New Combined Page Ranking Scheme in Information Retrieval System

Shikha Gupta[*], Vinod Jain [**], Pawan Bhadana[**]

[*,**] Department of Computer Science and Engineering,
B.S.Anangpuria Institute of Technology and Management, Village Alampur, Ballabgarh, Faridabad, Haryana, INDIA

Abstract- Available information is expanding day by day and this availability makes access and proper organization to the archives critical for efficient use of information. People generally rely on information retrieval (IR) system to get the desired result. In such a case, it is the duty of the service provider to provide relevant, proper and quality information to the user against the query submitted to the IR System. With time, many old techniques have been modified, and many new techniques are developing to do effective retrieval over large collections. This paper is concerned with the analysis and comparison of various available page ranking algorithms based on the various parameters to find out their advantages and limitations in ranking the pages. This paper proposed a new page ranking system that will combine the old tf/idf weight and a new document-context-weight to find the rank of the documents.

Index Terms- Information Retrieval System, Page Ranking, Context Based Ranking.

I. INTRODUCTION

Information retrieval systems are defined as some collection of components and processes which takes input in the form of a query from the user to the system, then compares it with the information which has been collected by the system, and then produce an output, which is some set of texts or information objects considered to be related to the query. It is the activity of obtaining the information resources which are relevant to an information need (query) from a collection of information resources. Data structure used by an IR system is Inverted index which is an index of {term, doc IDs} entries.

IR system consists of three main components : firstly the user in the system; then the knowledge resource on which the user has an access and with which he/she interacts; and, a person(s) and/or device(s) that supports and mediates the interaction of the user with the knowledge resource (the intermediary).

In an IR System the processes which are to be considered as important are:

Representation of the user’s information problem which is in the form of texts in the knowledge resource: e.g. indexing;
Comparison of representation of texts and information problem: e.g. retrieval techniques;
Interaction between the user and an intermediary: e.g. human-computer interaction or reference interview; and, sometimes,
Judgment of appropriateness of the text to information problem submitted by the user: e.g. relevance judgments; and
Modification of the representation of an information problem: e.g. query reformulation or relevance feedback.

II. EXISTING WORK AND LITERATURE SURVEY

Page Ranking

Ranking is a process of arranging the resulted documents in the order of their relevancy. An information retrieval process begins when the user enters a query into a system. Queries can be defined as formal statements of information needs, for example the search strings in web search engines. In information retrieval not only a single object uniquely identifies a query in the collection, rather, several objects may match the query, but, with different degrees of relevancy. Most of the IR systems compute a numeric score for each object in the database to determine how well each of them matches the query, and then it rank the objects according to this calculated value. After ranking, objects having top ranks are shown to the user. The user can then iterate the process by refining the query, if required.
Use of ranking :

1. To improve search quality.
2. To do effective retrieval over large collections.

3. Granting relevant,efficient, fast and quality information against the user query.

RELATED WORK

In this paper, a review of previous work on ranking is given. In the field of ranking, many algorithms and techniques have already been proposed but they all seem to be less efficient in efficiently granting the rank. The various algorithms are defined below.

.

Page Rank Algorithm

Page Rank Algorithm [11] is one of the most common ranking algorithms. It is a link analysis algorithm which provides a way of measuring the importance of pages. Its working is based on the number and quality of links to a page to make a rough estimate of the importance of the page. It is based on the assumption that more important pages are will receive more links from other pages. The numerical weight that it assigns to any given element E is referred to as the Page Rank of E and is denoted by PR (E).

HITS Algorithm

Hyperlink-Induced Topic Search[12] (HITS; also known as hubs and authorities) is a link analysis algorithm that rates pages. In links and out links of the web pages are processed to rank them. A good hub represents a page that pointes to many other pages, and a good authority represents a page that was linked by many different hubs. The scheme therefore assigns two scores for each page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages. HITS algorithm has the limitation of assigning high rank value to some popular pages that are not highly relevant to the given query.

Weighted Page Rank Algorithm

Weighted Page Rank algorithm (WPR)[1] is an extension to the standard Page Rank algorithm. The importance of both in-links and out-links of the pages are taken into account. Rank scores are distributed based on the popularity of the pages .Number of in-links and out-links are observed to determine the popularity of a page. This algorithm performs better than the conventional Page Rank algorithm in terms of returning a large number of relevant pages to the given query.

Weighted Links Rank Algorithm

Weighted links rank (WLRank) [13] algorithm is a variant of Page Rank algorithm. Different page attributes are considered to give more weight to some links, for improving the precision of the answers. Various page attributes which are considered for assigning the weight are: tag in which the link is contained, length of the anchor text and relative position in the page. The use of anchor text is the best attribute of this algorithm.

Distance Rank Algorithm

It [5] is an intelligent ranking algorithm based on learning. In this algorithm, the distance between pages is calculated. The distance can be defined as the number of ‘‘average clicks’’ between two pages. It considers distance between pages as a punishment and therefore aims at minimizing this distance so that a page with less distance will get a higher rank. The Advantage of this algorithm is that it can find pages with high quality and more quickly with the use of distance based solution. Also, the complexity of Distance Rank is low. The Limitation of this algorithm is that it requires a large calculation to calculate the distance vector.

Time Rank Algorithm

This algorithm[3] utilizes the time fact or to increase the accuracy of the web page ranking. In this the rank score is improved by using the visit time of the page. The visit time of the page is measured after applying original and improved methods of web page rank algorithm to know about the degree of importance to the users. Time factor is used in this algorithm to increase the accuracy of the page ranking. It isa combination of content and link structure. It provides satisfactory and more relevant results.

Query Dependent Ranking Algorithm

This algorithm [10] is used to point out a large variety of queries. The similarities between the queries are measured. The ranking of documents in search is conducted by using different models based on different properties of queries. The ranking model in this algorithm is the combination of various models of the similar training queries.

Categorization by context

This approach proposes [8] a ranking scheme in which ranking is done on the basis of context of the document rather than on the terms basis. Its task is to extract contextual information about documents by analyzing the structure of documents that refer to them. It uses context to describe collections. It is used to overcome the disadvantages of term based approach.

III. Proposed Architecture

The architecture of the proposed ranking system is given in figure-1. The indexer module creates a context based index. This index is used by ranking module to calculate the tf/idf weight and document-context-relation weight of the document. Finally the two weights sums up and find the final weight and rank of the document in the result list.

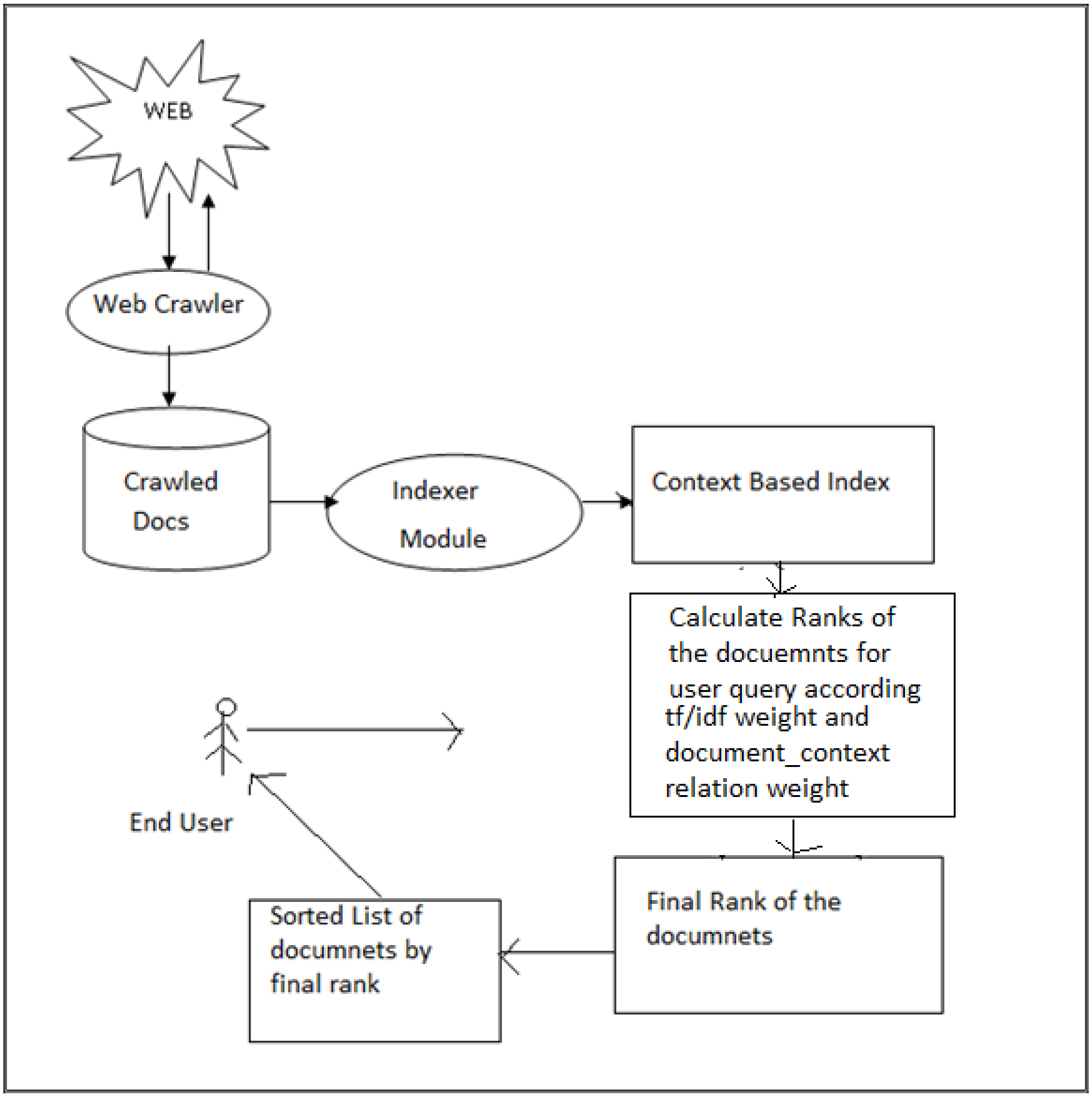
WEB
Web Crawler
Crawled Docs
Indexer Module
Context Based Index
Calculate Ranks of the docuemnts for user query according tf/idf weight and document_context relation weight
End User
Final Rank of the documnets
Sorted List of documnets by final rank

Figure-1 Architecture to combine tf/idf weight and proposed document-context-relation weight.

Description of various components

Crawler

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

Indexer

The job of the indexer is to parse the documents of the page repository and make entry of every token in the index. It also assign context to a document. Finding context of a document is not the area of concern of this paper. It is assumed that a component of the indexer will do this job.

Weight Calculation and Ranking Module

This module will calculate the tf/idf weight and document-context-relation weight of the document. These two weights will be combined to calculate the final rank of the document.

Document-context-relation

This weight will be calculated in the background by the ranking system for every document. This weight will indicate the goodness of the document in the given context. If a particular document is good in a context then this weight will be high for that document.

IV. CONCLUSION

A large number of algorithms are present today which can be used for ranking the pages in Informational Retrieval System. There will always be a scope of better ranking of pages as each algorithm has its associated advantages and disadvantages. In this paper a new architecture of IR system is proposed. The ranking module in the new proposed architecture will use the old tf/idf weight of the terms and also the context based index to find the final rank of the documents for a user query. The index will be a context based index that will be created using thesaurus. The design of formula for document-context-relation weight will be done in future.

REFERENCES

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

[2] Ricardo Baeza-Yates and Emilio Davis ,"Web page ranking using link attributes" , In proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, PP.328-329, 2004.

[3] H Jiang et al., "TIMERANK: A Method of Improving Ranking Scores by Visited Time", In proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, 12-15 July 2008.

[4] Jon Kleinberg, “Authoritative Sources in a Hyperlinked Environment”, In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998.

[5] Ali Mohammad ZarehBidoki and Nasser Yazdani, “DistanceRank: An Intelligent Ranking Algorithm for Web Pages”, Information Processing and Management, 2007.

[6] Dilip Kumar Sharma and A. K. Sharma,“ A Comparative Analysis of Web Page Ranking Algorithms”, in International Journal on Computer Science and Engineering, 2010.

[7] Giuseppe Attardi and Antonio Gullì, “Automatic Web Page Categorization by Link and Context Analysis”,

[8] ParulGupta and Dr. A.K.Sharma, “Context based Indexing in Search Engines using Ontology”, 2010 International Journal of Computer Applications.

[9] AbdelkrimBouramoul, Mohamed-Khireddine Kholladi1 and Bich-Lien Doan,, “ USING CONTEXT TO IMPROVE THE EVALUATION OF INFORMATION RETRIEVAL SYSTEMS” International Journal of Database Management Systems, May 2011.

[10] XiuboGeng, Tie-Yan Liu, Tao Qin, “Query Dependent Ranking Using K-Nearest Neighbor”, *SIGIR’08,* July 20–24, 2008, Singapore

[11] 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.

[12] 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.

http://en.wikipedia.org/wiki/PageRank#Description

AUTHORS

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)

Vinod Jain is working as a lecturer in information technology department at B.S.Anangpuraia Institute of Technology and Management, Faridabad since September 2008. 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)