Indexing from Manual to Web Era

Prepared by

Amera Ahmed El-Sayed Mostafa

Assistant Lecture in libraries and information Science Dept.

Faculty of Arts - Menofia University

List of content

0. Introduction

There have been substantial efforts in experimentation dealing with indexing and content analysis, especially comparative studies between manual and automatic indexing. Sparck Jones (1981)[1], for example, provided an excellent assessment of comparative studies for twenty years including the landmark Canfield and Canfield studies, conducted under the direction of Cleverdon in the United Kingdom. Automatic indexing research has shown remarkable progress within the last fifty years. Beginning with Luhn's experiments[2] in the late 1950's, automatic indexing algorithms became more sophisticated through the work of Salton and others in the late 1960's and early 1970's, and have been implemented in search engines in the late 1990's

It has been claimed that automatic indexing is comparable to, or sometimes even better than, manual indexing because the automatic indexing systems often provide completely satisfactory retrieval results (Salton, 1969)[3]. Automatic indexing has been developed so as to mimic human indexing. Compared to manual indexing, automatic indexing is preferred by professional users because automatic indexing can provide many features such as a higher level of exhaustively, much larger indexable matter, a higher level of specificity, and a much larger indexing vocabulary

[1] Sparck Jones, K. (1981). **Information Retrieval Experiment**. London: Butterworths (pp. 256-284).

[2] Dana Indra (2004) .**a comparison of manual indexing and atomatic indexing in the humanities** . Canada : national library of Canada

[3] Salton, G. (1969). **A Comparison Between Manual and Automatic Indexing Methods**.American Documentation, January.

(Anderson and Perez-Carballo, 2001)[4]. In addition to that, automatic indexing can be applied to large collections that change frequently.

Even though computer-generated indexes have been used by a majority of the World Wide Web' s search engines, there have been major problems with the search utilities. One of the difficulties (or drawbacks) is the relevancy of the items retrieved. The search utility that is usually based on automatic indexing provides a high level of retrieval with a low level of relevancy. This problem has become a major issue in searching the Internet using search engines.

The environment for information retrieval on the World Wide Web differs from that of "conventional" information retrieval in a number of fundamental ways. The collection is very large and changes continuously, with pages being added, deleted, and altered. Wide variability between the size, structure, focus, quality, and usefulness of documents makes Web documents much more heterogeneous than a typical electronic document collection. The wide variety of document types includes images, video, audio, and scripts, as well as many different document languages. Duplication of documents and sites is common. Documents are interconnected through networks of hyperlinks. Because of the size and dynamic nature of the Web, preprocessing all documents requires considerable resources and is often not feasible, certainly not on the frequent basis required to ensure currency. Query length is usually much shorter than in other environments-only a few words-and user behavior differs from that in other environments. These differences

[4] Anderson, J. D., and J. Perez-Carballo. (2001). **The Nature of Indexing: How Humans and Machines Analyze Messages and Texts for Retrieval.** Part I: Research, and theNature of Human Indexing. Information Processing & Management. 37: 231 -254.

make the Web a novel environment for information retrieval (Yang, Kiduk 2005)[5].

Importance of the study

The study provides an understanding of index term characteristics In range of paragraphs. This study indicates that the nature of two indexing systems is different. Automatic indexing generally works on the basis of algorithms, while human indexers assign terms using their background and knowledge[6]. The study demonstrates that the differences in their indexing characteristics would lead to different outcomes, but it would also be useful if their terms were combined to provide the best terms.

Techniques for automated indexing and information retrieval (IR) have been developed, tested, and refined over the past 40 years. With the introduction of the Web, and the capability to index and retrieve via search engines, these techniques have been extended to a new environment. They have been adopted, altered, and in some cases extended to include new methods. “In short, search engines are indispensable for searching the Web, they employ a variety of relatively advanced IR techniques, and there are some peculiar aspects of search engines that make searching the Web different than more conventional information retrieval” (Gordon & Pathak, 1999, p. 145)[7].

Indexing:

[5] Yang, Kiduk (2005) .**Information retrieval on the web** . annual review of information science and technology , V.39

[6] Borko, H. (1977). **Toward a Theory Indexing. Information Processing & Management**,13: 355-365

[7] Gordon, M., & Pathak, P. (1999). **Finding information on the World Wide Web: The retrieval effectiveness of search engines**. Information Processing & Management, 35, 141-180.

An index has an important role in information retrieval systems because it provides users with a systematic tool for locating the documents they need (Wellisch, 1994)[8]. Indexing is a process for producing a representation of the document content so that the documents are accessible to users. People use an information retrieval system to solve information problems. The query, as are presentation of an information problem, is compared to the index, as a representation of the documents, to locate the information that users need.[9] Identified problems in information retrieval, including how to represent document contents and information that users need, and how to devise the query-index comparison so that users can identify documents of interest to them.

Borko and Bernier (1978, p.8)[10] defined indexing as "...the process of analyzing the informational content of records of knowledge and expressing the informational content in the language of the indexing system." Considering this definition, indexing has at least two components, namely selecting concepts in a document, and expressing the concepts selected in an indexing language. More specifically, Chan (1994)[11] noted that indexing is not only about assigning the term, but also identifying interrelationship among the terms. She pointed out that indexing involves basically three steps: (1) determining subject content of the item, (2) identifying multiple

[8] Wellisch, H.H. (1995). **Indexing from A to Z** . New York: H.W. Wilson

[9] Chu, C.M., and I. Ajiferuke. (1989). **Quality of Indexing in Library and InformationScience Databases**. Online Review. 13: 11-35.

[10] Borko, H and C.L. Bernier. (1978). **Indexing Concepts and Methods**. New York:Academic Press.

[11] Chan, L. M. (1994). **Cataloging and Classification, an Introduction** (2nd ed.). New York: McGraw-Hill.

subjects and/or subject aspects and interrelationships, and (3) representing them in the language of the subject headings list (Chan, 1994, p. 166). The following section discusses two principal components in indexing.

- Determining Concepts

The first step in indexing is deciding what a text is about. This involves cognitive processes that are rarely explored in manual indexing studies. When the indexers assign index terms they usually do not read the whole document, but they do reading and skimming (Moens, 2000, p.56)[12]. There are particular parts of the document that are carefully read to extract the content of the document in the shortest period. These include summaries, conclusions, abstracts, first paragraphs of the section, and closing sentences of paragraphs, illustrations, diagrams, and tables and their captions. They skim the rest of the document to understand the document content as a whole.

There are several steps involved in deciding the content of documents. "Human indexers perceive (read, view, examine, listen to) a text, interpret the message encoded inthe text as they understand it (influenced by previous experience and current personal knowledge, including their interpretations of any instructions given them), and then describe their version of the message, plus any important text or document features, in accordance with rules and patterns for the types of index they are working on"

[12] Moens, M. (2000). **Automatic Indexing and Abstracting of Document Texts**. Boston: Kluwer Academic Publishers.

(Anderson and Perez-Carballo, 2001, p.237)[13].

Capturing the important content of a document usually arises from a linguistic cue that shows the thematic structure of the document on a micro and a macro level. On a macro level, the notion of the topic can be identified through a paragraph that has the most links to other paragraphs, or it can be identified in the first sentence of a paragraph. On a micro level, the topic is identified by a noun phrase that appears many times in the document, or indexers can scan texts for special words or phrases. These language cues are useful for developing automatic indexing algorithms.[14]

- Expressing the Concepts

After determining the main concepts of the document, an indexer selects a set of index terms to represent the document content. The concepts that have been identified in the first step are translated into a set of index terms. Expressing the concepts as index terms can be done into two ways regardless of whether they are assigned or derived terms

[13] Anderson, J. D., and J. Perez-Carballo. (2001). **The Nature of Indexing: How Humans and Machines Analyze Messages and Texts for Retrieval.** I: Research, and theNature of Human Indexing. Information Processing & Management. 37: 231 -254.

[14] Langridge, D.W. (1976). **Classification and Indexing in the Humanities**. London:Butterworth and Co.

(Silvester, 1998; Jonak, 1984; Moens, 2000)[15]. A derived-index term uses words taken from the item itself, while an assigned-index term is provided by an indexer who uses some intellectual effort to determine the subject matter of the document. A derived indexing system will assign natural language terms while an assigned indexing system will use controlled vocabularies.

1. Natural language

A natural language system can be based on human indexing, machine indexing, or no indexing at all (Lancaster, 1986)[16]. Natural language terms are usually extracted from the title, abstract, or from the entire document. In derived-indexing, human indexers or automatic indexing systems identify terms that can represent the document based on term frequency in the text, the location of its appearance (e.g., title, summary, or caption), and from its context (Lancaster, 2003, p.284)[17].

[15] Silvester, J.P. (1998). **Computer Supported Indexing: A History and Evaluation of NASA's MAI System**. In Encyclopedia Library and Information Science, Vol. 61,Supplement 24. Ed. by Allen Kent. New York: Marcel Dekker. p.76-90.

Jonak, Z. (1984). **Automatic Indexing of Full Texts**. Information Processing & Management, 20(5/6):385-394.

Moens, M. (2000). **Automatic Indexing and Abstracting of Document Texts**. previous citation …

[16] Lancaster, F.W. (1986). **Vocabulary Control fo r Information Retrieval**, 2nd ed. Arlington,VA: Information Resources Press.

[17] Lancaster, F.W. (2003). **Indexing and Abstracting in Theory and Practice**, 3rd ed.Champaign, IL.: Graduate School of library and Information Science, University of Illinois.

Uncontrolled vocabulary may consist of unpredictable words appearing in the text. For example, in derived-term systems, all descriptors are taken from the document text itself. Any terms that appear in the document, especially in the title or abstract, are candidates for index terms that are usually represented by single words and phrases(Lancaster,1986)[18]

Although natural language has a number of advantages, natural language also has weaknesses (Rowley, 1994)[19]. The weakness of using natural language terms is that the searchers need extra effort to search because they may have too many documents retrieved and they may face common problems in natural language search in which the terms have many synonyms and several narrow terms. Since the natural language terms are unrestricted vocabularies, this exhaustively may lead to loss of precision. Finally, incorrect association in natural language may also cause false drops.

2. Controlled Vocabulary

Lancaster (1986, p.9)[20] defined controlled vocabulary as, a type of index language in which the terminology is controlled.” Controlled vocabulary is sometimes called an index language or subject authority list because it contains the terms authorized for use. Human indexers use this tool to express the document content. They control the index language using terminology lists, instruction

[18] Lancaster, F.W. (1986). **Vocabulary Control fo r Information Retrieval.** 2nd ed. Arlington,VA: Information Resources Press.

[19] Rowley, J. (1994). **The Controlled versus Natural Indexing Languages Debate Revisited: A Perpsective on Information Retrieval Practice and Research**. Journal o f InformationScience, 20(2): 108-119.

[20] Lancaster, F.W. (1986). **Vocabulary Control for Information Retrieval.** 2nd ed. Arlington,VA: Information Resources Press.

manuals, and specially structured worksheets to record the Indexing products. The terms selected from the text are mapped into the appropriate classificatory index terms according to the indexers' perception of the text content. Moens (2000, p.57)[21] pointed out that the concept expressed in assigned indexing is often a combination of words and/or phrases in controlled language terms and natural language terms that frequently appear in the text. Indexers assign those words as index terms based upon their judgment of term importance.

There are a number of advantages to using controlled vocabulary. Users always deal with vocabulary (either controlled vocabulary or natural language) to perform queries in information retrieval systems. The problems of using the natural language include syntax, semantics, generalization, and viewpoint (Wall, 1978 in Perez, 1982)[22]. The problems of *syntax* that will affect the meaning of terms may include sequence or order of words in the text. The *semantic* problems are caused by the confusion generated from synonyms and homonyms, and other word variations, such as tenses, plurals, gender, prefixes and suffixes, and adjectival variation. The *generic* problems may include the loss of specificity or precision because the terms are too generic for a particular topic. Finally, the *viewpoint* problems include the potential differences between the indexer and the end user. These problems can be overcome by employing a controlled vocabulary in the system.

Controlled vocabulary should be able to bridge between indexers and searchers as Lancaster (1986) pointed out. A controlled

[21] Moens, M. (2000). **Automatic Indexing and Abstracting of Document Texts**. previous citation …

[22] Perez, E. (1982). **Text Enhancement: Controlled Vocabulary vs. Free Text.** SpecialLibraries, 73(3): 83-192.

vocabulary is needed to represent the user’s request with the same terms that represent the document contents, to bring together semantically related terms, and to make a search more efficient and effective. Controlled vocabulary is used to facilitate conducting a search by linking together terms which are related. Controlled vocabulary makes it easier to conduct a search if the appropriate terms are available in a definitive list.

There are a number of disadvantages to using the controlled vocabulary. Rowley (1994)[23] noted that in using controlled vocabulary, the searchers have to learn the language before hand because the searchers are not familiar with the controlled vocabulary. Unlike a natural language, constructing and maintaining a controlled vocabulary are expensive. The greater the number of terms in the vocabulary, the more expensive the cost for the vocabulary. Since constructing and maintaining a large controlled vocabulary can be costly, many controlled vocabularies are insufficient and out-of-date.

1. Objectives of the study

Summarized objectives of the study in the presentation of an important subject in the field of libraries and information which is indexing where the researcher developments subject indexing offered starting from the traditional form of indexing and even Web indexing at the present time, as well as Web 2 and the passing automated indexing and regulations and advantages and disadvantages, as well as offered to the study for a quick comparison between studies that addressed the comparison between traditional and automatic indexing indexing.

[23] Rowley, J. (1994). **The Controlled versus Natural Indexing Languages Debate Revisited: A Perpsective on Information Retrieval Practice and Research**. Journal o f Information Science, 20(2): 108-119

The introduction and growth of the World Wide Web (WWW, or Web) have resulted in a profound change in the way individuals and organizations access information. In terms of volume, nature, and accessibility, the characteristics of electronic information are significantly different from those of even five or six years ago. Control of, and access to, this flood of information rely heavily on automated techniques for indexing and retrieval , All of this is what made the researcher discuss the subject of indexing in all stages.

2. Methodology of the Study

The researcher display the information that represents where all previous efforts of research and studies, books and articles that revolve around the subject of manual or automatic indexing and finally display the indexing of the web in a range of paragraphs**.**

This study focused on the research and written articles for indexing in English only. And wrote this study in English that according to the importance of the English language lies the importance of the English language in our time in it's most prevalent among the world's languages, and may be the mother tongue, which can be used in all countries . English has become the global language, the first and most widespread in the world. . English is the language of modern times. English language science and technology and scientific research. English language and computer language study at universities and institutes

- Definitions Used in This Study

Definitions of the central concepts used in the study are given in this section. These definitions are used to provide a common understanding of the concepts discussed in the whole study :

Automatic indexing[24]: is a process of assigning index terms that is performed without, or with very modest, human intervention.

Exhaustively : refers to “the degree to which all the concepts and notions included in the text are recognized in its description, including the central topics and the ones treated only briefly.” (Moens, 2000, p.72)[25].

False drops[26] : refer to documents that are retrieved by a search but are not relevant to the question.

Indexing[27] : is the process of creating a document representation, mainly of its topic or content, although formal representation such as authorship, title, bibliographic context etc., is sometimes included in the term.

Index terms : are a set of words or phrases that are extracted or assigned from the text that represent the content of the text or act as access points for the text.(Tumey, 1999)[28].

Manual indexing : is indexing done by human indexers (as opposed to “automatic indexing”).

[24] Silvester, J.P. (1998). **Computer Supported Indexing: A History and Evaluation of NASA's MAI System**. In Encyclopedia Library and Information Science, Vol. 61,Supplement 24. Ed. by Allen Kent. New York: Marcel Dekker. p.76-90.

[25] Moens, M. (2000). **Automatic Indexing and Abstracting of Document Texts**. Boston: KluwerAcademic Publishers.

[26] Meadow, C.T., B.R. Boyce, and D.H. Kraft. (2000). **Text Information Retrieval Systems**.Second edition. New York: Academic Press.

[27] Wellisch, H.H. (1995). **Indexing from A to Z**. New York: H.W. Wilson.

[28] Turney, P.D. (1999). **Learning to Extract Keyphrases from Text. Ottawa**: NationalResearch Council of Canada.

Precision : for measuring a retrieval performance, is defined as the ratio of number of relevant retrieved to total number retrieved (Meadow, *et al.,* 2000, p. 322)[29]. *Precision* is also used to measure the indexing effectiveness in capturing contents o f the document

3.Manual Indexing

Manual indexing refers to a process o f assigning significant words or terms that is performed manually by trained subject experts and that involves a level of human intellect. Automatic indexing, on the other hand, refers to a process of assigning content identifiers controlled by automatic, machine-performed procedures (Salton, 1989, p.276)[30].

Although humans can easily determine concept abstraction and judge the value of a concept, manual indexing has the associated disadvantages of cost, processing time, and consistency (Salton and McGill, 1983; Kowalski, 1997)[31]. The cost of manual index processing time is more expensive than automatic indexing. Consistency is one of the main issues in human indexing. The following section will discuss indexing consistency in relation with indexing.

[29] Meadow, C.T., B.R. Boyce, and D.H. Kraft. (2000). **Text Information Retrieval Systems**. Second edition. New York: Academic Press.

[30] Salton, G. (1989). **Automatic Text Processing: The Transformation, Analysis, and Retrievalo f Information by Computer**. New York: Addison-Wesley.

[31] Salton, G. and M.J. McGill. (1983). **Introduction to Modem Information Retrieval**. NewYork: McGraw-Hill.

Kowalski, G. (1997). **Information Retrieval Systems: Theory and Implementation**. Boston: Kluwer Academic Publishers.

The general consensus among indexers and theoreticians is that human indexers perceive (read, view, examine, listen to) a text, interpret the message encoded in the text as they understand it (influenced by previous experience and current personal knowledge, including their interpretations of any instructions given them), and then describe their version of the message, plus any important text or document features, in accordance to rules and patterns for the type of index they are working on. Not much more detail than that is provided by experts in indexing. Here are examples of explanations provided by leading experts in human indexing.

Nancy Mulvany[32]In Indexing Books (1994), Mulvany says:
" I do not believe that indexing can be taught. . . . The ability to objectively and accurately analyze text and to produce a conceptual map that directs readers to specific portions of the text involves a way of thinking that can only be guided and encouraged, not taught. . . .Indexing cannot be reduced to a set of steps that can be followed" (p. vii-viii).

". . The indexer's ability to thoroughly digest the intentions of the author and anticipate the needs of the readers, thereby producing a knowledge structure that is sensible and usable, involves the application of abilities and skills that are inherent in some individuals and not in others" (p. 39).

"An indexer with a clear idea of the scope of the book itself and a general understanding of the subject matter and the audience will be in a position to distinguish between relevant and peripheral information.

[32]Mulvany, N. C. (1994). **Indexing books**. Chicago: University of Chicago Press.

"Distinguishing between relevant and peripheral information involves judgment. Careful exercise of such judgment is what sets a true index apart from a computer generated list of words" (p. 45).

Lois Mai Chan[33]Cataloging is the application of indexing procedures to a particular collection of documents. Classification is indexing that results in conceptual groupings of topics, rather than alphabetic arrays of headings. Chan has written widely on cataloging and classification. Here is what she saysabout subject analysis in her popular introductory textbook, Cataloging and Classification: An Introduction (1994):

"No matter what the subject access system within which a subject cataloger is working, subject analysis of a particular work or document involves basically three steps: (1) determining the overall subject content of the item being cataloged, (2) identifying multiple subjects and/ or subject aspects and interrelationships, and (3) representing both in the language of the subject headings list at hand".

"The most reliable and certain way to determine the subject content is to read or examine the work in detail" (p. 166).

Robert Fugmann[34]Writing on "recognizing and selecting the essence of a text", German indexing theorist Fugmann (1993) says:

"Essence recognition is a most fundamental and cognitive process in science. The kind of subjectivity which is inherent in this process does not detract from its fundamentality. To the contrary, all progress in cognition has been achieved through subjectivity. At

[33]Chan, L. M. (1994). **Cataloging and classification, an introduction** (2nd ed.). New York: McGraw-Hill

[34]Fugmann, R. (1993). **Subject analysis and indexing: Theoretical foundation and practical advice**. Frankfurt/Main: Index Verlag.

some time, a genius saw or hypothetically assumed lawful relations which up to then had been hidden to everybody" (p. 74).

4. Automatic Indexing

Automatic indexing is the process of assigning content identifiers that are controlled by automatic, machine-performed procedures (Salton, 1989, p.276)[35]. There have been a number of automatic indexing techniques developed to show that automatic indexing techniques can perform at least as well as manual ones. Kowalski (1997)[36] pointed out that the simplest processing method in automatic indexing is achieved when all words in the document are used as index terms. It becomes more complicated if the aim of the indexing is to emulate a human indexer and to use a limited number of index terms to represent the document content. The following sections describe the automatic indexing techniques that have been used to extract terms representing the text content.

- Automatic Indexing Systems

In the early research, automatic indexing techniques usually involved extracting words and phrases from text, selecting descriptive terms, and indicating relations between terms (Wilson, 1974)[37]. They also included techniques for generating index languages and for manipulating document descriptions in searches.

[35] Salton, G. (1989). **Automatic Text Processing: The Transformation, Analysis, and Retrieval**. New York: Addison-Wesley

[36] Kowalski, G. (1997). **Information Retrieval Systems: Theory and Implementation**.Boston: Kluwer Academic Publishers.

[37] Wilson, E. (1974). **Report on Automatic Indexing** Workshop, April 29th - 30th, 1974,George Hotel, Crawley. Canterbury: Computing Laboratory, the University of Kent.

Automatic indexing techniques have become more advanced with the current availability of computer processing combined with linguistics theory.

It was Luhn (1957)[38] who stated that frequency of word occurrence in an article could give significant information about the content of the document. He indicated that high-frequency words tend to be non-content bearing words because they are too common. Low-frequency words do not represent the document contents either. Luhn used Zipf 's law that states product logarithm of the frequency of a term against rank order is approximately constant, as null hypothesis to specify two cut-offs, an upper and a lower shown at Figure 1 . Luhn suggested that the words between upper and lower borders are significant enough to be used as content-bearing words. These findings are the basis of a number of classical weighting functions.

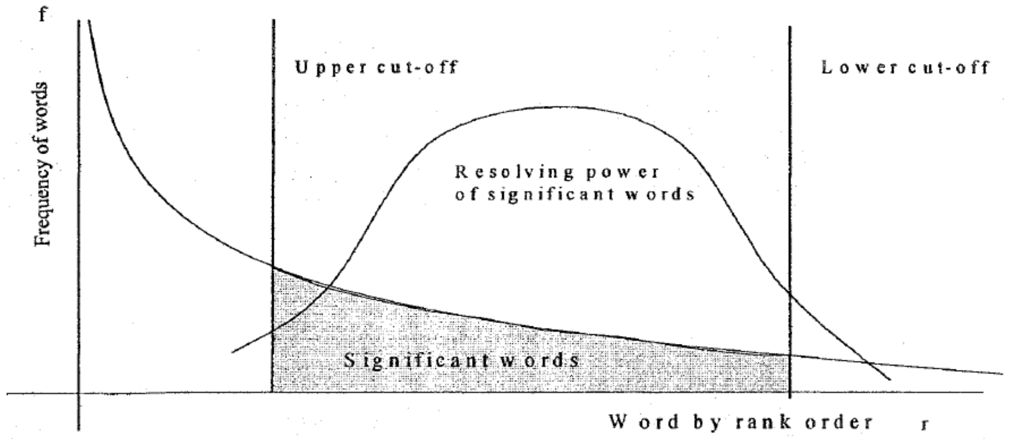

Figure 1. A plot of frequency of occurrence and rank order

[38] Luhn, H.P. (1957). **A Statistical Approach to Mechanized Encoding and Searching o f Literary Information**. IBM Journal of Research and Development, 1(4): 309-317.

(Adapted from Schultz , 1968 in Van Rijsbergen, 1979 , p.16)[39]

Procedures for automatic indexing consist of a number of steps that generally refer to conceptual analysis. Salton (1989, p. 303)[40] described procedures to extract terms from texts: (1) identification of individual words in the text; (2) removal of function words and highly frequent terms in the subject domain; (3) reduction of the remaining words to their stem form, called stemming; (4) optional formation of phrases as index terms; (5) optional replacement of words, word stems, or phrases by their thesaurus class terms; (6) computation of the weight of each remaining word stem or word, thesaurus class term, or phrase term. Ordering of the above steps is possible. For example, recognition of phrases can occur before removal of function words.

Identification of individual words in texts can be done by lexical analysis, which is defined as "a process of converting an input stream of characters into a stream of words or tokens" (Fox, 1992, p. 102)[41]. Lexical analysis begins when the text is stored electronically and can be viewed as a sequence of characters. A word or token is defined as a string o f characters separated by the white space and/or punctuation. Lexical analysis produces candidate index terms that can be further processed, and eventually selected as index terms. Lexical analysis is expensive because it examines every input character.

[39] Van Rijsbergen, C J. (1979). **Information Retrieval**, 2nd ed. London: Butterworths

[40] Salton, G. (1989). **Automatic Text Processing: The Transformation, Analysis, and Retrieval**. previous citation

[41] Fox, C. (1992). **Lexical Analysis and Stoplists. In Information Retrieval: Data Structures & Algorithms**. Ed. By William. B. Frakes and Ricardo Baeza-Yates. Englewood Cliffs, N.J.:Prentice Hall.

Automatic indexing systems can be categorized into statistical, syntactic, semantic, or knowledge-based systems (Silvester, 1998; Lancaster and Warner, 1993)[42]. Hersh (2003, p. 314)[43] categorized syntactical and semantic approaches into a linguistic method that uses natural language processing. Statistical systems refer to those that are based on counts of words or word stems, statistical association, correlation techniques that assign weights to word location, calculation regarding the likelihood of word co-occurrences, clustering of word stems and transformations, or any other computational method used to identify pertinent terms. Syntactical systems are concerned with grammar and parts of speech. A syntactical approach is often used for extracting phrases from the document. Semantic systems refer to those that focus on the context sensitivity of words in the text. The main goal of the method is to identify content-bearing words in the text. Finally, knowledge-based systems are concerned with a conceptual network such as building a thesaurus. While some improvements have been made, the syntactic and semantic techniques are still not easily applied in current systems. They are considered expensive methods to

[42] Silvester, J.P. (1998). **Computer Supported Indexing: A History and Evaluation of NASA's MAI System**. In Encyclopedia Library and Information Science, Vol. 61,Supplement 24. Ed. by Allen Kent. New York: Marcel Dekker. p.76-90.

Lancaster, F.W. and A.J. Warner. (1993). **Information Retrieval Today: Revised, Retitled, and Expanded Edition** . Arlington, VA: Information Resources Press

[43] Hersh, W.R. (2003). **Information Retrieval: A Health and Biomedical Perspective**. 2nd ed.New York, N.Y: Springer-Verlag.

implement and ineffective in constructing terms or phrases (Lancaster and Warner, 1993)[44].

1. Statistical Approaches

The basis of the statistical approach is the use of frequency of occurrence, which is used to calculate a number that indicates the potential relevance of an item. Another technique to improve retrieval performance is the application of the weighting technique (Harman, 1994)[45].

The weighting technique is used to determine which content identifiers should be used as more or less meaningful keywords. The weighting technique ranks terms or phrases according to their significance. Sparck Jones (1973)[46] indicated that the results of the SMART project (which used a weighted technique) showed that frequency weighted terms performed better than un-weighted ones.

There exist statistical models based on the frequency of term occurrence when it is used in combination with weighting techniques. Some of these include the Term Frequency Model, the Inverse Document Frequency Model, and the Term Discrimination Value Model. These are discussed in greater detail in the sections that follow.

[44] Lancaster, F.W. and A.J. Warner. (1993). **Information Retrieval Today: Revised, Retitled, and Expanded Edition** . Arlington, VA: Information Resources Press

[45] Harman, D. (1994). **Automatic Indexing. In Challenges in Indexing Electronic Text and Images**. Ed. by Raya Fidel, Trudi Bellardo Hahn, Edie M. Rasmussen, and Philip J.Smith, ASIS Monograph Series. Medford, NJ: Learned Information. Pp. 247- 275.

[46] Sparck Jones, K. (1973). **Indexing Term Weighting. Information Storage and Retrieval**,9(ll):619-6333.

2. Linguistic Approaches

In statistical models, it is assumed that terms that are statistically related are also semantically related. However, a major problem in such an approach is that the grammatical relationship among words in a sentence often lacks meaning (Hirschman, et al., 1975; Lancaster and Warner, 1993, p.263)[47]. For example, the statistical approach cannot distinguish the relationship between two words that co-occur in the sentence (that means two words are statistically related), whether they are in a subject-verb relation, a host-modifier relation, or no relation at all. The meaning of terms that are statistically related is not always identified in the relationship.

The linguistic method is another method that is often used to generate a multiword index. Although the use of single words using statistical methods in indexing and retrieval is successful, single words still have a number of problems because they cannot represent all the information in the document. Single words present problems in synonyms, polysemy, and phrases, as Hersh (2003, p.310)[48] indicated. Many single words have more than one synonym, that is, different words having the same meaning, such ascancer and carcinoma. Many single words also have polysemy, which is one word that has more than one meaning. For example, lead has two meanings, namely, (1) the verb indicating movement, and (2) an element. Common words often have many senses (Hersh, 2003, p. 310). Word order, for example, has an impact of the

[47] Hirschman, L., R.Grishman, and N. Sager. (1975). **Grammatically-Based AutomaticWord Class Formation**. Information Processing and Management, 1 l(l/2):39-57.

Lancaster, F.W. and A.J. Warner. (1993). **Information Retrieval Today: Revised, Retitled, and Expanded Edition** . previous citation ...

[48] Hersh, W.R. (2003). **Information Retrieval: A Health and Biomedical Perspective**. 2nd ed.New York, N.Y: Springer-Verlag.

meaning of an index term. For example, library school has a different meaning from school library. When single words are combined into a phrase, it will have a specific meaning. For example, the single words high, blood, and pressure will have a specific meaning if combined into high blood pressure.

These problems above show that ambiguity of human language is the biggest obstacle to computer-based understanding of the contents of texts. The ambiguity of the language offers a major challenge to computational linguistics in the need to devise algorithms to take this factor into consideration. To find other methods for automatic indexing, some researchers have explored methods based on linguistic principles that create terms using linguistic analysis rather than using a statistical method.

Hersh (2003, p.313) identified levels of languages that include phonology, morphology, syntax, semantics, pragmatics, and world knowledge. Linguistic methods in information retrieval include levels of morphology, syntax, and semantics. At the level of morphology, the main concerns are analysis of parts of words, usefulness in verb tenses, and norm singular/plural. The analysis at this level also includes affixes. At the syntax level, the analysis focuses on the relationship of words in a sentence to each other, how phrases are formed, and also what words modify each other. At a semantic level, the analysis focuses on the meanings of words, phrases, and sentences.

5. Comparative Studies of Automatic Indexing and Manual Indexing

Although studies in comparing automatic indexing and manual indexing have been conducted for years, the results remain unclear as to whether automatic indexing is comparable to manual indexing, especially in the field of the Humanities, an area rarely explored for automatic indexing.

Automatic indexing systems have been claimed to be comparable to manual indexing systems (Salton, 1969; Janssen and van der Meulen, 1977; Janos, 1975)[49]. However, it is still questionable whether automatic indexing is able to extract appropriate terms as well as humans can.

Sparck Jones and Bates (1977)[50] reported comparative tests of automatic indexing projects undertaken during 1974-1976. The findings showed that the use of statistical weighting is effective and leads to improvements in performance. Variations in input data properties were not typically major influences on performance. Automatic indexing using a statistical approach is thus competitive with manual indexing. The same results showed up in a study comparing the SMART system to MEDLARS. The study showed that simple automatic text analysis is comparable to controlled indexing performed conventionally.

Although such studies show that some simple automatic text techniques are comparable to manual indexing, other studies have

[49]Salton, G. (1969). **A Comparison Between Manual and Automatic Indexing Methods**.American Documentation, January.

Janssen, P.J.F.C., and van der Meulen, W.A. (1977). **Automatic versus Manual Indexing**. Information Processing & Management, 13(1): 13-21.

Janos, J. (1975). **Results of an Experiment with Automatic Indexing Based on the Analysis o f the Texts o f Abstracts**. Information Processing & Management, 11(3/4): 115- 122.

[50]Sparck Jones, K., and R.G. Bates. (1977). **Research on Automatic Indexing** 1974-1976,Vol. 1. Cambridge, UK: Computer Laboratory, University o f Cambridge.

found contradictory results. For example, (Dana 2004)[51] studied index terms based on term frequency in journal articles and compared the results with human-assigned index terms. She found that the documents could not be characterized by terms assigned based on term frequency.(Dana, 2004. fromSeo,1995) conducted a study to compare automatic syntactic-statistical indexing and manual indexing using Korean language texts. She developed an experimental database of 100 long form abstracts and 200 short form abstracts covering business subjects. She found that the terms generated by automatic indexing were less similar compared to those assigned by human indexers.

6.Indexing the Web

Given the size, breadth, and rate of change of the Web, it is not surprising that automated techniques for indexing its content dominate. Lynch (1997, online)[52] described the need for both human and automated indexing: “the librarian’s classification and selection skills must be complemented by the computer scientist’s ability to automate the task of indexing and storing information. Only a synthesis of the different perspectives brought by both professions will allow this new medium to remain viable.” However, despite the democratic nature of Web publishing and the potential for manual indexing of their documents by Web publishers, the practice is not widespread. There are two types of indexing methods on the Web, firstly indexing by Web publishers, and the second is indexing in search engines .

1. Indexing by Web Publishers

Individuals or organizations posting pages on the Web can selfindex their sites by providing significant keywords in contexts that are

[51]Dana Indra (2004) .**a comparison of manual indexing and atomatic indexing in the humanities** . Canada : national library of Canada

[52] Lynch, C. (1997, March). **Searching the Internet**. Scientific American. Retrieved December 20, 2001, from http://www.sciam.com/0397issue/03971ych.html

specifically indexed or even preferentially treated by search engines. In theory, at least, this provides a mechanism for an individual or organization to provide direction to search engines, which extract indexing information from their sites. Many articles and commercial services advise on valid (and sometimes less than ethical) ways for Web publishers to optimize their rankings by providing indexing information; see, for example, Stanley (1997)[53]

Meta Tag is a language codes HTML coding texts superior means of the most reliable web publishers for the preparation of metadata helps describe the substantive content of the pages to tack (such as: tag keywords, and tag Description) This information is stored in the text file for Web pages but does not display on the screen[54].

The problem of indexing Web pages is the ability of web publishers to address the arrangement by placing keywords in duplicate pages to trick the search engines, which is what is referred to a number of terms (such as: Engine Search Persuasion, Stuffing, Spam-Indexing, Keyword Spam).

Many Web documents may warrant indexing of a higher quality than that provided by Web search engines, but the evidence cited here suggests that human indexing of Web documents is relatively rare, at least in the publicly indexable Web, although the situation may be different in the "hidden Web" (that portion of the Web that is dynamic, stored in local databases, and created on demand)[55].

[53] Stanley, T. (1997b). **Moving up the ranks**. Ariadne, 12. Retrieved December 20, 2001, from http://www.ariadne.ac.uk/issue12/search-engines.

[54] Sullivan, D. (2001, July 2). **Search engine features for Webmasters**. Retrieved December 20, 2001, from http://www.searchenginewatch.com/webmasters/featureshtml

[55] Sullivan, D. (2001, December 11). **Search engine sizes**. Retrieved December 20,2001, from http://www.searchenginewatch.com/reports/sizes.html

Types of indexing by publishers :

- Metadata

Metadata (metacontent) is defined as the data providing information about one or more aspects of the data, such as: Means of creation of the data, Purpose of the data, Time and date of creation, Creator or author of the data, Location on a computer network where the data were created, Standards used[56]

For example, a digital image may include metadata that describe how large the picture is, the color depth, the image resolution, when the image was created, and other data. A text document's metadata may contain information about how long the document is, who the author is, when the document was written, and a short summary of the document. Metadata is data. As such, metadata can be stored and managed in a database, often called a metadata registry or metadata repository.

- Tagging

The HTML meta tags in particular provide an opportunity for Web publishers to specify their own metadata indicating the content of their pages, especially with the keywords and description tags. This information is stored with the Web page without being viewed on screen, and is available to search engines for indexing. It should be noted, however, that not all search engines index the meta tags[57].

- Folksonomy

A folksonomy is a system of classification derived from the practice and method of collaboratively creating and translating tags to annotate and categorize content; this practice is also known

[56] Ayse S. Altingvde, I . (2004) **Metadata-Based Modeling of Information Resources on the Web** . - Journal of the American Society for Information Science & Technology; Jan2004, Vol. 55 Issue 2, p97, 14p

[57] Kobayashi, M., & Takeda, K., (2000). **Information retrieval on the Web**. ACM Computing Surveys, 32(2), 144-173.

as collaborative tagging, social classification, social indexing, and social tagging. Folksonomy, a term coined by Thomas Vander Wal, is portmanteau of folk and taxonomy. Vander Wal explains some of the characteristics of folksonomies by identifying two types: broad and narrow. A broad folksonomy is the one in which multiple users tag particular content with a variety of terms from a variety of vocabularies, thus creating a greater amount of metadata for that content. A narrow folksonomy, on the other hand, occurs when a few users, primarily the content creator, tag an object with a limited number of terms. While both broad and narrow folksonomies enable the search ability of content by adding textual description - or access points - to an object, a narrow folksonomy does not have the same benefits as a broad folksonomy, which allows for the tracking of emerging trends in tag usage and developing vocabularies. Folksonomies became popular on the Web around 2004 as part of social software applications such as social bookmarking and photograph annotation. Tagging, which is one of the defining characteristics of Web 2.0 services, allows users to collectively classify and find information. Some websites include tag clouds as a way to visualize tags in a folksonomy. However, tag clouds visualize only the vocabulary but not the structure of folksonomies, as do tag graphs[58].

2. indexing in search engines

Descriptions of search engines and their methods and algorithms at the implementation level are scarce, although nonproprietary details are available or discernable. According to Gordon and Pathak (1999)[59] , “the precise algorithms that search engines use for retrieval are not publicized, but one can infer their approximate

[60] **Folksonomy** Available on-line http://en.wikipedia.org/wiki/Folksonomy

[59] Gordon, M., & Pathak, P. (1999). **Finding information on the World Wide Web: The retrieval effectiveness of search engines**. information Processing & Management, 35, 141-180.

workings by reading the Help, Hint or FAQ pages that accompany them as well as by being familiar with the field of IR.

Most major search engines have a centralized architecture, with the index and retrieval engines located on a single site. Search engines have a number of necessary components: the crawler (or robot) is a program that traverses the Web, following links and retrieving pages for indexing. The indexer module extracts the words (or some subset of words) and (in some cases) hyperlinks from each page and creates indexes to them. (Arasu et al. 2001 distinguish a collection analysis module that creates additional indexes)[60] . The retrieval engine consists of a query module that receives and fills users' queries and a ranking module that compares queries to information in the indexes, producing a ranked list of the results. The design of these components raises research questions related to optimizing the performance of the search engine.

- **The crawler**

Crawlers treat the Web as a graph and, using a set of known URLs as a seed set, usually traverse the graph either breadth first or depth first. Research on crawlers addresses both effectiveness and efficiency issues, although they may be interrelated because a more efficient crawling algorithm may save resources while improving the quality of the database. Research issues include how to prioritize URLs to obtain the best pages (due to resource limits on the proportion of Web pages that can be indexed). Cho, Garcia-Molina, and Page (1998)[61] presented a number of URL-ordering schemes

[60] Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., & Raghavan, S. (2000). **Searching the Web**. Stanford University Technical Report 2000-37. Retrieved December 20, 2001, from http://dbpubs.stanford.edu/pub/2000-3

[61] Cho, J., Garcia-Molina, H., & Page, L. (1998). **Efficient crawling through URL ordering**. Proceedings of the Seventh International World- Wide Web Conference rwwW7), published as Computer Networks and ISDN Systems, 30(1-7), 161-172.

based on metrics of page importance, showing that a good ordering strategy makes it possible to efficiently obtain a significant proportion of important pages. Najork and Wiener (2001, using Page Rank as a quality metric, found that a breadth first crawling strategy tends to deliver high-quality pages early in the crawl[62].

- **Indexer**

Includes three types of indexes : text index that includes keywords and titles semantic and sentences contained in the document indexed content. Where he works to extract all the words of all the pages, and record the unique determinants of sites and place the appearance of each word , Structure index that reflect the links between pages, and include information concerning the structure of hyperlinks to pages indexed and stored in a file known as the main index reptiles and rely on it in the pages to follow pulled through hyperlinks , And Utility index ; indexes entities other than entities encoded high-texts, such as PDF files index and indexes of images[63].

- **search engine program**

The role of search engine program starts when you write a keyword in the lucrative search (search box) as this program takes a keyword and looking for Web pages that meet the query, which is being indexer program in the index database, and then displays the search result of web pages requested by the user in browser window "browser window" is also in the process of arranging for these pages[64].

[62] Najork, M., & Wiener, J. L. (2001). Breadth**-first search crawling yields high-quality pages**. 10th Znternational World Wide Web Conference (WWW10). Retrieved December 20, 2001, from http://www10.0rg/cdr0m/papers/208

[63] Talim, J., Liu, Z., Nain, P., & Coffman, E. G., Jr. (2001). **Optimizing the number of robots for Web search engines**. Telecommunications Systems, 17,243-264.

[64] Spink, A., Wolfram, D., Jansen, B. J., & Saracevic, T. (2001). **Searching the Web: The public and their queries**. Journal of the American Society for Information Science and Technology, 52, 226-234

References

1- Anderson, J. D., and J. Perez-Carballo. (2001). **The Nature of Indexing: How Humans and Machines Analyze Messages and Texts for Retrieval**. Part I: Research, and theNature of Human Indexing. Information Processing & Management. 37: 231 -254.

2- Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., & Raghavan, S. (2000). **Searching the Web**. Stanford University Technical Report 2000-37. Retrieved 20, 2001, from http://dbpubs.stanford.edu/pub/2000-3 December

3- Ayse S. Altingvde, I . (2004) **Metadata-Based Modeling of Information Resources on the Web** . - Journal of the American Society for Information Science & Technology; Jan2004, Vol. 55 Issue 2, p97, 14p

4- Borko, H. (1977). **Toward a Theory Indexing. Information Processing & Management**,13: 355-365

5- Borko, H and C.L. Bernier. (1978). **Indexing Concepts and Methods**. New York:Academic Press

6- Chan, L. M. (1994). **Cataloging and classiffication, an introduction (2nd ed.)**. New York: McGraw-Hill

7- Cho, J., Garcia-Molina, H., & Page, L. (1998). **Efficient crawling through URL ordering**. Proceedings of the Seventh International World- Wide Web Conference rwwW7), published as Computer Networks and ISDN Systems, 30(1-7), 161-172.

8- Cousins, S.A. (1992). **Enhancing Subject Access to OPACs: Controlled VocabularyVersus Natural Language**. Journal o f Documentation, 48(3).

9- Croft, W.B. (1989). **Automatic Indexing.** In B.H. Weinberg (ed.), Indexing: The State of Our Knowledge and the State of Our Ignorance, New York City

10- Chu, C.M., and I. Ajiferuke. (1989). **Quality of Indexing in Library and InformationScience Databases**. Online Review. 13: 11-35.

11- Dana Indra (2004) **.a comparison of manual indexing and atomatic indexing in the humanities** . Canada : national library of Canada

12- Fairthorne, R. A. (1971). **Temporal structure in bibliographical classification**. Littleton, CO: Libraries Unlimited.

13- Fugmann, R. (1993). **Subject analysis and indexing: Theoretical foundation and practical advice**. Frankfurt/Main: IndeksVerlag.

14- **Folksonomy** Available on-line http://en.wikipedia.org/wiki/Folksonomy

15- Fox, C. (1992). **Lexical Analysis and Stoplists. In Information Retrieval: Data Structures & Algorithms**. Ed. By William. B. Frakes and Ricardo Baeza-Yates. Englewood Cliffs, N.J.:Prentice Hall.

16- Gordon, M., & Pathak, P. (1999). **Finding information on the World Wide Web: The retrieval effectiveness of search engines. information Processing & Management**, 35, 141-180

17- Harman, D. (1994). **Automatic Indexing. In Challenges in Indexing Electronic Text and Images**. Ed. by Raya Fidel, Trudi Bellardo Hahn, Edie M. Rasmussen, and Philip J.Smith, ASIS Monograph Series. Medford, NJ: Learned Information. Pp. 247-275.

18- Hersh, W.R. (2003). **Information Retrieval: A Health and Biomedical Perspective**. 2nd ed.New York, N.Y: Springer-Verlag.

19- Hirschman, L., R.Grishman, and N. Sager. (1975). **Grammatically-Based AutomaticWord Class Formation**. Information Processing and Management, 1 l(l/2):39-57.

20- Jonak, Z. (1984). **Automatic Indexing of Full Texts**. Information Processing & Management, 20(5/6):385-394.

21- Janssen, P.J.F.C., and van der Meulen, W.A. (1977). **Automatic versus Manual Indexing**. Information Processing & Management, 13(1): 13-21.

22- Janos, J. (1975). **Results of an Experiment with Automatic Indexing Based on the Analysis o f the Texts o f Abstracts**. Information Processing & Management, 11(3/4): 115- 122.

23- Kobayashi, M., & Takeda, K., (2000). **Information retrieval on the Web. ACM Computing Surveys**, 32(2), 144-173.

24- Kowalski, G. (1997). **Information Retrieval Systems**: Theory and Implementation.Boston: Kluwer Academic Publishers.

25- Lancaster, F.W. (1986). **Vocabulary Control fo r Information Retrieval**, 2nd ed. Arlington,VA: Information Resources Press.

26- Lancaster, F.W. (2003). **Indexing and Abstracting in Theory and Practice**, 3rd ed.Champaign, IL.: Graduate School of library and Information Science, University of Illinois.

27- Lancaster, F.W. and A.J. Warner. (1993). I**nformation Retrieval Today: Revised, Retitled, and Expanded Edition** . Arlington, VA: Information Resources Press

28- Lancaster, F. W. (1991). **Indexing and abstracting in theory and practice.** Champaign, IL: University of Illinois, Graduate School of Library and Information Science (A 2nd ed. was published in 1998).

29- Langridge, D.W. (1976). **Classification and Indexing in the Humanities**. London:Butterworth and Co.

30- Luhn, H.P. (1957). **A Statistical Approach to Mechanized Encoding and Searching o f Literary Information**. IBM Journal of Research and Development, 1(4): 309-317.

31- Lynch, C. (1997, March). **Searching the Internet. Scientific American**. Retrieved December 20, 2001, from http://www.sciam.com/0397issue/03971ych.html

32- Meadow, C.T., B.R. Boyce, and D.H. Kraft. (2000). **Text Information Retrieval Systems**.Second edition. New York: Academic Press.

33- Moens, M. (2000). **Automatic Indexing and Abstracting of Document Texts.** Boston: KluwerAcademic Publishers.

34- Mulvany, N. C. (1994). **Indexing books**. Chicago: University of Chicago Press.

35- Najork, M., & Wiener, J. L. (2001). **Breadth-first search crawling yields high-quality pages**. 10th Znternational World Wide Web Conference (WWW10). Retrieved December 20, 2001, from http://www10.0rg/cdr0m/papers/208

36- O'onnor, B. C. (1996). **Explorations in indexing and abstracting: pointing, virtue, and power**. Englewood, CO: LibrariesUnlimited 182 p.

37- Perez, E. (1982). **Text Enhancement: Controlled Vocabulary vs. Free Text. SpecialLibraries**, 73(3): 83-192.

38- Rowley, J. (1994). **The Controlled versus Natural Indexing Languages Debate Revisited: A Perpsective on Information Retrieval Practice and Research**. Journal o f InformationScience, 20(2): 108-119.

39- Salton, G. (1969). **A Comparison Between Manual and Automatic Indexing Methods**. American Documentation, January.

40- Salton, G. (1989). **Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer**. New York: Addison-Wesley.

41- Salton, G. and M.J. McGill. (1983). **Introduction to Modem Information Retrieval.** NewYork: McGraw-Hill.

42- Silvester, J.P. (1998). **Computer Supported Indexing**: A History and Evaluation of NASA's MAI System. In Encyclopedia Library and Information Science, Vol. 61,Supplement 24. Ed. by Allen Kent. New York: Marcel Dekker. p.76-90.

43- Soergel, D. (1985). Organizing **information: Principles of data base and retrieval systems**. Orlando: Academic Press.

44- Sparck Jones, K. (1973). **Indexing Term Weighting**. Information Storage and Retrieval,9(ll):619-6333.

45- Sparck Jones, K., and R.G. Bates. (1977). **Research on Automatic Indexing** 1974-1976,Vol. 1. Cambridge, UK: Computer Laboratory, University o f Cambridge.

46- Sparck Jones, K. (1981). **Information Retrieval Experiment**. London: Butterworths (pp. 256-284).

47- Spink, A., Wolfram, D., Jansen, B. J., & Saracevic, T. (2001). **Searching the Web: The public and their queries**. Journal of the American Society for Information Science and Technology, 52, 226-234

48- Stanley, T. (1997b). **Moving up the ranks**. Ariadne, 12. Retrieved December 20, 2001, from http://www.ariadne.ac.uk/issue12/search-engines

49- Sullivan, D. (2001, July 2). Search **engine features for Webmasters**. Retrieved December 20, 2001, from http://www.searchenginewatch.com/webmasters/featureshtml

50- Sullivan, D. (2001, December 11). **Search engine sizes**. Retrieved December 20,2001, from http://www.searchenginewatch.com/reports/sizes.html

51- Talim, J., Liu, Z., Nain, P., & Coffman, E. G., Jr. (2001). **Optimizing the number of robots for Web search engines**. Telecommunications Systems, 17,243-264.

52- Turney, P.D. (1999). **Learning to Extract Keyphrases from Text**. Ottawa: NationalResearch Council of Canada.

53- Van Rijsbergen, C J. (1979). **Information Retrieval**, 2nd ed. London: Butterworths

54- Weinberg, B.H. (1981). Word **Frequency and Automatic Indexing**. Ph.D. dissertation.NY: Columbia University

55- Wellisch, H.H. (1995). **Indexing from A to Z**. New York: H.W. Wilson.

56- Wilson, E. (1974). **Report on Automatic Indexing Workshop**, April 29th -30th, 1974,George Hotel, Crawley. Canterbury: Computing Laboratory, the University of Kent.