IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

# Effective Data Retrieval Using Relevance Feedback and Local Search Approximation Algorithms

**Mathumathi .B[1], Darsana K Gopidas[2]**

Assistant Professor, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India [1]

M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India [2]

**Abstract:** In search engines, data/Information Retrieval (IR) is affected with indexing and retrieving documents with user's requirement. The documents which are retrieved should most relevant to the users need. In order to improve the retrieval efficiency, RF (Relevance Feedback) methods are used. The query expansion with effective data retrieval is the main aim of our proposal. In this paper, we propose a local search approximation algorithm to re-weight the query terms and to re-rank the document retrieved by an IR and it solves the computational hard optimization problems. Additionally, we propose a new indexing method named as Set Inverted Index, which is a semantic extracted term based inverted index. This helps to summarize the documents and sum up with its semantic similarity. This indexing system outperforms than other inverted index methods.

**Keywords:** Information Retrieval systems, search engine, relevance feedback, local search approximation, inverted index.

## I. INTRODUCTION

Search engine is one of the most effective and prominent method to find information online. It has become an essential part of life for almost everyone to search desired information in the various fields such as business, entertainment, research etc [1]. In IR system, hundreds of thousands of documents are returned in response to a user query and this huge number of links makes it very difficult for the user to select the page of their interest and need.

The search engines build their database by crawling web pages periodically and indexing the Web pages that are suitable to be added in the database. When a query is submitted by the user to search engine, the appropriate match of the query term with database is carried out by the search engine with the help of complicated searching algorithms. The searching algorithms vary from one search engine to other. At last, the retrieved documents are ranked according to the relevancy of the document with the query [2]. In our proposed system the returned documents are kept in meaningful groups and labels based on semantic similarity, this will be more convenient to the user's to select among the different groups instead of selecting and viewing all links.

### 1.1 RETRIEVAL MODELS

There are several retrieval models to improve the retrieval process. The various information retrieval models classified into the following categories [3]:

1. User centric or cognitive models
2. System centric models
3. Alternative models

The user centric model also consider ways in which the query is formulated in the form of user information needs, the human computer interaction during the search process [4], the environment which the search is carried out and the way in which the information is used to meet specific information need in addition to retrieval mechanisms used in matching queries.

The system centric model is based on logical and mathematical principles such as probabilistic model, Boolean search and vector processing models [5]. In probabilistic model, the search is carried out by comparing the relevance probabilities of the documents while queries are compared with terms which are used to represent the documents in case of Boolean search model. The global similarity between queries and set of documents is compared case of vector processing model.

#### a. Best match searching and relevance feedback model

The purpose of best match searching is to create the ranked out which necessitates to calculate the relative significance of retrieved items which in turn requires weighting the search terms in one or the other way. A similarity consists of two main components:

1. A term weighting scheme that indicates the significance of a term by assigning numerical values to each index term in the document or query.
2. A similarity coefficient which uses these weights to compute the similarity between query and retrieved item. Each query term is compared against the each term in the database in case of best match search technique, the

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

measure of similarity is calculated between the term in the document and the query and finally all the items retrieved so far are sorted with decreasing similarity values. The ranking of the documents involves some sort of quantitative measurement [6]. The various weighting schemes are used to produce best results such as term frequency and collection frequency [7].

## II. PROBLEM DEFINITION

The current searching scenario has many challenges and hurdles due to the huge number of documents. The explicit relevance feedback of each user is a problematic one. But the earlier paper shows that's not an obstacle with pseudo RF. IR methods and Inverted indexes[8][9] are widely used to efficiently retrieve data related to the keyword queries in most search engines, with techniques designed to compress the inverted indexes. But, the interval trees are not good for keyword search because:

(1) An interval tree is needed for each word, which increases the index size;
(2) Interval trees cannot be easily compressed; and
(3) Interval trees cannot support multi-way merging and probing, which are important for accelerating calculations.

## III. LITERATURE SURVEY

**Aji, Y. Wang, E. Agichtein, and E. Gabrilovich[10]** introduced a novel term weighting scheme that uses Revision History Analysis (RHA) of the document edit history. Unlikely previous models it directly captures the document authoring process when accessible, and is particularly valuable for collaboratively generated substance, particularly Wikipedia contents. In this paper, web ranking with adequate techniques are left for future work.

**Blanco and P. Boldi [11]** proposed a way to extend the probabilistic relevance framework with a notion of virtual region based on the use of operators applied to the query. The method has room for improvements and further study should be undertaken to understand which operators are more useful and under which circumstances. This uses evidence of relevance of the document. This technique fits nicely in the previous IR model behind BM25 (which is a ranking function);

**I. Frommholz, B. Larsen, B. Piwowarski, M. Lalmas, P. Ingwersen, and K. van Rijsbergen [12]** have shown how incorporating reading level features for users and documents can provide a valuable new signal for relevance in Web search. In this paper, the authors explored three key problems in improving relevance for search using reading complexity: estimating models of user reading proficiency, estimating models of result difficulty, and combining relevance and difficulty signals to re-rank based on the difference between user and result reading level.

The authors also provided a large-scale analysis of log data to characterize certain aspects of user behavior and classes of features and queries that were likely to be executive in personalization using reading difficulty predictions. Using this process, the web search results could be ranked with the introductory material first, followed by increasingly technical material. Other advances such as level appropriate query suggestions, result snippets, and site recommendations are also possible. But this increases the total iteration for document retrieval.

## IV. PROPOSED SYSTEM

Query expansion is the process of refining the information needed by the user and it also has the task of query re-weight.the re evolved query terms will be applied again in the retrieval function. In our proposed system, the query expansion with effective data retrieval is proposed. In this paper, we propose a local search approximation algorithm to re-weight the query terms and to re-rank the document retrieved by an IR and it solves the computational hard optimization problems. Additionally, we propose a new indexing method named as Set Inverted Index, which is a semantic extracted term based inverted index. This helps to summarize the documents and sum up with its semantic similarity. This indexing system outperforms than other inverted index methods.

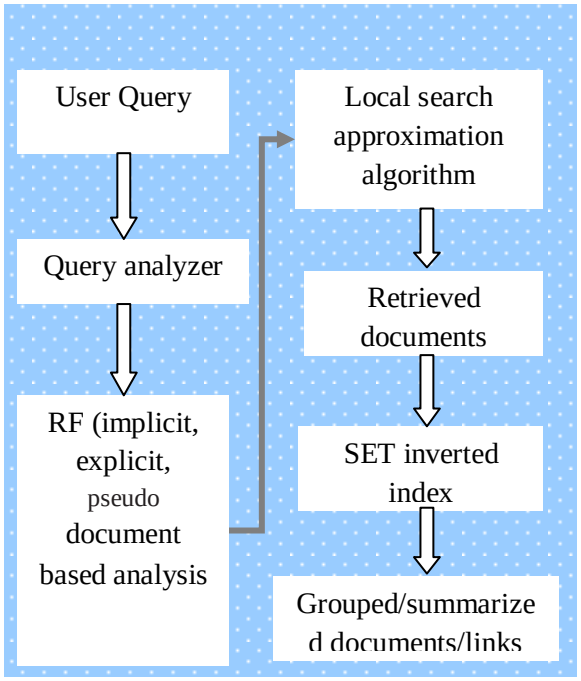


**Fig 1.0 proposed user query based document retrieval system**

The above fig 1.0 represents the overall process of the proposed system. Initially the user provides a query. For example if the users gives "what is the features of apple" , the query will be analyzed using query analyzer, this will eliminates the stop words and auxiliary words. After the

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

preprocessing stage the RF will be implemented. The user's implicit and explicit feedbacks are analyzed along with the pseudo documents. The pseudo documents are created from the list of key terms and associated words of the documents. Using the RF, the local search will be initiated. For optimal results, the local search approximation algorithm is used. This reduces the irrelevant documents at the time of search. Key word matching, key pair matching and sentence analysis are the major part of this searching process. After the successful document retrieval, we have a post processing technique to improve the results. In order to group the relevant documents together, the SET inverted Index (SII) has been used.

## V. IMPLEMENTATION

In this paper, we focused on the problem of automatically extracting and summarization of data records and indexing them according to the similarity and interval. This chapter describes the implementation process of our proposed system. Implementation is the realization of an application, or execution of plan, idea, model, design of a research. This section explains the software, datasets and modules which are used to develop the research.

### a. EXPERIMENTS

It presents the experimental results for SII (**S**et (**S**emantically Extracted **T**erm based) **I**nverted **I**ndex) over several datasets which are described below. Some links may provide complete extraction availability and some links are limited to the extraction process. Proposed system is implemented in Asp.Net with C#.

### i. Data Sets

The system used a dynamic dataset, which can be any number of user specified web documents created by the own. Pages from internet are the primary sources of datasets used for the experiments. The following datasets are used in the experiments to compare the performance of **SII** with the existing methods. The following dataset contains some URLs extracted from the website and description about the extracted URLs.

The extraction of links, tags, and values with semantic labels are implemented. Extraction process uses some notations and concepts in order to get the URL and the data. The system initially extracts the page source by applying the extract html function which is coded using C#.net. The system uses structure based data extraction from the web page source.

The web source may have more unwanted page contents, which the user doesn't wants. The data units in some composite text nodes are separated by blank spaces created by consecutive HTML entities like " " or some formatting HTML tags such as <SPAN>. Second, the data units of the same attribute across different SRRs may sometimes vary a lot in terms of appearance or layout.

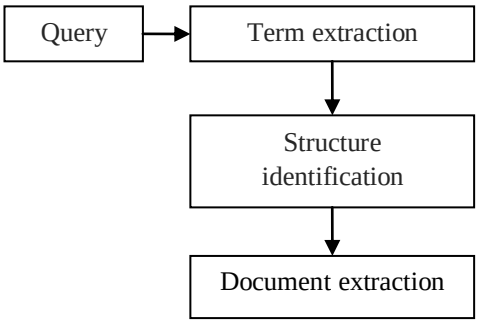


**Fig 2.0 query term extraction and structure identification process**

### ii. Data extraction process

In the project, links will be extracted from the keyword given by the user. The system will perform SII process and brings the results with index structure. The proposed system performs the following steps in result generating process. The following are the steps involved in the implementation process of the proposed system.

### iii. User Search

In this step, the user login into the process by giving their user name and other details, then it transfers to the query search page. Here the user specifies the query and searches for the information.

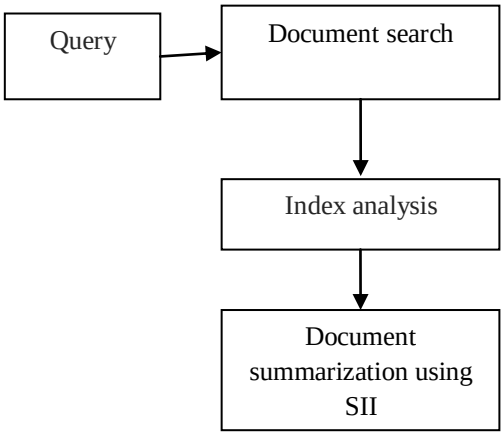


**Fig 3.0 Document summarization using SII**

### iv. . Local Monitoring

Local monitoring is the process of monitoring data about user searching query. It collects the details about the user searching result. Local monitoring information is obtained using the improved scan line algorithm.

| domain | data | datalink |
|---|---|---|
| american | United States | http://www.conservapedia.com/United States |
| android | operating system | http://www.conservapedia.com/operating system |
| android | Open Handset Alliance | http://www.conservapedia.com/Open Handset Alliance |
| android | Google | http://www.conservapedia.com/Google |
| android | [1] | http://www.google.com/[1] |
| android | [2] | http://www.google.com/[2] |

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

The system initially retrieves the documents which contains the keyword specified by the user.

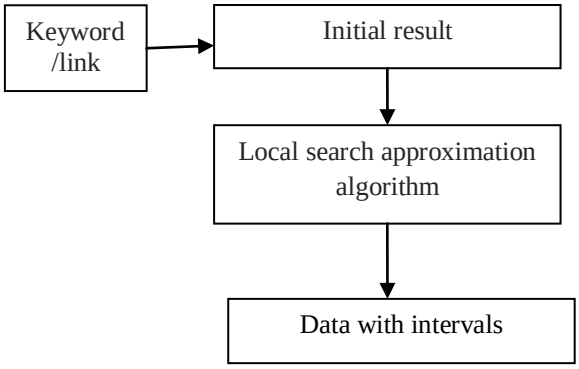


**Fig 4.0 Document summarization using SII**

**Relevant Data Extraction using Query expansion**

In this step, the data generated with respect to user's query are extracted. The search engine will generate many contents which may be relevant or irrelevant. Data Extraction methods will extract only the relevant data by caching the auxiliary data and eliminating redundant and irrelevant data. These relevant data will be added to table query record result set. In order to reduce the irrelevant results the system using ranking concepts for popular data retrieval. This helps to extracts the most reliable data from huge results. Ranking process has been used based on popularity, so frequent search words will be specified with high rank.

**Data Index Identification**

Data Index Identification module identifies all possible data sequence indexes in web results which is generated dynamically from searched data. This step mines every data index in a web page that contains similar data results. Before identifying all the indexes in a query result page, it is necessary to determine whether any data regions should be merged. After identifying the data regions, this step merges the data index sequences that contain similar data records.

**SET Inverted index generation**

The sequence index Segmentation module then segments the identified data indexes into data records according to the keyword patterns in the data indexes .Record inverted index segmentation first finds the appropriate index. The visual result gap has been identified and specified by the starting and ending indexes, this is simply known as inverted index.

A keyword search system usually supports union and the intersection operations on inverted lists. The union operation is a core operation to support OR query semantics in which every document that contains at least one of the query keywords is returned as a result. The intersection operation is used to support AND query semantics, in which only those documents that contain all the query keywords are returned.

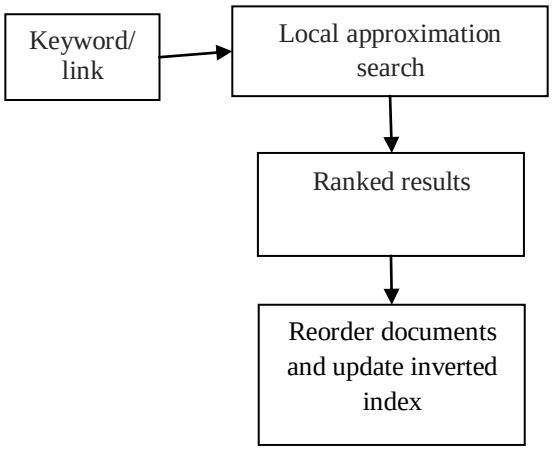


**Fig 5.0 Re ranked Documents**

Document reordering increases the performance of SET-inverted index for finding the exact keyword. This section first explains the necessity of document reordering. Then, since finding the best order of documents is NP-hard, a sorting-based method is used to find near optimal solutions. The following details show the performance results of the proposed system.

**Table 1.0 performance analysis table**

| Parameter (query size/ Time in ms) | Query expansion time | Document retrieval time |
|---|---|---|
| 10 | 800 | 950 |
| 20 | 1650 | 1900 |

From the above table 1.0, the following chart has been generated. The chart shows the time taken for query expansion and document retrieval in the proposed system.

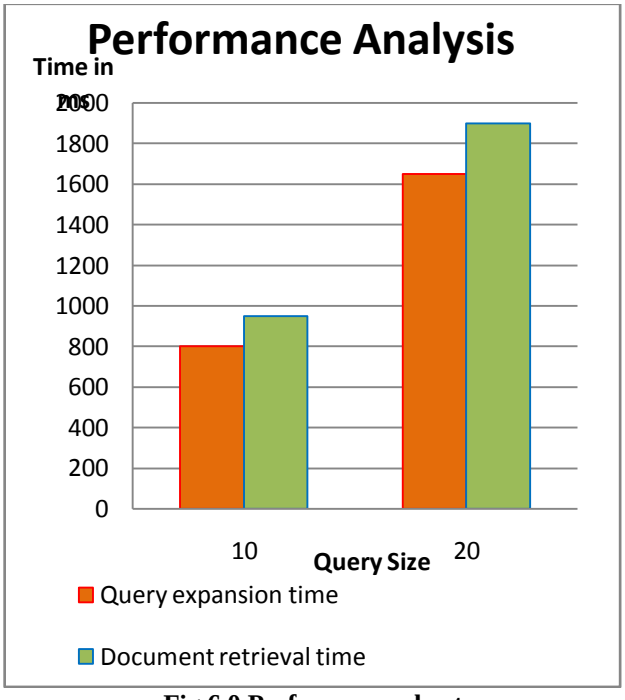


**Fig 6.0 Performance chart**

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

## VI. CONCLUSION

In this paper, a new RF algorithms and indexing methods have been proposed for effective document retrieval. The re-weight query terms by projecting the query vector on the subspace represented by the local search approximation is the main contribution, which helps to find the optimal solution to the problem of finding the appropriate document. First, the documents retrieved by an IR system to answer the original query are used to extract a feature matrix. This utilizes the local search approximation algorithm; second, some relevance assessments are obtained according to whether RF is implicit and explicit or pseudo. The quantum probability distributions can be estimated and the optimal solution of a distance between two quantum probability distributions can be calculated. The results and experiments show the proposed system yielded better result.

## REFERENCES

[1] Halpin, Harry, and Victor Lavrenko. "Relevance feedback between web search and the semantic web." IJCAI Proceedings-International Joint Conference on Artificial Intelligence. Vol. 22. No. 3. 2011.
[2] Salton, Gerard, and Chris Buckley. "Improving retrieval performance by relevance feedback." Readings in information retrieval 24.5 (1997): 355-363.
[3] Hiemstra, Djoerd. "Information retrieval models." Information Retrieval: searching in the 21st Century (2009): 1-19.
[4] Spink, Amanda. "A user-centered approach to evaluating human interaction with web search engines: an exploratory study." Information processing & management 38.3 (2002): 401-426.
[5] Salton, Gerard, Edward A. Fox, and Harry Wu. "Extended Boolean information retrieval." Communications of the ACM 26.11 (1983): 1022-1036.
[6] Fuhr, Norbert. "Probabilistic models in information retrieval." The Computer Journal 35.3 (1992): 243-255.
[7] J. J. Rocchio and G. Salton, "Information search optimization and interactive retrieval techniques," in Proce. Fall Joint Comput. Conf., 1965, pp. 293–305.
[8] G. Salton, "Associative document retrieval techniques using bibliographic information," J. ACM, vol. 10, pp. 440–457, 1963.
[9] Scholer, Falk, et al. "Compression of inverted indexes for fast query evaluation." Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2002.
[10] Aji, Y. Wang, E. Agichtein, and E. Gabrilovich, "Using the past to score the present: Extending term weighting models through revision history analysis," in Proc. 19th ACM Conf. Inf. Knowl. Manage., 2010, pp. 629–638.
[11] R. Blanco and P. Boldi, "Extending BM25 with multiple query operators," in Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2012, pp. 921–930.
[12] Frommholz, B. Larsen, B. Piwowarski, M. Lalmas, P. Ingwersen, and K. van Rijsbergen, "Supporting polyrepresentation in a quantum- inspired geometrical retrieval framework," in Proc. 3rd Symp. Inf. Interaction Context ins, 2010, pp. 115–124.