Pak. j. eng. technol. sci.
Volume 5, No 1, 2015, 27-47
ISSN: 2222-9930 print
ISSN: 2224-2333 online

Review of Indexing Techniques Applied in Information Retrieval

Najmus Saher Shah,
Department of Computer Science, College of Computer
Science and Information System

ABSTRACT

Indexing is one of the important tasks of Information Retrieval that can be applied to any form of data, generated from the web, databases, etc. As the size of corpora increases, indexing becomes too time consuming and labor intensive, therefore, the introduction of computer aided indexer. A review of indexing techniques, both human and automatic indexing has been done in this paper. This paper gives an outline of the use of automatic indexing by discussing various hashing techniques including fuzzy finger printing and locality-sensitive hashing. Two different processes of matching that are used in automatic subject indexing are also reviewed. Accepting the need of automatic indexing in a possible replacement to manual indexing, studies in the development of automatic indexing tools must continue.

KEYWORDS:

Indexing, Automatic indexing, Information Retrieval

1. INTRODUCTION

Technology has widely advanced in creating and sharing of data or information through various platforms. Informationis available in the form of large repositories of an organization, World Wide Web, catalogs, etc. To retrieve data from these large sets of data is not only timely but also as per requirement of a user. These two aspects are very important for the searcher of the data. Retrieving relevant data is a major task of an Information retrieval system. Before data could be retrieved, the system is needed to be developed that could match the words in a query with that in repositories. The development of a retrieval system consists of many steps. One of the initial steps in the development of retrieval is indexing. With the advancement in technology, storing data and that too not only homogeneous, but heterogeneous could be easily stored in huge amount which make a repository of organizations not only large but searching for data in these repositories become tedious. To solve this issue and to make searching quick and efficient such that extracted information is relevant to the user query, indexing of the data plays a vital role.

This paper reviews different methods of indexing data on large corpora.

2. INFORMATION RETRIEVAL SYSTEM

The process of representing information, storing, and accessing information is the task of information Retrieval field.(Yates & Neto, 1999).Information retrieval systems are used to extract relevant data from the web, emails, and library catalogs, etc. Relevant data that can be retrieved can betext, audio, video oran image. Information Retrieval also helps in retrieving and presenting documents that consist of unstructureddata, within collections of documents .Therefore, IR is able to satisfy the information need(Manning, Raghawan, & Schutze, 2008).

Creating information retrievals to mine these data consist of many stages, of which the very first one is to collect various resources of data into a summarized collection called Meta data.This Meta data is the collection of indexed documents. An article or a web page, chapters of a book or abstracts, sentences or emails etc. can be defined as a document (Gonzalez, 2008). The document can also be referred to thegranularity of the information presented to the user.

3. THE PROCESS OF RETRIEVAL

Three different processes must be managed by an IR system(Croft W. , 1993)(Hiemstra, 2001),illustrated in figure 1:

- Documents content representation
- Representation of the user's information need

Comparison of both representations

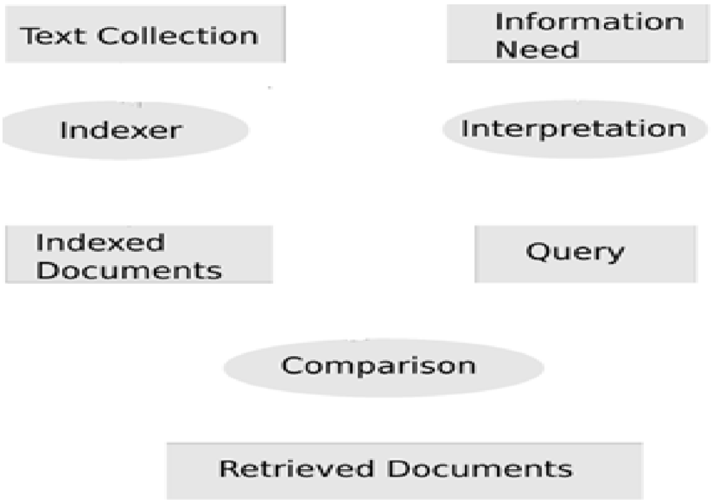

Figure 1: IR General Frame Work

4. INDEXING Documents content representation

Documents are indexed so that searching for information become faster and DBMS does not have to check number of data blocks in the database for data match in response to the user query. Thus, this practice makes the searching process faster.

Croft & Adriani(1997)defined the process of indexing which includes the assigning of keywords or descriptive terms representing the document. Application of indexing can be either manually or automatically. With the gradual growth of the repository, huge amount of time is required by manual indexing. Therefore, automatic indexing could not only be applied to larger repositories,but is much faster than manual with less error problems (Adriani & Croft, 1997),(Gerad, 1986). Indexing consists of several steps involving gathering documents to identifying them and also generating final data structures(Manning, Raghawan, & Schutze, 2008).

5. INDEX PROCESS

Step1: provision of indexable terms:

Effective keyword searching requires clear reflection of the content of the documents in the document collection.Documents' sources in a collection can be varied for example documents can be collected from hard drives, emails or web pages. Information from a web page is collected from web crawler. A web crawler is a computer program that searches for new URls in the directory of URLs and when found a new URL start collecting web pages for indexing. Since web pages are hyperlinked, therefore, to collect information from them, web pages have to be passed through a web crawler.

Step2:determination of character sequence in each document

Collected documents are in different forms and therefore it's necessary to normalized them. Some document may be encoded differently other than ASCII codes. The correct encoded mechanism has to be determined before normalizing text. The contents of web pages are in different formats like HTML.Some files can be in different formats like PDF, XML or Postscript files. For these API(Application Program Interface) is required which provides the manipulation of files in different formats.

Step 3: index granularity

To determine what the document for indexing is. It can be a single document or set of documents. For example a single email can be considered as a document or the thread of an email with an attached document. Therefore, it is necessary to have the knowledge of the structure of the document.

Step 4: Creating tokens

Words are transformed into tokens, by removing punctuation signs, hyphens or down-casing words. Stemming and removing stop words are the two most common techniques.

Step 5: creating inverted files

The last step is to create a data structure which consists of two parts: dictionary and posting files.

6. DEFINITIONS OF INDEXING

To point out, indicate or to guide; that's what an index does. It's a collection of an organized, systematic arrangement of language, signs or symbols that represent ideas(Cleveland & Cleveland, 2013).

Taylor & Joudrey (2009)described indexing as aprocess that helps in evaluating the content of the source of information and determining the aboutness of an item… Indexing is also concerned with describing the information resource in such a way that users are aware of the basic attributes of a document… and the location of the content.

Indexing itself isthe process of creating an index. Derived from the Latin root "indicare", it means to point or to indicate. Embedded in the root, the current meaning has hardly changed, in comparison with the initial meaning. An index is a means to an end and not the end itself(Obaseki, 2010).

Indexing has become an important tool in the area of Information Retrieval such that whenever information is to be systematized or organized, retrieved or used, the need for indexing grows.

Hanson (2004)described Indexing as a " finding device that connects a symbol for a topic(usually in the form of an image or a word) with whatever material is pertinent to that topic in a body of information stored in human memory, in print, or electronically".

7. LITERATURE REVIEW

There are different methods to index data by using different searchengines. An index process that is based on a weighted random walk algorithm was discussed by (Willis & Losess, 2013).In their paper they evaluate the role of thesaurus structure played in an indexing process. A thesaurus- content matching algorithm was introduced by them.Four different collections of documents were, indexed by four different thesauri.

Willis & Losess (2013) observed that some vocabulary performed much better during automatic indexing. From thesauri; manually selection process and term assignment are dependent on the vocabulary structure.Selection of terms by an indexer is based on document frequency concepts, matching of terms between documents, thesaurus and the relations of terms within the thesaurus.

Subject indexing broadly defined the identification of the important subjects of a document and representing them with an indexing language.Thus, Subject indexing is a process dependent on vocabulary hierarchy and thesaurus.The process of Subject indexing, Thesauri and index languages and related automatic indexing techniques have been overviewed by (Willis & Losess, 2013).A Random Walk on a Thesaurus which is a weighted random walk algorithmis described along with an algorithm for term matching using a thesaurus .

In Manual Subject Indexing when a thesaurus is used along with other indexing languages, subject indexing constitute of: a) a collection of document restricted to domain specific, b) list of terms and their relationships that represents domain's specific concepts comprised within a thesaurus, c)indexing guidelines and manuals asexternal sources.

Automatic subject indexing worked with controlled vocabulary by many researchers and many have worked with different document components; Document title, citation or abstract, some worked with subject heading, others worked with document's full text which is newly used.Two broad categories of automated indexing with controlled vocabulary are termed as 1) rule base (expert system)and 2)statistical (machine learning) as described by (Sebastiani, 2002),(Hlava, 2005).In automatic indexing, it's a common practice of matching the language of a document with the vocabulary terms. (Willis & Losess, 2013)

Many researchers like (Vleduts, 1987),(Humphrey & Miller, 1987),(Silvester, Genuardi, & Klingbeil, 1994) and (Hlava, 2005)described rule based approach .According to them controlled indexing terms which are based on text document can be made dependent on formal languages.

Document-Centric Approaches to Matching

Silvester, Genuardi, & Klingbeil (1994)followed the approach of (Vleduts, 1987).They parsed the document into a sentence based on boundaries like punctuations etc. *n*-grams method was used to generate all groupings of words in afive-word window. These groupings were then matched against the vocabulary, and the longest match is selected.

Medelyan (2009)described another approach consist of 4-steps. Firstly, he tokenized the documents into a set of all phrase n-grams whose size is of a certain length.Then theses n-grams, arranged alphabetically, and vocabulary terms are subjected to normalization by applying stop word removal, down casing, stemming and word recording. Joining of preferred and non-preferred terms were considered unnecessary,according to the nature of the vocabulary used for their research. The result is a matched collection of indexing terms.

A Thesaurus-Centric Approach to Matching

The algorithm that is used in this approach is the expanded version of both (Medelyan, 2009)and (Silvester, Genuardi, & Klingbeil, 1994)concepts. The input to the process are thesaurus and document text with a set of matched a term as output. Thedifference of this approach to the earlier ones isthat the thesaurus is matched against the n-grams collection of documents rather than matching n-gram documents to a thesaurus.For a small size of documents and thesaurus, this method is effective bur for lager size ones, all phrase n-grams should be generated in descending length. The output is always the longest match.Thesauri, can also be regarded as trees or graphs as it consists of many relationships.Each term is considered as a node and eachrelationship as an edge.

Algorithm for a Weighted Random Walk on a Thesaurus

Many scholars used this algorithm to explain the behaviors of web servers and stock markets. It is a type of Markov Chain and is most famously used by Google in Page Ranker algorithm to rank web pages.(Willis & Losess, 2013)

To find the effect of this algorithm,(Medelyan, 2009)used sets of vocabulary terms produced by a thesaurus centric matching process . An unidentified graph was constructed and 5 parameters were also introduced for Random walk with probabilities assigned to each thesaurus relationships.More important and frequently encountered points are clustered or separated more accurately against less frequent occurring.Thesaurus used isthe one, that have been used by(Medelyan, 2009).

The results of the algorithm supported the hypothesis of (Medelyan, 2009that,the thesaurus structure helps in term selection. Thesaurus also impacts the matching process individually and that each thesaurus relations contribute differently to subject indexing. There is a good representation of languages and concepts in the collection by a controlled vocabulary. The Vocabulary structure could also be able to play an important role in the indexing process but the amount of vocabulary structure is different for different thesaurus.

Stein & Martin (2007)discussed the twohash-based indexing approaches and compared the improvement in the performances of information retrieval when these are applied.Three tasks i) grouping, ii) similarity search iii) classification of the text retrieval were selected where hash –based indexing was applied.

Stein & Martin (2007) while discussing grouping specified, grouping playing an important role when searching on the web. The returned sets are unrefined and may consist of duplicates. Therefore, there is a need for refinement and duplication cleansing. Cluster Analysis has been used to solve this problem by categorizing search engines(Zamir & Etzioni, 1998),(zu EiBen & Stein, 2002).The performance of these retrieval tasks greatly improved when hash-based indexing is applied. A user input his query as a term query to which a result set was compiled as inverted index. The next hash index was applied for the grouping process which can either be categorized or duplicate elimination.

Stein & Martin (2007) explained similarity searching for exact document matching, in which user query is formulated as Boolean query,then the index structure is an inverted index. Search engines like Google, Yahoo are specialized in this. But, if a user query is a document query, hash-based index could be applied as index structure. In this process appropriate key words must be extracted from document query and then term queries must be created from these keywords.Afterwards, their result sets must be compared with document query.

Classification which is another important task of information retrieval can be effectively addressed with Bayes, discriminant analysis, support vector machines, or neural networks (Stein & Martin, 2007). With the increase in the number of classes,classifier with statistical techniques is now nearly impossible. The solution is a hash classifier.

Stein & Martin (2007) used two recently introduced approaches, fuzzy finger printing and locality-sensitive hashing, which can be applied to the vector space representation of a document.

Fuzzy fingerprinting is a hashing approach specially designed for text-based interval, but not necessarily restricted to it(Stein, 2005).Locality-sensitive hashing (LSH) is a generic framework for the randomized construction of hash functions(Indyk & Motwani, 1998).Both applications were applied on the collection sets. The aim was to find their accuracy in retrieving and evaluating their performances.

For measuring the accuracy of retrieving by hash function, (Stein & Martin, 2007)created hash-indexes for each document sets. For each document, precision and recall values were determined with respect to similarity threshold. During the application of vector space model, reference values for precision and recall were calculated by deploying a term weighting scheme and cosine similarity measure.For fuzzy finger printing, two or three fuzzification schemes were applied to adjust recall of the hash .To do locality sensitive hashing,between 10and 20 random vector sets were employed. This was done in relation to reducing the retrieval results.For near duplication detection,custom-built plagiarism collection was compiled by generating 3000 artificial documents. On comparison,it was found that both approaches performed better than the linear one.Fuzzy finger printing performs significantly better in detecting near duplication as compared to locality sensitive hashing.

For similarity search, (Stein & Martin, 2007)used Google, Yahoo and Alta Vista search engines to collect 100000 documents by searching on a specific topic.The recall performance of both the approaches shows that at high similarity, both performances are excellent. The precision for Fuzzy fingerprinting is significantly higher than the precision of locality-sensitive hashing. Since the size of result set of a document query is smaller, therefore precision performance may be negligible. Thus, both Fuzzy finger printing and locality-sensitive hashing can be applied for the tasks.

Sen (2011) evaluated the automatic indexing to manual indexing by comparing both when applied to biomedical literature.Helping Interdisciplinary Vocabulary Engineering, (HIVE) which is a jointly funded project by the University of North Carolina and the National Evolutionary Synthesis Center, North Carolina was selected as an indexing tool.HIVE is described as having an automatic generation approachthat integrates discipline-specific controlled vocabularies encoded with the Simple Knowledge Organization System (SKOS). SKOS is a World Wide Web Consortium (W3C) standard (WC3 , 2011). The advantages of HIVE, as described by its researchers, are in its cost, interoperability and usability(HIVE, 2011).

For large collections, manual indexing is labor intensive (Neveol, Rogozan, & Darmoni, 2009)as well as costly(Vasuki & Cohen, 2010)which has paved way for the development of automatic tools in an effort to reduce time and costs.

Same documents that have been indexed manually have been subjected to automatic indexing using HIVE. As a result, when the overall terms generated by both human and automatic indexing tool were considered, both were comparable but their performances differ in generating minor terms. HIVE cannot distinguish between major and minor terms. It also fails to identify any publication type terms. Matching of their terms reveals that most matches occurred for major terms, few for minor terms and none at all for publication types one as HIVE was unable to generate any publication type terms. On checking the output data for errors, it was found that both have made errors. Both errors figure were comparable over major terms, but if relates to minor terms, human indexer performs better than automatic indexing tool HIVE. Articles not in English were a problem to both of them. Human indexer has translated these articles with the help of translator to some extent. HIVE did not have the advantage of a translator. Thus, human indexer is much efficient as compared to automatic, but with the advancement in technology, automatic indexer will quickly overcome its problems and will be a great help in indexing large corpora.

Cohen et al. (2010) in their paper discussed indirect referencing and defined it asfinding important associations between terms that are connected yet do not happen together in any archive in a collection. Indirect inference is valuable in numerous applications including data recovery since archives that don't contain words in an inquiry may be pertinent to a client's data need. In this way, recovery frameworks that reach past question terms can enhance execution. The thought of getting term vectors from significant archive vectors rose up out of the perception that term vectors can be consistently retrained in random indexing and related models, and in addition the perception that creating positional term vectors utilizing pretrained term vectors expands the inferencing capacity of a permutation based variation of random indexing (random indexing can be adjusted to encode the relative position of terms utilizing vector permutations . Cohen et al. (2010) called this iterative, repeating preparing procedure Reflective Random Indexing (RRI) as the framework creates new inferences by considering what it has gained from an information set in a past iteration.RRI is able to derive meaningful indirect connections from large corpora such as the MEDLINE corpus of abstracts, and as the growth in complexity of the algorithm underlying random indexing is linear to the size of the data being processed, it scaled comfortably to accommodate the increasing size of the MEDLINE database. The database of MEDLINE is look after by US Ntional Library of medicine (NLM) which is an extensive source of biomedical bibliographic information.Thus, RRI promised to provide superior recovery of indirect relations in comparison to the original version which is very important in the area of literature based discovery.RRI holds RI’s alluring properties of adaptability and the potential for incremental upgrades and (Cohen et al., 2010) anticipated the further use of this system to be used in the biomedical domain and beyond.

Vasuki & Cohen (2010) evaluated the proposed system in combination with nearest neighbor search to forecast the MeSH (Medical Subject Headings) terms for indexing the citations of MEDLINE. MeSH is an arrangement of controlled set of keywords used for indexing the citations of MEDLINE. NLM currently use MTI(medical text Indexer) system which is based on vocabulary-based methods. The evaluation showed that the propsed system of RRI outperformed the MTI system indicating RRI to be a useful addition to methods for indexing.

Blei (2012) on surveying a suit of algorithms known as Probabilistic topic models that are used to discover theme first rather than persue words. As collective information keeps on being digitized and put away—as news, sites, Web pages, scientific articles, books, pictures, sound, video, Furthermore, social networks—it turns out to be harder to discover and find what we are searching for. We require new computational devices to help sort out, hunt, and comprehend these unlimited information.As opposed to discovering reports through decisive word pursuit alone, we may first discover the theme and after that analyze the reports identified with that theme. Since, interaction with electronic archives still not this way, but now lots of texts are available and humans do not have enough power to read and examine all of them with the browsing experience explained above. Therefore, probabilistic topic modeling has been developed by machine learning researchers with the objective to determine and interpret large corpora of documents which can be unstructured with thematic information.These are statistical methods that analyze the words from original text to determine themes that are within them and how those topics are joined with one another, and how they change after some time. They have been used to discover patterns in genetic data ,images and social networks.

8. CONCLUSIONS

Lots of work has been done in the field of indexing. Largely, the studies focused on text indexing. In this paper both human (manual) indexing and automatic indexingmethods have been reviewed and the literature review will help researchers in understanding both the techniques. After the review, it can be concluded that manual indexing is an efficient way to index data, but as the size of the corpora is increasing, manual indexing is becoming labor and cost incentive due to which the need of automatic indexer arises. Studies on both the methods have been conducted over the years and both have been evaluated against each other too. One of the studies applied to biomedical literature, though revealed that automatic indexing performance is still much better in generating minor terms in indexing as compare to automatic indexing tool HIVE. This shows that still a lot of research is required in this area. Researchers are investigating more in developing automatic indexing tools by accepting its importance and needdue to increase in data size.Thus, the scope of study in this field is still expanding.

References

Adriani, M., & Croft, W. B. (1997, May). Retrieval Effectiveness Of Various Indexing Techniques On Indonesian News Articles.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*,*55*(4), 77-84.

Cleveland, A., & Cleveland, D. (2013). *Introduction to Indexing and Abstracting* (4 ed.). Santa Barbara: ABC-CLIO.

Cohen, T., Schvaneveldt, R., & Widdows, D. (2010). Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of biomedical* informatics, 43(2), 240-256.

Croft, W. (1993, April). Knowledge-Based and Statistical Approaches to Text retrievals. *IEEE Expert:Intelligent Systems and thier Applications, 8*(2), 8-12.

Croft, W., & Adriani, M. (1997). Retrieval Effectiveness of Various Indexing Techniques on Indonesin News Articles. 1-7.

Gerad, S. (1986). Another look at Automatic Text-Retrieval Systems. *Communications of the ACM, 29*(7), 648-656.

Gonzalez, R. B. (2008). *Index Compression for Information Retrieval Systems.* PhD thesis,University of Coruna.

Hanson, F. (2004). From Classification to Indexing: How Automation Transforms the Way We Think. *Social Epistemology, 18,* 333-356.

Hiemstra, D. (2001). *Using language models for information retrieval.* University of Twente. Ph.D. thesis,Centre for Telematics and Information Technology.

HIVE. (2011, August). Retrieved from HIVE Web Interface: http://hive.nescent.org:9090/home.html

Hlava, M. M. (2005, August). Automatic Indexing: Comparing Rule- based and Statistical -Based Indexing Systems. *Information Outlook, 9*(8), 22-23.

Humphrey, S. M., & Miller, N. E. (1987). knowledge-based indexing of the medical literature:The indexing aid project. *Journal of the American Society of Information Science, 38*(3), 184-196.

Indyk, P., & Motwani, R. (1998). Approximate Nearest Neighbor:Towards Removing the Curse of Dimensionality. *Proceedings of the 30th Symposium on Theory of Computing,* (pp. 604-613).

Manning, C. D., Raghawan, P., & Schutze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

Medelyan, O. (2009). Human-competitive automatic topic indexing. The University of Waikato. Retrieved from http://hdl.handle.net/10289/3513

Medeyan, O., & Witten, I. (2008). Domain independent automatic key phrase indexing with small training sets. *Journal of American Society for Information Science and Technology, 59*(7), 1026-1040.

Neveol, A., Rogozan, A., & Darmoni, S. (2009). A recent advance in the automatic indexing of the biomedical literature. *Journal of Biomedical Informatics, 42*(5), 814-823. Retrieved from http://dx.doi.org/10.1016/ j.jbi.2008.12.007

Obaseki, T. I. (2010, March 19). Automated Indexing: The Key to Information Retrieval in the 21st Century.

Library Philosophy and Practice(e-journal), 338.

Ramakrishna, K., & Rani, D. (2013, January). Study of Indexing Techniques to Improve the Performance of Information Retrieval in Telugu Language. *International Journal of Emerging Technology and Advanced Engineering, 3*(1).

Salton, G. (1986, July). Another look at automatic text-retrieval systems. *Communications of the ACM, 29*(7), 648-656.

Sebastiani, F. (2002). Machine learning in automated text categorization . *ACM Computing Surveys, 34*(1). Retrieved from http://dx.doi.org/ 10.1145/ 505282.505283

Sen, B. (2011, September). The viability of autimatic indexing for biomedical literature. In:International Journal of Health Information Management Research. *15th International Symposium on Health Information Management Research*, (pp. 8-10). Zurich,Switzerland.

Silvester, J., Genuardi, M., & Klingbeil, P. (1994). Machine-aided indexing at NASA. *Information Processing & Management, 30*(5), 631-645.

Stein, B. (2005). Fuzzy-Fingerprints for Text-Based Information Retrieval. In K. Tochtermann, & M. Herman (Ed.), *Proceedings of the 5th International Conference on Knowledge Management(I-KNOW 05) .Journal of Universal Computer Science* (pp. 572-579). Know-Center.

Stein, B., & Martin, P. (2007). Applying Hash-based Indexing in Text-based Information Retrieval. *7th Dutch -Begian Information Retrieval Workshop*.

Stollak, M. J., Vandenberg, A., Burklund, A., & Weiss, S. (2011). Getting Social: The Impact of Social Networking Usage on Grades Among College Students. Proceedings of ASBBS, 18. Las Vegas.

Taylor, A., & Joudrey, D. (2009). *The Organization of Information* (3 ed.). CT:Libraries Unlimited.

Vasuki, V., & Cohen, T. (2010). Reflective random indexing for semi-automatic indexing of the biomedical literature. *Journal of Biomedical Informatics, 43*(5), 694-700. Retrieved from http://dx.doi.org/10.1016/j.jbi.2010.04.001

Vleduts, S. N. (1987). Concept recognition in an automatic text processing system for the life sciences. *Journal of the American Society for Information Science, 38*(4), 269=287.

WC3 . (2011, August). Retrieved from World Wide Web Consortium WC3: http://www.w3.org/2011 August

Willis, C., & Losess, R. (2013). A Random Walk on an Ontology:Using Thesaurus Structure for Automatic Subject Indexing. *Journal of the American Society for information Science and technology, 64*(7), 1330-1344.

Yates, R. B., & Neto, B. R. (1999). *Modern Information Retrieval.* Newyork: ACM Press.

Zamir, O., & Etzioni, O. (1998). Web Document Clustering: A Feasibility Demonstration. *SIGIR'98:Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 46-54). University of Washington,Seattle.U.S.A.

zu EiBen, S., & Stein, B. (2002). The AISEARCH Meta Search Engine Prototype. In A. Basu, & S. Dutta (Ed.), *proceedings of the 12th Workshop on Information Technology and Systems (WITS02).WITS 02.* Barcelona: Technical University of Barcelona.