Department of Library and Information Science

Bharathidasan University
Tiruchirappalli-620024

Name of the Programme: M.Lib.I.Sc

Course –– 2.2 Information Processing and Retrieval Systems

Course Code:P21MLS7

Unit-II: Indexing systems – General Theory of Indexing languages. Indexing: Pre coordination and Post coordination, Keyword Indexing, Evaluation of Indexing System, thesaurus and vocabulary control

Course Teacher : DR.C.RANGANATHAN
Professor
e-mail: cranganathan72@gmail.com

What is Index?

- The word **‘Index’** comes from the Latin word ‘**indicaire**’
- Meaning ‘to point out or to guide’

Index is “a systematic guide to the text of any reading matter or to the contents of other collected documentary material, comprising a series of entries, with headings arranged in alphabetical or other chosen order and with references to show where each item indexed is located”.

What is Indexing?

- Indexing is “a process by which the information is organized to enable its easy retrieval and access”.
- Subject Indexing “refers to the process of identifying and assigning labels, descriptors, or subject headings to an item so that its subject contents are known and the index, thus created, can help in retrieving specific items of information.” [UNESCO Handbook of Information Systems and Services,1977].

Definition

- **Online Dictionary for Library and Information Science**

defines indexing “as an artificial language consisting of subject headings or content descriptors selected to facilitate information retrieval by serving as access points in a catalogue or index, including any lead-in vocabulary and rules governing the form of entry, syntax, etc.

Types of Indexing

- Indexing is of two types, viz, Derived Indexing and Assigned Indexing.
- Derived indexing uses the same language as that used by the author. It is also known as Natural Language Indexing (NLI).
- Words/Terms used by the author in the text are used to provide access to users. Such a system of indexing suffers from a drawback which is, that approach and access through alternative terms are not possible.
- The users looking for information through such alternative terms are not able to find the information though the information may be available in the file.

Cont…,

- Assigned indexing is based on conceptual analysis of terms and words.
- The analysis is done to find out the concept and deciding the terms/words representing them and also the related concepts. It helps the user to reach to the required as well as related information.
- This assumes importance in view of the fact that the user may not be exactly sure of his/her information requirements. Even if he/she is sure of his/her requirements, he may not be able to express it exactly.
- Thus, a map of related concepts presented before him would help him to better understand, represent and reach to his required

Indexing Language

- To understand the concept of indexing language it will be proper if we first dwell on the concept of language.
- Language is the vehicle for communication; it is a carrier for thought and plays an important role in communication. It enables the thought to flow from the source to the sink or from the origin to the destination.
- The term language to refer to only the language that is used by humans. The characteristics of a language are vocabulary and rules for their arrangement (syntax).

Cont…,

- The languages may be Artificial and Natural.
- Natural languages refer to our languages, which we normally use for communication,

whereas,

- Artificial are those that we have designed for a specific purpose or are used in a specific sense or for limited use only. Shorthand is an example of this category which all of us have heard about.
- Similarly, we have examples of artificial languages in different disciplines e.g., in chemistry we have a language to indicate names of different elements, compounds and also the process of their transformation.
- Similarly, notation of a **classification scheme** is an artificial language.

Need

Indexing language is necessary because:

- different authors in documents may express same concepts in different ways.
- different concepts may be expressed by same terms.
- users may be interested in these concepts. They need to be informed that their documents exist discussing the concepts specifically.
- users might be interested to access documents on related concepts also, thus, related concepts need to be brought together showing their relationships.

Purpose

The purpose of an indexing language, therefore, is to:

- express the concepts discussed in documents;
- show the relationship among different concepts; and
- help in depicting a panoramic view of the related concepts.

Characteristics

- The purpose of an indexing language is to express the concepts of documents in an artificial language so that users are able to get the required information.
- The indexing language does this by depicting the relationships among the different related concepts.
- Thus, an indexing language consists of elements that constitute its vocabulary, rules for admissible expressions (i.e. syntax) and semantics. An indexing language should, therefore, have:

- Semantic structure
- Syntactic structure
- Syndetic structure.

Semantic Structure

- Semantics refers to the aspects of meaning. In the context of an indexing language, two kinds of relationships between concepts – hierarchical and non-hierarchical can be identified.
- The hierarchical relationships may be Genus-Species and Whole-Part Relationships.
- The Non-hierarchical relationships may be Equivalence or Associative relationships.

Hierarchical Relationships

- It is a permanent relationship.
- a) Genus-Species (Example: Telephone is always a kind of Telecommunication)
- b) Whole-Part (Example: Human Body - Respiratory system)
- c) Instance (Example: Television - Phillips TV).

Non-Hierarchical Relationships

- It may be of two kinds – Equivalence and Associate.
- **a) Equivalence**
- Synonym (Example: Defects - Flaws)
- Homonym (Example: Fatigue (of metals), Fatigue (of humans);
- **b) Associate**
- It refers to the relationships in which concepts are semantically related but do not necessarily belong to same hierarchy (e.g. Weaving and cloth).

Syntactic Structure

- As you know the word syntax refers to grammar. In the context of indexing language syntax governs the sequence of occurrence of terms in a subject heading viz., for the title export of iron, it may be Iron, Export or Export, Iron.

Syndetic Structure

- To show the relationships described at semantic structure, syndetic structure should be built in indexing language (viz., see, see also; use, use for).
- Syndetic structure in the indexing language aims to link related concepts otherwise scattered and helps to collocate related concepts.
- It guides the indexer and the searcher to formulate index entries and to search for his/her information.

Types of Indexing Languages

- Likewise the indexing languages may also be of different kinds, viz.:
- Natural indexing language,
- Free indexing language and
- Controlled indexing language.

Cont…,

- **Natural language indexing** uses the same vocabulary as those used by the author to represent the concepts. It is used in derived indexing.
- **Free indexing language** makes use of all possible terms to index documents irrespective of their use by authors. These would include all possible forms including synonyms, technical versus popular terms, words used in different areas, etc.
- **Controlled language** limits the use of terms based on the system used. It is used for assigned indexing. There are different examples of controlled languages,
- **Example.,** Lists of Subject Headings, Classification schemes, Thesauri, Thesaurofacet and Glossaurus

INDEXING PRINCIPLES AND PROCESS

Purpose of Indexing

- Indexing is regarded as the process of describing and identifying documents in terms of their subject contents. Here, the concepts are extracted from documents by the process of analysis, and then transcribed into the elements of the indexing systems, such as thesauri, classification schemes, etc.
- In indexing decisions, concepts are recorded as data elements organized into easily accessible forms for retrieval. These records can appear in various forms, e.g. back-of-the-book indexes, indexes to catalogues and bibliographies, machine files, etc.
- The process of indexing has a close resemblance with the search process. Indexing procedures can be used, on one hand, for organizing concepts into tools for information retrieval, and also, by analogy, for analysing and organizing enquiries into concepts represented as descriptors or combinations of descriptors, classification symbols, etc.

Subject Indexing : Main purposes

- To prescribe a standard methodology to subject cataloguers and indexers for constructing subject headings.
- To be consistent in the choice and rendering of subject entries, using standard vocabulary and according to given rules and procedures.
- To be helpful to users in accessing any desired document(s) from the catalogue or index through different means of such approach.
- To decide on the optimum number of subject entries, and thus economise the bulk and cost of cataloguing indexing.

Problems in Indexing

- Complexities in the subjects of documents—usually multi-word concept;
- Multidimensional users need for information;
- Choice of terms from several synonyms;
- Choice of word forms (Singular / Plural form);
- Distinguishing homographs;
- Identifying term relationships — Syntactic and Semantic;
- Depth of indexing (exhaustivity);
- Levels of generality and specificity for representation of concepts (specificity);

Cont..,

- Ensuring consistency in indexing between several indexers (inter-indexer consistency), and by the same indexer at different times (intra-indexer consistency);
- Ensuring that indexing is done not merely on the basis of a document's intrinsic subject content but also according to the type of users who may be benefited from it and the types of requests for which the document is likely to be regarded as useful;
- The kind of vocabulary to be used, and syntactical and other rules necessary for representing complex subjects; and
- Problem of how to use the 'index assignment data'.

Indexing Process

- Does the collection contain any categories of material that should not be indexed?
- Does the material require general, popular vocabulary in the index?
- What is the nature of collection?
- What is the characteristic of user population?
- The physical environment in which the system will function; and
- Display or physical appearance of the index.

Cont..,

- The processes of indexing consist of two stages:
- establishing the concepts expressed in a document, i.e. the subject; and
- Translating these concepts into the components of the indexing language.

Establishing the concepts expressed in document

- **Understanding the overall content of the document, the purpose of the author, etc**
- **Identification of concepts**
- **Selection of concepts**

Cont..,

Translating the concepts into the indexing language

- In the next stage in subject indexing is to translate the selected concepts into the language of the indexing system.
- At this stage, an indexing can be looked from two different levels:
  - **Document level**, which is known as **Derivative indexing**; and
  - **Concept level**, which is known as **Assignment indexing**.
- Derivative indexing is the indexing by extraction. Words or phrases actually occurring in a document can be selected or extracted directly from the document (keyword indexing, Automatic indexing, etc.).
- Here, no attempt is made to use the indexing language, but to use only the words or phrases, which are manifested in the document.

Cont..,

- Assignment indexing (also known as ‘concept Indexing’) involves the conceptual analysis of the contents of a document for selecting concepts expressed in it, assigning terms for those concepts from some form of controlled vocabulary according to given rules and procedures for displaying syntactic and semantic relationships .

Example

- Chain Indexing,
- PRECIS,
- POPSI,
- Classification Schemes, etc.).

Axiom

- a) Axiom of definability: Compiling information relevant to a topic can only be accomplished to the degree to which a topic can be defined.
- b) Axiom of order: Any compilation of information relevant to a topic is an order creating process.
- c) Axiom of the sufficient degree of order: The demands made on the degree of order increase as the size of a collection and frequency of searches increase.
- d) Axiom of predictability: It says that the success of any directed search for relevant information hinges on how readily predictable or reconstructible are the modes of expression for concepts and statements in the search file. This axiom is based on the belief that the real purpose of vocabulary control devices is to enhance representational predictability.
- e) Axiom of fidelity: It equates the success of any directed search for relevant information with the fidelity with which concepts and statements are expressed

  in the search file.

Type of Indexing System

- Pre-Coordinate Indexing Systems

- Cutter's Rules for Dictionary Catalogue
- Kaiser's Systematic Indexing
- Chain Indexing
- Relational Indexing
- Coates's Subject Indexing
- PRECIS
- COMPASS
- POPSI

Cont…,

- Post-Coordinate Indexing Systems

➢Uniterm Indexing- Term entry

➢Optical co incidence system- Peek a boo system

➢Edge-Notched Card- Item entry

➢Post-Coordinate Searching Devices

Cont…,

- Automatic Indexing
- Manual Indexing vs. Computerized Indexing
- Methods of Computerized Indexing
- File Organisation
- Indexing Systems using AI Techniques
- User Interface Design

Cont..,

- Non-Conventional Indexing: Citation Indexing
- Web Indexing

➢Operational Aspects of the Web

➢Pre-requisites for Web Indexing

➢Search Engines

➢Semantic Web

Chain Indexing

- Introduced by S.R.Ranganathan in 1948
- It was first mentioned in "Theory of library catalogue".
- 1951 onwards it was adopted BNB successful.
- But since 1971- it has be abandoned.
- CCC are using chain procedure successfully to derive subject headings.
- Unsought link, sought link, mission link,False link

  Connections symbol.

  Digit represent a phase relation.

  Digit represents Intra facet relation.

  Digit represent Intra array.

  Time isolate.
- "Chain procedure is used to derive class index entries in a Classified Catalogue, and specific subject entries, subject analytical, and 'see also' subject entries in a Dictionary Catalogue."
- The term 'chain' refers to a modulated sequence of subclasses or isolates.
-

Relational Indexing& Subject Indexing

- **J. E. L. Farradane** has identified the nine relationships in his **Relational Indexing**. These nine relationships and their respective relational operators are
  - a) Concurrence / 0
  - b) Self-activity/ *
  - c) Association / ;
  - d) Equivalence/ =
  - e) Dimensional / +
  - f) Appurtenance / (
  - g) Distinctness/)
  - h) Reaction / i)
  - Causation / :
- **E. J. Coates** applied his idea of **Subject Indexing** on British Technology Index (now Current Technology Index). He gave us the order of significance: Thing–Part–Material–Action.

PRECIS(Preserved Context Indexing System)

- Derek Austin in 1970.Adopted BNB from 1971 onwards.
- First Version of PRECIS was introduced in BNB in 1971.
- Detail Manual Published in 1974.
- In 1974 as an alternative procedure for deriving subject headings and generating index entries for British National Bibliography (BNB) which since 1952, was following Chain Indexing.
- Second edition of PRECIS-1984
- PRECIS consist of two inter-related sets of working procedures: Syntactical and Semantic.
- Syntactical relationships in PRECIS are handled by means of a set of logical rules and a schema of role operators and codes.
- Semantic, i.e. a priori relationship between indexing terms and their synonyms are regulated by machine-held thesaurus that serves as the sources of see and see also references in the index.

Cont..,

- The PRECIS is based on two principles:
  - a)Principle of Context Dependency
  - b) Principle of One-to One Relationship
- **Two-Line-Three-Part** entry structure is followed in PRECIS.
  - The first line consists of two units—the lead and qualifier—is called the heading, and
  - the other line is called the display.
- There are two kinds of **Role Operators**:
  - Primary Operators (earlier called Mainline Operators) and
  - Secondary Operators (earlier called Interposed Operators).

Cont…,

- An important feature of the revised version of PRECIS is the provision of Codes for bringing expressiveness in the resulting index entries.
- Three types of codes are there – Primary, Secondary, and Typographic codes.
- There are **three kinds of format** in PRECIS:
  - Standard Format,
  - Inverted Format and
  - Predicate Transformation.
  -
- In 1990, it was decided to revise UKMARC and to replace PRECIS by a more simplified system of subject indexing. As a result **Computer Aided Subject System** (COMPASS) was introduced for BNB from 1991 using the same kind of basic principles of PRECIS

POPSI-Postulate Based Permuted Subject Indexing

- In 1964, Ranganathan demonstrated a new line of thinking regarding verbal indexing based on facet analysis [Ranganathan, 1964].
- Since then, continuous research in this new line of thinking was going on, and ultimately the researches of **Dr G Bhattacharyya** on this new line of thinking led to three distinct but interrelated contributions in the field of subject indexing in India, which together constitute what has been designated as the **General Theory of Subject Indexing Language** (GT-SIL).
- A methodology for designing a specific purpose-oriented SIL, known as POstulate-based Permuted Subject Indexing or POPSI -in 1968

- The POPSI was developed through logical interpretation of the postulates for structural analysis of the names-of-subject forming part of the GT-SIL.

Cont..,

- It is based on a "General Theory of Subject Indexing Language"
- It is a "Deep Structure of Subject Indexing Language"
- Five Elementary Categories "DEPAM"
- [D (Discipline), E (Entity), A (Action), P (Property)] and modifier
- POPSI Nine Steps

  Step1- Verbal represented

  Step2- Display of components

  Step3- Short display

  Step4- Approach term

  Step5- Subject index entries derivation

  Step6- Display of subject index entry

  Step7- Cross reference

  Step8- Alphabetical arrangement of entries

  Step9- Guidance

2.POST CORDINATE INDEXING

POST CORDINATE INDEXING

Coordination at output stage

Term entry

Uniterm Indexing – Martimer Taub(1953)

Optical co incidence system/ Peek a boo system - W.E. Batten - c cordonnier

Item entry - Edge notched system, calvin moore

Cont…,

- The Post-coordinate indexing system is of two types:
- **Term Entry System**:
- Here, index entries for a document are made under each of the appropriate subject headings and these entries are filed according to alphabetical order. Under this system, the number of index entries for a document is dependent on the number of component terms associated with the thought content of the document. Here, terms are posted on the item. Searching of two files (Term Profile and Document Profile) is required in this system. For examples, Uniterm, Peek-a-boo, etc.
- **Item Entry System**:
- It takes the opposite approach to term entry system and prepares a single entry for each document (item), using a physical form, which permits access to the entry from all appropriate headings. Here, items are posted on the term. Item entry system involves the searching of one file (i.e. Term Profile) only. For example, Edge-notched card.

Cont..,

- **Martimer Taube** devised the **Uniterm indexing** system in 1953 to organize a collection of documents at the Armed Services Technical Information Agency (ASTIA) of Atomic Energy Commission, Washington.
- Sometimes this is also referred to as Unit Concept Indexing where concepts can also be represented by a single term. This system has two files: Numerical File & Uniterm File.
- For each uniterm, a term card is prepared and it is divided into 10 equal vertical columns (from 0 to 9) in which accession number of the document relevant to the uniterm are recorded according to the system of terminal digit posting.
- **Peek-a-boo** is the trade name of the optical coincidence card. It is also called '**Batten Cards**'.
- One card may punch 500 to 10000 accession numbers depending on the size of the card and size of the ruled squares. Instead of putting accession number (as is done in uniterm system),
- an item is indexed by punching a hole into the appropriate position that serves to represent the document number.

Cont..,

- Indexing on **Edge-Notched card** is based on punched card system. Their value is limited to very small collection.
- Features:
  - One card corresponds to one document.
  - Number of holes on a card varies from 75 to 128.
  - Each hole represents a particular index term.
- **False drop**
- Designation for unwanted retrieved document representations in information retrieval caused by post-coordinative search techniques.
- Example: Search for "female alcoholics". By Boolean search in which "female" is combined with "alcoholism" by logical "and" are retrieved both wanted records (such as "treatment of female alcoholics with psychotherapy") and unwanted records (such as "treatment of alcoholism by female doctors"). This last example represents a "false drop".

Cont..,

- Watters states that: "a false drop is a retrieved item that meets the syntactic requirements of a query but not the semantic requirements of a query...“
- The following devices are used to eliminate **false drops in post-coordinate indexing**:
  - a) Use of bound terms;
  - b) Links;
  - c) Roles; and
  - d) Weighting.

- When the assignment of the content identifier is carried out with the aid of modern computing equipment the operation becomes **automatic indexing**. (**Salton**).

Natural Language Indexing

- The idea of **analysing the subject of a document** through automatic counting of term occurrences was first put forward by H P **Luhn** of IBM in 1957.
- 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.

- **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. American Chemical Society established the value of KWIC after its adoption in 1961 for its publication 'Chemical Titles'.

Key Word Indexing

Two important other versions of keyword index are **KWOC** and **KWAC**.
KWAC also stands for 'key-word-and-context'. KWAC is also called enriched KWIC

Different versions of keyword indexing are:

a) KWIC (Keyword –In-Context) Index;

b) KWOC (Keyword Out-of-Context) Index;

c) KWAC (Keyword Augmented-in-Context) Index;

d) KWWC (Keyword-with-Context) Index;

e) KEYTALPHA (Key Term Alphabetical) Index;

f) WADEX (Word and Author Index);

g) DKWTC (Double KWIC) Index;

h) KLIC (Key-Letter-In-Context) Index;

i) KWIT (Keyword-In-Title) Index; and

j) SWIFT (Selected Words In Full Titles) Index.

KWIC (Keyword-In-Context) Index

- 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.

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,

- Control of damages of rice by insets 118
- Damages Control of damages of rice by insets 118
- Insets Control of damages of rice by insets 118
- Rice 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.

Other Versions

- **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. Keytalpha 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.

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:
- Chemical Titles;
- BASIC (Biological Abstracts Subject In Context);
- Keyword Index of Chemical Abstracts;
- CBAC (Chemical Biological Activities);
- KWIT (Keyword-In-Title) of Laurence Burkeley Laboratory;
- SWIFT (Selected Words in Full Titles); and
- SAPIR (System of Automatic Processing and Indexing of Reports).

Advantages

1) The principal merit of keyword indexing is the speed with which it can be produced;
2) 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;
3) Involves minimum intellectual effort;
4) Vocabulary control need not be used; and
5) Satisfies the current approaches of users.

Disadvantages

1) 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 controlled vocabulary is used, keyword indexing appears to be unsatisfactory for the subjects of Humanities and Social Sciences;
2) 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;
3) Search of a topic may have to be done under several keywords;
4) Search time is high;
5) Searchers very often lead to high recall and low precision; and
6) Fails to meet the exhaustive approach for a large collection.

Cont…,

- Different methods adopted in measuring the word significance in **computerized indexing** are
  - a) Weighting by location;
  - b) Relative frequency weighting;
  - c) Use of noun phrase;
  - d) Use of thesaurus;
  - e) Use of association factor; and
  - f) Maximum-depth indexing.
- Different levels of knowledge used in natural language understanding for the NLP (Natutal Language Processing)-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.

Citation Indexing

- There are three parts in **Science Citation Index**:
  - Citation Index
  - Source Index
  - Permuterm Subject Index
- The first application of Citation Indexing was evident in **Shepard's Citations**, developed by **Frank Shepard** in 1873 as an index to American legal cases.

- Institute for Scientific Information (ISI), Philadelphia in 1963, published the first **SCI** comprising the literature of 1961.

- The online version of the SCI, known as **SCISEARCH**, is published from 1974.
- ISI also brought out the **Social Science Citation Index** (SSCI) and **Arts and Humanities Citation Index** (A&HCI) in 1973 and 1978 respectively. The online version of the SSCI is known as SOCIAL SCISEARCH.

Cont..,

- **Meta Search Engine** are often referred to as multi-threaded search engines or federated search engines.
- **Uniform Resource Identifier** (URI): A URI is an identifier used to identify objects in a space. The most popularly used URI is the URL. URL is a subset of URI, they are not synonymous. A URI can be anything from the name of a person or an email address or URL or any other literal. The URI specifies a generic syntax. It consists of a generic set of schemes that identify any document / resource like URL, URN (Uniform Resource Name), URC (Uniform Resource Characteristic), etc.
- **Resource Description Framework** (RDF) - A generic framework for describing and interchanging metadata on the Internet. RDF metadata expresses the meaning of terms and concepts in the XML that is understandable to computers.

Cont..,

- **Semantics** is a study of meaning. In an indexing language, semantic relationship refers to the hierarchical and non-hierarchical relationships between the subjects and is governed by see and sees also references in an index file. Controlled vocabulary serves as the source for see and see also references.
- The term **Semantic Web**, introduced by **Tim Berners-Lee**, is the extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.
- **Syntax** - The grammatical structure consisting of a set of rules that govern the sequence of occurrence the terms in a subject heading.
- **Term Significance** - It refers to the order of significance developed by **E J Coates** in his subject indexing system. It states that the most significant term in a compound subject heading is the one that is most readily available in the memory of the enquirer and this leads to the order of significance as Thing-Part-Material-Action.

Cont…,

- **Uniform Resource Name** (URN) - A URI (name and address of an object on the Internet) that has some assurance of persistence beyond that normally associated with an Internet domain or host name.
- An **information retrieval system** (IRS) includes database, information base and Structured Query Language (SQL).
- **Structured Query Language** (SQL) is a query language used for accessing and modifying information in a database. The language was first created by IBM in 1975 and was called SEQUEL for "Structured English Query Language". Since then, it has undergone a number of changes, with a lot of influence from Oracle Corporation. Though SQL is now considered to be a standard language, there are still a number of variations of it, such as mSQL and mySQL. Many database products such as MS-Access, SQL Server and Oracle support SQL with proprietary extensions to the standard language.

Evaluation of Indexing systems

- Recall ratio = $\frac{a}{a+b} \times 100$

a = Retrieved Relevant

b = Not retrieved Relevant

- Precision ratio = $\frac{a}{a+c} \times 100$

c= retrieved not relevant

Vocabulary control

Thesaurus

Allen Kent

BT-NT-RT

THESAURUS

Introduction

- The word thesaurus is Greek word, the meaning of which is the collection, store or dictionary of the words. It is used as a tool to control over indexing language.
- The word thesaurus first used in 1957, since then its use has been increasing day by day in the information retrieval techniques and systems.
- Number of thesauri have been developed to use, and are now available in the printed form, from which have gotten a particular method of construction of the format of a thesaurus.

Definition

- A thesaurus may be defined as a compilation of description for use in an information retrieval system arranged in a systematic order and manifesting the various types of relationship existing between the terms.

- According to **UNISIST**, a thesaurus may be defined either in terms of its function or its structure. In terms of function a thesaurus is a terminological control device used in translating from natural language of documents, indexers or users into a more constrained system of language. In terms of structure a thesaurus is a controlled and dynamic vocabulary of semantically and generically related terms which covers a specific domain of knowledge.

Purpose of Thesaurus

1. To provide a map of a given field of knowledge indicating how concepts or ideas are related to one another thus providing complete understanding and composite knowledge of the structure of the field.

2. To provide a standard vocabulary for given subject field which ensures consistency and standardization of language.

3.To provide a system of references between terms which will ensure that only one term from a set of synonyms is selected as search term, and to provide guides to terms which are related to any indexed term in other ways, either by classification structure or otherwise

4. By providing, **see, see under** and **see also** cross references terms are properly linked together so as to avoiding duplicity of terms in search process.

5. To locate new concepts in a scheme of relationships with existing concepts in a way which makes sense to the users of the system?

6. To provide classified hierarchies so that a search can be broadened or narrowed systematically, if the first choice of search terms produces either too few or to many references to the material in store.

Types of Thesaurus

- **Regular Thesaurus**
- **Stems Thesaurus**
- **Structure of Thesaurus**

Structure of Thesaurus

- Structural organization of concepts with the symbols of hierarchical and correlating among the concepts.
- The words directing a specific concept, like synonyms etc,.
- In thesaurus, the relations among the terms are as follows:
  - Hierarchical
  - Semantic
  - Associated
  - Syntexical

Steps in thesaurus construction

- **Step-1: Initiation steps**
- **Step-2: Subject Scope of Determination**
- **Step-3: Collection of Terms**
- **Step-4: Organization of Terms**
- **Step-5: Draft Thesaurus Preparation**
- **Step-6: Testing of Draft Thesaurus**
- **Step-7: Finalization**
- **Step-8: Updating**

VOCABULARY CONTROL

VOCABULARY CONTROL

- Indexing may be thought of as a process of labelling items for future reference. Lancaster [1986] suggests that the process of subject indexing involves two quite distinct intellectual steps:
- The ‘conceptual analysis’ of the documents and 'translation' of the conceptual analysis into a particular vocabulary.
- The second step in any information retrieval environment involves a ‘controlled vocabulary’,
- i.e, a limited set of terms that must be used to represent the subject matter of documents.
- Similarly, the process of preparing the search strategy also involves

VOCABULARY CONTROL

- **Define:**

  Vocabulary control means to control over the search terms of the indexing system by way of preferred term used in the information retrieval process in order to achieve consistency in the subject indexing with the help of rules prescribed for selecting words and adding new words.

Concept:

- The controlled vocabulary exists to facilitate communication in the information retrieval process.

Need of vocabulary Control:

- The indexing language the actual term used for vocabulary control.
- In indexing languages where symbols or codes are used the difference of vocabulary with a national language is quite apparent.
- To control synonymous terms and also to show the relationship between concepts.
- The instruments of control has been the list of subject headings

Objectives of Vocabulary Control

a) To promote the consistent representation of subject matter by indexers and searchers, thereby avoiding the dispersion of related materials. This is achieved through the control (merging) of synonymous and near synonymous expressions and by distinguishing among homographs;

b) To facilitate the conduct of a comprehensive search on some topic by linking together terms whose meanings are related.

Role of vocabulary Control:

- Controlled Vocabulary in contrast to the use of natural languages is essential to ensure efficient retrieval performance of the system.
- A natural language contains a large amount of synonyms, quasi-synonyms, homonyms, eponyms, acronyms, ambiguous terms etc.,
- If the indexes represent the concepts discussed in a document using the terms of a natural language, retrieval will not be very successful.
- The controlled vocabulary also distinguish has among homographs, that is words with identical spelling but different meanings.

Cont..,

- Controlling synonyms, near synonyms and quasi, synonyms and distinguishing among homographs. The controlled vocabulary the avoid dispersion of like subject matter and the collaboration of unlike subject matter.
- Vocabulary control is to link together terms that are semantically related in order to facilitate the conduct of comprehensive searches.
- Syntactic relationship is transient and deal with documents.
- Vocabulary is well constructed it brings together terms that are equivalence hierarchical and associative relationship.

Function of vocabulary Control

- To provide for consistent representation of subject matter, thereby avoiding subject dispersion a input (indexing) and output (searching) stages by controlling.
- Synonyms
- Near synonyms
- Quasi-synonyms and by differentiation of homographs.
- To facilitate the conduct of broad (generic) searches by bringing together, in some way terms that are semantically and syntactically related.

Cont..,

- **Two stages:** conceptual analysis and translation into the language of the system.
- The first step involves an analysis of the request (submitted by the user) to determine what the user is really looking for, and
- The second step involves translation of the conceptual analysis to the vocabulary of the system.
- Thus there is a close resemblance between indexing and search process.

Methods of Vocabulary Control

a) Semantics

- **i) Synonyms:** Synonyms lead to the same concept being denoted by different terms. For control, one of the terms is used/accepted, the rest of the terms are linked to the accepted term. The terms are connected through the connecting device See or Use or Used For (UF),

  e.g., child medicine and pediatrics are synonyms. If child medicine were the accepted term, then there would be a link.

  Pediatrics

  See Child Medicine or

  Use Child Medicine

  and another link

  Child Medicine

  UF Pediatrics

Cont..,

- ii) **Variant Spellings:** A term may have variant spellings. One of these is chosen, and the other spelling would be linked to it, e.g.

Catalog

See Catalogue or

Use Catalogue

And another link

Catalogue

UF Catalog

Cont…,

iii) Homonyms: Homonyms are concepts where the same term may have different meaning and cause a problem in understanding the concept behind a term. Vocabulary control is achieved by providing the context in brackets along with the term,

e.g.,

Bridge (Road)

Bridge (Game)

- **b) Syntactics**

- All the points, discussed above under headings (i) to (iii), are related to the semantics of the language. Thus, they help to control the variations in semantics. Similarly, the flexibility in syntax has also to be controlled. These occur due to headings occurring in compound terms and phrases.

- **i) Compound Terms:** Terms representing a subject may occur as compound terms in a combination of a noun and an adjective. These may be expressed in different ways

  e.g.

  Academic libraries could also be expressed as ***Libraries in academic institutions or as Libraries, academic***.

- **ii) Phrase Headings:** There may be subjects that need to be expressed in phrases using conjunctions or prepositions. There are different ways in which these can be expressed,

  e.g., **Role of libraries in societal development.**

  It could also be expressed as ***Libraries- Role in societal development, or Societal development, role of libraries in.***

Thank you…

Thank you…

Thank you…