ISSN:0975-9646

Efficient Hybrid Ranking Algorithm for Search Engine

Nirali Arora[#1], Sharvari Govilkar[*2]

[#]*Computer Engineering Department, PIIT Mumbai University*
New Panvel India

Abstract— **Searching is a characteristic behaviour in our general life .Millions of users communicate with search engines on day to day basis .They are checking on the links of results ,clicking on ads and bookmarking the set of web pages .The largest of all challenges is to find best ranked pages. This paper focuses on series of ranking algorithms and provides a methodology to combine these algorithms in order to produce a more relevant ranked results. A new approach also called re-ranking is presented in order to produce filtered and relevant ranked results to improve user satisfaction and goals.**

Keywords— ***Information Retrieval, content based ranking, usage based ranking, link based ranking, re-ranking, hybrid ranking .***

I. INTRODUCTION

Ranking is a critical activity in reordering of the search engine results .Information retrieval is a framework that defines and explains the major phases of information retrieval process including document representation and user query representation and ranking of retrieved documents. The major challenge of the search engines is to build communication between its users. Most of the web users do not know how to represent the areas of interest in form of queries. The traditional web is of enormous size . A lot of search results that corresponding to a user's query are irrelevant to the user need. Most of the page ranking algorithms use Link-based ranking (web structure mining) or Content-based ranking to calculate the overall relevancy of the information according to the user's need, but those ranking algorithms might be not suitable enough to provide a good and useful ranked list. . There are four kinds of conventional IR models, namely Boolean model, Vector Space model, Probabilistic model, and Language model.

The organization of the paper is as follows :Section I gives introduction to ranking and its different algorithms, Section II gives an insight to the literature survey ,Section III explains the problem statement and Section IV explains about the proposed system and the modules and Section V concludes the paper .

The different types of ranking algorithms are as follows

A. *Content-based ranking algorithm*

In this approach is introduced to rank the relevant pages based on the content and keywords rather than keyword and link structure provided by search engines. Based on the user query, search engine results are retrieved. Every result is individually analysed and interpreted based on keywords and content.

B. *Usage based ranking algorithms*

Recommendation algorithms aims at proposing "forward next "pages before a user based on her current visit and the past users' navigational patterns. In the vast majority of related algorithms, only the usage data are used to produce recommendations, whereas the structural properties of the Web graph are ignored. Usage based ranking algorithms consider the past user navigation pattern and thus analyze user behaviours and accordingly recommends the data .

C.*Link based ranking algorithms*

The link analysis algorithm is based on link structure of documents. The quality of the results from search engines are generally lower in quality than what user expects and quality can be improved greatly if the pages are ranked according to a useful criteria based on the links between the pages ie each page may have the reference to other set of pages. They view the web as a directed graph where the web pages form the nodes and the hyperlinks between the web pages form the directed edges between the nodes. There are two famous link analysis methods

1. Page rank algorithm
2. Hits algorithm

II. LITERATURE SURVEY .

In this section we cite the relevant past literature that utilizes the various techniques for ranking. Ranking search results is a common problem in information retrieval. Most common techniques focus on the similarity of query and page as well as overall page quality. However, with increasing popularity of search engines the capturing of user behaviours insists to appear on surface. A lot of methods have been done on implicit techniques of user preference in field of information retrieval.

Madhurdeep kaur and Prof Charanjit Singh [1] presented a novel hybrid page rank algorithm based on both content and link structure of web pages. A set of repository of web pages in the database are selected. A link structure will be created that will explain how these pages are linked to each other. It uses a concept of similarity of content and query. Accordingly both the page rank and content rank will be compared.

Pijitra jomsri and Dulyawit Prangchumpol [2] devised and proposed a hybrid model ranking for research paper searching and bookmarking. The keywords used for ranking are the year of the published paper and type of research paper this is combined with the similarity ranking. The proposed approach uses the concept of indexing using tf-idf.

Sharmila Subudhi [3] proposed the hybrid page ranking algorithm which is based on prediction and calculation of different numbers of backlinks to a particular web page. It uses a concept of coordinate list and compressed sparse row which calculates the number of backlinks to a page using sparse matrices.

Lissa Rodrigues and Shree jaiswal [4] proposed a hybrid approach of enhanced ratio rank and page level keyword algorithm. This approach uses the combination of link based mining and page level keyword search. In enhanced ratio rank both inlinks and outlinks are considered along with the times the user visits a particular link.

Veningston, Dr R Shanmugalakshmi[5] proposed a enhanced web search re-ranking algorithm by incorporating user profile information obtained from various sources .this paper also discusses the problems of present web search .The user profile is created by using the URLs of the initial results. Reranking strategy is based on calculating the personalized score of web page and generating the personalized result set.

Anuradha R Kale, Prof VT Gaikwad[6] proposed reranking method which employs the semantic similarity to improve the quality of search results .By fetching the top n results returned by the search engine and by using semantic similarities between the candidate and query to rerank the results. The search based on lexical semantics instead of keyword matching is more emphasized to eliminate the potential weakness of keyword based search. The ranking position is converted into an importance score and a combination of similarity score with importance score is used to get the ranks.

III. Problem Statement

The problem is on combining different ranking algorithms in order to design an effective hybrid ranking approach that uses a combination of content based, link based and usage based ranking algorithms such that it meets user specific needs and goals. In order to narrow down the ranked list even further to meet the user specific search goals the reordering of the set of top n ranked pages is imperative using re ranking algorithm.

IV. Proposed System

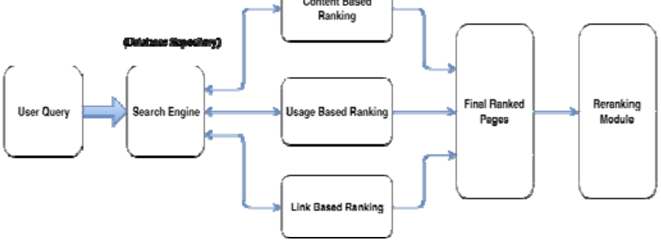

Fig 1 . Overview of proposed system

The proposed approach uses the combination of various modules to produce a set of ranked documents.

1) The user specifies the query in the natural English language. There are the set of 30-35 documents on various diverse topics kept in the backend. Apart from the static documents the data is also extracted from the different web sources (search engines such as Google, MSN etc)
2) When the user enters the query in system API the query is forwarded to the search engine and the top n results are extracted from the search engine.
3) Preprocessing is an important step in the process all the dirty data from the set of documents and the query are preprocessed to form a clean and structured query and data.
4) The top 'n' retrieved results from the search engine module are reordered using content based ranking. A page score is assigned on the basis of keyword matching and user query. The values of these scores are assigned in descending order. [10]
5) The results of the search engine which are retrieved are then passed through a link based ranking module which assigns the page score on the basis of enhanced normalized page ranking algorithm [17]
6) The results of both content and link are passed to usage based ranking module accordingly [14] which then assigns the score to each of pages
7) A score corresponding to each module is averaged together to form a composite score and the pages are ranked and filtered in descending order of composite scores. [14]
8) The set of ranked pages are displayed to the user.
9) The set of the ranked pages are again ranked using re-ranking module that uses the user profile creation and then finally the re-ranked list is displayed to the user. [5]

A. Query generation , data extraction and data cleaning module

1) User is asked to specify a query in a natural language. Data extraction is the first stage of our project. A set of 25-35 static documents on various topics would be stored in the backend.

We would extract dynamic data from a set of search engines and web sites since the data on web is highly unstructured it is necessary to clean it using the set of pre-processing based techniques.

2) Pre-processing is an important step in data cleaning. Both query and data obtained are pre-processed using the following algorithm accordingly Tokenization is a critical activity in any information retrieval model, which simply separates all the words, numbers, and their characters etc. from given document and these identified words, numbers, and other characters are called tokens. Along with token generation this process also evaluates the frequency value of all these tokens present in the input documents.

3) Pre-processing involves the set of all documents are gathered and passed to the word extraction phases in which all words are extracted.

4) In next phase all the infrequent words are listed and removed for example remove words having frequency less than two. Intermediate results are then passed to the stop word removal phase. In this phase remove those English words which are useless in information retrieval these

English words are known as stop words. For example, stop words include “the, as, from, of, and, or, to etc. this phase is very essential in the Tokenization because it has some advantages: It reduces the size of indexing file and it also improves the overall efficiency and make effective.

5) Next phase in tokenization is stemming. Stemming phase is used to extract the sub-part i.e. called as stem/root of a given word. For example, the words continue, continuously, Continued all can be rooted to the word continue. The main role of stemming is to remove various suffixes as result in the reduction of number of words, to have exactly matching stems, to minimize storage requirement and maximize the efficiency of IR Model.

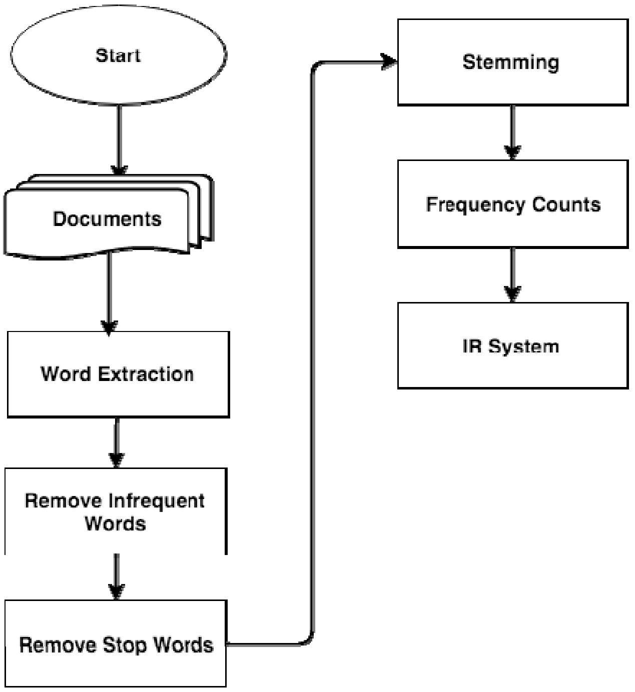

Fig 2 .Steps of data cleaning

```
Tokenization Algorithm [16]
Input(Documents)
Output (Tokens)
Begin
Step1: Collect input documents where i=1,2,3…n
Step 2: For each input Di;
        Extract word(EWi)
Step 3: For each EWi;
        Apply Stop word (SWi)=EWi;//apply stop word elimination accordingly
        Stemming (Si)=SWi;//to create stems
Step 4: For each Si;
        Freq_count(Wci)=Si;
Step 5: Tokens (Si ) will be passed to an IR System
End
```

B .*Link Based Ranking Module*

Results that are retrieved from search engine and set of clean documents in the database repository are given to link based ranking module.

In this proposed approach an optimized page rank algorithm based on normalization technique has been proposed. In this proposed scheme the page rank of all web pages are being normalized by using a mean value factor, which reduces the time complexity of the conventional page rank algorithm.

The proposed algorithm is based on normalized page rank approach is as follows[4]

```
Inputs: A set of web pages with inlinks and outlinks
Output: reordered results with page rank score
1) Initially assume PAGE RANK of all web pages to be any value, let it be1.
2) Calculate page ranks of all pages by following
Formula PR(A) = .15 + .85 (PR(T1)/C(T1) + PR(T2)/C(T2) +……. + PR(Tn)/C(Tn)) Where
T1 through Tn are pages providing incoming links to
Page A PR(T1) is the Page Rank of T1
        PR(Tn) is the Page Rank of Tn
        (Tn) is total number of outgoing links on Tn
3) Calculate mean value of all page ranks by following formula: -
  Summation of page ranks of all web pages / number of web pages
4) Then normalize page rank of each page
Norm PR (A) = PR (A) / mean value
   Where norm PR (A) is Normalized Page Rank of page A
        And PR (A) is page rank of page A
5) Assign PR(A)= Norm PR (A)
6) Repeat step 2 to step 4 until page rank values of two consecutive iterations are same The pages which have the highest page rank are more significant pages.
```

C *Content Based Ranking Module*

In the proposed work a new approach is introduced to rank the relevant pages based on the content and keywords rather than keyword and page ranking provided by search engines.

Based on the user query, search engine results are retrieved. Every result is individually analyzed based on keywords and content. User Query is pre-processed to identify the root words. Every root words are considered for Dictionary construction and Dictionary is built with synonyms for the user query. Every result page keywords and content words are pre-processed and compared against the dictionary. If a match is found, then particular weight is awarded to each word. Finally, the total relevancy of the particular link against user request is computed by summarizing all the weights of the keyword and content words. The page which contains total relevancy value nearest to 1 are ranked as first page and 0 are ranked as last page. Part of speech tagging is an imperative part of content based ranking. A POS tagger is a piece of software that reads the text in some language and assigns part of speech to each word such as noun, adjective etc. Out of all these only noun and adjectives are stored into the server log.

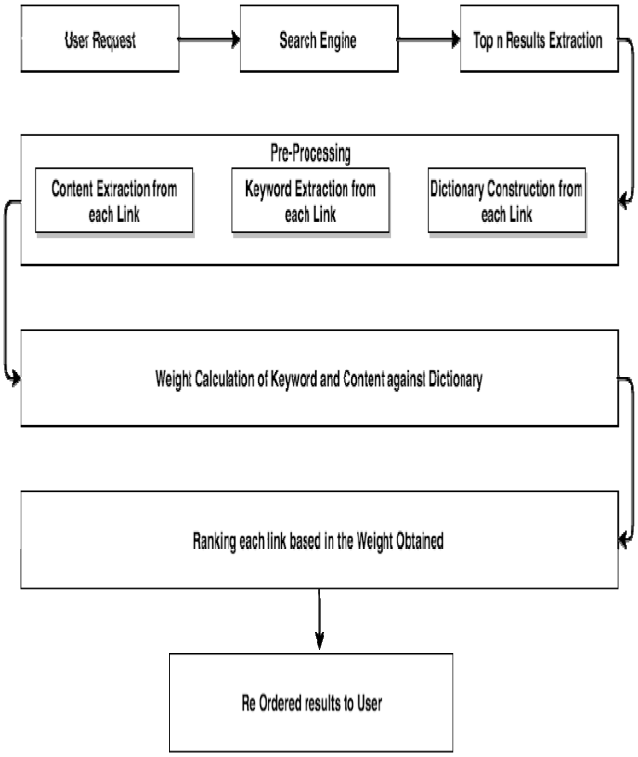
User Request
Search Engine
Top n Results Extraction
Pre-Processing
Content Extraction from each Link
Keyword Extraction from each Link
Dictionary Construction from each Link
Weight Calculation of Keyword and Content against Dictionary
Ranking each link based in the Weight Obtained
Re Ordered results to User

Fig 3 . Content based ranking .

Algorithm for Content Based Ranking : Relevancy and Weight based approach[10]
Input: Extracted Web/document contents
Output: Reordered Web Content on content score of each page

```
Step 1. Extract Search Engine results SRi for the user query where 1<i<N
Step 2. Pre-process user query and extract root words RWj where 1<j<N
Step 3.Construct Dictionary D for the user query RWj
Step 4. Extract and Pre-process the keywords KWi for the search results SRi
Step 4.a. Compute Keyword Strength S(KWi) =1/Σ KWi
Step 5. Extract and Pre-process the Content words CWi for the search results SRi
Step 5.a: Compute Content Words
Strength S(CWi ) =1/Σ CWi
Step 6. Compare each keyword KWi against Dictionary D.
Step 6.a. If match is found, then award strength S(KWi) to particular keyword
Step 6.b. Else award 0 as a strength for particular keyword.
Step 7. Compare each content word CWi against Dictionary D
Step 7 If match is found then award Strength S(CWi) to particular content word
Step 7.b. Else award 0 as a strength for particular content word.
Step 8. Calculate Total Strength for Keyword TKS(SRi) = ∑S(KWi)
Step 9. Calculate Total Strength for Content Word TCS(SRi) = ∑S(CWi)
Step 10. Compute Total Relevancy for the particular link TRi = TKS(SRi)*(Wt) + TCS(SRi)*(1- Wt)where 0 < Wt < 1
Step 11. Repeat step 4 to 10 for all Search Results(SR)
Step 12. Sort the result set SR based on TRi in Descending order.
```

The Topmost Search Result SRi is the most relevant for the user query and bottom most search result is the least relevant for the User query

D Usage based ranking module

Usage based ranking is never done in alienation .It depends on content based ranking and link based ranking . Usage-based ranking algorithms score documents by how often they are viewed by Internet users . For Usage-based ranking, there are very limited works to utilize the usage data in the web information retrieval systems, especially in the ranking algorithm. In this stage we calculate several parameters which are the inputs to our algorithm .The two important parameters are frequency of visit to a particular page and time spend on a particular page. It's uses the combination of the previous modules to provide a new weight called usage based weight for the pages, then ranking those pages. Re-ordered result list obtained through link based and content based ranking are presented to the user and user is asked to select any one of the results. The user is totally unaware about the proposed and objectives of the project. User enters his choice. The user behaviour is carefully monitored using web server log. The required data is now stored in the server log. Depend on user's choice the results are now re-ordered by using the approach of usage ranking. For storing the history of user access behaviour the server log is maintained. The web server log is implemented using a WAMP server. The server log stores the time spend by a user on a particular page using Ajax and jQuery commands or by using event handlers in cookies. A web server log maintains the session using session variables accordingly.

Algorithm of usage based ranking module
Input: Set of results retrieved from search engines, query q.
Output: Set of reordered results on the basis on uscore (usage ranking score)

```
Step 1: For each corresponding query a match of same query is checked in web server log.
Step 2: The match of query indicates that data is previously accessed
Step 3: Based on the records of query corresponding results are retrieved and sorted in descending order of Number of clicks and time spend on a particular page.
Step 4: Accordingly, on the descending order of sorted results the scores are assigned in order of serial numbers.
Step 5: For subsequent matches the serial numbers which are page scores are incremented accordingly.
```

E. Composite score creation for ranked list display

From the previous modules the web scores from link based ranking ,content based ranking and usage based ranking would be obtained and accordingly these scores are arranged together in ascending order and a median of the scores obtained are considered as composite score that would be used to produce a final ranked list for display .

F. Reranking module [6]

Re-ranking of the results reflects the most relevant results for the user. It is a process of re-ordering the retrieved results based on user preferences. Re-ranking computation performs the following two processes. They are,

- Calculating personalized score for page
- Generating personalized result set

Click entropy [6] of the query is computed as follows,

Click entropy $(q)= \sum P(p|q)\log P(p|q)$ *………eqn(1)*

Where, summation is the collection of the web pages click on query q

And log P(p|q) is the percentage of clicks on page p among all clicks of q.

Smaller click entropy means that majorities of users agree with each other on smaller number of web pages.

Re ranking on historical clicks (P clicks)

Pclick $(q,p,u)=(|clicks(q,p,u)|/|clicks\ (q,*,u)|+\beta)$ *…eqn(2)*

Where (clicks (q,p,u)): is number of clicks on web page p by user u on query q in the past .

Clicks(q,*,u) is total number of clicks on query q by user u .
B is the normalizing factor.

Matching and reranking algorithm [6]

The proposed algorithm aims to provide re-ranking of the retrieved documents so as to bring most relevant document in first few pages on which users would be much interested in

Inputs

a. User profile generated using the following parameters
1)Url (id)
2) Title(T)
3)Body text of each web page (FT)
4)Snippet Summary (SS)
5)Extracted terms from the snippets (SS terms)
6)Whether URL is clicked previously (Pclick)
7) Dwell time (Dtime)
b) A set of search results
c)A input Query

Output: Re-ranked search results

```
Variable used
--- //set of documents
q-input query string-temporary similarity document set
algorithm matching reranking ()
for each document =calculate sim(,q)
for each document
score=match (Profile (id,T,Fterms,SSterms ,Pclick,Dtime);
end
end
```

V Conclusion

The World Wide Web (WWW) is rapidly and exponentially growing on all aspects and is a massive, explosive, diverse, dynamic and mostly unstructured data repository. As on today WWW is the huge information repository for knowledge reference. There are a lot of challenges in the Web: Web is large, Web pages are highly semi structured, and Web information tends to have diversity in terms of meaning, degree of quality of the information extracted and the conclusion of the knowledge which is extracted from the information. So it is important to understand and analyse the underlying data structure of the Web for efficient information retrieval. Thus web search ranking algorithms play a vital role in ranking of the web pages so that we could retrieve the web pages that are relevant. With the rapid growth of the information sources we are drowning in data but starving for knowledge ,therefore it has become necessary for the user to use information retrieval techniques and combination of different ranking algorithms to find and extract and filter the desired information. Many of the existing Information Retrieval Systems still relies on various approach of ranking algorithms, like Content-based ranking algorithms, Link-based ranking, or a few of them based on utilize user behaviours via usage-based ranking algorithm. Unfortunately, those ranking algorithms still have some drawbacks to a ranked list provided from some search engines. A combination of these algorithms enables to filter out the redundant results

To gather the useful information from the web page, the user interested in, the log is maintain on the database of WAMP server. From this server log, interest of particular user can be found easily. The inference and analysis of user search goals can have a lot of benefits in improving search engine relevance and user experience which can be achieved by web usage mining while web content mining removes persistence and redundancy problem. This work is the combination of content ,link and usage mining which works better than using any one of them.

Acknowledgment

A very special thanks to the computer department of Pillai college of Engineering Panvel for giving us the opportunity to conduct the research .This paper wouldn't have been possible without the efforts of our principal Dr. RIK Moorthy

References

[1] Madhurdeep Kaur and Charanjeet Singh *A Hybrid page rank algorithms using content link based algorithms* in International Journal of Computer Applications (ijca) 2012

[2] Pijistra Jomsri and Dulyawat Prangchumpol *A hy brid model ranking search result for research paper searching on social bookmarking in* 2013 1st International Conference on Industrial Networks and Intelligent Systems (INISCom)

[3] Sharmila Subudhi *in Designing a hybrid page ranking algorithm for semantic web search engine* in 2013 International journal of scientific research(ijsr)

[4] Lissa Rodrigues ,Shree Jaiswal *Hybrid approach of Enhanced ratio rank and page level keyword algorithm ,in 2013 International Journal of Computer Applications*

[5] Venington K and Dr R.Shanmugalakshmi *Enhancing personalized web search re-ranking algorithm by incorporating user profile published in IEEE 2013*

[6] Hema Dubey and BN Roy *An Improved Page rank Algorithm Based on optimized Normalization technique* in IJCSIT 2013

[7] Samriti Gupta,Alka Jindal *Contrast of Link based Web Ranking Techniques* at ieee 2008 proceedings

[8] Azam Feyznia,mohsin Kahanti *A link analysis based ranking method for semantic web documents* at ieee proceedings of 2010 [5] Jun fang,Lei Guo, *Calculation of weight of entities in ontologies by using usage based information* in ieee proceedings of *2011*

[9] Mercy paul Selvan,ChandraSekar ,A.Priya Darshan *Survey on web page ranking algorithms* in ijca (International Journal of Computer Applications) proceedings of 2012

[10] Guosheng Kang,JianxunLiu *Active Web Service Recommendation Based on Usage History* in ieee proceedings of service computing 2012

[11] Ashlesha Gupta,Ashutosh Dixit,AK Sharma *Relevant document crawling with usage pattern and domain profile based page ranking* in ieee proceedings of 2013

[12] P.Sudhakar,G.Poonkuzhal R.KishoreKumar *A content based ranking for search engines* in the IAENG proceedings of 2013

[13] Shital C Patil,RR keole *Content and usage based ranking for enhancing search engine delivery 2014* in volume 3 issue 7International journal of Science and Research (IJSR)

[14] Safaa I. Hajeer, Rasha M. Ismail, Nagwa L. Badr, M. F. *Taiba An Efficient Hybrid Usage-Based Ranking Model for Information Retrieval Systems and Web Search Engines.* in ieee proceedings of 2015 6th International Conference on Information and Communication Systems (ICICS)

[15] Vikram Singh and Balwinder Saini for *Effective Tokenisation algorithm for Informaton Retreival Systems* in 2015 International journal of computer science and Information technologies

[16] Introduction to Information Retreival "(Html) Cambridge University press 2008

.