VOL. 12, NO. 2, JANUARY 2017 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

SEMANTIC SEARCH USING LATENT SEMANTIC INDEXING AND WordNet

Anita R., Subalalitha C. N., Abhilash Dorle and Karthick Venkatesh
Department of Computer Science and Engineering SRM Institute of Science and Technology University, Kattankulathur, Chennai, India
E-Mail: anita_kalai@yahoo.co.in

ABSTRACT

Semantic Search and Information Retrieval forms an integral part of various Search Engines in use. Famous search engines such as, Yahoo, Google, Lycos etc. use the concept of semantic search, where the only comparator for the objects under study is semantic similarity between the objects. The general method involves document-to-document similarity search. This sort of search involves the sequential search of documents one after the other, which involves numerous noise effects. An efficient way of improving this technique is the Latent Semantic Indexing (LSI). LSI maps the words under study on a conceptual space. The conceptual space depends on the queries and the document collection. It uses a mathematical function to figure out the similarity between the words, something called as Singular Value Decomposition. It utilizes the words under study and the ones that are being compared and produces appropriate results. The results obtained are free of semantics like synonymy, polysemy etc. Integrating Word Net, a large lexical database of English language is an efficient way to increase the search result. The word under consideration is linked to the application and the semantic similarities of the word are found out. Documents similar to these similarities are then indexed and listed. The proposed model is tested with standard set of Forum for Information Retrieval (FIRE) documents and a comparison with the term based search has been done.

Keyword: singular value decomposition, term document matrix, latent semantic indexing, WordNet.

1. INTRODUCTION

This research aims to describe a new approach to automatic indexing and retrieval techniques. The inherent flaw with existing key word search is that the conceptual meaning of the document is ignored during the process of retrieval. The key words are obtained from the query provided by the user is taken and the documents are mined for these key words. If a document possesses the key words, it is retrieved as material relevant to the provided query. This technique is not entirely reliable, as it does not take into account the conceptual meaning of documents. Individual words do not provide direct evidence about the meaning of the document. Apart from this, this technique fails to take into account that multiple words may have the same meaning; it focuses only on documents containing the specified key words. Documents that do not contain the key words but are semantically relevant to the provided query are not retrieved.

Latent Semantic Indexing is a technique that can be employed to overcome this problem. This is an indexing and retrieval method that makes use of a mathematical technique called Singular Value Decomposition to figure out patterns in the relationship between the terms used and the meaning they convey. It works on the principle that words that are used in the same context usually have analogous meanings. A term-document matrix is created and singular-value decomposition is performed on this. This ensures that the arrangement of the matrix mirrors the important associative patterns in data. Points that are closer to each other refer to documents that are semantically similar.

This system is further improved by integrating 'WordNet' with this. 'WordNet' is a lexical database of the English language and contains synonyms of words arranged in the form of 'Synsets'. Integrating this to the system makes it easier to identify words with similar meanings and makes the search more relevant and effective.

The main purpose of this research is to explore the use of Latent Semantic Indexing as an improvement over general key word search and to improve the process by the use of the 'WordNet'. The deficiency in standard key word search is that it ignores the conceptual meaning conveyed by the document and it works by scouring the documents for key words, thus not taking into consideration the meaning associated with the documents. Latent Semantic Indexing is the answer to this problem as it employs a mathematical technique to form patterns regarding the semantic relationship between documents. The use of 'WordNet' further enhances the system as it makes it easy to examine and evaluate relationships between words and analyze similarity of documents. We hope to arrive at an effective solution that can identify relevant documents without difficulty.

The rest of the paper is organized as follows. Section 2 gives the back ground information of the proposed approach. Section 3 describes the works that are related to the proposed work. Section 4 illustrates the proposed approach. Section 5 discusses the evaluation of the proposed approach and Section 6 lists the conclusions of the paper.

2. BACKGROUND

A. Latent semantic indexing

Documents of text generally comprise a high amount of redundancies and ambiguities and these results in significant noise effects. This leads to higher dimensionality, which contributes to make indexing and retrieval very inefficient. Although the data lies in a high

dimensional space, it is advantageous to lower the dimension to increase accuracy and efficiency of data analysis. Reducing dimensionality goes a long way in improving the representation of data. This happens because the data is understood in terms of concepts rather than words, where a concept is a linear combination of terms. The enhancement here is that individual terms are not searched for; instead a search is performed on the concepts.

B. Singular value decomposition

A matrix algebra technique called Singular Value Decomposition is used to generate a source space with a lot less dimensions. This mathematical method reduces dimensionality of the matrix. It makes use of the inherent higher order structure in association of terms within documents by the largest singular vectors. A new truncated matrix is obtained due to the application of Singular Value Decomposition and a new low dimensional space is obtained. The low rank approximation is used to project the documents possessing a high dimension into a low dimensional space. This way, the performance of the retrieval is immensely enhanced. The following are the steps to perform Singular Value Decomposition.

- A term document matrix A is taken with rank r and dimension tXd.
- Singular Value Decomposition is used to factor this matrix into three matrices such that $A = USV^T$
- Left and Right Singular matrices of A are taken to be the columns of matrices U and V. S is the diagonal matrix.
- A k-dimensional SVD of A gives $A_k = U_k S_k V^T_k$.
- This contains the first k columns of U and V
- Rank is reduced from r to k.

C. Singular value decomposition

WordNet is a lexical database that exists for many languages. In this project, we employ an English language WordNet that groups English words into 'synsets'. These are tree like structures that contain lists of synonyms. WordNet also provides short definitions and examples of their usage. It can be seen as an amalgamation of a dictionary and a thesaurus. WordNet is widely used in automatic text analysis and artificial intelligence applications. The database and the software are available for free download from the WordNet website. Both the lexicographic data and the compiler are available as part of this free download.

WordNet includes several lexical categories such as nouns,s, verbs, adjectives and adverbs. On the other hand, it ignores prepositions, determiners and other such function words.

The following categories are used to group words in WordNet:

- **Nouns**
- **Hypernyms:** Y is a hypernym of X if every X is a (kind of) Y (canine is a hypernym of dog)
- **Hyponyms:** Y is a hyponym of X if every Y is a (kind of) X (dog is a hyponym of canine)
- **Coordinate terms:** Y is a coordinate term of X if X and Y share a hypernym (wolf is a coordinate term of dog, and dog is a coordinate term of wolf)
- **Meronym:** Y is a meronym of X if Y is a part of X (window is a meronym of building)
- **Holonym:** Y is a holonym of X if X is a part of Y (building is a holonym of window)
- Verbs
- **Hypernym:** the verb Y is a hypernym of the verb X if the activity X is a (kind of) Y (to perceive is an hypernym of to listen)
- **Troponym:** the verb Y is a troponym of the verb X if the activity Y is doing X in some manner (to lisp is a troponym of to talk)
- **Entailment:** the verb Y is entailed by X if by doing X you must be doing Y (to sleep is entailed by to snore)
- **Coordinate terms:** those verbs sharing a common hypernym (to lisp and to yell)

3. RELATED WORK

A novel technique for conceptual similarity search of text using word chaining seems to be more efficient at document-to-document similarity search when compared to the standard inverted index [1]. Also, the quality of the results is not compromised, which is quite important. By using a compressed version of all the documents, the storage space required for this process is minimized. This helps enormously in increasing the efficiency of the search. This paper is extremely useful in that it clearly demonstrates how their proposed system performs better than standard similarity search techniques on the basis of quality, storage and efficiency. Another notable achievement of the paper is that it is among the first ones to talk about the performance and quality aspects in the same paper.

Latent Semantic Indexing (LSI), which is often cited as a method to better connect the conceptual meaning of a query to the conceptual meaning in documents [2]. This method does not rely on a few key words in a query and distrusts the vague terms used in queries and documents most of the time.

This paper tackles the computational challenges one face when working with distributed search. Apart

from this, the merger of results from multiple collections of documents is examined and these topics haven't been the focus of intense engineering study, for which the authors have to be commended. They opine that the practical need to understand and analyze data combined with the humanistic need to obtain search results that are personalized will help emphasize the importance of solving the problems mentioned.

The Latent Semantic Indexing approach to information retrieval is described in detail. The successive application is compared by including a number of Eigen values k that differs by ten percent of the maximum of k. A statistical technique is built to analyze and compare the obtained ranking for every value of k, from which it is seen that for each value of k, the rankings differ substantially. However, it is not asserted that any specific value of k is best used and it is found that the rankings with high values of k fail to maintain consistency. This finding represents a potential problem in Latent Semantic Indexing, even for values that differ only by a mere ten percent of the maximum value of k.

Assessing the similarity of words and concepts is probably among the more important topics in Natural Language Processing and Information Retrieval systems. WordNet is a lexical database of the English language that is widely used for the purposes of Natural Language Processing and improving similarity search [3]. In recent times, the focus on employing WordNet to enhance similarity search has been quite high. The Depth of Subsumer of the words and the Shortest Path between them have been used to calculate the closeness of between each pair word. By providing weights to the edges of WordNet hierarchy, the authors of the paper have tried to improve the measure of semantic similarity. A new formula for weighing edges has been put forth and this has been used to calculate the distance between two words. Particle swarm optimization has been used to tune parameters belonging to the transfer functions. Experiments were performed to demonstrate that the resultant correlation has been made better.

There have been few works reported related to Information Access system that uses WordNet, LSI either in isolation or combination.

An indexing technique that uses only WordNet which has been implemented using 1400 documents [4]. This work handles simple sentences and needs to be formatted in Standard Generalised Markup Language (SGML).

A concept based information access using ontologies and LSA has been attempted to test the ambiguity queries [5]. They work focuses on handling the polysemy relationship. The method has been proved efficient for handling short queries and short documents.

Expansion of query to search the relevant emails using the LSI and WordNet has been proposed [6]. This method is exclusively built for searching digital forensics emails.

A simple Query-Document retrieval system has been built and tested on Wikepedia articles [7]. LSI has been used for document representation. The LSI has been used for ranking the retrieved pages.

The proposed system builds an Information retrieval system using both LSI and WordNet. Many semantic relations of WordNet such as, Synonymy, Antonymy, Hyepernymy, Polysemy, Homonymy, Holonymy, Meronymy, etc., are handled thereby reducing ambiguity and increasing precision. The LSI also adds to the quality of retrieval through its concept space construction done using Singular Value Decomposition. The proposed technique also differs from the existing works by handling large corpora of 2000000 Forum for Information and Retrieval (FIRE) documents.

4. PROPOSED WORK

Architectural design represents the structure of data and program components that are required to build a computer-based system. It considers the architectural style that the system will take, the structure and properties of the components that compromise the system, and the interrelationships that occur among all architectural components of a system.

a) **Stop words recognizer:** In this module, the system removes the stop words. Words like "a", "an", "the", "above", "over" etc. are called as Stop Words. The removal then gives us the key words. The keywords are stored and worked upon.
b) **Term document matrix:** In this module, a term document matrix is created. The matrix consists of terms, i.e. key words and documents which are under consideration.
c) **Singular value decomposition:** The module performs singular value decomposition on the term document matrix. A mathematical application that truncates the matrix and reduces the rank.
d) **Word net:** This module integrates the previous modules by relating it with the word net application. It finds out the semantic similarities and provides an increased number of relevant documents.

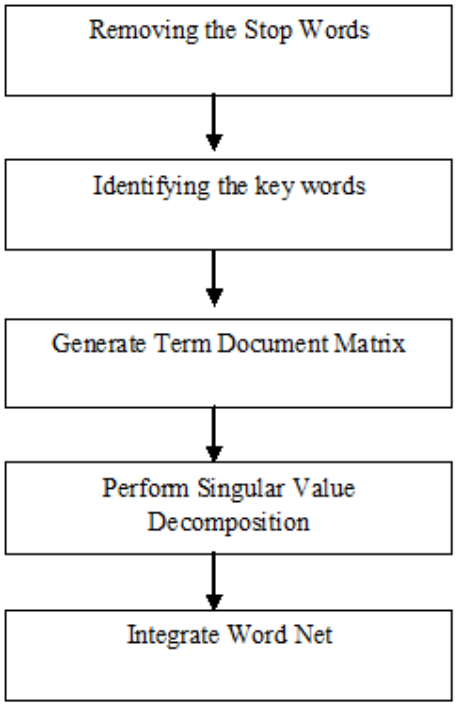

Figure-1. Flow chart for the system.

The decomposition of the systems is described in Figure-1. It clearly specifies all the functions and the responsibilities that are performed by each and every module and how they are combined to form a single system.

A. Stop words recognizer

The stop word recognizer recognizes the stop words and removes them from the documents under considerations. Stop words are also described as the words that are most common in a language. Although there is no single list of stop words, but there are several lists that are widely accepted.

Words like *the, a, an, of, above, even* etc. form the list. These stop words are present in the documents. Each and every word is compared to the list of the stop words; if the word matches the word is removed. After several iterations the stop words in the file are removed.

B. Term document matrix

The term document matrix comprises of the terms and the documents. Terms represent the key words and the documents are the documents under consideration.

Once the stop words are removed, the iterations are over; the system forms a term document matrix with rows defined by the terms and the columns identified by the documents. The words when compared with the list of stop words removes the stop words, if the words do match with any of the words in the stop word list then those words are inserted into another file. The frequency of these keywords is found out and eventually represented in the term document matrix.

The term document matrix also helps to identify a relevant document, as the frequency of the words is represented. Figure-2 is an example for term document matrix.

	D1	D2	D3	D4	D5	D6	D7
binary	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
computer	0.0198	0.0000	0.0198	0.0198	0.0000	0.0000	0.0000
computer system	0.0000	0.1405	0.0000	0.0000	0.1405	0.1405	0.1405
engineering	0.1138	0.0000	0.1138	0.1138	0.0000	0.0000	0.0000
eps	0.1733	0.0000	0.1733	0.1733	0.0000	0.0000	0.0000
generation	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
graph	0.0000	0.0559	0.0000	0.0000	0.0559	0.0559	0.0559
human	0.1336	0.0000	0.1336	0.1336	0.0000	0.0000	0.0000
interface	0.0198	0.0000	0.0198	0.0198	0.0000	0.0000	0.0000
intersection	0.0000	0.0105	0.0000	0.0000	0.0105	0.0105	0.0105
machine	0.0198	0.0000	0.0198	0.0198	0.0000	0.0000	0.0000
management	0.0595	0.0000	0.0595	0.0595	0.0000	0.0000	0.0000
minors	0.0000	0.0454	0.0000	0.0000	0.0454	0.0454	0.0454
opinion	0.0000	0.1405	0.0000	0.0000	0.1405	0.1405	0.1405
ordered	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
random	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
response	0.0000	0.1405	0.0000	0.0000	0.1405	0.1405	0.1405
survey	0.0000	0.1859	0.0000	0.0000	0.1859	0.1859	0.1859
system	0.2871	0.0000	0.2871	0.2871	0.0000	0.0000	0.0000
testing	0.1138	0.0000	0.1138	0.1138	0.0000	0.0000	0.0000
time	0.0000	0.1405	0.0000	0.0000	0.1405	0.1405	0.1405
user	0.0000	0.1405	0.0000	0.0000	0.1405	0.1405	0.1405
user interface	0.0595	0.0000	0.0595	0.0595	0.0000	0.0000	0.0000

Figure-2. Term document matrix.

C. Singular value decomposition

The advantage of LSI is that it is independent of the language. It works with any language and gives the output with the same efficiency. It overcomes the problems caused by synonyms. This language independency is because of the mathematical approach that is used by the process. It uses common algebra techniques to figure out the correlation in the process.

The key step is LSI is decomposing the matrix using a mathematical technique called Singular Value Decomposition. The SVD algorithm preserves as much information as possible about the relative distances between the document vectors, while collapsing them down into a much smaller set of dimensions.
The term matrix document can be defined as *ABC**.

Where
A = *m x m*
B = *m x n*
C = *n x n*
m = *terms,*
n = *documents*

D. WordNet

Word Net is a large database of English words arranged lexically. This arrangement consists of nouns, adjectives, verbs and adverbs grouped distinctly. The mathematical functions involved in LSI removes the noise, the synonyms, the polysemy etc. To increase the bandwidth of the search the system is connected to the Word Net.

The application gives out the semantic similarities of the words. These similarities are then looked for in the documents. Once the process of LSI is completed on these words as well the number of documents retrieved is increased.

5. EVALUATION

As mentioned earlier, the proposed system is evaluated using 2000000 FIRE corpus. The system is tested with the standard set of 15 queries as given in the FIRE evaluation. For each query, precision and recall are calculated. The efficiency of the IR system using both LSI and WordNet are brought out by comparing the IR system that uses only term based indexing.

The precision and Recall are calculated as follows.

Precison = No of correct documents retrieved/Total no of documents retrieved.

Recall = No of correct documents retrieved/ Total no of documents actually relevant to the query.

Figure-3 shows the recall values for both the IR system built using with and without LSI+WordNet for 8 queries Q1, Q2…Q8.

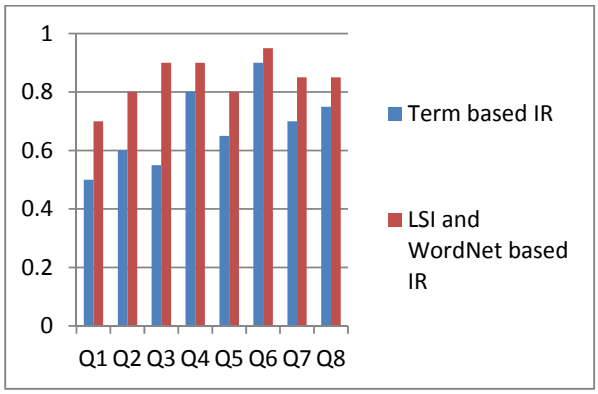

Figure-3. Comparison of term based IR and LSI+WordNet based IR.

It can be observed from the above graph that the proposed model shows better performance than the term based IR. This is merely due to the semantic query-document mapping induced in the proposed model which in turn eliminates the irrelevant documents. Furthermore it increases the precision and recall of the retrieved documents. The term based IR on the other hand was not able to retrieve documents having semantic similarity. For instance, the query could not retrieve the documents indexed with the synonyms and other semantic relations.

6. CONCLUSIONS

Latent Semantic Indexing is a useful technique to implement retrieval of documents on the basis of conceptual meaning and semantic analysis. This is a mathematical approach to indexing and retrieval and it makes sure that the relevancy of the obtained documents is optimized. Not all the documents we get using generic key word search are relevant and this property is often compromised. This isn't the case with Latent Semantic Indexing as the documents that do not contain the key words in the query are also acquired, provided they satisfy the criteria for relevancy. WordNet is excellent software that enabled us to improve upon the system and enhance the search results.

REFERENCES

[1] Aggarwal C. C. and Yu P. H 2001. On effective conceptual indexing and similarity search in text data. Proceedings of the 2001 IEEE International Conference on Data Mining (pp. 3-10). San Jose.

[2] H. Kettani and G.B. Newby. 2010. Instability of Relevance-Ranked Results Using Latent Semantic Indexing for Web Search. System Sciences (HICSS), 2010 43rd Hawaii International Conference on. pp. 1-6.

[3] M.G. Ahsaee, M. Naghibzadeh and S.E.Y. Naieni. 2012. Weighted Semantic Similarity Assessment Using WordNet Computer and Information Science (lCCIS), 2012 International Conference on. 1: 66-71

[4] R. Mihalcea and D. Moldovan. 2000. Semantic indexing using WordNet senses. ACL Workshop on IR & NLP.

[5] Ozcan R., Aslangdogan Y.A. Concept Based Information Access Using Ontologies and Latent Semantic Analysis. Technical Report CSE-2004-8. University of Texas at Arlington.

[6] Lan Du Huidong Jin de Vel, O. Nianjun Liu. 2008. A Latent Semantic Indexing and WordNet based Information Retrieval Model for Digital Forensics. In Proc. of Intelligence and Security Informatics, 2008. ISI 2008, IEEE International Conference, Taipei. pp. 70-75.

[7] B. Bai, J. Weston, D. Grangier, R. Collobert, O. Chapelle and K. Weinberger. 2009. Supervised semantic indexing. In CIKM.