A Comparative Study of Vector Space Model and Probabilistic Model for Information Retrieval

[1]Satya Prakash Awasthi
Department of Computer Science,
SHRI VENKATESHWARA UNIVERSITY,
Gajraula, UP, INDIA,

[2]Sandeep Gupta
Department of Computer Science Engineering, JIMS Engineering Management Technical Campus,
Greater Noida, UP, INDIA

Abstract — The method of personalization can either focus on individuals and their interaction with documents, or on the identification of shared patterns of behavior and the segmentation of the user population in to groups of common interests. One problem in the personalization and the filtering process is the choice of a suitable model for effective user profile and document representation and manipulation. In terms of resemblance between users and records it should be able to facilitate the identification of appropriate document. A comparative study is presented in this paper between probabilistic model and vector space model .Notations and definitions necessary to identify the concepts and relationships that are important for modeling objects and processes in the context of vector spaces.

Keywords — Information retrieval, Vector space model, Probabilistic model.

I. INTRODUCTION

Information Retrieval is an organizational discipline; Storage, retrieval and display of bibliographic Information. Information Retrieval Systems are designed with the aim of offering, in response to a user query, references to documents which would contain the data required by the user[1]. A major difference between information retrieval systems and other types of information system is the intrinsic uncertainty of Information Retrieval, whereas for database systems an information need can always(at least for standard applications) be mapped precisely on to a query formulation, and there is a precise definition of which elements of the data base constitute the answer, the condition is much more difficult in Information Retrieval: here neither a query presumed can be assumed to reflect a unique need for an information , nor is there a definite method for deciding whether or not a data base objects is a response.

Web search outcomes are impacted by the fact that Web Search systems design often lack understanding of user's needs. In order to be able to provide the data a user is looking for there is a critical need to know how individuals are using web, how they are looking for data and what tools or methods they use to locate appropriate records[2]. Thanks to the Internet boom, searching for documents through textual queries has become a prevalent practice for a large proportion of society.

Search engines are a sort of information retrieval system. In reaction to the use's query a rank or sorted list of documents was obtained. Evidence of resemblance between the request and each component must be an information system. Information models are important because they represent three distinct mathematical models, with their own techniques of documents representations and resemblance calculation between documents and user's profile. It will focus on two models; the vector space model and the probabilistic model [3]. The vector space model is an algebraic model in which documents and user's profile are presented as vectors operations. Probabilistic model is used to obtain the records.

Vector Space Model:

Vector Space Model is an algebraic model used for data processing, information retrieval, indexing and classification of significance. The Vector Space Model [4] is a way to represent and compare values-based records and queries. This model can be used to rank the resemblance between documents, not just to respond if document contains and does not contain the necessary phrases. Each component of a vector reflects one term, and has a value. The valuation is a real number that shows how applicable a word is to the specified document or request. Processing of Vector Space Model can be split into two phases. Indexing of the documents with word weighting and classification of relevance ranking.

The fundamental assumptions of implementing the vector space model is that the different items of information retrieval are presented as vector space components [5]. The fundamental assumption of implementing the model of vector space is that the

different items of information retrieval are represented as vector space components. Documents, queries, ideas and so on in particular. All vectors are in the room of the vector. Are all vectors in vector space?

Let us first consider the issue of representation of documents in terms of index terms. Let t_1, t_2, ….t_n be the terms used to represent documents, Corresponding to each term ti, suppose there exists a vector ti in the space. Without loss of generality, it is assumed that ti s is vectors of unit length. Now, suppose that each document Dr, $1 \leq r \leq m$, is a vector expressed in terms of t_{is}. Let the document vector Dr be $Dr=(a_{1r},a_{2r},\ldots\ldots a_{nr})$ where a_{irs} are real numbers reflecting the importance of term i in Dr.

Every vector in this sub space, and in particular all documents vectors are linear combinations of the term vectors. Thus D can be expressed as

$$Dr = \sum_{i=1}^{n} a_{ir} t_i \qquad (1)$$

The coefficients air, for $1 \leq i \leq n$ and $1 \leq r \leq m$ are the components of Dr along the t_{is}.

A set of vectors y_1 , y_2, …….y_k are linearly dependent. If there exists some scalar a_1, a_2,……..a_k. so

$$a_1 y_1, a_2 y_2, \ldots\ldots a_k y_k = 0$$

Important concepts and relationship for applying vector space model in Information Retrieval:

For reasons of clarity, it is assumed that the number of terms is equal to the dimension of the subspace of interest, and the number of documents is exactly the same as the number of terms. i.e. n' = n = m

We know that the term Vector t_1, t_2, …..t_n are normalized. However, that the vectors in each set are not assumed to be pair wise orthogonal.

A- Computation of similarity measures between documents and query.

From equation-1, we have

$$Dr = \sum_{i=1}^{n} a_{ir} t_i \qquad (r = 1,2,\ldots n) \qquad (2)$$

For any query q, the corresponding query vectors has the expression.

$$q = \sum_{i=1}^{n} q_i t_i \qquad (3)$$

In general case , the scalar product , which we suppose is the measure of similarity between two vectors D and q is

$$Dr.q = \sum_{i,j=1}^{n} a_{ir} q_i t_i t_j \qquad (4)$$

Projections versus Components:

Components of documents along the term vectors are related to the corresponding projections via the term – term similarities. By multiplying equation-3 by tj, (j=1 , 2 , 3….n) on both sides . by this

$$Drt_i = \sum_{i=1}^{n} a_{ir} t_i t_j \qquad (j_r = 1,2\ldots.n) \qquad (5)$$

Since t_is are unit vectors, the scalar product $t_j.D_r$ is the projections of D_r on to t_j. Equation -5 can be rewritten in a matrix form[6-9].

$P=G_iA$

Where $(P)_{jr} = t_j . D_r$

$(G_1)_{ij} = t_j.t_i$ and

$(A)_{ir} = a_{ir}$

That is G_1 is the matrix generations between term vectors, and the rth column of A represents the components of D_i along the vector t_is.

Vector Space Model in Information Retrieval choices and implications:

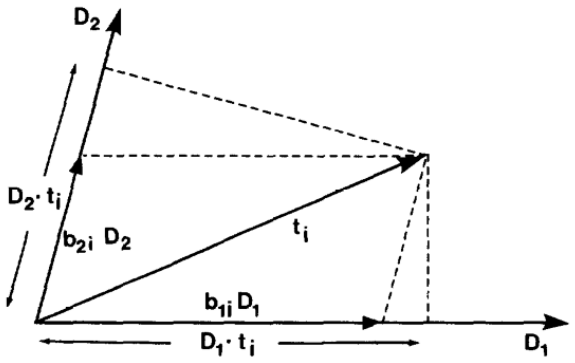

Figure-2: Two dimensional vector space with D r S as basis

Now we will discuss a spectrum of options available in terms of how the model can be applied.

Some of the model elements identified are the component matrix A,B the projection Matrix P, and the term – term and document – document correlation matrices Gi and Gij. Since these are Related, Not all of these matrices need be known[10-12].

The Standard Vector Model:

D is interpreted to correspond to the component matrix A(i.e. dir is air, the component of Dr along ti) . In the standard vector model the assumption that $t_i . t_j = 1$, if i = j and 0 otherwise, is made (in other words, this assumption means G_i =I)

From equation-6, it follows

If G_i= I then P = A

Thus , the above specifications imply the interpretation

D = A = p, i.e. d_{ir}s are both projections and components of the documents along the term vectors.

Where $q = (q_1 , q_2 ,\ldots\ldots q_n)$ is the query vector and q_is are the components of q along the term vectors. However , document – document correlation can be obtained[13] by

$G_d = P'A = D'D$

An Alternative Interpretation:

In the Standard vector model , Gi is assumed to be an identity Matrix which then leads us to the interpretation that D = P.

Let us represent the document query similarity Dr.q for r= 1 , 2,…..n as a vector Rq =(D1q, D2q, ……..Dnq) which can be written as

$$P = \begin{bmatrix} t_1 \cdot D_1 & t_1 \cdot D_2 \\ t_2 \cdot D_1 & t_2 \cdot D_2 \end{bmatrix}$$

$$= \begin{bmatrix} t_1 \cdot (a_{11}t_1 + a_{21}t_2) & t_1 \cdot (a_{12}t_1 + a_{22}t_2) \\ t_2 \cdot (a_{11}t_1 + a_{21}t_2) & t_2 \cdot (a_{12}t_1 + a_{22}t_2) \end{bmatrix}$$

$$= \begin{bmatrix} t_1 \cdot t_1 & t_1 \cdot t_2 \\ t_2 \cdot t_1 & t_2 \cdot t_2 \end{bmatrix} \cdot \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = G_i A. \quad \square$$

Where R'q and q' denote the transpose of matrices Rq and q respectively.

$P = G_iA$

$R_q = q (G_iA) = q^p$

The Dual of the Standard Vector Model:

We can instead interchange the role of documents and terms to obtain a new interpretation. Now suppose that D is interpreted as B. The components of terms along document vectors, then

$t_i.t_j =$

Furthermore, Let G_d be assumed to be an identify matrix[14]. So by this.

$t_i.t_j =$

For this special case , P = B' = D implying that

$G_i = PP'$

Probabilistic Information Retrieval

Vector Space Model provides similarity rating without taking in to account a standard of the assurance for the significance of output relevance. There are several models centered on probability theory aimed at determining the probability of a document being applicable to a query.

In this the user information is depicted here by a collection of potentially weighted keywords provided by the users or caused by the system.

The findings obtained by probabilistic information retrieval systems are based on estimate and probabilities. The first hypothesis is that the conditions are differentiated between appropriate and non-relevant document. A probabilistic Information System will rank documents and sorts them in reducing order of probability of relevance to the need for information. The outcomes are as precise as the probability calculated.

In decreasing order of calculated probability of relevance to the data necessity, the basic probability model returns records. After the indexing process every term can have allocated a value that shows the probability that a document comprising this term is relevant to the concept mentioned by the term. In the retrieval phase the documents have calculated a value which is the sum of probabilities from terms that exists in both a document and in the query. The documents are then retrieved in order according to this value. For this version of probability information retrieved, the document representation could be the same as in the Boolean model, as information only needs to be stored if either document contains a term or not.

If P is the probability that a document which contains a term and it is relevant to the query and m is the probability that the document contains the term but it is not relevant, the the weight of the term is calculated as

$$w_i = \log \frac{Pi\ (1 - mi)}{mi(1 - Pi)}$$

Where

$$P = \frac{\text{numer of relevant documents containing term}}{\text{total number relevant of ducuments}}$$

If

n_i = Number of documents containing term i

r_i = Number of relevant documents containing term i

N = Total number of documents

R = Number of relevant documents

The p and m can be expressed as

$$P_i = \frac{ri}{R}$$

$$m_i = \frac{ni - ri}{N - R}$$

and w_i can be expressed as

$$w_i = \log \frac{Pi\ (1 - mi)}{mi(1 - Pi)} = \log \frac{ri\ (N - R - ni + ri\ )}{(ni - ri)(R - ri\ )}$$

Usually it is assumed that the probability P is constant and that m can be created by the values from Inverse Document Frequency vector.

II. CONCLUSION

The way in which the vector space and probabilistic model has been introduced and used in the literature has led to a situation where many important concepts are ignored or poorly understood. In this work, we critically review the vector space and probabilistic model for information retrieval by using notation that more clearly brings out problems and challenges associated with the use of the vector space model.

We believe that this work will lead to the harnessing of the real power inherent in the probabilistic and vector space model as a formal framework for developing information retrieval systems.

III. ACKNOWLEDGEMNT

I offer my earnest thanks to my guide Dr. Sandeep Gupta, Associate Professor (CSE Department) for his constant help, worth full guidance and encouragement during the work. I would like to thanks to SVU, Uttar Pradesh for giving me such platform for taking my research work to some heights.

REFRENCES

[1] Robertson, S. E.; Maron, M. E.; Cooper, W. S. "Probability of Relevance: A Unification of Two Competing Models for Document Retrieval." Information Technology: Research and Development. 1(1):1-21; 1982.

[2] Salton, G.; McGill, M. H. Introduction to Modern Information Retrieval. New York: McGraw-Hill; 1983.

[3] Salton, G. (ED.). The SMART Retrieval System-Experiments in Automatic Document Processing. Englewood Cliffs, NJ: Prentice• Hall; 1971.

[4] Salton, G. Dynamic Library and Information Processing. Englewood Cliffs, NJ: Prentice-Hall; 1975.

[5] Koll, M. "An Approach to Concept Based Information Retrieval." ACM-SIGIR Forum, Xlll:32-50; 1979.

[6] Greub, W. H. Linear Algebra. New York: Academic; 1963.

[7] Salton, G. "Automatic Term Class Construction Using Rele• vance-A summary of Work in Automatic Pseudoclassification." Information Processing and Management. 16(1):1-15; 1980.

[8] Raghavan, V. V.; Yu, C. T. "Experiments on the Determination of the Relationships Between Terms." ACM Trans. on Database Systems. 4(2):240-260; 1979.

[9] Wong, S. K. M.; Ziarko, W.; Wong, P. C. N., "Generalized Vec• tor Space Model in Information Retrieval II," Proc. of ACM• SIGIR Conference on Research and Development in Information Retrieval, June 1985, Montreal, Canada.

[10] Salton, G. "Experiments in Automatic Thesauras Construction for Information Retrieval." Information Processing 71, Amster• dam, The Netherlands: North-Holland; 1972:115-123.

[11] Sparck-Jones, K. Automatic Keyword Classifications. London: Butterworths; 1971.

[12] Minker, J.; Wilson, G. A.; Zimmerman, B. H. "An Evaluation of Query Expansion by the Addition of Clustered Terms for a Document Retrieval System." Info. Stor. and Retrieval. 8:329-348; 1972.

[13] Can, F.; Ozkarahan, E. A. "Concepts of the Cover-Coefficient· Based Clustering Methodology," Proc. of ACM-SIGIR Confer• ence on Research and Development in Information Retrieval, June 1985, Montreal, Canada.

[14] Can, F.; Ozkarahan, E. A. "Similarity and Stability Analysis of the Two Partitioning Type Clustering Algorithms." Journal of the American Society of Information Science. 36(1):3-14; 1985.