IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-2, Issue-6 E-ISSN: 2347-2693

Combined Approach for Page Ranking In Information Retrieval System Using Context and TF-IDF Weight

Shikha Gupta[1*], Vinod Jain[2] and Pawan Bhadana[3]

[1*,2,3]*Department of Computer Science and Engineering*,
B.S.Anangpuria Institute of Technology and Management, India
shikha.0909@gmail.com; jainvinod81@gmail.com; pawanbhadana79@gmail.com

www.ijcaonline.org

Received: 17/05/2014 Revised: 29/05/2014 Accepted: 19/06/2014 Published: 30/06/2014

***Abstract*—** Ranking in Information Retrieval System has been researched extensively in recent years. IR System is aimed at providing users the most relevant documents in minimum possible time. Therefore, providing fast and efficient result to the user is a major issue in determining the performance of the IR systems. Ranking of the pages is done after they have been indexed. Most of the existing architectures of IR system shows that they rely on keyword-based queries and the indexing is done based on the terms of the document and also consists an array of the posting lists, each posting list being associated with a term and containing the term along with the identifiers of the documents containing them. This paper proposes a ranking structure where ranking is done on the basis of a combination of the context of the document and on term basis. Context based indexing is considered in which all the available context along with the list of related terms of that context are stored. List of documents of particular contexts are stored in context repository. The indexing of the documents are done with respect to their context. To rank these documents a combination of context based weight (how much a document is relevant with a context) and TF-IDF weight (how much the user query is relevant to a document without considering context) are used. The ranking is done in decreasing order of their total weight.

***Keywords*—**Information Retrieval System, Page Ranking, Context, TF, IDF.

II. INTRODUCTION

A. *Information Retrieval System*

Information Retrieval is an art of storing, presenting, organizing and providing easy access to information items, which are collection of components and processes. The information should be represented and organized in a way to meet the information access needs of the user. The information retrieval can be defined as:

Information retrieval (IR) searches unstructured material (usually text documents) to find an object which satisfies the information need from the available large collections which are usually stored on computers.

The input is taken from the user in the form of a query to the system, and then compared with the information already collected by the system, to produce an output, in the form of set of information objects or texts which are considered to be relevant to the query. IR system uses Inverted index as the data structure which is an index having entries {term, doc IDs}.

B. *Searching*

An information retrieval process starts when the user fires a query into the system. Queries are considered as formal statements of information needs, like search strings in the web search engines. Searching can be defined as the process of finding the documents that matches the user query. Searching can be done in two ways: simple search and context based search. In simple search, all the documents are taken into account whereas in context based search, only documents with relevant context are searched.

C. *Ranking*

In ranking all the resulted documents are arranged according to their relevancy with the query. In information retrieval system it's not just one object that uniquely identifies the query from the collection, instead, multiple objects may match the user query, but, with different order of relevancy. Most IR systems calculates a numeric value (weight) for each object present in the database to estimate the relevancy of the object with the query, and then ranking of these objects are done with respect to their calculated value. For different type of searching, different values are assigned. When simple search is done, then only TF-IDF value is used for ranking and in context based search, a combination of TF-IDF weight and context based weight is used. Objects having top ranks (large weights) are shown to the user after ranking. The process can then be iterated by the user by refining the query, but only if required. Page rank derived its importance from the fact that an information retrieval system takes the assigned value into account while deciding which results should appear at the top –so that they can be easily seen.

D. *Context Based Approach*

Context based approach is used in ranking scheme where ranking is done based on the context of the document instead of terms basis. The contextual information about documents

Corresponding Author: *Shikha Gupta,* shikha.0909@gmail.com

is extracted by analyzing the structure of those documents that refers to them. Contexts are used to describe collections. The disadvantages of term based approach like polysemy and synonymy are overcome by using this approach.

E. TF-IDF Weight

TF-IDF stands for term frequency–inverse document frequency, which is a numerical value that is used for estimating the importance of a word in a document in the collection of corpus. It is mostly used in information retrieval and text mining as a weighting factor. The value of tf-idf increases proportionally with the number of times a word appears in a document, but it decreases by the number of documents the word appears, which helps in controlling the fact that there are few words which are more common than other words.

TF-IDF weight is calculated by values for each word in a document using the formula as defined in equation 1 :

$$\text{TF-IDF weight} = TF * \log(IDF) \quad (1)$$

Where, IDF = n/DF

Here, n = total number of available documents

DF = Document Frequency

Words having high TF-IDF weight show a strong relationship of the word with the document in which they appear, showing that if that word appears in a query, then that document could be of use to the user.

II. RELATED WORK

In this paper, a review of some previous work on ranking is given. In this field, large number of algorithms and techniques are already available but they seem to be less efficient in efficiently ranking the documents.

Page Rank Algorithm- Page Rank Algorithm [5] is among the most common page ranking algorithms. This link analysis algorithm provides a way of measuring how important a page is. It works by roughly estimating the importance of the page using the quality and number of links to that page. It uses the assumption that an important page will receive more number of links from other pages. It assigns a numerical weight to any given element E which is referred as the Page Rank of E and is represented by PR (E).

HITS Algorithm- Hyperlink-Induced Topic Search [6] is also known as hubs and authorities. It is a link analysis algorithm which rates pages. Ranking is done by processing in links and out links of pages. A page pointing to many other pages represents a good hub, and page that was linked by large number of hubs represents a good authority. Each page is therefore assigned two scores in this scheme: its authority value, which is the estimated value of the content of a page, and its hub value, which is the estimated value of its number of links to other pages. The limitation of this algorithm is that sometimes it assigns high rank value to those pages that are popular but not much relevant to the given query.

Weighted Page Rank Algorithm- A modification to the standard Page Rank algorithm is the Weighted Page Rank algorithm (WPR) [7]. The ranking is done by considering the importance of both in-links and out-links of the pages. Popularity of the pages are used for distributing the rank scores and the popularity is decided by observing the total number of its in-links and out-links. Its performance is better than the standard Page Rank algorithm by providing more relevant pages to the given query.

Categorization by context- This approach proposes [1] a new ranking scheme where ranking is done based on the context of the document instead of terms basis. The contextual information about documents is extracted by analyzing the structure of those documents that refers to them. Contexts are used to describe collections. The disadvantages of term based approach are overcome by using this approach.

The given paper discusses the ranking system in which ranking is done on the basis of the context of the terms in the documents along with the basis of terms itself.

III. PROPOSED WORK

A. Architecture of Context based Ranking System

This paper proposes a new ranking scheme which uses a combination of context weight and TF-IDF weight. Context weight assigns a weight to each document which signifies the relevancy of that document to the specified context and TF-IDF weight assigns weight to the documents according to its relevancy with respect to the user query. Context weight is static and is pre-determined. TF-IDF weight is dynamic and is determined after a query is fired. The proposed architecture of the context based ranking in information retrieval system is shown in figure1.

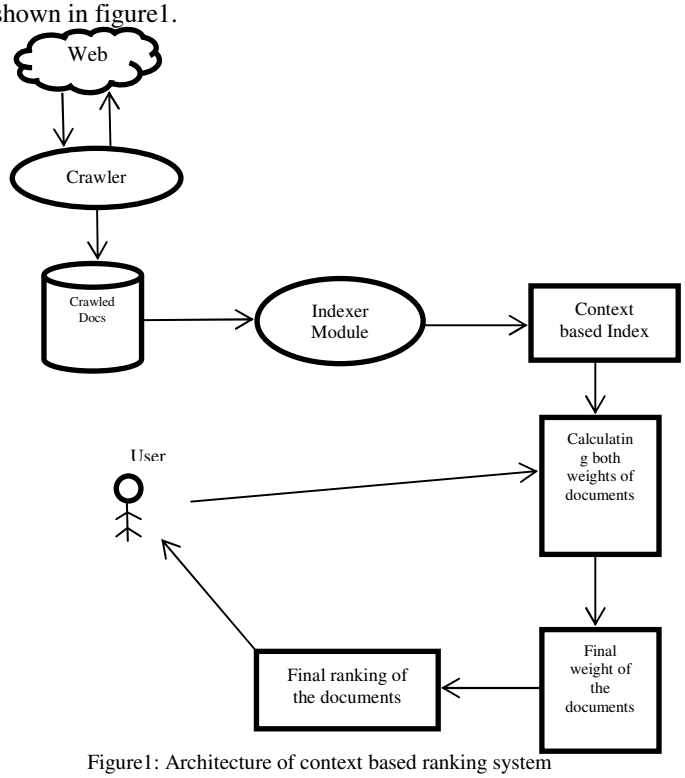

Figure1: Architecture of context based ranking system

B. *Description of various components*

1) Crawler : The job of the crawler is to store the pages in a repository. This repository stores all the documents to be indexed and searched for a user query.

2) Crawled docs : It is the set of documents that have been collected by the crawler.

3) Indexer : The indexer parses all the documents of the page repository and makes an entry of every token in the index. It will also assign context to the documents. This paper is not concerned with finding the context of a document.

4) Context Based Index : This index stores all the tokens of the documents in the corpus. The index also stores the frequency of the tokens in different documents.

5) Calculate both weights of the documents : Ranking of the pages is done by using combination of two weights.

6) Context based weight : It assigns a static weight to all the documents which will help in ranking of these documents. Algorithm for calculating of this weight is discussed later.

7) TF-IDF weight : It assigns a dynamic weight to only those documents which contains the user query and is dynamic. Algorithm for calculating of this weight is discussed later.

8) Final weight of the documents : After assigning both the weights to the documents, final weight is calculated by combining these two weights.

9) Final ranking of the documents : The documents are now displayed in decreasing order of their total weight.

C. *Proposed Algorithm for calculating TF-IDF Weight*

```
Read query from the user (list-1)
For (All Documents available in the Corpus)
{
   Pick a document from the document list and parse that document to find the list of tokens.(list-2)
                Find TF = frequency of user query in list-2

}
Find DF = Number of documents in which list-1 is present
Find IDF = n/DF
        Where, n = total no. of documents in the corpus
Assign TF-IDF Weight to all the documents, using-
            TF-IDF Weight = TF*log(IDF)
```

D. *Proposed Algorithm for calculating Context Based Weight*

```
Read query from the user
Read the Context name in which the user wants to search documents for the query
        Obtain list of related words of that context (list-1)
For (all documents related to that Context)
{
        Pick a document from the document list of that context and parse that document to find the list of tokens (list-2)
        Find count = common words in list-1 and list-2
        Set Context weight to that document as
        Context weight = count
}
```

E. *Calculation of Final Weight*

It is the order in which the document list is displayed to the user. The results will be displayed in the descending order of the final rank weight. The formula for calculating the final rank weight is as follows:

Final rank weight = TF-IDF weight + Context weight

IV. P ERFORMANCE ANALYSIS

A. *Performance Analysis of Proposed and the Existing Ranking System*

It can be proved that this new technique provides more relevant documents to the user. To measure the correctness of the defined algorithms a term called precision is used. Precision is defined as a measure of the ability of a system to provide only relevant items to the user.

Precision = (no of relevant items retrieved) / (total no of items retrieved)

When the simple searching technique is used then large number of documents are available as output and there may be documents with high TF-IDF weight which are not relevant as shown in fig 2.

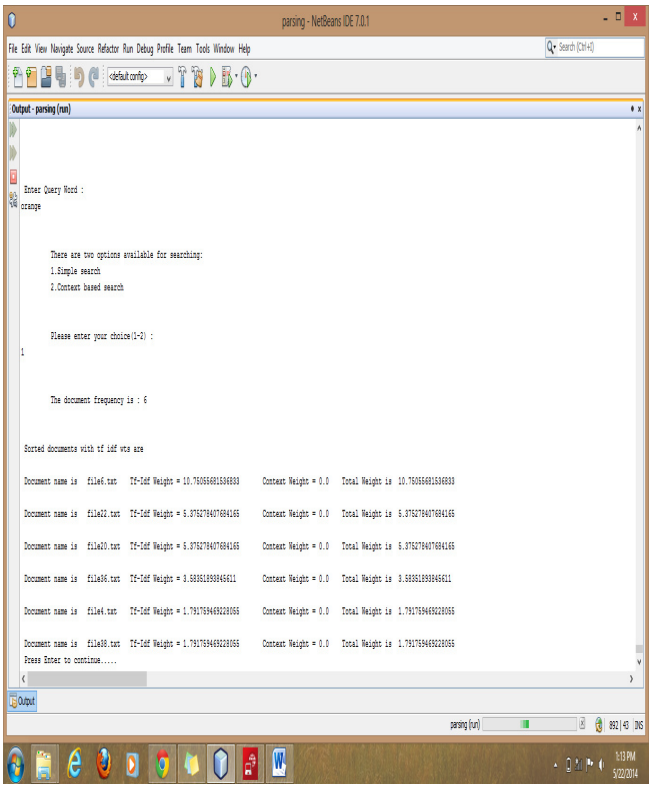

Fig 2: Simple search

Precision = 2/6 = 0.33

When context based searching is done, then number of relevant documents increased, as it uses a combination of Context based and TF-IDF weight as shown in fig 3.

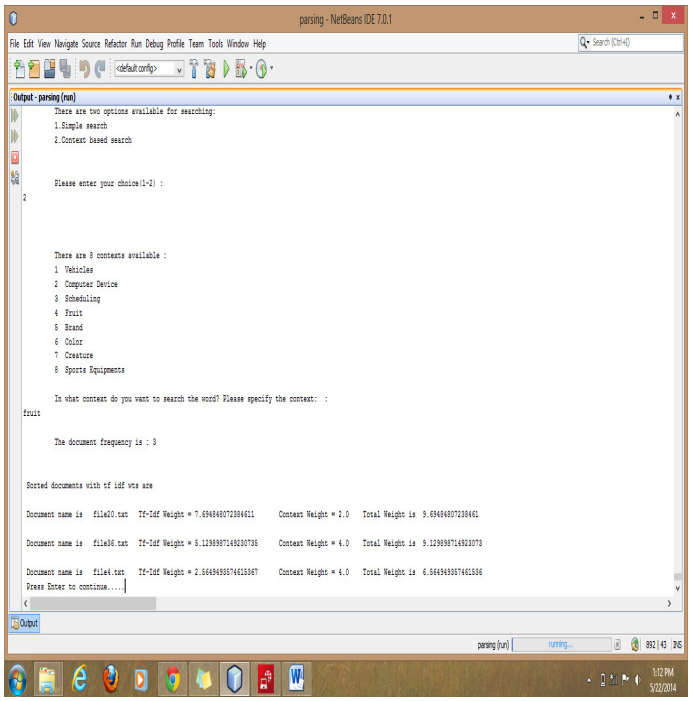

Fig 3: Context Based Search

Precision = 2/3 = 0.66

B. Comparing Performance of Proposed and Existing Ranking System

Performances of both the current and existing systems have been graphically visualized in figure 4.

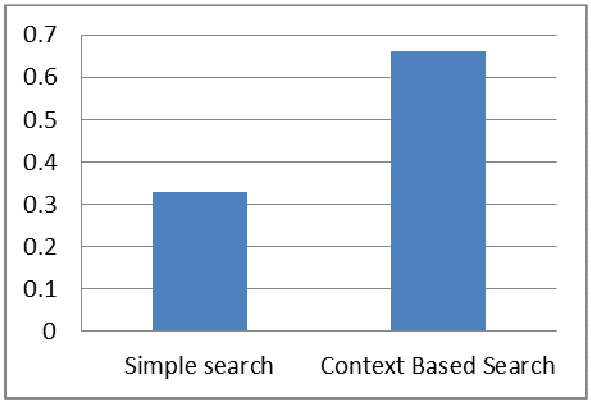

Fig 4: Graph showing performance of proposed and existing ranking system

V. CONCLUSION

This paper presents a ranking architecture where ranking is done on the basis of the context of the document along with the term basis. Using this approach in Information Retrieval system improves the ranks given to the obtained documents before displaying them to the user. The system is now improved and will now provide more relevant and useful pages to the user as a result of the fired user query. In future the work can be tested on a corpus with large number of documents in different context.

REFERENCES

[1] Parul Gupta and Dr. A.K.Sharma, "Context based Indexing in Search Engines using Ontology", International Journal of Computer Applications Vol. 1, - No. 14, ISSN 0975-8887.

[2] Sunita Rani, Vinod Jain and Geetanjali Gandhi, "Context Based Indexing and Ranking in Information Retrieval Systems", International Journal of Computer Science and Management Research, Vol. 2, Issue 4 April 2013, ISSN 2278-733X.

[3] Shikha Gupta, Vinod Jain and Pawan Bhadana, " New Combined Page Ranking Scheme in Information Retrieval System", International Journal of Scientific and Research Publications, Vol. 4, Issue 4, April 2014, ISSN 2250-3153.

[4] Dilip Kumar Sharma and A. K. Sharma," A Comparative Analysis of Web Page Ranking Algorithms", in International Journal on Computer Science and Engineering, Vol. 02, No. 08, 2010, 2670-2676.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Technical Report,Stanford Digital Libraries SIDL-WP-1999-0120, 1999.

[6] Sergey Brin and Larry Page, "The anatomy of a Large-scale Hypertextual Web Search Engine", In Proceedings of the Seventh International World Wide Web Conference, 1998.

[7] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", In proceedings of the 2rd Annual Conference on Communication Networks & Services Research, PP. 305-314, 2004.

AUTHORS PROFILE

Shikha Gupta is a M.Tech scholar in computer science and engineering at B.S.Anangpuraia Institute of technology and Management, Faridabad. (shikha.0909@gmail.com). She has published one more paper in this field titled "New Combined Page Ranking Scheme in Information Retrieval System".

Vinod Jain is working as Assistant Professor in information technology department at B.S.Anangpuraia Institute of Technology and Management, Faridabad. He has completed master of computer application (MCA) in June 2004 and Master of Technology in 2012. His area of research include IR systems and Genetic Algorithms. (jainvinod81@gmail.com)

Pawan Bhadana is working as associate professor and head in department of computer science and information technology at B.S.Anangpuraia Institute of technology and Management, Faridabad. He has published many research papers in the area of MANETs, mobile sensor networks and information retrieval systems. (pawan.bhadana79@gmail.com)