Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 3, March 2016, pg.287 – 290

Document Ranking using Customizes Vector Method – A Review

Priyanka Mesariya[1], Nidhi Madia[2], Abhishek Kumar[3]

[1]Department of Computer Engineering & Gujarat Technology University, India
[2]Department of Information & Technology & Gujarat Technology University, India
[3]Computer Science INFLIBNET, UGC-MHRD Gandhinagar, Gujarat, India
[1] priyanka.mesariya@gmail.com; [2] nidhimadia.ce@socet.edu.in; [3] abhishek@inflibnet.ac.in

Abstract— Information retrieval (IR) system is about positioning reports utilizing client's question and get the important records from extensive dataset. Archive positioning is fundamentally looking the pertinent record as per their rank. Document ranking is basically search the relevant document according to their rank. Vector space model is traditional and widely applied information retrieval models to rank the web page based on similarity values. Term weighting schemes are the significant of an information retrieval system and it is query used in document ranking.

Keywords— "Information Retrieval", "Term Frequency", "Inverse Frequency", " Vector Space Model", "Cosine Similarity".

I. INTRODUCTION

In the information retrieval (IR) system documents are ranked optimally by using user's query to find out the relevant documents from large data base or form dataset [21].When the user gives a query, the index is consulted to archives the most relevant documents. The relevant documents are then ranked significance of their degree of relevance. Majority of internet users rely on search engines for extracting information by providing a query from any walk of life. These queries are processed by the search engines and a certain information retrieval or mining algorithm is applied to obtain the cluster of documents related to the query. After the retrieval of these documents, an important task is to present these documents in a list where documents at the top are the ones considered more relevant for the user. This task is called ranking of documents [15].Information retrieval system is a set of documents to discover convenient information equivalent to a user's query. In information retrieval basically data can be fetching from web structure information that can be type of content, pictures, graph etc. Several components make this task challenging :(i) normally unstructured information is in document database; (ii) reports are typically composed in unconstrained characteristic dialect; iii) regularly, the documents cover extensive variety range of subjects.

II. INFORMATION RETRIEVAL MODELS

A. Boolean Model:

This model consist documents and queries are indicate as set of index terms. The advantage of this model is its simplicity. The Boolean model allows for the use of logical operators of Boolean algebra, AND, OR and NOT for query formulation. It has major drawback that The Boolean model suffers with too many documents retrieval during exact matching and system can't get ranked list of documents. The retrieval function in this model treats a document as either relevant or irrelevant [10].

B. Vector Space Model:

Vector-Space model Due to partial matches under the Vector-Space model, it gives better results from the Boolean model of information retrieval [3]. The vector space model can best be characterized by its endeavour to rank documents by the equivalence between the query and each document [7]. Its term-weighting scheme improves retrieval performance and gives the scoring and ranking of the document's which is relevant to user query. Its partial matching strategy allows retrieval of documents that approximate the query conditions .Its cosine ranking formula archives the documents according to their degree of sameness to the query [10].

C. Probabilistic Model:

The Probabilistic model archives in which almost correct match of documents is found [13] the most important characteristic of the probabilistic model is its attempt to rank documents by their probability of relevance given a query [16].disadvantages of this model is the need to guess the initial relevant and non relevant sets of Term frequency is not considered unconventional suspicion for index terms [10].

III.VECTOR SPACE MODEL

It is a model for representing text documents or any other items as vectors of identifiers [17]. It is utilized as a part of information filtering, information retrieval, indexing and relevancy rankings. relevance rankings of documents in a keyword search can be calculated, using the suppositions of document equivalence theory, by comparing the deviation of angles between each document vector and the main query vector where the query is represented as the equivalent of vector as the documents. The vector space model technique can be partitioned into three stages. The main stage is the document indexing where content relevance 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 of the documents archives as per the query comparability value [4] [7]. Index terms are assigned positive and non-binary weights Documents and queries are shows as vector

$$d_j = (w_{1,j}, w_{2,j}, \ldots . w_{t,j})$$
$$q = (w_{1,q}, w_{2,q}, \ldots . w_{n,q},) \qquad (1)$$

A. Term –Count Model

This model required database collection to retrieve documents, input query and index term. The terms are single words or keywords. If words are select as terms, the dimensionality of the vector is the large amount of words in the vocabulary. Relevance ranking of documents in a keyword search can be calculated using the inference of document equivalence [7]. In Term- count model weight of terms has been computed using Term frequency given by.

Weight = W(i,j)= TF(i,j)
Where TF(i,j)= Frequency of term j in documents i.

B. Tf-Idf Vector Space Model

In Information retrieval Tf-Idf is known as term frequency and inverse document frequency. It is a common method to assess how a word is required a document. It is commonly used as weighting factor in information retrieval. Tf-Idf is also a very interesting method to convert the textual representation of information into a Vector Space Model (VSM). The weight of term in document vector can be determined using method [14]. The weight of term is measured sometimes term j obtain in the document i (the term frequency) and tdf (the inverse document frequency) [7]. The weight of a term j in the document i is given by

$$w_{i,j} = freq_{i,j} \times idf$$
And $idf\ i = \log \frac{N}{n_i}$ (2)

C. Cosine similarity

Cosine similarity could be of equivalence of two vectors of associate scalar product area that measures the

circular function of the angle between them. This method apply on information retrieval and text mining formula for this :

$$\text{Cos } (\theta) = \frac{A.B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \qquad (3)$$

Where A_i *and* B_i are components of vector *A* and *B* respectively

IV. RELETED WORK

In 2013 the researcher Jiaul H. Paik is represents a novel TF-IDF term weighting scheme. The suggest term weighting scheme has two feature of within document term frequency assign to discover the importance of a term. In that Experiments done at the huge amount of TREC news and web collection data and proposed that the out performs five state of the art retrieval model is significance and consistent. And its shows that proposed model better than the existing models [6].

T.Suganya and M.Ravichandran proposed a method in e-learning rank ordering the documents according to their individual term relevance degree using possibility approach and vector-based technique method. This proposed system provides highly relevant learning materials to the learner and it recommends the items based on individual term relevance with respect to the query specified by the user [5].

In proposed vector space model used in XML document ranking suggest by Weimin He and Teng Lv. They proposed effectively rank Xml document and also differentiate the framework with Lucene to demonstrate their extended TF*IDF is successful and it is effective ranking than existing XML search engine Lucene [15].

Premalatha.R and Srinivasan.S emphasis, on Information retrieval for Tamil literary document using the model vector space. they approaches in text processing in information retrieval. In their system explore that can be divided into three categories, Main topic search, Subtitle search and Keyword search. So the system would explore necessitate information rapidly mainly using the vector space model, that illustrate documents as vectors. It would be applicable for all Tamil literates and understudies to look and learn [2]

In their research Vaibhav Kant Singh and Vinay Kumar Singh is proposed a vector space model using in information retrieval .In that individual document and user query is represented as a vector based against the vocabulary and Calculating similarity measure and than Ranking the documents for relevance and other variant of VSM that Term weighting, Normalized term frequency(tf) and Inverse document frequency (idf) is shows in their system[1].

Bo Yu and Guoray Cai is recommend a dynamic document ranking scheme join thematic and geographic pertinence measures on a for each query premise. They have been using Dempster-Shafer's theory to gather the two different sources of ranking verification and evaluate the different web document data set. Which can be either news stories or blog and it can be fetch from the web data [19].

Dik Lun Lee, Huei Chuang and Kent E.Seamons is proposed that Using various interpretation of the vector-space model for text retrieval queries, they optimal balance between processing efficiency and retrieval effectiveness as expressed in relevant document rankings. They using six different vector method and their retrieval effectiveness [20]

V. CONCLUSIONS

In this paper we conclude that on the basis of query document ranking is utilized for search relevant document so that information retrieval is a process of searching and retrieving the knowledge based information from collection of documents. for that their distinctive model is used with its advantages. Vector space model is used in information filtering, information retrieval and relevancy ranking of documents. Tf-Idf model in such circumstance where the query terms are presented in each document also provided similar results.

REFERENCES

[1] Singh, Vaibhav Kant, and Vinay Kumar Singh. "VECTOR SPACE MODEL: AN INFORMATION RETRIEVAL SYSTEM." Int. J. Adv. Engg. Res. Studies/IV/II/Jan.-March 141 (2015): 143.

[2] Premalatha, R., and S. Srinivasan. "Text processing in information retrieval system using vector space model." Information Communication and Embedded Systems (ICICES), 2014 International Conference on. IEEE, 2014

[3] Khan, Junaid. "Comparative study of information retrieval models used in search engine." Advances in Engineering and Technology Research (ICAETR), 2014 International Conference on. IEEE, 2014.

[4] Singh, Jitendra Nath, and Sanjay K. Dwivedi. "Comparative Analysis of IDF Methods to Determine Word Relevance in Web Document." International Journal of Computer Science Issues (IJCSI) vol 11 (2014): 59-65.

[5] T. Suganya and M. Ravichandran, "Ranking Documents in IR Using Vector Based Ordering In E-Learning," vol. 3, no. 3, 2014.
[6] Paik, Jiaul H. "A novel TF-IDF weighting scheme for effective ranking."Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013.
[7] J. N. Singh, "A Comparative Study on Approaches of Vector Space Model in Information Retrieval," pp. 37–40, 2013
[8] Asadi, Nima, and Jimmy Lin. "Document vector representations for feature extraction in multi-stage document ranking." Information retrieval 16.6 Springer (2013): 747-768.
[9] Bhatia, Parul Kalra, Tanya Mathur, and Tanaya Gupta. "Survey Paper on Information Retrieval Algorithms and Personalized Information Retrieval Concept." International Journal of Computer Applications 66.6 (2013).
[10] Sharma, Manish, and Rahul Patel. "A Survey on Information Retrieval Models, Techniques And Applications." International Journal of Emerging Technology and Advanced Engineering, ISSN (2013): 2250-2459.
[11] Sharma, Aditi, Nishtha Adhao, and Anju Mishra. "A survey: Static and dynamic ranking." International Journal of Computer Applications 70.14 (2013): 7-12.
[12] Wang, Shuaiqiang, et al. "Adapting vector space model to ranking-based collaborative filtering." Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012.
[13] Raman, Shivangi, Vijay Kumar Chaurasiya, and Swaminathan Venkatesan. "Performance comparison of various information retrieval models used in search engines." Communication, Information & Computing Technology (ICCICT), 2012 International Conference on. IEEE, 2012.
[14] Zhang, GuanHong. "Sentence alignment for web page text based on vector space model." Computer Science and Information Processing (CSIP), 2012 International Conference on. IEEE, 2012.
[15] He, Weimin, and Teng Lv. "Extending vector space model for XML ranking."Applications of Digital Information and Web Technologies (ICADIWT), 2011 Fourth International Conference on the. IEEE, 2011.
[16] Barrera, Araly, and Rakesh Verma. "A ranking-based approach for multiple-document information extraction." University of Houston (2010).
[17] Khankasikam, Krisda. "A comparison of information retrieval models applied to Thai digital library." Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on. Vol. 1. IEEE, 2010.
[18] Xu, Huaiyu, et al. "An intelligent project agency for web-3D virtual trading community based on google earth." Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on. IEEE, 2009.
[19] Yu, Bo, and Guoray Cai. "A query-aware document ranking method for geographic information retrieval." Proceedings of the 4th ACM workshop on Geographical information retrieval. ACM, 2007.
[20] Dik L. Lee, Huei Chuang, Kent Seamons, "Document Ranking and the Vector-Space Model", IEEE Software vol.14, no. 2, pp. 67-75, March/April 1997, doi:10.1109/52.582976
[21] Salton, Gerard, Anita Wong, and Chung-Shu Yang. "A vector space model for automatic indexing." Communications of the ACM 18.11 (1975): 613-620.
[22] R. Azhar-ul-haq, "A Review :Ranking documents using Ranking Algorithms & Techniques."

