International Journal of Engineering Inventions
ISSN: 2278-7461, www.ijeijournal.com
Volume 1, Issue 4 (September 2012) PP: 01-05

A Literature Survey on Latent Semantic Indexing

Ashwini Deshmukh[1], Gayatri Hegde[2]
[1,2]*Computer Engineering Department,Mumbai University,New Panvel,India.*

Abstract*—Working of web engine is to store and retrieve web pages. One of the various methods such as crawling, indexing is used. In this paper we will be discussing about latent semantic indexing which uses indexing technique. When a user enters a query into a search engine the engine examines its index and provides a listing of best-matching web pages according to its criteria, usually with a short summary containing the document's title and sometimes parts of the text. The index is built from the information stored with the data and the method by which the information is indexed. In this study, we perform a systematic introduction and presentation of different information retrieval techniques and its comparison with latent semantic indexing (LSI). A comparison table has been made to give a better look of different models. Few limitations of LSI are discussed.

***Keywords* — Data mining, Latent Semantic Indexing, retrieval techniques, Singular Value Decomposition,**

I. INTRODUCTION

Latent semantic indexing is a retrieval technique that indexes and uses a mathematical technique called (SVD)Singular Value Decomposition, which identifies the pattern in an unstructured collection of text and find relationship between patterns. Latent semantic indexing (LSI) has emerged as a competitive text retrieval technique. LSI is a variant of the vector space model in which a low-rank approximation to the vector space representation of the database is computed [1]. There are usually many ways to express a given concept, so the literal terms in a user’s query may not match those of a relevant document. In addition, most words have multiple meanings, so terms in a user’s query will literally match terms in documents that are not of interest to the user [2]. Textual documents have many similar words such documents can be classified depending upon similarities in terms using various techniques. The most classic technique is Vector Space Model (VSM) and SMART system, but due to few disadvantages of this methods improved techniques like LSI & PLSI is used for semantic search of documents. The first session of the paper gives an overview about different techniques for information retrieval and its comparison with LSI. Second session discuss the various applications of LSI.

II. LATENT SEMANTIC INDEXING (LSI)

Latent Semantic Indexing is an information retrieval technique that indexes and uses mathematical technique SVD, which identifies the pattern in an unstructured collection of text and finds relationship between them. A well known method for improving the quality of similarity search in text is called LSI in which the data is transformed into a new concept space [6]. This concept space depends upon the document collection in question, since different collections would have different sets of concepts. LSI is a technique which tries to capture this hidden structure using techniques from linear algebra. The idea in LSI is to project the data into a small subspace of the original data such that the noise effects of synonymy and polysemy are removed. The advantageous effects of the conceptual representation extend to problems well beyond the text domain LSI has emerged as a competitive text retrieval technique. LSI is a variant of the vector space model in which a low-rank approximation to the vector space representation of the database is computed.

2.1 Algorithm for LSI

To perform Latent Semantic Indexing on a group of documents, following steps are performed:
Firstly, convert each document in your index into a vector of word occurrences. The number of dimensions your vector exists in is equal to the number of unique words in the entire document set. Most document vectors will have large empty patches, some will be quite full. Next, scale each vector so that every term reflects the frequency of its occurrence in context. Next, combine these column vectors into a large term-document matrix. Rows represent terms, columns represent documents. Perform SVD on the term-document matrix. This will result in three matrices commonly called U, S and V. S is of particular interest, it is a diagonal matrix of singular values for document system.
Set all but the *k* highest singular values to 0. *k* is a parameter that needs to be tuned based on your space. Very low values of *k* are very lossy, and net poor results. But very high values of *k* do not change the results much from simple vector search. This makes a new matrix, S'. Recombine the terms to form the original matrix (i.e., U * S' * V(t) = M' where (t) signifies transpose). Break this reduced rank term-document matrix back into column vectors. Associate these with their corresponding documents. Finally this results into Latent Semantic Index. The first paragraph under each heading or subheading should be flush left, and subsequent paragraphs should have

III. INFORMATION RETRIEVAL TECHNIQUES

There exist various information retrieval techniques.
1.1 SMART System

In the Smart system, the vector-processing model of retrieval is used to transform both the available information requests as well as the stored documents into vectors of the form:

$$D_i = (w_{i1}; w_{i2};:::; w_{it}) \ (1)$$

Where D_i represents a document (or query) text and w_i is the weight of term T_k in document D_i. A weight zero is used for terms that are absent from a particular document, and positive weights characterize terms actually assigned[4].The SMART system is a fully-automatic document retrieval system, capable of processing on a 7094 computer search requests, documents available in English, and of retrieving those documents most nearly similar to the corresponding queries. The machine programs, consisting of approximately 150,000 program steps, can be used not only for language analysis and retrieval, but also for the evaluation of search effectiveness by processing each search request in several different ways while comparing the results obtained in each case [5]. Steps for information retrieval SMART systemare as shown in Fig 1: [11]

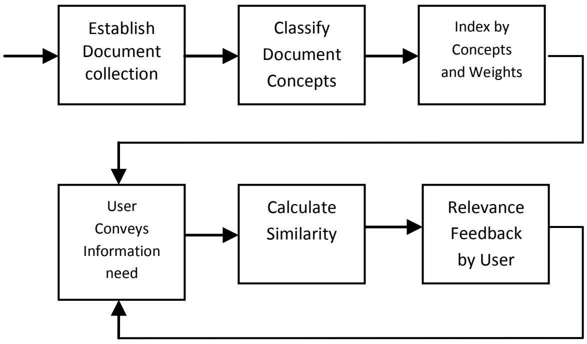

Establish Document collection
Classify Document Concepts
Index by Concepts and Weights
User Conveys Information need
Calculate Similarity
Relevance Feedback by User

Figure1. Information Retrieval SMART System[11]

3.2 Vector Space Model (VSM)

The representation of a set of documents as vectors in a common vector space is known as the vector space model and is fundamental to a host of information. Vector Space Model's retrieval operations ranges from scoring documents on a query, document classification and document clustering.The vector model is a mathematical-based model that represents terms, documents and queries by vectors and provides a ranking[6].VSM can be divided in three steps:

- Document indexing where content bearing words are extracted
- Weighting of indexed terms to enhance retrieval of document
- Ranks the document with respect to the query according to the similarity measure [14] Vector space model is an information retrieval technique in which , a word is represented as a vector. Each dimension corresponds to a contextual pattern. The similarity between two words is calculated by the cosine of the angle between their vectors [7].

3 3. Concept Indexing (CI)

In term-matching method similarity between query and the document is tested lexically [7]. Polysemy (words having multiple meaning) and synonymy (multiple words having the same meaning) are two fundamental problems in efficient information retrieval. Here we compare two techniques for conceptual indexing based on projection of vectors of documents (in means of least squares) on lower-dimensional vector space Latent semantic indexing (LSI)Concept indexing (CI)Indexing using concept decomposition(CD) instead of SVD like in LSI Concept decomposition was introduced in 2001:

First step: clustering of documents in term-document matrix *A* on *k* groups

Clustering algorithms: Spherical k-means algorithm is a variant of k-means algorithm which uses the fact that vectors of documents are of the unit norm Concept matrix is matrix whose columns are centroids of groups c_j– centroid of *j*-th group.

Second step: calculating the concept decomposition

Concept decomposition D_k of term-document matrix *A* is least squares approximation of *A* on the space of concept vectors:

$$D_K= C_K Z \ (2)$$

Where *Z* is solution of the least squares problem given as :

$$Z = (C_k^T C_k)^{-1} C_k^T A \ (3)$$

Rows of C_k = terms

Columns of *Z* = documents

3.4 . Probabilistic Latent Semantic Analysis (PLSA)

It is a statistical technique for the analysis of two-modes and co-occurrence data. PLSA evolved from latent semantic analysis, adding a sounder probabilistic model. Compared to standard latent semantic analysis which stems from linear algebra and downsizes the occurrence tables (usually via a singular value decomposition), probabilistic latent semantic analysis is based on a mixture decomposition derived from a latent class model. This results in a more principled approach which has a solid foundation in statistics. Probabilistic Latent Semantic Analysis (PLSA) is one of the most popular statistical techniques for the analysis of two-model and co-occurrence data [8]. Considering observations in the form of co-occurrences (w,d) of words and documents, PLSA models the probability of each co-occurrence as a mixture of conditionally independent multinomial distributions:

$$P(w,d) = \sum_{c} P(c)P(d|c)P(w|c) = P(d)\sum_{c} P(c|d)P(w|c) \qquad (4)$$

The first formulation is the *symmetric* formulation, where 'w' and 'd' are both generated from the latent class 'c' in similar ways (using the conditional probabilities P(d|c) and P(w|c)), whereas the second formulation is the asymmetric formulation, where, for each document d, a latent class is chosen conditionally to the document according to P(c|d), and a word is then generated from that class according to P(w|c).The starting point for Probabilistic Latent Semantic Analysis is a statistical model which has been called aspect model [13] The aspect model is a latent variable model for co-occurrence data which associates an unobserved class variable z 2 Z = fz1; : : zKg with each observation. A joint probability model over D W is as shown is Fig 2.

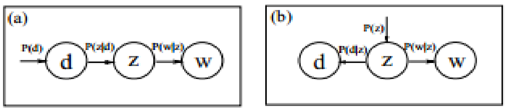


Figure 2: Graphical model representation of the aspect model in the asymmetric (a) and symmetric (b) parameterization

IV. COMPARING VARIOUS TECHNIQUES

While composing a document, different expressions are used to express same or somewhat different data and the use of synonyms or acronyms is very common. vector space model regards each term as a separate dimension axis .various search has been done which proves that VSM creates a hyper dimension space vector of which 99% of term document matrix are empty [3]. Limitations of VSM like long documents are poorly represented because they have poor similarity values, Search keywords must precisely match document terms; word substrings might result in a "false positive match" and more makes it difficult to used in some cases ,hence a more improved method for semantic search and dimension reduction LSI is introduced. SMART system prepossess the document by tokenizing the text into words ,removing common words that appear on its stop-list and performing stemming on the remaining words to derive a set of terms .it is a UNIX based retrieval system hence cataloging and retrieval of software component is difficult. LSI is superior in performance that SMART in most of cases[9].

Table1: Document Table 1

D1	Survey of textmining: clustering, classification, and retrieval
D2	Automatic text processing: the transformation analysis and retrieval of information by computer
D3	Elementary linearalgebra: A matrix approach
D4	Matrixalgebra& its applications statistics and econometrics
D5	Effective databases for text&document management
D6	Matrices, vectorspaces, and informationretrieval
D7	Matrixanalysis and appliedlinearalgebra
D8	Topological vectorspaces and algebras

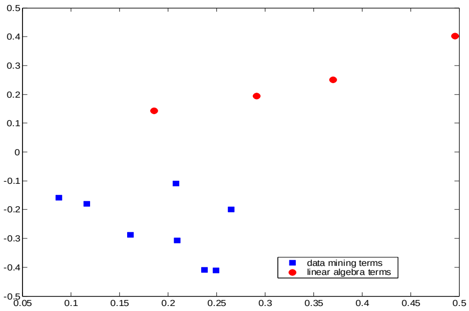
Table2: Document Table 2

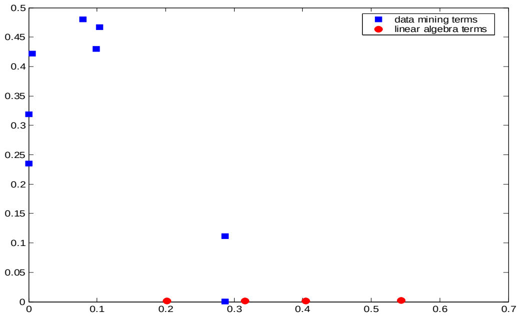
D9	Informationretrieval: data structures &algorithms
D10	Vectorspaces and algebras for chemistry and physics
D11	Classification, clustering and dataanalysis
D12	Clustering of large data sets
D13	Clustering algorithms
D14	Document warehousing and textmining: techniques for improving business operations, marketing and sales
D15	Datamining and knowledge discovery

Table 3: Terms

Document	Linear Algebra Term	Neutral term
Text	Linear	Analysis
Mining	Algebra	Application
Clustering	Matrix	Algorithm

Classification	Vector	
Retrieval	Space	
Information		
Document		
Data		

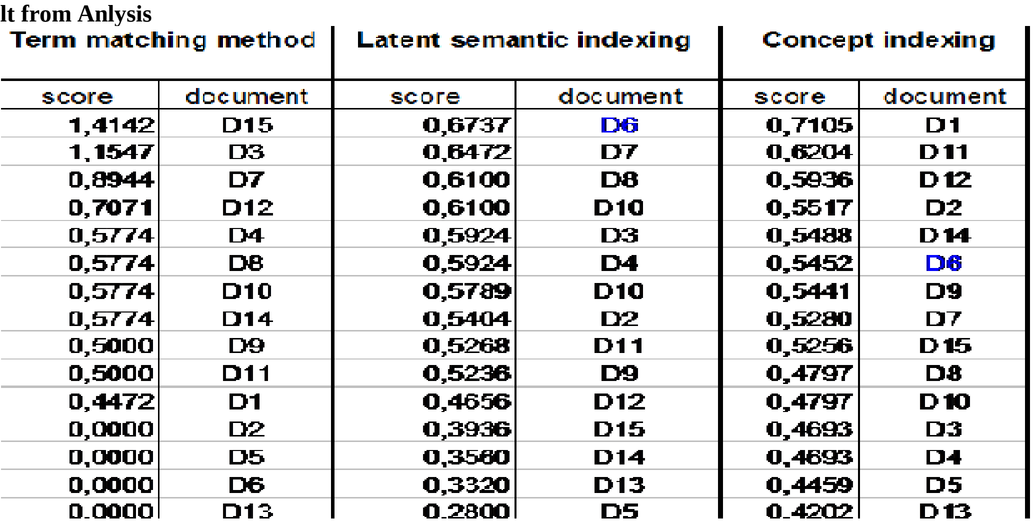




Projection of terms by SVD

Projection of terms by CD

4.1 Result from Anlysis



Term matching method		Latent semantic indexing		Concept indexing	
score	document	score	document	score	document
1,4142	D15	0,6737	D6	0,7105	D1
1,1547	D3	0,6472	D7	0,6204	D11
0,8944	D7	0,6100	D8	0,5936	D12
0,7071	D12	0,6100	D10	0,5517	D2
0,5774	D4	0,5924	D3	0,5488	D14
0,5774	D8	0,5924	D4	0,5452	D6
0,5774	D10	0,5789	D10	0,5441	D9
0,5774	D14	0,5404	D2	0,5280	D7
0,5000	D9	0,5268	D11	0,5256	D15
0,5000	D11	0,5236	D9	0,4797	D8
0,4472	D1	0,4656	D12	0,4797	D10
0,0000	D2	0,3936	D15	0,4693	D3
0,0000	D5	0,3560	D14	0,4693	D4
0,0000	D6	0,3320	D13	0,4459	D5
0.0000	D13	0.2800	D5	0.4202	D13

V. APPLICATIONS OF LSI

LSI is used in various application .as web technology is growing new web search engines are developed to retrieve as accurate data as possible .Few areas where LSI is used is listed as follows :-

LSI is used to clustering algorithm in medical documents as such documents as it includes many acronyms of clinical data [3].Used as retrieval method for analysing broken web links. The best results are obtained applying KLD when the cache page is available, otherwise a co-occurrence method. The use of Latent Semantic Analysis has been prevalent in the study of human memory, especially in areas of free recall and memory search it was found ,a positive correlation between the semantic similarity of two words (as measured by LSA) and the probability that the words would be recalled one after another in free recall tasks. To Identify Similar Pages in Web Applications This approach is based on a process that first computes the dissimilarity between Web pages using latent semantic indexing, and then groups similar pages using clustering algorithms.[12] To compare content of audio we use LSI. The latent semantic indexing (LSI) is applied on audio clips-feature vectors matrix mapping the clips content into low dimensional latent semantic space. The clips are compared using document-document comparison measure based in LSI. The similarity based on LSI is compared with the results obtained by using the standard vector space model [15]. LSI to overcome semantic problem on image retrieval based on automatic image annotation. Statistical machine translation used to automatically annotates the image. This approach considers image annotation as the translation of image regions to words [15]. A bug triage system is used for validation and allocation of bug reports to the most appropriate developers. Automatic bug triage system may reduce the software maintenance time and improve its quality by correct and timely assignment of new bug reports to the appropriate developers [16].

VI. LIMITATIONS OF LSI

Difficult to find out similarities between terms.

2. A bag-of-word lead to unstructured information.
3. A compound term is treated as 2 terms.
4. Ambiguous terms create noise in the vector space
5. There's no way to define the optimal dimensionality of the vector space
6. There's a time complexity for SVD in dynamic collections.

Table 4. Comparison between various techniques

Parameters	VSM	SMART	CI	LSI	PLSI
Basic working	Represents terms, documents by vectors.	Vector space model	Uses concept Decomposition (CD)	Uses Singular Value Decomposition (SVD)	Improved LSI with Aspect Model
Size of the matrix	Hyper dimension space is created	Hyper dimension space is created	Smaller matrix for presenting doc	Smaller matrix for presenting doc	Small
Performance	Good with small documents	Good with small documents	good	Good	Very good

VII. CONCLUSION

The representation of documents by LSI is economical. It is basically an extension of vector space model .It tries to overcome problem of lexical matching by conceptual matching. Thus by this study we can conclude that LSI is a very effective method of retrieving information from even for large documents, it has few limitations, detailed solution for those is out of paper's study. Paper mainly focuses on various methods of information retrieval and their comparison .Few disadvantages of LSI that can be overcome by PLSI. Detail description of PLSI has not been included will be done as a future study for this paper.

REFERENCES

1. Jing Gao, Jun Zhang Clustered SVD strategies in latent semantic indexing USA Oct 2004

2. Scott Deerwester Graduate Library School Susan T. Dumais George W. Furnas Thomas K. Landauer Indexing by Latent Semantic Analysis

3. Choonghyun Han, MA Jinwook Choi Effect of Latent Semantic Indexing for Clustering Clinical Documents., MD, PhD 978-0-7695-4147-1/10 $26.00 © 2010 IEEE

4. Gerard Salton and Chris Buckley. Term-weighting approaches in automatic text retrieval formationProcessing and management, 24(5):513{523, 1988}.

5. G. Salton The SMART System — Retrieval Results and Future Plans

6. Silva. I. R.; Souza, J. N; Santos K.S. Dependence among terms in vector space model"2004. IDEAS '04. Proceedings. International

7. Jasminka Dobša Comparison of information retrieval techniques Latent semantic indexing (LSI) and Concept indexing (CI) Varaždin

8. Chuntao Hong; Wenguang Chen; Weimin Zheng; Jiulong Shan; Yurong Chen; Yimin Zhang Parallelization and Characterization of Probabilistic Latent Semantic Analysis Parallel Processing, 2008. ICPP '08. 37th International Conference

9. Reea Moldovan, Raduioan Bot, Gertwanka Latent Semantic Indexing for Patent Documents.

10. Diana Inkpen Information Retrieval on the Internet.

11. Eric Thul The SMART Retrieval System Paper Presentation & Demonstration Session 2005.10.13

12. Stockholm Thomas Hofmann Probabilistic Latent Semantic Analysis, UAI'99.

13. John Blitzer Amir Globerson Fernando Pereira Distributed Latent Variable Models of Lexical Co-occurrences.

14. Yan JunhuDependable, Autonomic and Secure Computing, 2009. DASC '09. Eighth IEEE International Conference "A Kind of Improved Vector Space Model".

15. Biatov K; Khehler J; Schneider D, Semanting computing Audio clip content comparison using latent semantic indexing 2009 .ICSE'09. IEEE International Conference

16. Ahsan, S.N.; Ferzund, J.; Wotawa, F Automatic Software Bug Triage system (BTS) based on latent semantic indexing.Software Engineering Advances, 2009. ICSEA '09. Fourth International Conference