International Journal of Hybrid Information Technology
Vol.12, No.2 (2019), pp.13-20
http://dx.doi.org/10.21742/ijhit.2019.7.1.03

An Efficient Algorithm for Informational Retrieval using Web Usage Mining

Preeti Rathi[1] and Nipur Singh[2]

[1] Research Scholar Department of Computer Science Kanya Gurukul Campus Dehradun, India
[2]Department of Computer Science Kanya Gurukul Campus Dehradun, India
[1]mcapreeti.rathi@gmail.com, [2] nipursingh@gkv.ac.in

Abstract

Retrieval of information from the database and web log files is a very time consuming process. There are many techniques and models to retrieve data from the web. There are two types of data available on the web i.e. structured and unstructured. If data is structured then retrieval of information is an easy task. Otherwise, firstly apply the algorithm to unstructured data and then models will be applied. Vector space and Boolean models are used for IR. In this paper, we compare both Boolean model & Vector Space model techniques to retrieve data from the web (log files) and proposed a new algorithm based on time, frequency, memory consumption, etc.

Keywords: *Log files, IR, Vector space model, Boolean model, Web, TF, IDF*

1. Introduction

Information Retrieval (IR) is the process to retrieve relevant information from huge information resources. In information, retrieval searches can be based on full text rather than content-based. IR is the science of searching for information in a document, searching for documents themselves, and also searching for metadata describe data, and for databases of texts, images or sounds [13]. An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information which need by the user. For example search information in search engines. In IR a query does not uniquely identify a single object because it is a collection of documents. Some objects may match the query, with different degrees of relevancy.

There are different searches approaches for information retrieval system which deals with following features. These features are as follows.

- Storage and organization of information.
- Representation of information.
- Accessing of information.

Retrieval techniques - Those techniques which are helpful for users to locate information effectively and efficiently are known as retrieval techniques [11]. These techniques help users to find out the required information easily. There are two categories of retrieval techniques.

1.1. Basic retrieval techniques

Article history:
Received (May 28, 2019), Review Result (August 2, 2019), Accepted (September 15, 2019)

Print ISSN: 1738-9968, eISSN: 2652-2233 IJHIT

1.1.1. Boolean model

The boolean model used Boolean constraints i.e. AND, OR, NOT. There are two types of the Boolean model

(1) Unranked boolean retrieval model: In this model retrieval of documents, that satisfy the constraints in no particular order.

(2) Ranked Boolean Retrieval Model: In this model retrieval of documents, that satisfy the constraints and rank them based on these constraints.

(3) Advantages of the boolean retrieval model

It is easy to implement.

It is easy to understand why the document is retrieved or not.

Users can determine whether the query is too specific or too broad.

(4) Disadvantages of the boolean retrieval model

The problem is for the user to generate a virtuous Boolean query.

Exact matching may retrieve specific or broad documents.

1.1.2. Vector space model

The vector space model is also known as the term vector model. It is an algebraic model for representing text as vectors i.e. index term. This model is basically used in filtering, retrieval, indexing of information and relevancy rankings.

(1) Advantages of vector space retrieval model

- It is a simple model based on linear algebra.
- Term weights are not binary.
- It allows computing the similarity between queries and documents.
- It allows ranking documents according to their possible relevance.
- It allows partial matching of queries.

(2) Disadvantages of vector space retrieval model

- They have poor similarity values between long documents i.e. a small scalar product and a large dimensionality.
- Search keywords must precisely match otherwise result is a false positive match.

1.2. Advanced retrieval techniques

(1) Fuzzy search - Fuzzy searching, search those terms which are spelled incorrectly in data as well as in the query. Fuzzy search is related to truncation. Truncation is planned to retrieve different forms of terms when they share some parts in common. Fuzzy searching is designed to find terms that are spelled incorrectly at data entry or query points [12].

(2) Weighted search - Weighted searching, is a searching technique to retrieve information based on weights that are assigned to each term in a document. Weights are numerical value which is used to calculate the frequency of the term in a document.

The rest of the paper has been organized as, related work discussed in section 2, information retrieval process in section 3, and proposed work are discussed in section 4. After the proposed work experimental results and comparison of our proposed algorithm’s results with previously proposed techniques are discussed in section 5, and finally, conclusion and future scope discuss in section 6.

2. Related work

Now day's retrieval of information from the web is a very crucial task. There are various algorithms to retrieve information from the web. There are various algorithms exits for ranking of page discuss by the authors are as follows:

Author, Abdur Rehman et al. [2] gives a technique named relative discrimination criterion. This technique is based on ranking to calculate the rank of each term is present class and other class related to this class. This paper also compared the performance of RDC based on the following parameters i.e. information gain, CHI Squared, Odds ratio.

Author, Kyu-Huwan Jung et at. [3] discover a new method to calculate rank named as multi-support vector domain description. This method is used both support vector machine and ranking support vector domain description.

Author Berry et al. [4] proposed a vector space model for retrieving information from a large number of datasets. In this paper, the author used a matrix approach using rows and columns to determine the value of documents. In every query examine the frequency of each term.

Author Singh and Dwivedi [5] have discussed various approaches of the Vector space model that calculates the similarity between hits of information retrieval. These approaches are based on normalization. There is a various model like term count model, term frequency-inverse term frequency model.

Author, Hiemstra and De Vries [6] projected the language model using information retrieval models known as a Boolean model, Probabilistic model and Vector Space model.

Author, Lv and Zhai [7] discovered an adaptive feedback approach for information. It is a learning approach to calculate the optimal balance coefficient for each query and document. This approach used a ranking algorithm to rank the documents.

Author, Jitendra Nath Singh and Sanjay Kumar Dwivedi [8] discuss a comparative analysis of different types of vector space models. Authors describe the similarity function between the documents and query by using the term count model and TF-IDF vector space model based on normalization frequency. Page ranking algorithm is used for ranking the page according to the access of each page i.e. the frequency of web pages[8].

Author, Nicole Lang Beebe et al.[9] proposed a ranking algorithm based on digital forensic domain and support vector machine used to categorized a group of characters known as a string. M57 datasets were used to apply this algorithm and named as Relevancy Ranking Algorithm.

Author, Han Xu et al. [10] present a ranking algorithm to rank a scientific document using the PageRank algorithm. The author also discovered the ranking algorithm named Random Literature Explorer. This algorithm is more efficient and gives better results in comparison to the exits algorithm.

In the information retrieval process, first, we fetch the information from the log files, then apply the pre-processing techniques to personalize the retrieval information according to the user's requirement. We rank each page using the PageRank algorithm and match the information according to the user's need, after that data is fetched using a query model and finally get feedback from the user's side to retrieve useful information.

3. Information retrieval process

In the information retrieval process, first, we fetch the information from the log files, then apply the pre-processing techniques to personalize the retrieval information according to the user's requirement. We rank each page using the PageRank algorithm and match the information according to the user's need, after that data is fetched using a query model and finally get feedback from the user's side to retrieve useful information.

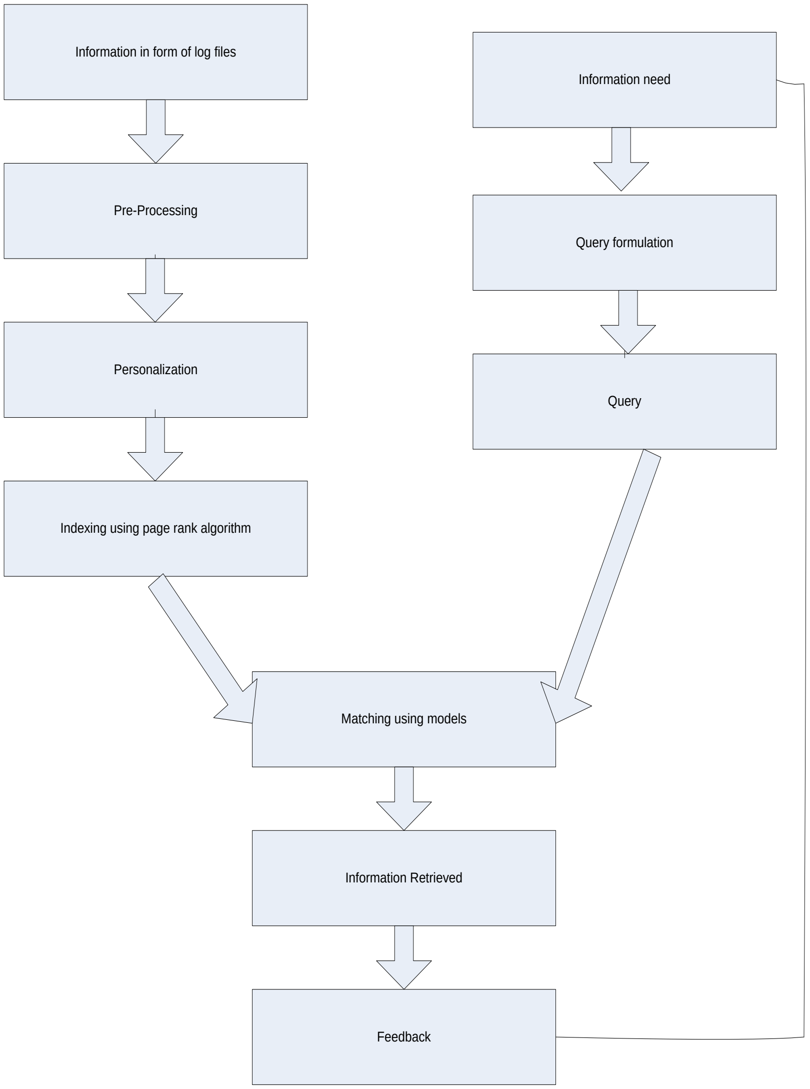

Figure 1. Framework of information retrieval process

4. Proposed work

We have proposed an algorithm for information retrieval using the keyword feedback, time and frequency, then have calculated the term frequency and inverse term frequency.

Page rank of page A is defined as follows, this is the basic page rank algorithm:

$$PR(A) = (1-d) + d\,(PR(W_1)/C(W_1) + \ldots + PR(Wn)/C(Wn)) \quad [14]$$

Where,

$PR(A)$ - Page Rank of page A

W_i $(1<=i<=n)$ - are pages linked to page A

$C(W_i)$ - number of outbound links on page W_i

d - damping factor which can be set between 0 & 1

To calculate page rank for the n webpages, firstly we initialize all webpages with equal page rank. Then step by step we calculate Page Rank for each Webpage one after the other.

The proposed algorithm (efficient retrieval algorithm)

For document D
Let L= (S1, S2,………,SN)
Where L is a collection of log files and S1, S2,…………..,SN are terms of log files.
The weight of a term Si in document dj is the number of times that Si appears in dj, denoted by fij .
N= total number of documents
dfi = total number of documents where Si appears.

Input: Log files

Output: Retrieval Information

1. Personalized log files.
2. Apply Page Rank Algorithm i.e.

$$PR\ (A) = (1\text{-}d) + d\ (PR\ (W_1)/C\ (W_1) + \ldots + PR\ (Wn)/C\ (Wn)) \tag{1}$$

3. Write a query to retrieve data from log files.
4. For each log files assign weight as a frequency of accessing an information

$$W_{ij} = \begin{cases} 0, & S_i \ of \ document \ d_j \nexists D \\ 1, & S_i \ of \ document \ d_j \in D \end{cases} \tag{2}$$

Also assign,

$$d_{ij} = \begin{cases} 0, if \ searched \ term \ is \ not \ in \ page \ i \ of \ document \ j \\ 1, if \ searched \ term \ is \ in \ page \ i \ of \ document \ j \end{cases} \tag{3}$$

5. Calculate term frequency, inverse term frequency and term weight,

$$tf_{ij} = \frac{f_{ij}}{Max\,\{f_{1j}, f_{2j}, \ldots\ldots\ldots f_{sj}\}} \tag{4}$$

$$idf_i = log\,\frac{N}{df_i} \tag{5}$$

$$Q_{ij} = tf_{ij}\ X\ idf_i \tag{6}$$

6. Calculate Cosine Similarity

$$Cosine(d_j, q) = \frac{\sum_{i=1}^{V} W_{ij}\ X\ W_{iq}}{\sqrt{\sum_{i=1}^{V} {W_{ij}}^2}\ \ X\ \sqrt{\sum_{i=1}^{V} {W_{iq}}^2}} \tag{7}$$

To find out the performance of our algorithm we have calculated precision and recall measure as [1]:

- Precision- The ability to retrieve top ranked documents that are most relevant.
- Recall - The ability of the search to find all the relevant items from documents.

$$Precision = \frac{relevant\ retrieved\ document}{total\ number\ of\ retrieved\ documents} \tag{8}$$

$$Recall = \frac{relevant\ retrieved\ document}{total\ number\ of\ relevant\ \ documents} \tag{9}$$

5. Experiment result & comparison

In this section, we discuss the result of the proposed algorithm based on high performance and low memory consumption and comparison of the Efficient Retrieval Algorithm with previously proposed techniques. Page rank of each page as given below:

Page Rank of page 1 is:	*0.31666666666666665*
Page Rank of page 2 is:	*0.05*
Page Rank of page 3 is:	*0.31666666666666665*
Page Rank of page 4 is:	*0.06666666666666667*
Page Rank of page 5 is:	*0.25*

If the page is present in the document then the assigned value of this term is 1 otherwise 0.

Table 1. Absence and presence of page in a document

Terms	D1	D2	D3	D4	D5
T1	0	1	1	1	0
T2	0	1	0	1	1
T3	1	0	1	1	1
T4	0	0	0	0	1
T5	1	1	1	1	0

Table 2. Term frequency and inverse term frequency

Terms	f_{ij}	tf_{ij}	idf_i
T_1	0.6	0.75	0.430676
T_2	0.6	0.75	0.430676
T_3	0.8	1	0.345694
T_4	0.2	0.25	0.247694
T_5	0.8	1	0.345694

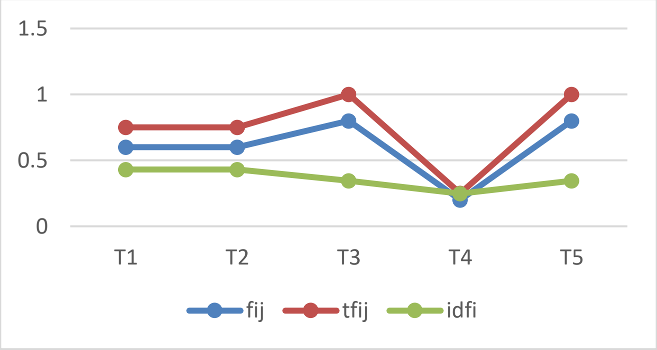

Figure 2. TF and IDF graph

[Table 3] shown the weight of each term.

Table 3. Term weight

Term	Weight
T_1	0.323007

T_2	0.323007
T_3	0.345694
T_4	0.061923
T_5	0.345694

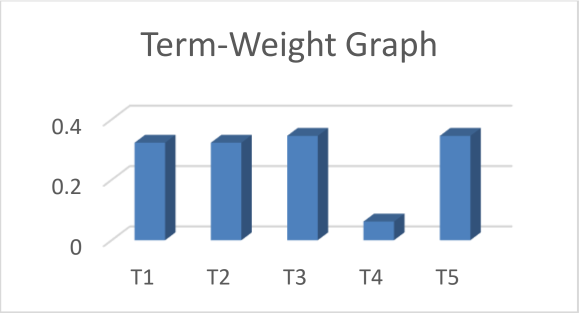

Figure 3. Term weight graph

[Table 4] shown the recall precision measure calculated by the recall precision formula.

Table 4. Recall Precision

I	$P(r_i)$	R_i
0	100%	0%
1	100%	10%
2	80%	20%
3	71%	30%
4	62%	40%

Table 5. Comparison of various algorithms.

Parameter	FCM	Genetic Algorithm	Proposed Algorithm
Memory consumption	75%	70%	50%
Time	24 sec	22 sec	sec

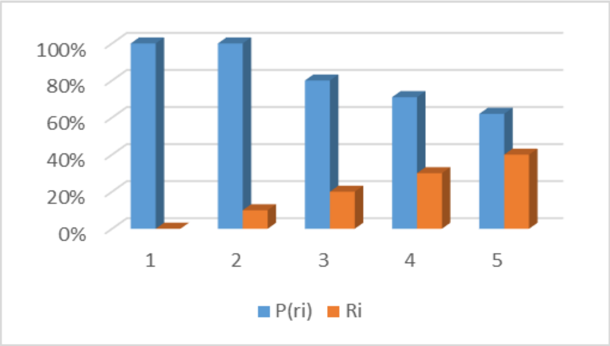

Figure 4. Recall precision graph

6. Conclusion & future scope

In this paper, we proposed an algorithm named as Efficient Retrieval Model to retrieve information from log files. The proposed algorithm takes less time and low memory consumption as compared to the previously proposed algorithm. We can enhance the algorithm for better performance by considering other parameters such as content etc.

References

[1] M.François Sy, S.Ranwez, and J.Montmain,"User centered and ontology based information Retrieval system for life sciences," BMC Bioinformatics, **(2105)** DOI: 10.1186/1471-2105-13-S1-S4

[2] Abdur Rehman, Kashif Javed, Haroon A. Babri, and Mehreen Saeed, "Relative discrimination criterion - A novel feature ranking method for text data," Expert Systems with Applications, Elsevier, vol.42, no.7, pp.3670-3681, **(2015)** DOI: 10.1016/j.eswa.2014.12.013

[3] Kyu-Hwan Jung and Jaewook Lee, "Probabilistic generative ranking method based on multi-support vector domain description," Information Sciences, Elsevier, vol.247, pp.144-153, **(2013)** DOI: 10.1016/j.ins.2013.05.001

[4] M.W. Berry, Z. Drmac and E. R. Jessup, "Matrics, vector spaces, and information retrieval," Society for Industrial and Applied Mathematics, vol.41, no.2, pp.335-362, **(1999)** DOI: 10.1137/S0036144598347035

[5] J.N. Singh and S.K. Dwivedi, "Analysis of vector space model information retrieval". In Proceedings of the National Conference on Communication Technologies and its Impact on Next Generation computing (CTNGC' 12), International Journal of Computer Application (IJCA), **(2012)**

[6] D. Hiemstra and A.P. De Vires, "Relating the new language models of information retrieval to the traditional retrieval models," CTIT Technical Report TR- CTIT- 00- 00, pp.1-14, **(2000)**

[7] Y. Lv and C. Zhai, "Adaptive relevance feedback in information retrieval," In Proceeding of CIK '09, November 2-6, Hong Kong, China, **(2009)** DOI: 10.1145/1645953.1645988

[8] Jitendra Nath Singh, "A comparative study on approaches of vector space model in information retrieval," International Journal of Computer Applications (0975-8887), **(2013)**

[9] Nicolas Couellan, Sophie Jan, Tom Jorquera, and Jean-Pierre George, "Self-adaptive support vector machine: a multi-agent optimization perspective," Expert Systems with Applications, Elsevier, vol.42, no.9, pp.1-15, **(2015)** DOI: 10.1016/j.eswa.2015.01.028

[10] Han Xu, Eric Martin, and Ashesh Mahidadia, "Contents and time sensitive document ranking of scientific literature," Journal Informetrics, Elsevier, vol.8, no.3, pp.546-561, **(2014)** DOI: 10.1016/j.joi.2014.04.006

[11] Dr. M. Balamurugan, "A trend analysis of information retrieval models," International Journal of Advanced Research in Computer Science, ISSN no.0976-5697, vol.8, no.5, May-June, **(2017)**

[12] Kiran Prakash Bachchhav, "Information retrieval: search process, techniques and strategies," IJNGLT, vol.2, no.1, February, (**2016)**

[13] Srinagnaya G., "A technical study on information retrieval using web mining techniques," IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication systems (ICIIECS) **(2015)** DOI: 10.1109/ICIIECS.2015.7192894

[14] Sanjay and Dharmender Kumar, "A review paper on page ranking algorithms," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) vol.4, no.6, June, **(2015)**