria:

- The extent to which the system includes relevant matter
- The time taken from when the search request is made, to the time an answer is given
- The look and feel of the presentation output
- The effort required by the user to get the answers to his search requests
- The percentage of relevant material that is actually retrieved in a search request
- The percentage of material that is relevant vis-à-vis the material that is actually retrieved in a search request

Further analysis of Claverdon's view gives following check points on which any ISAR can be evaluated:

- Coverage
- Cost Benefit Analysis
- Time

The foremost issue that should be considered is the time a user takes to get a satisfactory answer to his question. The system should be such that a user gets her/his answer immediately. Any search on an ISAR system will depend on and vary from user to user and time measurement should be supported by a long observation.

NOTES

Other evaluation criteria are as follows:

- Completeness and Relevance
- Novelty Ratio
- Noise

Critical Aspects of Recall and Precision

Without scanning the entire collections completely, the total number of relevant documents in collections cannot be determination. This is a herculean task. In addition, the degree of relevancy should also be considered when measuring the relevancy. We have seen earlier that recall and precision are inversely related. So, both cannot be at the highest degree at the same time. Following are some factors that influence recall and precision:

- Queries that do not match the information needs
- Indexing factors
- Search strategy factors
- Vocabulary factors

The following hindrances have to be removed to ensure relevant material is retrieved from an ISAR system, which will also impact recall and precision ratios:

- Lack of specificity
- Lack of exhaustively
- Lack of specific terms
- Inadequate hierarchical cross-reference structure
- Defects in hierarchy
- Failure to cover all reasonable approaches to retrieval

Systems Criteria for Evaluation

The evaluation of an ISAR system focuses around the users, information sources, intermediaries, the tools, techniques, methodologies and overall environment. A detailed structure has been provided for it by experts B.C. Vickery and Alina Vickery.

They have set a framework for evaluation of ISAR system based on quality and value as beneficial aspects of the evaluation system. The aim of any evaluation will be to check if user demands are satisfied by providing the correct information. Their thoughts are structured in Table 10.1.

Table 10.1 Criteria for Evaluation

S. No.	Aspects	Parameters
1.	Criteria for Evaluation	Quality and value
2.	A Framework of Evaluation	Correlation of information provision, information content and information use
3.	Relevance and Assessment	Source and receiver coexistence
4.	Service Qualities	Simplicity, ease of use, personal attention, internal decor and presentability, success/failure analysis
5.	Evaluating Performance	Choice of characteristics for performance Measure of performance
6.	System Efficiency, Cost, Manpower and Cost-Effectiveness	Labour, expenditure on documents consumables, equipment external charges, service overhead management and development
7.	Coverage in Acquisition Search	Messages acquired/Messages omitted
8.	Retrieval from Store	Relevant/Not relevant, Precision of retrieval
9.	Evaluation of an Information System	Study of Hits/Non Hits Relevance of Hits Search Qualities
10.	Operational Current Awareness Service	User assessment of processing, indexing and people studies
11.	Online Search Service	Value of computer-based information searches Value of Internet-based information searches
12.	Experimental Study of Retrieval	Intellectual analysis and its role on operations in retrieval
13.	Availability on Demand	Compatibility between Database and the User's demands
14.	Variables affecting Availability	Information and studies
15.	Document-delivery Test	Actual demand potential demand time. Time lag studies
16.	The Effect of Service Delay	Impact on access Impact on queuing Impact on service
17.	Degradation of Performance	Queue imputes Set priorities in the queries
18.	Value of Information	Reference Introduce self-service personal knowledge synthesis Measurement of relevance
19.	The Perceived Value of Information Services	Matrix of value for professional literature

NOTES

Modern ISAR systems make use of advanced technologies like full text and multi-media contents. Access to physical collections as well as electronic collections can be achieved with the help of hybrid systems. All said and done, the criteria for evaluation, as suggested by B.C. Vickery and Alina Vickery, are still applicable in modern ISAR systems with some adjustments to adapt to the new techniques, new technologies and new types of documents.

NOTES

Check Your Progress

1. What is successive fraction strategy?
2. List the check points which must be used to evaluate the ISAR as per Claverdon.

10.4 RELEVANCE AND FAILURE ANALYSIS

The first really structured and systematic technique which was ever developed to undertake the task of failure analysis was Failure Mode and Effects Analysis (FMEA)—also "failure modes". The reliability engineers developed the system first in the 1950s while working on finding solution for malfunctions coming up in the military systems. The first step of any reliability test today is the FMEA. The process constitutes analysing different components, assemblies, and subsystems as possible to recognize the different failure modes, the reason for their working and its consequences. Different FMEA worksheets are prepared to record the different aspects of each component. There are several variations of this kind of worksheets. The broad types include quantitative and qualitative worksheets.

The following are some of the distinct types of FMEA analyses:

- Process
- Functional
- Design

When critical analysis is done along with the FMEA, then it is known to be FMECA (failure mode, effects, and criticality analysis).

The logic that FMEA follows is inductive reasoning or forward logic single point of failure analysis and is the crucial part of the task in different fields as reliability engineering, quality as well as safety engineering.

Another benefit of the FMEA activity is that it can also identify the potential failure modes that might occur based on previously occurred failures in similar processes or products. Consequently, it is a popularly used tool in development and manufacturing industries as a part of the phases of the product life cycle. *Effects analysis* refers to studying the consequences of those failures on different system levels.

Functional analyses are needed as an input to determine correct failure modes, at all system levels, both for functional FMEA or Piece-Part (hardware) FMEA. An FMEA is primarily used to design the solution mechanism which will proactively diffuse any risky situations as well as reduce the chances of failure or both.

Even though we have said earlier that FMEA follows a forward logic principle, the probability function can only be done through the help of a failure

mechanism which necessarily requires a deductive analysis which includes elimination strategy.

The FME(C)A as a design tool is known to have two stage process or analysis one consisting of the study of failure modes and its effects and the other being the process of critical analysis. Fruitful development of an FMEA necessitates that the analyst includes all crucial failure modes for every contributing element or part in the system. The FMEAs is not restricted to a single level or stage and can be performed at different levels of the system including the system, subsystem, assembly, subassembly or part levels. It is important that the FMECA is undertaken simultaneously with the hardware design development otherwise it will not have any influence on the design decisions. The time and effectiveness which the problems related to design are identified and dealt with are the critical to its success. In fact, the time factor can be considered of utmost importance simply because if the problems are identified post the design process it will have no significance. So, the earliest detection and incorporation of the solution in the design process is the fundamental benefit of using the FMECA. This is because the problem is corrected as soon as it is seen. FMECA should be performed at the system level as soon as preliminary design information is available and extended to the lower levels as the detail design progresses.

Remark: Another reliability analysis that deserves a mention here is the fault tree analysis (FTA); which uses a *deductive* (backward logic) to undertake the failure analysis. It may tackle and control multiple failures within the item and/or external to the item including maintenance and logistics. But this mechanism is generally used at the higher levels. So, in the reliability test, the FMEA may be used at the basic levels after which the Fault tree analysis may be brought in. Further, Interface hazard analysis, Human error analysis and others may be added for completion in scenario modelling.

The FMECA can be used at individual levels of the system to ensure that the failure modes are not spilling onto other levels and affecting the functioning of the entire system. So to say that the analysis may first be only effective at the functional level, until further becoming sophisticated enough to deal with the hardware level problems and then moving on to the operational level failure modes that might surface. In addition, each part failure postulated is considered to be the only failure in the system (i.e., it is a single failure analysis). In addition to the FMEA can be accomplished without a CA, but a CA requires that the FMEA has previously identified system level critical failures. When both steps are done, the total process is called a FMECA.

Ground rules

A set of project selected procedures constitute the ground rules of each FMEA. These are nothing but the assumptions on which the analysis is based; the hardware that is being incorporated as well as eliminated from the analysis and along with

NOTES

NOTES

the reasoning behind it. Also included in the ground rules are descriptions about the indenture level of the analysis, status of the basic hardware, and the criteria for system and mission success. It is important that these rules are identified and set before the actual FMEA begins, of course the modifications might be done later on. The following explains a typical set of ground rules:

1. All inputs (including software commands) to the item being analyzed are present and at nominal values.
2. There is availability of Nominal power
3. Only a single failure mode exists at a time.
4. All consumables are present in appropriate quantities.

Benefits

The advantages of a systematically implemented FMECA effort are:

1. It becomes a uniform method of detecting failure mode along with its impact which basically provides a ranking as per the impact it has on the overall system along with the frequency with which it occurs.
2. It becomes a documented source of choosing a design which has a high probability of successful and efficient operation and safety.
3. It becomes the set method for early planning of different tests.
4. It becomes the early check point for failure modes which highly affect the mission success and/or safety. They also provide a method of verifying that switching between redundant elements is not jeopardized by postulated single failures.
5. An effective method for evaluating the effect of pre-decided alteration to the design and/or operational procedures on mission success and safety.
6. A basis for in-flight troubleshooting procedures and for locating performance monitoring and fault-detection devices.

From the aforementioned lists, early identifications of SFPS, input to the troubleshooting procedure and locating of performance monitoring / fault detection devices are probably the most critical advantages of the FMECA. In addition, the FMECA procedures are straightforward and allow orderly evaluation of the design.

History

The early mentions of the FMECA was found in the US Armed Forces Military's Procedures in 1949; which was later revised in 1980 as MIL-STD-1629A. Later on around the 1960s, contractors for the U.S. National Aeronautics and Space Administration (NASA) were also using different versions of FMECA or FMEA under different names. Some of the examples of NASA programs using FMEA includes Apollo, Viking, Voyager, Magellan, Galileo, and Skylab. This was also adapted by the Civil Aviation industry later on.

By 1970s, it was being used by other industries including the diverse field of geological surveys, environmental protection, space and food industries.

The FMEA system made its way to the automotive industry by the mid 1970s. The Ford Motor Company introduced FMEA to the automotive industry for safety and regulatory consideration after the Pinto affair.

There are different industries like the software, healthcare, semiconductor processing, food service, etc. There have also been innovation in the field like the Design Review Based on Failure Mode (DRBFM) which was developed by Toyota. The method is now supported by the American Society for Quality which provides detailed guides on applying the method. One of the disadvantages of the FMEA process is that even though it recognizes and catches the failure mechanisms of the product, it fails to model them sans a specific specialized software. This is why it does not contribute much to the root analysis, critical analysis, virtual qualification and related processes. This problem is solved through the use of Failure Modes, Mechanisms and Effect Analysis (FMMEA).

Probability (P)

In the failure analysis, the likelihood of the failure occurrence along with its causes are important. The manner in which this probability is judged is through the calculation of the FEM as well as considering the failures of the similar products/processes that must have occurred in the past. A failure cause is looked upon as a design weakness. It is important that the identification and documentation of all the likely causes are done. Based on the factors aforementioned, the failure modes are given *Probability Ranking*. This can be in the form of the following:

Rating	Meaning
A	Extremely Unlikely (Virtually impossible or No known occurrences on similar products or processes, with many running hours)
B	Remote (relatively few failures)
C	Occasional (occasional failures)
D	Reasonably Possible (repeated failures)
E	Frequent (failure is almost inevitable)

Severity (S)

This process involves ascertaining the worst-case scenario adverse end effect (state). A very good way to note and observe these effects down from the perspective of the user and what they might see or experience when the failure occurs. This could be in the form of full loss of function, degraded performance, functions in reversed mode, too late functioning, erratic functioning, etc. To judge these severity levels, every end effect is allotted a Severity number (S) from, say, I (no effect) to V (catastrophic), based on its overall consequences. These numbers prioritize the failure modes (together with probability and detectability). Below a

NOTES

typical classification is given. Other classifications are possible. The following is the example of the severity rating scale.

NOTES

Rating	Meaning
I	No relevant effect on reliability or safety
II	Very minor, no damage, no injuries, only results in a maintenance action (only noticed by discriminating customers)
III	Minor, low damage, light injuries (affects very little of the system, noticed by average customer)
IV	Critical (causes a loss of primary function; Loss of all safety Margins, 1 failure away from a catastrophe, severe damage, severe injuries, max 1 possible death)
V	Catastrophic (product becomes inoperative; the failure may result in complete unsafe operation and possible multiple deaths)

Detection (D)

This process observes and records the system which helps in the recognition of the failure mode and how much time it actually takes whether it is detected by the maintainer or the operator. This is crucial to deal with future failures and in emergency situations where multiple failures happen at once. The failure modes to be noted here can also be in the form of latent failure mode (failures occurring gradually but not in a big way) or dormant failure mode (which comprises the failure modes which do not directly affect the system or is covered up by back up systems). In the detection process, it should be understood how the detection is being discovered and noted by the operators in the normal system functioning as well as during diagnostic action by the maintenance crew or automatic system. Based on the observations, the failure modes might be given a dormancy and/or latency period may be entered. The following table showcases the rating which can be used in this case:

Rating	Meaning
1	Certain – fault will be caught on test - e.g. Poka-Yoke
2	Almost certain
3	High
4	Moderate
5	Low
6	Fault is undetected by Operators or Maintainers

Dormancy or Latency Period

It is nothing but a presentation of the average time it takes a failure mode to stay undetected.

Indication

In situations, where the failure mode is not fatal to the system, the operators or the maintenance crew should observe the secondary failure that might occur to ascertain whether or not an indication will be clear and visible to all operators and what corrective action is advisable in such situations.

The following three are the indications to the operator:

- Normal. As the name suggests, this just shows that the system is working in a normal state.
- Abnormal. It showcases that a failure has occurred and that the system has malfunctioned.
- Incorrect. An erroneous indication to an operator due to the malfunction or failure of an indicator (i.e., instruments, sensing devices, visual or audible warning devices, etc.).

Perform Detection Coverage Analysis for Test Processes and Monitoring (From ARP4761 Standard)

This can be called as the test of the test. This is to judge whether the system which checks the detection of the failure mode actually detects it and the failure rate of the failure modes that are detected. The possibility that the detection means may itself fail latently should be accounted for in the coverage analysis as a limiting factor (i.e., coverage cannot be more reliable than the detection means availability). The integration of the detection coverage in the FMEA can lead to each individual failure that would have been one effect category now being a separate effect category due to the detection coverage possibilities. A different way to include detection coverage is for the FTA to conservatively assume that no holes in coverage due to latent failure in the detection method affect detection of all failures assigned to the failure effect category of concern. The FMEA can be revised if necessary for those cases where this conservative assumption does not allow the top event probability requirements to be met.

After these three basic steps the Risk level may be provided.

Risk level (P*S) and (D) []

Risk is the combination of End Effect Probability And Severity where probability and severity includes the effect on non-detectability (dormancy time). This may influence the end effect probability of failure or the worst case effect Severity. The exact calculation may not be easy in all cases, such as those where multiple scenarios (with multiple events) are possible and detectability / dormancy plays a crucial role (as for redundant systems). In that case Fault Tree Analysis and/or Event Trees may be needed to determine exact probability and risk levels.

NOTES

NOTES

Preliminary Risk levels can be selected based on a Risk Matrix like shown below, based on Mil. Std. 882.^[24] The higher the Risk level, the more justification and mitigation is needed to provide evidence and lower the risk to an acceptable level. High risk should be indicated to higher level management, who are responsible for final decision-making.

Probability / I Severity -->	II	III	IV	V	VI	
A	Low	Low	Low	Low	Moderate	High
B	Low	Low	Low	Moderate	High	Unacceptable
C	Low	Low	Moderate	Moderate	High	Unacceptable
D	Low	Moderate	Moderate	High	Unacceptable	Unacceptable
E	Moderate	Moderate	High	Unacceptable	Unacceptable	Unacceptable

- After this step the FMEA has become like a FMECA.

Timing:

The FMEA should be updated whenever:

- A alteration is made in the design
- Newer regulations are being implemented
- A different original cycle of the product/process is beginning
- Changes are made to the operating conditions
- Problems are identified through the Customer feedback

Uses:

- Development of system requirements that minimize the likelihood of failures.
- Development of designs and test systems to ensure that the failures have been eliminated or the risk is reduced to acceptable level.
- Development and evaluation of diagnostic systems
- To help with design choices (trade-off analysis)

Check Your Progress

3. What are some of the distinct types of FMEA analysis?
4. Why is important that FMECA be undertaken simultaneously with hardware design process?
5. What are the rankings used to judge the severity levels of failure modes?

10.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The strategy which deals with the regularity of indexed collection, where various simple keywords are stored in a successive manner is known as successive fraction strategy.
2. Claverdon's view gives following check points on which any ISAR can be evaluated:
 - Coverage
 - Cost Benefit Analysis
 - Time
3. The following are some of the distinct types of FMEA analyses:
 - Process
 - Functional
 - Design
4. It is important that the FMECA is undertaken simultaneously with the hardware design development otherwise it will not have any influence on the design decisions.
5. To judge the severity levels of failure modes, every end effect is allotted a Severity number (S) from, say, I (no effect) to V (catastrophic), based on its overall consequences.

NOTES

10.6 SUMMARY

- Language model is the simplest form of automatic information retrieval. Users enter a text or keywords that are used to search the inverted indexes of the document keywords. This approach retrieves documents based on the presence or absence of exact single word strings as specified by the logical representation of the query. Because of this type of searching approach, the user might miss many relevant documents because it does not capture the complete or deep meaning of the user's query. Linguistic and knowledge-based approaches have also been developed to address this problem by performing a morphological, syntactic and semantic analysis to retrieve documents more effectively.
- A language model for an Information Retrieval (IR) is defined using an ordinary fuzzy language approach. The ordinary fuzzy language model is presented, and is used for modelling the ambiguity and subjectivity that appear in the information retrieval system. The user's query and Information Retrieval responses are modelled linguistically using the concept of fuzzy or

NOTES

unclear linguistic variables.

- Building an effective Information Retrieval system requires proper knowledge of information retrieval techniques and tools used for it. Many strategies are used for providing better services to the user.
- There are different kinds of strategies, ranging from general to efficient strategies. For building an information retrieval system, various design system ranging from weighing system to morphological normalization required. Nowadays, combination of strategies is widely used. The combination of various strategies can lead to significant improvement in retrieval effectiveness. The researchers are focusing more on the relative impact of retrieving more relevant documents with improved rankings.
- Very few IR strategies delivers relevant document, whereas the combination of strategies gives tremendous results. Conceptual IRS use domain knowledge and domain indices to solve the query of the user. In a search where no direct match of query is found, the relationship with others is used to enhance the Recall. The query might be found in this, or to narrow down the search, the document may be best suited to the query, which is known as precision.
- According to experts like Claverdon, an ISAR system should be evaluated on the following criteria:
 - o The extent to which the system includes relevant matter
 - o The time taken from when the search request is made, to the time an answer is given
 - o The look and feel of the presentation output
 - o The effort required by the user to get the answers to his search requests
 - o The percentage of relevant material that is actually retrieved in a search request
 - o The percentage of material that is relevant vis-à-vis the material that is actually retrieved in a search request
- The first really structured and systematic technique which was ever developed to undertake the task of failure analysis was Failure Mode and Effects Analysis (FMEA)—also "failure modes".
- The logic that FMEA follows is inductive reasoning or forward logic single point of failure analysis and is the crucial part of the task in different fields as reliability engineering, quality as well as safety engineering.
- In the failure analysis, the likelihood of the failure occurrence alongwith its causes are important. The manner in which this probability is judged is through the calculation of the FEM as well as considering the failures of the similar products/processes that must have occurred in the past.

10.7 KEY WORDS

- **Symmetrical threshold:** It is a tool in information retrieval system which identifies a new threshold semantic used to express qualitative restrictions on the documents retrieved for a given term.
- **FMEA:** It refers to the process which constitutes analyzing different components, assemblies, and subsystems as possible to recognize the different failure modes, the reason for their working and its consequences.

NOTES

10.8 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What are the three possible semantics in the language model of IR?
2. State the two important factors to be considered while providing information retrieval system.
3. What are the criteria on which the ISAR system must be evaluated as per Claverdon?
4. List some of the factors which influence recall and precision.
5. Briefly explain the logic that FMEA follows.
6. Enumerate the typical set of ground rules for FMEA.

Long-Answer Questions

1. Describe the language model used as a search strategy in IR system.
2. Explain the concept of selection and prevalence of various strategies for the IR system.
3. Discuss the different criteria on which the recall and precision of the IR system is undertaken.
4. Explain the stages and historical development of the FMEA.
5. Describe the meaning and ranking of probability, detection and indication concepts included in the FMEA.

10.9 FURTHER READINGS

- Alberico, R. and Micco M. 1990. *Expert systems for reference and Information retrieval*. West Port: Meckler.
- Atchison, J. & Gilchrist, A. 1972. *Thesaurus construction: a practical manual*. London: Aslib.

NOTES

Chowdhry, G.G.2003. *Introduction to modern Information retrieval*. 2nd Ed. London: Facet Publishing.

Cleaveland, D. B. 2001. *Introduction to Indexing and abstracting*. 3rd Ed. Englewood, Colo: Libraries Unlimited.

Ghosh, S.B. and Biswas, S.C. 1998. *Subject Indexing systems: Concepts, methods and techniques*. Rev. ed. Calcutta: IASLIC.

Pandey, S.K. Ed.2000 .*Library Information retrieval*. New Delhi: Anmol.

BLOCK - V
WEB TECHNOLOGY

Boolean Logic

UNIT 11 BOOLEAN LOGIC

NOTES

Structure

- 11.0 Introduction
- 11.1 Objectives
- 11.2 Boolean Logic: An Overview
 - 11.2.1 Limitations of Boolean Logic
- 11.3 Processing the Boolean Query or Search Operations: rules for operations
- 11.4 Answers to Check Your Progress Questions
- 11.5 Summary
- 11.6 Key Words
- 11.7 Self Assessment Questions and Exercises
- 11.8 Further Readings

11.0 INTRODUCTION

George Boole, the mathematician from 19th century is known to be the founder of the Boolean Logic. As per his stream of algebra, he believed that there must be a reduction of all algebraic equations or values to either 1 or 0. This theory of Boole found a match in the computer science stream's specific theory in which the numerical value is reduced to being perceived either as true or false. Following the logic, Boole was of the opinion, that every fraction of a whole or every individual bit in mathematics holds a value of either 1 or 0. Similarly, the value of true and false can be assigned to every fraction of a whole or bit of a single number.

The mathematical equation forming the basis behind the Boolean logic which is also found in computer science is what makes the logic a part of the field of computer science today.

One reason why the Boolean Logic is now part of computer science is perhaps because both may be interpreted in terms of mathematic equations, in a perfectly logical method. This can be understood from the fact that the earlier computer systems were known to be systems undertaking arithmetic operations automatically with some logic. But exactly how and why the logic worked was not known. Even in today's world, the role of logic is often time undermined.

George Boole, wrote the book, *Mathematical Analysis of Thought and An Investigation of the Laws of Thought*. It discussed the modern logic in the best possible manner but to call the book 'Laws of Thought' was considered to

NOTES

be a bit of a dramatic statement. This is because, even after several years of the publication of the book we are not certain as to what really governs our thoughts, Artificial intelligence would not have been a difficult field if we knew all the answers. It will not be wrong to say that George Boole's idea was both simplistic and revolutionary.

11.1 OBJECTIVES

After going through this unit, you will be able to:

- Discuss the Boolean logic
- Describe the limitations of the Boolean logic
- Explain the processing query expression: rules for operations

11.2 BOOLEAN LOGIC: AN OVERVIEW

We will begin this section by discussing the life background of Boolean.

Where Charles Babbage was known as the father of computer, his contemporary, Boolean is often dubbed the forefather of the information age. He was an Englishman by birth, and was the first professor of mathematics in Ireland's new Queen's College (now University College) Cork in 1849. He expired in 1864 when he was 49 years of age. Claude Shannon is considered to have been influenced by the works of Boolean and credited with bringing to the forefront the Boolean logic and its importance, without whom it would have been lost forever. Boolean logic today is deemed the practical foundation of digital circuit design and the theoretical grounding of the digital age.

Boolean logic is very easy to explain and to understand.

- The first step is the consideration of a statement P which is either true or false, and not anything in between (this called the law of the excluded middle).
- Then as a second step, with the help of fundamental operators like Or, And and Not, other statements, which are also true and false, will be formed based on the combination of the earlier statements.

It can be observed here that the rules mentioned here share similarities with the rules followed in the English language.

For instance, if P is true then Not(P) is false So, if "today is Sunday" is true then "Not(today is Sunday)" is false.

The logical expression is generally translated into English as "today is Not Sunday" and this makes it easier to see that it is false if today is indeed Sunday.

You can probably already sense that these types of discussions are often confusing to follow given the complicated logic involved but this is what is the inherent characteristic of the Boolean logic. Let us see how these arguments are expressed symbolically.

Truth Tables

Certain rules are adhered to when undertaking the task of combining the expressions. These are expressed in the table called truth tables and for the three fundamental operators these are:

P	Q	P AND Q
F	F	F
F	T	F
T	F	F
T	T	T

P	Q	P OR Q
F	F	F
F	T	T
T	F	T
T	T	T

P	NOT P
F	T
T	F

You would have notice here that the operator ‘AND’ is used similarly in English as well as in Boolean logic but this is not the case with the operator ‘OR’.

For instance, in case a person asks whether you’d like to have ‘Sandwich or Burger’, the answer which is expected is not ‘both’.

However, as per the Boolean logic and its use of ‘OR’, if A is true and B is true, then the combined phrase A or B is also true.

If one were to look for an operator in Boolean logic which has the same meaning as that of English ‘OR’, one can use the expression ‘Exclusive or’. This is written as XOR or EOR. The following represents the truth table for the operator:

P	Q	P XOR Q
F	F	F
F	T	T
T	F	T
T	T	F

NOTES

And this one really would stop you having both the tea and the coffee at the same time (notice the last line is True XOR True = False).

Practical truth tables

NOTES

Let us now try to understand the value of the truth tables.

One can definitely say that apart from the trivial level of basic selection, it isn't the foundation for everyday reasoning. Boolean logic is used by most of the people, except for may be politicians for the most non-essential level choice problems.

But despite that, the Boolean logic is critical to designing machines which become the interface and interacts with the outside world. For instance, suppose you decide to design a security system which only works at night and responds to a door being opened. By installing a light sensor, an indication through the signal indicate the truth of the statement:

This is to say that consider $P = \text{It is daytime}$.

This follows that $\text{Not}(P)$ is true when it is night-time and you can witness the Boolean logic in operation already.

Revealing the truth of the statement is the basis of the application of the Boolean logic:

$R = \text{Burglary in progress from } P \text{ and}$

$Q = \text{Window open}$

A little raw thought soon gives the solution that

$R = \text{Not}(P) \text{ And } Q$

That is the truth of the statement is revealed that a "Burglary in progress". This is evident from the following truth table:

P	Q	NOT(P)	NOT (P) AND Q
F	F	T	F
F	T	T	T
T	F	F	F
T	T	F	F

Hence, it must be clear that the alarm only goes off when it is night-time and a window opens.

11.2.1 Limitations of Boolean Logic

No invention in the world can be considered to be perfect. Each has a list of its benefits and limitations. This is especially so when undertaking the theories into the

practical world. Boolean logic is no different either. The Boolean logic is crucial in the subject of information retrieval too.

In this context, therefore, let us now consider what John Papiewski had to say about this topic on his blog Sciencing.

Boolean logic is known as the formal and mathematical approach applied to the process of decision making. But unlike other mathematical theories, the symbols and numbers are replaced here with decision states in the form of yes or no, one and zero. The Boolean system of logic has found its way from the circuit switching in engineering in the early 1990s to its recent applications in first telephone networks and now digital computing.

Boolean Algebra

It is referred to as a system representing the combination of two-valued decision states and arriving at a two-valued outcome. But standard numbers, such as 13.6 are not used here. Instead, binary variables that can have two values, zero and one, which stand in for "false" and "true," respectively are the foundation of the Boolean algebra. Operations here combine binary variables to come up with a binary result. For example, the "AND" operation gives a true result only if both of its arguments, or inputs, are also true. "1 AND 1 = 1," but "1 AND 0 = 0" in Boolean algebra. The OR operation gives a true result if either argument is true. "1 OR 0 = 1," and "0 OR 0 = 0" both illustrate the OR operation.

Digital Circuits

As mentioned earlier, the Boolean algebra benefited electrical designers in the 1930s who worked on telephone switching circuits. Following the Boolean algebra, they set a closed switch equal to one, or "true," and an open switch to be zero, or "false." This beneficial logic seamlessly applies to the digital circuits comprising computers. A high voltage state here is represented through a "true" and a low voltage state is representative of "false." Using high and low voltage states and Boolean logic, engineers developed digital electronic circuits that could solve simple yes-no decision-making problems.

Yes-No Results

It must be understood here that only Black-or-White results are received through the functioning of the Boolean logic on its own. There is never a chance of receiving a "maybe." This restriction is crucial since the logic can only be utilized for situations where the variables are strictly mentioned as true or false (and the values are the only outcome possible).

NOTES

NOTES**Web Searches**

The Boolean logic is used by Web searches for the critical function of filtering results. If you do a search on "scooty dealers," for example, a search engine will have hundreds of millions of matching web pages. If you add the word "Noida," the number drops significantly. This is to say that the search engine is using the Boolean logic to only retain variables which match. And so in this case 'Scooty' And 'dealer' and 'Noida' must be present in the search results. The operation 'or' can also be added in the form of "scooty" and "dealer" AND ("Chicago" OR "Saket") which gives you pages for scooty dealers in Noida or Saket. This elimination trait is very helpful for sifting large volumes of data.

Difficulty

Another problem with the Boolean logic is that its language is a little complex for the beginners. For instance, the 'And' operation is oftentimes confused with its use in English. They expect that the search query for "scooty" AND "dealer" to give more results than just "scooty," as the AND implies adding to results. In order to pinpoint the specific meaning, the Boolean logic utilizes parentheses: "scooty OR car AND dealer" gives you a list of anything to do with scooty added to a list of car dealers, whereas "(scooty OR car) AND dealer" gives a list of scooty dealers and car dealers. The users of this language must learn the language in good time so that its relative losses in the meantime does not prove to be a loss.

Check Your Progress

1. Boolean is known as the forefather of which field?
2. List the decision states which are referred to be the Boolean logic.

11.3 PROCESSING THE BOOLEAN QUERY OR SEARCH OPERATIONS: RULES FOR OPERATIONS

In this section, we attempt to understand the concept of Boolean Query and the rules for operations. The Boolean Query is recognized by mostly all the search engines in the world because it is such an easy and effective search algorithm.

The Boolean search operators like AND, OR, NOT and near are used to combine words and phrases which along with the add and subtract symbols are useful in forming the Boolean query. The operators in fact limit and widen the search query for the user. It is useful to know here that most of the search engines of the world are foundationally made on the Boolean parameters. The search query becomes additionally effective if the user is familiar with its correct usage.

Boolean Search Operators

Let us have a look at the basic rules for operations of the Boolean operators:

- The plus symbol corresponds to the Boolean search operator and
- The minus symbol is equal to the Boolean search operator not
- The basis or the default setting of most of the search engine is based on the Boolean operator or and they give responses for any of the words entered by the user
- When using the Boolean operator ‘near’, the user gives the search engine the special request that the search result must contain the words in the specific order with which they are being posed. For example, if "The Lion King" is being entered, then the specific phrase or sequence is important.

Now let us understand how mathematics assists with the web searches:

- **Minus symbol:** When the minus operator is being used to divide the terms in the search query, it is essentially being conveyed to the search engine to look for and do not include the two respective words. For instance, in the search query, "Thor-Loki", the search results must contain the term "Thor" and must not include extra information like "Loki".
- **Plus symbol:** This operator is used when the search query gives directions to the search engine to present the results containing both of the terms or atleast one. For instance, while searching for "Football+ ISL", the search query must include either both or atleast one of these words in the search query. The add operator basically represents the addition of the specified items of the search query. It resembles the use of the "and" symbol.
- Ofcourse, the search operators may be combined for a more targeted yet complicated search. For instance, the user may type "Thor-Loki-Hela" to indicate that the search result must contain the word "Thor" and must ignore the terms Loki and Hela.

Database Search Tips: Boolean operators

- Overview
- Boolean operators
- Truncation
- Keywords vs. subjects
- Fields
- Phrases
- Stop words .

NOTES

NOTES

Why use Boolean operators?

- To give direction to the search query especially when the topic contains multiple search terms.
- To connect different pieces of information to specify exactly what is being searched
- Example: second creation (title) AND wilmut and campbell (author) AND 2000 (year)

Using AND

The operator AND is used in a search to:

- Restrict the search results
- Specify that all the terms used in the search query must be present in the search result
- Example: early man AND stone age AND Neolithic age

Remember: In majority of the search databases, the operator AND is implied.

- For example, Google automatically puts an AND in between your search terms.
- But it must also be kept in mind that even though all your search terms are included in the results, they may not be connected together in the manner the user desires.
- For example, this search: "School students tiffin ideas" is seen by the search engine as: school AND students AND tiffin AND ideas. The words may be present together or individually throughout the resulting records.
- The phrase can be made more effective by phrasing the search query in a slightly better manner.
- For example: "school students" AND "tiffin ideas". This way, the phrases show up in the results as you expect them to be.

Using OR

Use OR in a search to:

- Join two or more similar concepts (synonyms)
- Widen the search results, telling the database that ANY of your search terms can be present in the resulting records
- example: interval OR lunchbreak OR recess

All three circles represent the result set for this search. It is a big set because any of those words are valid using the OR operator.

Using NOT

This is used in a search query to

- eliminate words from your search
- limit your search, telling the database to ignore concepts that may be implied by your search terms
- example: cloning NOT sheep

Search order

Databases work on the commands given by the users and return results following those command. The logical order in which words are connected should be kept in mind when using Boolean operators:

- Databases generally recognize AND as the primary operator, and will connect concepts with AND together first.
- If you use a combination of AND and OR operators in a search, enclose the words to be "OR" together in parentheses.

Examples:

- ethics AND (cloning OR reproductive techniques)
- (ethic* OR moral*) AND (bioengineering OR cloning)

Check Your Progress

3. What does the Boolean operator 'NEAR'?
4. How is the combination of the Boolean operator 'AND' and 'OR' used together?

11.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Boolean is often dubbed the forefather of the information age.
2. The decision states which are referred to be the Boolean logic includes yes or no, one and zero.
3. When using the Boolean operator 'NEAR', the user gives the search engine the special request that the search result must contain the words in the specific order with which they are being posed.
4. If a combination of AND and OR operators is to be used in a search, the words to be used must be enclosed with the operator "OR" together in parentheses.

NOTES

11.5 SUMMARY

NOTES

- George Boole, the mathematician from 19th century is known to be the founder of the Boolean Logic. As per his stream of algebra, he believed that there must be a reduction of all algebraic equations or values to either 1 or 0.
- Where Charles Babbage was known as the father of computer, his contemporary, Boolean is often dubbed the forefather of the information age. He was an Englishman by birth, and was the first professor of mathematics in Ireland's new Queen's College (now University College) Cork in 1849.
- Boolean logic is very easy to explain and to understand.
 - o The first step is the consideration of a statement P which is either true or false, and not anything in between (this called the law of the excluded middle).
 - o Then as a second step, with the help of fundamental operators like Or, And and Not, other statements, which are also true and false, will be formed based on the combination of the earlier statements.
- Certain rules are adhered to when undertaking the task of combining the expressions. These are expressed in the table called truth tables.
- Apart from the trivial level of basic selection, it isn't the foundation for everyday reasoning. Boolean logic is used by most of the people, except for may be politicians for the most non-essential level choice problems.
- Boolean logic is known as the formal and mathematical approach applied to the process of decision making. But unlike other mathematical theories, the symbols and numbers are replaced here with decision states in the form of yes or no, one and zero. The Boolean system of logic has found its way from the circuit switching in engineering in the early 1990s to its recent applications in first telephone networks and now digital computing.
- It must be understood here that only Black-or-White results are received through the functioning of the Boolean logic on its own. There is never a chance of receiving a "maybe."
- Another problem with the Boolean logic is that its language is a little complex for the beginners. For instance, the 'And' operation is oftentimes confused with its use in English.
- The Boolean search operators like AND, OR, NOT and NEAR are used to combine words and phrases which along with the add and subtract symbols are useful in forming the Boolean query. The operators in fact limit

and widen the search query for the user. It is useful to know here that most of the search engines of the world are foundationally made on the Boolean parameters.

11.6 KEY WORDS

- **Truth tables:** It refers the table which contain rules are adhered to when undertaking the task of combining the expressions. These are expressed in the table called truth tables.
- **Boolean algebra:** It is referred to as a system representing the combination of two-valued decision states and arriving at a two-valued outcome.
- **Boolean search operators:** It refers to the three fundamental operators like AND, OR, NOT and near used to combine words and phrases which along with the add and subtract symbols are useful in forming the Boolean query.

11.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What are the steps involved in the Boolean operation?
2. How does the Boolean logic share similarities with the English language?
3. Explain the yes-no results that are derived from Boolean operation.
4. What are the basic rules for operations of the Boolean operations?
5. Why are Boolean operators used?

Long-Answer Questions

1. Discuss the value of the Boolean logic.
2. What are the limitations of the Boolean logic?
3. Explain the rules followed in the Boolean Query and search operation.

11.8 FURTHER READINGS

- Alberico, R. and Micco M. 1990. *Expert systems for reference and Information retrieval*. West Port: Meckler.
- Atchison, J. & Gilchrist, A. 1972. *Thesaurus construction: a practical manual*. London: Aslib.
- Chowdhry, G.G. 2003. *Introduction to modern Information retrieval*. 2nd Ed. London: Facet Publishing.

NOTES

NOTES

Cleaveland, D. B. 2001. *Introduction to Indexing and abstracting*. 3rd Ed. Englewood, Colo: Libraries Unlimited.

Ghosh, S.B. and Biswas, S.C. 1998. *Subject Indexing systems: Concepts, methods and techniques*. Rev. ed. Calcutta: IASLIC.

Pandey, S.K. Ed.2000 *.Library Information retrieval*. New Delhi: Anmol.

UNIT 12 RECENT TRENDS IN IRS

Structure

- 12.0 Introduction
- 12.1 Objectives
- 12.2 An Overview of Recent Trends in IRS
- 12.3 Internet Information Retrieval and Web based Information Retrieval Trends
- 12.4 Answers to Check Your Progress Questions
- 12.5 Summary
- 12.6 Key Words
- 12.7 Self Assessment Questions and Exercises
- 12.8 Further Readings

NOTES

12.0 INTRODUCTION

Science and technology are two fields where research and development constantly take place almost on a daily basis. So, it is basic and natural to consider that library and information science, and the related field of information retrieval will not lag far behind. Over the centuries, or should we say over the last few decades, developments have been taking place in the science of information retrieval systems as well.

To be more precise, information retrieval has now become a common feature in many other sectors as well, such as the news sector, which in other words is known as the information and broadcasting sector.

In this unit, we will discuss the recent trends in the information retrieval science including the internet and web-based information retrieval.

12.1 OBJECTIVES

After going through this unit, you will be able to:

- Discuss the recent trends in information retrieval systems
- Describe the trends in internet and web based information retrieval

12.2 AN OVERVIEW OF RECENT TRENDS IN IRS

The science journal called ‘Science Digest’ constantly publishes developments in the field of information retrieval science systems. In this section we are giving excerpts from one of the recent articles published in this journal ‘Science Digest’.

The Boolean framework and the standard sets of index terms are the implicit assumption behind the operational retrieval systems which also reveal their characteristics and performance of the search formulation and the retrieval system.

NOTES

But this assumption is not the best way as the document retrievals process is uncertain by its inherent nature. In fact, the failure of the retrieval systems to accurately display the search queries can be pinpointed to the inability of the Boolean model to accurately tackle the retrieval indecisions.

Even though, Boolean algorithm is the most popular choice for information search systems, it is has yet not been established as the best possible framework yet. There are several problems associated with the Boolean system including frequent null output and output overload, inhospitable request formalism, and lack of provision for differing emphasis on different facets of the search. There is available a comprehensive literature on some other nontraditional design principles that can tackle these problems.

It is generally understood that the conventionally designed commercial document retrieval systems will have an efficient and rather perfect performance with the fulfillment of the following two (logical) conditions for every search: (1) There exists a document property (or combination of properties) that belongs to those (and only those) documents that are relevant. (2) That property (or combination of properties) can be correctly guessed by the searcher.

But it is also known that while the first assumption is false, the second assumption is impossible to satisfy; which proves the failure of the conventional systems to work at their most efficient.

Additionally, the function of Recall performs consistently poor with the growth in size of the document library. It must be considered here that the relationship between document properties (whether they be humanly assigned index terms or words that occur in the running text) and relevance is at best probabilistic, and so the best way to approach the design problem is through the probabilistic principles. For conventional IR system, searches should be permitted to attach probabilistically interpreted weights to their query terms at the front end design. This may result in improved performance.

Efforts to improve the effectiveness of Boolean retrieval during the last decade, has been concentrated with studies featuring SMART and SIRE systems. While the SMART system is great for extended Boolean logic, automatic Boolean query construction, and Boolean feedback. Ranking of the output of Boolean queries has shown more efficiency with SIRE.

As mentioned earlier, the foundation of the traditional information retrieval systems are Boolean query formulations which are based with the user requirement in mind. However, it has been noted that that several retrieval system users fail to construct useful Boolean expressions, and require the help of trained search intermediaries in the search formulation process.

The ranking of output documents if made more efficient with weight mechanism would improve tremendously the current operational information retrieval systems However, a number of previous attempts to improve conventional Boolean retrieval systems along these lines have not been entirely successful given their inherent inconsistencies and ambiguities.

Through varied experiments over the years, it has been observed that the extended probabilistic output ranking methods that can be used in retrieval systems which contain simply sets of index terms have worked positively for showcasing demonstrated the value of a systematic statistical use of relevance feedback information. Similarly, it is expected that the system would also boost and correct the inadequacies of the conventional Boolean systems.

Another manner in which the information retrieval system can be improved has been the use of ranking of documents and relevance feedback. This includes positioning the Boolean database queries into Conjunctive Normal Form, a conjunction of disjunctions, and by making the assumption that the disjunctions represent a hyperfeature, documents to be retrieved can be probabilistically ranked and relevance feedback incorporated, improving retrieval performance.

Other trend that has been observed information retrieval system has been the problems related to the integration of database management systems and information retrieval systems. In a database management environment, the records are formatted. The use of precise description of the attributes of the record characteristics and the user needs. This is different from an information retrieval system where textual form is used for identifying references to documents, from which the needs of the user information are met. The descriptors in this case are not as precise as the database management environment. In integrating the two systems, the suggestion of providing the choice of the degree of structuredness of the search, in the hands of the user based on his ability to specify his or her needs is recommended.

The user utilizing the retrieval system is judged by different behavioural responses. Two are very important to be discussed here. First, represents the user's knowledge-perception influencing the search and its expression semantically. Second, the user is expected to link the terms together syntactically, following the rules of the Boolean logic. Since some users might be better at the first or second steps, it has been suggested that a system could be used to assist the user invisibly through the use of enhancement that allows Boolean syntax to be imposed automatically (and invisibly to the user) on an input set of search terms. MEDLINE, is a system which attempts to undertake the task of providing such assistance.

Three formal characteristics usually define a document retrieval system. These are the syntax used to describe documents (keywords or vectors of weights, for instance), the type of machine-processable queries which are considered a valid function for the system (unordered sets of keywords, keywords with Boolean connectives or weighted vectors, for example), and the retrieval rules that are integrated to rank or retrieve documents. In recent times, there have been studies which suggest the use of entirely new forms of queries and retrieval rules with which to make the traditional Boolean document retrieval systems more flexible with queries and ranked document output. But this demands the documents to be explained radically different from their current state.

The ad hoc inquiry in the document retrieval system is explored through the full potential of the relational model. In addition, relational DBMS's also provide

NOTES

NOTES

effective tools for managing document data bases by providing facilities for, inter alia, concurrency control, data migration and reorganization routines, authorization mechanisms, enforcement of integrity constraints, dynamic data definition, etc. Examples of ad hoc inquiry can be seen through the SQL.

Check Your Progress

1. State the two logical conditions necessary for the perfect performance of a conventionally designed commercial document retrieval systems.
2. What are the foundations of the traditional information retrieval systems?
3. List the three formal characteristics that usually define a document retrieval system.

12.3 INTERNET INFORMATION RETRIEVAL AND WEB BASED INFORMATION RETRIEVAL TRENDS

While we have been discussing information retrieval trends in recent times in the previous section, we should also remember that the internet itself is a wide world out there. The Internet can be used by almost every industry today from food, to the health care provider and hospital industry to the travel industry. Every sector in business and the industrial world practically benefitted due to the internet. Of course, one must never forget the social media networking sector, which includes such social media as facebook, twitter, tumblr and more recently LinkedIn.

It is also important for internet users and the students of library and information science that the two terms internet information retrieval and web-based information retrieval are perhaps synonyms for the same field of science. Every database of information is linked to a specific web page or web site. Therefore, it must be understood that one field cannot be disassociated or separated from the other.

While information retrieval as a whole, and the use of the internet and various web pages and web sites in specific essentially has brought development into our world, it must be remembered that it has also brought in many unseen but nevertheless grave concerns. We have endeavored to explain the issues and trends in the internet information and web-based information retrieval systems in the content that follows on the following pages.

But by far the most significant sectors that makes use of the internet in more than one way would be the information and broadcasting industry, and the news agencies, newspapers, radio stations and television news channels.

One of the most famous sources of information that are relevant in the contemporary time are the mainstream media outlets. It can be seen from the manner in which the media affects the political activities and brand reputations today.

New media uses the Information Retrieval (IR) system since a very long time but there is always scope for improvement and loopholes that need to be plugged.

Let's see how IR system applies to the social media platform and news in the recent times. Social networking services work on different devices including the likes of desktops and on laptops, on mobile devices such as tablet computers and smartphones which essentially indicates the use of different formats. They may feature digital photo/video/sharing and "web logging" diary entries online (blogging). Moving beyond the individual interactions, the social media can be referred to as "websites that facilitate the building of a network of contacts in order to exchange various types of content online". Of course, new social ties are established and maintained through these platforms.

Social networking sites allow users to share ideas, digital photos and videos, posts, and to inform others about online or real-world activities and events with people in their network. The social media allows long distances to not be an issue helping the maintenance of social networking. The success of social networking services can be judged by their imposing presence in society today, with Facebook having a gigantic 2.13 billion active monthly users and an average of 1.4 billion daily active users in 2017.

Now, let's get a little technical to study the major categories of data input that are used on social networking services, these could be data descriptions like age or occupation or religion, means to connect with friends (usually with self-description pages), and a recommendation system linked to trust. Depending on the utility or purpose, social network services can be categorized as:

- Used for socializing with existing friends (e.g., Facebook)
- Used for non-social interpersonal communication (e.g., LinkedIn, a career- and employment-oriented site)
- Used for helping users to find specific information or resources (e.g., Goodreads for books)

A challenge of definition

To define social media is a big problem and complex issue because of the variety and evolving range of stand-alone and built-in social networking services in the online space introduces a challenge of definition.

Attempting definition

The following are the commonalities unique to current social networking services:

- (1) social networking services are interactive Web 2.0 Internet-based applications,
- (2) social networking services facilitate the development of social networks online by connecting a user's profile with those of other individuals or groups
- (3) users create service-specific profiles for the site or app that are designed and maintained by the SNS organization

NOTES

- (4) user-generated content (UGC), such as user-submitted digital photos, text posts, "tagging", online comments, and diary-style "web logs" (blogs), is the lifeblood of the SNS organism,

NOTES

Offline and online social networking services

Table 12.1 Differences between offline and online social networking services

Characteristic	Offline social network	Online social network
Degree centrality	While the number of cognitively manageable ties is limited to about 150 (Dunbar 2003), most people report having 14-56 ties at average (Granovetter 1983; van Tilburg 1995; Christakis and Fowler 2009)	Huge number of ties technologically possible, but average number is limited, e.g., Facebook: 395 (Tong et al. 2008), LinkedIn: 149 (Utz 2016), XING: 121 (Buettner 2016c), Twitter: 150-250 (Gonçalves et al. 2011; Hofer and Aubert 2013)
Symmetry	Usually symmetric (reciprocal behavior, cf. Buettner (2009))	Symmetric (e.g., Facebook, LinkedIn, XING, cf. Buettner (2016d)) and asymmetric (e.g., Twitter, cf. Buettner and Buettner (2016))
Affect	Positive (92-97 %) and negative (3-8 %) tie relationships (Kane et al. 2014) can be managed using high sophisticated coordination mechanisms such as argumentation and negotiation (Buettner 2006a, 2006b; Landes and Buettner 2012; Buettner 2016a)	Except through blocking (e.g., Twitter) or hiding (e.g., Facebook) limited support to deal with negative tie relationships
Strength	2-8 strong ties and 12-48 weak/latent ties on average (Granovetter 1983; Christakis and Fowler 2009)	9-37 strong ties and 68-131 weak/latent ties on average (Levin and Cross 2004; De Meo et al. 2014; Utz 2016)
Dynamic of change	Low due to manual interaction (Freeman 1977; Miritello et al. 2013)	High because of technological support (Miritello et al. 2013; Kane et al. 2014)

Usenet, ARPANET, LISTSERV, and bulletin board services (BBS) are the examples of many early online services which were actually effort to help the social networks via computer-mediated communication. Prototypical features of social networking sites were present in many online services such as America Online, Prodigy, CompuServe, ChatNet, and The WELL.

Generalized online communities including the likes of Theglobe.com (1995), Geocities (1994) and Tripod.com (1995) were the earliest known form of global communities on the Web. These early communities focused on bringing people together to interact with each other through chat rooms, and motivated users to share personal information and ideas via personal webpages through easy-to-use publishing tools and free or inexpensive webspace. And some communities – such as Classmates.com – did it differently by making email addresses the linking factor between people.

The highlighting feature of social networking sites in the late 1990s was towards allowing users to compile lists of "friends" and search for other users based on similar interests. Further innovation began to revolutionize the field through the development of more advanced features for users to find and manage friends. Examples include platforms like Open Diary, a community for online diarists, invented both friends-only content and the reader comment, two features of social networks important to user interaction.

SixDegrees.com in 1997, followed by Open Diary in 1998, Mixi in 1999, Makeoutclub in 2000, Hub Culture and Friendster in 2002, were all the newer generation of social media platforms and soon became part of the Internet mainstream. However, thanks to the nation's high Internet penetration rate, the first mass social networking site was the South Korean service, Cyworld. In 1999, it was launched as a blog-based site and by 2001 the different social networking features were added to it. Another landmark achieved by the website was that it also became one of the first companies to profit from the sale of virtual goods. Friendster was followed by MySpace and LinkedIn a year later, and eventually Bebo. Friendster became very popular in the Pacific Islands. Orkut became the first popular social networking service in Brazil (although most of its very first users were from the United States) and quickly grew in popularity in India (Madhavan, 2007). Attesting to the rapid increase in social networking sites' popularity. Myspace was getting more page views than Google by 2005. In early 2009, Facebook, launched in 2004, became the largest social networking site in the world. Facebook was first introduced as a Harvard social networking site, which expanded to other universities and eventually, anyone. The term social media was introduced and soon became widespread.

Social impact

Web-based social networking services make it possible to connect people who share interests and activities across different borders including social, cultural and political. Gift economy and reciprocal altruism are seen to rise encouraged through cooperation which is introduced in the e-mail and instant messaging and the subsequent creation of online communities. Information is suited to a gift economy, as information is a nonrival good and can be gifted at practically no cost. The social features of these platforms are not restricted to only the technological features of the social network platforms alone, the level of network sociability should determine by the actual performances of its users. In fact, there are several benefits of that people are seeking from the social media and the internet in general including personal integrative, social integrative, and tension free needs along with cognitive, affective needs. Internet technology is quickly becoming a supplement to fulfill needs, it is in turn influencing every day life, including relationships, school, church, entertainment, and family. Potential employees' personalities and behaviour are easily revealed through the social media companies.

Much research is still required in the terms of identity, privacy, social capital, youth culture, and education from platforms like Facebook and other social

NOTES

NOTES

networking sites. The lines between work and home lives are often being blurred by the social media sites being relied upon by the users too much. In the times of breaking news, Twitter users are more likely to stay invested in the story. This can also be realized with the increasing trend of considering social media as the primary source of news.

A 2015 study shows that 85% of people aged 18 to 34 use social networking sites for their purchase decision making. While over 65% of people aged 55 and over rely on word of mouth. Several websites are beginning to tap into the power of the social networking model for philanthropy. Fragmented industries and small organizations benefit from these models without the resources to reach a broader audience with interested users. Social networks are providing a different way for individuals to communicate digitally. These communities of hypertexts allow for the sharing of information and ideas, an old concept placed in a digital environment.

Research has provided us with mixed results as to whether or not a person's involvement in social networking can affect their feelings of loneliness. Studies have indicated that how a person chooses to use social networking can change their feelings of loneliness in either a negative or positive way. Each social networking user is able to create a community that centers around a personal identity they choose to create online. In 2016, news reports stated that excessive usage of SNS sites may be associated with an increase in the rates of depression, to almost triple the rate for non-SNS users. According to a recent article from *Computers in Human Behavior*, Facebook has also been shown to lead to issues of social comparison. *Computers in Human Behavior* emphasizes that these feelings of poor mental health have been suggested to cause people to take time off from their Facebook accounts; this action is called "Facebook Fatigue" and has been common in recent years.

Typical features

According to Boyd and Ellison's (2007) article, "Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life", social networking sites share a variety of technical features that allow individuals to: construct a public/semi-public profile, articulate a list of other users that they share a connection with, and view their list of connections within the system.

Additional features

There is a trend towards more interoperability between social networks led by technologies such as OpenID and OpenSocial. In most mobile communities, mobile phone users can now create their own profiles, make friends, participate in chat rooms, create chat rooms, hold private conversations, share photos and videos, and share blogs. Some companies provide wireless services that allow their customers to build their own mobile community and brand it.

Emerging trends

While the popularity of social networking consistently rises, new uses for the technology are frequently being observed. The concept of "real-time web" and

"location-based" are the primary dominant trends in social networking sites. Real-time allows users to contribute contents, which is then broadcast as it is being uploaded—the concept is analogous to live radio and television broadcasts. Twitter, Facebook, Pinterest all have these features.

Cloud computing is also gaining popularity among companies to handle the social networking concepts. Shared business needs are quickly becoming the connecting links between individuals based on social interest. Examples include sites like LinkedIn, Monster.com, which have integrated a "socialized" feel to their career center sites. These more business related sites are generally referred to as "Vocational Networking Sites" or "Vocational Media Networks", with the former more closely tied to individual networking relationships based on social networking principles.

Foursquare, Gowalla and similar sites have gained popularity as it allowed for users to check into places that they are frequenting at that moment. Another example is Clixtr.

According to Jody Nimetz, author of *Marketing Jive*, there are five major uses for businesses and social media: to create brand awareness, as an online reputation management tool, for recruiting, to learn about new technologies and competitors, and as a lead generation tool to intercept potential prospects. These companies are able to drive traffic to their own online sites while encouraging their consumers and clients to have discussions on how to improve or change products or services.

Niche networks

The niche social network offers a specialized space that's designed to appeal to a very specific market with a clearly defined set of needs. Where once the streams of social minutia on networks such as Facebook and Twitter were the ultimate in online voyeurism, now users are looking for connections, community and shared experiences. These tap into shared interests in similar specific activities and hobbies. Fishbrain for fishing and Strava for cycling are examples of the same.

Science

Social networking is allowing scientific groups to expand their knowledge base and share ideas, and without these new means of communicating their theories might become "isolated and irrelevant". Researchers use social networks frequently to maintain and develop professional relationships. They are interested in consolidating social ties and professional contact, keeping in touch with friends and colleagues and seeing what their own contacts are doing. The scale of dissemination of information through these social networks and the result increase in the social contact is one of the intriguing aspects of social networks. Social networks like Academia.edu, LinkedIn, Facebook, and ResearchGate give the possibility to join professional groups and pages.

NOTES

Impact of Social Media

Education

NOTES

The European Southern Observatory uses social networks to engage people in astronomical observations.

The advent of social networking platforms may also be impacting the way(s) in which learners engage with technology in general. For a number of years, Prensky's (2001) dichotomy between Digital Natives and Digital Immigrants has been considered a relatively accurate representation of the ease with which people of a certain age range—in particular those born before and after 1980—use technology. However there are certain research which have disclosed that the use of social networks among students have been known to negatively affect their academic life. Given the fact that their use constitutes distractions, as well as that the students tend to invest a good deal of time in the use of such technologies.

Albayrak and Yildirim (2015) examined the educational use of social networking sites. They investigated students' involvement in Facebook as a Course Management System (CMS) and the findings of their study support that Facebook as a CMS has the potential to increase student involvement in discussions and out-of-class communication among instructors and students.

Curriculum use

Curriculum uses of social networking services also can include sharing curriculum-related resources. Educators tap into user-generated content to find and discuss curriculum-related content for students. Responding to the popularity of social networking services among many students, teachers are increasingly using social networks to supplement teaching and learning in traditional classroom environments as they can provide new opportunities for enriching existing curriculum through creative, authentic and flexible, non-linear learning experiences.

Professional use

Professional use of social networking services refers to the employment of a network site to connect with other professionals within a given field of interest. SNSs like LinkedIn, a social networking website geared towards companies and industry professionals looking to make new business contacts or keep in touch with previous co-workers, affiliates, and clients.

Learning use

Educators and advocates of new digital literacies are confident that social networking encourages the development of transferable, technical, and social skills of value in formal and informal learning. The use of SNSs allow educators to enhance the prescribed curriculum. When learning experiences are infused into a website students utilize everyday for fun, students realize that learning can and should be a part of everyday life.

Constraints

In the past, social networking services were viewed as a distraction and offered no educational benefit. Blocking these social networks was a form of protection for students against wasting time, bullying, and invasions of privacy. Cyberbullying has become an issue of concern with social networking services. Social networking services often include a lot of personal information posted publicly, and many believe that sharing personal information is a window into privacy theft.

Positive correlates

Many researchers have tried to establish the fact that people can derive a sense of social connectedness and belongingness in the online environment.

Employment

A rise in social network use is being driven by college students using the services to network with professionals for internship and job opportunities. It allows alumni, students and unemployed individuals look for work. They are also able to connect with others professionally and network with companies.

In addition, employers have been found to use social network sites to screen job candidates.

Grassroots organizing

Social networks are being used by activists as a means of low-cost grassroots organizing different movements.

Business model

Few social networks charge money for membership. In part, this may be because social networking is a relatively new service, and the value of using them has not been firmly established in customers' minds. Companies such as Myspace and Facebook sell online advertising on their site. Their business model is based upon large membership count, and charging for membership would be counterproductive. Revenue is typically gained in the autonomous business model via advertisements.

Hosting service

A social network hosting service is a web hosting service that specifically hosts the user creation of web-based social networking services, alongside related applications.

Trading networks

A social trade network is a service that allows traders of financial derivatives such as contracts for difference or foreign exchange contracts to share their trading activity via trading profiles online. These services are created by financial brokers.

Spamming

Spamming on online social networks is quite prevalent. A primary motivation to spam arises from the fact that a user advertising a brand would like others to see them and they typically publicize their brand over the social network. Detecting

NOTES

NOTES

such spamming activity has been well studied by developing a semi-automated model to detect spams.

Social interaction

People use social networking sites for meeting new friends, finding old friends, or locating people who have the same problems or interests they have. More and more relationships and friendships are being formed online and then carried to an offline setting.

Potential for misuse

The relative freedom afforded by social networking services has caused concern regarding the potential of its misuse by individual patrons. Online social networks have also become a platform for spread of rumors, one such study has analyzed rumors in retrospect. One of the approaches to detect rumors (or misinformation) is to compare the spread of topic over social network (say Twitter) with those spread by reliable and authorized news agencies

Privacy

Privacy concerns with social networking services have been raised growing concerns among users on the dangers of giving out too much personal information and the threat of sexual predators. Users of these services also need to be aware of data theft or viruses.

Another debate lies in the design of algorithmic systems to target specific audiences on social networking sites. With multiple formats for marketing, Facebook offers a variety of direct marketing options for advertisers to reach their intended audience. Users who "like" a business page will be subscribed to receive these business' updates on their home News Feed. Banner ads and suggested posts are paid for by marketers and advertisers to reach their intended audience. Like other methods of marketing, emotional connections are critical to reaching the user.

The debate questions to what extent the design of these systems is compromising the needs, privacy and information of the users. This debate was further ignited in early 2018. On April 10, 2018 Mark Zuckerberg testified before congress on questions defining Facebook's policy, information handling and data design systems.

Data mining

Through data mining, companies are able to improve their sales and profitability. With this data, companies create customer profiles that contain customer demographics and online behavior. A recent strategy has been the purchase and production of "network analysis software". This software is able to sort out through the influx of social networking data for any specific company.

Impact on employability

Many people use social networking sites to express their personal opinions about current events and news issues to their friends. If a potential applicant expresses

personal opinions on political issues or makes potentially embarrassing posts online on a publicly available social networking platform, employers can access their employees' and applicants' profiles, and judge them based on their social behavior or political views. Cases like these have created some privacy implications as to whether or not companies should have the right to look at employees' social network profiles.

Notifications

There has been a trend for social networking sites to send out only "positive" notifications to users. This allows users to purge undesirables from their list extremely easily and often without confrontation since a user will rarely notice if one person disappears from their friends list.

Access to information

Many social networking services, such as Facebook, provide the user with a choice of who can view their profile. This is supposed to prevent unauthorized users from accessing their information. Log in or provide a password is used to edit information on a certain social networking service account, the social networking sites. It will help prevent unauthorized users from adding, changing, or removing personal information, pictures, or other data.

Unauthorized access

There are different forms where user data in social networks are accessed and updated without a user's permission.

Risk for child safety

Citizens and governments have been concerned with misuse of social networking services by children and teenagers, in particular in relation to online sexual predators. For instance, there is a study which suggests the children are not too far from inappropriate content on YouTube. Social networking can also be a risk to child safety in another way; parents can get addicted to games and neglect their children. Law enforcement agencies have published articles with their recommendations to parents about their children's use of social networking sites.

Trolling

Social networking sites such as Facebook are occasionally used to emotionally abuse, harass or bully individuals, either by posting defamatory statements or by forwarding private digital photos or videos that can have an adverse impact on the individuals depicted in the videos. Such actions are often referred to as "trolling". Confrontations in the real world can also be transferred to the online world. Trolling is a prominent issue in the 2010s, and as the Internet and social media is consistently expanding and more individuals sign up to social networking sites, more people come under fire and become the target of trolls.

Online bullying

Online bullying, also called cyberbullying, is a relatively common occurrence and it can often result in emotional trauma for the victim. There are not many limitations

NOTES

NOTES

as to what individuals can post when online. Individuals are given the power to post offensive remarks or pictures that could potentially cause a great amount of emotional pain for another individual.

Interpersonal communication

Interpersonal communication has been a growing issue as more and more people have turned to social networking as a means of communication. The convenience that social network sites give users to communicate with one another can also damage their interpersonal communication.

Psychological effects of social networking

As social networking sites have risen in popularity over the past years, people have been spending an excessive amount of time on the Internet in general and social networking sites in specific. This has led researchers to debate the establishment of Internet addiction as an actual clinical disorder. Social networking can also affect the extent to which a person feels lonely.

Patents

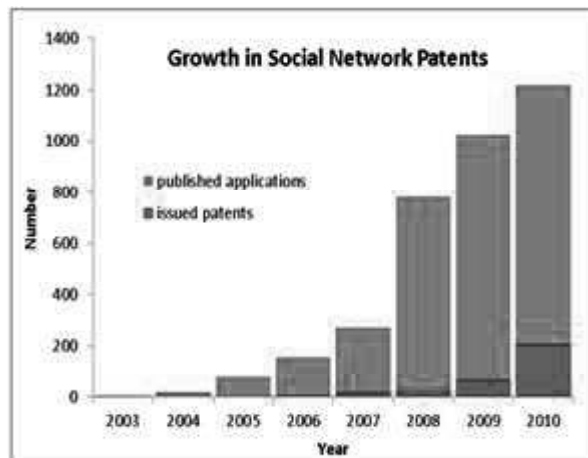


Figure 12.1 Growth in Social Network Patents

Number of US social network patent applications published per year and patents issued per year

A networked computer system provides various services for assisting users in locating, and establishing contact relationships with, other users. For example, in one embodiment, users can identify other users based on their affiliations with particular schools or other organizations. The system also provides a mechanism for a user to selectively establish contact relationships or connections with other users, and to grant permissions for such other users to view personal information of the user. The system may also include features for enabling users to identify contacts of their respective contacts. In addition, the system may automatically notify users of personal information updates made by their respective contacts.

The patent has garnered attention due to its similarity to the popular social networking site Facebook.

Worker's rights

Companies are concerned with the potential damage comments online can do to public image due to their visibility and accessibility online.

Virtual identity suicide

There is a growing number of social network users who decide to quit their user account by committing a so-called virtual identity suicide or Web 2.0 suicide. The number one reason for these users was privacy concerns (48%), being followed by a general dissatisfaction with the social networking website (14%), negative aspects regarding social network friends (13%) and the feeling of getting addicted to the social networking website (6%).

Breaking up

Sites such as Facebook are becoming increasingly popular tools for methods of ending relationships and friendships, proving that although new media is being used as a tool for connecting with individuals, it is now creating new problems associated with disconnecting from others. Instead of the traditional phone call or face-to-face interaction between individuals, people are now starting to end relationships by simply changing their relationship status, knowing full well that their partner will soon see it. New media websites have made our private lives much more public, especially when it comes to breaking up, since updates are able to be immediately viewed by everyone in our networks (which tend to be more people than we would normally tell personally); for example, having friends comment on your newly changed "single" relationship status, and having to explain what happened can be distressing.

This creates further problems, as it is even more crucial to 'save face' after one's relationship has been broken when one is connected to new media technologies. Many people find that the only way to really move on from a past relationship is to cut the person out of their life completely. Social media has made this process much more complicated and difficult.

Social overload

The increasing number of messages and social relationships embedded in SNS also increases the amount of social information demanding a reaction from SNS users. Consequently, SNS users perceive they are giving too much social support to other SNS friends. This dark side of SNS usage is called 'social overload'. It is caused by the extent of usage, number of friends, subjective social support norms, and type of relationship (online-only vs offline friends) while age has only an indirect effect. The psychological and behavioral consequences of social overload include perceptions of SNS exhaustion, low user satisfaction, and high intentions to reduce or stop using SNS.

Social anxiety

Smart phones and social networking services enable us to stay connected continuously with people around us or far away from us, which however is sometimes the root of our anxiety in social life. The eager to know what everyone

NOTES

NOTES

was saying and the tendency to see if anyone shared new things are typical "symptoms" of this anxiety called FOMO. There is a study that examined possible connections between FOMO and social media engagement indicating that FOMO was associated with lower need satisfaction, mood and life satisfaction.

Another type of social anxiety is the FOBM (fear of being missed). It comes from the situation that we can't produce share-content for people to consume.

Effects on personal relationships and social capital

The number of contacts on a social platform is sometimes considered an indicator of social capital. Online platforms and social media services altered the old definition of friendship. Indeed, friendship "redoubleth joys, and cutteth griefs in halves" as stated by Francis Bacon. However, nowadays we see that Facebook friends for instance encourage negative feelings, such as envy, revenge and sadness.

Investigations

Social networking services are increasingly being used in legal and criminal investigations. Information posted on sites such as MySpace and Facebook has been used by police (forensic profiling), probation, and university officials to prosecute users of said sites.

Business application

Social networks connect people at low cost; this can be beneficial for entrepreneurs and small businesses looking to expand their contact bases. These networks often act as a customer relationship management tool for companies selling products and services. Applications for social networking sites have extended toward businesses and brands are creating their own, high functioning sites, a sector known as brand networking. For these types of applications the term "enterprise social software" is becoming increasingly popular.

Educational application

Social networks focused on supporting relationships between teachers and their students are now used for learning, educator professional development, and content sharing. HASTAC is a collaborative social network space for new modes of learning and research in higher education, K-12, and lifelong learning; Ning supports teachers; TermWiki, TeachStreet and other sites are being built to foster relationships that include educational blogs, eportfolios, formal and ad hoc communities, as well as communication such as chats, discussion threads, and synchronous forums. These sites also have content sharing and rating features. Social networks are also emerging as online yearbooks, both public and private.

Medical and health applications

The advantage of using a dedicated medical social networking site is that all the members are screened against the state licensing board list of practitioners. Social networks are creating new trend to help its members with various physical and mental ailments.

Social networks are beginning to be adopted by healthcare professionals as a means to manage institutional knowledge, disseminate peer to peer knowledge and to highlight individual physicians and institutions.

Social and political applications

Social networking sites have recently showed a value in social and political movements. In the Egyptian revolution, Facebook and Twitter both played an allegedly pivotal role in keeping people connected to the revolt. Another important aspects is that social media helps with in political applications is attracting the younger generations involved in politics and ongoing political issues. To engage with the youth, who perhaps are the least educated in politics and the most in social networking sites, political applications of social networking sites are crucial, particularly

Social media driven political campaigns are getting increasing successful including examples like Obama's social media campaign, Donald Trump's presidential electoral campaign in 2016 and PM Narendra Modi's 2014 campaign.

Check Your Progress

4. State the major categories of data input used by social media.
5. What was the highlighting feature of the social networking sites in the late 1990s?

12.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. It is generally understood that the conventionally designed commercial document retrieval systems will have an efficient and rather perfect performance with the fulfillment of the following two (logical) conditions for every search: (1) There exists a document property (or combination of properties) that belongs to those (and only those) documents that are relevant. (2) That property (or combination of properties) can be correctly guessed by the searcher.
2. The foundation of the traditional information retrieval systems are Boolean query formulations which are based with the user requirement in mind.
3. Three formal characteristics usually define a document retrieval system. These are the syntax used to describe documents (keywords or vectors of weights, for instance), the type of machine-processable queries which are considered a valid function for the system (unordered sets of keywords, keywords with Boolean connectives or weighted vectors, for example), and which retrieval rules are integrated to rank or retrieve documents.
4. The major categories of data input that are used on social networking services, these could be data descriptions like age or occupation or religion, means

NOTES

NOTES

to connect with friends (usually with self-description pages), and a recommendation system linked to trust.

5. The highlighting feature of social networking sites in the late 1990s was towards allowing users to compile lists of "friends" and search for other users based on similar interests.

12.5 SUMMARY

- Science and technology are two fields where research and development constantly take place almost on a daily basis. So, it is basic and natural to consider that library and information science, and the related field of information retrieval will not lag far behind.
- The Boolean framework and the standard sets of index terms are the implicit assumption behind the operational retrieval systems which also reveal their characteristics and performance of the search formulation and the retrieval system. But this assumption is not the best way as the document retrievals process is uncertain by its inherent nature. In fact, the failure of the retrieval systems to accurately display the search queries can be pinpointed to the inability of the Boolean model to accurately tackle the retrieval indecisions.
- Efforts to improve the effectiveness of Boolean retrieval during the last decade, has been concentrated with studies featuring SMART and SIRE systems. While the SMART system is great for extended Boolean logic, automatic Boolean query construction, and Boolean feedback. Ranking of the output of Boolean queries has shown more efficiency with SIRE.
- The ranking of output documents if made more efficient with weight mechanism would improve tremendously the current operational information retrieval systems However, a number of previous attempts to improve conventional Boolean retrieval systems along these lines have not been entirely successful given their inherent inconsistencies and ambiguities.
- Another manner in which the information retrieval system can be improved has been the use of ranking of documents and relevance feedback. This includes positioning the Boolean database queries into Conjunctive Normal Form, a conjunction of disjunctions, and by making the assumption that the disjunctions represent a hyperfeature, documents to be retrieved can be probabilistically ranked and relevance feedback incorporated, improving retrieval performance.
- Three formal characteristics usually define a document retrieval system. These are the syntax used to describe documents (keywords or vectors of weights, for instance), the type of machine-processable queries which are considered a valid function for the system (unordered sets of keywords, keywords with Boolean connectives or weighted vectors, for example), and which retrieval rules are integrated to rank or retrieve documents.

- The Internet can be used by almost every industry today from food, to the health care provider and hospital industry to the travel industry. Every sector in business and the industrial world practically benefitted due to the internet. Of course, one must never forget the social media networking sector, which includes such social media as facebook, twitter, tumblr and more recently LinkedIn.
- The following are the commonalities unique to current social networking services:
 - (1) social networking services are interactive Web 2.0 Internet-based applications,
 - (2) social networking services facilitate the development of social networks online by connecting a user's profile with those of other individuals or groups
 - (3) users create service-specific profiles for the site or app that are designed and maintained by the SNS organization
 - (4) user-generated content (UGC), such as user-submitted digital photos, text posts, "tagging", online comments, and diary-style "web logs" (blogs), is the lifeblood of the SNS organism,
- Much research is still required in the terms of identity, privacy, social capital, youth culture, and education from platforms like Facebook and other social networking sites. The lines between work and home lives are often being blurred by the social media sites being relied upon by the users too much. In the times of breaking news, Twitter users are more likely to stay invested in the story. This can also be realized with the increasing trend of considering social media as the primary source of news.
- While the popularity of social networking consistently rises, new uses for the technology are frequently being observed. The concept of "real-time web" and "location-based" are the primary dominant trends in social networking sites. is Real-time allows users to contribute contents, which is then broadcast as it is being uploaded—the concept is analogous to live radio and television broadcasts. Twitter, Facebook, Pinterest all have these features.
- The effects of the social media is varied and still under study.

NOTES

12.6 KEY WORDS

- **Database management system:** It refers to the software that handles the storage, retrieval, and updating of data in a computer system.
- **Social media:** It refers to the websites that facilitate the building of a network of contacts in order to exchange various types of content online.

- **Niche social network:** It refers to the networks that offer a specialized space that's designed to appeal to a very specific market with a clearly defined set of needs.

NOTES

12.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What are the shortcomings of the traditional retrieval systems?
2. Comment on the integration of database management systems and information retrieval systems.
3. What are the categories of social media based on the utility or purpose?
4. Define social media in technical terms.
5. What are the shared characteristics of all social media?

Long-Answer Questions

1. Discuss the recent trends in the information retrieval system.
2. Trace the comprehensive history of social media and its use.
3. What are the ways in which companies use social media to further their business?
4. Discuss the psychological effects of social media.

12.8 FURTHER READINGS

- Alberico, R. and Micco M. 1990. *Expert systems for reference and Information retrieval*. West Port: Meckler.
- Atchison, J. & Gilchrist, A. 1972. *Thesaurus construction: a practical manual*. London: Aslib.
- Chowdhry, G.G. 2003. *Introduction to modern Information retrieval*. 2nd Ed. London: Facet Publishing.
- Cleaveland, D. B. 2001. *Introduction to Indexing and abstracting*. 3rd Ed. Englewood, Colo: Libraries Unlimited.
- Ghosh, S.B. and Biswas, S.C. 1998. *Subject Indexing systems: Concepts, methods and techniques*. Rev. ed. Calcutta: IASLIC.
- Pandey, S.K. Ed. 2000. *Library Information retrieval*. New Delhi: Anmol.

UNIT 13 AUTOMATIC INDEXING AND WEB ONTOLOGY

NOTES

Structure

- 13.0 Introduction
- 13.1 Objectives
- 13.2 Automatic Indexing
- 13.3 Web Ontology Language (OWL) Source Wikipedia
- 13.4 Answers to Check Your Progress Questions
- 13.5 Summary
- 13.6 Key Words
- 13.7 Self-Assessment Questions and Exercises
- 13.8 Further Readings

13.0 INTRODUCTION

Computerized indexing is a highly systematic technique used in the indexing systems. It requires professional inputs of related data and previous records and results to arrive at different relations which affect the functioning of the retrieval system. Ontologies affect the classification of knowledge into various systems. These languages are an integral part of the indexing system. In this unit, we will discuss the components of the automatic indexing system and the levels that are a part of the web ontology.

13.1 OBJECTIVES

After going through this unit, you will be able to:

- Discuss the concept of automatic indexing
- Describe the constituents of web ontology

13.2 AUTOMATIC INDEXING

In many literatures of Library and Information Science, the term ‘automatic indexing’ is interchangeably used with the term ‘computerized indexing’. A fully automatic indexing system would be one in which indexing is conducted by computers, an internally generated thesaurus is prepared, and search strategies are developed automatically from a natural language statement of information need. Salton provides the following definition of automatic indexing: When the assignment of the content identifier is carried out with the aid of modern computing equipment the operation becomes automatic indexing. It has been suggested that the subject of a document can be derived by a mechanical analysis of the words in a document and by their

NOTES

arrangement in a text. In fact, all attempts at automatic indexing depend in some way or other on the text of the original document or its surrogates. The words occurring in each document are examined and substantive words are selected through statistical measurements (like word frequency calculation, total collection frequency, or frequency distribution across the documents of the collection) by the computer.

However, the use of computers in generating indexes of documents started from KWIC indexing developed by H.P. Luhn.

The idea of analyzing the subject of a document through automatic counting of term occurrences was first put forward by H P Luhn of IBM in 1957. He proposed that:

- (a) The frequency of word occurrence in a text of the document furnishes a useful measure of word significance;
- (b) The relative position of a word within sentence furnishes a useful measurement for determining the significance of sentences; and
- (c) The significance factor of a sentence will be based on a combination of these two requirements.

The basic idea behind Luhn's automatic indexing was based on word extraction, that is, keywords were extracted from the text by counting the frequency of occurrence of words in a given document. Here, the computer was used to scan the text with the object of counting the words or phrases that occur most frequently in a machine-readable document, and the extraction programs select the words or phrases that occur most frequently to represent the subject-matter of the document. A 'stop word' list was first used to eliminate the common and non-substantive words. The system pioneered by Luhn was relatively effective and the words or phrases selected by computer were quite similar to those, which would be extracted by a human indexer.

In the early 1960s, some other attempts were made at implementing automatic indexing systems. These consisted in using the computer to scan document texts, or text excerpts such as abstracts, and in assigning as content descriptor words that occurred sufficiently frequently in a given text. A less common approach uses relative frequency in place of absolute frequency. In relative frequency approach, a word is extracted if it occurs more frequently than expected in a particular corpus. Thus in a document on 'Aerodynamics' the word 'Air Craft' and the word 'Wing' might be rejected, even though they are the most frequently occurring words in the document, and the word 'Flutter' might be selected even though, in absolute terms, it is not a high frequency words. Other approaches to automatic indexing use other types of extraction criteria in place of, or along with the statistical criteria, word position in the document, word type, or even the emphasis placed on words in printing—(e.g. boldface and italics)—may all be used as the basis for selection. Subsequently linguistics led the way by pointing out

that a number of linguistic processes were essential for the generation of effective content identifiers characterizing natural language texts.

An ideal computerized indexing is one that has the ability to create and modify new subject terms mechanically, by minimizing or without the help of human intellectual efforts. As computer can understand only machine code, so it is necessary to translate the information into machine code and in a fixed machine- readable format. Usually, the titles and abstracts are used for the purpose of computerized indexing. However there are two assumptions:

- (a) There is a collection of documents; each contains information on one or several subjects.
- (b) There exists a set of index terms or categories from which one or several of them can describe/represent the subject content of every document in the collection.

NOTES

Manual Indexing vs. Computerized Indexing

The differences between the manual and computerized indexing is given below in Table 4.7:

Table 13.1 Manual Indexing vs. Computerized Indexing

Manual Indexing	Computerized Indexing
Identifying and selecting keywords	Keywords and/or phrases denoting the subject the tile, abstract and full text of the matter of the document are extracted only from document to represent its content the title and abstract rather than the document's full text.
Content analysis of the document	The computer does content analysis by purely a mental process and carried out following the human instructions in the form by the human indexer of a computer programming.
Human indexer makes inferences	Computer cannot think and draw inferences judgment in selecting index terms like human indexer and as such, it can select judiciously. or match keywords, which are provided as input text.
Human indexer selects and excludes	It is possible to instruct a computer through index terms on the basis of semantic, proper programming to select, or exclude a syntactical as well as contextual term by following the rules of semantic, considerations. syntactical and contextual connotations, like human indexer.
Scanning, analyzing the critical views	Computer cannot do this. It involves less understanding the concepts and using intellectual effort, indexer's own subject knowledge and previous experience do indexing.
Selected index terms less in number	Selected index terms are more in number
It is time consuming	It takes less time.
It is expensive	Index entries can be produced at lower cost.
It is very difficult to maintain consistency	Consistency in indexing is maintained. in indexing.

NOTES

Methods of Computerized Indexing

In the context of image database systems, a method for indexing and retrieval of images by their color content and the spatial distribution of color is disclosed. The method is implemented as a software tool that runs on a personal computer and enables the computer to find a desired image from an image database. The user interface allows the user to describe the desired image, and the tool searches the repository for any images that satisfy the description. The description of the composite image may include information on shapes, texture, presence or absence of objects, and color. The process of search and retrieval of images makes use of a method that determines whether a specific color, or a combination of colors, is present in an image. Further, the method determines the location where the specific color or color combination can be found. By analysing the results of comparisons between the features found in the composite image and in the stored images, the tool retrieves those images that satisfy the description of the query, image. Calculation of color-related features makes significant use of the calculations that are made in the process of image compression which is an essential step prior to image storage. The process of search and retrieval on the basis of features relating to color are provided via a library of routines that perform the necessary tasks. We have discussed some of these concepts in the earlier units. Listed below are the methods for computerized indexing:

Keyword Indexing

An indexing system without controlling the vocabulary may be referred as 'Natural Language Indexing' or sometimes as 'Free Text Indexing'. Keyword indexing is also known as Natural Language or Free Text Indexing. 'Keyword' means catch word or significant word or subject denoting word taken mainly from the titles and / or sometimes from abstract or text of the document for the purpose of indexing. Thus keyword indexing is based on the natural language of the documents to generate index entries and no controlled vocabulary is required for this indexing system. Keyword indexing is not new. It existed in the nineteenth century, when it was referred to as a 'catchword indexing'. Computers began to be used to aid information retrieval system in the 1950s. The Central Intelligence Agency (CIA) of USA is said to be the first organization to use the machine-produced keywords index from Title since 1952. H P Luhn and his associates produced and distributed copies of machine produced permuted title indexes in the International Conference of Scientific Information held at Washington in 1958, which he named it as Keyword-In-Context (KWIC) index and reported the method of generation of KWIC index in a paper. American Chemical Society established the value of KWIC after its adoption in 1961 for its publication 'Chemical Titles':

o KWIC (Keyword-In-Context) Index

As told earlier, H P Luhn is credited for the development of KWIC index. This index was based on the keywords in the title of a paper and was

produced with the help of computers. Each entry in KWIC index consists of following three parts:

- (a) **Keywords:** Significant or subject denoting words which serve as approach terms;
- (b) **Context:** Keywords selected also specify the particular context of the document (i.e. usually the rest of the terms of the title).
- (c) **Identification or Location Code:** Code used (usually the serial numbers of the entries in the main part) to provide address of the document where full bibliographic description of the document will be available.

NOTES

The operational stages of KWIC indexing consist of the following:

- (a) Mark the significant words or prepare the 'stop list' and keep it in computer. The 'stop list' refers to a list of words, which are considered to have no value for indexing/retrieval. These may include insignificant words like articles (a, an, the), prepositions, conjunctions, pronouns, auxiliary verbs together with such general words as 'aspect', 'different', 'very', etc. Each major search system has defined its own 'stop list';
- (b) Selection of keywords from the title and / or abstract and / or full text of the document excluding the stop words;
- (c) KWIC routine serves to rotate the title to make it accessible from each significant term. In view of this, manipulate the title or title like phrase in such a way that each keyword serves as the approach term and comes in the beginning (or in the middle) by rotation followed by rest of the title;
- (d) Separate the last word and first word of the title by using a symbol say, stroke [/] (sometime an asterisk '*' is used) in an entry. Keywords are usually printed in bold type face;
- (e) Put the identification / location code at the right end of each entry; and finally
- (f) Arrange the entries alphabetically by keywords.

Let us take the title 'control of damages of rice by insets' to demonstrate the index entries generated through KWIC principle:

- o Control of damages of rice by insets 118
- o Damages of rice by insets / Control of 118
- o Insets / Control of damages of rice by 118
- o Rice by insets / Control of damages of 118

In the computer generated index, the keywords can be positioned at centre also.

Variations of KWIC

Two important other versions of keyword index are KWOC and KWAC, which are discussed below:

NOTES

- **KWOC (key-word out-of-context) Index**

The KWOC is a variant of KWIC index. Here, each keyword is taken out and printed separately in the left hand margin with the complete title in its normal order printed to the right. For examples:

- o Control of damages of rice by insets 118
- o Damages Control of damages of rice by insets 118
- o Insets Control of damages of rice by insets 118
- o Rice Control of damages of rice by insets 118

Sometime, keyword is printed as heading and the title is printed in the next line instead of the same line as shown above. For examples,

- o Control of damages of rice by insets 118
- o Damages Control of damages of rice by insets 118
- o Insets
- o Control of damages of rice by insets 118
- o Rice
- o Control of damages of rice by insets 118

- **KWAC (key-word Augmented-in-context) Index**

KWAC also stands for 'key-word-and-context'. In many cases, title cannot always represent the thought content of the document co-extensively. KWIC and KWOC could not solve the problem of the retrieval of irrelevant document. In order to solve the problem of false drops, KWAC provides the enrichment of the keywords of the title with additional keywords taken either from the abstract or from the original text of the document and are inserted into the title or added at the end to give further index entries. KWAC is also called enriched KWIC or KWOC. CBAC (Chemical Biological Activities) of BIOSIS uses KWAC index where title is enriched by another title like phrase formulated by the indexer.

- o **Other Versions:** A number of varieties of keyword index are noticed in the literature and they differ only in terms of their formats but indexing techniques and principle remain more or less same. They are
 - (i) KWWC (Key-Word-With-Context) Index, where only the part of the title (instead of full title) relevant to the keyword is considered as entry term.
 - (ii) KEYTALPHA (Key-Term Alphabetical) Index. It is permuted subject index that lists only keywords assigned to each abstract. Keytalpa index is being used in the 'Oceanic Abstract'.

- (iii) WADEX (Word and Author Index). It is an improved version of KWIC index where the names of authors are also treated as keyword in addition to the significant subject term and thus facilitates to satisfy author approach of the documents also. It is used in 'Applied Mechanics Review'. AKWIC (Author and keyword in context) index is another version of WADEX.
- (iv) DKWTC (Double KWIC) Index. It is another improved version of KWIC index.
- (v) KLIC (Key-Letter-In-Context) Index. This system allows truncation of word (instead of complete word), either at the beginning (i.e., left truncation) or at the end (i.e. right truncation), where a fragment (i.e., key letters) can be specified and the computer will pick up any term containing that fragment. The Chemical Society (London) published a KLIC index as a guide to truncation. The KLIC index indicates which terms any particular word fragment will capture.
- o **Uses of Keyword Index:** A number of indexing and abstracting services prepare their subject indexes by using keyword indexing techniques. They are nothing but the variations of keyword indexing apart from those mentioned above. Some notable examples are:
 - o Chemical Titles;
 - o BASIC (Biological Abstracts Subject In Context);
 - o Keyword Index of Chemical Abstracts;
 - o CBAC (Chemical Biological Activities);
 - o KWIT (Keyword-In-Title) of Laurence Berkeley Laboratory;
 - o SWIFT (Selected Words in Full Titles); and
 - o SAPIR (System of Automatic Processing and Indexing of Reports).

Advantages

- The principal merit of keyword indexing is the speed with which it can be produced;
- The production of keyword index does not involve trained indexing staff. What is required is an expressive title coextensive to the specific subject of the document;
- Involves minimum intellectual effort;
- Vocabulary control need not be used; and
- Satisfies the current approaches of users.

Disadvantages

- Most of the terms used in science and technology are standardized, but the situation is different in case of Humanities and Social Sciences. Since no

NOTES

NOTES

controlled vocabulary is used, keyword indexing appears to be unsatisfactory for the subjects of Humanities and Social Sciences;

- Related topics are scattered. The efficiency of keyword indexing is invariably the question of reliability of expressive title of document as most such indexes are based on titles. If the title is not representative the system will become ineffective, particularly in Humanities and Social Science subjects;
- Search of a topic may have to be done under several keywords;
- Search time is high;
- Searchers very often lead to high recall and low precision; and
- Fails to meet the exhaustive approach for a large collection.

Other Methods of Automatic Indexing

Since the KWIC indexing methods various methods generating automatic indexes have been tried. In fact, all attempts at computerized indexing were based on two basic methods: Statistical analysis; and Syntactic and semantic analysis. These are discussed below:

(a) Statistical Analysis

The statistical analysis methods are based on the hypothesis that occurrence of a word in the text indicates its importance. On the basis of this hypothesis a prediction can be made about the subject terms that can be assigned to the document. The computer program can list all the words in a document. The words are grouped by number of occurrences and arranged alphabetically within each frequency. Generally articles, conjunctions, prepositions and pronouns are excluded using a 'stop list' file. Words having same stem can be counted either as the same or as different words. The following methods are adopted in measuring the word significance:

- **Weighting by Location:** A word appearing in the title might be assigned a greater weight than a word appearing in the body of the work.

(b) Relative Frequency Weighting

This is based upon the relation between the number of times the words is used in the document being indexed and the number of times the same word appears in the sample of other documents.

(c) Use of Noun Phrase

Noun and adjective noun phrases can be selected as index terms and these are selected from the title or abstract of the document.

(d) Use of Thesaurus

A thesaurus can be used to control synonyms and otherwise related terms. In this way, the count of some word types increases as is the separation between 'good' and 'poor' index terms.

(e) Use of Association Factor

By means of statistical association and correlation techniques, the degree of term relatedness, that is, the likelihood that that two terms will appear in the same document, is computed and used for selecting index terms.

(f) Maximum-depth Indexing

This procedure indexes a document by all of its content words and weights these words, if desired, by the number of occurrences in the document. In this way, the problem of selecting term is avoided.

(h) Syntactic and Semantic Analysis

Among the linguistic techniques of interest, the syntactical and semantic analyses are most important in the development of information analysis system needed for computerized indexing. According to Salton, most information analysis systems are based on the recognition of certain key elements, often chosen from a pre-constructed list of acceptable terms, and on the determination of rules by which these basic elements are combined into larger units. The syntactical analysis identifies the role of the word in the sentence, that is, its grammatical class (i.e., parts of speech) and relation among words in the sentence. Whereas semantic analysis helps to establish the paradigmatic or class relations among terms so as to associate words with simple concepts. The main objective of semantic analysis is to identify subject and content bearing words of the document or surrogate text.

Among the linguistic techniques of interest, the following were considered to be of significant:

- (a) Use of hierarchical term arrangements relating to the content terms in the given subject area can help to expand the standard content description by adding superordinate and/or subordinate terms to a given content description.
- (b) Use of synonym dictionaries or thesauri can help to broaden the original context description through a complete class of related terms.
- (c) Use of syntactical analysis systems capable of specifying syntactic roles of each term and of forming complex content descriptions consisting of term phrases and large syntactic units. A syntactic analysis scheme makes it possible to supply specific content identification.

Use of semantic analysis systems in which the syntactic units are supplemented by semantic roles attach to the entities making up a given content description. Semantic analysis systems utilize various kinds of knowledge extraneous to the documents, often specified by pre-constructed 'semantic graphs' and other related constructs.

Advanced linguistic techniques, that is, the application of computer to analyse the structure and meaning of language led by Noam Chomsky. The linguistic model proposed by Chomsky distinguishes between surface structure and deep structure

NOTES

NOTES

of a language. By means of transformational grammar, a structure can go through a series of transformations that will exhibit the deep structure. Chomsky found that a purely syntactic transformation could provide a semantic interpretation of the sentence.

Advantages

The advantages of computerized indexing are manifold like level of consistency in indexing can be maintained; index entries can be produced at a lower cost in the long run; indexing time can be reduced; and better retrieval effectiveness can be achieved.

Disadvantages

The main criticism against computerized indexing centers round the fact that a term occurs several times in a document may not always be regarded as a significant term.

File Organization

Organization of files in manually operated libraries are expensive, time consuming, labourious and error-prone. Moreover, manual organization of data/files often leads to duplication of data or data redundancy. Various files maintained in traditional library system are sequential in nature. For example, catalogue cards are arranged in order of search keys (author, title, subject, series etc.) in the form of access points transcribed at the top of the cards.

In computerized indexing system, data elements are stored in suitable digital media and it is possible to manipulate, retrieve and view data elements quite easily. The efficiency of such computerized indexing system largely depends on its file structure.

File organization forms an important element in computerized indexing. File organization is a technique for physically arranging the records of a file on secondary storage devices. This is the technique for organization of the data of a file into records, blocks and access structures.

A file contains data that is required for information processing. These data are about entities. An entity is anything about which data can be stored (e.g., book).

The essential properties of an entity are called attributes (e.g. author, title, edition, etc. are attributes of the entity book). Each attribute of an entity is represented in storage by a data item. A data item is assigned a name in order to refer to it in storage, processing and retrieval. Data items are usually grouped together to describe an entity. The data representation in storage of each instance of an entity is commonly called as a record. A collection of related records is called a file.

When data are stored on auxiliary storage devices (e.g. hard disk), the chosen method of file organization will determine how the data can be accessed. The organization of data in a file is influenced by a number of factors, but the most important among them is the time required to access a record and to transfer the data to the primary storage or to write a record or to modify a record.

A schematic view of file organization techniques is represented in Figure

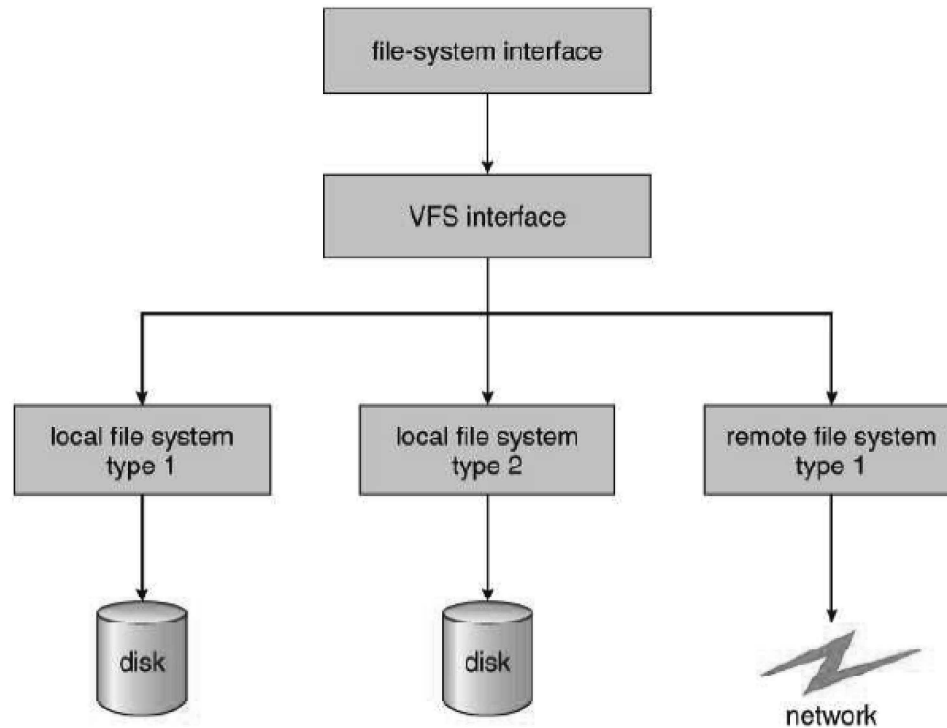


Fig. 13.1 Schematic View of File Organization

- (1) **Sequential File Organization:** In this technique, records are stored in some predetermined sequence, one after another. One field, referred to as the primary key, usually determines their sequence of order.
- (2) **Direct File Organization:** This technique supports direct access (also called random access), in which records can be accessed instantaneously and in any order from the data scattered throughout the disk.

Relative Addressing: It is the simplest method of finding a record. Here, a record's primary key is associated with a specific physical storage location and contents of the records are stored in this address.

Hashing: It is a method for determining the physical location of a record. Here, the record key is processed mathematically, and another number is computed that represents the location where the record will be stored.

Indexing: It is a procedure for locating a record in a file stored randomly throughout the disk. Here, a primary index associates a primary key with the physical location in which a record is stored.

NOTES

NOTES

Ordered Index: It is based on a sorted ordering of values.

Primary Index: The records in the indexed file can be stored in some sorted order. If the file containing the records is sequentially ordered, the index whose search key specifies the sequential order of the file is called the primary index

Secondary Index: Indexes whose search key specifies an order that is different from the sequential order of the file are called secondary indexes.

B-Tree Indexes: It takes the form of a balanced tree in which every path from the root to the tree leaf is of the same length. It eliminates the redundant storage of search key values.

Hashed Index: It is based on the values being uniformly distributed using a mathematical function called hash function.

Indexing Systems Using AI Techniques

Artificial Intelligence (AI) is a branch of computer science concerned with study and creation of computer systems that exhibits some form of intelligence, systems that can learn new concepts and tasks, systems that reason and draw useful conclusions about the world around us, systems that can understand a natural language or perceive or comprehend a visual scene and systems that perform other types of feats that require human types of intelligence. The research in AI applications to information retrieval is gaining much importance. The recent development of the AI applications in the areas of which only those concerned with the subject indexing system is discussed below:

NLP-Based Subject Indexing System

Natural Language of Processing (NLP), one of the areas of AI, has gained momentum in research activity that explores how natural language text that is entered into a computer system can be manipulated and transformed into a form more suitable for further processing for improved storage and retrieval of information. NLP offers a potential viable alternative to statistical techniques in computerized indexing. Most automatic retrieval systems based on NLP techniques, convert the contents of the document files and user's queries in an internal form and the task of matching takes place at that level of the system, which is known as natural language interfaces, or front-end systems. At the core of any NLP technique there lays an important issue of natural language understanding. The process of building computer programs that understand natural language requires knowledge of how the words are formed, how the words in turn form clauses and sentences. In general, the different levels of knowledge that is to be used in natural language understanding falls into the following groups:

- (1) **Morphological Knowledge:** This level gives knowledge of word formation and deals with the morphological structure of words like the word root, prefix, suffix and infixes. The basic unit in a written word is a morpheme.

- (2) **Lexical Knowledge:** A lexicon consists of words considered as valid in the given domain. The lexicon may also contain syntactic markers or certain categories, which can be useful in processing. This level deals with thesaurus look up, spelling variations, acronyms and abbreviations, etc.
- (3) **Syntactic Knowledge:** Syntax deals with the structural properties and validity of input sentences—how a right combination of words in a particular sequence constitutes a valid sentence.
- (4) **Semantic Knowledge:** Semantics deals with the meaning of words and that of sentences. Different methods of representation of meaning have been developed over the years.
- (5) **Pragmatic Knowledge:** A given concept may occur in a number of different meanings, and to decide about the correct meaning in a given context the NLP system needs pragmatic knowledge. Pragmatic level deals with sentences in a particular context. This requires a higher-level knowledge, which relates to the uses of sentences in different contexts. This knowledge is useful because it helps to eliminate ambiguities and supplements the semantic representation.
- (6) **World Knowledge:** In order to carry out effective communication, both the communicator and communicatee should have background knowledge, either to send or to receive a message, without any noise. This background knowledge is considered as the world knowledge of a particular domain.

Chomsky's classification of the types of grammar formalism may be correlated with the above mentioned levels of knowledge. A NLP system having pragmatic and world knowledge constitutes his 'type 0' grammars; systems with semantic knowledge or context sensitive grammars constitute his 'type 1' grammars; systems with syntactic knowledge or context free grammars constitute 'type 2' grammars; and systems with most restrictive or regular grammars constitute 'type 3' grammars. One method to formalize our knowledge is to provide a series of rewrite rules (known as grammars) that will generate the legal sentences of the language. Of the above mentioned grammars, context free grammars or 'type 2' grammars, introduced by Noam Chomsky, are well understood from the computational point of view. In particular, the Phrase structure grammars (PSGs) have been used extensively in developing parsers. Another variant of it, called Definitive Clause Grammars (DCGs) is the basis for a programming language. PROLOG (Programming in Logic) and LISP (List Processing) are most popular languages in artificial intelligence programming.

For syntax analysis in NLP, valid sentences of a language are recognized and their underlying structures are determined. A central component of the syntactic phase is the parser (the term parse is derived from the Latin phrase *pars orationis* which means 'part of speech'), a computational process that takes individual sentences or connected texts and converts them to some representational structure useful for further processing. Parsing refers to the use of syntax to determine the

NOTES

NOTES

functions of the words in the input sentences in order to create a data structure that can be used to get at the meaning of the sentence. The major objective of the parsing is to transform the potentially ambiguous input phrase into an unambiguous form as an internal representation. A few of the notable computational models of the parsers are Finite State Transition Networks (FSTN), Recursive Transition Networks (RTN), Augmented Transition Network (ATN), Definite Clause Grammar (DCG), etc. There are many computational models of semantic grammars like Conceptual Dependency Grammars (CDG), Modular Logic Grammars (MLG), Lexical Functional Grammars (LFG), Case Grammars (CG), etc.

NLP has progressed to the point at which assignment indexing by computers should not be possible. Vleduts-Stokolov, for example, described an experiment performed at BIOSIS in which terms from a limited vocabulary of 600 biological concept headings were assigned automatically to journal articles. The assignment is achieved by matching article titles against a semantic vocabulary containing about 15,000 biological terms, which, in turn, were mapped to the concept headings. In NLP-based subject indexing, phrases are automatically extracted from texts to serve as content indicators. Nonsyntactic methods are based on simple text characteristics, such as word frequency and word proximity, while syntactic methods selectively extract phrases from parse trees generated by an automatic syntactic analyser.

It seems very likely that we will see increased emphasis on the use of natural language in information retrieval in future. This claim seems justified due to the following factors:

- (a) The continued growth in the availability of machine-readable databases many of which are in natural language form;
- (b) The continued expansion of online system, which is likely, eventually, to put terminal in the offices and houses of scientists and other professionals. Bibliographic searching in one of many possible applications of these terminals and natural language mode of searching seems imperative in this type of application;
- (c) A number of evaluation studies have indicated that natural language may offer several advantages over controlled vocabularies in many retrieval systems;
- (d) Natural language indexing systems have been shown to work, and to work well, in the legal field, the scientific information dissemination centres, the defence and intelligence communities; and
- (e) New development in computer storage devices will make the storage of very large text files increasingly feasible.

The development of the AI techniques in other two important areas concerned with the subject indexing system includes the following:

Knowledge Representation-based Subject Indexing System

Knowledge representation connotes a way of codifying knowledge. The arrangement and representation of knowledge as reflected in classification schemes and thesauri may be treated as one form of codification of knowledge. The codification helps computer to store, process and inference the codified knowledge. A computer system capable of understanding a message in natural language would seem to require both contextual knowledge and the processes for the inferences assumed by the generator. Some progress has been made towards computer systems of this sort, for understanding spoken and written fragments of language. Fundamental to the development of such systems are certain ideas about structures of contextual knowledge representation and certain techniques for making inferences from that knowledge. Based on the idea that knowledge is symbolic, the methods of using NLP tools and techniques include use of symbols and symbol structures, semantics and roles, categories and relations present in the subject statement. The work in the area of 'semantic primitives' or 'semantic factoring' leads to several means of representing knowledge for reasoning, including symbolic equations, frames and semantic nets.

Expert System-based Subject Indexing

An Expert system is the embodiment, within the computer, of a knowledge-based component derived from an expert in such a form that the computer can offer intelligent advice or take an intelligent decision about the processing function. An expert system consists of knowledge acquisition, knowledge base system, inference machine, and user interface. This is one of the major areas of AI with a wide application to information processing and retrieval. The researches in this area lead to the development of expert systems, that should be fruitful in solving the problems of indeterminacy in both indexing and term selection for retrieval. Expert systems were designed to assist the user in query formulation and selection of relevant documents. A number of studies were undertaken in devising expert systems to aid the process of subject indexing. One notable example is the Medical Indexing Expert System or MedIndEx System (previously referred to as the Indexing Aid System) being developed at National Library of Medicine (NLM), USA. Referred to as a 'prototype interactive knowledge-based system', it uses an experimental frame language and is designed to interact with trained MEDLINE indexers by prompting them to enter MeSH terms as 'slot fillers' in completing document-specific indexing frames derived from the knowledge-base frame. Another expert system for machine-aided indexing used in the Central Abstracting and Indexing Service (CAIS) of the American Petroleum Institute (API) incorporates a rule-based expert system founded on the API thesaurus, which has been in operation since 1985. Terms automatically generated by the system (by matching abstracts against a knowledge-base derived from the API thesaurus) and by API's human indexers are reviewed by a human editor, and edited terms are added to print indexes and the online index. At present the base contains 14,000 text rules.

NOTES

NOTES

User Interface Design

User interface performs two major tasks – search or browse an information collection and display of search results. It is also designed to perform other related tasks such as sorting, saving and/or printing of search results and modification of search query. Obviously, the success of any information retrieval system largely depends on the efficient and effective design of user interface.

Shneiderman proposes the following guiding principles for the design and development of user interface for computerized information retrieval:

- Strive for consistency in terminology, layout, instructions, fonts and colour
- Provide shortcuts for skilled users
- Provide appropriate and informative feedback about the sources and entity that is being searched for
- Design of message or notification system to indicate end of search process
- Permit reversal of actions so that users can undo or modify actions
- Allow users to monitor the progress of a search
- Permit users to specify the parameters to control a search
- Help users to recall their search queries
- Provide extensive online help and error-handling facilities
- Error messages should be clear and specific
- Provide facilities to enter long queries and use of Boolean, relational and positional operators
- Provide alternative interface for novice and expert users
- Provide multilingual interface if required

Information retrieval systems vary in terms of design, objectives, characteristics, contents, and users. However, the above guidelines with the help of following technical features will help to design an effective user-centered information retrieval system :

1. The presentation layer must reside within the client. Windows clients must contain an applications layer.
2. The system proposed must store all help files on the client for immediate retrieval/customization.
3. The system proposed must enable the operator to request and receive context- sensitive help for the command in use with a single keystroke.
4. The client(s) proposed must enable the operator to page backwards and forwards through the help text, and include hypertext links to related topics, screen shots, examples, etc., and the ability to access an index of help topics.

5. The clients proposed must enable the operator to transfer any data field from one screen to another.
6. Wizards must guide the operator through a series of steps necessary to complete a defined process, without the use of commands or traditional menus.
7. Windows clients must perform all edit checking of user input before sending input to the server.
8. The clients proposed must display a form with all appropriate fields on the screen when an operator initiates a command.
9. The clients proposed must enable the user to type data onto the screen at the current cursor position.
10. The clients proposed must enable the user to use delete and insert character keys to correct mistakes.
11. The clients proposed must retain work forms on the screen, until the operator changes to another command.
12. If an error is detected, the system must report the error on the screen, leaving the form and the operator's input otherwise intact.
13. The system proposed must display an explanatory error message whenever the operator has provided inappropriate input.
14. The system proposed must set up a session with a user, which records user's search history and allows users to re-execute their previous searches against the same server or a different server or servers.
15. The system proposed must allow restriction of access to local or remote databases based upon the user's login and password.
16. The system must retain a user's authorization as he or she moves from one database to another within one session.
17. The system must allow both access by specific IP without sign-on and sign- on capability for users from outside the allowed IP range.
18. The proposed Web client must support creation and execution of simple or complex searches.
19. The proposed Web client must support browsing (SCAN. and selecting terms from standard vocabulary device.
20. The proposed Web client must support hypertext searching for related items and should support display of cross-reference information.
21. The proposed Web client must support sorting of search results to user- defined sequence.
22. The proposed Web client must support record export to printer, local file or e-mail.
23. The proposed Web client must support linking to library holdings information on multiple servers via the Z39.50 protocol.

NOTES

NOTES

24. The proposed Web client must support execution of previous searches from stored search history.
25. The proposed Web client must allow authentication of users by user ID and optional personal identification number, and permit authenticated users to:
 - (a) Access their own user accounts to view the status of charges, holds, fines, and bills associated with their account; and
 - (b) Obtain access to additional databases, such as licensed journal citation and other reference databases closed to anonymous users.
26. The system proposed must enable an authorized user to:
 - (a) Save records to the review file one at a time, or in stated ranges;
 - (b) View records stored in a review file using Vendor's client software;
 - (c) Remove records from the review file;
 - (d) Duplicate records in the review file; and
 - (e) Print records in the review file;

Check Your Progress

1. State the basic idea behind Luhn's automatic indexing.
2. What are two assumptions of the computerized indexing?
3. What does an expert system consist of?

13.3 WEB ONTOLOGY LANGUAGE (OWL)

SOURCE WIKIPEDIA

Web Ontology Language (OWL) is referred to as a family of knowledge representation language for authoring different ontologies. The structure of knowledge for different domains or the information about the classification of networks and taxonomies is known as Ontologies. They are used to improve the web searches accuracy. This is a more specific tool to probably allow search engines to retrieve only those web pages which discuss the concept in particular rather than all the pages mentioning ambiguous keywords. For instance, nouns are used as the different classes of objects and verb is used to represent the relation between different objects. The ontologies might share similarities with class hierarchies in object-oriented programming but there are many significant distinctions between them like regular updates are essential to the web ontologies since it contains description of information on the internet unlike other data structure. These are also more flexible since their sources are heterogenous. Class hierarchies are mostly static and have stable data sources like the corporate databases.

Formal semantics guide the OWL languages. In fact, the Resource Description Framework, which is the XML standard for objects of the World Wide Web Consortium (W3C) is the crucial foundation to the OWL languages. The subject has garnered varied interest ranging from the medical, commercial and academic researchers.

Different semantic editing platforms like Pellet, RacerPro, FaCT++ and Hermit have assisted in the meaningful expansion of the OWL in 2007 and in 2009.

There are different serializations and species which share the same names in the OWL. This is to mention the point that unless taking about rather exclusive features, the common word OWL word is used to represent the subparts of the family in common use. And the OWL and OWL2 represent particular specifications from the years 2004 and 2009 respectively

Different Ontology Languages for the Web:

The year 2000 was when the development of DAML (Distributed Agent Markup Language) by DARPA led by James Hendler begun in the United States. It was later next year that the Joint EU/US Committee on Agent Markup Languages suggested the merger of the DAML with OIL. This project was now starting to be developed by a EU/US ad hoc Joint Working Group on Agent Markup Languages with funding from the DARPA (under the DAML program) and the European Union's Information Society Technologies (IST) funding project. Considered to be a above the RDFS, the DAML+OIL worked with formal semantics based on a description logic (DL). DAML+OIL is the major design influence for OWL.

Semantic web standards

Providing a common framework that permits the sharing of data and its reuse across application, enterprise, and community boundaries is the main function of the semantic web standards.

Acronym

Even though the phrase Web Ontology Language is abbreviated as WOL, the word Owl is made more popular only because its pronunciation and appeal as a word is far more easy as would its representation on logos which will also be the symbol for honor and wisdom along with honoring William A. Martin's *One World Language* knowledge representation project.

Adoption

A study conducted in the year 2006 of ontologies available on the web collected 688 OWL ontologies. Of these, 199 were OWL Lite, 149 were OWL DL and 337 OWL Full (by syntax). They discovered during this study that 19 ontologies had in excess of 2,000 classes, and that 6 had more than 10,000. The same survey collected 587 RDFS vocabularies.

NOTES

NOTES

An ontology is an explicit specification of a conceptualization.

A set of "individuals" and a set of "property assertions" which relate these individuals to each other is referred to as the interpretation of the OWL data. An ontology comprises of a set of axioms which restricts sets of individuals (called "classes") and the types of relationships permitted between them. Utilizing the explicitly provided data, the role of the axioms is to provide semantics by allowing systems to infer additional information.

OWL ontologies can import other ontologies, adding information from the imported ontology to the current ontology.

Example

An ontology defining families could perhaps include axioms declaring that a "hasMother" property is merely present between two individuals when "hasParent" is also present, and that moreover individuals of class "HasTypeOBlood" are never related via "hasParent" to members of the "HasTypeABBlood" class. If it is thus declared that the individual Harriet is related through "hasMother" to the individual Sue, and that Harriet is a member of the "HasTypeOBlood" class, then it can be deduced that Sue is not a member of "HasTypeABBlood".

Species

OWL sublanguages

Three variants of the OWL are included in the W3C-endorsed OWL specification each characterized by different levels of expressiveness. In the increasing order of expressiveness these are OWL Lite, OWL DL and OWL Full. It is crucial to remember here that the simpler previous form of these sublanguages form the syntactic extension of the next. Observe here that the following set of relations hold. Their inverses do not.

- Every legal OWL Lite ontology is a legal OWL DL ontology.
- Every valid OWL Lite conclusion is a valid OWL DL conclusion.
- Every legal OWL DL ontology is a legal OWL Full ontology.
- Every valid OWL DL conclusion is a valid OWL Full conclusion.\

OWL DL

The maximum expressiveness possible while retaining computational completeness (either φ or $\neg\varphi$ holds), decidability (there is an effective procedure to determine whether φ is derivable or not), and the availability of practical reasoning algorithms are provided by OWL DL. It comprises of all OWL language constructs, under the condition that certain restrictions apply (for example, number restrictions may not be placed upon properties which are declared to be transitive). OWL DL gets its name from its description logic, since OWL DL is a field of research that has studied the logics that form the formal foundation of OWL.

OWL Lite

Simple constraints and need for classify hierarchy drove the development of OWL Lite. For example, it just about permits cardinality values of 0 or 1. It was hoped that it may be able to simpler to provide tool support for OWL Lite than its more expressive relatives, allowing quick migration path for systems using thesauri and other taxonomies. In reality, little more than syntactic inconvenience crop up due to most of the expressiveness constraints placed on OWL Lite like most of the constructs available in OWL DL can be built using complex combinations of OWL Lite features, and is equally expressive as the description logic. Its development is no less difficult than that of tools for OWL DL, and further OWL Lite is not widely used.

OWL Full

Using totally distinct semantics from OWL Lite or OWL DL, OWL full was designed keeping in mind the use and preservation of the RDF Schema. A class can be treated simultaneously as a collection of individuals and as an individual in its own right in OWL full which is not allowed in OWL DL. OWL Full permits an ontology to support the meaning of the pre-defined (RDF or OWL) vocabulary. Since no reasoning software is successful in undertaking complete reasoning for it, OWL full is often called undecidable.

OWL2 profiles

In OWL 2, there are three sublanguages of the language. OWL 2 EL is a fragment that has polynomial time reasoning complexity; OWL 2 QL is designed to enable easier access and query to data stored in databases; OWL 2 RL is a rule subset of OWL 2.

Syntax

Several different syntaxes are compatible with the OWL family of languages supports. A distinction between the high level syntaxes used for specification from exchange syntaxes more suitable for general use is very important here.

High level

These are close to the ontology structure of languages in the OWL family.

OWL abstract syntax

The OWL ontology structure and semantics is specified through the high level. A series of *annotations*, *axioms* and *facts* are showcased by the OWL ontology. Annotations lift machine and human oriented meta-data. It is only in the axioms and facts that the information about the classes, properties and individuals that create the ontology are stored. Every class, property and individual is either *anonymous* or identified by an URI reference. Data either about an individual or about a pair of individual identifiers (that the objects identified are distinct or the same) are stated with the help of facts. Axioms specify the characteristics of classes

NOTES

and properties. This style shares similarities with the frame languages, and are different from the popular syntaxes for DLs and Resource Description Framework (RDF).

NOTES

Many researchers like Sean Bechhofer, *et al.* consider that though this syntax is hard to parse, it is quite concrete. They also think that the name *abstract syntax* may be somewhat misleading.

OWL2 functional syntax

OWL2 uses this syntax to specify semantics, mappings to exchange syntaxes and profiles on the.

Exchange syntaxes


OWL RDF/XML Serialization	
	
Filename extension	.owx, .owl, .rdf
Internet media type	application/owl+xml, application/rdf+xml ^[34]
Developed by	World Wide Web Consortium
Standard	OWL 2 XML Serialization October 27, 2009; 8 years ago, OWL Reference February 10, 2004; 14 years ago
Open format?	Yes

Fig.13.2 OWL RDF/XML Socialization

RDF syntaxes

Syntactic mappings into RDF are specified for languages in the OWL family. There have been the development of various RDF serialization formats which leads to a

syntax for languages in the OWL family through this mapping. RDF/XML is normative.

OWL2 XML syntax

OWL2 specifies an XML serialization that closely models the structure of an OWL2 ontology.

Manchester Syntax

A compact, human readable syntax with a style close to frame languages is the Manchester syntax. One can find that for OWL and OWL2 variations are also available. It must however be noted that not all OWL and OWL2 ontologies can be expressed in this syntax.

Examples

The W3C OWL2 Web Ontology Language provides syntax examples.

Tea ontology

Let's discuss a Tea class for an ontology for tea. First, there is a need for an ontology identifier. A URI is essential for the identification of every OWL ontology (<http://www.example.org/tea.owl>, say). This example provides a sense of the syntax.

OWL2 Functional Syntax

```
Ontology(<http://example.org/tea.owl>
  Declaration( Class( :Tea ) )
)
```

OWL2 XML Syntax

```
<Ontology ontologyIRI="http://example.org/tea.owl" ...>
  <Prefix name="owl" IRI="http://www.w3.org/2002/07/
owl#" />
  <Declaration>
    <Class IRI="Tea" />
  </Declaration>
</Ontology>
```

Manchester Syntax

```
Ontology: <http://example.org/tea.owl>
```

```
Class: Tea
```

RDF/XML syntax

```
<rdf:RDF ...>
  <owl:Ontology rdf:about="" />
  <owl:Class rdf:about="#Tea" />
</rdf:RDF>
```

RDF/Turtle

```
<http://example.org/tea.owl> rdf:type owl:Ontology .
:Tea rdf:type owl:Class .
```

NOTES

NOTES

Relation to description logics

The description logic concepts (DL) correspond to the OWL classes, similarly OWL properties to DL *roles*, while the same terminology is used for *individuals* in both the OWL and the DL.

In the beginning, IS-A was quite simple. Today, with the increase in the knowledge representation systems, there are almost as many meanings for this inheritance link as there are knowledge-representation systems.

Lack of clear definitions were harmful to the early attempts to build large ontologies. It is important to mention here that members of the OWL family have model theoretic formal semantics, and so have strong logical foundations.

OWL DL and OWL Lite semantics are based on DLs. They integrate a syntax for describing and exchanging ontologies, and meaning is given to them by formal semantics. For example, OWL DL corresponds to the description logic, while OWL 2 corresponds to the logic. Sound, complete, terminating reasoners (i.e. systems which are guaranteed to derive every consequence of the knowledge in an ontology) exist for these DLs.

Relation to RDFS

RDF Scheme is compatible with OWL Full and capable of supporting the meanings of existing Resource Description Framework (RDF) vocabulary. The extensions of the RDFS meaning define the OWL Full ontologies and OWL Full is a semantic extension of RDF.

Open world assumption

[The closed] world assumption implies that everything we don't know is *false*, while the open world assumption states that everything we don't know is *undefined*.

— *Stefano Mazzocchi, Closed World vs. Open World: the First Semantic Web Battle*

The open world assumption is used by the languages in the OWL family. Under the open world assumption, if a statement cannot be proven to be true with current knowledge, the conclusion that the statement is false cannot be certainly drawn.

Contrast to other languages

SQL is a query and management language for relational databases. This is a relational database consists of sets of tuples with the same attributes. Prolog is a logical programming language, both of which use the closed world assumption.

Terminology

Languages in the OWL family are capable of creating classes, properties, defining instances and its operations.

Instances

It refers to an object. It is a description logic *individual*.

Classes

It is a collection of objects. It may comprise of individuals, *instances* of the class. Any number of instances is possible and they may belong to one or more or no classes at all.

A class may be a *subclass* of another when a class inherits the characteristics from its parent *superclass* it is known as a subclass.

Extension or intention define the Class and their members in OWL. Class assertion can be explicitly assigned to an individual can be explicitly assigned a class. For insatnce, we can add a statement Queen elizabeth is a(n instance of) human, or by a class expression with ClassExpression statements every instance of the human class who has a female value to the sex property is an instance of the woman class.

Example

Let's call *human* the class of all humans in the world is a subclass of owl:thing. The class of all women (say *woman*) in the world is a subclass of *human*. Then we have owl:

The membership of some individual to a class could be noted

ClassAssertion(*human George_Washington*)

and class inclusion

SubClassOf(*woman human*)

The first means "George Washington is a human" and the second "every woman is human".

Properties

The directed binary relation that specifies some attribute which is true for instances of that class is the characteristic defined by 'properties'. It sometimes act as data values, or links to other instances. They exhibit logical features, for example, by being transitive, symmetric, inverse and functional and may also have domains and ranges.

Datatype properties

Instances of classes and RDF literals or XML schema datatypes are known as datatype properties. For example, modelName (String datatype) is the property of Manufacturer class. They are formulated using *owl:DatatypeProperty* type.

Object properties

The relations between instances of two classes are defined by object properties. For instance, ownedBy may be an object type property of the Vehicle class and may have a range which is the class Person. They are formulated using *owl:ObjectProperty*.

Operators

NOTES

NOTES

Various operations on classes such as union, intersection and complement are supported by the languages in the OWL family support. They also allow class enumeration, cardinality, and disjointness.

Metaclasses

It refers to the classes of classes. They are allowed in OWL full or with a feature called class/instance punning.

Public ontologies

Libraries

Biomedical

- OBO Foundry
- NCBO BioPortal
- NCI Enterprise Vocabulary Services

Miscellaneous

- MMI ORR (Marine Metadata Interoperability) Ontology Registry and Repository
- ESIP COR Earth Science Information Partners Community Ontology Repository
- ESIP Semantic Portal
- AgroPortal Agricultural Semantic Portal
- *"SchemaWeb - RDF Schemas Directory". September 2005. Archived from the original on 10 August 2011.*

Standards

- Suggested Upper Merged Ontology
- TDWG
- PROV-O, the ontology version of the W3C's PROV-DM
- Basic Formal Ontology

Browsers

The following tools include public ontology browsers:

- Protégé OWL

Search

- Swoogle

Other tools

OptiqueVQS -- an ontology-based visual query formulation tool

Limitations

- No direct language support for n-ary relationships. For example, modelers may wish to describe the qualities of a relation, to relate more than 2

individuals or to relate an individual to a list. This cannot be done within OWL. They may need to adopt a pattern instead which encodes the meaning outside the formal semantics.

Check Your Progress

4. What is referred to as the interpretation of the OWL data?
5. What is the sequence of OWL in the increasing order of their expressiveness?
6. What is the function of facts in web ontology?

NOTES

13.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The basic idea behind Luhn's automatic indexing was based on word extraction, that is, keywords were extracted from the text by counting the frequency of occurrence of words in a given document.
2. Two assumptions used in computer indexing are:
 - (a) There is a collection of documents; each contains information on one or several subjects.
 - (b) There exists a set of index terms or categories from which one or several of them can describe/represent the subject content of every document in the collection.
3. An expert system consists of knowledge acquisition, knowledge base system, inference machine, and user interface. This is one of the major areas of AI with a wide application to information processing and retrieval.
4. A set of "individuals" and a set of "property assertions" which relate these individuals to each other is referred to as the interpretation of the OWL data.
5. In the increasing order of expressiveness these are OWL Lite, OWL DL and OWL Full.
6. Data either about an individual or about a pair of individual identifiers (that the objects identified are distinct or the same) are stated with the help of facts.

13.5 SUMMARY

- In many literatures of Library and Information Science, the term 'automatic indexing' is interchangeably used with the term 'computerized indexing'. A fully automatic indexing system would be one in which indexing is conducted by computers, an internally generated thesaurus is prepared, and search

NOTES

strategies are developed automatically from a natural language statement of information need.

- In the early 1960s, some other attempts were made at implementing automatic indexing systems. These consisted in using the computer to scan document texts, or text excerpts such as abstracts, and in assigning as content descriptor words that occurred sufficiently frequently in a given text. A less common approach uses relative frequency in place of absolute frequency.
- In the context of image database systems, a method for indexing and retrieval of images by their color content and the spatial distribution of color is disclosed. The method is implemented as a software tool that runs on a personal computer and enables the computer to find a desired image from an image database.
- The process of search and retrieval of images makes use of a method that determines whether a specific color, or a combination of colors, is present in an image. Further, the method determines the location where the specific color or color combination can be found.
- In computerized indexing system, data elements are stored in suitable digital media and it is possible to manipulate, retrieve and view data elements quite easily. The efficiency of such computerized indexing system largely depends on its file structure.
- File organization forms an important element in computerized indexing. File organization is a technique for physically arranging the records of a file on secondary storage devices. This is the technique for organization of the data of a file into records, blocks and access structures.
- Artificial Intelligence (AI) is a branch of computer science concerned with study and creation of computer systems that exhibits some form of intelligence, systems that can learn new concepts and tasks, systems that reason and draw useful conclusions about the world around us, systems that can understand a natural language or perceive or comprehend a visual scene and systems that perform other types of feats that require human types of intelligence.
- Web Ontology Language (OWL) is referred to as a family of knowledge representation language for authoring different ontologies. The structure of knowledge for different domains or the information about the classification of networks and taxonomies is known as Ontologies.
- There are different serializations and species which share the same names in the OWL. This is to mention the point that unless taking about rather exclusive features, the common word OWL word is used to represent the subparts of the family in common use. And the OWL and OWL2 represent particular specifications from the years 2004 and 2009 respectively.

13.6 KEY WORDS

- **Fully automatic indexing system:** It is system in which indexing is conducted by computers, an internally generated thesaurus is prepared, and search strategies are developed automatically from a natural language statement of information need.
- **Web Ontology Language (OWL):** It is referred to as a family of knowledge representation language for authoring different ontologies. The structure of knowledge for different domains or the information about the classification of networks and taxonomies is known as Ontologies.
- **Semantic web standards:** It refers to the standards which provide a common framework that permits the sharing of data and its reuse across application, enterprise, and community boundaries is the main function of the semantic web standards.

NOTES

13.7 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What are the criticisms against the automatic indexing?
2. Briefly explain some of the automatic indexing techniques.
3. Write a short note on the adoption of OWL.
4. What are the different Ontology languages for the web?

Long-Answer Questions

1. Compare manual and automatic indexing.
2. Explain the development of user interface in the automatic indexing procedure.
3. Describe the indexing systems using AI techniques.
4. Explain the OWL sublanguages.
5. Discuss the properties of OWL with example.
6. Describe the OWL syntax.

13.8 FURTHER READINGS

- Alberico, R. and Micco M. 1990. *Expert systems for reference and Information retrieval*. West Port: Meckler.
- Atchison, J. & Gilchrist, A. 1972. *Thesaurus construction: a practical manual*. London: Aslib.

NOTES

Chowdhry, G.G.2003. *Introduction to modern Information retrieval*. 2nd Ed. London: Facet Publishing.

Cleaveland, D. B. 2001. *Introduction to Indexing and abstracting*. 3rd Ed. Englewood, Colo: Libraries Unlimited.

Ghosh, S.B. and Biswas, S.C. 1998. *Subject Indexing systems: Concepts, methods and techniques*. Rev. ed. Calcutta: IASLIC.

Pandey, S.K. Ed.2000 .*Library Information retrieval*. New Delhi: Anmol.

UNIT 14 SEQUENTIAL FILE ACCESS AND STRUCTURE OF INDEX

NOTES

Structure

- 14.0 Introduction
- 14.1 Objectives
- 14.2 Structure of Sequential File Access
- 14.3 Inverted File and its structure
- 14.4 Matching Criteria for Index Files
- 14.5 Answers to Check Your Progress Questions
- 14.6 Summary
- 14.7 Key Words
- 14.8 Self Assessment Questions and Exercises
- 14.9 Further Readings

14.0 INTRODUCTION

File access and the manner in which the files are maintained form a significant subject of library science. It is important to learn about the manner in which the files are stored, how they are manipulated so that the indexing process is undertaken smoothly and the information retrieval is done in an efficient manner. In this unit, the concept of sequential file access, file organization, components, the indexing and matching system and related concepts are discussed in detail.

14.1 OBJECTIVES

After going through this unit, you will be able to:

- Discuss the concept of Sequential file access
- Explain inverted file,
- Describe the structure of an index file and the matching criteria,

14.2 STRUCTURE OF SEQUENTIAL FILE ACCESS

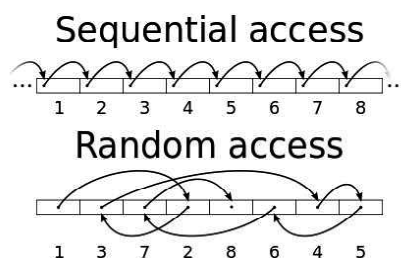


Fig. 14.1 Structure of File Access

NOTES

We have already learnt a little bit about sequential access in the previous unit, let's gain more clarity about the concept by beginning our discussion with comparison between sequential access and random access.

If the elements of a computer storage are accessed in a preset sequence or order, it is called a sequential access. In many situations like while accessing a tape, it might be the only choice of access of data, while other times it is deliberately used in processes where data elements are to be processed in a specific sequence only.

Even the sequencing as concept might seem like a uniform concept of accessing data, it is defined in many dimensions and therefore, its meaning changes with each description and so does the results. For instance, in the spatial dimension, several factors like request size, strided distance, backward accesses, re-accesses have a bearing on the sequentiality. While interval time, multi-stream and related factors affect the threshold impact.

If in a data structure, the values are met only in a specific sequence, it is referred to having a sequential order. Examples of such data structure includes linked lists. The main factor which must be kept in mind while indexing into a sequential order list is that it necessitates the use of $O(n)$ time, where n is the index. This critical factor makes binary search, quicksort and such algorithms not efficient compared to their naive alternatives; these can only work when random access is provided. Other algorithms like mergesort do not face such difficulty.

The file system, the computing world is responsible for the manner in which data is both stored and retrieved. Its importance can be understood by the fact that without it, it will not be possible to understand where one section of data begins and where it ends. The file system assists the users in the isolating the data by differentiating the data into pieces and naming them. Since all paper paper-based information systems are name a group of data: 'file, this system follows the same protocol. The file system then represents that rules related to the logic and structure which is employed to manage the groups of information and their names.

The differences between the varied kinds of file system is based on characteristics like properties of speed, size, structure and logic, flexibility, security, etc. Certain file systems are designed to be used for specific applications. For example, the ISO 9660 file system is designed specifically for dealing with optical discs.

There are varied kinds of storage systems that can be used for file systems. Hard disk drive is one of the most common form of storage devices used today. Flash memory, magnetic tapes, and optical discs are some of the other storage device media. In some cases, such as with tmpfs, the computer's main memory (random-access memory, RAM) is used to create a temporary file system for short-term use.

File systems have the option to utilize either the local or the external storage devices in the form of a network protocol (for example, NFS, SMB, or 9P clients). Virtual file systems are also popular today computed on request (such as procfs and sysfs) or are merely a mapping into a different file system used as a backing

store. The file system manages access to both the content of files and the metadata about those files. It is responsible for arranging storage space; reliability, efficiency, and tuning with regard to the physical storage medium are important design considerations.

The origin of the term

File system in the pre-computer days was used as a term to refer to the method of storing and retrieving paper documents. But since the 1964 the word came to be used for computerized filing as a general use.

Structure of a Sequential File

Layers are an essential part of the file system. Generally, it is seen that a file system contains two or three layers. It is not always necessary that the layers will be combined or solely separated.

The file system arranged logically forms the foundation of the first layer or the application program interface. Different operations — OPEN, CLOSE, READ, etc., are passed through the requested operation to the layer below it for processing. The logical file system ‘manage[s] open file table entries and per-process file descriptors.’ This layer provides ‘file access, directory operations, [and] security and protection.’

The virtual file system is the second optional layer. The file system implementation allows the concurrent use of varied physical file system which gets its support from the virtual file system.

The third layer is known as the *physical file system*. This layer deals with the physical operation of the storage device (e.g. disk). The reading writing and placement of the physical blocks along with other activities like buffering and memory management at specific system location on the storage medium. The storage device activities work through the interaction between the physical file system interacts and the device drivers or with the channel.

Aspects of file systems

Space management

Note: this only applies to file systems used in storage devices.

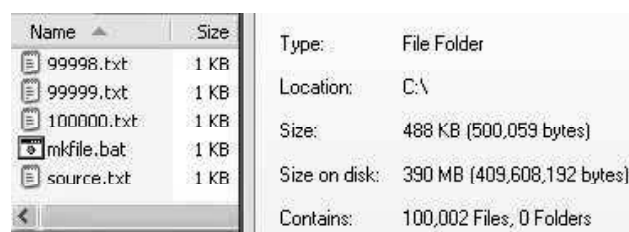


Fig. 14.2 File Information Showing Space Management

An example of slack space, demonstrated with 4,096-byte NTFS clusters: 100,000 files, each five bytes per file, which equal to 500,000 bytes of actual data but require 409,600,000 bytes of disk space to store.

NOTES

NOTES

To allocate granular space to the file system multiple physical units on the device is generally brought into use. Different activities like keeping a tab of the locations at which different files are saved, as well as organizing the files and directories constitute as the functions of the file system.

Slack space refers to the unused space when a file is not an exact multiple of the allocation unit. For a 512-byte allocation, the average unused space is 256 bytes. For 64 KB clusters, the average unused space is 32 KB. The size of the allocation unit is chosen when the file system is created. The reduction of the amount of unusable space can be brought it by selection of the allocation size based on the general size of the files expected to be in the file system.

Reasonable storage is provided by the default allocation. Ofcourse, the size must be carefully selected otherwise the problem of overhead spaces might also arise.

File system fragmentation happens when unused space or single files are not contiguous. The general path of the use of a file system is that files are created, modified and deleted. Space is allocated when a file is created. Allocating initial space is permitted or demanded by certain file systems. This includes specifying not only the initial space allocation but also the subsequent incremental allocations as the file develops. The allotted space become free once the files are deleted leading to the creation of alternating used and unused areas of various sizes which is called free space fragmentation. When a contiguous space is not allotted initially to a file the space must be assigned in fragments. On exceeding the allotted space, another allocation must be assigned elsewhere and the file becomes fragmented.

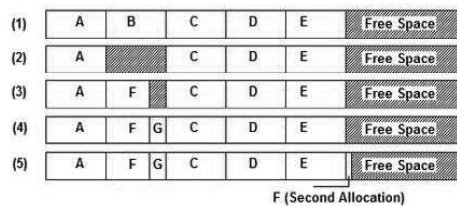


Fig. 14.3 Fragmentation of File System

Filenames

The identification of a storage location in the file system is done through a file name. There are the certain restrictions on the length of filenames. Some may also be case sensitive (i.e., the names MYFILE, MyFile, and myfile refer to three separate files) while others not (i.e., the names MYFILE and myfile refer to the same file).

Most modern file systems allow filenames to contain a wide range of characters from the Unicode character set. The use of special characters however are restricted. disallowing them within filenames; those characters might be used to indicate a device, device type, directory prefix, file path separator, or file type.

Directories

Storage of file system information is done in the form of separate collections known as directories or commonly as folders. These allow for grouping of files. It is done

through the combination of corresponding file name with an index in a table of contents or an inode in a Unix-like file system. These directory elements may be in the flat form or stored as hierarchy. The first file system to support arbitrary hierarchies of directories was used in the Multics operating system. Unix-like systems support arbitrary directory hierarchies, as do, for example, Apple's Hierarchical File System, and its successor HFS+ in classic Mac OS, the FAT file system in MS-DOS 2.0 and later versions of MS-DOS and in Microsoft Windows, the NTFS file system in the Windows NT family of operating systems, and the ODS-2 (On-Disk Structure-2) and higher levels of the Files-11 file system in OpenVMS.

Metadata

Additional bookkeeping information is typically associated with each file within a file system. The length of the data contained in a file may be stored as the number of blocks allocated for the file or as a byte count. The file's timestamp stores the information of the time that the file was last modified. The information related to the file creation time, the time it was last accessed, the time the file's metadata was changed, or the time the file was last backed up may also be stored. File's device type (e.g. block, character, socket, subdirectory, etc.), its owner user ID and group ID, its access permissions and other file attributes (e.g. whether the file is read-only, executable, etc.) are also other types of metadata information.

A separate file containing all the metadata associated with the file—including the file name, the length of the contents of a file, and the location of the file—are saved for each file.

The directory table is also created with names of all files in the directory in one place. This is at most times stored like any other file. There are also chances that only certain information is able to be stored in the directory table while others use inode to store the remaining information separately.

Most file systems also store metadata not associated with any one particular file. Such metadata includes information about unused regions—free space bitmap, block availability map—and information about bad sectors. Often such information about an allocation group is stored inside the allocation group itself.

Extra attributes such as NTFS, XFS, ext2, ext3, some versions of UFS, and HFS+, using extended file attributes. Are also saved. User defined attributes including information about the author of the document, the character encoding of a document or the size of an image are also provided by some file systems.

Different data collections under one file name is also allowed by certain file systems referred to as *streams* or *forks*. Examples include the likes of forked file systems on the Macintosh, and Microsoft supports streams in NTFS. Some file systems maintain multiple past revisions of a file under a single file name. It is then crucial to note that the most recent version of the file is retrieved through the filename while prior saved version can also be accessed using a special naming convention such as "filename;4" or "filename(-4)" to access the version four saves ago.

NOTES

File system as an abstract user interface

There are also chances that the storage device is used by the file system not to store but simply organize and represent access to any data, whether it is stored or dynamically generated (e.g. procfs).

NOTES

Utilities

Initialize, alter parameters of and remove an instance of the file system are the utilities provided by the file systems. Some include the ability to extend or truncate the space allocated to the file system.

Directory utilities may also include capabilities to create additional links to a directory (hard links in Unix), to rename parent links and to create bidirectional links to files. Creation, deletion and renaming of the files can be done through *directory entries*, which are also known as *dentries* (singular: *dentry*), and to alter metadata associated with a directory.

File utilities create, list, copy, move and delete files, and alter metadata. They may be able to truncate data, truncate or extend space allocation, append to, move, and modify files in-place. Depending on the underlying structure of the file system, they may provide a mechanism to prepend to, or truncate from, the beginning of a file, insert entries into the middle of a file or delete entries from a file.

Utilities to free space for deleted files, if the file system provides an undelete function, also belong to this category.

Operations such as reorganization of free space, secure erasing of free space, and rebuilding of hierarchical structures by providing utilities to perform these functions at times of minimal activity are also supported by the directories. An example is the file system defragmentation utilities.

Supervisory activities which may involve bypassing ownership or direct access to the underlying device constitute as some of the file system utilities. Also included are activities like high-performance backup and recovery, data replication and reorganization of various data structures and allocation tables within the file system.

Restricting and permitting access

The control of access to data of the files system can be done through different mechanisms. Generally, the intent is to prevent reading or modifying files by a user or group of users. Of the other reason is to make sure data is modified in a controlled way so access is controlled only till a specific program. Passwords stored in the metadata of the file or elsewhere and file permissions in the form of permission bits, access control lists, or capabilities are some examples of the same. The activities of the intruders cannot be stopped through these activities though.

Ways of encrypting file data are sometimes part in the file system. This is highly effective since there is no requirement for file system utilities to separately gain knowledge about the encryption seed to effectively manage the data. Encryption demands a reliance on the fact that an attacker can copy the data and

use brute force to decrypt the data. Losing the seed means losing the data.

Maintaining integrity

A significant duty of a file system is to make sure that, despite the actions by programs accessing the data, there is consistency of the structure. This comprises of actions taken if a program altering data stops suddenly or neglects to inform the file system that it has finished its tasks. This may comprise of updating the metadata, the directory entry and handling any data that was buffered but not yet updated on the physical storage media.

Additional failures which the file system must tackle include media failures or loss of connection to remote systems.

Special routines in the file system is necessary in situations of an operating system failure or "soft" power failure.

The file system must also be capable of repairing damaged structures which might occur due to an operating system failure for which the OS was unable to notify the file system, power failure or reset.

The file system must also record events to allow analysis of systemic issues as well as problems with specific files or directories.

User data

Management of user data is the most important function of the file system which includes functions like storing, retrieving and updating data.

Some file systems accept data for storage as a stream of bytes which are collected and stored in a manner efficient for the media. When a program retrieves the data, it specifies the size of a memory buffer and the file system transfers data from the media to the buffer. It is also allowed that a runtime permit the user program to define a *record* based on a library call specifying a length. When the user program reads the data, the library retrieves data via the file system and returns a *record*.

Some file systems permit the specification of a fixed record length which is used for all writes and reads. This facilitates locating the n^{th} record as well as updating records.

Key is the identification marker for each record which makes for a more sophisticated file system. Without regard to their location, the user program can read, write and update records. This necessitates complicated management of blocks of media generally dividing key blocks and data blocks. Pyramid structure assists with the development of very efficient algorithms for locating records.

NOTES

Check Your Progress

1. What is free space fragmentation?
2. List some of the systems which support arbitrary directory hierarchies.

14.3 INVERTED FILE AND ITS STRUCTURE

NOTES

Also referred to as postings file or inverted file), inverted file index is an index data structure storing a mapping from content, such as words or numbers, to its locations in a database file, or in a document or a set of documents (named in contrast to a forward index, which maps from documents to content). Its purpose is to allow fast full text searches, at a cost of increased processing when a document is added to the database. The inverted file may be the database file itself, rather than its index. It is the most popular data structure used in document retrieval systems, used on a large scale for example in search engines. Additionally, several significant general-purpose mainframe-based database management systems have used inverted list architectures, including ADABAS, DATACOM/DB, and Model 204.

There are two main types of inverted indexes: A word-level inverted index (or full inverted index or inverted list) additionally contains the positions of each word within a document and a record-level inverted index (or inverted file index or just inverted file) contains a list of references to documents for each word. The latter form offers more functionality (like phrase searches), but needs more processing power and space to be created.

Uses of an Inverted File Index

The inverted index data structure is a central component of a typical search engine indexing algorithm. One of the missions of a search engine implementation is to optimize the speed of the query: find the documents where word X occurs. After the development of a forward index it is next inverted to develop an inverted index. Sequential iteration through each document is required for querying the forward index, to each word to verify a matching document. The technical realism is not always expected of the time, memory, and processing resources to perform such a query are not always technically realistic. Instead of listing the words per document in the forward index, the inverted index data structure is developed which lists the documents per word.

The creation of the inverted index, can resolve the query by jumping to the word ID (via random access) in the inverted index.

Concordances to important books were manually assembled in pre computer times. These were effectively inverted indexes with a small amount of accompanying commentary that required a tremendous amount of effort to produce.

Structure of an Index File

A file is a sequence of records stored in binary format. A disk drive is formatted into several blocks that can store records. File records are mapped onto those disk blocks.

When a file is created using Heap File Organization, the Operating System allocates memory area to that file without any further accounting details. File records

can be placed anywhere in that memory area. It is the responsibility of the software to manage the records. Heap File does not support any ordering, sequencing, or indexing on its own.

Sequential File Organization

Practically, it is not possible to store all the records sequentially in physical form. For unique identification, every file record contains a data field (attribute). Some sequential order based on the unique key field or search key is followed in sequential file organization.

Hash File Organization

Hash function computation is used in some file records. The output of the hash function determines the location of disk block where the records are to be placed.

Clustered File Organization

Clustered file organization is not considered ideal for large databases. In this mechanism, related records from one or more relations are kept in the same disk block, that is, the ordering of records is not based on primary key or search key.

File Operations

Operations on database files can be divided into two categories:

- **Update Operations**
- **Retrieval Operations**

Update operations alter the data values through insertion, deletion, or update. Retrieval operations, on the other hand, do not modify the data but retrieve them after optional conditional filtering. In both types, a significant role is played by selection. The following are the operations, which can be done on files:

- **Open:** Two modes, **read mode** or **write mode** can be used to open a file. In read mode, the operating system does not give permission to anyone to change data. In other words, data is read only. Files opened in read mode can be shared. Write mode permits data modification. Files opened in write mode cannot be shared.
- **Locate:** A file pointer tells the current position where the data is to be read or written. It can be adjusted. It can be moved forward or backward using find (seek) operation.
- **Read:** It is by default, that when files are opened in read mode, the file pointer is at the beginning of the file. This location of the file pointer can be specified by the operator as well.
- **Write:** Write mode enables the users to edit the contents. It can be in the form of deletion, insertion, or modification. The file pointer location can again be directed.
- **Close:** This is the most significant operation from the operating system's approach. When a request to close a file is generated, the operating system

NOTES

- o saves the data (if altered) to the secondary storage media
- o releases all the buffers and file handlers associated with the file
- o removes all the locks (if in shared mode),

NOTES

The organization of data within a file plays a crucial role here. The method to locate the file pointer to a needed record inside a file is dependent on whether the records are sequentially or clustered arranged.

A database index is a data structure that makes efficient the speed of data retrieval operations on a database table by discounting the additional writes and storage space to sustain the index data structure. Indexes are used to quickly locate data without the need to recognize every row in a database table every time a database table is accessed. Indexes can be made using one or more columns of a database table, providing the basis for both rapid random lookups and efficient access of ordered records.

An index is a copy of selected columns of data from a table that can be searched very efficiently that also comprises a low-level disk block address or direct link to the complete row of data it was copied from. Some databases extend the power of indexing by letting developers create indexes on functions or expressions. For instance, an index could be made on `upper(last_name)`, which would only store the upper-case versions of the `last_name` field in the index. Other option is partial indices, where index entries are made only for those records that fulfill some conditional expression. A further aspect of flexibility is to allow indexing on user-defined functions, as well as expressions formed from an assortment of built-in functions.

The data is available in random order, but the **logical ordering** is specified by the index. The data rows may be spread throughout the table despite the effect of the value of the indexed column or expression. The non-clustered index tree comprises of the index keys in sorted order, with the leaf level of the index containing the pointer to the record (page and the row number in the data page in page-organized engines; row offset in file-organized engines).

In a non-clustered index,

- The indexed columns are typically non-primary key columns used in JOIN, WHERE, and ORDER BY clauses
- The physical order of the rows is not the same as the index order.

There can be more than one non-clustered index on a database table.

Clustered

Clustering modifies the data block into a specific distinct order to match the index, resulting in the row data being stored in order. Therefore, only one clustered index can be made on a given database table. Clustered indices can greatly increase the overall speed of retrieval, but usually only where the data is accessed sequentially in the same or reverse order of the clustered index, or when a range of items is selected.

Since the physical records are in this sort order on disk, the next row item in the sequence is immediately before or after the last one, and so fewer data block reads are required. The basic foundational feature of a clustered index is the ordering of the physical data rows in line with the index blocks that correspond to them. Some databases divide the data and index blocks into separate files, others put two different data blocks within the same physical file(s).

Cluster

When multiple databases and multiple tables are joined, it is called **cluster** (not to be confused with clustered index described above). The records for the tables sharing the value of a cluster key shall be stored together in the same or nearby data blocks. This may make better the joins of these tables on the cluster key, since the matching records are stored together and less I/O is required to locate them. The cluster configuration specifies the data layout in the tables that are parts of the cluster. A cluster can be keyed with a B-Tree index or a hash table. The value of the cluster key defines the data block where the table record is stored.

Column order

The sequence that the index definition refers to the columns in is crucial. It is possible to retrieve a set of row identifiers using only the first indexed column. However, it is not possible or efficient (on most databases) to retrieve the set of row identifiers using only the second or greater indexed column.

For instance, imagine a phone book that is organized by city first, then by last name, and then by first name. If you are given the city, you can easily extract the list of all phone numbers for that city. However, in this phone book it would be very difficult to locate all the phone numbers for a given last name. You would have to look inside every city's section for the entries with that last name. Some databases are capable of this, others just won't utilize the index.

Applications and limitations

Indexes are useful for many applications but come with some limitations. Consider the following SQL statement: `SELECT first_name FROM people WHERE last_name = 'Smith'`; To process this statement without an index the database software must look at the `last_name` column on every row in the table (this is known as a full table scan). With an index the database simply follows the B-tree data structure until the Smith entry has been found; this is much less computationally expensive than a full table scan.

Consider this SQL statement: `SELECT email_address FROM customers WHERE email_address LIKE '%@wikipedia.org'`; This query would yield an email address for every customer whose email address ends with "`@wikipedia.org`", but even if the `email_address` column has been indexed the database must perform a full index scan. This is because the index is built with the assumption that words go from left to right. With a wildcard at the beginning of the search-term, the database software is unable to use the underlying B-tree data structure (in other words, the WHERE-clause is *not sargable*). This problem can

NOTES

NOTES

be solved through the addition of another index created on `reverse(email_address)` and a SQL query like this: `SELECT email_address FROM customers WHERE reverse(email_address) LIKE reverse('%@wikipedia.org')`; This puts the wildcard at the right-most part of the query (now `gro.aidepikiw@%`), which the index on `reverse(email_address)` can satisfy.

When the wildcard characters are used on both sides of the search word as `%wikipedia.org%`, the index available on this field is not used. Rather only a sequential search is performed, which takes $O(N)$ time.

14.4 MATCHING CRITERIA FOR INDEX FILES

In this section, we will learn about the basics of matching and lookup functions.

The matching of the index of the information to the search query when done efficiently ensures the successful execution of the retrieval function.

Lookup Functions

There are many ways to do a simple lookup in Excel, using functions such as `VLOOKUP` or `HLOOKUP`. In this example, we need to do a complex lookup:

- there are multiple criteria, instead of just one
- we need to find a product code, which is to the left of the criteria

`VLOOKUP` won't work here, so we'll use the `INDEX` and `MATCH` functions together, to get the results that we need.

	A	B	C	D
1	Code	Item	Size	Price
2	SW001	Sweater	Small	10
3	JK001	Jacket	Small	30
4	PN001	Pants	Small	25
5	SW002	Sweater	Med	12
6	JK002	Jacket	Med	35
7	PN002	Pants	Med	30
8	SW003	Sweater	Large	14
9	JK003	Jacket	Large	40
10	PN003	Pants	Large	35
11				
12	Item	Size	Price	Code
13	Jacket	Med	35	JK002
14				

Fig. 14.4 Using Index and Match Functions

INDEX and MATCH

To do this complex lookup with multiple criteria, we'll use the `INDEX` and `MATCH` functions.

- The `INDEX` function can return a value from a specific place in a list
- The `MATCH` function can find the location of an item in a list.

When INDEX and MATCH are used together, they create a flexible and powerful lookup formula.

Simple INDEX and MATCH

Before using INDEX and MATCH with multiple criteria, let's see how they work together in a simpler formula.

In this lookup formula, we need to find "Sweater" in a column B of a price list, and get its price from column C.

- The item name that we need a price for is entered in cell A7 – Sweater.
- This INDEX and MATCH formula is entered in cell C7, to get the price for that item:

=INDEX(\$C\$2:\$C\$4,MATCH(A7,\$B\$2:\$B\$4,0))

	A	B	C	D	E
1	Code	Item	Price		
2	SW001	Sweater	10		
3	JK001	Jacket	30		
4	PN001	Pants	25		
5					
6	Item	Price	Code		
7	Sweater	10			
8					

Fig. 14.5 Simple Index and Match Function

How INDEX MATCH Formula Works

Here's how that simple INDEX / MATCH formula finds the sweater price:

- the MATCH function can find "Sweater" in the range B2:B4. The result is 1, because "Sweater" is in the first row of that range.
- the INDEX function can tell you that in the range C2:C4, the first row contains the value 10.

So, by combining INDEX and MATCH, you can find the row with "Sweater" and return the price from that row.

Match for Multiple Criteria

In the previous example, the match was based on one criterion -- the Item name. For the next lookup, there are 2 criteria -- Item name and product Code.

In the screen shot below, each item is listed 3 times in the pricing lookup table. To get the right price, you'll need to specify both the item name and the size. We want to find the price for a large jacket.

NOTES

NOTES

	A	B	C	D	E
1		Code	Item	Size	Price
2	1	SW001	Sweater	Small	10
3	2	JK001	Jacket	Small	30
4	3	PN001	Pants	Small	25
5	4	SW002	Sweater	Med	12
6	5	JK002	Jacket	Med	35
7	6	PN002	Pants	Med	30
8	7	SW003	Sweater	Large	14
9	8	JK003	Jacket	Large	40
10	9	PN003	Pants	Large	35
11					
12			Item	Size	Price
13			Jacket	Large	40
14					

Fig. 14.6 Match for Multiple Criteria

MATCH True or False

In the lookup formula, we need the MATCH function to check both the Item and Size columns.

To show how that will work, I'll add temporary columns on the worksheet, to check the item and size columns -- is the item a Jacket, and is the Size a Large?

Enter this formula in F2, and copy down to F10: **=C2=\$C\$13**

- If the Item in column C is a Jacket, the result in column E is TRUE. If not, the result is FALSE

Enter this formula in G2, and copy down to G10: **=D2=\$D\$13**

- If the Size in column D is Large, the result in column F is TRUE. If not, the result is FALSE

		F2						
		=C2=\$C\$13						
	A	B	C	D	E	F	G	
1		Code	Item	Size	Price			
2	1	SW001	Sweater	Small	10	FALSE	FALSE	
3	2	JK001	Jacket	Small	30	TRUE	FALSE	
4	3	PN001	Pants	Small	25	FALSE	FALSE	
5	4	SW002	Sweater	Med	12	FALSE	FALSE	
6	5	JK002	Jacket	Med	35	TRUE	FALSE	
7	6	PN002	Pants	Med	30	FALSE	FALSE	
8	7	SW003	Sweater	Large	14	FALSE	TRUE	
9	8	JK003	Jacket	Large	40	TRUE	TRUE	
10	9	PN003	Pants	Large	35	FALSE	TRUE	
11								
12			Item	Size	Price			
13			Jacket	Large	40			
14								

Fig. 14.7 Match for True or False

MATCH Both True

We need the price from the row where both results are TRUE. We'll use a formula to calculate that for us:

Enter this formula in H2, and copy down to H10: **=F2*G2**

In Excel, TRUE is equal to 1, and FALSE is equal to zero. When you multiply the values,

- If **either** value is FALSE (0), the result is zero
- If **both** values are TRUE (1), the result is 1

Only the 8th row in our list of items has a 1, because both values are TRUE in that row.

NOTES

	A	B	C	D	E	F	G	H
1		Code	Item	Size	Price			
2	1	SW001	Sweater	Small	10	FALSE	FALSE	0
3	2	JK001	Jacket	Small	30	TRUE	FALSE	0
4	3	PN001	Pants	Small	25	FALSE	FALSE	0
5	4	SW002	Sweater	Med	12	FALSE	FALSE	0
6	5	JK002	Jacket	Med	35	TRUE	FALSE	0
7	6	PN002	Pants	Med	30	FALSE	FALSE	0
8	7	SW003	Sweater	Large	14	FALSE	TRUE	0
9	8	JK003	Jacket	Large	40	TRUE	TRUE	1
10	9	PN003	Pants	Large	35	FALSE	TRUE	0
11								
12			Item	Size	Price			
13			Jacket	Large	40			
14								

Fig. 14.8 Match for Both True

Lookup With Multiple Criteria

We could use a MATCH formula to find the position of a 1 in column G, in the screen shot above. The 8th row of data (worksheet row 9), has the 1, and that row will give us the correct price for a large jacket.

But, instead of adding extra columns to the worksheet, we will use an array-entered INDEX and MATCH formula to do all the work.

Here is the **array-entered*** formula that we'll use in cell E13, to get the correct price:

o =INDEX(E2:E10,
 MATCH(1,
 (C13=C2:C10)*(D13=D2:D10),0))

NOTES

*Press **Ctrl + Shift + Enter**, instead of just pressing the Enter key. That will automatically add curly brackets around the formula.

Check Your Progress

3. List some of the general purpose main-frame based database management system which use the inverted list architecture.
4. What is the output of the hash system?
5. What is a database index data?

14.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Space is allocated when a file is created. Allocating initial space is permitted or demanded by certain file systems. This includes specifying not only the initial space allocation but also the subsequent incremental allocations as the file develops. The allotted space become free once the files are deleted leading to the creation of alternating used and unused areas of various sizes which is called free space fragmentation.
2. Unix-like systems support arbitrary directory hierarchies, as do, for example, Apple's Hierarchical File System, and its successor HFS+ in classic Mac OS, the FAT file system in MS-DOS 2.0 and later versions of MS-DOS and in Microsoft Windows, the NTFS file system in the Windows NT family of operating systems, and the ODS-2 (On-Disk Structure-2) and higher levels of the Files-11 file system in OpenVMS.
3. Several significant general-purpose mainframe-based database management systems which have used inverted list architectures, include ADABAS, DATACOM/DB, and Model 204.
4. The output of the hash function determines the location of disk block where the records are to be placed.
5. A database index is a data structure that makes efficient the speed of data retrieval operations on a database table discounting the additional writes and storage space to sustain the index data structure.

14.6 SUMMARY

- If the elements of a computer storage are accessed in a preset sequence or order, it is called a sequential access. In many situations like while accessing

a tape, it might be the only choice of access of data, while other times it is deliberately used in processes where data elements are to be processed in a specific sequence only.

- Even the sequencing as concept might seem like a uniform concept of accessing data, it is defined in many dimensions and therefore, its meaning changes with each description and so does the results.
- To allocate granular space to the file system multiple physical units on the device is generally brought into use. Different activities like keeping a tab of the locations at which different files are saved, as well as organizing the files and directories constitute as the functions of the file system.
- File system fragmentation happens when unused space or single files are not contiguous. The general path of the use of a file system is that files are created, modified and deleted.
- Also referred to as postings file or inverted file), inverted file index is an index data structure storing a mapping from content, such as words or numbers, to its locations in a database file, or in a document or a set of documents (named in contrast to a forward index, which maps from documents to content).
- Practically, it is not possible to store all the records sequentially in physical form. For unique identification, every file record contains a data field (attribute). Some sequential order based on the unique key field or search key is followed in sequential file organization.
- Operations on database files can be divided into two categories "
 - o Update Operations
 - o Retrieval Operations
- The sequence that the index definition refers to the columns in is crucial. It is possible to retrieve a set of row identifiers using only the first indexed column. However, it is not possible or efficient (on most databases) to retrieve the set of row identifiers using only the second or greater indexed column.
- There are many ways to do a simple lookup in Excel, using functions such as VLOOKUP or HLOOKUP.
- The INDEX function can return a value from a specific place in a list
- The MATCH function can find the location of an item in a list.
- When INDEX and MATCH are used together, they create a flexible and powerful lookup formula.

NOTES

14.7 KEY WORDS

- **Sequential access:** It refers to the type of file access where the elements of a computer storage are accessed in a preset sequence or order.

NOTES

- **Inverted file index:** It refers to an index data structure storing a mapping from content, such as words or numbers, to its locations in a database file, or in a document or a set of documents (named in contrast to a forward index, which maps from documents to content).
- **Cluster:** It refers to the joining of multiple databases and multiple tables.

14.8 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Write a short note on the structure of a file systems.
2. What is metadata?
3. Briefly discuss the utilities of file system.
4. Write a short note on the applications and limitations of index files.

Long-Answer Questions

1. Explain the aspects of file systems.
2. Describe the different file operations.
3. Explain the structure of an index file.
4. Discuss the uses of an inverted index file.
5. Explain the concept of matching criteria for index files.

14.9 FURTHER READINGS

- Alberico, R. and Micco M. 1990. *Expert systems for reference and Information retrieval*. West Port: Meckler.
- Atchison, J. & Gilchrist, A. 1972. *Thesaurus construction: a practical manual*. Aslib: London.
- Chowdhry, G.G. 2003. *Introduction to modern Information retrieval*. 2nd Ed. Facet Publishing: London.
- Cleaveland, D. B. 2001. *Introduction to Indexing and abstracting*. 3rd Ed. Libraries Unlimited: Englewood, Colo.
- Ghosh, S.B. and Biswas, S.C. 1998. *Subject Indexing systems: Concepts, methods and techniques*. Rev. ed. IASLIC: Calcutta.
8. Pandey, S.K. Ed. 2000. *Library Information retrieval*. Anmol: New Delhi.

Master of Library & Information Science 323 11

INFORMATION PROCESSING AND RETRIEVAL

I - Semester



ALAGAPPA UNIVERSITY

[Accredited with 'A+' Grade by NAAC (CGPA:3.64) in the Third Cycle
and Graded as Category-I University by MHRD-UGC]

KARAIKUDI – 630 003

DIRECTORATE OF DISTANCE EDUCATION



ISBN 978-93-5338-209-4



9 789353 138209 4