Proceedings of ***BITCON-2015*** *Innovations For National Development*
National Conference on : Information Technology Empowering Digital India

Research Paper

VECTOR SPACE MODEL: AN INFORMATION RETRIEVAL SYSTEM

Vaibhav Kant Singh, Vinay Kumar Singh

Address for Correspondence

[1] Assistant Professor in the Department of Computer Science & Engineering , Institute of Technology, Guru Ghasidas Vishwavidyalaya, Central University, Bilaspur, Chhattisgarh, India

[2]Senior Assistant Programmer, in the Department of Development, Guru Ghasidas Vishwavidyalaya, Central University, Bilaspur, Chhattisgarh, India

ABSTRACT

In this paper we will be Examining the Vector Space Model, an Information Retrieval technique and its variation. The rapid growth of World Wide Web and the abundance of documents and different forms of information available on it, has recorded the need for good Information Retrieval technique. The Vector Space Model is an algebraic model used for Information Retrieval. It represent natural language document in a formal manner by the use of vectors in a multi-dimensional space, and allows decisions to be made as to which documents are similar to each other and to the queries fired. This paper attempts to examine the Vector Space Model, an Information Retrieval Technique that is widely used today. It also explains existing variation of VSM and proposes the new variation that should be considered.

KEYWORDS Vector Space Model, Information Retrieval, Stop words, Term weighing, Inverse Document Frequency, Stemming.

1.0 INTRODUCTION

The growth of E-libraries with the advancement in computing technology as a result of growth of Internet use has made Electronic data as the major source for the extraction of useful information for the field of Research. It has created a platform for simulation of decision support systems. The Information Retrieval Systems aims at finding right kind of information from documents, unidentified pieces of information from the information base and also tries to search for hidden relationships among the informations retrieved.

It has been seen that information can be stored both in structured and unstructured manner. For Example it can be stored in form of Database systems representing structuring of data whereas storage of HTML and Text files show unstructured behavior.

The art and science of searching for information in documents, searching for documents themselves, searching for metadata which describes documents, or searching within databases for text, sound, images or data is Information Retrieval.

A good IR system should be able to perform quickly and successfully the following process:

- Accept a user query,
- Understand from the user query what the user requires,
- Search a database for relevant documents,
- Retrieve the documents to the user, and
- Rank them in an order based on the relevance of the document to the users initial query.

An Information Retrieval model[Baeza] is a quadruple (D, Q, F, R) where

- D is a set of representations for the documents in the collection
- Q is a set of representations for the user information needs (queries)
- F is a framework for modeling document representations, queries, and their relationships
- R: Q → D (R) is a ranking function which associates a real number with a query qi of Q and document representation dj of D.

Major Information Retrieval Techniques are:

- Boolean Matching/Model
- Extended Boolean Model
- Vector Space Model
- Probabilistic Modeling
- Fuzzy Set Model
- Latent Semantic Indexing
- Neural Networks

2.0 VECTOR SPACE MODEL

The VSM is an algebraic model used for Information Retrieval. It represent natural language document in a formal manner by the use of vectors in a multi-dimensional space.

The Vector Space Model (VSM) is a way of representing documents through the words that they contain. The concepts behind vector space modeling are that by placing terms, documents, and queries in a term-document space it is possible to compute the similarities between queries and the terms or documents, and allow the results of the computation to be ranked according to the similarity measure between them. The VSM allows decisions to be made about which documents are similar to each other and to queries

3.0 Experiment

(a) How it works

- Each document is broken down into a word frequency table
- The tables are called vectors and can be stored as arrays
- A vocabulary is built from all the words in all documents in the system
- Each document and user query is represented as a vector based against the vocabulary
- Calculating similarity measure
- Ranking the documents for relevance

The vector space model provide the user with a guide to documents that might be more similar and of greater significance by calculating the distance or angle measure between the query and terms or document. Vector space modeling is based on the assumption that the meaning of a document can be understood from the document's constituent terms. Documents are represented as "vectors of terms d = (t1, t2, … tn) where ti (1<= i <= t) is a non-negative value denoting the single or multiple occurrences of term **i** in document **d**." [URL:6]. Each unique term in

the document represents a dimension in the space. "Similarly, a query is represented as a vector Q = (t1, t2, …, tn) where term ti (1<=i<=n) is a non-negative value denoting the number of occurrences of ti (or, merely a **1** to signify the occurrence of term) in the query". Once both the documents and query have their respective vectors calculated it is possible to calculate the distance between the objects in the space and the query, allowing objects with similar semantic content to the query should be retrieved. Vector space models that don't calculate the distance between the objects within the space treat each term independently. Using various similarity measures it is possible to compare queries to terms and documents in order to emphasize or de-emphasize properties of the document collection. A good example of this is, "the dot product (or, inner product) similarity measure finds the Euclidean distance between the query and a term or document in the space".

Consider the following two documents

- Document A: "A man and a woman."
- Document B: "A baby."

Step-1: Each document is broken down into a word frequency table

Document A: "A man and a woman."

A	Man	And	Woman
2	1	1	1

Document B: "A baby."

A	Baby
1	1

The tables are called vectors and can be stored as arrays

Step-2: A vocabulary is built from all the words in all documents in the system

The vocabulary contains all words used: a, man, and, woman, baby

Step-3: The vocabulary needs to be sorted: a, and, man, woman, baby

Step-4: Each document is represented as a vector based against the vocabulary

Vector for Document A: "A man and a woman."

A	And	Man	Woman	Baby
2	1	1	1	0

Vector: (2,1,1,1,0)

Vector for Document B: "A baby."

A	And	Man	Woman	Baby
1	0	0	0	1

Vector: (1,0,0,0,1)

Step-5: Queries can be represented as vectors in the same way as documents

For example, Woman = (0,0,0,1,0)

(b) Similarity measures/coefficient

Using a similarity measure, a set of documents can be compared to a query and the most similar documents are returned. The similarity in VSM is determined by using associative coefficients based on the inner product of the document vector and query vector, where word overlap indicates similarity.

There are many different ways to measure how similar two vectors are, like Inner Product, Cosine Measure, Dice Coefficient, Jaccard Coefficient.

The most popular similarity measure is the cosine coefficient, which measures the angel between a document vector and query vector.

(c) The cosine measure

The cosine measure calculates the angle between the vectors in a high-dimensional virtual space

For two vectors d and d' the cosine similarity between d and d' is given by:

(D * D') / | D | * | D' |

Here d X d' is the vector product of d and d', calculated by multiplying corresponding frequencies together

Step-5: Calculate the similarity measure of query with every document in the collection

For Document A, d = (2,1,1,1,0) and d' = (0,0,0,1,0)

dXd' = 2X0 + 1X0 + 1X0 + 1X1 + 0X0=1

$|d| = \text{sqrt}(2^2+1^2+1^2+1^2+0^2) = \text{sqrt}(7)=2.646$

$|d'| = \text{sqrt}(0^2+0^2+0^2+1^2+0^2) = \text{sqrt}(1)=1$

Similarity = 1/(1 X 2.646) = 0.378

For Document B, d = (1,0,0,0,1) and d' = (0,0,0,1,0)

dXd' = 1X0 + 0X0 + 0X0 + 0X1 + 1X0=0

$|d| = \text{sqrt}(1^2+0^2+0^2+0^2+1^2) = \text{sqrt}(2)=1.414$

$|d'| = \text{sqrt}(0^2+0^2+0^2+1^2+0^2) = \text{sqrt}(1)=1$

Similarity = 0/(1 X 1.414) = 0

(d) Ranking documents

- A user enters a query
- The query is compared to all documents using a similarity measure
- The user is shown the documents in decreasing order of similarity to the query term

Step-6: Rank the in descending order and display to user

Document A	0.378
Document B	0

3.1 Variation in VSM

(e) Stop words

- Commonly occurring words are unlikely to give useful information and may be removed from the vocabulary to speed processing
- Stop word lists contain frequent words to be excluded
- The top 20 Stop words according to their average frequency per 1000 words: The, of, and, to, a, in, that, is, was, he, for, it, with, as, his, as, on, be, at, by, I, etc. [4]

3.1.1 Term weighting

- Not all words are equally useful.
- A word is most likely to be highly relevant to document A if it is: Infrequent in other documents and Frequent in document A
- The cosine measure needs to be modified to reflect this
- Considering the frequency of every word in a document can do this.
- It is given by tf = log (1 + n(d, t) / n(d)), where n(d, t) is number of occurrences of the term t in document d; and n(d) is the number of terms in document d.

Rank	Word	Per 1000	Rank	Word	Per 1000
1	the	70	11	for	9
2	of	36	12	it	9
3	and	29	13	with	7
4	to	26	14	as	7
5	a	23	15	his	7
6	in	21	16	on	7
7	that	11	17	be	6
8	is	10	18	at	5
9	was	10	19	by	5
10	he	10	20	I	5

3.1.2 Normalized term frequency (tf)

- A normalized measure of the importance of a word to a document is its frequency, divided by the maximum frequency of any term in the document
- This is known as the tf factor.
- Document A: raw frequency vector: (2,1,1,1,0), tf vector: (1, 0.5, 0.5, 0.5, 0)
- This stops large documents from scoring higher

3.1.3 Inverse document frequency (idf)

- A calculation designed to make rare words more important than common words
- The idf of word i is given by idf(i) = log (n / n(i)), Where N is the number of documents and n(i) is the number that contain word i
- IDF provides high values for rare words and low values for common words

3.1.4 tf-idf

- The tf-idf weighting scheme is to multiply each word in each document by its tf factor and idf factor
- Different schemes are usually used for query vectors
- Different variants of tf-idf are also used
- Increases with the number of occurrences *within* a doc
- Increases with the rarity of the term *across* the whole corpus.

(f) Stemming

Stemming is the process of removing suffixes from words to get the common origin. In statistical analysis, it greatly helps when comparing texts to be able to identify words with a common meaning and form as being identical. For example, we would like to count the words stopped and stopping as being the same and derived from stop. Stemming identifies these common forms.

(g) Synonyms and Multiple meaning of word

- There are many ways to describe something. For example, car and automobile may describe the same thing.
- Words often have multiple meanings.

(h) Concept based VSM

It considers the semantic of the document instead of only considering the terms contained in document.

(i) Proximity

If the terms occur close to each other in the document, the document would be ranked higher than if they occur far apart.

The words “Information” and “Retrieval” that comes together defines a document on “Information Retrieval” than the document containing two words “Information” and “Retrieval” scattered.

(j) File Attribute

For temporal data, the file attributes, like last modified date, plays an important role in deciding the document relevance.

(k) Hyperlink

The hyperlink coming in the page and going out of the page (particularly IN LINK) can give useful information about page.

(l) Position of word

The word in header (for example, HTML tag) can be considered more content bearing word than word in Body of document. Similarly, the word occurring in the beginning of document can be more indicative about document content than word coming later in the document; or the word found nearer to word ‘Abstract’ may define the content of document well.

(m) User Profile

User profile can be used to improve the process using adaptive approach. The response of the user can also be considered to improve the process. This can be possible if and only if we can measure trustworthiness (dependent variable) of user response based on some dependent variables like education, age, gender, income, number of time he visited page, etc.

(n) User defined weight to terms in query

There can be some way by which user can provide weight to his/her query.

3.2 Major Problems with VSM

- There is no real theoretical basis for the assumption of a term space
- It is more for visualization that having any real basis
- Most similarity measures work about the same regardless of model
- Terms are not really orthogonal dimensions
- Terms are not independent of all other terms

CONCLUSION

From the above findings it has been observed that the usefulness of Information Retrieval System is going to increase in near future is going to increase from what it is at this point because with the advancement of hardware technology and growth of Internet we are having huge amount of data having potential to produce unused information that too when we are having the Hardware Technology to handle the problem in an efficient manner

In this work we have given an insider to the working of Vector Space Model Techniques used for efficient retrieval techniques. It is bare fact that each systems has its own strengths and weaknesses what we have sort out in our work for Vector Space Model is the model is Easy to Understand, cheaper to implement considering to the fact that the system should be cost effective i.e. should follow the Space/Time constraint. Also very popular. Although the system is having all these properties it is facing some major drawbacks. The Drawbacks are the system yields no theoretical findings, Weights associated with the vectors are very arbitrary and this system is an independent system. Thus require separate attention. Though, it is promising technique, current level of success of the Vector Space Model techniques used for Information Retrieval is not able to satisfy user needs and need extensive attention.

REFERENCES

1. A.B.Singh, G.M. Mollick, Vaibhav Kant Singh “A Production System Approach Acquisition on E-Governance” , *Proceeding of Conference on Datamining and E-Governance,GGU Bilaspur*,Feb 2007.
2. V.K. Singh, Vinay Kumar Singh “Rural Development Using concept of Data mining” Tribal Situation In Middle India: Development and vision, Bilaspur, Oct 2007.
3. Vinay Kumar Singh, Vaibhav Kant Singh “The Huge Potential of Information Technology” Proceeding of Global Leadership: Strategies and Challenges for Indian Business, Department of Management Studies GGU, Bilaspur, Feb 2007.
4. Abraham Silberschatz, Henry F. Korth, and S. Sdarshan, “Database System Concepts”, 4th edition, McGraw-Hill.

Note: This Paper/Article is scrutinised and reviewed by Scientific Committee, BITCON-2015, BIT, Durg, CG, India