A New Hybrid Method for Web Pages Ranking in Search Engines

***Sajjad Najafi*[1], *Farhad Soleimanian Gharehchopogh*[2]**

1- Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran.

2- Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran.(bonab.farhad@gmail.com)

Received (2019-06-15) Accepted (2019-08-01)

Abstract: *There are many algorithms for optimizing the search engine results, ranking takes place according to one or more parameters such as; Backward Links, Forward Links, Content, click through rate and etc. The quality and performance of these algorithms depend on the listed parameters. The ranking is one of the most important components of the search engine that represents the degree of the vitality of a web page. It also examines the relevance of search results with the user's query. In this paper, we try to optimize the search engine results ranking by using the hybrid of the structure-based algorithms (Distance Rank algorithm) and user feedback-based algorithms (Time Rank algorithm). The proposed method acts on multiple parameters and with more parameters it tries to get better results while keeping the complexity and running time of the algorithms. Average distance and average attention time have been evaluated on web pages and by using the obtained data, proposed method performance has been evaluated. We compare proposed method with several famous algorithms such as Time Rank, Page Rank, R Rank, WPR and sNorm(p) in this field by applying Precision@N (P@N), Average Precision (AP), Mean Reciprocal Rank (MRR), Mean Average Precision (MAP), Discounted Cumulative Gain (DCG) and Normalized Discounted Cumulative Gain (NDCG) criteria. The results indicate better performance in comparison with existing algorithms.*

Keywords: *Ranking, Search Engine, Click through Rate, Distance Rank, User Attention Time.*

How to cite this article:

Sajjad Najafi, Farhad Soleimanian Gharehchopogh. A New Hybrid Method for Web Pages Ranking in Search Engines. J. ADV COMP ENG TECHNOL, 5(4) Autumn 2019 : 233-244

I. INTRODUCTION

In the last several years, web application has been unimaginable. The amount of content stored and shared on the web is increasing quickly and continuously. So, problems and difficulties such as finding and properly managing all the existing amount of information arise [1]. Understanding and analysing the structure of Web data is important for data recovery [2]-[3]. With the increasing amount of data available on the World Wide Web, web mining became one of the most valuable sources of information to recover exploration of knowledge [4]. Web mining uses data mining techniques for discovering and retrieving information from web documents. Web mining is used to deal with complex and varied data on the web in the form of structured, semi-structured, and non-structured even [5]. The use of web mining is ranking web pages based on content, structure, and their usage. Many ranking algorithms have been proposed in recent years that each of them is used to evaluate the web page ranking. Web mining can be classified in three categories of Web content mining, web structure mining and web usage mining [6]-[7].

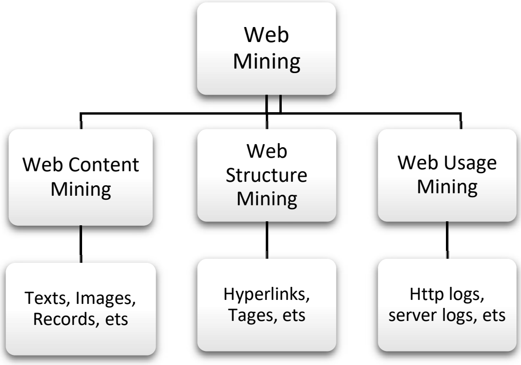

Fig. 1. Web mining classification.

Web graph structure composes of web pages as nodes and hyperlinks as connected edges between pages, and the content of web pages can be organized in a tree structure. The goal of Web mining is trying to produce a summary of web pages. Web structure mining categorizes web pages and produces information on the similarities and relationships between different pages. Web usage mining is the process of extracting useful information from data which is derived from user behavior on the web. This type of web mining extracts saved data in report access servers, the user profiles and additional data. The process of extracting useful information from web documents and content are called web content mining. Web documents can include text, images and sounds, movies or structured documents such as tables and lists. Web content mining is associated with data mining because most data mining techniques can be used in web content mining. Also Web content mining is associated with text mining, because most web content is text [3]-[6].

The extracted data from the web structure mining, web usage mining and as well as web content mining are three types of data which were extracted from the website. Each of the available ranking algorithms, use any of these data and ranks Web pages. Then the ranking algorithms can be categorized into three content-based algorithms, structure based algorithms and Web usage-based algorithms. Content-based algorithms by examining the content of web pages such as text, images and documents ranks pages. Among content-based algorithms it can be noted of Salton vector space model [8]. Structure-based algorithm does action ranking by evaluating the information on the web Structure such as tags and links, among these algorithms we can refer to page rank [9] and hits [10]. User-oriented ranking algorithms analysis Web Pages using data obtained from the analysis of users' behavior that is extracted from log proxy servers. In fact, these methods by realizing the users' interests do ranking according to them. Some of these methods include CVM, NM and IA [11] and the document development methodology [12]. Considering that each of the ranking algorithms is using a certain method, each has its own weaknesses and strengths. Of course, the weakness of ranking algorithms is followed by an error in the ranking of web pages and discontent users [13], [14] and [15]. In our proposed method by hybridization the two algorithms with different methods it tried to cover each algorithm weaknesses and raise the ranking carefulness.

We've prepared structure of the paper as follows. In section 2 of this paper, we will review several famous ranking algorithms. In section 3, we will continue to provide the proposed method. In section 4 we will discuss the results obtained by the proposed method. In section 5, we will compare the proposed method with the algorithms discussed in section 2 and the last section presents the conclusions and future work of the paper.

II. RELATED WORKS

Salton vector space model [13] is the most important text-based approach that is a standard technique for ranking documents corresponding to the query. In this model, document and the user query is a vector with aspect of the number of words. Each vector with a specific formula changes into a weighted vector and then cosine of the angle between two weighted vectors as the degree of similarity is calculated. After calculating the similarity of documents with user query, related documents are arranged in descending order. Statistical properties are used for weighting the keywords within the Web pages. Mr Salton uses the words repetition for calculating the weight. In this way that If a keyword repeats serval times in a document, that word describes it well.

Page Rank algorithm is query-independent algorithm and was presented for the first time in 1998 by larry page and Sergey Brin [9] and it is used in Google's search engine. The algorithm works on the links between web pages. So that if page A refers to page B, it means that the content of page B for page A is valuable. And the link of A is valuable for B, if page A is the important page, in fact in this algorithm for the calculation rank of a page, the number and value of the incoming links are important. Links from one page to another are like a vote; not only the number of votes are important but the value of pages that votes are important,too. The algorithm calculates the ranking at the time of Indexing. For this reason, the results relationship is less.

HITS ranking algorithm is a query related algorithm which was proposed in 1998 by Kleinberg [10]. In this algorithm, Web pages are defined in two forms of Hub an Authorities. This means that a page has content relevant is authority and a hub page has some links to authority pages. In other words, a page with a high authority is mentioned with the number of pages with high hub also a page with a high hub refers to the many pages with high authority. This method also has problems because it is not easy to distinguish between hub and authority because in HITS one page can be a good hub and at the same time can be a good authority. As well as some sites because of raising ranking point out regularly to each other which is called rank spamming.

Weighted page rank algorithm [16] is the modified form of page rank technique and is used to generalize the page rank algorithm. The algorithm obtains ranks of web pages by examining the incoming and outgoing links of pages. Each page with outbound links offers a value rankings focused on fame. These calculations offers high ranking for more known pages and not for non-relevant pages. To any non-relevant page has been given a ranking Based on its frame. And the importance of a page is selected by observing the number of input and output connections. In this algorithm while a series of pages may be unrelated to the user's query, but also have the highest ranking, because it has too many outgoing and incoming links.

Distance rank algorithm is an exact ranking that is suggested in [17] and focused on reinforcement learning. In this approach the distance between the pages has been considered as a main element in calculating pages rank. The advantage of this approach is that by using distance-based order it can rank pages more quickly, and limitation of this approach is that if the new page is placed between two pages, crawlers should do a more calculation to compute the distance vector.

Time rank algorithm is proposed to raise the ranking quality by using the visiting time of the web page [18]. The authors of this method have calculated Pages visiting time by applying unique and strengthened strategies to improve the ranking of web pages. This method uses the time element for Accuracy of the rankings of web pages. The main problem in the method described by Jiang et al [18] is that if unrelated pages stay open for a long time, their visit time will increase and thus the value of their rankings will increase too, while are unrelated to the users' query.

In the methods based on click-through rate such as provided by Jachymz [11] or the document development methodology [12] as well as the method presented in [19] ranking takes place by user's behavior in the previous search. And the ranking for a particular query is done according to the user clicks on the results list, which is provided for a new search. It should be noted that in these methods click action on a web page shows the importance of that page in terms of users. The main problem with this method is that if a page with fake content for any reason to be clicked by users even if the user after opening the web page found it fake and closed it Web page rank goes up because of clicks done.

Reachability based algorithms is an approach that uses the concept of access and hear wavelet for the ranking of web pages [20]. In this way, ranking web pages are represented through the development of a structured signal from the input and output links and the accessibility of web pages from Web graph. So, the ranking of web pages has been developed using hear wavelet. This approach does not involve any iterative technique damping factor or initialization of the page ranks.

The methods presented by Shakeri et al. [14]-[21] are a special case of the weighted output link model. On each page, user reads the content of a single page and is likely to scroll through the output links. These algorithms are a combination of connected-base and content-based methods. The hybrid algorithms are recursive and repeated so as to converge to constant values. Also this ranking algorithms is dependent on queries. In other words, at first the graph originates from documents related to queries forms, and then the above algorithms run. Therefore, the above algorithms will run online. The main problem with these online algorithms is that it reduces the speed of the system in the user's response. Shakeri and Zhai [21], developed by the method [14], have provided a general framework for hybridization text and link information under the title "Release of Dependency". Which uses a possible vision to distribute scores. In this framework, the possible scores divided between the pages linked to the input and output links that are interrelated with the desired page. In contrast to the method [14], the page scores are converted to a probable value before being published and then published.

C-Rank [22] is a hybrid algorithm and combines content and link information in a very effective way using the concept of contribution. However, C-Rank suffers from very high costs to reflect the highly dynamic and extremely frequent changes in the Web, because it re-computes all of the C-Rank scores used for ranking from scratch to reflect the changes. As a result, C-Rank may be considered inappropriate to provide users with accurate and up-to-date search results. Jangwan Koo and et al proposed incremental C-Rank [23], which is designed to update the C-Rank scores of only a carefully chosen portion of the Web pages rather than those of all of the Web pages without any accuracy loss.

sNorm(p) [24] a variant of SALSA [25], combines the p-Norm from the family of vector norms to SALSA, to increase the effect of higher hub weight in a calculation of authority weight and decrease the effect of low hub weight in the calculation of authority weight. This algorithm combines automatically link-based and content-based methods to enhance the accuracy and quality of ranking compared to the salsa algorithm.

That all this algorithms are content-based and link-based. These methods compared to methods of content-based connection offer better performance. However, the main problem is the online algorithms and the algorithms are done in order which makes the system speed be reduced.

III. PROPOSED METHOD

Content-based algorithms are associated with spamming Problems and structure-based algorithms are grappling with the problem of rich-get-richer. To fix these problems many hybridization algorithms have been proposed, for example, paksima and khajeh content and structure based method in [26], Shakeri's proposed hybrid algorithms that is based on content and structure [14]-[21] and Derhami et al proposed hybrid algorithm using user feedback and reinforcement learning [15]. We offer a hybrid method hybridization Distance Rank algorithm [17] and an algorithm based on the average attention time [18]. Distance Rank algorithm is a structure-based algorithm. In this algorithm after calculating the average distance of each page it uses this data to discover the pages associated with the page you want, and the average distance is considered as the standard for rating documents and web pages.

Algorithm in [18] which is an algorithm based on users' behavior, with the development of Firefox, records the length of time each user spends on a web page, and sends it to a server. Use this data to calculate the quality and value of each page and estimate each page’s rank. each of these two methods have its potential which has been mentioned in other sources [4]-[5]. But alongside these abilities each has weaknesses that our proposed method tries by maintaining runtime and complexity of the algorithm, to cover these weaknesses, and increase the accuracy of the rankings as much as possible. In the following we will explain both distance based and attention time based algorithms and then will consider the proposed method.

1. Distance Rank algorithm

Distance rank algorithm [17] is based on reinforcement learning uses logarithmic distance between pages for ranking. The distance between i and j when page i refers to j, is logarithmic output i (number of links) to better understand this concept Bidoki [17] has offered the following definition:

1) If page i points to page j then the weight of the link between i and j is equal to Log O(i) where O(i) shows i's out degree.

2) The distance between two pages i and j is the weight of the shortest path (the path with the minimum value) from i to j and denote it with d_ij. So instead of the normal distance, the new definition of the distance called the average clicks is provided.

The above definitions show that if the loaded page i has d_i distance, the distance of each of her children from the root (referenced pages by i) if they don't have another input, can be calculated using Equation (1).

$$d_j=d_i+\log O(i) \qquad (1)$$

d_j is the distance of the j's child on page i from the root and logO(i) is the received punishment of transition from i to j. In distance rank algorithm if a page has too many links, has less distance than others. And If the referring pages to it have a little distance, the distance of this page also will be little. To display this content a third definition is given as follow:

3) If d_{ij} shows the distance between two-page i and j, d_j denotes the average distance of page j and is defined as the Equation (2) where N shows number of web pages:

$$d_j = \frac{\sum_{i=1}^{N} d_{ij}}{N} \qquad (2)$$

And finally to calculate the amount of distance ranks have provided Equation (3). After calculating the mean vector on all pages, pages are arranged in ascending order according to their distance and thus pages with less will have more rank.

$$d_j=\min_i(d_i+\log (O(i)\,)\,) \qquad (3)$$
$$i\in B(j)$$

2. Time Rank Algorithm

The main idea of this method is based on the amount of time users spend browsing a document that Jiang et al [18]. have used as a clue to calculate the quality and value of a document to help it provide better rankings. To calculate the user's attention, a Firefox browser add-on installed which collects data and transfer it to server. And also in this method Equation 4 has been used to calculate new page attention time, the new pages and documents attention time have been calculated by comparing the content similarity of the new page with previous pages. After identifying the user's attention to a product number of documents, the proposed method can predict the user's attention to a new document via document similarity analysis that can be used both for the text and the images. Jiang et al [18]. offered Equation 4 to predict due time.

$$t_{att}(u,d_x) = \frac{\sum_{i=1}^{k}\left(t_{att}(u,d_i)Sim^{\gamma}(d_i,d_x)\delta(d_i,d_x)\right)}{\sum_{i=1}^{k}\left(Sim^{\gamma}(d_i,d_x)\delta(d_i,d_x)\right)+\epsilon} \qquad (4)$$

γ is the controller of sim (,) and helps to estimate due time. ∊ is small and positive values that will prevent the division by zero error. Function δ (.) filters through the effects of those pages their similarity is below the specified threshold and is defined as Equation (5).

$$\delta(d_i,d_x) = \begin{cases} 1 & \text{If } Sim^{\gamma}(d_i,d_x) > 0.01 \\ 0 & \text{Otherwise} \end{cases} \qquad (5)$$

The Time Rank algorithm [18] is an accurate and powerful method for ranking web pages because it performs rankings according to the users' tastes. The main disadvantage of the method presented in [18] is to open a non-relevent page by the user and to leave that page unrelated to the user's time of doing something other than browsing the open web page, which can damage accurately and reliably of ranking. In our proposed method, we have used a more precise tool to collect user's time-consuming data to cover this weakness as much as possible.

3. Hybrid Method

Our proposed method is a Combination of these algorithms. This means that the proposed method to calculate the value and quality of each page evaluates three parameters of average distance, average attention time and the number of visits. In this case, we can say that our proposed method is a Combination of structure-based algorithms (Distance Rank algorithm) and user feedback-based algorithms (Time Rank algorithm). In our proposed method, a web page with less average distance and high attention time and with a high number of visits will have better rank than other pages. In order to improve time rank problems such as: ensuring the accuracy of attention time data, and accurate record of users' attention time, we used a Firefox add-on called Mind the Time that the add-on stops attention time registration operation after one minute of the latest user activity. The implementation of the proposed method is as follows the function of the distance-based algorithm runs to calculate the average distance each of the pages and the calculated values will be recorded as a value for any web page. During the early stages of labor, ranking is done just by using the value calculated by this function. After each ranking and displaying the results to the user, by using the plugin "mind the time", the amount of time that each user spends on each page, is stored. The plug-computing data for each user (for example: A, B, C,) is done, for each page there is a binary (web page , the time spent by the user):

A:{(P1,T1a),(P2,T2a),(P3,T3a)}
B:{(P2,T2b),(P3,T3b)}
C:{(P1,T1c),(P3,T3c)}

If A to C users do a specific search and P1 to P3 web pages related to their search, T is the amount of time each user has spent on the web page. On the server side using the data from the plugin "mind the time" value of time spent on each page, Pi is calculated with the Equation 6.

$$\mathrm{T}p_i = \mathrm{T}_{\mathrm{A}_{\mathrm{p_i}}} + \mathrm{T}_{\mathrm{B}_{\mathrm{p_i}}} + \mathrm{T}_{\mathrm{C}_{\mathrm{p_i}}} \qquad (6)$$

In equation 6, if T_{pt} is the total amount of time that users spend on web page Pi then T_{Api} is the amount of time spent by the user A on page Pi and the number of references for each web page or npi times is the number of users who have visited that page (Equation 7).

$$\mathrm{n}_{\mathrm{p_i}} = \text{\# of users who visit the web page} \qquad (7)$$

After calculating the number of visits and time spent on each web page, these values like the average distance per page will be recorded. in the Next time, ranking web pages will be done by Using three data, average distance, spent time and the number of visits. It should be noted that any search for a particular query, The amount of time spent by users and the number of visits will be updated and for each new search, the ranking accuracy grows. Increasing ranking accuracy and reducing ranking errors (such as giving high rankings to non-related pages) is expected in the ranking by our hybrid method with multi parameters assessment.

IV. UNITS RESULT AND DISCUSSION

1. Dataset

In order to evaluate the proposed method, we have evaluated a few web pages that are listed in table 1 with the proposed method. To show the relationship between these pages and the better understanding of input and output links between this pages, we draw the graph according to the links between these pages in Figure 2.

TABLE I
DATASET WEB PAGES

Web page address	Title of the Web page
http://www.iaurmia.ac.ir/	Islamic Azad University of Urmia
http://www.iau.ac.ir/	Islamic Azad University
http://amozesh.iaurmia.ac.ir /login.aspx	Islamic Azad University sing up and Information System
http://thesis.iaurmia.ac.ir/	Automation System of Student Thesis
http://transfers.stu.iau.ir/	Student Transfer System
http://bp.swf.ir/	Students Welfare Fund Portal

By using the proposed method in [17], we calculated the average distance for each page from other pages according to links between the pages and Figure 2. The results are shown in Table 2.

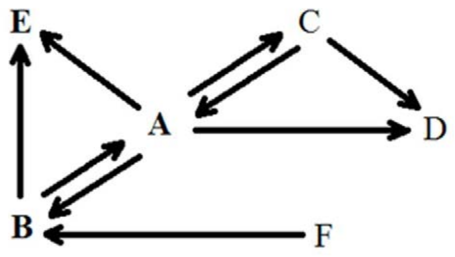

Fig. 2. Web Graph

To get the user's attention time we used the conventional browser add-ons Mind the Time. And we installed it on four systems of Azad University of Urmia computer site for three days (from the date of 2018/08/07 to 2018/08/09) to record user attention time when they are browsing the web pages. And after extraction of the number of visits to each page, user attention time data has been achieved. Table 2 shows the data.

TABLE 2
DATASET

Web page address	Total attention time	Number of visits	Average attention time	Average distance
http://www.iaurmia.ac.ir/	1.2	1	2	2
http://www.iau.ac.ir/	1.4	1	1	1
http://amozesh.iaurmia.ac.ir /login.aspx	1.8	.07	66	335
http://thesis.iaurmia.ac.ir/	1.8	2.33	3	37
http://transfers.stu.iau.ir/	1.6	8	1	8
http://bp.swf.ir/	2.2	5	1	5

In Table 2, numbers related to column 2 (total attention time) is total time that users spent on web page, numbers related to column 3 (number of visits) is total number of users visits for each page, numbers related to column 4 (average attention time) is obtained from diving numbers of column 2 on numbers of column 3 and column numbers "mean distance" indicates the average distance of each of the pages from other pages.

2. Measuring Standards

To evaluate the proposed method, compared with existing algorithms, criterion standard were used. The criteria used in this article are P@n, MRR, MAP, AP, DCG@n and NDCG@n by Kalervo Jarvelin and Jaana Kekalainen which were presented in [27]. Benchmark P@n (Equation 8) criteria represents the number of documents related to the first document which is provided to n. The main objective of this measure is to calculate the accuracy of the system from the user.

$$P@n = \#of\ relevant\ docs\ in\ top\ n\ results/n \qquad (8)$$

According to Table 3 of the proposed method in three stages (precision in the two, three and four first results,) is appropriate.

AP (average accuracy) that is calculated for each query is equal with P@n. In Equation 10, if i be related with query, then rel(i)=1 and otherwise rel(i)=0. And N indicates the number of results for a query.

$$AP = \frac{\sum_{i=1}^{N} P@i \cdot rel(i)}{\#\ \text{total relevant docs for this query}} \qquad (9)$$

MRR (Mean Reciprocal Rank) and MAP (Mean Average Precision) where $|q|$ is a total number of queries, *ri* is the rank position of a first relevant web page in the list of web pages retrieved for the ith query. A value of MRR and MAP range is from 0 to 1. A system that have higher MRR and MAP value as compared to other is more preferable.

$$MRR = \frac{1}{|q|} \sum_{i=1}^{|q|} \frac{1}{r_i} \qquad (10)$$

$$MAP = \frac{1}{|q|} \sum_{i=1}^{|q|} AP_i \qquad (11)$$

DCG (Discounted Cumulative Gain) In this criteria rel(i) shows the degree (3=highly relevant, 2=more relevant, 1=less relevant and 0=irrelevant) of relevancy of i with the relevant query. and NDCG (Normalized Discounted Cumulative Gain) values range is from 0 to 1. NDCG values is calculated by dividing a DCG value for a query by Ideal DCG (IDCG) values for that query.

$$DCG@n = \sum_{i=1}^{n} \frac{2^{rel(i)}-1}{\log(i+1)} \qquad (12)$$

$$NDCG@n = \frac{DCG@n}{IDCG@n} \qquad (13)$$

3. Evaluation Results

In the ranking of web pages, The best ranking in terms of users is that all of the results presented in several first results are related to the user's query and the web pages that have the most relevance to the user's query are at the top of the Search list results. In this regard, considering the criteria for measuring the accuracy and quality of the ranking algorithms presented in the previous section, placing an even non-relevent page can reduce the accuracy and quality of the ranking algorithm's.

After getting the Table 2 data, web pages can be ranked by evaluating three parameters average attention time, average distance and the number of visits. Chart 1 compares our proposed method and Distance Rank [17], Time Rank [18], PageRank [9], HITS [10], Weighted Page Rank

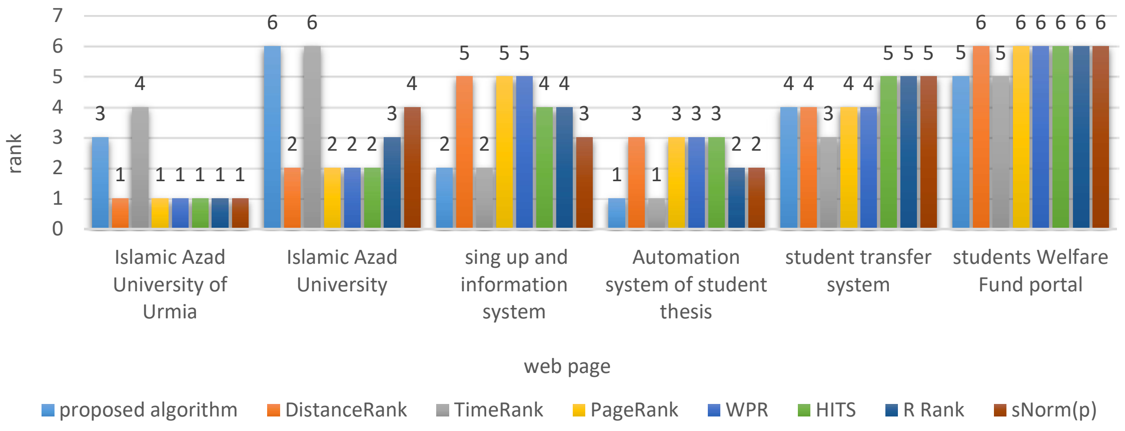

Fig. 3. comparing the ranking with " Urmia Azad thesis " query

[16], Reachability based algorithm [20] and sNorm(p) [24] ranking results with "Urmia Azad university" and "Urmia Azad thesis" queries. figure 3 shows ranking results with "Urmia Azad thesis" query. According to figure 3, the similarity between page rank and distance rank algorithms would be observable. The similarity between page rank and distance rank algorithms as well as the difference in the Time Rank and proposed method ranking with other algorithms indicates differences in practicing of these algorithms with that of Distance Rank, PageRank, HITS and WPR algorithms.

In Figure 3, the numbers are indicating the rank of each page and each of the color bars indicates the rank of the web page in the ranking with different algorithms.

In order to better understanding Figure 3 consider web page of " Automation system of student thesis " in the ranking of the proposed method, Due to the high number of visits to this web page as well as the large number of incoming and outgoing links, It has been awarded first place, While in the ranking with Distance Rank, PageRank, HITS and WPR algorithms third rank belong to this web page and first place in ranking algorithms of the web page belongs to "Islamic Azad University of Urmia" As the data in Table 2 show visiting the web page and the number of visits " Automation system of student thesis " and " sing up and information system " is high. It is understood that rank of a web page that many users are concerned to be higher when compared to other web pages.

V. COMPRESSION

Given that users of search engines focus on just the first few results offered by the search engine, the amount of relevance of the first few results with the relevant query is very important. Therefore, in order to calculate the quality of the algorithms, often the first results are evaluated and compared with other algorithms.

By using the evaluation criteria, we compared our proposed method with algorithms Distance Rank [17], Time Rank [18], PageRank [9], HITS [10], Weighted Page Rank [16], Reachability based algorithm [20] and Snorm(p) [24] in two queries ("Urmia Azad university" and "Urmia Azad thesis"). The results obtained were as Table (3):

TABLE 3
COMPARING WITH P@N CRITERIA

Algorithm	*n=2*		*n=3*		*n=4*	
	Q1	*Q2*	*Q1*	*Q2*	*Q1*	*Q2*
Proposed Method	1	1	1	1	1	0.75
Distance Rank	0.5	0.5	0.66	0.66	0.75	0.5
Page Rank	0.5	0.5	0.66	0.66	0.75	0.5
Time Rank	1	1	1	0.66	1	0.75
WPR	0.5	1	0.66	0.66	0.75	0.75
HITS	0.5	0.5	0.66	0.66	0.75	0.75
R Rank	1	1	0.66	0.66	0.75	0.75
sNorm(p)	1	1	0.66	1	0.75	0.75

According to Table 3, the proposed method in three stages (precision in the two, three and four first results,) is appropriate. In comparing the algorithms with the p@n criterion, we considered three modes (comparison in two, three and four first results). In Table 3, numbers represent the accuracy of the algorithm. These numbers range from zero to one, and the number one represents the highest degree of accuracy. In Table 3, the magnitudes near zero indicate that the algorithm is less accurate. But P@n does not have enough accuracy. Because in this criteria, only a document being related or not, can be considered.

TABLE 4
Comparing with AP criteria

Algorithm	Query1	Query2
Proposed Method	1	1
Distance Rank	0.74	0.75
Page Rank	0.74	0.75
Time Rank	1	0.91
WPR	0.74	0.91
HITS	0.74	0.8
R Rank	0.82	0.91
sNorm(P)	0.82	1

According to Table 4, the numbers represent the degree of relevance of the results in n of the first result (three and four first results). The larger number shows that the first three results obtained from the ranking with the algorithm compared with other algorithms have the highest relevance with user queries. Table 4 shows the average accuracy of the algorithms. In this Table, the numbers range from 0 to 1, and the largest number represents the highest average of accuracy among the algorithms. According to table, our proposed method has a high accuracy than other algorithms. And has better performance than other algorithms.

TABLE 5
Comparing with MAP and MRR criteria

Algorithm	MAP	MRR	
		Query1	Query2
Proposed Method	1	1	1
Distance Rank	0.75	0.25	0.29
Page Rank	0.75	0.25	0.29
Time Rank	0.95	0.5	0.75
WPR	0.83	0.33	0.41
HITS	0.77	0.33	0.33
R Rank	0.86	0.5	0.5
sNorm(P)	0.91	0.5	0.5

In clomun 2 of Table 5, the numbers represent Mean Average Precision of the each algorithms. And in MRR clomuns, the numbers represent the degree of relevance of the n results(three and four first results). Table 5 shows that our proposed method has better Precision and high accuracy than other algorithms. A system with higher MRR and MAP value is always preferable over models having lower values.

TABLE 6
Comparing with DCG@n and NDCG@n criteria

Algorithm	DCG@n		NDCG@n	
	Q1	Q2	Q1	Q2
Proposed Method	1	1	1	1
Distance Rank	0.75	0.75	0.25	0.29
Page Rank	0.75	0.75	0.25	0.29
Time Rank	0.95	0.95	0.5	0.75
WPR	0.83	0.83	0.33	0.41
HITS	0.77	0.77	0.33	0.33
R Rank	0.86	0.86	0.5	0.5
sNorm(P)	0.91	0.91	0.5	0.5

And according to Table 6, the numbers represent degree of relevancy between n first results (three and four first results) with the relevant query. the numbers range is from 0 to 1, and the largest number represents the highest relevancy among the algorithms. The results of comparison with criteria DCG@n and NDCG@n, shows that our proposed method lists the web pages that have the most relevancy with the relevant query at the beginning.

VI. CONCLUSION AND FUTURE WORKS

In this paper, we chose and hybrid two high-precision algorithms from structure-based and user attention time based algorithms. We reduced the ranking errors and increased the accuracy of the ranking by overcoming the defects of this algorithms. In addition, add-on used in our proposed method is more accurate than proposed method by Jiang et al [18]. and this increases ranking accuracy. The results indicate a better performance of the proposed method in comparison with other algorithms. Therefore, it can be concluded that by hybridization a structure-based algorithm like Distance Rank with a user attention time based algorithm, it might be a powerful method for ranking search engine results that ranks search engine results according to interests of users and reduces ranking errors as possible.

Nowadays, time-based algorithms are of paramount importance due to the proper performance from the other algorithms. The main problem of these algorithms is that installing the add-on on the Web browser by the users themselves is not a reasonable idea and accessing to this data is not possible easily for search engines. Due to the ability and the performance of this algorithms and also search engines need to optimize their ranking algorithms, the user attention time data access methods for search engines entails much more work and research.

REFERENCES

1. S. Khalatbari and S.A. Mirroshandel, "Automatic construction of domain ontology using wikipedia and enhancing it by google search engine," Journal of Information Systems and Telecommunication, vol. 3, no. 4, 2015, pp. 248-258.

2. S. Chawla, "A novel approach of cluster based optimal ranking of clicked URLs using genetic algorithm for effective personalized web search," Applied Soft Computing, vol. 46, 2016, pp. 90-103.

3. F.S. Gharehchopogh, & Z.A. Khalifelu, (2011, October). Analysis and evaluation of unstructured data: text mining versus natural language processing. In 2011 5th International Conference on Application of Information and Communication Technologies (AICT) (pp. 1-4). IEEE.

4. Divjot and J. Singh, "Effective Model And Implementation Of Dynamic Ranking In Web Pages," 2015 Fifth International Conference on Communication Systems and Network Technologies, 2015, pp. 1010-1014.

5. R. Chaudhary and M. Bhusry, "A New Contrive to Evaluate Web Page Ranking," Ajay Kumar Garg Engineering College Ghaziabad, India, 2014, pp. 1-6.

6. P. Kumari, P. Ranout , A. Sharma and P. Sharma, "Web Mining - Concept, Classification and Major Research Issues: A Review," 1, 2, 3 & 4 Deptt. of Computer Science and Engineering, Career Point University, Hamirpur,(H.P.) INDIA, 2016, pp. 41-44.

7. S. Viralkumar M, R. J. Patel and . N. Kumar Singh, "Web Mining: A Survey on Various Web Page Ranking Algorithms," International Research Journal of Engineering and Technology (IRJET), vol. 03, no. 04, 2016, pp. 1206-1211.

8. G. M. Salton, A. Wong and. C. Yang, "A Vectore Space Model for Automatic Indexing," Information Retrieval and Language Processing, vol. 18, 1975, pp. 613-620.

9. Langville, A.N., & Meyer, C.D. (2011). Google's PageRank and beyond: the science of search engine rankings. Princeton University Press..

10. J.M. Kleinberg, "Authoritative sources in a hyperlinked environment," Journal of the ACM, vol. 46, no. 5, 1999, p. 604–632.

11. T. joachims, "optimizing search engine using clickthrough data," department of computer science, 2002, pp. 133-142.

12. C. K. a. K. Ramamohanarao, "Long-Term Learning for Web Search Engines," Department of Computer Science & Software Engineering, vol. 2431, 2002, p. 263–274.

13. F. Soleimanian Gharehchopogh, M. Mahmoodi Tabrizi and I. Maleki, "Search Engine Optimization based on Effective Factors of," International Journal of Computer & Mathematical Sciences, vol. 2, no. 1, 2014, pp. 9-13.

14. A. Shakery and C. Zhai, "Relevance Propagation for Topic Distillation," Department of Computer Science University of Illinois at Urbana-Champaign, 2003, pp. 673-677.

15. V. derhami, j. Paksima and h. Khajeh, "Web pages ranking algorithm based on reinforcement learning and userfeedback." Journal of AI and Data Mining, vol. 3, no. 2, 2014, pp. 157-168.

16. W. Xing and A. Ghorbani, "Weighted PageRank Algorithm," Proceedings of the Second Annual Conference on Communication Networks and Services Research, 2004, pp. 305-314.

17. A. M. Zareh Bidoki and N. Yazdani, "DistanceRank: An intelligent ranking algorithm," Information processing & management, vol. 44, no. 2, 2007, p. 877–892.

18. Y. Z. H. J. F. C. L. Songhua Xu, "A User-Oriented Webpage Ranking Algorithm Based on User Attention Time," Proceeding AAAI'08 Proceedings of the 23rd national conference on Artificial intelligence, 2008, vol. 2, pp. 1255-1260.

19. G.-R. Xue, H.-J. Zeng, Z. Chen, Y. Yu, W.-Y. Ma, W. Xi and W. Fan, "Optimizing Web Search Using Web Click-through Data," Proceedings of the thirteenth ACM international conference on Information and knowledge management, 2004, pp. 118-126.

20. S. Hariharan, S. Dhanasekar and K. Desikan, "Reachability Based Web Page Ranking Using Wavelets," 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15) , 2015, vol. 50, pp. 157-162.

21. A. Shakery and C. Zhai, "A Probabilistic Relevance Propagation Model for," Department of Computer Science University of Illinois at Urbana-Champaign Illinois, 2006, pp. 550-558.

22. Kim, D.J., Lee, S.C., Son, H.Y., Kim, S.W. and Lee, J.B., 2014. C-Rank and its variants: A contribution-based ranking approach exploiting links and content. Journal of Information Science, 40(6), pp.761-778.

23. Koo, Jangwan, Dong-Kyu Chae, Dong-Jin Kim, and Sang-Wook Kim. "Incremental C-Rank: An effective and efficient ranking algorithm for dynamic Web environments." Knowledge-Based Systems 176 (2019): 147-158.

24. Goel, Shubham, Ravinder Kumar, Munish Kumar, and Vikram Chopra. "An efficient page ranking approach based on vector norms using sNorm (p) algorithm." Information Processing & Management 56, no. 3 (2019): 1053-1066.

25. Lempel, R. and Moran, S., 2000. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. Computer Networks, 33(1-6), pp.387-401.

26. Paksima, Javad and Homa Khajeh "The surfer model with a combined approach to ranking the web pages," journal of information systems and telecommunication (JIST) , vol. 4, no. 3, 2016, pp. 200-209.

27. K. Jarvelin and J. Kekalainen, "IR evaluation methods for retrieving highly relevant," Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2002, pp. 41-48.