A Comparative Study on Approaches of Vector Space Model in Information Retrieval

Jitendra Nath Singh
Department of Computer Science,
Babasaheb Bhimrao Ambedkar University
Lucknow, India

Sanjay Kumar Dwivedi
Department of Computer science
Babasaheb Bhimrao Ambedkar University
Lucknow, India

ABSTRACT
The vector space model is one of the classical and widely applied information retrieval models to rank the web page based on similarity values. The retrieval operations consist of cosine similarity function to compute the similarity values between a given query and the set of documents retrieved and then rank the documents according to the relevance. In this paper, we are presenting different approaches of vector space model to compute similarity values of hits from search engine for given queries based on terms weight. In order to achieve the goal of an effective evaluation algorithm, our work intends to extensive analysis of the main aspects of Vector space model, its approaches and provides a comprehensive comparison for Term-Count Model, Tf-Idf model and Vector space model based on normalization.

Keywords
Vector space model, Information Retrieval, Tf-Idf, Term-Frequency, Cosine Similarity.

1. INTRODUCTION

Information retrieval systems are designed to help users to quickly find useful information on the web. The field of information retrieval attained peak popularity for many years, number of researchers contributed through their efforts and achieved several remarkable milestones in order to facilitate the internet users with easiest searching in very small slots of time. In the recent years, number of doubts emerged that demand for adequate solutions to satisfy searchers. Performance evaluation of search engines [8] and their differentiation are very popular issue, lots of possible solutions have been proposed but still satisfactory results are not achieved. Researchers followed the statistical way to evaluate the search engines and to select best search engine from various available all. However, for the task of seeking information, these statistical techniques have indeed proven to become the most effective and efficient ones so for. Statistical model, that consisting Vector Space Model [1], [2], [6], [7] and Probabilistic model [2] helped much and became the base line for their framework and algorithms. The Boolean model [2] that is also known as "exact match" model is still being used by most of the online services. In the process of information retrieval two key problems still exist: First, information retrieval process fetch some irrelevant documents together with relevant document. second, search engines are not capable to perform retrieval of all relevant documents [3].

1.1 INFORMATION RETRIEVAL MODELS

The goal of information retrieval (IR) is to provide users with those documents that will satisfy their information need. To make the information retrieval efficient, the documents are typically transformed into a suitable representation. Now such type of information is retrieved efficiently with the help of IR models [2]. The models are categorized according the properties of the models.

The following major models have been developed to retrieve information: the Set-Theory model, the Statistical model, which includes the vector space and the probabilistic model. Set-theoretic models represent documents as sets of words or phrases. Similarities are usually derived from set-theoretic operations on those sets. Common models are: Standard Boolean, Smart Boolean and Extended Boolean models. The Boolean model [2] is based on Boolean logic and classical set- theory. In that both the documents to be searched and the user's query are conceived as sets of terms. Retrieval is based on whether or not the documents contain the query terms.

The Boolean model represents documents by a set of index terms, each of which is viewed as a Boolean variable and valued as True if it is present in a document. Boolean model cannot rank documents in decreasing order of relevance.

The vector space and probabilistic models are the two major example of the statistical retrieval approach.

Both models use statistical information in the form of term frequencies to determine the relevance of the documents with respect to a query. Although they differ in the way they use the term frequency, both produce as their output a list of documents ranked by their estimated relevance.

The vector space model [2] has been widely used in the traditional IR field.

Search engines use similarity values computed by VSM model to rank the web documents.

This paper is organised as follows.

In Section 2, we present concept of vector space model and its three approaches. In Section 3, we present our discussion and results. In Section 4, we present comparison of methods. Finally, in section 5, we conclude the paper.

2. VECTOR SPACE MODEL (VSM)

The vector space model represents documents and queries as vectors in multidimensional space, whose terms are used as dimensions to build an index to represent the documents.
Each dimension corresponds to separate term. If a term occurs in the document, its value in the vector is non-zero. It is used in information retrieval, indexing and relevant ranking and can be successfully used in evaluation of web search engines. The vector space model procedure can be divided into three stages. The first stage is the document indexing where content bearing terms are extracted from the document text. The second stage is the weighting of the indexed terms to enhance retrieval of documents relevant to the user. In the last stage, rank the documents with respect to the query according to a similarity values. A common similarity measure known as cosine measure determines angle between the document vector and the query vector. The angle between two vectors are considered as a measure of divergence between the vectors, cosine angle is used to compute the numeric similarity, determines angle between the document vector and the query vector when they are represented in V-dimensional Euclidian space where V is the size. The similarity function [4] between documents vectors Di and query Q is given by,

$$\text{Cosine}\theta = \text{Sim}(Q, D_i) = \frac{\sum_{j=1}^{V} w_{Q,j} \times w_{i,j}}{\sqrt{\sum_{j=1}^{V} w_{Q,j}^2} \times \sqrt{\sum_{j=1}^{V} w_{i,j}^2}} \qquad (1)$$

Where $w_{i,j}$ is the weight of term j in document i and $w_{Q,j}$ is the weight of term j in the query. The denominator in this equation, called the normalisation factor, discards the effect of documents length on document scores. The weight $w_{Q,j}$ is defined similar way as $w_{i,j}$ but computing the weight of query vector (the first term in the denominator of Eq.1) requires access to every document, not just the terms specified in the query and for second term in the denominator, it is very difficult to calculate weight of each term in documents. So to avoid the problem of these normalisation factors, we use the square root of the number of terms in a document as normalisation factor. With this approximation, the similarity function [12] between documents vectors Di and query Q is given by

$$\text{sim}(Q, D_i) = \frac{\sum_{j=1}^{V} w_{Q,j} \times w_{i,j}}{\sqrt{\text{number of terms in document Di}}} \qquad (2)$$

The term weighting scheme plays an important role in similarity measure. So based on weighting scheme different approaches [11] of vector space models have been derived (Tf only, Idf only, and combination of these two).

2.1 Term –Count Model

In this model, and any other models, we need database collection to retrieve documents, input query and index term. The terms are single words or keywords. If words are chosen to be terms, the dimensionality of the vector is the number of words in the vocabulary. Relevance ranking of documents in a keyword search can be calculated, using the assumptions of document similarities, by comparing the deviation of angles between each document vector and the original query vector where the query is represented as the same kind of vector as the documents using Eq2. In Term- count model [9] weight of terms has been computed using local information only (i.e. Term frequency) given by.

$$\text{WEIGHT} = W_{i,j} = TF_{IJ} \qquad (4)$$

Where TF_{IJ} = Frequency of term j in documents i.

2.2 Tf-Idf (Classical) Vector Space Model.

In information retrieval or text mining, the term frequency – inverse document frequency also called as Tf-Idf, is a well-known method to evaluate how important is a word in a document. It is often used as weighting factor in information retrieval.Tf-Idf [5] is also a very interesting way to convert the textual representation of information into a Vector Space Model (VSM). Unlike the Term Count Model, Tf-Idf incorporates local and global information as shown in Eq3; hence weighting schemes use both local and global information. The observation here is that if cosine angle is near to 1, the documents are more similar to query terms. On the basis of cosine values calculated, these models rank the retrieved documents on the similarity score. The weight of term in document vector can be determined using $\mathbf{Tf \times Idf}$ method. The weight of term is measured how often the term j occurs in the document i (the term frequency $tf_{i,j}$) and Idf (the inverse document frequency). The weight of a term j in the document i is given by.

$$w_{i,j} = Tf_{i,j} \times Idf \quad \& \quad Idf = \log \frac{D}{df_j} \qquad (3)$$

Where D is the number of documents in the document collection and df_j number of documents containing the query term,

2.3 Vector Space Model Based on Normalised Frequency

In this model our objective is to improve the position of a document in the search engine ranking result. In this process normalization is a way of penalizing the term weights for a document and query. We apply commonly used normalization techniques given by.

2.3.1 *Maximum Tf Normalization:*

The most popular normalization technique is normalization of individual Tf weights for a document by maximum Tf in the document.
The normalized frequency of a term j in document i is given by:

$$f_{i,j} = \frac{Tf_{I,J}}{\max tf_{I,J}} \qquad (5)$$

$f_{I,J}$ = normalized frequency.

$Tf_{I,J}$ = Frequency of term j in documents i.

The normalized frequency [9], [10] of a term j in a query Q is given by.

$$f_{Q,j} = 0.5 + 0.5 * \frac{Tf_{Q,J}}{\max Tf_{Q,J}} \qquad (6)$$

$Tf_{Q,J}$ = frequency of term j in query Q

f_{Qj} = normalized frequency

It is restricting the Tf factors to maximum value of 1.0.

3. EXPERIMENTAL RESULTS

In the experiments, we applied the three approaches of vector space model to measure the relevance of web page. The experiments were based on an accepted number of TREC pattern queries. These queries contain 1, 2 or 3 terms. The query set contains 10 queries i.e. query Id from 1 to 10 given by.

1.Iodine in Blood
2.Student Job
3.Survey Maps
4.Global Warming
5.Segrass
6.Surface Area Evaporation
7.Corn price
8.Food Service
9.Loan Proposal
10.Weather Radar

We applied keyword based search on Google search engine and considered only top five documents from number of documents retrieved.

Ranking of documents depends on the similarity values computed by three approaches of VSM.The similarity values are based on occurrence of keywords. We only considered the title, snippet and meaningful words from retrieved documents and discarded the stop words like a, an, the, in, of, for etc.

We observed the relevance of documents based on term appearances in the documents and similarity values. In this experiment, we observed that all query terms were presented in all top five documents. So Idf value became constant using Eq3. So we considered Idf value 1 in such condition.

Based on experiments, we observed that two models, term-count model and Tf-Idf model show that they are producing higher rank to the documents that have more repetitive words. But it is not necessary that documents having more repetitive words are more relevant than short one. Due to that it affects documents ranking. It is also observed in these two models that if any document containing maximum number of query terms but length of document does not increase accordingly, it computed similarity value above 1 which is not very appropriate. So we need to normalise frequency of terms in documents and queries, since normalisation decreases the differences in similarity values between documents and also improves the ranking of documents. Vector space model based on normalised frequency is as next approach of vector space models as discussed above.

DISCUSSION & COMPARISON OF METHODS

Based on the experiments using three approaches of the vector space model, we have made certain observations. Comparison of three approaches of VSM is based on similarity value computed in Table1 using these three approaches and Eq2 as discussed above.

The Term-count model provided higher similarity values for long documents (documents containing more query terms) as compared to small documents, since these contain many words that are very often repeated thus for long documents similarity values will be higher. In Tf-Idf model, Tf-Idf weight is numerical statistics which shows the importance of words in the document.

If we use Eq1 as discussed in previous study [11], we need to compute weight of each term in the documents. Idf can be successfully used for filtering stop-words (a, an, the etc.) or most common words. That is why Tf-Idf model is better than term-count model in such situations. But we used Eq2, we did not need to compute weight of each term in the documents. It facilitates the computation easy. It is also observed that all query terms presented in all five documents, so Idf value was considered constant i.e. 1. Tf-Idf in such situations only depends on number of terms presented in the documents hence term-count and Tf-Idf models provided same similarity values as shown in Table 1. The VSM model based on normalized frequency provided low similarity values as compared to term-count and Tf-Idf models for those documents that contain more repetitive words because when the term frequency in the documents and the query is normalized, it reduces weight of terms in the documents and queries. The low document weight decreases the similarity score and it is observed that it changes the rank of such documents.

Table 1 show the experimental results based on all three approaches of VSM and cosine function given by Eq2.

Table 1. Similarity values based on weight

Q ID	TERM-COUNT MODEL					TF-IDF MODEL					NORMALISED MODEL				
	D1	D2	D3	D4	D5	D1	D2	D3	D4	D5	D1	D2	D3	D4	D5
1	.7958	.8944	.9176	1.0	.8944	.7958	.8944	.9176	1.0	.8944	.3127	.4472	0.3058	0.3401	0.4472
2	0.834	0.8728	0.6396	1.0	1.0	0.834	0.8728	0.6396	1.0	1.0	0.4170	0.4364	0.3198	0.3333	0.2984
3	1.0	.9428	.9176	.6708	.9128	1.0	.9428	.9176	.6708	.9128	.3211	.3925	.3823	.3354	.2282
4	1.0	.6324	.5773	.8728	1.0	1.0	.6324	.5773	.8728	1.0	.5163	.6324	.5773	.4364	.4082
5	.4714	0.5	0.75	.4714	.4714	.4714	0.5	0.75	.4714	.4715	.2357	0.25	0.25	.2357	.2357
6	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	.4760	.6123	.4263	.5214	.5214
7	1.0	1.0	.7559	1.0	.6546	1.0	1.0	.7559	1.0	.6546	.2738	.3000	.3779	.3553	.3273
8	.4264	.8164	.8164	.9428	.4082	.4264	.8164	.8164	.9428	.4264	.4264	.4264	.4264	.4714	.4082
9	1.0	1.0	1.0	.8528	.8164	1.0	1.0	1.0	.8528	.8164	.4264	.3553	.3779	.4264	.4082
10	.9428	1.0	1.0	.8528	.8340	.9428	1.0	1.0	.8528	.8340	.3142	.4264	.5163	.2842	.2780

two approaches of vector space model favour long documents where the documents containing more appearance of query terms but normalisation model penalises such documents.

5. CONCLUSIONS

In this paper, we performed extensive analysis of the three approaches of vector space model with an accepted number of TREC pattern queries. We computed the similarity values of top five documents by using three approaches of vector space model. A higher similarity value is observed by term-count model for long documents, Tf-Idf model in such situation where the query terms are presented in each document also provided similar results as in term-count model. The normalization model does not provide similar results in these situations; it normalize term frequency in query and documents due to that similarity values change for long documents that is why this model penalizes the long documents. We concluded that the first two approaches of vector space model may favour for long documents with large number of unique terms while last one penalizes the long documents and changes the ranking of documents. For future work, we will focus on updating one of these approaches of vector space model according to the search scenario that could be use in better evaluation of search engine.

6. REFERENCES

[1] Shalton, G; Wong, A; Yang, C.S.: A vector space Model for automatic indexing, Communications of the ACM, Volume 18 and Issue 11:1975.

[2] Sanjay Kumar Dwivedi, Jitendra Nath Singh, and Rajesh Gotam Information Retrieval Evaluative Model, FTICT 2011: Proceedings of the 2011, International conference on Future Trend in Information & Communication Technology, Ghaziabad, India: 2011.

[3] Yi Shang Longzhuang Li: Precision Evaluation of Search Engines, World Wide Web: 2002.

[4] D.L. Lee, H. Chuang, and K. Seamons. Document ranking and the vector space model, IEEE Transactions on Software, 14(2): 1997.

[5] Chris Buckley. The importance of proper weighting methods, In M. Bates, editor, Human Language Technology. Morgan Kaufman: 1993.

[6] Longzhuang Li, Yi Shang A new statistical method for performance evaluation of search engines. ICTAI: 2000.

[7] Longzhuang Li, Yi Shang A new method for automatic performance comparison of search engines. World Wide Web: 2000.

[8] Chu, H. & Rosenthal: "Search engines for the World Wide Web: A comparative study and evaluation methodology". In Proceedings of the 59th Annual Meeting of the American Society for Information Science, Baltimore, 1996.

[9] Gerald Salton and Chris Buckley. Term weighting approaches in automatic text retrieval. Information Processing and Management, 24(5): Is-sue 5. 1988.

[10] G. Salton and C. Buckley, "Improving Retrieval Performance by Relevance Feedback," J. Amer. Soc.for Information Science, Vol. 41, No. 4, 199

[11] Jitendra Nath Singh & Sanjay Kumar Dwivedi: Analysis of Vector Space Model in Information Retrieval.

Proceedings (IJCA) on National Conference on Communication Technologies & its impact on Next Generation Computing 2012 CTNGC (2):14-18:2012.

[12] Lee, D.l.; Huei Chuang; Seamons, K.; "Document ranking and the Vector-space model," Software, IEEE, vol.14, no.2, Pp.67-75, Mar/April.1997