IJICTR
International Journal of Information & Communication Technology Research

Volume 12- Number 1 – Winter 2020 (32 -41)

Web Domains Ranking with Real User Traffic Based on the Big Data Platform

Leila Rabiei, Mojtaba Mazoochi*, Maryam Bagheri
ICT Research Institute (Iran Telecom Research Center)
Tehran, Iran
(l.rabiei, mazoochi, m.bagheri)@itrc.ac.ir

Received: 22 August 2019 - Accepted: 12 December 2019

Abstract—**Disseminating information through the World Wide Web as the most popular medium has resulted in creating a huge number of web pages and so growing the dimension of the web. In this era of big data, an efficient website ranking to satisfy the web user requirements in different areas such as marketing and E-commerce is a major challenge in the current Internet. In this context, the role of ranking algorithms as a tool to provide services such as measuring the website visibility and comparing the website position to the competitors is crucial. In this paper, we propose an architecture for web domain ranking which includes processing capability required for handling Big Data available on the web. The proposed architecture presents a new method for web domain ranking that is independent of the link structure of the web graph. The proposed method provides web domain ranking based on the number of unique visitors, the number of user sessions, and session duration.**

Keywords-Web domain ranking; Web domain importance metric; Web traffic; Traffic analysis component; Big data;

I. INTRODUCTION

Due to the dimension of the World Wide Web, search engines encounter critical challenges such as providing relevant results to the users. A search engine can do its responsibility by scanning its index of web pages which is made using a web crawler. In fact, a web crawler builds up a huge index of many web pages by traversing the web graph and fetching the URLs. The search engines try to sort the results based on their usefulness to the user. To this end, search engines apply ranking algorithms such as PageRank [1] to weigh web pages based on their relevancy to search queries. A web user may receive millions of results in response to his or her simple search queries which is too big to be explored to find the desired result. Therefore, providing the accurate, up-to-date, and authoritative results within the top few pages possesses a special importance [2].

One of the most crucial factors to measure the quality of a website is web traffic which is used as a measure of the popularity and importance of web pages and websites. Websites analytics tools like Google analytics and Alexa use different measurements to present website ranking. Google analytics tracks the users' website activity such as session duration, pages per session, and bounce rate. Gathering information is done by Google Analytics Tracking Code which is added by website owner to every page of the website. If JavaScript is enabled in the browser, the tracking code runs when the client browses the page and collects visitor data and sends it to a Google data collection server [3]. Alexa Rank is designed as an estimation of a website's popularity. Alexa rank is calculated from a combination of unique visitors and page views on a website over a 3-month period. Traffic Ranks are updated daily. Page views are the total number of URL requests of Alexa users for a website, and unique visitors are determined by the number of

* Corresponding Author

unique Alexa users who visit a website on a day. In addition, data normalization is utilized to correct occurred biases [4].

Website ranking can be considered as a tool that measures the popularity of different websites based on some criteria which are defined and measured based on the websites traffic. The detailed data collected from websites is the widely used source of big data. The limitations of conventional data mining methods to mine useful patterns from the web for reliable website ranking has resulted in introducing a term called Big Data analytics. Big Data is defined as a huge and complex collection of data sets which are too large or complex to be processed using traditional database management tools. In fact, because of the complex nature of Big Data, the traditional static Business Intelligence tools can no longer be efficient while powerful technologies and advanced algorithms are required. The mining of Big Data offers many attractive opportunities. However, several challenges may arise when exploring Big Data sets or extracting the knowledge. These challenges can be related to data capture, storage, sharing, analysis, searching, visualization, and management. In addition, there are security and privacy issues especially in distributed data driven applications [5]. Big Data analysis on WWW can be done by employing Hadoop which is a scalable open-source platform for processing Big Data [6,7]. Hadoop can rapidly process large data sets because of its parallel clusters and distributed file system. Hadoop distributed file system (HDFS) is a data storage system which distributes large data across the cluster [8].

In this paper, we propose an architecture for web domain ranking. The architecture consists of three subsystems, namely, traffic data collection subsystem, web domain ranking and traffic analyzing subsystem, and visualization subsystem. The first subsystem receives online and offline data and stores them on system servers. The main responsibility of this subsystem is collecting traffic data. The second subsystem processes the Log files received from the first subsystem based on Big Data and parallel processing. In this subsystem, a new method for web domain ranking is proposed based on the number of unique visitors, the number of user sessions, and session duration. The proposed method can be applied for domains that are blocked in some countries or domains that are down for some days. The third subsystem provides information about the rank and statistics of websites visits to the end user. Since this system deals with Big Data, modern technologies according to the modular structure of the system are applied in implementation of this system. It is worth to mention that the system is implemented according to the OWASP rules, and servers and systems hardening are performed. The experiments have been done with real users over an extended period of time. In summary, the main contributions of this paper are as follows:

- A comprehensive multilayer architecture for web domains ranking with real user traffic based on the big data platform is introduced. It should be noted that this architecture is implemented in practice and is fully functional.
- Unlike many existing systems, the proposed ranking system uses various data sources including script log, extension log and traffic log.
- A new method for web domains ranking is proposed that is independent of the link structure of the web graph. According to our knowledge, no exact method for web domains ranking with real user traffic has been introduced yet.

The remainder of this paper is organized as follows: In Section II, the related works are briefly described. In Section III, the proposed system for web domain ranking is presented. Experimental results are demonstrated in section IV. Finally, a conclusion is provided in Section V.

II. Related Works

Retrieving the relevant information from the web based on the user query is the most important responsibility of search engines. For this purpose, web ranking algorithms are used by search engines to provide the most preferred results. Websites ranking methods can be classified into the following four categories:

- Web Graph / Link analysis based methods: Web page importance is calculated by considering the links to or from other pages [9,1,10-12].
- Content analysis based methods: The idea is considering the keywords relevancy or visiting time of a web page [13,14]
- Comparison based methods: Web ranking is done by comparing the feature vector or score vector of domains [15,16].
- Score based methods: Ranking of pages is based on computing a score which is a combination of some weighted parameters [17-20]

In order to optimize the search engine results, most of the web ranking algorithms are designed based on the context of user queries [2,21,10]. Link analysis is the most widely used method in these algorithms to measure the web page importance which can be calculated by using the link graph of the web. Two well-known link analysis algorithms that have been considered as the basis of lots of developed web ranking algorithms are HITS [9] and PageRank [1] algorithms.

PageRank algorithm which is the heart of Google search engine has been the basis of many web ranking algorithms [22-26]. The main idea of PageRank is based on this assumption that more important websites are expected to receive more links from other websites. So, it counts the number and quality of links to a web page to estimate the importance of the website. A discrete-time Markov chain model simulating a web surfer's random walk on the web graph is defined in which the states are pages, and the transitions are the links between pages. As a result, page importance is calculated as the stationary probability distribution of the Markov chain.

The HITS algorithm is an iterative algorithm that considers two types of web pages namely hubs and authorities within a sub graph of relevant pages. A web page which provides important and trustworthy information on a given topic and so pointed by many hyperlinks is an authority page, while a hub page is the

page point to various hyperlinks and authority pages. Therefore, two scores are assigned to each page. First, the authority score that estimates the value of the content of the page and can be calculated as the sum of the scaled hub values that point to that page. Second, the hub score that estimates the value of its links to other pages by calculating the sum of the scaled authority values of the pages it points to.

The authors in [27] have proposed a stochastic method based on the idea of PageRank and HITS for link-structure analysis, which examines random walks on graphs derived from the link-structure.

In [13], the time factor of the new data source tag is utilized for page ranking. The authors in [14] have proposed a ranking algorithm based on the visit time of the web page. In [15], a generalized Kendall distance is defined to compare the underlying scores with application in comparing web page ranking. The defined metric relies on the margins separating the scores. In [16], a website traffic comparison model via SVM is suggested which can determine the partial order of the traffic information of any two websites. The authors in [17] have proposed a clickstream based metric for Web page importance estimation which is independent of the link structure of the web graph. In [18], the authors have discussed non-textual factors of documents ranking and presented a new document ranking method. In [19], web page importance score is computed based on analyzing user surfing behavior attributes, dwell time, and click count. Then, the ranks are assigned by implementing a Learning Automata.

Also, there are some web traffic analysis services, such as Alexa and Comscore [28,29]. Unfortunately, these services do not use different data sources for ranking. Also, due to the lack of access to a country's traffic log, they cannot provide accurate ranking for the users of that country. In this paper, a traffic ranking system is introduced that uses various data sources including traffic logs for ranking websites.

III. THE PROPOSED RANKING SYSTEM

In this section, we introduce a comprehensive multilayer architecture for web domains ranking system with real user traffic based on the big data platform and also propose some new methods for ranking. The architecture consists of three subsystems; data collection subsystem, web domain ranking subsystem, and visualization subsystem. In the first subsystem the Logs can be collected online or offline. In the second subsystem the Logs are processed and website ranking and traffic analysis are stored in the relational database. Finally, in the third subsystem related statistics are displayed in the users' panels.

A. Traffic Data Collection subsystem

This subsystem which is shown in Fig. 1 consists of three data sources; script, extension, and network Log. This subsystem consists of three layers; data layer, component layer, and security layer. Data layer receives online and offline data and stores them on system servers. Online data sources include scripts and add-ons (or extensions). The script is provided to the websites' owners. It can be inserted on the pages of the websites. The browser plugin which is provided to the users can be downloaded and added to the browser. Thereafter, each visit of that website results in sending a request to the ranking system, and so the visit and its data are recorded. Offline data sources include network logs received from an Internet service provider in the country.

In order to send websites traffic data to a central server, a script written in JavaScript must be uploaded by website owners to all pages on their websites. Each time the page is refreshed, the script sends a request to the central server through which the data items are logged. The script consists of two parts. The first part is placed on each page of a website and loads the second part of the script which is a JavaScript file stored on the system server by creating a dynamic tag. The second part of the script initializes a set of parameters and sends them to the server g. This file contains the JavaScript code which uses cookies to measure the required parameters of visitors' browsers. Since the cookies are created on the client side, the CORS problem does not exist.

In component layer, registering the Log on the server is done through the access log of the web server. In order to retrieve all data through this Log, the required values are contained in the URL as GET parameters. So, the server-side processing load is negligible.

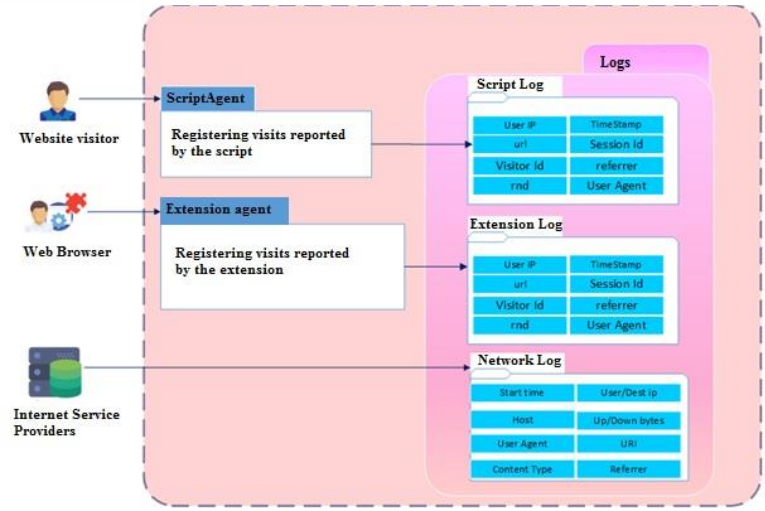

Figure 1. Architecture of Traffic Data Collection Subsystem.

For each visit Log, different parameters are sent to the server which can be classified into two categories, header parameters that are sent with HTTP/HTTPS requests, and adjusted parameters that are defined as Query String in the HTTP/HTTPS request. Header parameters include Time stamp, IP address of visitors, and User-Agent which provides information about the browser, version and type of system, device type (mobile, tablet, or personal computer), screen size, and data. Adjusted parameters include URL address, session ID which can be applied to determine the average session length and the number of pages viewed per session, Visitor ID that is a unique code created on the client side and stored in the cookie, and can be used to provide statistics about single visitors and online users, Referrer field which is used to determine how a user enters a website and what search keywords are used, rnd field which is a random number to prevent caching requests, t field that is used to determine hashchange events or page loading events, and title field.

B. Web Domain Ranking and Traffic Analyzing Subsystem

The second part of the designed ranking system is web domain ranking and traffic analyzing subsystem. The architecture of this subsystem and its layers are shown in Fig. 2 and Fig. 3, respectively. There are four servers in hardware cluster. Managing and monitoring this Hadoop cluster is done by Ambari. In this subsystem, the component layer consists of four parts including pre-processing, criteria calculating, websites ranking, and fraud detection. According to Fig. 4 and Fig. 5, log files which are received from the traffic data collection subsystem are processed in data layer, and the output is inserted in the database which contains statistics of websites visits and rankings. In order to present a trustable web domain ranking, after log processing, fraud detection such as bot-driven fraud is done. After computing defined criteria, web domains scores are calculated and finally ranks are assigned to the web domains based on the computed scores. The required processing tasks, including preprocessing, fraud detection, criteria calculation and websites ranking, are based on Big Data and parallel processing. In data layer, analyzing Big Data on WWW is done by employing HDFS.

This subsystem contains an algorithm to calculate web domain ranks. Now, we propose a method for daily web domain ranking. We pursue the following goals: First, controlling sudden variations in daily domain ranking caused by some special events, second, avoiding similar domain ranks generation, and third, giving preference to domains that are visited more days in a 3-month period.

The traffic parameters that can be calculated on both HTTP/HTTPS networks and HTTP/HTTPS web servers are Unique visitors, Pageviews, Sessions/Visits, and Session/Visit duration. The Pageviews parameter is not an appropriate parameter for calculating the ranking of a website, because the number of web pages can affect this parameter. In addition, mobile applications with push notification ability or mobile applications that use Ajax and Web services (such as social network, messenger, games, etc.), send out a large number of requests in the form of HIT, while separating these requests with Pageviews may not be completely feasible. Also, there are various methods of cheating for Pageviews, i.e., ClickFraud and Blackhat SEO techniques. As a result, the following traffic parameters are considered to calculate the ranking of web domains:

- Unique Visitors
- Sessions / Visits

Session / Visit Duration

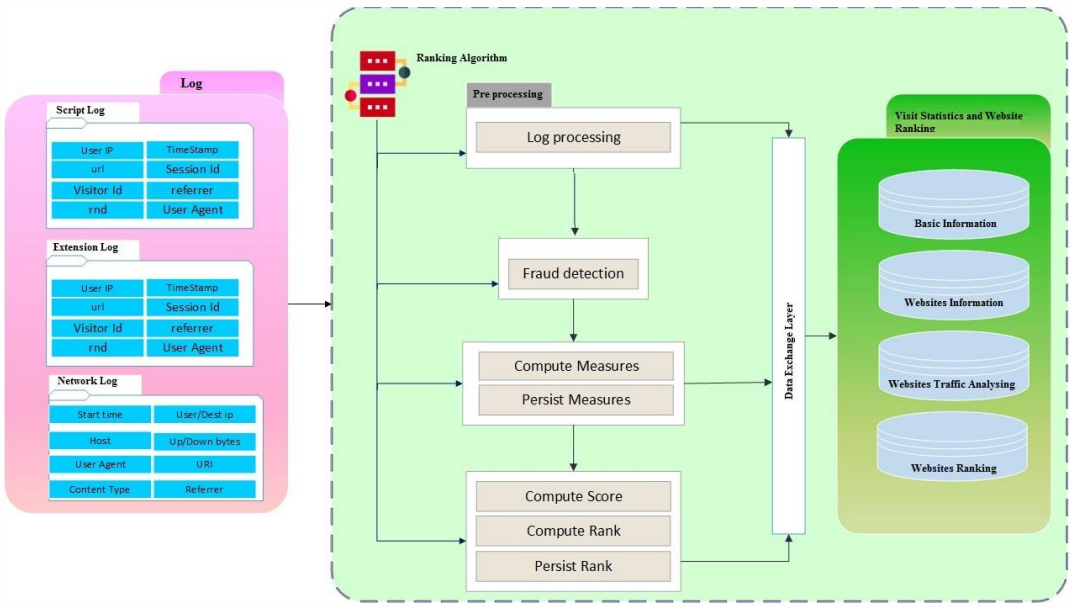

Figure 2. Architecture of web domain ranking and traffic analyzing subsystem.

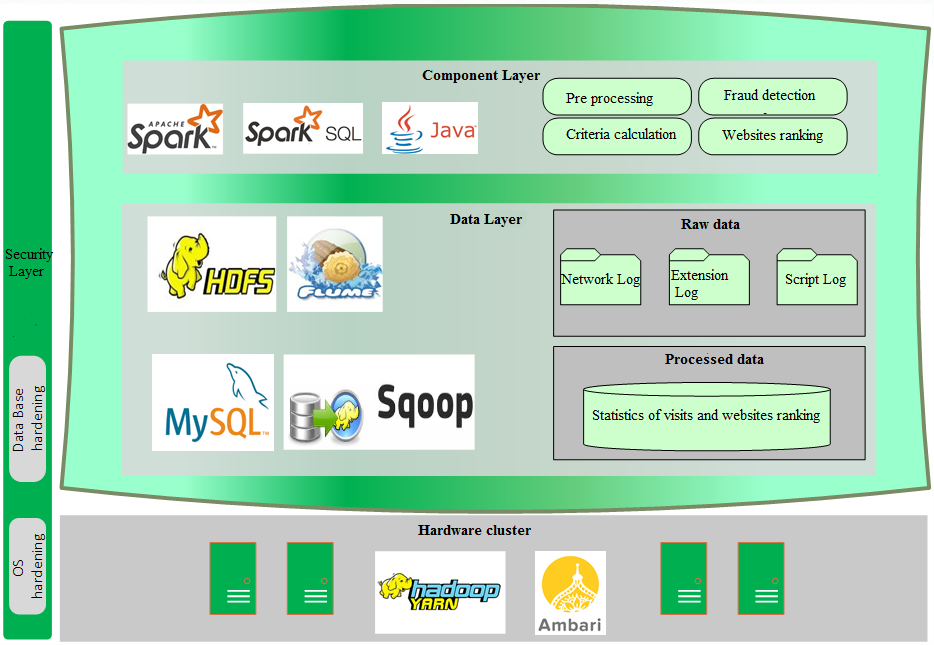

Figure 3. Layers of web domain ranking and traffic analyzing subsystem.

Most of the existing importance metrics in the context of web page ranking are based on the link analysis of the web graph or the context similarity between the queries and indexed web pages. Therefore, they are confronted with spam activities and precision drawbacks. To cope with this problems, our proposed method is inspired by the LogRank method proposed in [17] which is a link independent approach. In LogRank approach, the web page rank is defined as the total page-stay durations from different user sessions per each single page multiplied by the number of distinct user sessions containing visits to that page. The reader may note that the LogRank method is designed to calculate the importance of web pages not web domains. The predominant parameter for determining the page rank in the LogRank algorithm is the number of distinct user sessions. However, this parameter cannot be applied alone to determine the web domain rank. For example, if a domain has a little unique visitor creating a lot of sessions during a day, then the LogRank algorithm might assign the same rank of a domain with a lot of distinct users creating a few sessions that cannot be desirable. As a result, the LogRank algorithm cannot be applied to identify the rank of web domains.

To present a reliable domain ranking, we are confronted with some problems that affect the log view of domains. For instance, some domains might be down or blocked, or the domain name might be changed.

We propose a new notion of web domain importance. To this end, we have an assumption that a domain is more important if visitors spend more time within a unique session on it. To compare the number of unique visitors, the number of created sessions during a day, and the number of domain views during a day of a specific domain with other domains, we use the combination of two parameters: the number of unique visitors, and the number of created user sessions. The first proposed method for ranking web domains is shown in (1).

$$LR2_{D_i} = \left(\frac{\left|UQV_{D_i}\right|+\left|S_{D_i}\right|}{2}\right) \times \frac{\sum_{j=1}^{\left|S_{D_i}\right|} T_{D_i}}{MAX\{\sum_{j=1}^{\left|S_{D_k}\right|} T_{D_k} : k \in K\}} \quad (1)$$

where, D_i is the considered domain, $\left|UQV_{D_i}\right|$ is the number of unique visitors of D_i, $\left|S_{D_i}\right|$ is the whole number of user sessions visiting domain D_i, and T_{D_i} is the user visiting time of domain D_i in a distinct session. This formulation states that the rank of each domain is related to the number of unique visitors, the number of user sessions, and the amount of page views time.

Although (1) can be considered as an improved version of the method presented in [17], the following challenges may be arisen:

- Sudden jumps in daily rankings due to specific events can have a significant impact on the overall visibility.
- There are a lot of equal ranks in the lower quartile ratings of domains.

So, to solve these problems, we try to calculate the daily rank of domains based on an interval of daily ranks.

Let $D = \{d_1, d_2, d_3, \dots, d_n\}$ be the list of unique domains in the log, and $ST(d_i) = \{sd_i t_j : d_i \in D, 1 \le j \le DataRange, sd_i t_j \ge 1\}$ be the list of calculated scores of domains for each day in log.

Method1: In the first method, we have focused on normalizing existing and missing data. In other words, if the scores of some days are not available, the cumulative score is calculated by averaging over existing days' scores. This method is formulated in (2).

$$SCORE_{method_1}(d_i) = \frac{\sum_{j=1}^{j \le |ST(d_i)|} sd_i t_j}{|ST(d_i)|} \quad (2)$$

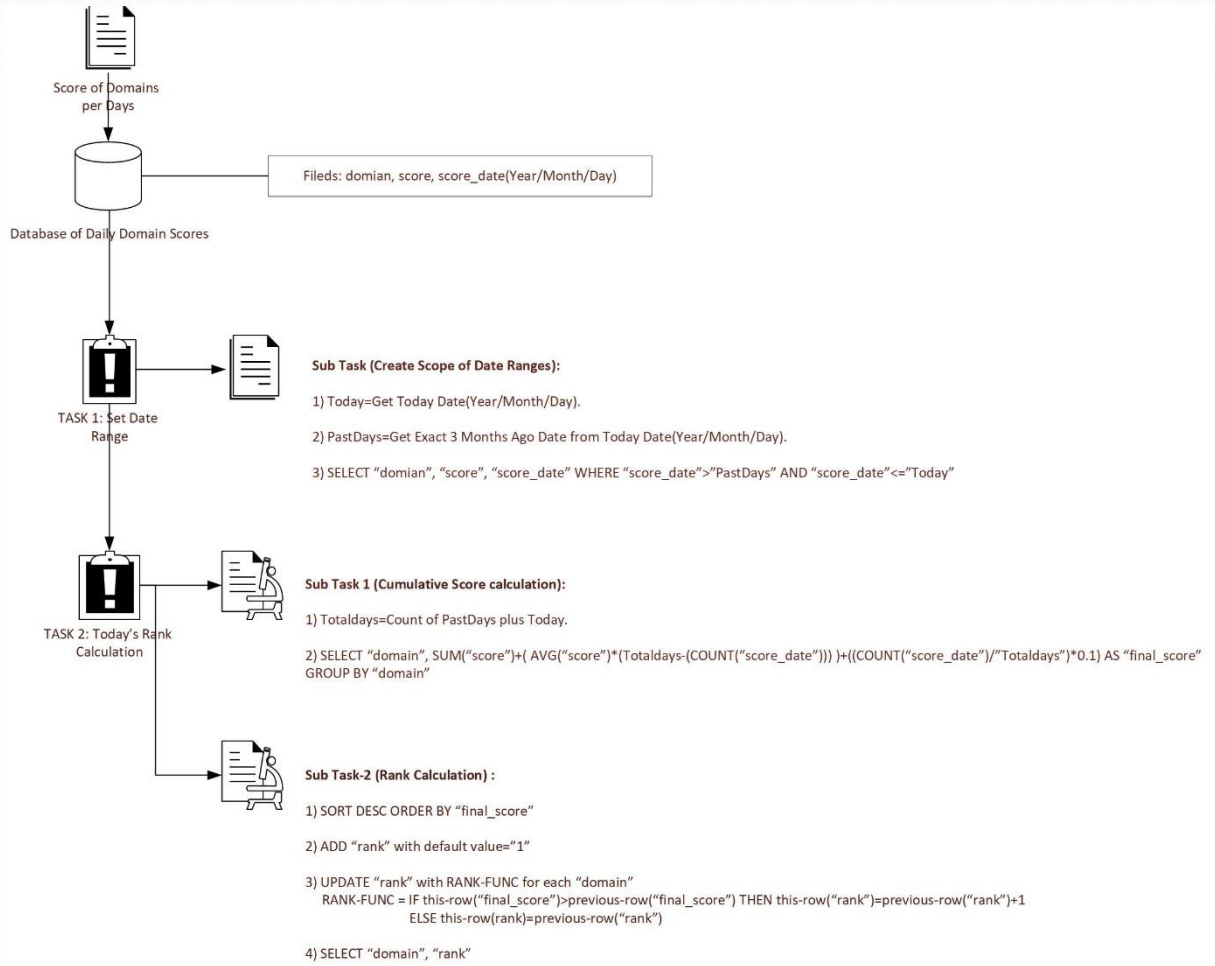

Figure 4. Calculating daily score of domains.

Since there is not a noticeable distinction between domains with different number of available daily score, the fairness is not hold. In addition, lots of similar scores might be generated.

Method 2: in this method, we try to decrease the number of similar scores by giving priority to domains with more daily scores available. We have done it by adding a weighted term of $\frac{|ST(d_i)|}{DateRange}$.

$$SCORE_{method_2}(d_i) = \frac{\sum_{j=1}^{j\leq|ST(d_i)|} sd_i t_j}{|ST(d_i)|} + \left(\frac{|ST(d_i)|}{DateRange} \times 0.1\right) \quad (3)$$

Method 3: In order to hold fairness, sum of available daily scores is considered as the cumulative score of a domain as the follows:

$$SCORE_{method_3}(d_i) = \sum_{j=1}^{j\leq|ST(d_i)|} sd_i t_j \quad (4)$$

Method 4: similar to Method 2, a weighted term of $\frac{|ST(d_i)|}{DateRange}$ is added to (4) to decrease the number of similar scores

$$SCORE_{method_4}(d_i) = \sum_{j=1}^{j\leq|ST(d_i)|} sd_i t_j + \left(\frac{|ST(d_i)|}{DateRange} \times 0.1\right) \quad (5)$$

Method 5: The main proposed strategy based on Method (2) is to calculate the daily rank of domains based on an interval of daily rankings. The cumulative score of domain d_i over the DateRange is calculated as (6).

$$SCORE_{method_5}(d_i) = \begin{cases} \sum_{j=1}^{j\leq|ST(d_i)|} sd_i t_j + \left(\frac{|ST(d_i)|}{DateRange} \times 0.1\right) & |ST(d_i)| = DateRange \\ \sum_{j=1}^{j\leq|ST(d_i)|} sd_i t_j + \left(\frac{\sum_{j=1}^{j\leq|ST(d_i)|} sd_i t_j}{|ST(d_i)|} \times (DateRange - |ST(d_i)|)\right) + \left(\frac{|ST(d_i)|}{DateRange} \times 0.1\right) & otherwise \end{cases} \quad (6)$$

In this method, for days with no score information available, the average of other days' scores is considered.

The flowchart of calculating daily score of domains is shown in Fig. 4. The proposed algorithm works as follows:

1. Daily scores of domains have been calculated and stored in a data base.
2. Daily scores of each domain during the past 3-month is read from the data base.
3. Cumulative score of each domain is calculated based on (2).
4. Domain ranking is assigned based on the obtained scores.

C. *Visualization Subsystem*

In this subsystem the visitor request is received by the web server, and the user information such as ranking information and visit statistics is extracted from the database, and the output will be displayed to the user. As Fig. 5 shows, there are four output display panels for different users in this subsystem: A panel for guest users to display the statistics of websites, a panel for websites administrators, a panel for system administrator for managing tasks and system monitoring, and a panel for web browsers.

IV. EXPERIMENTAL RESULTS

The proposed web domain ranking system is designed and developed by agile software development. The Scrum methodology, which is an iterative model of agile framework, is used. Now, to estimate the system components characteristics, we need to estimate the number of unique domains, the amount of raw data input, and the number of daily requests to the script server. To this end, an available Log of one of mobile operators has been used. According to this Log, the number of domains is about one million, and the amount of input data to the system is approximately 1.5 TB. To provide Big Data processing the Hadoop cluster includes 3 worker nodes with SSD Hard, and 2 master nodes. The number of replications is three for worker nodes, so the amount of data is about 4.5 TB, and if $\frac{amount\ of\ raw\ data}{amount\ of\ processed\ data} = 50$, amount of input data to the system is 500 GB per day. The number of requests is estimated as 615 requests in a second, and the number of system website visits is estimated as 230 visits in a second.

The average value of utilized traffic parameters for nine websites from 08/08/2018 to 1/20/2019 are shown in Figures 6-9. The numbers in vertical axis are in logarithmic scale. Fig. 10 and Fig. 11 show the rank score and rank number of these websites, respectively.

In order to compare five proposed methods for web domain importance metric, we have used three 1-2 hours log files for 3 days. Scores of domains in these three days are collected for the proposed methods, and the results are shown in Table I. The number of similar scores and the standard deviation of cumulative scores which is a measure of scores' dispersion are two considered parameters.

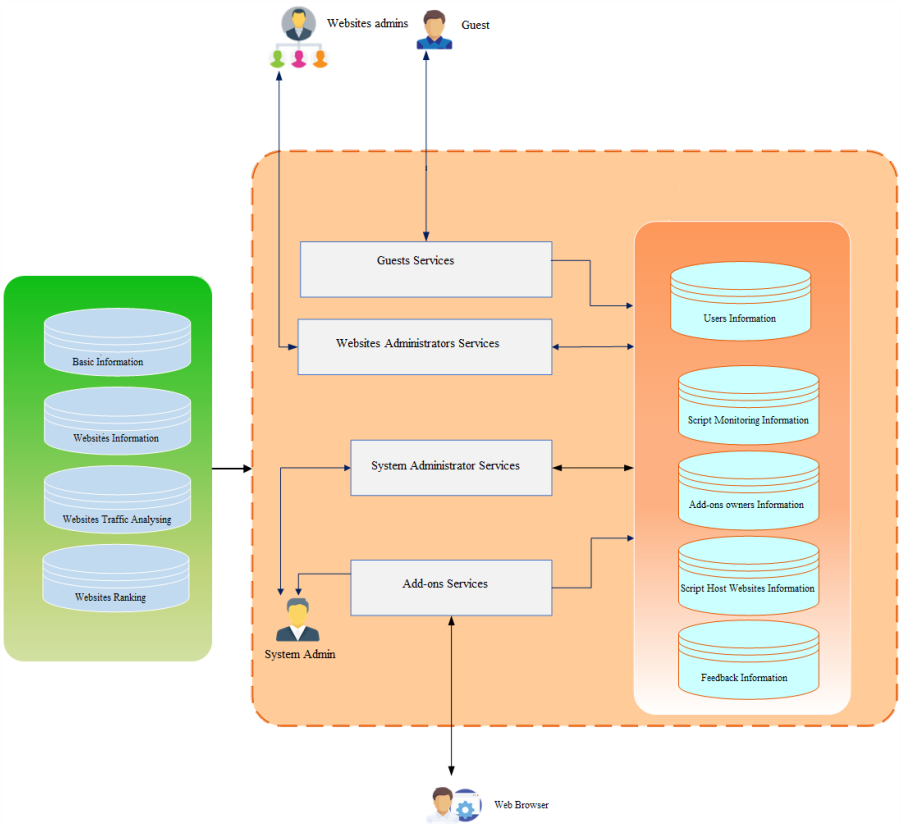

Figure 5. The architecture of visualization subsystem.

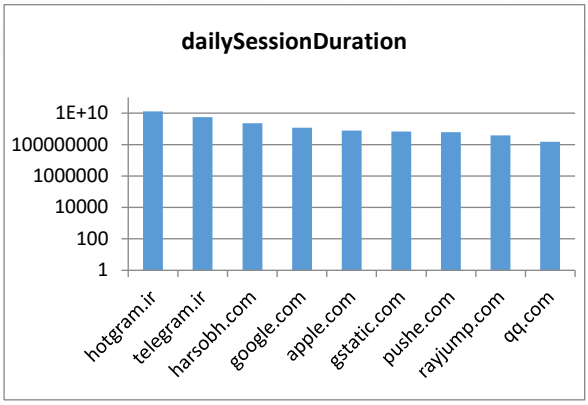

Figure 6. Daily session duration.

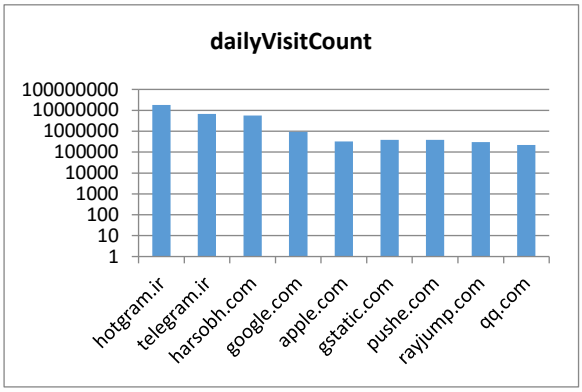

Figure 7. Daily visit count.

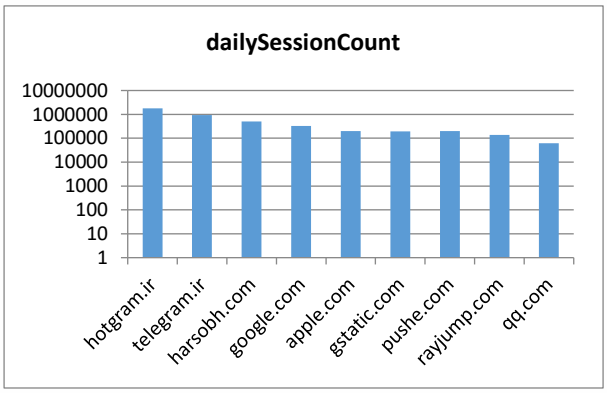

Figure 8. Daily session count.

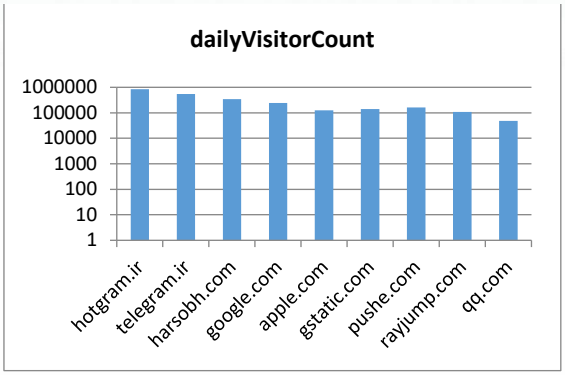

Figure 9. Daily visitor count.

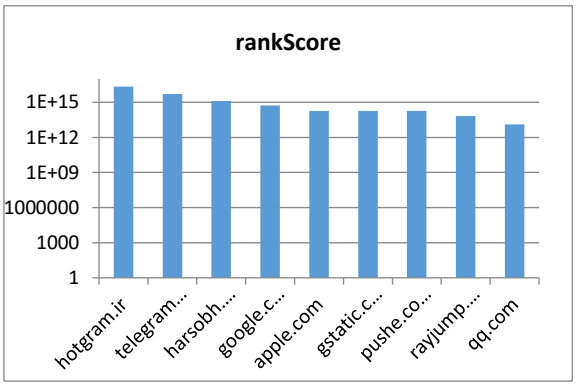

Figure 10. Rank score.

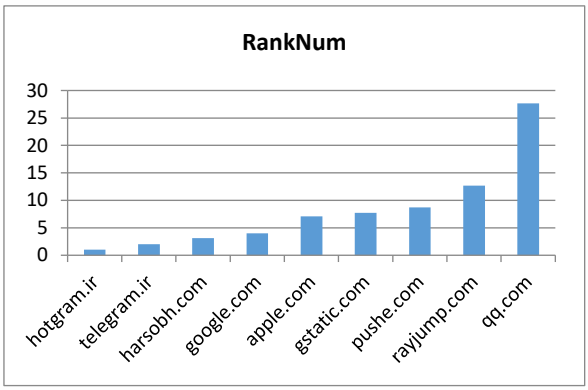

Figure 11. Rank number.

TABLE I. COMPARISON OF PROPOSED METHODS

Method	Number of similar scores	Standard deviation
Method 1	17430	948685163.5
Method 2	16919	948685163.5
Method 3	17797	2846055380
Method 4	16919	2846055380
Method 5	16919	2846055331

The number of unique domains in the log is 21382. According to table I, method 2, 4, and 5 have provided the least number of similar cumulative scores of domains. Since methods 1, and 3 consider the average of scores, the daily scores increase ineffectively. Regards the standard deviation, this table demonstrates that methods 4, and 5 have provided the best results among these methods. The most important feature of method 5 is considering the history of scores during the last 3 months, therefore the overall ranks of websites are more realistic. For example, if a domain draws attention from users temporarily because of a special event, its overall rank is not affected dramatically by this event.

V. CONCLUSION AND FUTURE WORK

Although there are some web traffic analysis services, such as Alexa and Comscore, but these services do not use different data sources for ranking. Also, due to the lack of access to a country's traffic log, they cannot provide accurate ranking for the users of that country. In addition, according to our knowledge, no exact method for web domains ranking with real user traffic has been introduced yet. In this paper, we have proposed a comprehensive multilayer architecture for web domains ranking with real user traffic based on the big data platform that

uses various data sources including script log, extension log and traffic log. Five new methods for web domains ranking were proposed that were independent of the link structure of the web graph. The proposed methods provide daily web domain ranking based on the number of unique visitors, the number of user sessions, and user sessions duration which includes processing capability required for handling Big Data available on the web. The proposed methods are able to present domain rank even for blocked or down web domains.

The proposed web domain ranking system has been designed and developed by agile software development. In order to compare five proposed methods, we used three 1-2 hours log files for 3 days. Scores of domains in these three days are collected for the proposed methods, and the results are shown in Table I. The number of similar scores and the standard deviation of cumulative scores which is a measure of scores' dispersion are two considered parameters. The experimental results demonstrated the efficiency of the proposed methods. Also, we showed that methods 4, and 5 provided the best results among these methods.

For the future work, we intend to implement a bot detection algorithm to detect traffic generated by bots. Also, log correlation and data fusion techniques can be used to improve performance of the web domains ranking system. In addition, the content analysis of the top domains identified by the ranking system can be used to determine the interest of web users.

REFERENCES

[1] Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web. In. Stanford InfoLab, (1999)

[2] Signorini, A.: A survey of Ranking Algorithms. Department of Computer Science, University of Iowa (2005).

[3] https://developers.google.com/analytics/resources/concepts/gaConceptsTrackingOverview.

[4] https://www.alexa.com/siteinfo/wikipedia.org.

[5] Oussous, A., Benjelloun, F.-Z., Lahcen, A.A., Belfkih, S.: Big Data technologies: A survey. Journal of King Saud University-Computer and Information Sciences 30(4), 431-448 (2018).

[6] Khullar, R., Sharma, T., Choudhury, T., Mittal, R.: Addressing Challenges of Hadoop for BIG Data Analysis. In: 2018 International Conference on Communication, Computing and Internet of Things (IC3IoT) 2018, pp. 304-307. IEEE

[7] Malhotra, D., Malhotra, M., Rishi, O.: An Innovative Approach of Web Page Ranking Using Hadoop-and Map Reduce-Based Cloud Framework. In: Big Data Analytics. pp. 421-427. Springer, (2018)

[8] White, T.: Hadoop: The definitive guide. " O'Reilly Media, Inc.", (2012)

[9] Kleinberg, J.M.: Hubs, authorities, and communities. ACM computing surveys (CSUR) 31(4es), 5 (1999).

[10] Borodin, A., Roberts, G.O., Rosenthal, J.S., Tsaparas, P.: Link analysis ranking: algorithms, theory, and experiments. ACM Transactions on Internet Technology (TOIT) 5(1), 231-297 (2005).

[11] Sethi, S., Dixit, A.: A novel page ranking mechanism based on user browsing Patterns. In: Software Engineering. pp. 37-49. Springer, (2019)

[12] Sangamuang, S., Boonma, P., Natwichai, J., Chaovalitwongse, W.A.: Impact of minimum-cut density-balanced partitioning solutions in distributed webpage ranking. Optimization Letters, 1-13 (2019).

[13] Jie, S., Chen, C., Hui, Z., Rong-Shuang, S., Yan, Z., Kun, H.: Tagrank: a new rank algorithm for webpage based on social web. In: 2008 International Conference on Computer Science and Information Technology 2008, pp. 254-258. IEEE

[14] Jiang, H., Ge, Y.-X., Zuo, D., Han, B.: TIMERANK: A method of improving ranking scores by visited time. In: 2008 International Conference on Machine Learning and Cybernetics 2008, pp. 1654-1657. IEEE

[15] Bhamidipati, N.L., Pal, S.K.: Comparing scores intended for ranking. IEEE Transactions on Knowledge and Data Engineering 21(1), 21-34 (2009).

[16] Ren, P., Yu, Y.: Web site traffic ranking estimation via SVM. In: International Conference on Intelligent Computing 2010, pp. 487-494. Springer

[17] Ahmadi-Abkenari, F., Selamat, A.: A clickstream based web page importance metric for customized search engines. In: Transactions on Computational Collective Intelligence XII. pp. 21-41. Springer, (2013)

[18] Kiewra, M., Nguyen, N.T.: Non-textual document ranking using crawler information and web usage mining. In: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems 2005, pp. 520-526. Springer

[19] Makkar, A., Kumar, N.: User behavior analysis-based smart energy management for webpage ranking: Learning automata-based solution. Sustainable Computing: Informatics and Systems 20, 174-191 (2018).

[20] Sankar, S., KUNNATUR, S., Dhamdhere, K.: Ranking search results using diversity groups. In. Google Patents, (2018)

[21] Sharma, D.K., Sharma, A.: A comparative analysis of web page ranking algorithms. International Journal on Computer Science and Engineering 2(08), 2670-2676 (2010).

[22] Baeza-Yates, R., Davis, E.: Web page ranking using link attributes. In: Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters 2004, pp. 328-329. ACM

[23] Kumar, G., Duhan, N., Sharma, A.: Page ranking based on number of visits of links of web page. In: 2011 2nd International Conference on Computer and Communication Technology (ICCCT-2011) 2011, pp. 11-14. IEEE

[24] Duhan, N., Sharma, A., Bhatia, K.K.: Page ranking algorithms: a survey. In: 2009 IEEE International Advance Computing Conference 2009, pp. 1530-1537. IEEE

[25] Xing, W., Ghorbani, A.: Weighted pagerank algorithm. In: Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004. 2004, pp. 305-314. IEEE

[26] Sharma, A., Duhan, N., Kumar, G.: A novel page ranking method based on link-visits of web pages. International Journal of Recent Trends in Engineering and Technology 4(1), 58-63 (2010).

[27] Lempel, R., Moran, S.: SALSA: the stochastic approach for link-structure analysis. ACM Transactions on Information Systems (TOIS) 19(2), 131-160 (2001).

[28] https://www.alexa.com.

[29] https://www.comscore.com.

Leila Rabiei received her B.Sc. degree in Computer Engineering from Islamic Azad University of Tehran, Iran, and her M.Sc. degree in Computer Engineering from Iran University of Science and Technology, Tehran, Iran. She is currently works as a researcher and project manager in ICT Research Institute (ITRC), Tehran, Iran. Her research interests include big data analysis, data mining and social networks analysis.

Mojtaba Mazoochi received his B.Sc. degree in Electrical Engineering from Tehran University, Iran in 1992. He received his M.Sc. degree from Khajeh Nasir Toosi University of Technology, Iran in 1995 and his Ph.D. degree from Islamic Azad University, Tehran, Iran in 2015 in Electrical Engineering (Telecommunication). He is an assistant professor and deputy of IT faculty in ICT Research Institute (ITRC), Tehran, Iran. His research interests include data analytics, social networks analysis, quality of service (QoS), and network management.

Maryam Bagheri received her B.Sc. degree in 2009 and M.Sc. degree in 2011 both in Information Technology from Computer Engineering Department, Sharif University of Technology (SUT), Tehran, Iran. She currently works as a researcher in ICT Research Institute (ITRC), Tehran, Iran. Her research interests include Social Computing, Data Analytics and Data Mining.