INFS 324
INDEXING AND ABSTRACTING

Session 3 – Subject Indexing I

Lecturer: S. Nii Bekoe Tackie, School of Information and Communication Studies,
Department of Information Studies
Contact Information: snbtackie@ug.edu.gh

UNIVERSITY OF GHANA
College of Education
School of Continuing and Distance Education
2014/2015 – 2016/2017

Session Overview

Subject indexing refers to indexing that is based on the conceptual analysis of the contents of a document. It means the identification of concepts or ideas or topics to represent the subject matter of a document in the index. As a matter of fact this whole course is about subject indexing. I have therefore divided this segment into two Sessions for ease of understanding. Thus in this Session, I will be taking you through what will make for an effective indexing system and how to measure the effectiveness of an indexing system.

Session Objectives

- Understand the need for subject analysis in order to be able to produce an effective index.
- Identify the parameters of an effective indexing system
- Explain the influence of the parameters on each other.
- Appreciate the intellectual effort that is involved in the determination of concepts to use in the index.

Session Outline

The key topics to be covered in the Session are:

- Indexing and Subject Analysis
- Parameters of an Effective Indexing System
- Identifying Indexable Concepts

Reading List

Topic One:

Indexing and Subject Analysis

Subject Analysis

- As you saw earlier on when I was discussing the types of indexes, subject indexing is the creation of indexes from the conceptual analysis of the contents of documents.
- Most users of information approach information sources with queries concerning a particular subject or topic.
- Hence indexers analyze the intellectual contents of documents in order to represent them in the index to facilitate easy access to the relevant document or information.

Subject Analysis(Cont.)

- Authors put their ideas in documents so to index a document the thought content (ideas embodied) of the document has to be analyzed by the indexer in order to determine the importance of what the author has said.

- Subject Analysis is basic to indexing. Without it we cannot create an index.
- In analyzing the subject matter of a document, the indexer selects concepts that will be used in describing the document in the index.
- The indexer names the concept he has selected in his own words or in the words of the author of the document.
- At a later stage in the indexing process, the concepts may be expressed in terms of a particular indexing language.

Subject Analysis(Cont.)

- During subject analysis, it would be seen that a document contains a large number of concepts.
- It would be realized also that the amount of information about each concept differs.
- It is the duty of the indexer to decide which concepts to include and which to exclude from the index.

Subject Analysis(Cont.)

- The decision that results from subject analysis depends on the indexing policy of the institution that is creating or compiling the index.
- The greater the number of concepts included in the index, the more detailed the index is.
- The number of concepts recognized in the index is described as the **exhaustivit**y of the indexing.

Subject Analysis(Cont.)

- **Depth Indexing**

  - When a high degree of exhaustivity is employed in the indexing, the policy that is being followed is called **depth indexing**.

  - Depth indexing recognizes main themes as well as sub-themes.

  - Depth indexing is employed in information units where the needs of the users can be fairly easily foreseen.

  - It is often used for technical reports and other documents which are relatively short.

Subject Analysis(Cont.)

- **Summarization**

  - Juxtaposed to depth indexing is summarization.

  This is the policy where only dominant overall themes are recognized for the purpose of indexing.

  - In other words summarization of a document is the expression of the total contents of the document by a brief description

  For example:

  a document that discusses 'Psychology' would be recognized as the overall theme or concept.

  Such subjects like 'clinical psychology', 'abnormal psychology', 'and child psychology ', 'industrial psychology' etc. would not be recognized.

TOPIC TWO

- PARAMETERS OF AN EFFECTIVE INDEXING SYSTEM

Introduction

- A number of parameters underscore the effectiveness of an indexing system. They are five in number.
- But before I discuss the parameters with you, let me introduce you to what is called the **SOUGHT TERM**:
- This is what a searcher is looking for, when using an index.
- It may be one word or it may be a phrase, for example, ‘psychology’ or ‘clinical psychology’ etc.

Parameters…(Cont.)

- The end result of indexing invariably is to provide a term that may be used to gain easy access to documents or information contained in a database.
- Thus the sought term is affected by the parameters that we are about to look at.
- The parameters are exhaustivity, specificity, recall, precision and fallout.
- Now let us look at what they are and how they affect the indexing system.

Parameters…(Cont.)

EXHAUSTIVITY

- It is the extent to which the indexing system allows for the analysis of the content of a document to its barest minimum.
- That is how fully the subject matter of a document has been represented in the index.
- In order to achieve this objective the indexer has to select as many keywords as possible to represent the author’s ideas in the document.

Parameters…(Cont.)

SPECIFICITY

- This refers to the extent to which the indexing system allows for precision when searching for information within the index.
- That is how broad or specific the terms or keywords selected in a particular situation, are.

- For example:

‘Orange’ is a more specific or precise term to use for a search than ‘Citrus fruits’ when an information seeker is searching for information on oranges.

Parameters…(Cont.)

RECALL

- It is a measure of the efficiency of an indexing system in retrieving information or documents.
- Recall is reckoned in percentages.
- It is measured by the relevant terms retrieved over the number of relevant terms in the system multiplied by a hundred.
- Thus if there are fifty relevant terms in the index and twenty are retrieved, the recall efficiency of the system would be calculated as $\frac{20 \times 100}{50}$

Parameters…(Cont.)

PRECISION

- This refers to the number of relevant terms retrieved over the total number of terms retrieved multiplied by hundred.
- Thus if five out of the twenty terms retrieved proved to be useful, the precision rate of the system will be reckoned as

$$\frac{5 \times 100}{20}$$

- This is also a measure of the efficiency of the system.

Parameters…(Cont.)

FALLOUT

- Fallout ratio is another parameter used to measure the efficiency of the indexing system.
- It is the ability of the system to suppress or not to retrieve irrelevant terms.
- It is also reckoned as total irrelevant terms retrieved over the total relevant terms in the system multiplied by a hundred.

Parameters…(Cont.)

Interactions of the Parameters

- The parameters affect each other in the way they behave in an indexing system and eventually affect the effectiveness of an index.
- Now let me explain how they affect each other.

Recall and Precision are affected by Exhaustivity and Specificity.

- A high level of exhaustivity ensures a higher recall because there is greater representation of the subject matter of the document.
- That is to say that there are more terms which increase the possibilities of retrieving more relevant terms.
- However, a higher degree of exhaustivity tends to lower precision because of the possibility of retrieving terms or concepts that have received only a partial or narrow treatment in the document.

Parameters…(Cont.)

- In the same vein, the more specific is the term, the better the precision.
- However, when the level of specificity is increased, it lowers the level of recall.
- That is to say, when very precise terms are used only those terms would be retrieved in a search
- Other terms which may be imprecise and so are not retrieved may contain equally important or vital information on the subject matter.

- In practice indexers attempt to achieve a balance between recall and precision because it is simply not possible to achieve 100% recall and 100% precision at the same time.
- Thus Lancaster (2003), proposes that an intermediate performance level of 50% to 60% variation for both recall and precision is acceptable.

Topic Three:

IDENTIFYING INDEXABLE CONCEPTS

Introduction

- A document may treat several (topics) concepts with varying degrees of information on each concept.
- It is the indexer's responsibility to analyze the thought content (subject matter) of the document in order to determine the concepts that would be represented in the index.
- This is often a difficult task for the indexer. In this section, I will discuss how the indexer goes about the task of selecting concepts to be included in the index
- The purpose of analyzing the subject content of the document is to enable the indexer to identify index able concepts.
- In analyzing the subject content the main aim is to ensure that no important information has been overlooked.

Introduction(Cont.)

- The indexer may have to relate the content of the document to the users of the index.
- This is because:

  -Not all the items of information in the document are worth indexing.

  -Again different items may have different amounts of information.

  -Some textual documents are not worth a detailed analysis e.g. catalogues, brochures, trade publications etc.

Questions to Pose

- There are questions the indexer has to pose to help him identify concepts.
- Some of these are:

  -To what extent is the document about a particular subject?

  -Is there enough information about this particular concept in the documents?

  -Would the user searching for information on this concept be satisfied with this document?

  -Is there any possibility that the concept will feature in a search query?

Questions to Pose(Cont.)

British Standards, BS 6529 (Chowdury, 2004)

has the following questions to the general factors that should be considered in determining concepts to be represented:

- Does the document deal with a specific product, condition or phenomenon?
- Does the subject contain an action concept, an operation or a process?
- Is the object or patient affected by the action identified?
- Does the document deal with the agent of this action/

Questions to Pose(Cont.)

- Does it refer to particular means for accomplishing the action eg. Special instruments, techniques or methods?
- Were these factors considered in the context of a particular location or environment/
- Are any independent or dependent variables identified?
- Was the subject matter considered from a special view point not normally associated with that field, eg. A sociological study of religion?

Intellectual Involvement of the Indexer

- It is important to note from the foregoing that the determination of the concepts to use requires intellectual involvement of the indexer.
- As a result it is possible that different indexers may analyze the contents of a document in different ways resulting in different index entries for the same document.
- This assertion is illustrated by Cleverdon (1984) thus;

  - If two people or group of people construct a thesaurus in the same subject area, only 60% of the index terms may be common to both thesaurus

  - If two experienced indexers index the same document using the same thesaurus only 30% of the index terms may be common.
- This is a problem related to manual indexing systems. Automatic indexing systems avoid this problem of inconsistency(Chowdury,2004).

Intellectual Involvement of the Indexer(Cont.)

- The above are general guidelines.
- In practice, however, an indexer may have guidelines to help him make decisions about which concepts to include in the index and which ones to exclude from the index
- For example:

- Commonwealth Agricultural Bureau International (CAB International) has the following guidelines about concepts that should be indexed.

  – Organisms e.g. snakes, tigers
  – Geographical Locations e.g. Kumasi, Tamale, UK
  – All relevant concepts like techniques, behaviour
  – Bibliographical terms like conferences, books, theses etc.

References