Fuzzy Page Ranking Scheme In IR System

Shikha Gupta

Assistant Professor, Advanced Institute of Technology and Management

Abstract— **In recent years, ranking in Information Retrieval System has been researched extensively. 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 the relevance of the document with respect to the context of the query. 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 along with finding the context of the document using the concept of membership value of fuzzy sets. 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 searched. 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 calculation of relevancy is done through the concept of fuzzy logic. The ranking is done in decreasing order of their membership value.**

Index Terms— **Fuzzy Set, Information Retrieval System, Membership Value, Page Ranking, TF-IDF.**

I. INTRODUCTION

In this paper, the concept of page ranking in Information Retrieval System is discussed. A large number of techniques are available but all of them have some limitations involved, so this paper will cover the most common available techniques and will provide an idea to overcome the limitations.

A. Information Retrieval System

Information Retrieval System is a system which retrieves the relevant information from a large collection of information. It works in a series of steps. The steps followed are:

- The Information Retrieval System collects all the available data and store it in a repository.
- It then takes an input from the user, called as query.
- As the query is fired, it processes it and finds the relevant documents from all the available documents.
- All the relevant documents are then returned to the user arranged in the order of their relevancy, with the most relevant documents displayed at the top.

B. Page Ranking

Page Ranking is used to rank all the relevant documents according to their relevancy. The IR system will return a number of documents relevant to the query. These documents are then arranged in a particular series depending on their degree of relevancy to the query. The relevancy may vary from user to user and from context to context. Therefore, an appropriate mechanism is required to rank these documents.

C. Fuzzy System

Fuzzy System as the name suggest is fuzzy, i.e., which is not certain and may vary according to the user and the domain. It uses membership value to define the relevancy of any term in a set or context. The membership value can take the value from 0-1 depending on how much the term belongs to that set. A term with higher membership value is more relevant with respect to the context or set than a term with lesser value.

D. Context Based Approach

Context Based Approach is used to rank the pages depending on their context rather than just the terms used. This approach is introduced because term based approach have some disadvantages like polysemy and synonymy associated with it.

E. TF-IDF Weight

TF-IDF stands for Term Frequency-Inverse Document Frequency. It calculates a numeric score for each word to find its relevancy in a document. The TF-IDF weight is directly proportional to term frequency, i.e., number of times the word appears in a particular document. It is inversely proportional to Document Frequency (directly to Inverse document frequency), i.e., number of documents in which the word appears. The TF-IDF weight is less for words which are more common and does not help in calculating the relevancy with the context. The weight is high for words which are more relevant.

TF-IDF Weight = TF*log(IDF)

Here,

IDF = x/DF

Where,

X = number of all the available documents

DF = Document Frequency

II. LITERATURE

A large number of page ranking algorithms are available which are used to provide ranking to the relevant documents. Some of the most commonly used algorithms are discussed in this paper.

A. Page Ranking Algorithm

One of the most commonly used algorithms is Page Ranking Algorithm. It provides ranking based on a numerical value, which is calculated by the number of links pointing to that page. A page with higher number of links pointed to it is considered more important than others.

The problem with this algorithm is that it does not consider the pages according to their context or meaning but it just searches the pages based on the terms and number on incoming links.

B. Weighted Page Rank Algorithm

This algorithm is an extension to the above defined Page Ranking Algorithm. It considers both in-links and out-links of the pages to determine the weight. Popularity of the pages are used to determine the importance of that page, which is determined by the total number of in-links and out-links. It performs better than page ranking algorithm.

The problem with this approach is that it also does not consider the context of the pages.

C. *Categorization by context*

This approach considers the concept of context of a document. It finds the context of the document and then provides only the relevant documents to the user. It overcomes the disadvantages of the previous algorithm. For example: The word Apple can be used in context of fruit as well as a brand name. Another example is Mouse which can be an animal or a hardware device of computer. This approach will categorize the available documents with respect to their context and will provide the documents of that context only.

The problem that arises with this algorithm is how to find how much a document is relevant to a particular context and to the requirement of the user as the relevancy with the query may vary from user to user.

D. Page Ranking Using Fuzzy Set

The above problems can be resolved by using a new algorithm named Fuzzy Logic based Weighted Page Content Rank Algorithm (FLWPCR). This algorithm will improve the order of the resultant pages as all the important and relevant pages will appear at the top of the list. The user can get the relevant pages easily. But the problem with this approach is that it requires more time to process when a keyword is searched by the user. This problem is solved by integrating the WPCR algorithm with the clustering algorithm. With this new approach, the processing time is reduced. This technique will provide more relevant documents to the user thus, improving the result.

III. CONCLUSION

This paper concludes that the concept of Fuzzy Logic can be used to check the relevancy of the documents as it provides flexibility and also makes the Information Retrieval System adaptive. The membership value is calculated for estimating the relevancy of the documents to the query and to the user needs. Uncertainty and imprecision are highly involved in IR System and Fuzzy Logic will handle them with the close approximate. For each document the Context weight and TF-IDF Weights are calculated and using these values the membership value is determined which finds the relevancy. By using this concept the limitations of other existing techniques can be overcome and a better result will be provided to the user.

ACKNOWLEDGMENT

Shikha Gupta thanks all her teachers who guided her during her learning age in graduation and post-graduation.

REFERENCES

[1] Shikha Gupta, Vinod Jain , Pawan Bhadana, "New combined page ranking scheme in Information Retrieval system", International Journal of Scientific and Research Publications, Volume 4, Issue 4, April 2014 1 ISSN 2250-3153.

[2] Shikha Gupta, Vinod Jain and Pawan Bhadana, "Combined Approach for Page Ranking In Information Retrieval System Using Context and TF-IDF Weight", Volume 2, Issue 6, E-ISSN:2347-2693.

[3] 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.

[4] Vinoth Kumar.G.S, Janet.J, Kamal.N, "Efficient Page Ranking using Fuzzy Logic Based Weighted Page Content Rank Algorithm", International Journal of Advanced Research in Computer Science Engineering and Information Technology, Volume: 2 Issue: 3 08-Apr-2014,ISSN_NO: 2321-3337.

[5] 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, 2670-267.