An Indiscriminate Semantic Similarity Search Using Latent Semantic Indexing in Data Mining

Nudrat Fatima*[1], Saba Khalid*[2], Halima Sadia*[3]

[*] *CSE Department Integral University, Lucknow*

Abstract: **In this paper we propose a Semantic model of an information system that provides precise definitions of fundamental concepts like Query, subquery, and coupling.Queries are mapped to this space with documents being retrieved based on similarity Model. In this paper, the performance in document retrieval is investigated and compared with traditional term matching techniques with the help of LSI, CHOLSKEY Transform, LU-Decomposition, Natural language processing for improving searching redundancy with stemming and stop words processing. After that we have to apply for singular value decomposition for gaining the motivation for fetching records and information from document which would be same in the meaning and texture. While users want to search through information based on conceptual content, natural languages have limited the expression of these concepts for individual words contained in user's queries, may not explicitly specify the intended user's concept, which may result in the retrieval of some irrelevant documents.**

Keyword:**Semanticcrawler,LSI,NLP,Cholskey transform, SVM,LU-Decomposition .**

I. INTRODUCTION

Various recent applications such as document summarization,passage retrieval and question answeringrequire LSA seems to be a promising technique in overcoming these natural language problems between terms and documents that are mapped and closely related to each other. Queries are then mapped to this space with documents being retrieved based on similarity Model.

LSA performance in document retrieval is investigated and compared with traditional term matching techniques. They improve the performance on average,but also introduce some instability and thusincreased variance (Levow et al., 2005). Most NLP applications such as information extraction, machine translation, sentiment analysis and question answering, require both syntactic and semantic analysis at various levels. The motivation for a LU decomposition is based on the observation that systems of equations involving triangular coefficient matrices are easier to deal with.

A) Indexing of Documents

Indexing is an essential part of the IR systems for two reasons. First, it optimizes the query performance and improves the respond times considerably by storing terms in an inverted file structure. Basically, it stores the text positions for occurrences of each term. Secondly, a number of processing tasks are carried out during the indexing phase similar to the query processing phase, which further improves the performance.

B) Searching

In this phase, the query terms are searched against the inverted index. All the documents that contain the occurrences of the query terms are retrieved. Depending on the application, the retrieval can be done even for the partially matched documents.

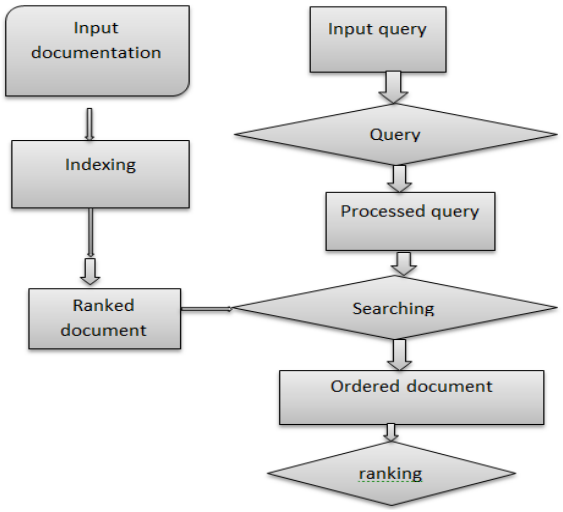

Fig 1: Indexing process

C)Ranking

The documents retrieved in the previous step are given scores according to the matching quality between the query terms and the documents. The documents are sorted according to this score, so that the most relevant documents are presented to the user on top of the retrieval list. Ranking process is highly dependent to the IR model.

semantic analysis at various levels. Traditionally, NLP research has focused on developing algorithms that are either language-specific and/or perform well only on closed-domain text.

II. LITERATURE REVIEW

There is a lot of literature on Web fetching algorithms(Kleinberg 1999; Najork and Wiener 2001; Bharatand Henzinger 1998). Focused fetching was defined andformalized by (Chakrabarti, van den Berg, and Dom 1999),which introduced taxonomic classification or text analysis todetermine document relevance and distillation or link analysisuser queries. A substring search, even when implemented using sophisticated algorithmslike suffix trees or suffix arrays (Manber and Myers 1990), is not adequate forsearching very large text collections. Many different methods of text retrieval havebeen proposed in the literature, including early attempts such as clustering (Salton1971) and the use of signature files (Faloutsos and Christodoulakis 1984). In practice,inversion (Berry and Browne 1999;Wittenet al. 1999) is the only effective techniquefor dealing with very large sets of documents. The method relies on the constructionof a data structure, called an inverted index, which associates lexical items totheir occurrences in the collection of documents.to identify authoritative sources of relevant documents.Shark is one of the earlier focused crawlers, and more sophisticatedvariations exist (Peng, Zhang, and Zuo 2008;Pandey and Olston 2008). Also, extensive follow up workhas compared focused fetching against a variety of otherfetching techniques across a variety of domains (Menczeret al. 2001; Davison 2000; Cho, Garcia-Molina, and Page1998). It has been shown in recent work (Almpanidis,Kotropoulos, and Pitas 2007) that uses a latent semantic indexing(LSI) classifier, which combines link analysis withtext content, that a variant of the simple Shark-search can besurprisingly efficient when compared to more sophisticatedand expensive techniques such as LSI and PageRank (PR).

III. PROBLEM DEFINITION

The problems of interoperability between interrelating computer systems have been well documented. A good classification of the different kinds of interoperability problems can be originate in [Sheth, 98] who classifies the system, syntactic, structural and semantic stages of heterogeneity. the syntactic level refers to different languages and data representations; the structural level comprises different data models and the semantic level refers to the sense of terms using in the interchange. LSI uses a term-document matrix to identify the occurrence of terms within a set of documents, applies term weighting based on term frequencies to reflect the fact that some terms are more important than others in a body of text, and then performs a Singular Value Decomposition (SVD) on the matrix to determine patterns in the relationships between the terms and concepts used in the documents.

IV. SYSTEM MODEL

The user interest model in this work is constructed based on the knowledge of the domain. Knowledge structure of education domain has been considered here. The knowledge representation database, semantic, is organized into a three level hierarchical structure as shown in Figure 1.

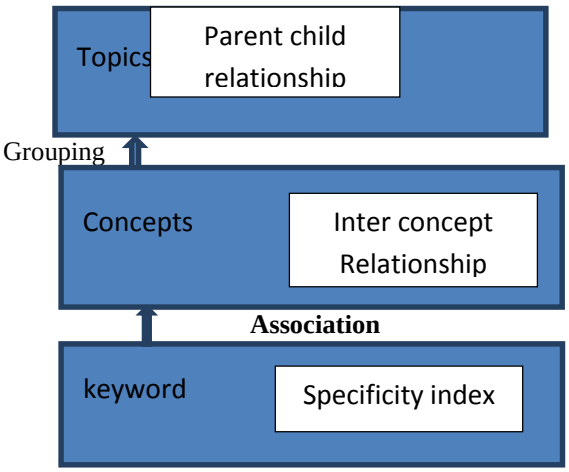

Fig.2 Representation of semantic

There is a need to model the requirement of the user in order to filter the web documents with respect to the need of the user. As stated earlier, this system looks at the requirement of the user in the school level topics. The features related to this domain are as follows.

• In every class there are some predefined syllabi that can be treated as the learning objective for the class. Generally, a student starts with a small subset of the concept space specified in the syllabi. Gradually, the requirement of the student tends

to be saturated to the whole concept space enclosed by the syllabi. This type of modeling is very much similar to the Overlay Model.

• The syllabus for a class represents the learning objective of the student in that class. A portion of the concept space may be of higher interest value compared to others.

V. PROPOSED IMPLEMENTATION

LSI

The goal of the tests was to understandwhether the use LSI to index a gathering of articles and the words "program" and "code" appear together regularly enough, the search algorithm will notice that the two terms are semantically close. A search for "program" will consequently return a set of articles covering that phrase, but also articles that contain just the word "code". LSI does not understand the word distance, but by investigative a sufficient amount of documents, it knows the two terms are related. It then uses that data to provide an extended set of results with better recall than a plain keyword search.

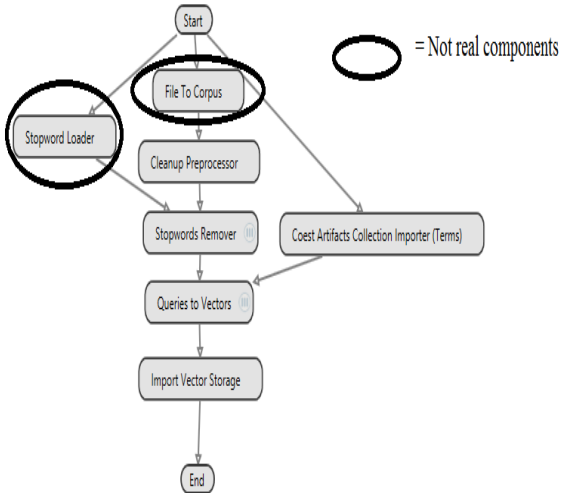

Fig 3: Load Query vectors

1. Load Query Vectors into notepad
2. Load the query vectors into notepad using the Import Vector Storage component.
3. Load LSI Space into notepad
4. Load the LSI Space Matrix into Trace using the Import Matrix component.
5. Multiply Query Vectors Times the LSI Space
6. Using the Vectors * Matrix to paper each vector into the LSI Space. This will result in a set of vectors one for each query, with the same length as the number of columns within the LSI Space (or the rank of the LSI Space).
7. Save the Query Vectors
8. Save the resulting queries using the Export Vector Storage component.
9. Deriving patterns automatically from a corpus
10. Using Singular Value Decomposition to smooth frequency data
11. Generating synonyms that are used to explore word pairs with similar meanings
12. The algorithm requires a search engine with a very large corpus of text, a broad coverage thesaurus of synonyms, and an efficient implementation of Singular Value Decomposition (SVD).

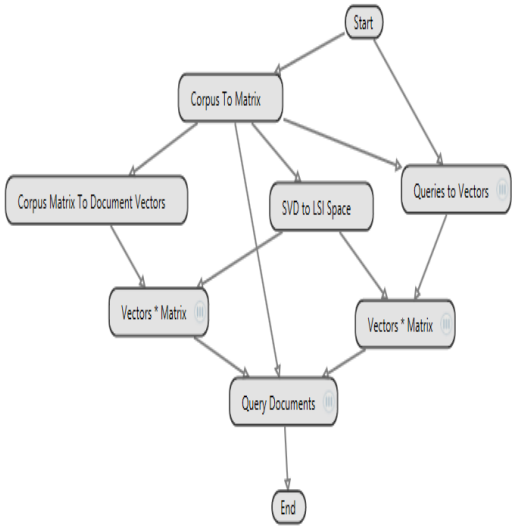

Fig4: SVD with LSI Query Structure

LSI takes as input a set of word and constructs a matrix that can be used to find the relational similarity between any word

and sentence. We proposed a sequential clustering algorithm that scales linearly with the number of patterns, to efficiently cluster a larger number of patterns.

LU Decomposition

If *A* is a square matrix and it can be factored as where *L* is a lower triangular matrix and *U* is an upper triangular matrix, then we say that *A* has an **LU-Decomposition** of *LU*.

If A is a square matrix and it can be reduced to a row-echelon form, U, without interchanging any rows , then A can be factored as where *L* is a lower triangular matrix.

LU decomposition of a matrix is not unique.There are three factorization methods:

Crout Method: diag (U) = 1;

Doolittle Method: diag (L) = 1;

Choleski Method: diag (U) = diag (L) ;

To solve several linear systems with the same A, and A is big, we would like to avoid repeating the steps of Gaussian elimination on A for every different B.

NLP (Natural language processing)

- Reading in the meetings transcripts
- Parsing by speech
- (stopwords, stemming etc are options – benefits less obvious for text reuse)
- Writing a swalign function
- Calling that function and applying it to the file list(s)
- Investigating whether you learn anything from the speeches that produce the highest similarity scores

2 - Test the argument that the Federal Reserve's response to the financial crisis was hindered by a preoccupation with inflation.

For starters, produce a graph of attention to inflation over time (using quarterly meetings).

Do not assume that the only relevant word is 'inflation,' and do not assume that raw word counts are the best measure of attention.

Then use an NLP approach to test whether there was "growing concern" at the Fed about the mortgage crisis and when this occurred. Overlay these results on the inflation graph.

The most obvious approach would be to create a dictionary of terms (unigrams, bigrams) related to the mortgage crisis. But you could also consider an approach that captures the amount of discussion focusing on the mortgage crisis (as opposed to keyword frequencies), or even the 'anxiousness' of the discussion etc.

VI. RESULT

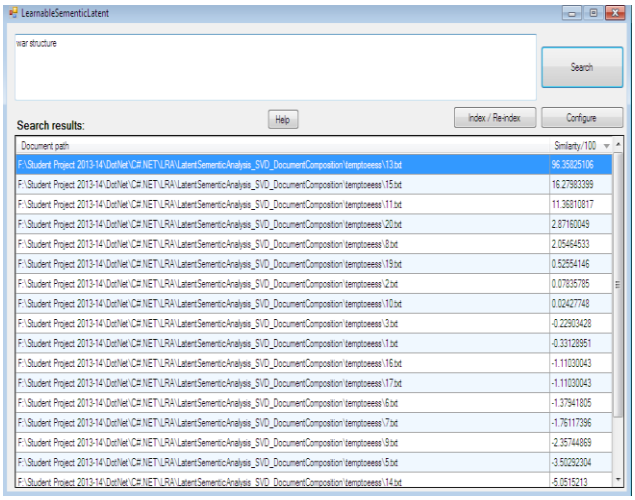

Figure 1: semantic search fetching sentence based similarity

The performance evaluation in the paper is being carried out by using standard semantic of recall and precision For each sentence, interpolated average percentage is computed.
In Figures 2 and 3, we have presented the results on evaluationof interpolated sementic search for each of the content queries in the database. Figure 3 represents the highlightof sentance for each query with respect to SVM, SVD and LUd at , where we have recorded themaximum average percentage in Figure 3.

• Recall A measure of the ability of a system to present all relevant items
recall = number of relevant items retrieved/ number of relevant items in collection

• Precision. A measure of the ability of a system to present only relevant items

precision = number of relevant items retrieved /total number of items retrieved

Table

Query	Similarity %
Similarity	**96.2764**
Non-similarity	**-12.546**

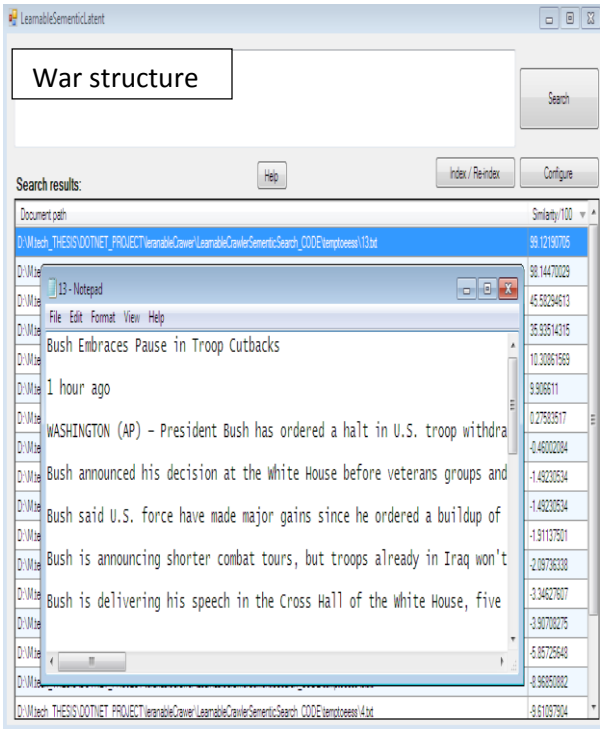

Figure 2: Semantic Crawl data on the basis of sentence

Table

Query	Similarity %
similarity	**98.1764**
Non-similarity	**-4.5896**

Library

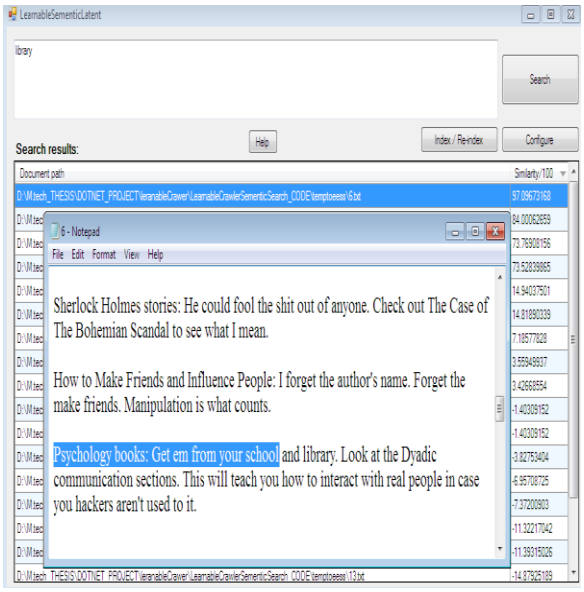

Figure 3: Highlight content in semantic sementic search

The visual analysis of Figure clearly shows how each query has performed for a method with respect to the other. Amongst the many queries, sementic has performed better.

Analysis of Recall & precision

	Precision	Recall
Doc 1	**0.980**	**0.970**
Doc 2	**0.911**	**0.981**
Doc 3	**0.922**	**0.918**
Doc 4	**0.931**	**0.930**

Fig 5: Rank of Document

Discussion

In this discussion, semantic with LSI indexing has been discussed as the best document ranking, which satisfies the literature result. Through the implementation of different indexing performance and classifier available in semantic crawler, it is demonstrated preprocessing and documents indexing are two important stages to advance the mining quality.

VII.CONCLUSION

In this work we have motivated and defined the concept of a specific focused crawler has been implemented. A domain specific crawler is useful for saving time and other resources since it is concerned with a particular domain. Hence this

easily fetch the frequent information from big data and set the data redundancy for doubly focused users for market. It was found that there is a requirement to build a crawler that takes into account the context of the words or phrases being searched for. LSI (Latent semantic Indexing) model is one such promising model in the field of information crawler. LSI uses a mathematical technique known as Singular Value Decomposition. This model has the ability to extract the conceptual content of a body of text by looking for relationships between the information those who want to capture frequently grape for users of the text finally the LSI based crawler. Hence it is clear that the performance of Latent and singular value based crawler is the most efficient and accurate.

Future Scope Therefore, this future work will include automatic semantic mapping, bridging axiom production from machine learning and natural language processing, pattern reuse and consistency testing for merged ontologies.

REFERENCES

[1.]Almpanidis, G.; Kotropoulos, C.; and Pitas, I. 2007. Combiningtext and link analysis for focused fetching-an application for vertical search engines. Information Systems.

[2.]Bharat, K., and Henzinger, M. R. 1998. Improved algorithms fortopic distillation in a hyperlinked environment.In Proceedings ofthe 21st annual international ACM SIGIR conference.

[3.]Chakrabarti, S.; van den Berg, M.; and Dom, B. 1999. Focusedfetching: a new approach to topic-specific web resource discovery.Computer Networking 1623–1640.

[4]Dumais, S., and Chen, H. 2000. Hierarchical classification ofWeb content.In Proceedings of the 23rd annual internationalACM SIGIR, 263.

[5.]Pandey, S., and Olston, C. 2008. Crawl ordering by search impact.In Proceedings of the international conference on Web search andweb data mining, 3–14. ACM.

[6]Ron Bekkerman, Ran El-Yaniv, and Naftali Tishby.2003. Distributional word clusters vs. words for textcategorization.

[7.] Irina Matveeva, Gina-Anne Levow, Ayman Farahat, andChristian Royer. 2005. Generalized latent semanticanalysis for term representation. In Proc. of RANLP.