ISSN No. 0976-5697

Volume 4, No. 10, Sep-Oct 2013

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

# Ontology Based Information Retrieval Using Vector Space Model

Nayana S. Zope
(Lecturer): dept. of Information Technology
GF's Godavari Collage of Engineering
Jalgaon, India
nayanazope24@gmail.com

Ankita K. Kolhe
B.E Information Technology
Pune University
kolheankita15@gmail.com

Prashant Bharambe
Senior QA Engineer
BOARDVantage Menlo park, CA USA
bharambeprashant143@gmail.com

***Abstract:*** Information retrieval (IR) is the science of searching for documents, for information within documents and for metadata about documents, as well as that of searching relational databases and the World Wide Web. In this paper, after a brief review on ranking models, a new ontology based approach for ranking HTML/TXT documents is proposed and evaluated in various circumstances. Our approach is applying the vector space model method. Increasing growth of information volume in the internet causes an increasing need to develop new semi) automatic methods for retrieval of documents and ranking them according to their relevance to the user query. This combination reserves the precision of ranking without losing the speed. Our approach exploits natural language processing techniques for extracting phrases and stemming words. The annotated documents and the expanded query will be processed to compute the relevance degree exploiting statistical methods. The outstanding features of our approach are (1) combining HTML, TXT, PDF documents, (2) finding frequency of each and every word, (3) removing stop keywords, (4) applying porter stemming algorithm, to remove the suffix of every word and (5) allowing input variable document using vector dimensions. A ranking system called Information Retrieval using Vector Space Model (IRVSM) is developed to implement and test the proposed model.

***Keywords:*** Ontology, Parsing, Indexing, Stemming, Vector Space Model, Document Ranking.

## I. INTRODAUCTION

Ontology formally represents knowledge as a set of concepts within a domain, and the relationships between those concepts. The main reason is that the semantic of documents is not recognized correctly and users do not express their information needs clearly. Information retrieval is a sub-field of computer science that deals with the automated storage and retrieval of documents. In this paper such operations as lexical analysis and stop lists, stemming algorithms, thesaurus construction, and relevance feedback and other query modification techniques. It provides information on VSM (Vector Space Model) operations, Porter Stemming algorithms, ranking algorithms. Information science and library science professionals are interested in text retrieval technology, refer as [1] [2] [3].

Information Retrieval (IR) our basic task is to find the subset of a collection of elements that is relevant to a query. An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy. Retrieval models that would capture relevance very well, but are computationally prohibitively expensive are not suitable for an information retrieval system. There is overlap in the usage of the terms data retrieval, document retrieval, information retrieval, and text retrieval, but each also has its own body of literature, theory, praxis and technologies. Information retrieval is an activity, and like most activities it has a purpose. A user of a search engine begins with an information need, which he or she realizes as a query in order to find relevant documents. This query may not be the best articulation of that need, or the best bait to use in a particular document pool. It may contain misspelled, misused, or poorly selected words. It may contain too many words or not enough. Nevertheless, it is usually the only clue that the search engine has concerning the user's goal, refer as [4] [5].

## II. DOCUMENT RETRIEVAL

Document retrieval is defined as the matching of some stated user query against a set of free-text records. These records could be any type of mainly unstructured text, such as newspaper articles, real estate records or paragraphs in a manual. User queries can range from multi-sentence full descriptions of an information need to a few words.

Document retrieval is sometimes referred to as, or as a branch of, Text Retrieval. Text retrieval is a branch of information retrieval where the information is stored primarily in the form of text. Text databases became decentralized thanks to the personal computer and the CD-ROM. Text retrieval is a critical area of study today, since it is the fundamental basis of all Internet engines, refer as [6].

## III. VECTOR SPACE MODEL

Vector space model (or term vector model) is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms. It is used in information filtering,

information retrieval, indexing and relevancy rankings. Its first use was in the SMART Information Retrieval System.

Working of Vector Space Model

The vector space model procedure can be divided in to 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 document relevant to the user. The last stage ranks the document with respect to the query according to a similarity measure. The vector space model has been criticized for being ad-hoc for a more theoretical analysis of the vector space model.

### *A. Boolean Vector Space Model:*

Text documents can be conveniently represented in a high-dimensional vector space where terms are associated with vector components. More precisely, a text document can be represented as a sequence of terms, d= (ω (1), ω (2)… ω (|d|)), where |d| is the length of the document and ω (t) ∈ V. A vector-space representation of d is then defined as a real vector, where each component is a statistic related to the occurrence of the $j^{th}$ vocabulary entry in the document. The simplest vector-based representation is Boolean, i.e. $x_j \in \{0, 1\}$, indicates the presence or the absence of term $\omega_j$ in the document being represented. In the Fig [1] is represented a vector-space documentation in the Boolean model and in the Term-Weighted model.

| 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0……...…….. | 0 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|

Figure 1. Document Vector- Boolean Model.

Vector-based representations are sometimes referred to as a 'bag of words' , emphasizing that document vectors are invariant with respect to term permutations, since the original word order ω (1), . . . ω (|v|) is clearly lost, , refer as [7].

Representations of this kind are appealing for their simplicity. Moreover, although they are necessarily lossy from an information theoretic point of view, many text retrieval and categorization tasks can be performed quite well in practice using the vector-space model. Note that typically the total number of terms in a set of documents is much larger than the number of distinct terms in any single document, |V|>>|d|, so that vector-space representations tend to be very sparse. This property can be advantageously exploited for both memory storage and algorithm design.

### *B. Term Weighted Vector Space Model:*

In Boolean vector models each coordinate of a document vector is zero (when the corresponding attribute is absent) or unity (when the corresponding attribute is present). Term weighting is a widely used refinement of Boolean models that takes into account the frequency of appearance of attributes (such as keywords and key phrases), the total frequency of appearance of each attribute in the document set, and the location of appearance (e.g., in the title, section header, abstract, or text).

An important family of weighting schemes combines term frequencies (which are relative to each document) with an absolute measure of term importance called inverse document frequency (IDF). IDF decreases as the number of documents in which the term occur increases in a given collection. So terms that are globally rare receive a higher weight.

Formally, let D={$d_1$, $d_2$…$d_n$} be a collection of documents and for each term $\omega_j$ let $n_{ij}$ denote the number of occurrences of $\omega_j$ in $d_i$ and $n_j$ the number of documents that contain $\omega_i$ at least once. Then we define

$$TF_{ij} = \frac{n_{ij}}{|d_i|},\ IDF_j = \log\frac{n_j}{n}$$

Here the logarithmic function is employed as a damping factor.

The TF-IDF weight of j $\omega_i$ in $d_i$ can be computed as

$$x_{ij} = TF_{ij} \cdot IDF_j$$

Or, alternatively, as

$$x_{ij} = \frac{TF_{ij}}{\max_{\omega_k \in d_i} TF_{ik}} \cdot \frac{IDF_j}{\max_{\omega_k \in d_i} IDF_k}$$

### *C. Similarity Coefficients:*

The similarity in vector space models is determined by using associative coefficients based on the inner product of the document vector and query vector, where word overlap indicates similarity. The inner product is usually normalized. The most popular similarity measure is the cosine coefficient, which measures the angle between the document vector and the query vector.

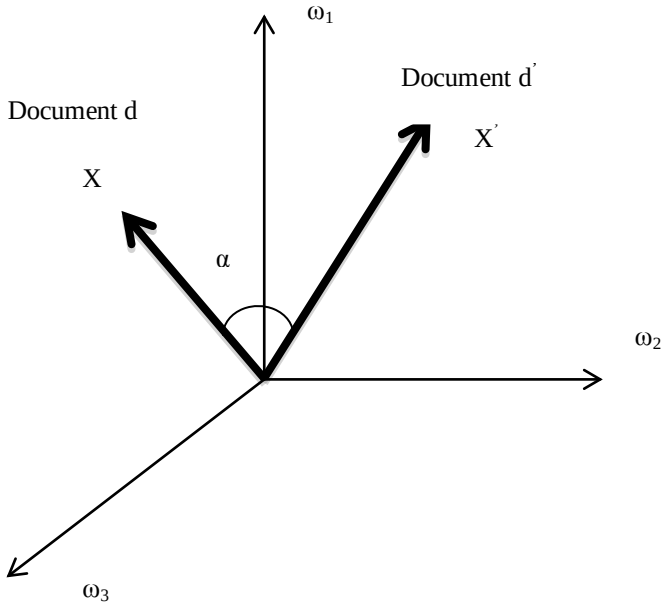


Figure 2. Cosine measure of document similarity

Document retrieval is now accomplished by computing the similarity between a query vector, q, and a document vector, d, represented Fig [2] (this measure is simply the cosine of the angle formed by the vector-space representations of the two documents, d and q) using the formula:

$$sim(q,d) = \frac{\sum_t x_{t,d} \cdot x_{t,q}}{\sqrt{\sum_t x_{t,d}^2}\sqrt{x_{t,q}^2}}$$

& then ranking the found documents in decreasing order with respect to this measure, refer as [8] [9].

## IV. ONTOLOGY PROCESSOR

The ontology processor is responsible for assigning weight to ontology's links. This weighted ontology is used in the query processing.

In our proposed model the weight of an ontology link is computed by multiplication of similarity measure and specificity measure. Similarity measure of each link (relation) indicates the similarity between two related concepts $C_j$ and $C_k$ and is computed by equation. The idea behind this measure is that two concepts will be similar if they are related to same concepts.

$$W(C_j, C_k) = \frac{\sum_{i=1}^{m} n_{i,j,k}}{\sum_{i=1}^{m} n_{i,j}}$$

In this equation $n_{i,j}$, is the number of related concepts to concept $c_j$ by relation i (the sum of in-degree and out-degree of $c_j$ according to relation (i) and m is the number of selected relations. So $\sum_{i=1}^{m} n_{i,j}$, is the total number of related concepts to concept instance $c_j$ and $\sum_{i=1}^{m} n_{i,j}$, is the total number of related concepts to both concept instances $c_j$ and $c_k$.

Ontology provides a common vocabulary for researchers who need to share information in the domain. some of the reasons to create ontology are: to share common understanding of the structure of information among people or software agents. A ranking algorithm calculates the similarity degree of each document to user query. Then documents are sorted by this degree and will be presented to the user. Ranking algorithms exploit different information to estimate the similarity degree. Most conventional algorithms are keyword based and use statistical information such as term frequency, document length, etc. to calculate the relevance degree. By the creation of the web, ranking algorithms apply hyperlink structures too. This problem has caused the conceptual approaches to appear in the recent years. These approaches try to extract and compare the concepts of the documents and the query refer as [10] [11].

## V. CONCLUSION

In this paper, we introduced an ontology-based model for ranking documents according to their relevancy to the user's query. The proposed model improves the precision of existing statistical models using concept instances in the document and query's vectors instead of words. Therefore, the relevancy degree of the retrieved document is increased. To complete this effort following improvements are proposed as further works: Ontology's classes usually have comment property. During query expansion, it is possible to search query's phrases and words in the comment property. If they were found, that class would have expanded the query. Document annotation and query expansion accomplish by applying ontology. So using special purpose ontologies would have great effect in test results. As our model depends on annotation, designing an appropriate annotation algorithm increases the relevancy of retrieved documents. More precise approximation of phrases' weight coefficient in comparison to words' requires more tests. For computing weight coefficient of relations in improved SA algorithm, it is possible to implement a system based on relevance feedback.

## VI. REFERENCES


[1] LANCASTER, F.W., Information Retrieval Systems: Characteristics, Testing and Evaluation, Wiley, New York (1968).

[2] Frakes, W.B. (1984) "Term Conflation for Information Retrieval", Cambridge University Press Published.

[3] Frakes, W.B. & Baeza-Yates, R (1992) "Information Retrieval: Data Structures and Algorithms", Englewood Cliffs, NJ, Prentice Hall.

[4] C.H. Chang and C.C. Hsu, ªExploiting Hyperlinks for Automatic Information Discovery on the WWW,º Proc. 10th IEEE Int'l Conf. Tools with Artificial Intelligence,Chien Tan Youth Activity Center, Taipei, Taiwan, Nov. 1998.

[5] Feiyue Ye ; Sch. of Comput. Eng. & Sci., Shanghai Univ., Shanghai, China ; Hongxin Cao ; Xiangfeng Luo "A Text Representation and Retrieval Method Based on Concept Algebra" with Computer and Information Technology (CIT), 2012 IEEE 12th International Conference on 27-29 Oct. 2012.

[6] Gautam, D. ; Dept. of Comput. Sci. & Eng., Chosun Univ., Gwangju ; Miyoung Cho ; Pankoo Kim ," Document Retrieval Based on Key Information of Sentence" Published in: Advanced Communication Technology, 2008. ICACT 2008. 10th International Conference on (Volume:3 )on 17-20 Feb. 2008.

[7] Dawson, J.L. (1974) "Suffix Removal for Word Conflation, Bulletin of the Association for Literary and Linguistic Computing, 2(3): 33-46.

[8] Wibowo, A. ; Inf. Dept., Petra Christian Univ., Surabaya, Indonesia ; Handojo, A. ; Halim, A. "Application of Topic Based Vector Space Model with WordNet" Published in: Uncertainty Reasoning and Knowledge Engineering (URKE), 2011 International Conference on (Volume:1 ) on 4-7 Aug. 2011.

[9] Kannan, P. ; Dept. of Comput. Sci., Pondicherry Univ., Pondicherry, India ; Bala, P.S. ; Aghila, G. "A comparative study of multimedia retrieval using ontology for semantic web" published in Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on 30-31 March 2012.

[10] Guo Chengxia ; Sch. of Inf., Shanghai Ocean Univ., Shanghai, China ; Huang Dongmei "Research on Domain Ontology Based Information Retrieval Model" Published in: Intelligent Ubiquitous Computing and Education, 2009 International Symposium on 15-16 May 2009.

[11] Dridi, O. ; Nat. Sch. of Comput. Sci., RIADI Lab., Univ. of Manouba, Manouba,"Ontology-based information retrieval: Overview and new proposition" Published in: Research Challenges in Information Science, 2008. RCIS 2008. Second International Conference on 3-6 June 2008.