ISSN:0975-9646

A Novel Techinque For Ranking of Documents Using Semantic Similarity

Rajni Kumari Rajpal
Department Of Computer Science & Engineering
Raipur Institute Of Technology
Raipur(C.G.), India

Mr.Yogesh Rathore
Sr. Lecturer,
Department Of Computer Science & Engineering
Raipur Institute Of Technology
Raipur(C.G.), India

Abstract— As the volume of information is in internet is increasing staggeringly therefore it is required to develop new methods for document retrieval and then ranking them according to their relevance value as per the user query. The quantity and complexity of information available over the internet is rapidly increasing. Information Retrieval helps in searching meaningful and relevant documents present on semantic web. The concept of semantic similarity is used in variety of fields like artificial intelligence, natural language processing, cognitive science, Biomedical Informatics etc. Even after many developments maintaining updated version of documents ordering as per user's request is still a big challenge. Our approach uses natural language processing techniques for preprocessing of documents that includes standard representation of documents, removal of stop words, stemming etc. Then the keywords are extracted using a technique. Later a correlation value between the query keywords and document keywords is calculated for ranking documents by using WordSimilarity-353 Test Collection values. The novel approach suggested in this paper not only dependent on the syntactic structure of the document but the semantic structure also. It includes both lexical and conceptual matching. The combination of conceptual, linguistic and ontology based matching can significantly improve the performance of the retrieval system. Through experiments we have found that this semantic similarity based ranking methodology gives much better results as compared to traditional methods.

Keywords—natural language processing, information retrieval, semantic similarity, ontology

I. Introduction

Natural language processing uses computational methods for representing naturally occurring texts. It helps computers to derive meaning from human or natural language input. It has many stages phonological analysis, morphological analysis, syntactic analysis, semantic analysis, pragmatic analysis and discourse analysis. Natural language processing is used in informational retrieval, summarization, machine translation etc.

Due to the rapid growth of internet , it has become very difficult to organize data .Traditional methods uses lexical information for retrieval but in many cases semantic information can be more useful as to improve the quality of search result. Main aim of search engine is to provide relevant, accurate information as per their query within limited time and rejecting the irrelevant documents. So in order to make them work effectively. Search engines contains huge amount of information which can be provided to user for any specific query. But it is very difficult to decide which documents and information should be given priority that meets user's expectations. Some common problems with searching systems are:

a) Identifying hyponyms and synonyms for keywords present in the query.
b) Poor precision and poor recall
c) Information excess
d) To understand the user's actual requirement and information needs
e) To know how user's use and process information
f) To represent the documents and the query properly
g) To analyze the relevance judgments technique and rank the documents accordingly

Information retrieval is a subfield of computer science that deals with the representation, storage, and access of from large database collections. Information Retrieval (IR) is the method by which a collection of data is represented, stored, manipulated and searched for the purpose of knowledge discovery as a response to a user query. It consists of methods for representing data and returning relevant information to the user. In between these two process, filtering, searching, matching and ranking functions are carried out. For a best information retrieval system only the relevant information as per the user's need should be provided to the user's, rest should be ignored.

Natural language processing plays a very important role in efficient information retrieval system. Natural language processing can help to overcome the problems of IR Systems by performing lexical and syntactic analysis of text , by assignment of logical structures to sentences or by using semantic analysis and hence may provide precise information. Through research it has been proved that that natural language processing provides a potential contribution of natural language processing (NLP) to IR[1][2][3][4].Flank also suggested A Layered Approach for Information retrieval using NLP Retrieval. [17]

Various techniques of Natural Language Processing have been used in Information Retrieval. It includes Simple methods like stop word removal, porter stemming, etc. which usually yields substantial improvements while higher-level processing like word sense disambiguation, parsing results in less accurate results and also leads to increase in the processing and storage cost. This So optimization of NLP has to be done for Information Retrieval for best results[5]. Today's information repository is very diverse, information comes from different sources in various formats using different languages. Interpreting the actual meaning of this information is very difficult. This task can be highly subjective and time consuming. To relate concepts or entities semantic similarity should be calculated.

The semantic information retrieval (IR) improves the performance of the system drastically and overcomes the problem of low degree of recall or precision obtained from traditional keyword matching techniques. It helps to retrieve semantically or lexically related terms that are not present explicitly within the query. Using semantic similarity, the syntactic structure of the document as well as the semantic structure of the document and the query is considered. Hence it includes the lexical as well as the conceptual matching so as to improve the performance of the system. It helps in better understanding of the user's requirement by semantically interpreting the query given by the user on the basis of relevance and ranking value calculated. Semantic similarity plays a vital role in various tasks in natural language processing such as word sense disambiguation, language modelling, document clustering and extraction of synonyms.

The remainder of this paper is organized in the following sections. Section II contains LITERATURE REVIEW, Section III consists of METHODOLOGY used for semantic similarity ranking. Section IV consists of the RESULT and CONCLUSION of the proposed method.

II. Literature Review

Okkyung Choi et al. [6] suggested SW-IQS (Semantic Web Based Information Query System), using ontology server to improve the efficiency and accuracy of information, which is a combination of classification techniques and agents. The system is based on the RDF (Resource Description Framework) and the Ontology whose efficiency and accuracy is verified through a new method of similarity measure by using semantic metadata. To measure the performance of system a new semantic vector model based on cosine similarity measurement of non-binary weights using RDF semantic meta-information was used. Using this system, ranking of the documents extracted from the web and documents present in the Content DB Server was more precise and relevant than other methods.

Elias Iosif et al. [8] suggested two unsupervised web-based similarity metrics. The first type considers only the page counts returned by a web search engine. The second is fully text-based that searches the contexts of downloaded documents which includes the words of interest. Page-count-based similarity metrics is based on word co occurrence factor to measure some kind of semantic relationship between words. Co-occurrence measures like Jaccard, coefficient, Dice coefficient etc. were used for calculating semantic similarity. Fully text-based similarity metrics uses a cosine similarity to measure the semantic distance between words and to generate semantic classes. Through experiments showed that the page-count-based metrics produced low to mid correlation. Fully text-based metric that uses a binary weighting scheme produces good correlation scores. The best performance result of 0.71 was achieved which is the highest correlation value among the unsupervised metrics. But there was low performance for Type 2 (OR) queries.

Byun et al.[20] introduced a corpus independent framework for quick ordering of documents in a dynamic corpus for retrieving relevant documents. System depends on robust judgment of relevancy scores for key phrases. Keyword extraction is an important prerequisite step in most of information retrieval tasks like text clustering and classification, text summarization etc. The performance was calculated by comparing the performance of their co-occurrence based key phrase extraction method with four other existing methods like tf, tf-idf etc. For each document, six sets of key phrases are retrieved.

Sridevi et al.[9] presented a method for the ontology based semantic annotation of web pages with annotation weighting scheme taking the advantage of the relevance of structured document fields. The retrieval model is based on the structural elements, which are used to re-rank the documents retrieval by using the ontology based distance measure. The relevance based concept similarity is combined with the annotation-weighting scheme to improve the relevance. Document Representation is done and ontology distance is calculated using word net. Semantic similarity between concepts is calculated using Resnik method. The query term is then matched with the concept in ontology by using Bayes learning method. This method has been tested on USGS Science directory collection. This method can resulted in precision improvement by 11% from 0.3761 to 0.4119 in relevant measure.

Danushka [10] proposed a supervised ranking-based method to identify relationally similar word pairs to a given word pair using information retrieved from a search engine. Each pair of words is represented by a vector of lexical patterns. Patterns are then extracted and then selected. A ranking Support Vector Machine is trained so as to recognize word pairs with similar semantic relations with given word pair. A dataset containing 374 SAT multiple-choice word-analogy questions was used for training. Performance achieved using this ranking-based approach is better than several previously proposed relational similarity methods achieving an SAT score of 46.9.however Other rank learning algorithms can be used for much better results.

C.S.Bhatia et al. [11] suggested the concept of semantic web which is an extension of current Web that can provide information along with meaning and therefore makes it possible for computers and people to interact effectively.

Semantic Web gives an idea of Ontology learning. Authors proposed a methodology for ontology learning to extract semantics using Grammatical Rule Extraction. decomposition of the text into smaller parts like title, text, etc was done. Naive Bayes method was used to identify re-occurring properties within HTML documents. Texts within documents are classified according to ontology, semantic querying can then be done for better results. They elaborated the ontology learning method to identify complex concepts, and the relationships between concepts. This method can be extended to include documents in other formats like PDF and Postscript also.

Wei He et al.[12] proposed a method for measuring semantic relatedness between words by using lexical context in which first for each word of a word-pair, a lexical context is created using Word Net, which constitutes words that are highly related to the target word. In next step semantic relatedness between a word and lexical context of another word for an original word-pair is calculated using Web Dice coefficient. Semantic relatedness between two original words is thus found by taking into consideration the scores obtained from relatedness between each word and lexical context of another word. Performance of the system was verified through experiment using Miller-Charles benchmark dataset achieving a Pearson correlation coefficient of 0.912. An alternative better method needs to be used to extract lexical context for a word.

Ruofan et al.[13] proposed a re-ranking method which exploits semantic similarity to improve the quality of search results. Top N results returned by search engine are fetched, then semantic similarity is calculated between the candidate and the query to re-rank the results. The ranking position is converted into an importance score for each candidate. Then semantic similarity score is combined with this initial importance score and finally new ranks are generated. Analysis of the combination ratio between these two parts is done to choose the best one. They used NDCG(Normalized Discounted Cumulative Gain) for evaluation of the re-ranking results. This method enhances the search performance and satisfies users' need to a certain extent.

ZENG et al.[14] presented a semantic Web service ranking mechanism based on semantic space vector model. Semantic similarity can be calculated using three dimensional semantic space vector model which can be ranked in according to the semantic similarity measure between user's request and candidate services. So as to improve accuracy of semantic Web service matchmaking and attain high performance as per user's expectations. Optimization of semantic vector space model has to be done.

S.Lavanya et al[15] proposed an approach for machine learning called Latent Structural Support Vector Machines (LSSVM) to compute the similarity measure so as to overcome the limitation of SVM to handle the missing data which occur frequently in statistical data analysis. The LS-SVM makes use of latent variables. The proposed system compares the similarity scores obtained from SVM and LS-SVM. Support Vector Machine uses observed data from dataset. Handling of missing values in the training data can be done by the LS-SVM which gives higher accuracy for classification of synonymous and non-synonymous words. In this the input given to the machine is word pairs collection and the output is the effective clustering of words into synonymous and non-synonymous words groups.

Rashmi et al. in [16] have presented a system that relies on ontology. When user enters a query meaningful concepts are extracted and these concepts and domain ontology are used to expand query. For all the terms original as well as expanded, SPARQL query is built which is then fired on the knowledge base to find appropriate RDF triples. The Web documents that are relevant to the user requested concepts and individuals specified within these triples are retrieved. The retrieved documents are ranked as per their relevance with respect to the user's query. The query expansion makes use of query concepts as well as synonyms of these concepts and the new terms relate with the terms of actual query within a threshold. But it doesn't covers all domains as it focuses on Sports Domain.

A ranking scheme is proposed in [18] by poonam et al. based on semantic similarity between the documents and the query given by user that relies not only on the syntactic structure of the document but considers the semantic structure of the document and the query also. This approach also includes the lexical as well as the conceptual matching. This altogether improves the performance of the system. The proposed ranking model takes into account the concepts and relationship between the concepts which exists both in the query and document. But A better ranking strategy could be designed by using the semantic analysis of web pages using deep statistical analysis relevance of documents and it was not made scalable for semantic web.

III. METHODOLOGY

The methodology we used is for semantic similarity and ranking is depicted in Figure 1. All Documents in database are sent to Preprocessing Module to extract the required parts of the document.

A. **Preprocessing of Documents** :- Before the documents are processed for ranking value calculation it needs to be converted into a common format. This representation helps in processing to be done easily and quickly. Later all the steps of processing can be applied on the common represented documents. main aim of preprocessing is representing documents efficiently both in terms of space and time and to have a good performance in information retrieval .Document preprocessing is a complex process that leads to the representation of each document by a select set of index terms. In our method we first convert all the words to lowercase, remove special symbols and numbers and focus only on text part of the document. The preprocessing includes:-

a) Word token extraction
b) Erase infrequent words and Stop words
c) Stemming
d) Frequency counts

a. Word token extraction

Sentences present within the text documents are broken up into words known as called tokens, the frequency of each word is then calculated. Lexical analysis is used to separate the input alphabet into Word characters, Word separators. Punctuation marks are treated as word separators in most of the cases. Letter cases are also ignores as case matching is not of any importance. We used white space as the delimiter to extract the words.

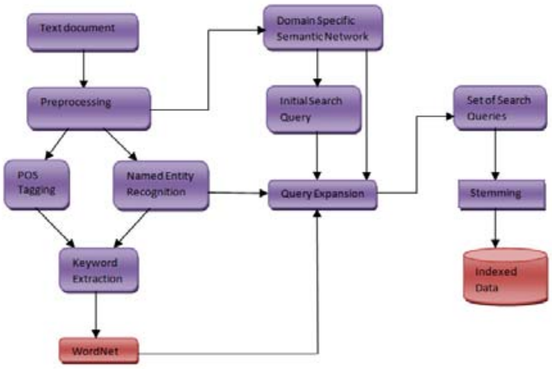

Figure 1: Model for Ranking Documents using Semantic Similarity

b. Erase infrequent words and Stop words

There are many words in a document which are of little value in identifying the documents matching a user need. These words should be are excluded from the vocabulary. Using a stop list significantly reduces the number of postings that a system has to store. Stop-words are generally common words that are not of much interest and importance in a retrieval system. They are generally prepositions, pronouns, conjunctions , articles etc. Stop-words are a part of natural language and should be removed from a text so as to increase the performance measures in terms of memory and time and are of no use in analyzing and categorizing data words. For some search systems stop words are short function common words like the, is, at, which and on. Stop words can create problems when searching for phrases that include them. Some search engines removes most common lexical words which can be lexical words from query so as to improve performance. we have created a stop word list which is checked during processing and if it is a stop word it is eliminated from the text.

c. Stemming

Stemming is done to find out the root of a word. This method is used to remove suffixes of words, to have exactly matching stems, to reduce memory space and time. There are various stemming algorithms. We have used “M.F. Porters Algorithm” to perform effective stemming. In Natural Language morphological variants of words have same semantic interpretations and hence they should be considered as similar for IR applications. Many stemming Algorithms have been evolved which reduces a word to its root word or stem. These stems are used as key terms of the document rather than by the original words. It helps to reduce the number of unique terms needed to represent a set of documents. The smaller the dictionary size, lesser storage space and processing time is saved. Stemming can be performed using Algorithms that stripe off suffixes as per the substitution rules or by using Large dictionaries that is able to provide the stem of each word.

d. Frequency counts

We count the number of occurrences of each word and then retrieve the top ten words which can be the keywords of the document. The result so obtained is then intersected with the above intersection results so as to obtain the final keywords.

B. **Parsing and parts of speech**

The word parsing comes from Latin term “pars” which means part of speech. It is the technique of analyzing a string of symbols in natural language or computer languages as per the rules of a grammar. Sentence parsing is performed to interpret the meaning of a sentence. Under computational linguistics parsing refers to the formal analysis of a sentence or other string of words into its individual constituents by a computer resulting a parse tree that shows the syntactic relation with each other and may contain semantic information. The Stanford Parser has a good accuracy but more training is possible for applying it on domain specific texts. Part-Of-Speech Tagger (POS Tagger) is a piece of software that takes text input in some language and allots parts of speech such as noun, verb, adjective, etc. to each word based on both its definition, as well as its context i.e. its relationship with adjacent and related words in a phrase, sentence, or paragraph. Using Stanford parts of speech Tagger we extract only nouns, verbs and adjectives for keyword identification.

C. **Named Entity Recognition**

Named-Entity Recognition(NER) is a part of information extraction that helps to identify, locate and classify words present in the text document into predefined categories such as the names of persons, locations, organizations, time expressions, quantities, monetary values etc. In this step the Entities are recognized from the sentences and are then stored for use in. In our method, we have considered only Location, Person and Organization for better results. Information Retrieval Based on Extraction of Domain Specific Significant Keywords and Other Relevant Phrases from a Conceptual Semantic Network Structure.

D. **Keyword Extraction**

For keyword extraction we use an algorithm which is based on the extraction of verb and nouns of a particular type. The first noun type of a sentence will extract that type of nouns from the sentence and same is the procedure for verb extraction for object identification. In this way we are able to get some limited type of nouns and verbs which is then intersected with the output of Stanford POS Tagger and NER and top ten frequent words so as to have effective keyword set. Better the keyword better will be the ranking of the document for relevancy measure.

E. **Word Net**

It is a lexical based database of English words. Nouns, verbs, adjectives and adverbs are arranged into sets of cognitive synonyms called synsets, each of which conveys a distinct concept. Synsets are interconnected by means of conceptual-semantic and lexical relations. This resultant network of meaningful words and concepts can be navigated with the help of browser. Word Net is freely available for download. Its structure and data organization techniques has made it a useful tool for natural language processing and computational linguistics.

Word Net resembles a thesaurus, in which it groups words together based on their meanings. But there are some differences as Word Net interlinks not just word but also specific senses of words which helps to find semantically related words that are found in close proximity to one another within network. Word Net labels the semantic relations among words. We use WordSimilarity-353 Test Collection[21] dataset and the documents present in WebSim Proposed by Bollega[19] which contains some documents related to WordSimilarity-353 Test Collection[21] word pairs as well as training documents. Once we have obtained the query keywords and document keywords, this method matches the word pair present in WordSimilarity-353 Test Collection[21] dataset against the words in the document. This will first generate feature vectors for the all word pairs in the Miller-Charles dataset and then scale those features to [0,1] range. If the synonyms of these word pairs or words itself occurs in the document, its frequency is checked and is multiplied by the value present in miller dataset for that particular pair to get the correlation coefficient. We then compare the similarity scores produced by WebSim against the human ratings in the MC dataset and compute correlation coefficients. Pearson correlation coefficient is the correct measure for this dataset. The maximum value among all the calculated values for all words pairs for a particular document is the set as the rank for that document as per the user's query.

F. **Domain Specific Semantic Network**

The corpus and the search space that we use are domain specific. so we build a model of Concept based Semantic Network structure (CSN) manually. The CSN contains various conceptual terms and phrases related to the respective domains and connections among them which are extracted from Word Net. These non-stop terms and phrases are used as initial keywords to perform search [7]

G. **Initial Search Query**

Significant keywords are extracted from the search text with the help of the CSN to which results in initial search query

H. **Search Query Expansion**

Using the Conceptual Semantic Network, possible connections with other conceptual terms related to the initial set of keywords are analyzed and are then added to form an alternate set of queries hence semantic information is added. The synonyms are also added to expand query using semantic information.

I. **Set of Search Queries**

After the system analyzes the text it discovers the domain dependent terms from the Word Net. The System continues to discover similar associations using WORDNET adds synonym set of those keywords as per their parts of speech tag. A collection of sets containing 1…m number of sets is created each having n number of keywords.

The cardinality of the sets that appears inside the superset will range from 2….n. It includes keywords within the search text, keywords from the conceptual network connections which are associated with them.

J. **Indexing**

Indexing is done so as to increase the efficiency by extracting a selected ser of words from the resulting document which can be used for indexing the document. Indexing is a vital process so as to ease the User's ability to find documents on a particular subject of their interest and need. It is done so as to decrease the searching time so that information is displayed within a short period of time. Indexing can be done manually or automatically. Manual indexing is not practical and efficient whereas automatic indexing is more useful as it can be updated easily at any time. It is done on the basis of the class to which document may belong to so as to reduce the search time for best results. And ranks as per the query are stored in index files so that for similar queries in the future can be directly retrieved.

IV. Result and Conclusion

In this paper, we propose a method to make the search quality better by adding semantic similarity as a key component to the ranking result obtained by keyword based method. Since a different technique is used for keyword extraction, hence it leads to better results. The ranking technique proposed in this paper uses the concepts and relationship between the concepts that exists both in the document and the user query to improve the retrieval of relevant documents. Input documents from Dataset WebSim 3.0[19] is used and similarity value is used. We extract the most similar documents to the query entered by the user by using WordSimilarity-353 Test Collection Dataset[21] which contains the similarity values between some words to compute the similarity value and rank the documents accordingly. Through experiments on the documents present in the dataset we have found that the ranking is better than keyword based techniques as well as other techniques most of the times. Hence adding semantic similarity factor leads to better performance of the system in terms of result set as well as time by interpreting the actual need of the user. In most of the cases a correlation higher than 0.75 is achieved. The documents ordered are more similar to the user's order as per relevance with respect to query.

Future work of the proposed system will be to use more larger dataset defining the similarity measure so as to cover every English word so as to have the relevant documents in all domain and eliminate the irrelevant documents having lesser rank values from the result set. We will also try to make our approach scalable for the web documents.

References

[1] B. Hui. , "Applying NLP to IR: Why and how.", Technical report, Department of Computer Science, University of Waterloo,April,1998

[2] D. D. Lewis and K. Sparck Jones, "Natural language processing for information retrieval", Communications of the ACM, 39(1):92–101,1996

[3] A. F. Smeaton. , "Using NLP or NLP resources for information retrieval tasks", In T. Strzalkowski, editor, Natural language information retrieval, pages 99–111. Kluwer Academic Publishers, Dordrecht, NL, 1999

[4] K. Sparck Jones. What is the role of NLP in text retrieval? In T. Strzalkowski, editor, Natural language information retrieval, pages 1–21. Kluwer Academic Publishers, Dordrecht, NL, 1997.

[5] Natural Language Processing in Information Retrieval (2004) by Thorsten Brants , Google Inc In Proceedings of the 14th Meeting of Computational Linguistics in the Netherlands

[6] Okkyung Choi, Sangyong Han, Ajith Abraham, “ Semantic Web based Information Query System for the Integration of Semantic Data”, Proceedings of the International Conference on Next Generation Web Services Practices 2005

[7] Mohammad Moinul Hoque, Prakash Poudyal, Teresa Goncalves and Paulo Quaresma, “Information Retrieval Based on Extraction of Domain Specific Significant Keywords and Other Relevant Phrases from a Conceptual Semantic Network Structure ”,FIRE 2013

[8] Elias Iosif and Alexandros Potamianos, “Unsupervised Semantic Similarity Computation using Web Search Engines”,2007 IEEE/WIC/ACM International Conference on Web Intelligence)

[9] Sridevi.U.K and Nagaveni .N, “Ontology based Similarity Measure in Document Ranking”, 2010 International Journal of Computer Applications Volume 1 – No. 26

[10] Danushka Bollegala , “A Supervised Ranking Approach for Detecting Relationally Similar Word Pairs”, 2010 5th International Conference (Information and Automation for Sustainability (ICIAFs)

[11] C.S.Bhatia and Dr. Suresh Jain, “Semantic Web Mining: Using Ontology Learning and Grammatical Rule Inference Technique”, Conference on Technologies and Applications of Artificial Intelligence 2011

[12] Wei He, Xiaoping Yang, Dupei Huang, “Measuring Semantic Relatedness between Words Using Lexical Context”, 2011 Seventh International Conference on Computational Intelligence and Security

[13] Ruofan Wang, Shan Jiang , Yan Zhang and Min Wang, “Re-ranking Search Results Using Semantic Similarity”, Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) 2011

[14] ZENG ZhiHao, HU JiPing, DONG Ting, WANG Yu, “Semantic Web Service Similarity Ranking Proposal Based on Semantic Space Vector Model”, Second International Conference on Intelligent System Design and Engineering Application 2012

[15] S.Lavanya, S.S.Arya, “An Approach for Measuring Semantic Similarity between Words Using SVM and LS-SVM”, International Conference on Computer Communication and Informatics (ICCCI), Jan. 10 – 12, 2012, Coimbatore, INDIA

[16] Rashmi Chauhan, Rayan Goudar, Robin Sharma, Atul Chauhan, “Domain Ontology based Semantic Search for Efficient Information Retrieval through Automatic Query Expansion”, 2013 International Conference on Intelligent Systems and Signal Processing (ISSP)

[17] Flank, “A Layered Approach for Information retrieval using NLP Retrieval”, Proceedings 17th International Conference on Computational Linguistics, 1998.

[18] Poonam Chahal, Manjeet Singh, Suresh Kumar, “Ranking of Web Documents using Semantic Similarity”, International Conference on Information Systems and Computer Networks (ISCON) 2013

[19] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka “A Web Search Engine-based Approach to Measure Semantic Similarity between Words”, IEEE Transactions on Knowledge and Data Engineering (TKDE), Vol. 23, no. 7, pp. 977-990, July, 2011

[20] Byung-Hoon Park, Nagiza F. Samatova, Rajesh Munavalli, Ramya Krishnamurthy, Houssain Kettani, and Al Geist, “ Rapid and Robust Ranking of Text Documents in a Dynamically Changing Corpus”, Computer Systems and Applications, AICCSA 2008, IEEE/ACS International Conference.

[21] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/