# CONTENT AND USER CLICK BASED PAGE RANKING FOR IMPROVED WEB INFORMATION RETRIEVAL

Dhiliphan Rajkumar.T[1],Suruliandi.A[2] and Selvaperumal.P[3]

[1] Research Scholar,Department of Computer Science & Engineering Manonmaniam Sundaranar University,Tirunelveli-627012, India
[2] Professor,Department of Computer Science & EngineeringManonmaniam Sundaranar University,Tirunelveli-627012, India and
[3]Research Scholar, Department of Computer Science & Engineering, Manonmaniam Sundaranar University,Tirunelveli-627012, India

## *ABSTRACT*

*Search engines today are retrieving more than a few thousand web pages for a single query, most of which are irrelevant. Listing results according to user needs is, therefore, a very real necessity. The challenge lies in ordering retrieved pages and presenting them to users in line with their interests. Search engines, therefore, utilize page rank algorithms to analyze and re-rank search results according to the relevance of the user's query by estimating (over the web) the importance of a web page. The proposed work investigates web page ranking methods and recently-developed improvements in web page ranking. Further, a new content-based web page rank technique is also proposed for implementation. The proposed technique finds out how important a particular web page is by evaluating the data a user has clicked on, as well as the contents available on these web pages. The results demonstrate the effectiveness of the proposed page ranking technique and its efficiency.*

## *KEYWORDS*

*Web mining, World Wide Web, Search Engine, Web Page, Page Ranking.*

## 1.INTRODUCTION

### 1.1.Back Ground

Information retrieval today is a challenging task. The amount of information on the web is increasing at a drastic pace. Millions of users are searching for information they desire to have at their fingertips. It is no trivial task for search engines to feed users relevant information. Today, web use has increased across fields such as e-commerce, e-learning, and e-news. Naturally, finding user needs and providing useful information are the goals of website owners [1]. Consequently, tracing user behaviour has become increasingly important. Web mining is used to discover user behaviour in the past, including the content of the web and web pages a user wants to view in the future. A search engine is a piece of software that act as an interface for users to retrieve the data desired from the World Wide Web (WWW). As of now, the WWW is the largest information repository for knowledge, and the quality of a page is defined based on user clicks. It is precisely for that purpose that a lot of page ranking algorithms have been introduced. The page ranking is an algorithm developed by Sergey Brin and Lawrence Page in 1998 and used by Google search engine. The algorithm assigns a numerical value to each element in the WWW for

DOI:10.5121/ijcsa.2015.5608

the purpose of measuring the relative importance of each page, the idea being to give a higher page rank value to a page that is frequently visited by users. The ranking can be done in three ways: keyword popularity, keyword-to-web page popularity, and web page popularity. Keyword-based ranking focuses on the most popular keyword first. The keyword-to-web page popularity records which pages have been selected for a user's search query. The final one determines how frequently a web page is selected by a user worldwide. Page ranking is a great concept that helps determine the importance of a given page among a number of similarly indexed pages. It is basically used to calculate scores of elements of a search system. Traditionally, that concept has been widely accepted for web page ranking and organization of results. Therefore, a number of data retrieval and search systems utilize this concept for organizing their results. The results returned, for the same query at different times in search engines, are the same. In recent times, search engines have been using the page ranking concept so that the values on the web page are not identical at all times, as interest in the page might then vary or change. Table 1 displays the results of user satisfaction with relevant search engines. User expectations are satisfied by using search engines.

Table 1.Foresee Results for the American Customer Satisfaction Index (ACSI) For Search Engines

| **Search engine** | **2012** | **2013** | **2014** | **2015** | **Change in % for last two years** |
|---|---|---|---|---|---|
| GOOGLE | 82 | 77 | 83 | 78 | -5 |
| BING | 81 | 76 | 73 | 72 | -1 |
| YAHOO | 78 | 76 | 71 | 75 | 4 |
| MSN | 78 | 74 | 73 | 74 | 1 |
| AOL | 74 | 71 | 70 | 74 | 4 |

Table 1 makes it clear that there is a change in terms of users' satisfaction values with reference to various search engines, including a drop in Google and Bing percentage values. Hence user interest has to be given due importance since it clearly plays a major role. In the proposed work, user interest is considered and results provided for the given queries. The proposed page rank method first analyzes the contents of documents, employs user search queries for the results, and then optimizes them. Thus the proposed page rank model is a content-based page rank methodology for search result optimization.

## 1.2.Overview of Web Mining

A discussion on page ranking presupposes knowledge of web mining, a data mining technique used to extract information from web documents. The major tasks of web mining are resource finding, information selection, preprocessing, generalization and analysis. First, data are extracted from online or offline text data available on the web. The next step is automatic selection and preprocessing from the retrieved web resources. The third step is an automatic discovery of a general pattern at individual or multiple sites. Finally, the results are validated and the analysis arrived at plays a major role in pattern mining. Types of web mining include web content mining, web structure mining and web usage mining.

### 1.2.1.Web Content Mining

The discovery of useful information forms web content/data/documents, and the process is also known as text mining, which is scanning and mining the text, pictures and graphs of a web page to determine the relevance of the content to a search query, and is related to data mining because lots of techniques used in data mining are also used in web content mining. It is the process of retrieving information from the WWW into a more structured form, and provides results lists to search engines in order of the highest degree of their relevance to the keywords in a query. Also, it does not provide information about the structure of the content that users are searching for or the various categories of documents found.

### 1.2.2.Web Structure Mining

This is the process of discovering a model of the link structure of web pages. For the purpose of generating data, similarities and relationships are established using hyperlinks. Both page rank and hyperlink analysis fall into this category, the idea being to generate a structured summary of a website and a web page. Accordingly, web structure mining can be divided into two kinds to minimize two chief problems the WWW comes up against as a result of the vast amount of information at its disposal. The first problem has to do with irrelevant search results, and the next is the inability to index the vast volume of information provided on the web.

### 1.2.3.Web Usage Mining

This refers to the automatic discovery and analysis of patterns in a click stream and associated data collected or generated as a result of user interactions with web resources on one or more websites. The goal is to capture, model and analyze user behavioral patterns as well as the profiles of users interacting with the web. The patterns discovered are usually represented as a collection of pages, objects or resources frequently accessed by a group of users with common needs or interests. Table 2 represents web mining categories with views of data, main data, representations and methods of web content mining, web structured mining and web usage mining.

Table 2. Web Mining Categories

| | **Web Mining** | | |
|---|---|---|---|
| **Features** | **Web Content Mining** | **Web Structured Mining** | **Web Usage Mining** |
| **View of Data** | Structured, Unstructured | Link Structure | Interactivity |
| **Main Data** | Text document, Hypertext document | Link Structure | Server and Browser Logs |
| **Representation** | Collection of words, phrases, Contents and relations | Graph | Relational Table Graph |
| **Method** | Machine Learning Statistical(including NLP) | Proprietary algorithms | Machine Learning Statistical and Association rules |

## 1.3.Literature Review

Search results today are based on web popularity, as a consequence of which a user does not get the right results. Results that show up on the first page have fewer chances of holding user interest. Consequently, in order to provide users the needed results first, a new concept of page ranking was developed to re-rank the results of users' interests. Agichetein et al. [1] proposed a new concept for ranking by incorporating user behavior information. They used the Lucene search API for processing hits made by users, and keywords are mapped into the database thereby calculating the popularity score, ensuring that peak results come first. The drawback with this method is that while the results do not change for every user, the need of each user is not the same. Therefore, some users find good results while others do not. Ahmadi- Abkenari et al [2] proposed a method in which log ranks are used to calculate website significance values. The drawback here is that the log data may be old, but since the results clicked change over time they are unlikely to be good for current users. Anivban kundu[3] in his work two database is used ie. Global, local and ranking procedures, implemented in a hierarchical fashion. Here, inbound and outbound web page sessions for matched URLs, and the number of HITS for matched URLs, are calculated. A major problem, however, is that it is time-consuming. Anuradha et.al [4] proposed ANT Ranking, which is the ANT colony algorithm for ranking web pages. Here, too, the value of page ranking is calculated using clicks on the web. User interest on a web page changes dynamically if the user has visited the page. But, with this method, only those pages a user is interested in are considered for the purpose of page ranking. The base of page ranking was proposed by Brin and Page [5], where each web keyword search is used for page ranking and the process extended by counting links from all pages. The problem is that they do not develop or model a t for tracing user behaviour. Gomez-Nieto et. al [7], in their similarity processing, snippet-based visualization of web search results, considered only web snippets to process page ranking Liao et al [14] used user search behaviour to rank pages. They insist that a task trail performs better than sessions and a query trail in determining user satisfaction. Also, it increases users' web page utility compared to the query trail of other sessions. Shalya Nidhi et al. [17] state that an effective content-based web page ranking is possible when both web structure and content mining are mixed in order to get the relevant web page. Here, web structure mining is used to get the link structure and web content is used to get the content of the web page so that the relevant page can be fetched as per a user's query. In our proposed work too, we intend to use both web structure and content mining.

## 1.4.Motivation and Justification

The problem with ranking is the need to display a results list, based on user requests or preferences. In a traditional approach, a search engine always returns the same rank for the same query submitted at different times or by different users. As massive volumes of data are available, the principal problem is the need to scan page after page to find the desired information a user needs. In current search engines, the difficulty is with ordering the search results and then presenting the most relevant page to the user first. Motivated by this fact, a page ranking algorithm is proposed in this paper by considering both content and user click-based results. It is expected that users get effective and relevant results for a given query at a faster rate with page ranking, Justified by these facts, page ranking is done, taking into consideration each individual's keywords, URLS and the keyword-to-URLs. In this way, the user can browse one group of web pages or another and search for needed results faster.

## 1.5.Organization of the paper

The rest of the paper is organized as follows. Section 2 describes different page ranking algorithms. Section 3 discusses the proposed technique. In Section 4, quantitative analysis of page ranking algorithms are discussed. In Section 5, the results are analyzed and discussed. Section 6 focuses on the conclusion of the research.

# 2.PAGE RANKING ALGORITHMS

The three major algorithms for ranking pages i.e. page ranking, weighted page ranking and HITS (Hyperlink-Induced Topic Search) are presented below.

## 2.1.Page Ranking

Page ranking, developed by Brin and Page[5] at Stanford University, is used by Google to calculate the relative importance of web pages. In the beginning, a citation analysis was used. But the problem was that incoming links that were treated as citations could not provide good results so Page et al. proposed a new technique in which the page rank value of a web page is computed by simply counting the number of pages linked to it. These links are called back links.

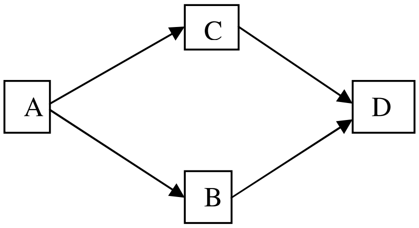

Figure 1. An example of Back links

In Figure 1, page A is a backlink of page B and page C, while page B and page C are backlinks of page D. If a backlink comes from an important page, then that link is given higher weightage. The link from one page to another is considered a vote. The formula they proposed for calculating the page rank is

$$PR(A)=(1-d)+d(PR(T_1)/C(T_1)+\ldots+PR(T_n)/C(T_n)) \qquad 1$$

where PR(Ti) is the page rank of the page Ti which links to page A. C(Ti) is the number of links put on page Ti. D is the damping factor which is usually set to 0.5, and is used to avoid other pages having too much influence.

## 2.2.Weighted Page Rank

This algorithm, proposed by Wenpu Xing and Ali Ghorbani, is an extension of the page ranking algorithm [7]. It assigns larger rank values to more popular pages instead of dividing the rank value of a page evenly among its outlink pages. Each outlink page gets a value proportional to its popularity, with popularity from the number of inlinks and outlinks recorded as $W^{in}(v, u)$ and $W^{out}(v, u)$ respectively. $W^{in}(v, u)$ in the equation below is the weight of link (v, u), used to make calculations based on the number of inlinks of page u and the number of inlinks of all the reference pages of page v.

$$W^{in}(v, u)=\frac{I_u}{\sum P\in R(v) I_p} \qquad 2$$

where $I_u$ and $I_p$ represent the number of links of page u and page p. R(v) denotes the reference page list of page v, and $W^{out}(v, u)$ in the equation is the weight of link (v, u), calculated based on the number of outlinks of page u and the number of outlinks of all the reference pages of page v

$$W^{out}(v, u)=\frac{O_u}{\sum P\in R(v) O_p} \qquad 3$$

where $O_u$ and $O_p$ represent the number of outlinks of page u and page p. Based on the weight of inlinks and outlinks, the page ranking formula is modified as

$$PR(u)=(1-d)+d\sum_{v\in B_{(u)}} PR(v)W^{in}(v,u)W^{out}(v,u) \qquad 4$$

## 2.3.HITS (Hyperlink-Induced Topic Search)

HITS ranks web pages by means of analysis inlinks and outlinks, and was proposed by Klien Berg[11]. In this algorithm, web pages pointed to by many hyperlinks are called authorities and the points to many hyperlinks are called hubs. A web page may be a good hub and a good authority at the same time. Here the WWW is treated as a directed graph G (V,E), where V is a set of vertices representing pages and E is a set of edges corresponding to links. Figure 2. shows the hubs and authorities in web. The two methods involved here are sampling and iterative..

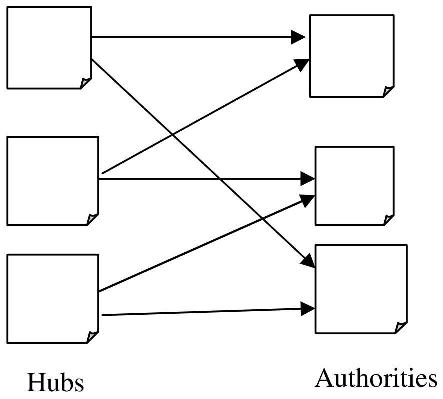

Figure 2. Hubs and Authorities

In the sampling method, a set of relevant pages for a query are collected. In the iterative method, hubs and authorities are found using the output of sampling. For calculating the weight of hubs ($H_p$) and the weight of authorities ($A_p$)

$$H_p=\sum_{q\in I(p)} A_q \qquad 5$$

$$A_p=\sum_{q\in B(p)} H_q \qquad 6$$

Here $H_q$ is the hub score of a page, $A_q$ is the authority score of a page, I(p) is the set of reference pages of page p, and B(p) is the set of reference pages of page p. The authority weight of a page is proportional to the sum of authority weight of the pages that it links to. Problems with HITS include the fact that hubs and authorities are not easily distinguished and fail to produce relevant results to user queries because of their equivalent weights. Finally, it is not efficient in real time and is therefore not implemented in a real-time search engine.

# 3.PROPOSED PAGE RANKING ALGORITHM

Figure 3. represents the overall working of the proposed page ranking algorithm using a content and user click-based method. To evaluate a page rank for the web, a user click and content-based page rank algorithm is proposed. Content analysis is evaluated based on the available content on a webpage. The entire page rank model can be simulated using three key steps: first, a user query interface by which a user sends a request for a query. The extracted results are pre-processed using Porter's stemming algorithm, and user-clicked contents are traced using user clicks through a collector. The contents of user-needed data are extracted, analyzed and similarities measured using the cosine formula. Then the listed results are again rearranged using a ranking based on content and user clicks and, finally, the listed results are re-ranked and their performance evaluated using precision, recall, fallout and f-measure.

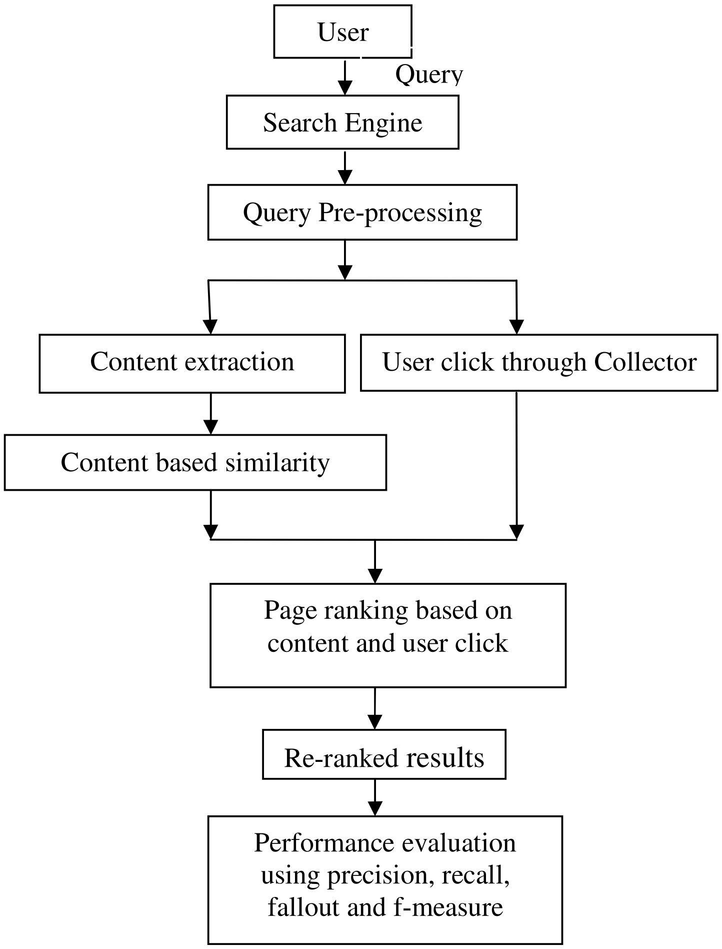
User
Query
Search Engine
Query Pre-processing
Content extraction
User click through Collector
Content based similarity
Page ranking based on content and user click
Re-ranked results
Performance evaluation using precision, recall, fallout and f-measure

Figure 3. Block Diagram for the Proposed Page Ranking Algorithm

## 3.1.Query interface

This can be a graphical user-interface design for providing a system inputs. Inputs can be search terms required to search for data extraction from data sources.

## 3.2.Pre-processing

Data in the real world is dirty and incomplete, lacking both attribute values as well as certain attributes of interest, or containing only noisy and inconsistent aggregate data. There is no quality in data-extracted results and no quality mining. Further, the queries asked by users are particularly short. After retrieving the results for a user's query, the snippets are mixed with unwanted content and are pre-processed using information-retrieval techniques. The aim is to generate highly relevant results for a given search query, which can be achieved by stemming and stop word removal.

### 3.2.1.Stemming

Stemming is a term used in the information-retrieval process for reducing inflected words to their word stem or base. Stemming algorithms are used to transform words in a text and improve the information-retrieval system. The goal is to obtain a one-word description of similar - but not identical - words. The word obtained in the end has neither meaning nor is grammatically correct, but it contains a description of and bears a similarity to all the other words it represents. For example: “implementing” and “implemented” are described by the word “implement.”

### 3.2.2.Stop Word Elimination

After stemming, it is necessary to remove unwanted words. Stop words are words which are filtered before or after processing data. A stop word is a word that does not have a meaning, so eliminating stop words offers better results in a phrase search. In all languages, certain words are considered stop words, of which there are more than 500 types. For example, words such as “on,” “and,” “the” and “in,” among others, do not provide useful information. After pre-processing these snippets, the results are considered for further processing.

## 3.3.Web Page Content Retrieval

In this phase after the removal of unwanted data, the contents that are frequently occurred in the snippets are extracted and the relationship between the extracted words is analyzed here. Also the user clicked contents are collected in the user click through collector.

### 3.3.1.Extracting Content from Web Snippets

Content from the refined results is extracted by finding frequent item sets in data mining. When a user types a query, a set of relevant web snippets are returned and if a keyword or phrase exists frequently in web snippets relating to a particular query, it represents important content related to the query because it exists with the query in the top documents. To measure interest in a particular keyword or phrase $k_i$ extracted from web snippets:

$$\mathrm{support}(\mathrm{k_i}) = \frac{\mathrm{sf(k_i)}}{\mathrm{n}}.\, |\mathrm{k_i}| \qquad 7$$

Table 3 Frequent Words Extracted for The Query “Web Mining”

| Word | Support |
|---|---|
| Data | 0.5 |
| Techniques | 0.108696 |
| Patterns | 0.130435 |
| Analysis | 0.108696 |
| Information | 0.173913 |
| Knowledge | 0.108696 |
| Usage | 0.217391 |

where $sf(k_i)$ is the snippet frequency of the keyword or phrase $k_i$ , $n$ is the number of web snippets returned, and $|ki|$ is the number of terms in the keyword or phrase $ki$ If the support of a keyword or phrase $k_i$ is greater than the threshold $s$, then $k_i$ acts as a concept for the query $q$. Table 3 shows a sample of frequent words extracted for the query “web mining. The maximum length of a concept is limited. This process not only reduces processing time but also avoids extraction of meaningless content.

### 3.3.2.Content-Based Similarity

In extracting content-based similarity, a signal-to-noise formula is used to establish the similarity between keywords $k_1$ and $k_2$. The two keywords from a query $q$ are similar if they coexist frequently in web snippets arising from the query $q$.

$$\mathrm{sim}(\mathrm{k}_1, \mathrm{k}_2) = \frac{\mathrm{n}*\mathrm{df}(\mathrm{k}_1\mathrm{U}\mathrm{k}_2)}{\mathrm{df}(\mathrm{k}_1)*\mathrm{df}(\mathrm{k}_2)} / \log \mathrm{n} \qquad 8$$

where $n$ is the number of documents in mass, $df(k)$ is the document frequency of the keyword $k$, and $df(k_1 U k_2)$ is the joint document frequency of $\mathrm{k}_1$ and $\mathrm{k}_2$. The similarity $sim(k_1,k_2)$ obtained using the above formula always lies between [0, 1]. In search engine contexts, two concepts $k_i$ and $k_j$ can coexist in web snippets

$$sim_{R,snippet}(k_i, k_j) = log\,\frac{n*sf_{snippet}(k_i \cup k_j)}{sf_{snippet}(k_i)*sf_{snippet}(k_j)} /\ \log n \qquad 9$$

Table 4 Relationship between the Words Extracted For The Query “Web Mining”

| ID | Concept1 | Concept2 | Relations |
|---|---|---|---|
| 1 | Data | Techniques | 0.108695652174 |
| 2 | Data | Patterns | 0.0652173913043 |
| 3 | Data | Analysis | 0.0869565217391 |
| 4 | Data | Information | 0.0869565217391 |
| 5 | Data | Knowledge | 0.108695652174 |
| 6 | Data | Usage | 0.0652173913043 |
| 7 | Techniques | Data | 0.108695652174 |
| 8 | Techniques | Patterns | 0.0434782608696 |
| 9 | Techniques | Analysis | 0.0217391304348 |
| 10 | Techniques | Information | 0.0217391304348 |
| 11 | Techniques | Knowledge | 0.0217391304348 |

| | | | |
|---|---|---|---|
| 12 | Techniques | Usage | 0 |
| 13 | Patterns | Data | 0.0652173913043 |
| 14 | Patterns | Techniques | 0.0434782608696 |
| 15 | Patterns | Analysis | 0.0217391304348 |
| 16 | Patterns | Information | 0 |
| 17 | Patterns | Knowledge | 0 |
| 18 | Patterns | Usage | 0.0652173913043 |
| 19 | Analysis | Data | 0.0869565217391 |
| 20 | Analysis | Techniques | 0.0217391304348 |
| 21 | Analysis | Patterns | 0.0217391304348 |
| 22 | Analysis | Information | 0.0217391304348 |
| 23 | Analysis | Knowledge | 0.0652173913043 |
| 24 | Analysis | Usage | 0.0434782608696 |

where $sf_{snippet}(k_i \cup k_j)/ sf_{snippet}(k_i \cup k_j)$ are joint snippet frequencies of the concept $k_i$ and $k_j$ in web snippets. $sf_{snippet}(k_i). sf_{snippet}(k_j)$ is the snippet frequency of the concepts $k_i$ and $k_j$ respectively for finding essential features from data word frequency using the following formula

$$word\ frequency = \frac{no.of\ times\ word\ in\ a\ document}{total\ words\ in\ document} \tag{10}$$

Now, using Euclidean distance, the similarity between the extracted data and the available set of data is computed and, according to the distance obtained, the data is placed in a similar set where it actually belongs. Table 4 is the relationships between the words extracted for the query “Web Mining”. In this, the relationships made between the frequently extracted concepts are evaluated.

### 3.3.3.User Click-through Collectors

The relationship that exists between concepts is processed by considering a user’s click-through. User-clicked queries are called user-positive preferences and others are user-negative preferences. When a user clicks on a query, the weight of the extracted concept is incremented by 1 to show user interest. Other concepts related to the user’s query are also incremented to a similar score. If the concept is closely related to the user’s positive result, then it is incremented to a higher value. Otherwise, it is incremented to a small fraction close to zero, by means of which a user log is created. After finding the data needed using the search engine, it is ranked according to its relevance to the user’s query, requiring that a new kind of system be implemented a consequence. Thus the proposed work is designed in the manner set forth below. The given ranking system is implemented in components of different text processing and weight estimation techniques.

## 3.4.Page Ranking Based on Content and User Click

User-interested results for a query are stored in a database and, over time, collected using user clicks through collector. This is termed a query log and provides useful information about searchers' queries and what users are interested in. A problem peculiar to a query log is that it has no relational information other than a query and a click. Considerable portions of queries are rare, with few clicks or even no clicks at all for certain queries [20]. In the proposed work, a combined user profile method is applied to Google search results and the retrieved results are re-ranked, based on user-interested results with level re-ranking being used for this particular group. Both web structure and web content mining are used to get users the relevant results anticipated. Web content mining is used here to get the linking structure of a web page and trace the content and similarity between each item of the contents of the said web page. Initially, a user visits a web page at random but this change over time. Here, user interest is calculated using clicks on web links. The quality of user interest changes dynamically with the number of user visits to the web page. The probability $P_i$ of a user-taken decision equals user interest relative to the sum of all user-interested values. To find the probability of a user choosing a web page 'i' is

$$P_i = \frac{U_i}{\sum_{i=S} U_i} \qquad 11$$

Here $u_i$ is the user interest and S is the similarity of contents in the web page.

As user interest changes, accordingly user interest value $u_i$ is updated, along with the time spent by the user when the page was visited for the query. As a result, the relevance of the page to the user increases greatly and the probability of its being chosen also increases correspondingly. When the user visits the page, the quantum of user interest is updated. The volume of the web page increases, proportional to its quality.

$$U_i^{t+1} = U_i^t + \Delta U_i^t \qquad 12$$

where $u_i$ user interested value at time in sec and and $\Delta U_i^t$ the amount of user interest saved at time t left by the user. It can be changed, depending on user interest in terms of clicks.

# 4.QUANTITATIVE ANALYSIS OF PAGE RANKINGALGORITHMS

Different page ranking algorithms are discussed and quantitative differences between each represented in Table 5. The mining technique used, input parameters, time complexity and limitations are discussed.

Table 5 Quantitative Analysis Of Different Pager Ranking Algorithms

| Quantitative Parameters | Different Pager Ranking Algorithms | | | |
|---|---|---|---|---|
| | Page Rank | Weighted Page Rank | HITS | Proposed Page Ranking |
| Mining technique used | WSM | WSM | WSM and WCM | WSM and WCM |
| I/P Parameters | Back links | Front and Back links | Front, Back links and content | Front, Back link, content and user click |
| Complexity | O(log N) | <O(log N) | <O(log N) | ≤O(log N) |
| Limitation | Query independent | Query Independent | Topic drift and efficiency problem | More computation time |
| Working | Results are stored according to importance of pages | Results are stored according to importance of pages | Compute hub and authority scores of highly relevant pages | Consider the user clicked links(VOL), content of the snippets and user relevant pages are stored |

From the table, it is clear that a lot of new ideas have been implemented to provide good results in content and user click-based page ranking, which are proved by providing experimental results like the relevancy rule, precision, recall, fall out and f-measure

# 5.EXPERIMENTAL RESULTS AND DISCUSSION

## 5.1. Performance Metrics

The proposed method is compared with existing methods using performance measures like precision, recall, fallout and f-measure. They are computed as below:

### 5.1.1. Precision

Precision is the fraction of documents retrieved that are relevant to a user's information needs. It takes all retrieved documents into account.

$$\text{Precision}=\frac{relevant\ document \cap retrieved\ documents}{retrieved\ documents} \qquad 13$$

### 5.1.2.Recall

Recall is the fraction of documents successfully retrieved and relevant to a query. Also called sensitivity, it can be looked at as the probability that a relevant document is retrieved by the query.

$$\text{Recall}= \frac{relevant\ document \cap retrieved\ documents}{relevant\ documents} \qquad 14$$

It is easy to achieve a recall of 100 percent by retrieving all documents in response to a query. Hence recall alone is not enough, and the number of non-relevant documents is required to be measured as well.

### 5.1.3.Fallout

Fallout is the proportion of non-relevant documents retrieved from all the non-relevant documents available. It can be looked at as the probability that a non-relevant document is retrieved by a query.

$$\text{Fallout}=\frac{non\ relevant\ documents \cap relevant\ documents}{non\ relevant\ documents} \qquad 15$$

It is easy to achieve fallout of 0 percent by returning zero documents in response to a query.

### 5.1.4. F-measure

F-measure is the harmonic mean of precision and recall, and provides good results when precision and recall provide good results.

$$\text{F-measure}= 2.\frac{precision*recall}{precision+recall} \qquad 16$$

### 5.1.5.Relevancy Rule

The relevancy of a page for a given query depends on the category and position the needed data holds in a page list. The larger the relevancy values, the better the results. Relevancy is calculated using the formula

$$K=\sum_{i\in R_{(p)}}(n-i)*W_i \qquad 17$$

where i denotes the i page in the result page-list R(p), n represents the first n pages chosen from the list R(p), and $W_i$ is the weight of page I with four categories of very relevant pages (VRP), relevant pages (RP), weakly relevant pages (WRP) and irrelevant pages (IR).

## 5.2.Dataset

This section provides a performance analysis of the proposed technique, and a comparative study is also provided with the traditional method of page rank estimation. The Bing API is used for preparing a dataset of user queries, with Bing search results for 30 days in June 2015 being considered. Default snippet counts are set to 100. User-clicked contents, as well as users' positive and negative preferences are collected and re-ranked based on content and user-clicked data, along with bookmarked contents from user log files. For evaluation some of the queries used may be ambiguous, entity names and general terms and are shown below. Table 7 shows the queries used for evaluation of search results.

Table 7 Queries Used For Evaluation

| Types | Queries |
|---|---|
| Ambiguous | Apple ,Tiger, Penguin, Jaguar |
| Entity names | Dell Disney, Raja |
| General terms | Sun , Music, Network |

## 5.3.Experimental Results

Experiments are conducted using different queries to check the performance of the retrieved results based on the metrics like precision, recall, fallout, f-measure and relevancy rule and are shown from Table 8 to Table 11 respectively. Precision values vary in accordance with changes in user interest. The given precision values define the relevancy of search results obtained during experimentation. Search recall values, which are measurements of accuracy, are measured in this section. Fallout values for the queries are evaluated and the values which are less are optimal because here the error rate is considered. F-measure is calculated by considering the precision and recall values estimated and the results are listed below.

Table 8 Precision Values for Different Page Ranking Algorithms

| Queries | Precision Values For Different Page Ranking Algorithms | | | |
|---|---|---|---|---|
| | PR | WPR | HITS | Proposed Page Ranking |
| Apple | 0.862 | 0.904 | 0.961 | 0.991 |
| Data mining | 0.886 | 0.870 | 0.952 | 0.926 |
| PHP | 0.799 | 0.893 | 0.904 | 0.919 |
| Web | 0.589 | 0.791 | 0.842 | 0.993 |

| | | | | |
|---|---|---|---|---|
| Jaguar | 0.647 | 0.751 | 0.835 | 0.893 |
| Google | 0.798 | 0.812 | 0.847 | 0.885 |
| Network | 0.719 | 0.729 | 0.787 | 0.945 |
| Tiger | 0.731 | 0.771 | 0.753 | 0.932 |

Table 9 Recall Values for Different Page Ranking Algorithms

| **Queries** | **Recall Values For Different Page Ranking Algorithms** | | | |
|---|---|---|---|---|
| | **PR** | **WPR** | **HITS** | **Proposed Page Ranking** |
| Apple | 0.977 | 0.979 | 0.970 | 0.989 |
| Data mining | 0.993 | 0.937 | 0.950 | 0.96 |
| PHP | 0.968 | 0.943 | 0.971 | 0.979 |
| Web | 0.974 | 0.987 | 0.958 | 0.98 |
| Jaguar | 0.978 | 0.876 | 0.984 | 0.99 |
| Google | 0.941 | 0.969 | 0.940 | 1.0 |
| Network | 0.969 | 0.947 | 0.947 | 1.0 |
| Tiger | 0.945 | 0.940 | 0.945 | 0.952 |

Table 10 Fallout Values for Different Page Ranking Algorithms

| **Queries** | **Fallout Values for Different Page Ranking Algorithms** | | | |
|---|---|---|---|---|
| | **PR** | **WPR** | **HITS** | **Proposed Page Ranking** |
| Apple | 0.023 | 0.021 | 0.030 | 0.011 |
| Data mining | 0.007 | 0.063 | 0.050 | 0.04 |
| PHP | 0..032 | 0.057 | 0.029 | 0.021 |
| Web | 0.026 | 0.013 | 0.042 | 0.02 |
| Jaguar | 0.022 | 0.124 | 0.016 | 0.01 |
| Google | 0.059 | 0.031 | 0.060 | 0.00 |
| Network | 0.031 | 0.053 | 0.053 | 0.00 |
| Tiger | 0.055 | 0.060 | 0.055 | 0.048 |

Table 11 F-Measure Values for Different Page Ranking Algorithms

| **Queries** | **F-Measure Values for Different Page Ranking Algorithms** | | | |
|---|---|---|---|---|
| | **PR** | **WPR** | **HITS** | **Proposed Page Ranking** |
| Apple | 0.938 | 0.917 | 0.962 | 0.99 |
| Data mining | 0.926 | 0.910 | 0.951 | 0.943 |
| PHP | 0.869 | 0.864 | 0.895 | 0.947 |
| Web | 0.847 | 0.737 | 0.895 | 0.986 |
| Jaguar | 0.849 | 0.743 | 0.902 | 0.938 |
| Google | 0.872 | 0.875 | 0.890 | 0.938 |
| Network | 0.831 | 0.816 | 0.859 | 0.991 |
| Tiger | 0.848 | 0.822 | 0.837 | 0.941 |

According to the results obtained in table 8, the performance of the proposed technique is optimal, when compared to other, traditional page ranking approaches. It is easy to achieve a recall of 100 percent by retrieving all documents in response to a query. Hence recall alone is not enough, and the number of non-relevant documents is required to be measured as well. Hence fallout is considered and from the table 10, that the proposed page ranking has fewer non - relevant documents retrieved than other page ranking algorithms. Table 11 makes it clear that the proposed method provides good results compared to other existing methods because precision and recall for the existing methods are fewer, compared to the proposed method. Hence the f-measure value for content and user click-based page ranking gives good results.

Relevant pages retrieved for different queries are calculated by varying the number of snippets by 50,100 and 150 and relevancy values measured. Values are calculated for different page ranking algorithms like proposed page ranking, weighted page ranking (WPR), page ranking (PR) and Hyperlink Induced Topic Search (HITS).

-

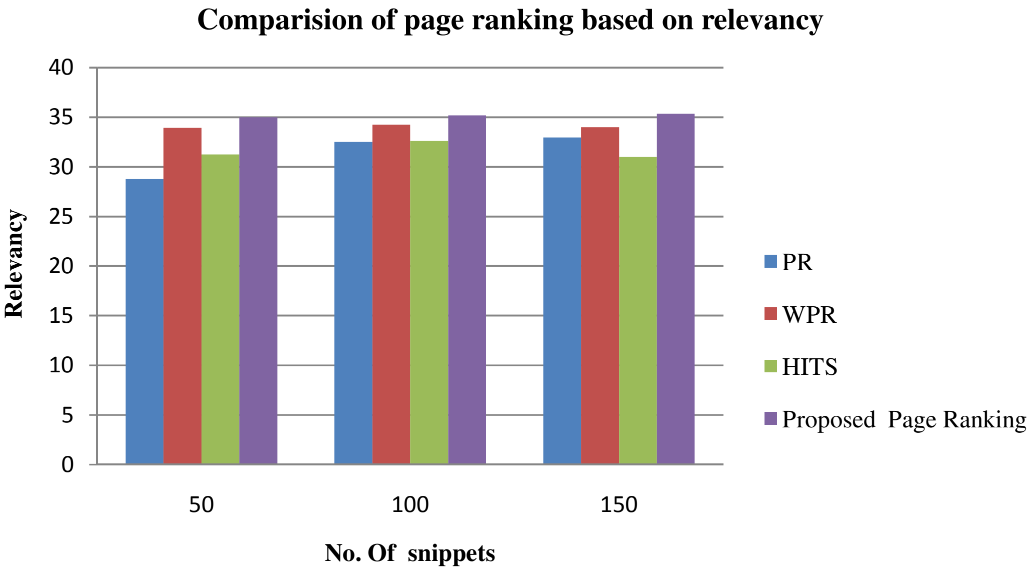

Figure 4. Comparison of Different Page Ranking Algorithms Based On Relevancy

From Figure 4 it is inferred that the relevancy rate of proposed page ranking is comparatively higher than page ranking algorithms like PR, WPR, and HITS across various numbers of snippets. It can also be inferred that the relevance rate for WPR is comparatively better than that of other page ranking algorithms.

## 6.CONCLUSIONS

Page ranking is essential to ascertain the importance of a web page for a given search query and rank the results according to their relevance. In this study, different page ranking techniques are investigated and a new kind of page rank algorithm proposed and designed. This page rank algorithm provides a rank according to the relevance of a user's query and the contents available in web pages. In addition, the relevancy of search results is measured in terms of precision, recall and F-measure. These results demonstrate the efficacy of relevant ranks for the search results available. The proposed work is intended to provide an efficient page rank technique using an analysis of web page content. The page rank technique presented ranks results according to the importance of a web page, user search query and the content available in the web page. The proposed technique is efficient but not generalized, providing efficient, scenario-specific page ranking. Therefore, in the near future, the proposed technique is to be extended to derive a generalized framework for page rank estimation.

# REFERENCE

1. Agichtein, Eugene, Eric Brill, and Susan Dumais. "Improving web search ranking by incorporating user behavior information." In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 19-26. ACM, 2006.
2. Ahmadi-Abkenari, Fatemeh, and Ali Selamat. "Application of the Clickstream-Based Web Page Importance Metric in Web Site Ranking and Rank Initialization of Infant Web Pages." International Journal of Advancements in Computing Technology 4, no. 1 (2012).
3. Anirban Kundu, RanaDattagupta and Sutirtha Kumar Guha,"Hierarchical Ranking in Search Engine Environment: An Overview", Jorunal of Convergence Information Technology(JCIT), Vol9, no.5,2014
4. Anuradha, G., G. Lavanya Devi, and MS Prasad Babu. "ANTRANK: An Ant Colony Algorithm For Ranking Web Pages." International Journal of Emerging Trends &Technology in Computer Science (IJETTCS) 3, no. 2 (2014): 208-211.
5. Brin, Sergey, and Lawrence Page. "Reprint of: The anatomy of a large-scale hypertextual web search engine." Computer networks 56, no. 18 (2012): 3825-3833.
6. Evans, Michael P. "Analysing Google rankings through search engine optimization data." Internet research 17, no. 1 (2007): 21-37
7. Gomez-Nieto, Erick, Frizzi San Roman, Paulo Pagliosa, Wallace Casaca, Elias S. Helou, Maria Cristina F. de Oliveira, and Luis Gustavo Nonato. "Similarity Preserving Snippet-Based Visualization of Web Search Results." Visualization and Computer Graphics, IEEE Transactions on 20, no. 3 (2014): 457-470.
8. Greenstein, Shane. "A Big Payoff." IEEE Micro 2, no. 32 (2012): 64.
9. Ishii, Hideaki, Roberto Tempo, and Er-Wei Bai. "A web aggregation approach for distributed randomized PageRank algorithms." Automatic Control, IEEE Transactions on 57, no. 11 (2012): 2703-2717.
10. Ishii, Hideaki and Roberto Tempo, "The Page Rank Problem, Multi-agent Consensus, and Web Aggregation", IEEE Control System Magazine, June 2014.
11. Jain, Rekha, and Dr GN Purohit. "Page ranking algorithms for web mining." International journal of computer applications 13, no. 5 (2011): 0975-8887.
12. Killoran, John B. "How to use search engine optimization techniques to increase website visibility." Professional Communication, IEEE Transactions on 56, no. 1 (2013): 50-66.
13. Lamberti, Fabrizio, Andrea Sanna, and Claudio Demartini. "A relation-based page rank algorithm for semantic web search engines." Knowledge and Data Engineering, IEEE Transactions on 21, no. 1 (2009): 123-136.
14. Liao, Zhen, Yang Song, Yalou Huang, Li-wei He, and Qi He. "Task Trail: An Effective Segmentation of User Search Behavior." Knowledge and Data Engineering, IEEE Transactions on 26, no. 12 (2014): 3090-3102.
15. Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. "The PageRank citation ranking: bringing order to the Web." (1999).
16. Papagelis, Athanasios, and Christos Zaroliagis. "A collaborative decentralized approach to web search." Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 42, no. 5 (2012): 1271-1290.
17. Shalya, Nidhi, Shashwat Shukla, and Deepak Arora. "An Effective Content Based Web Page Ranking Approach." International Journal of Engineering Science and Technology (IJEST) 4, no. 08 (2012).
18. Tyagi, Neelam, and Simple Sharma. "Weighted Page rank algorithm based on number of visits of Links of web page." International Journal of Soft Computing and Engineering (IJSCE) ISSN (2012): 2231-2307.
19. Xing, Wenpu, and Ali Ghorbani. "Weighted page rank algorithm." In Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on, pp. 305-314. IEEE, 2004.

## Authors

Dhiliphan Rajkumarhas a strong passion in Web Mining, Pattern recognition and Social networking. He is currently pursuing Ph.D degree in Computer Engineering, ManonmaniamSundaranar University, India.

**Dr. A. Suruliandi** received his B.E. (1987) in Electronics and Communication Engineering from Coimbatore Institute of Technology, Coimbatore, Bharathiyar University, Tamilnadu, India. He received M.E. (2000) in Computer Science and Engineering from Government College of Engineering Tirunelveli. He also received Ph.D. in Computer Science (2009) from Manonmaniam Sundaranar University as well.He is having more than 27 years of teaching experience. He is having more than 80 publications in International journals and conferences. His research interests include Pattern recognition, Image processing, Remote sensing and Texture analysis.

Selvaperumal has a strong passion in Web Mining, Data Mining, Machine learning, NLP and Art ificial Intelligence. He is currently pursuing Ph.D degree in Computer Engineering, ManonmaniamSundaranar University, India.