Volume 4, No. 8, May-June 2013

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Semantic Approach for Query Explication

Shruti Gupta
M.Tech Scholar, Jagan Nath University, Jaipur
shruti.1324@gmail.com

Mahesh Verma
HOD, CSE, JaganNath Gupta Institute of Engineering& Technology
mahesh.verma@jnit.org

Abstract The main focus is to add a new dimension to Internet-Searching and that is to apply semantic aspects towards it. An essential requirement of this work is one has to recognize the difference between what a user might say or do and what she or he actually meant or intended. In more simple words, "the search must be what user wish, not what he/she types". Quering the search engine for any particular topic would retrieve the results from the internet and presented to the web users. Since there are large number of web pages on the internet and thus result obtained are also vast. User gets more than enough web links as a result produced by search engine and wastes their precious time in navigating through unwanted links, searching the needed one. The main reason for this is that the Search Engine do the indexing of the pages on the basis of text entered by user. In order to overcome this shortcoming we need to implement a method that will allow the user to find the relevant words, starting from the few words that they may actually know [5].In other words, we need to focus on the semantic of words entered by user and for this purpose a new approach that is based on some algorithms which considers semantic aspects should be included. One of such technique for the semantic analysis is the Latent Semantic analysis and Probabilistic Latent Semantic Analysis.

Keywords: Information retrieval, Query-Expansion, Latent Semantic Analysis, Probabilistic Latent Semantic Analysis,

I. INTRODUCTION

Many applications that handle information on the internet would be completely inadequate without the support of information retrieval technology The quest for information on a particular topic drives them to search for it, and in the pursuit of their info the terms they supply for queries varies from person to person depending on the knowledge they have[1][13]. One of the most popular and widely used algorithms for extracting documents which are similar to a query document is TF-IDF [8], [6]**.** It measures the similarity between documents by comparing their word-count vectors. The similarity metric weights each word by both its frequency in the query document (Term Frequency) and the logarithm of the reciprocal of its frequency in the whole set of documents (Inverse Document Frequency). But this approach was not successful as the retrieval according to meanings of the words was not possible. It computes document similarity directly in the word-count space, which can be slow for large vocabularies. It assumes that the counts of different words provide independent evidence of similarity. It makes no use of semantic similarities between words. To remedy these drawbacks, numerous models for capturing low dimensional, latent representations have been proposed and successfully applied in the domain of information retrieval. A simple and widely-used method is Latent Semantic Analysis (LSA**)**, which extracts low-dimensional semantic structure using SVD decomposition to get a low-rank approximation of the word-document co-occurrence matrix [5][13]. This allows document retrieval to be based on "semantic" content rather than just on individually weighted words.

II. MODEL OF INFORMATION RETRIEVAL

An information retrieval system is a software programme that stores and manages information on documents, often textual documents but possibly multimedia. There are three basic processes an information retrieval system has to support: the representation of the content of the documents, the representation of the user's information need, and the comparison of the two representations. The processes are visualized in Figure 1. In the Figure, squared boxes represent data and rounded boxes represent processes.

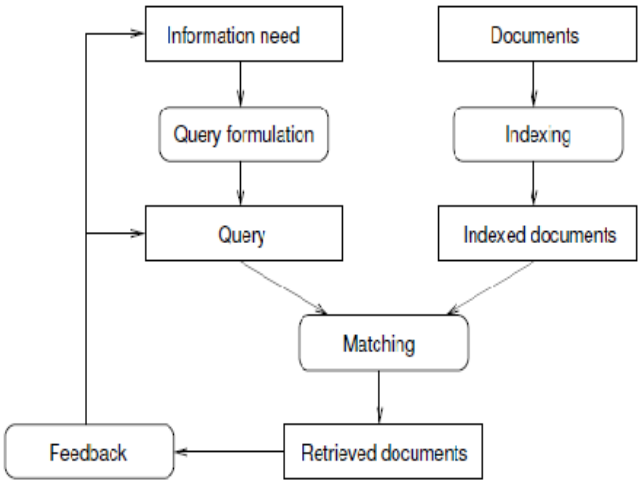

Figure 1 Model of Information Retrieval

III. SEMANTIC APPROACH

Efforts to incorporate semantic information into text processing systems date back nearly half a century. Over the

years, designers have followed various approaches to integrating some degree of semantic processing into their information retrieval systems [3].

a. Auxiliary Structures
b. Local Co-Occurrence Statistics
c. Latent Semantic Indexing

A. Auxiliary Structures:

Controlled vocabularies, or auxiliary structures, such as dictionaries and thesauri, allows broader terms, narrower terms, and related terms to be included into queries [3]. Controlled vocabularies are one way to overcome some other most severe constraints of Boolean free-text keyword queries is multiple words that have similar meanings (synonymy), and words that have more than one meaning (polysemy). These two problems are often the cause of discrepancy in the vocabulary of documents and the users of text retrieval systems.

B. Co-Occurrence Statistics:

Information retrieval systems using this method count the number of times pairs of terms appear together that is co-occur, within a range of terms or sentences (for example, ±5 sentences or ±50 words) within a document. This approach is simple, but it only focuses on a small portion of the semantic information in a collection of text.

C. Latent Semantic Analysis (LSA):

It is a mathematical/statistical method that can be used to decide similarity of meaning of terms and paragraphs by analyzing huge text [3].It does not use any artificial intelligence method or a natural processing method. It tries to explore the meaning of the words and about the topic. It has an added advantage of the semantic structure i.e. detection of the relevant documents on the basis of the queries. Each cell contains the frequency with which the word of its row appears in the passage denoted by its column [4], [12].

D. Steps in LSA:

a. The first step is to represent the text as a matrix in which each row stands for a unique word and each column stands for a text passage or other context. Each cell contains the frequency with which the word of its row appears in the passage denoted by its column.
b. Next, Singular value decomposition (SVD**)** [10], [11] is applied to the matrix.

In SVD, a large term by document matrix is disintegrated into a set of orthogonal matrices and and a diagonal matrix. Queries are represented as pseudo-document vectors formed from weighted combinations of terms and documents. The SVD of matrix A is written as [10].

$$A = U S V^T \quad (1)$$

Where A is t×d term by document matrix is orthogonal matrix, S is a Diagonal Matrix, V is a orthogonal matrix and k is the rank. By changing all but the top *k* rows of S to zero rows, a low rank approximation to A called Ak is obtained

$$A_{k=} \; U_k S_k V_k^T \quad (2)$$

Where*Uk* is the *t*×*k* term-by-concept matrix, S*k* is *k*×*k* concept-by-concept matrix, *Vk* is *k*×*d* concept-by-document matrix. The rank of *A* has been lowered from *r* to *k*. This low rank approximation removes redundancy from original data and allows us to uncover latent semantics is *relations* among terms as well as documents. Queries are formed into pseudo-documents that specify the location of the query in the reduced term-document space.

$$qc = qTUkSk\text{-}1 \quad (3)$$

Or

$q_c = q'^*M$, where M is the product of U_k and S^{-1}_k

IV. CORPORA AND EXPERIMENT

A. Dataset and Queries:

The experiments are carried out on 6 corpora taken from different fields. The Documents collected are the Wikipedia entries and total of 14 queries are used. The experiments are carried out on the Mat lab software which is primarily used for the complex numerical calculations.

Table1: Queries

Queries ID	Terms
Q1	Ball
Q2	Cricket
Q3	Bat
Q4	Coffee store
Q5	California
Q6	Company
Q7	Encyclopedia
Q8	Irvine
Q9	Run
Q10	Species
Q11	Starbucks
Q12	Store
Q13	University
Q14	Users

Table2: Documents

Document ID	Documents
Doc1	Wikipedia Cricket Bat
Doc2	Wikipedia Coffee
Doc3	Wikipedia Bat
Doc4	Wikipedia Starbucks
Doc5	Wikipedia Homepage Starbucks
Doc6	Wikipedia Irvine

B. Work Done:

The steps involved in the experiment are as follows:-

a. Creating a TFID matrix for the document
b. Plotting the matrix obtained onto a x-y axis
c. Calculating the Singular Value Decomposition
d. Plotting the new axis obtained after SVD
e. Applying the dimensionality reduction on the matrix
f. Show what SVD has Captured
g. Executing Query and Result of Query

Table3: Tf-IDf Matrix

Terms	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Df
Ball	6	0	0	0	0	0	1
Cricket	10	0	0	0	0	0	1
Bat	24	0	0	0	0	0	2
Coffee	0	0	0	6	0	0	1
California	0	5	0	0	0	0	2
Company	0	0	0	0	0	8	1
Encyclopedia	1	1	1	0	0	1	4
Irvine	0	0	0	6	0	0	1
Run	5	0	0	0	0	0	1
Species	0	3	2	0	0	0	2
Starbucks	0	0	0	0	4	14	2
Stores	0	0	0	0	0	7	1
university	0	0	0	6	0	0	1
Users	0	0	0	5	0	0	1

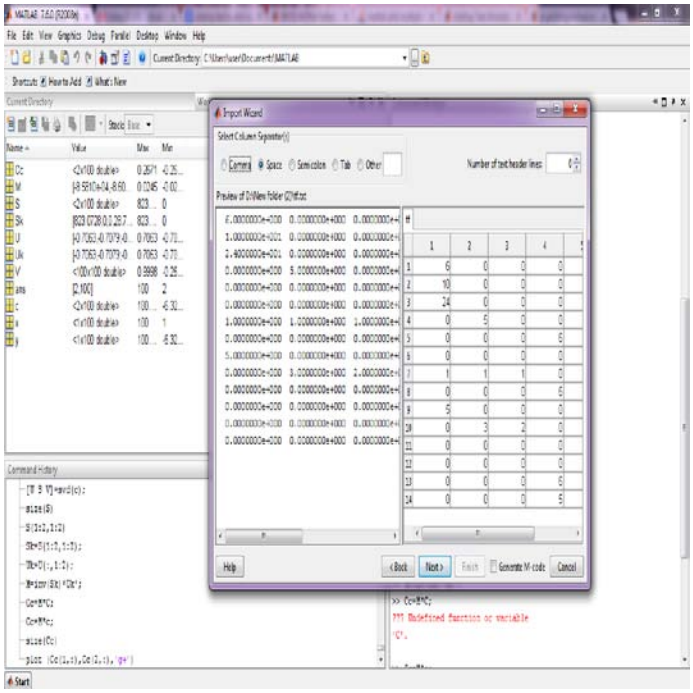

Figure2: Slideshot of Matlab

Then the SVD is performed on this particular matrix the result obtained after the matrix is:-

Table 4: Matrix of M after SVD

```
>> M = inv(Sk)*Uk'

M =

  Columns 1 through 10

   -0.0248   -0.0299   -0.0257    0.0000   -0.0006   -0.0005   -0.0018    0.0000   -0.0230   -0.0017
   -0.0000   -0.0000    0.0000   -0.0280    0.0000    0.0000    0.0000   -0.0280   -0.0000   -0.0000
    0.0006    0.0008    0.0005    0.0000   -0.0263   -0.0305   -0.0021    0.0000    0.0006   -0.0024

  Columns 11 through 14

   -0.0004   -0.0005    0.0000    0.0000
   -0.0000    0.0000   -0.0280   -0.0259
   -0.0241   -0.0290    0.0000    0.0000
```

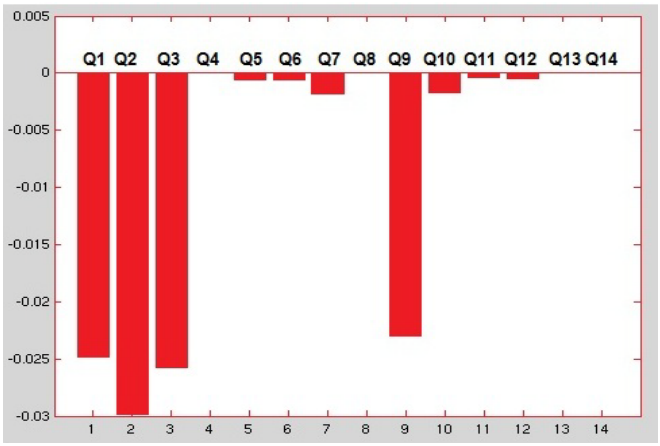

Figure 3 Graphical Representations for 1st Row for M matrix

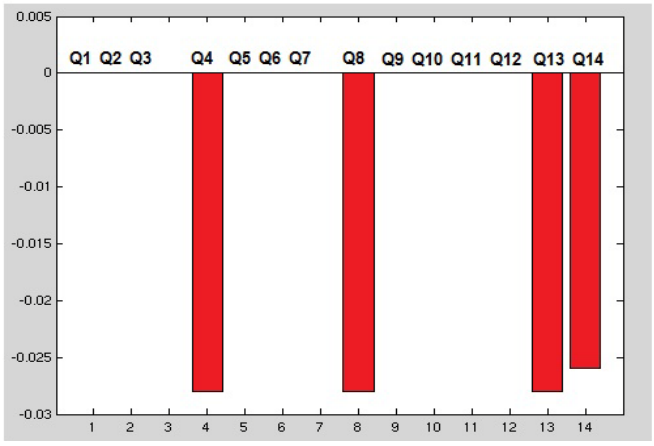

Figure4 Graphical Representations for 1st Row for M matrix

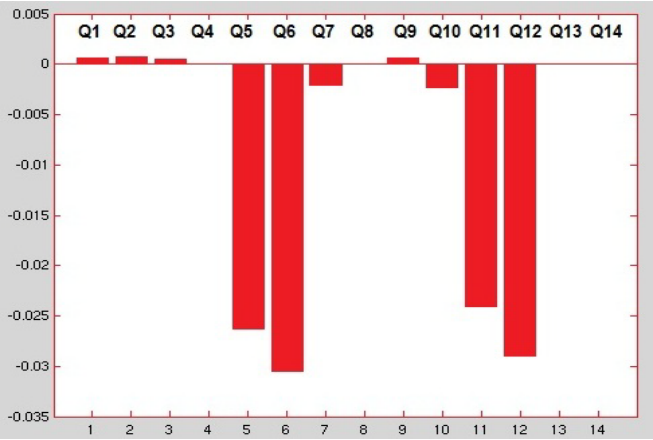

Figure 5 Graphical Representations for 1st Row for M matrix

IV. RESULTS

The query for the "Coffee Store" gave the results

q_c =M*q'= -0.0023
0.0000
-0.1166

Now finding the correlation between query and document.

Sim= (q.d)/|q|.|d|

The Similarity between the query and Document is measured by the Similarity measure formula for which is Shown above

•Wiki Star bucks (0.1166) • Wiki: star bucks(0.1166) • Wiki: Cricket Bat (-0.0004) • Wiki: coffee (0.1166)
• Wiki: bat (0.0091) • Wiki: Irvine (0.0000)

Doc5=Doc4=Doc2>Doc3>Doc6

The results of all the queries are calculated likewise shown in the above query just the query vector changes and according to it the similarity is calculated.

V. CONCLUSION

LSI has proven to be an optimal solution for a wide range of conceptual matching problems one consequence of lsi processing is the establishment of associations between terms that occur in similar contexts. as a result, queries against a set of documents that have undergone lsi will return results that are conceptually similar in meaning to the query even if they don't share a specific word or words with the query.

VI. REFERENCES

[1]. R.B. Yates, B R.Neto, Modern Information Retrieval Pearson Education, 1999.

[2]. N.J. Belkin, W.B.Croft, Information Filtering and Information Retrieval: Two sides of the same coin. Communications of the ACM, 35, 1992, 29–38.

[3]. Price, R., and Zukas, A., Application of Latent Semantic Indexing to Processing of Noisy Text, Intelligence and Security Informatics, Lecture Notes in Computer Science, Volume 3495, Springer Publishing, 2005, pp. 602-603.

[4]. S. C. Deerester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis Journal of the American Society of Information Science, 41(6):391–407, 1990.

[5]. A. Kontostathis, W.M. Pottenger, A Mathematical View of Latent Semantic Indexing: Tracing Term Co-Occurrences, Lehigh University Technical Report, LU- CSE-02-006, 2002

[6]. G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5):513–523, 1988.

[7]. Salton, G. and M. McGill (1983). Introduction to Modern Information Retrieval. McGraw-Hill.

[8]. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022

[9]. Salton Developments in automatic text retrieval. Science, 253, 1991

[10]. Jing Cao, Jun Zhang, Clustered SVD Strategies in Latent Semantic Indexing Information Processing and Management, 2005 Article in Press.

[11]. K. April, William M. Pottenger, "A Framework for Understanding Latent Semantic Indexing Performance" Journal of Information Processing and Management, 2005 Article in Press

[12]. Ch. Aswani Kumar, Ankush Gupta, Mahmooda Batool and Shagun Trehan "Latent Semantic Indexing-Based Intelligent Information Retrieval System for Digital Libraries" Journal of Computing and Information Technology - CIT1 4, 2006, 3, 191–196 doi:10.2498/cit.2006.03.02 191

[13]. Shruti Gupta ,"Query Explication by semantic approach"Internationla Journal of Science and Research, 2013.