J. of Comp. and I.T. Vol. 6(2), 15-22 (2015).

Vector Space Modelling and Latent Semantic Indexing for textual information retrieval

MUZAFAR RASOOL BHAT

Department of Computer Science, Islamic University of Science and Technology, Awantipora, Sriagar (India)
muzafar.rasool@islamicuniversity.edu.in

(Acceptance Date 24th December, 2012)

Abstract

Vector space model of a collection eases process of information retrieval. This process can be further simplified by decomposing term document matrix using singular value decomposition. Use of appropriate number of singular values produced by singular value decomposition can improve the relevance of retrieved results. This paper compares the results produced by two main algebraic model based algorithms namely VSM and LSI using metrics based on Precision and Recall tests.Commonly used collections like CACM, and CISI is used in this research.

Key words : information retrieval, SVD, VSM, LSI.

I. Introduction

In vector space modelling, documents and queries are represented as vectors. Measure of similarity of a query vector and document vector is represented as a scalar value. Ranked retrieval is performed using common matrix operations on document vector and query vector. Common models include Vector Space model, Generalized Vector Space Model and Latent Sematic Indexing etc.

Informally, query retrieval for corpus of text documents can be described as the task of searching this corpus for specific instances of text. Using set theoretic notations, this Process can beexplained as follows:

Suppose we have a set (collection) of documents D, with the user entering a query $q = w1, w2, ., wn,$ a sequence of words $wi$for i =1, 2,…n. We wish to return a subset D^* of D such that for each $d \in D^*$, we maximize the following probability:

$$P(d \mid q, D) \quad (1)$$

(Berger & Lafferty, 1999).

Information retrieval systems (Ho& Funakoshi, 1998) can be formulated as a quadruple $\delta = (J, D, Q, \alpha)$, where $J = \{t1, t2, \ldots, t_M\}$ is a

set of indexing terms; D = {d1, d2, d3, …, d^N} is a set of documents, Q is the set of queries where Q_K is subset of J; and $\alpha : Q \times D \rightarrow R^+$ is a ranking function that evaluates the relevance between a query and a document. Given a query $q \in Q$, for any document $d_{j1}, d_{j2} \in D$, if $\alpha(q, d_{j1}) > \alpha(q, d_{j2})$ then d_{j1} is considered more relevant to query q than d_{j2}. In a general form, a document can be treated as a set of index term-weight pairs $d_j = (t_{j1}, w_{j1}; t_{j2}, w_{j2}; \ldots t_{jn}, w_{jn})$, where $t_{jk} \in J$ and $w_{jk} \in [0, 1]$ reflects the relative importance of index term t_{jk} in document d_j. A query $q \in J$ and $w_{qk} \in [0, 1]$. The information retrieval task is to yield a subset $A = \{d_{j1}, d_{j2}, \ldots, d_{jm}\}$ of D to the query q with a ranking order of $\alpha(q,d_{jk})$[1].

Numerous approaches have been suggested to this problem. Few of them include algorithms based on set theoretic operations, algebraic and probabilistic model based algorithms. The focus of my study is Evaluation of performance of algorithms based on algebraic models. Evaluation measures Precision and Recall have been used in this study.

Rest of the paper is organized as follows. Section II formally reviews the available literature on algorithms based on algebraic models preceded by formal introduction of the problem in section I. A detailed description of VSM and LSI algorithms is given in section III. Section IV deal with the experimental evaluations. Discussion on results obtained from section IV starts in section V. Section VI concludes, discussing open issues that need further research.

II. Literature Review :

Standardized evaluation of IR began as early as 1992 with the initiation of the annual text retrieval conference (TREC) sponsored by Defence Advanced Research Projects agency (DARPA) and National Institute of Standards and Technology[2]. TREC participants Index a large text collection and are provided search statements and relevance judgments in order to judge the success of their approaches. In this paper, algebraic model based algorithms VSM and LSI are evaluated on CACM and CISI datasets. These algorithms vary in terms of their complexity in implementation and efficiency of retrieval. Retrieval of an algorithm is measured in terms of Precision and Recall.

For detailed overview as well as understanding supporting mathematical theory of algebraic model based algorithms, one can refer to M. W. Berry *et al*.(1995) and M. W. Berry *et al*.(1999) where author has in detail explained use of linear algebra for information retrieval and its implementation using Matrices and Vector Spaces[3,4]. Wen Zhang et. al studies TFIDF, LSI and multiword algorithms for text classification. Author while describing text classification, explains information retrieval as a major part of text classification. Author analyses TFIDF, LSI and multiword based on two kinds of properties of indexing terms i.e statistical and semantic property. Author concludes that TFIDF has better statistical quality[5].

Existing methods for text-retrieval tasks can be primarily divided into two categories i) keyword oriented and matrix oriented categories. Key word oriented category manipulates key words directly using certain data structures and retrieval algorithms. However matrix oriented methods change keyword representation of documents into a

term-by-document matrix and few decomposition techniques like Q.R factorization and SVD for improving resulting term-by-document matrix of a given collection of documents. Matrix methods generally show better performance than literal matching as claimed by 3,4. 4 Further illustrates representation of a document using vectors besides comparing matrix methods in text-based information retrieval using VSM.

Michael W. Berry *et.al*. in detail explains matrix formation from a given text collection as well as use Vector Space. Author further describes the process of retrieving information using VSM. April Kontostathis in his work on exploring the essential dimensions Latent Semantic Indexing (LSI) starts his work with detailed explanation of VSM. Mechanism of ranking documents with respect to their relevance with a given query is also clarified by the author[6]. [6] Further explains variety of well-known term waiting functions. Survey on information retrieval by Ed Greengrass explains in detail Information Retrieval (IR), use of VSM and state of the art, both research and commercial, in this field besides explaining probabilistic methods of analysing and retrieving documents[7].

In a typical IR scenario, while users formulate queries of specific words, they are generally interested in the concepts or topics implied by these keywords. They generally expect that documents and queries could be matched using higher level features than words. For this purpose Deerwester *et al*. 1990 has proposed latent semantic analysis to convert high dimensionality word-space representation of a document to a low dimensionality vectors of topics[8]. [8] Discussed initially the algebraic foundation of LSI. Work carried out by [8] was further discussed by Berry, *et. al*. in[3,4]. Available Literature on LSI describes SVD, a decomposing process that after finding Eigen values and Eigen vectors of a given term-document matrix, calculate singular values of it as well as its constituent matrices U, S, V. Proper interpretation of LSI in geometric context is also available in existing literature.[4] Argues that real power of extracting the hidden thematic structure or latency of LSI comes from SVD.

Although researchers have advanced the use of LSI and have also proposed theoretical understanding of it, however[9] initially studied values produced by LSI [6]. Besides other advancements in LSI, like PLSI,[10] describe LSI in terms of a subspace model and propose a statistical test for choosing the optimal number of dimensions for a given collection. April Kontostathis explores the appropriate k dimensions to which SVD can be truncated to. The optimal K can be chosen by running a set of queries with known relevance to documents in a collection and the value of K for which retrieval performance is best, can be chosen as optimal K [6]. [12, 13] claim that optimal value of K lies in the range of 100 – 300 dimensions. Despite research studies carried out, optimal dimensionality reduction parameter (k) for each collection remains elusive.

Given the available literature on VSM and LSI and their mathematical understanding as well as theoretical approximation, different experimental results drawn from different datasets motivate that VSM behaves a similar way as LSI except for optimal dimension

reduction parameter K, where LSI shows better performance. LSI as claimed to have better semantic quality has however lot of computational effort involved in singular value decomposition. For collections that are dynamic in nature, SVD updating is also serious issue [3, 4].

III. Description of Algebraic Model Based Algorithms :

various information retrieval systems that have been developed in the recent past, systems that work on algebraic model based algorithms, model the data using matrices. User's query is modelled as a vector (a column vector or a row vector). Relevant information that user wants to extract from the data is extracted by simple vector operations. For collections of data (datasets) that are larger in size may result in bigger matrices. Familiar algebraic operations like orthogonal factorizations can be used to reduce the size of that matrix. These basic algebraic operations have led to few information retrieval algorithms that are commonly known as algebraic model based information retrieval algorithms. These algorithms include I) VSM and II) LSI.

Before proceeding to next section, a formal description of these algorithms follows. The purpose of this description is to show how fundamental mathematical concepts from linear algebra can be used to manage and index large text collections.

1. Vector Space Model (VSM) :

A matrix with a single row or column is referred as row or column vector or it is simply called as vector. In vector space information retrieval model, each document in a collection is represented using vector. Each component (*i.e.* each entry in this vector) reflects a particular key word associated with the document. For example for a kth document if $(k,m)_{th}$ component is a non-zero value, it implies that mth indexing term is present in kth document. Moreover value assigned to that component reflects the importance of the indexing term in representing the document. This value is typically the function of the local frequency and global frequency of a given term using some standard normalization procedures. A collection containing d number of documents described by t terms is represented as a t x d term-by-document matrix. The d vectors representing d documents form the columns of the term-document matrix $[a_{ij}]$, where a_{ij} is the weighted frequency at which term i occurs in document j. Similarly rows of the term-by-document matrix $[a^{ij}]$ are the term vectors. Therefore semantic content of the collection is only contained in the column space of matrix $[a_{ij}]$.

A similar approach is used to model a query using a column vector generally called as query vector. Query is set of terms, may be with weights, and may contain words that may or may not be used to describe the collection of documents. All such words that are relevant to the collection are identified. Each component of the query vector reflects the presence/ absence of the indexing term in the query. Length of the query vector is equal to the number of the terms that describe the collection.

From information retrieval perspective, VSM explores the geometric relationship between the document vector [dij] and query

vector [qij] by measuring the angle that these vectors make with each other. If a document vector d that makes a minimum angle with query vector q then document d is treated are most relevant to query q. Angle between document vector [dij] and query vector [qij] and is computed using following equation

$$\text{Cos(theta)}_k = d_k^t * q / (\|aj\|_2 * \|q\|_2) = \frac{(\sum_{i=1}^{|D|} aij\ qi)}{\sqrt{\sum_{i=1}^{|D|} aij * aij}\sqrt{\sum_{i=1}^{t} qi * qi}}$$

Where d_k is column vector that represents kth document of the collection, q is the query vector. Transpose of the document vector is taken for multiplication compatibility of two vectors. $\| X \|$ represents Euclidean norm of vector X and is computed as $\|X\| = \sum_{i}^{|X|} xi * xi$

The Cos(theta)_k provides a measure of the similarity of document dk to query q. These results are ordered in descending order. Thus minimum value for Cos(theta)_m means that mth document is most relevant to the query q. This equation can be used to rank documents for each query. Retrieval system using VSM algorithm ranks the documents in-order of their relevance to a given query q.

2. Latent Semantic Indexing (LSI) :

Term document matrix resulted from a given dataset with documents forming the column vectors and terms that describe the collection, forms the row vectors of it. This term-document matrix (say A) can be decomposed using mathematical technique commonly known as Singular Value Decomposition (SVD). This decomposition process decomposes term-document matrix A into three matrices: a term by dimension matrix T, a singular matrix S and document by dimension matrix D. where number of dimensions "r" is the rank of term-document matrix A. A is decomposed into T, S and D matrices.

Matrix A can be recomputed using following equation

$$A = T * S * D^T \qquad (2)$$

The objective of this decomposition is not to re-compute A using T, S and D respectively, however A can be approximated by reduced dimensions of T, S and D respectively. This rank reduced matrix becomes basis for LSI. This dimension reduction of term-document matrix is accomplished by removing k+1 to r columns of T, k+1 to r columns and rows of S and k+1 to r rows of D^T. This dimension reduction is thought to reduce the noise in term-document matrix. Further researchers claim that this process reveals the latent structure present in the collection. Queries are converted into vectors. Relevance of Query with a document is computed by measuring angle between reduced document vectors and query vector q.

$$\text{Cos(Q)}_k = d_k^t * q / (\|aj\|_2 * \|q\|_2) = \frac{(\sum_{i=1}^{|k|} aij\ qi)}{\sqrt{\sum_{i=1}^{|k|} aij * aij}\sqrt{\sum_{i=1}^{t} qi * qi}}$$

This equation provides a similarity score for each document for a given query. As with vector space retrieval, the scores are sorted in descending order. Document vector that makes minimum angle with the query vector is treated as most relevant document to a given query.Optimal dimensionality reduction parameter (k) can be chosen by running a set of queries with known relevant

document sets for multiple values of k. The k that results in the best retrieval performance is treated as optimal k for a collection. Optimal values for dimension reduction parameter k lie typically in the range of 100-300[11].

IV. Results

a) CACM Results

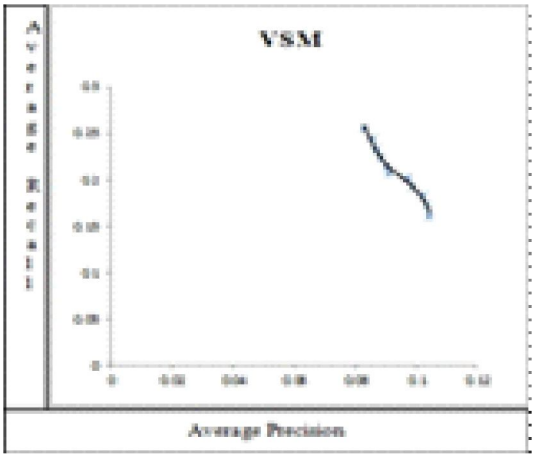

Figure 1.1: Average Precision, Recall curve of VSM Algorithm

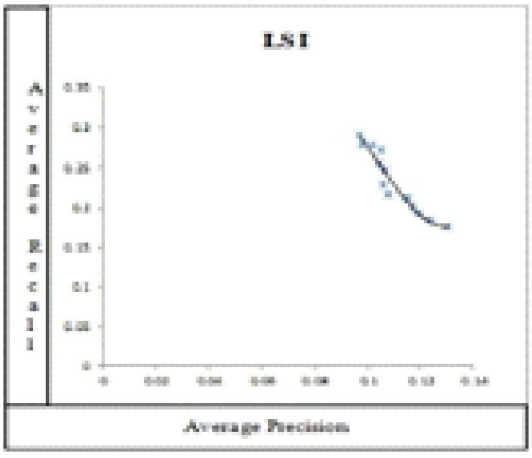

Figure 1.2: Average Precision, Recall curve of LSI Algorithm

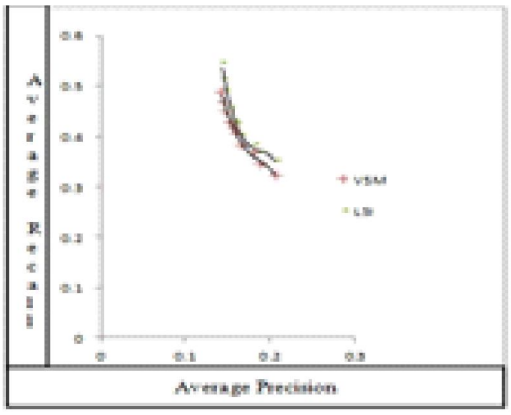

Figure 1.3: Average Precision, Recall curves of VSM and LSI Algorithms

b) CISI Results

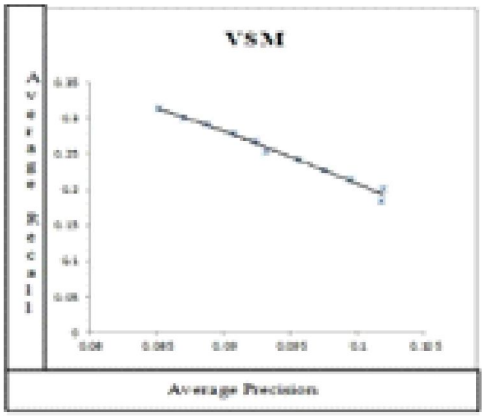

Figure 2.1: Average Precision, Recall curve of VSM Algorithm

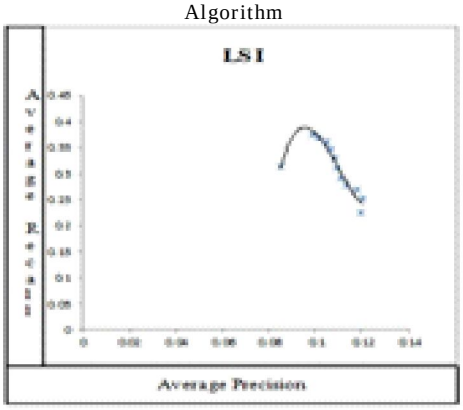

Figure 2.2: Average Precision, Recall curve of LSI Algorithm

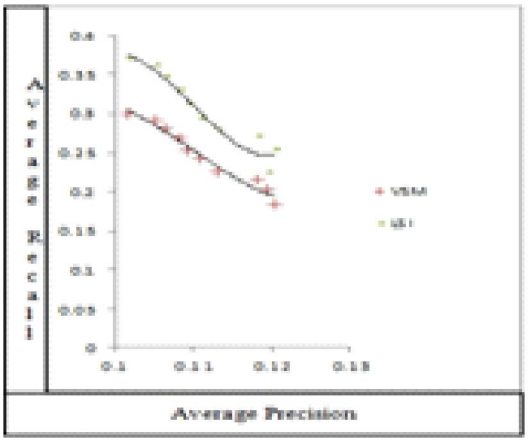

Figure 2.3: Average Precision, Recall curves of VSM and LSI Algorithms

V. Discussion and Conclusion

Analysing Average Precision versus Average Recall results VSM and LSI, it is found that although VSM uses complete term-document matrix resulted from collections CISI and CACM, however is less efficient than LSI. Efficiency of LSI can be attributed to better semantic quality (capability of the indexing term to better describe the text) of resulting indexing terms after decomposing term-document matrix and using appropriate number of singular values. Moreover use of inappropriate number of singular values for computation of reduced term-document matrix in-case of LSI affects retrieval of relevant documents. However optimal number of use of singular values varies from collection to collection thus remains elusive.

Although efficiency of an information retrieval algorithm is attributed two kinds of properties of indexing terms i) Statistical property ii) Semantic property. But due to lack of a standard measure to gauge these properties mathematically, these qualities are merely considered by intuition than supporting theory. Moreover taking into consideration the synonymy and polysemy of an English text, we need a method that mathematically takes this aspect into consideration as well. Furthermore new models like Fuzzy model and Probabilistic models may be used to further the research vis-à-vis information retrieval as these two models work completely in a different way than set-theoretic models and algebraic models.

VI. References

1. Wen Zhang, Taketoshi Yoshida and Xijin Tang, “A comparative study of TF*IDF, LSI and multi-words for text classification,” Expert Systems with Applications Vol. *38*, pp. 2758–2765 (2011).
2. Amit Singhal, “Modern Information Retrieval: A Brief Overview,” Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, (2001).
3. Michael W. Berry, Susan T. Dumais and Gavin W. O’Brein. “Using Linear Algebra for Intelligent Information Retrieval,” SIAM Review. Vol. *37* No. 4. (Dec., 1995), pp. 573-595.
4. Michael W. Berryt, Zlatko Drmacf and Elizabeth R. Jessupt “Matrices, Vector Spaces, and Information Retrieval”, SIAM Review, Vol. 41, No. 2 (Jun., 1999), pp. 335-362.
5. Wen Zhang, Taketoshi Yoshida and Xijin Tang “TFIDF, LSI and Multi-word in Information Retrieval and Text Categorization,” IEEE International Conference on Systems, Man and Cybernetics (SMC

2008).

6. April Kontostathis, "Essential Dimensions of Latent Semantic Indexing (LSI)," Proceedings of the 40th Hawaii International Conference on System Sciences (2007).
7. Greengrass, Ed, "Information Retrieval : A Survey by Ed Greengrass." Information Retrieval, November (2002).
8. Scott Deerwester, Susan T Dumais, George W Furnas and Thomas K Landauer "Indexing by Latent Semantic Analysis," Journal of the American Society for Information Science (1986-1998), Sep 1990.
9. Schutze, Hinrich. "Dimensions of meaning," Supercomputing'92., Proceedings. IEEE, (1992).
10. H. Zha and H. Simon, "A subspace-based model for Latent Semantic Indexing in information retrieval," In Proceedings of the Thirteenth Symposium on the Interface, pages 315–320 (1998).
11. http://ir.dcs.gla.ac.uk/resources.html.
12. Kontostathis, April. "Essential dimensions of latent semantic indexing (LSI)."System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on. IEEE (2007).
13. Letsche, Todd A., and Michael W. Berry. "Large-scale information retrieval with latent semantic indexing." Information sciences 100.1, 105-137 (1997).