Data and Knowledge Extraction Based on Structure Analysis of Homogeneous Websites

Mohammed Abdullah Hassan Al-Hagery
Qassim University, Faculty of Computer, Department of IT
Buraydah, KSA

***Abstract*—The World Wide Web includes several types of website applications. Mainly these applications are related to business, organizations, companies, and others. There is a lack to get raw data sets to study the behavior of the internal structure of each type of these websites. Where websites structures include treasure of links, and sub-links, in addition to some embedded features associated with the internal structure of each website. The objective of this paper is to analysis a set of homogeneous websites to establish raw data sets. These data sets can be employed for several research purposes. It also can be used to extract some invisible aspects/features within the structure. Several steps are required to accomplish this objective; first, to propose an algorithm for structure analysis, second, to implement the proposed algorithm as a software tool for the purpose of extraction and establishment of raw data sets (real data set), third, to extrapolate a set of rules or relations from these data sets. This data set can be employed for researches purposes in the field of web structure mining, to estimate important factors related to websites development processes, and websites ranking. The results comprise creation of Oriented Data Sets (ODS) for research purposes and also for deducing a set of features represents a type of new discovered knowledge in this ODS.**

Keywords—Hyperlinks Analysis Tools; Features Extraction; Oriented Data Sets generation; Knowledge Discovery in Oriented Data Sets

I. INTRODUCTION

The internet is becoming a main communication tool between various people, companies, and organizations in society. It has become the dominant force in the world in various areas where the information revolution represents the strength of economy and uplifts the level of people's life. In addition, the internet has entered many practical areas in our life such as business management, purchase operation, follow up economy of knowledge stocks, trading prices, currencies index, banking services, E-governments, distance learning, hospitals, and other areas. The internet websites contain huge amount of data. It includes simple and complex data within web links. Although there are many companies and organizations established some software tools and methods used to analyze these web links or web contents. Most of them are focusing on particular attributes, relevant to a specific objective. All these tools cannot be employed to collect the required data for this research.

This research is leading to a new idea focusing mainly on the analysis of web structure components that depend on the dynamic hyperlinks. This task is used to extract different useful features of a website. Few numbers of these features will be applied as needed in this research, for example, in estimation of websites' sizes or in ranking processes relevant to websites structure. The extracted data and features in this paper will be directed for serving websites development process as an extension for this research.

II. LITERATURE REVIEW

There are many researches and studies that focus on web structure analysis and web structure mining, but with different objectives. All these works are based on using many tools and methods that serve the goals of such research and studies. Some of these works will be discussed here. Many research works have been undertaken and different solutions have been suggested to the problem of searching, indexing or querying the web, taking into account its structure as well as the meta-information included in the hyperlinks and the text surrounding them [10], [5], [12] and [16]. There are a number of algorithms proposed based on link analysis. Dean and Henzinger [7] proposed an algorithm to exploit only the hyperlink-structure (i.e. graph connectivity) of any Website and does not examine the information about the content or usage of pages or structure components. Brin and Page [4], addressed the question of how to build a practical large-scale system which can exploit the additional information present in hypertext. They aimed to speed up Google considerably through hardware distribution, software, and algorithmic improvements. Their target was to handle the several hundred queries per second.

Jon and Kleinberg [15], developed a set of algorithmic tools for extracting information from hyper link structures from web environments. They did some experiments that demonstrated their effectiveness in a variety of contexts on the WWW. The central issue they addressed within their framework was the distillation of broad search topics, through the discovery of "authoritative" information sources on such topics. They proposed and tested an algorithmic formulation of the notion of authority, based on the relationship between a set of relevant authoritative pages and the set of "hub" pages" that joined them together in the link structure.

Xing and Ghorbani developed a Weighted PageRank (WPR) algorithm. It was an extension to the standard PageRank algorithm. The developed algorithm takes into account the importance of both the in-links and the out-links of the pages and distributes rank scores based on the popularity of the pages. The results showed that the WPR performs better than the conventional PageRank algorithm in

terms of returning larger number of relevant pages to a given query [20].

Taherizadeh and Moghadam proposed an approach to integrate web content mining into web usage mining. The textual content of web pages is captured through extraction of frequent word sequences, which are combined with web server log files to discover useful information and association rules about users' behaviors [19].

Kao and Lin have proposed an algorithm called DRank to diminish the bias of PageRank-like link analysis algorithm that attains better performance than Page Quality. In their algorithm, they modeled web graph as a three-layer graph which includes Host Graph, Directory Graph, and Page Graph by using the hierarchical structure of URLs and the structure of link relation of Web pages [14]. In addition, Kumar and Singh introduced a study on hyperlink analysis. They analyzed the links in order to retrieve web information. They used Google search engine and different algorithms for link analysis, such as PageRank, Weighted PageRank, and Hyperlink-Induced Topic Search algorithms [17].

Jeyalatha and Vijayakumar [13] proposed and implemented a web link extraction tool to deal with web structure using Java and standard interface. They used the Breadth First Search strategy. This work is mainly focusing on performing a quick check on search links, analyzes the structure information from the web that includes document structure & hyperlinks, to Crawl HTML files, and counts the number of occurrences of the keywords in those files. The research helps web users, faculty, students and Web administrators in a university environment.

Mishra, et al. introduced their work based on PageRank created to rank the results of a search system based on a user's topic or query. More than one algorithm was proposed in their work [18]. Derouiche et al. in [8], presented a novel approach for extracting structured data from websites, and the goal was to harvest real-world items from the structured web. They proposed an alternative approach to automatic information extraction and integration from structured Web pages.

Based on the topology of the hyperlinks, Web Structure mining categorizes the web pages and generates the information like similarity and relationship between different web sites. This type of mining can be performed at the document level (intra-page) or at the hyperlink level. It is important to understand the web data structure for information retrieval [18]. The web structure analysis can be performed by several ways, based on the type of required information or knowledge, then creating or using appropriate software tools to accomplish the analysis task. Some tools were employed to identify the relationship among web pages based on their contents or direct link connection, and other tools were created for different objectives. Birla et al., reported that the web is a treasure of information and data, where large amounts of data are available in different formats and structures. Finding the useful data from the web is a complex task [3].

Some of features extraction techniques are based on extracting content based features. However, many such solutions have been handcrafted and thus not guaranteed to work optimally under all data environments. Anand in his research explored an evolutionary algorithm based feature extraction techniques. This work explores Evolutionary algorithms based feature extraction techniques where the extracted features are used to describe user or item profiles [1]. On the other hand, Benslimane et al. proposed in their research idea a novel approach for reverse engineering data-intensive web application into ontology-based semantic web. They analyzed the HTML pages structure to identify its components, interrelationships, and extract a form model schema [2]. As discussed above and although, there is a number of tools and crawlers which can be used to analyze the websites, but it seems clear that the collected data sets by these tools were focusing to solve specific problems, not for everything we need, so one may not be able to take advantage of them, for example in the field of prediction and websites development and ranking as it is in the data sets of this search.

III. Websites Structure Anlysis

Web requirements include three classes: functional requirements, non-functional requirements, and other requirements. The traditional information retrieval system focuses on information provided by the text of web documents. Web mining technique provides additional information through hyperlinks where different documents are connected. The web may be viewed as a directed labeled graph where nodes are the documents or pages and the edges are the hyperlinks between them. The directed graph structure in the web is called as web graph [11]. A web can be imagined as a large graph containing several hundred million or billions of nodes or vertices and a few billion arcs or edges [17]. Link mining is divided into four parts; external structure mining, internal structure mining, URL mining, and web usage mining [6]. The structure analysis can be applied in several areas, such as query ranking, webpages importance, pages classification, and clustering. The objective of these types of analysis is to find the most related pages, redundancies, and measuring the similarity degree among pages.

IV. Problem Statement

There are many important features within the web structure that are invisible. This research is mainly focusing on the structure analysis of homogeneous websites based on link analysis, as a step forward for web structure mining. The challenge for this type is to deal with the structure of hyperlinks within the web itself. Link analysis is an old area of research. However, with the growing interest in web mining, the research of structure analysis had increased and these efforts had resulted in a newly emerging research area called Link Mining [9], [6]. There is a difficulty to find a tool to produce the appropriate data relevant to the website hyperlinks contents and structure. Although there are many algorithms used to analyze the web links, these algorithms cannot be used to collect the required data for a specific objective, as in the objective of this research.

V. Objectives

The objective of this research is to propose and implement an algorithm in order to analyze hyperlinks of many homogeneous websites, to collect ODS to be used for research

purposes, and also to extract some hidden features. The extracted features/rules can be used to help websites' developers to employ the results of this research in measurement and estimation models relevant to websites domain.

VI. Research Methodology

The analysis process concentrates on follow-up of dynamic hyperlinks of each website in order to discover interconnected links to find some important features. The research steps include; algorithm design, implementation, links analysis, generation of ODS, and features extraction.

A. The proposed algorithm

The proposed algorithm is designed to process websites' hyperlinks and to extract the required ODS & embedded features. Each website has its own structure and attributes values. The analysis results of hyperlinks are different from one website to another. The ODS have established based on the proposed algorithm. It has constructed for several homogeneous websites including many attributes. The Pseudo-codes of this algorithm are illustrated in the following two segments.

1. *Start*
2. *Identify a list Li of Homogeneous websites links (HWL),*
3. *Li_length := N, Li={R1, R2,… RN} // N is size sample*
4. *Create ODS with size N;*
5. *Define a set of oriented attributes to be extracted from each Page*
6. *For each web Page i= 1; i > 0 and i ≤ N do*
7. *Begin*
8. *Read First Link(Ri);*
9. *Initlize (Oriented_attrib_Record);*
10. *S_Analysis (Ri , Oriented_attrib_Record); // Proc- call*
11. *Save (Oriented_attrib_Record, ODS[i]);*
12. *End;*
13. *Finish*

The proposed algorithm can be used to extract the same attributes of any type of website not only the homogeneous websites.

1. *S_analysis(Root_R:Link, Data_Rec: record): Record;*
2. *If (Current_Root_link=end of last branch then exit*
3. *If (Current_Root_link<>Terminal leave OR external Link)*
4. *Begin*
5. *Scan all possible pages connected with Current_Root_link*
6. *Calculate the attributes of the Current Page*
7. *Update the data record*
8. *If there are links extensions for the current Page then*
9. *Begin*
10. *J:= the number of sub links of Current_Root_link;*
11. *Repeat*
12. *Get (Link j);*
13. *S_Analysis(Link j , Data_rec); // Recursive Call*
14. *J:= J -1;*
15. *Until (j<=0)*
16. *End // if*
17. *End*
18. *Refresh & Return(Data_rec);*
19. *Finish*

B. Web Analysis Processes

The Web Analysis task includes many steps as shown in Figure 1. The steps are started by a user who enters the main hyper link/root of a website. The tool receives the root as an input and starts the analysis. It follows-up the sub links to the depth of the website in several directions and extracts the required data from all internal links. The extraction process stops when the following of current link is external link or final leaves.

The algorithm listed above illustrates the process of following-up the links. These algorithms developed as S/W tools that include several steps. These tools implemented by the PHP programming language based on its facilities enabling the tool to go to the depth of the website and to follow-up the discovery process of all hyperlinks. It is able to deal with web structure components and its contents. The main challenge of this research from the first step is to get the required data sets, for the purpose of showing this work into presence and leading it to success. The research data set is established and organized through several steps as real data sets. The proposed algorithm and the designed tools were essentially developed based on the needs of this research to achieve the desired objectives (web structure analysis, raw data establishment, and extraction of a special type of knowledge). The PHP language was applied through the local host XAMP. It is a suitable software environment for this work.

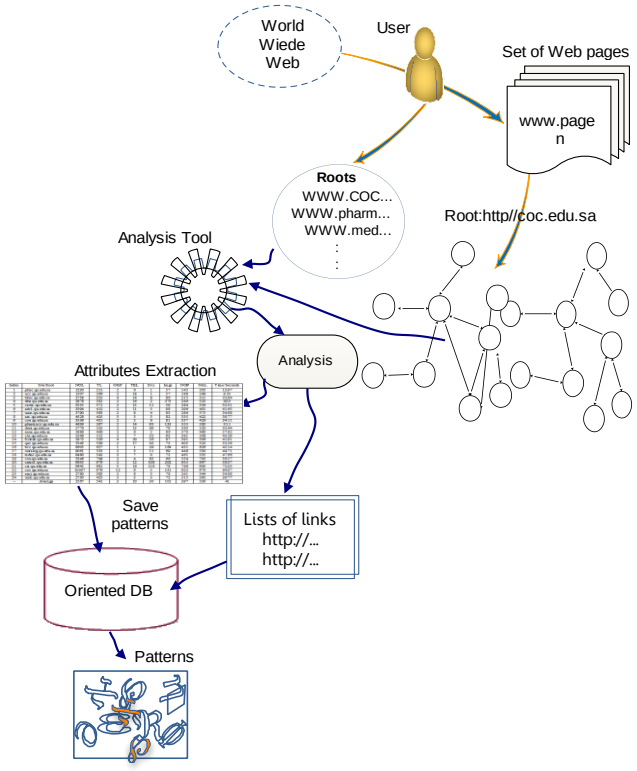

Fig. 1. Websites structure analysis

VII. Research Results

There are two types of results; first, set of raw data, denoted by ODS. Second, set of embedded features or rules. The following two sections A and B obtain more details about extracted results.

A. Creation of raw Data sets

Raw data sets consist of several attributes related to each website structure, such as, Total links, External Links, Number of Leaves, Active Links, No of Pages, Images, Docs, Other Files, Analysis Time/seconds, and etc. as presented in Table I and in Figure 2, these attributes were employed for new features extraction. The ODS extracted from 24 websites related to educational field from the websites of Qassim University. The contents are shown in Table I. This sample includes the following attributes; Site Root, Total number of Leaves (NOL), Total number of links (TL), Total number of External Links (TEL), Total number of Active Links (NAL), Total number of Pages (NOP), Images (Imgs), Docs (Dcs), Other Files (OthF), and analysis Time estimated in seconds. Figure 2 shows a part of the analysis results of one website at Qassim University.

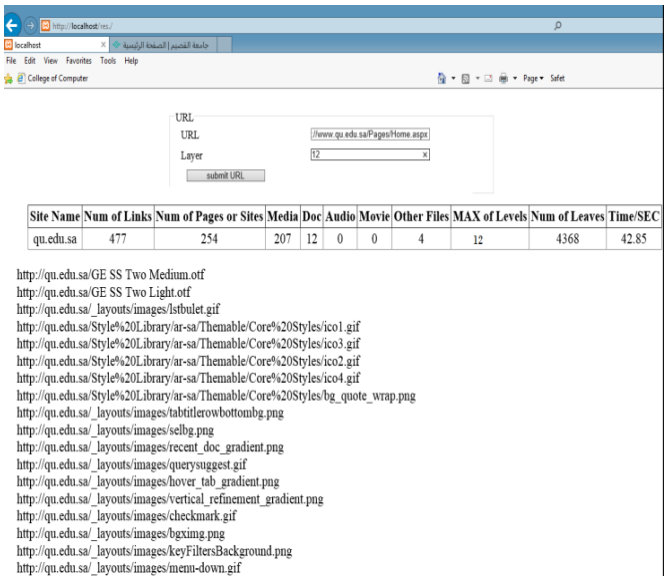

Fig. 2. Some extracted attributes of a single website

TABLE I. A Set of Extracted Attributes of 24 QU Colleges

I	Site Root	NOL	TL	OthF	TEL	Dcs	Imgs	NOP	NAL	Time/Seconds
1	phuc.qu.edu.sa	1230	211	2	6	1	57	145	205	13.97
2	ucc.qu.edu.sa	1507	266	2	18	1	57	188	248	9.55
3	bhsc.qu.edu.sa	2758	325	4	14	8	86	215	311	22.94
4	dhe.qu.edu.sa	2878	541	2	16	2	275	246	525	40.3
5	cavm.qu.edu.sa	3555	371	2	15	11	59	284	356	32.31
6	adc1.qu.edu.sa	3394	412	2	11	5	98	296	401	41.65
7	uasc.qu.edu.sa	3790	383	2	8	4	83	286	375	34.88
8	asc.qu.edu.sa	4528	425	3	3	3	82	335	422	39.77
9	coe.qu.edu.sa	3539	431	2	5	6	81	337	426	34.11
10	pharmacy.qu.edu.sa	4639	597	2	14	93	133	355	583	31.1
11	dent.qu.edu.sa	5776	533	2	13	69	79	363	520	33.44
12	enuc.qu.edu.sa	5886	466	2	6	2	86	370	460	57.82
13	chr.qu.edu.sa	5546	515	2	7	32	93	381	508	43.29
14	fcohsb.qu.edu.sa	5973	526	4	20	26	87	391	506	41.91
15	qec.qu.edu.sa	5343	539	2	17	42	73	405	522	32.28
16	bcc.qu.edu.sa	8605	627	1	1	29	164	431	626	42.14
17	nursing.qu.edu.sa	9061	533	2	3	11	69	448	530	44.71
18	mduc.qu.edu.sa	9480	542	2	7	0	72	461	535	47.66
19	cos.qu.edu.sa	5346	709	2	4	83	66	554	705	56.17
20	cams1.qu.edu.sa	9635	979	2	12	108	202	655	967	58.07
21	csi.qu.edu.sa	9445	981	3	16	118	78	766	965	72.05
22	coc.qu.edu.sa	10567	979	12	9	5	131	822	970	64.87
23	asoj.qu.edu.sa	2780	383	2	8	3	78	261	344	34.88
24	uasc.qu.edu.sa	2730	425	3	3	3	72	315	292	39.77
–	Average	5567	541	3	10	30	101	397	530	41

B. Feature Extraction

A set of properties organized as logical rules discovered from ODS contents. This set consists of five mathematical rules. These rules represent a special type of advanced knowledge. It was structured based on the internal relations of hyperlinks. The formulas from 1 to 5 show this type of results. Before the extraction of the required features from the ODS contents, the set theory was used to represent the basic components, as follows.

Set of External links: $EL_i := \{EL_1, EL_2, EL_3, \ldots EL_n\}$

Set of Media:=Set of Images+Set of Videos+Set of Audios:=

$\{Im_{i+} V_i + Ao_i\}$, *Where*

$Im_{i:=}\{ Im_1, Im_2, Im_3, \ldots Im_n \}$

$V_{i:=}\{ V_1, V_2, V_3, \ldots V_n \}$

$Ao_{i:=}\{ Ao_1, Ao_2, Ao_3, \ldots Ao_n \}$

Set of Files_Doc, $f_i\_D := \{F_1, F_2, F_3, \ldots F_n\}$

Set of Other Files, $Othr_i := \{Othr_1, Othr_2, Othr_3, \ldots Othr_n\}$

Set of Leaves (Paths), $Lv_i:=\{ Lv_1\ Lv_2\ Lv_3 \ldots Lv_n\}$

The analysis of each component takes a period of time (T), the total analytical time required to analysis a sample of websites is T_i. Where, $T_{i:=}\sum(\{T_1, T_2, T_3, \ldots T_n\})$, also the web structure organized in different levels and different tracks, where, Track Length, $Tr_L:= \{Tr_1, Tr_2\ Tr_3, \ldots Tr_n\}$.

Five features are extracted, to reflect the internal behavior of websites structure. This in turn constitutes a special type of knowledge organized as a set of relations/rules, concluded based on the contents of ODS and set theory as follows:

N_ of Total_Links=Active_links+External_links (1)

Active_links =Total_No of Pages + No of Other attributes (2)

External_Links= Total_Links – Active_Links (3)

The additional features (AF) are represented in rule 4.

$$(AF) = \sum_{i=1}^{n} Audio_i + \sum_{j=1}^{k} Vedio_j + \sum_{c=1}^{l} Movies_c + \sum_{o=1}^{m} Doc\_Fs_o + \sum_{q=1}^{z} Others\_Fs_q + \sum_{r=1}^{i} \mathrm{Im}ages_r \quad (4)$$

No of Pages= Active_links –Other attributes (5)

VIII. Results Discussion and Interpretation

The analyzed attributes belong to a set of 24 homogeneous websites of academic colleges within the Qassim University websites, as presented in Table I. The figures from 3 to 6 show some of the inter-relationships in terms of the attributes that are provided for comparisons and extracted from 24 websites. Figure 3 illustrates the relation between the total links, the external links, and the active links. These details are shown in rule number 3. Figure 4 presents the distribution of the websites' components, such as number of active links, documents, images, number of pages, and other files.

In Figure 4, the values range from largest values to smallest, starting with active links and then ending with other files.

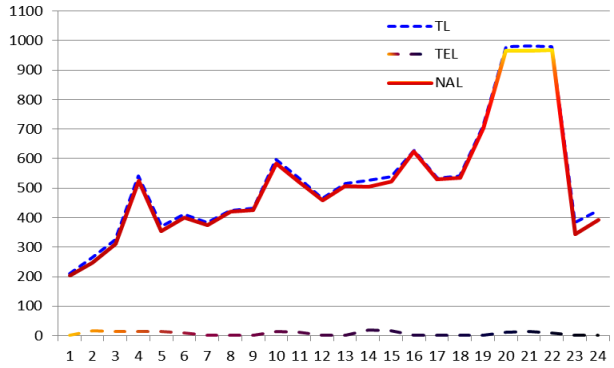

Fig. 3. Rule1 & Rule 3 (total links and external links)

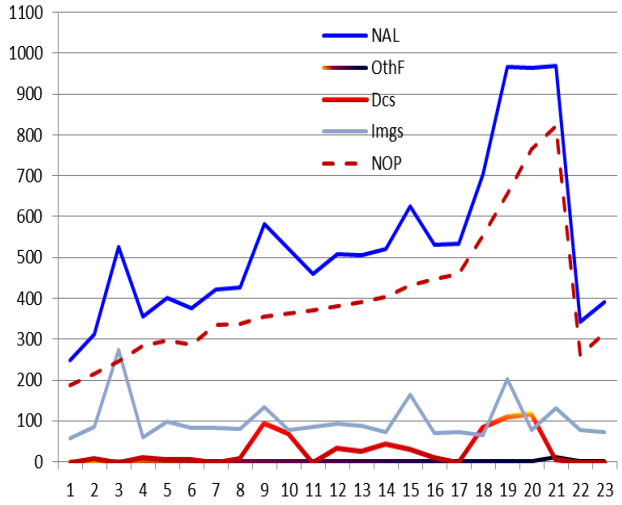

Fig. 4. Rule number 2

As well as Figure 5 represents the rule no 4 and shows the relation between three ingredients of websites structure, such as images, docs, and other files. The attributes "images" have obtained the highest level, while "other files" have obtained the lowest level in this comparison. Figure 6 represents the rule number 5. It shows the relation between the active links, documents, images, and other files.

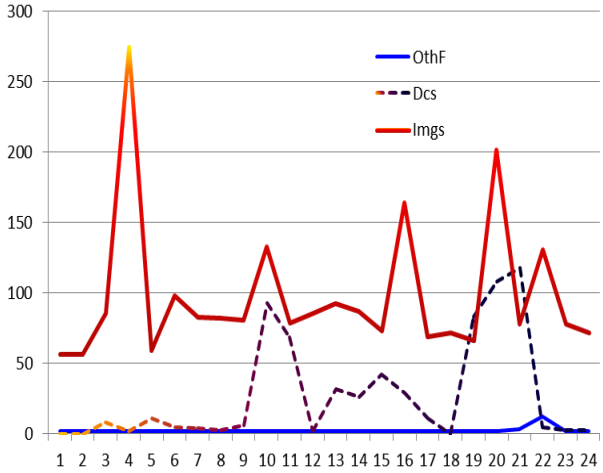

Fig. 5. Rule no.4

It found that, the number of active links is higher than the number of pages in each website structure, because active links encompass many attributes; the number of pages, docs, all types of media files, and other files. Also, it found that, some attribute values increased in some colleges and decreased in others based on the nature of the college. So, the internal relations among the website components are reflecting the development requirements of any website. The relations obtained above in the five formulas are new features that have been deduced based on the contents of ODS.

These formulas represent five association rules. These rules are reflexing the behavior of hyperlinks components and the nature of websites structure.

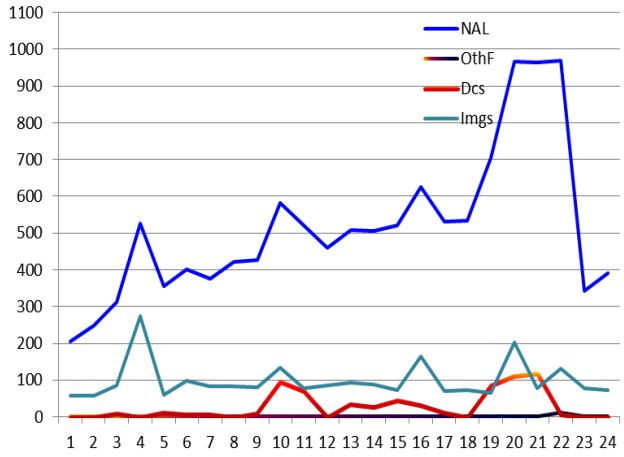

Fig. 6. Representation of Rule 5 components

These formulas can be used later for further analysis and websites estimation, such as structure size estimation.

IX. Conclusions

The research established an algorithm to produce a suitable repository of research data. The proposed algorithm can be applied for any type of websites to analyze the whole hyperlinks and extract many attributes related to the website structure and its contents. About 127991 links were analyzed in this research. It is covered about 69% of the academic colleges of Qassim University. The raw data sets prepared for scientific research purposes. In addition, the research results provide a detailed description of the internal relations of website structure components, where five rules were included in this situation based on the produced ODS. This research has achieved two objectives, based on the analysis of educational websites.

There are many benefits that can be derived from these results, including the ability of developers and users to get a comprehensive understanding of the components of the internal structure of each website and discover the complex relationships between various components. In addition, developers can discover basic relationships that can be employed in the planning and development process of new websites, as well as the results of this research helps developers to build new standards models to estimate different aspects related to the websites developments stages.

X. Future Work

Future research can include the establishment of big ODS based on the proposed algorithm and the implemented tool. These data can be provided as big repositories. These repositories can be analyzed to explore invisible features within different types of websites' structures. The results of this research will help developers in some fields such as websites measurement and estimation models, especially for the early prediction during the development life cycle, for example, these results will be applied in the measurement field, in one of specialized researches which, approved for support by the Deanship of Scientific Research in Qassim University.

References

[1] Anand D., "Improved Collaborative Filtering using Evolutionary Algorithm based Feature Extraction, International Journal of Computer Applications, vol. 64, no.20, 2013.

[2] Benslimane S., Malki M., Rahmouni M., and Rahmoun A., "Towards Ontology Extraction from Data-Intensive web sites: An HTML Forms-Based Reverse Engineering Approach," The International Arab Journal of Information Technology (IAJIT), vol. 5, no. 1, pp. 34-44, January 2008.

[3] Birla B. and Patel S., "An Implementation on web log mining," International Journal of Advanced Research in Computer Science and Software Engineering, vol.4, no. 2, pp. 68-73, 2014.

[4] Brin S. and Page L., "The anatomy of a large scale hypersexual web search engine," Computer Network and ISDN Systems, pp.107-117, 1998.

[5] Chakrabarti S., Dom E., Gibson D., Kleinberg J., Kumar R., Raghavan P., Rajagopalan S., and Tomkins A.,"Mining the link structure of the world wide web," IEEE Comput., vol.32, pp. 60-67, 1999.

[6] Chopra P. and Ataullah M., "A survey on improving the efficiency of different web structure mining algorithms," International Journal of Engineering and Advanced Technology (IJEAT), vol. 2, no.3, 2013.

[7] Dean J. and Henzinger M., "Finding related pages in the world wide web," Elsevier Science B.V, pp. 389-401, 1999.

[8] Derouiche N., Cautis B., and Abdessalem T., "Automatic extraction of structured web data with domain knowledge," IEEE 28th International Conference on Data Engineering, 2012.

[9] Getoor L., "Link Mining: A New Data Mining Challenge," SIGKDD Explorations, vol. 4, no. 2, 2003.

[10] Gibson D., Kleinberg J., and Raghavan P., "Inferring web communities from link topology," Proceeding of the of the 9th ACM Conference on hypertext and hypermedia, June 20-24, ACM Press, PA, USA, pp: 225-234, 1998.

[11] Horowitz E., Sahni S., and Rajasekaran S., "Fundamentals of Computer Algorithms," Galgotia Publications Pvt. Ltd, pp.112-118, 2008.

[12] Iraklis V., Michalis. V., Maria H., Benjamin N., and Inria F., "A closer view on web content management enhanced with link semantics," IEEE Trans, 2004.

[13] Jeyalatha S. and Vijayakumar B., "Design and implementation of a web structure mining algorithm using breadth first search strategy for academic search application," 6th International Conference on Internet Technology and Secured Transactions, Abu Dhabi, United Arab Emirates, 11-14 December 2011.

[14] Kao H. and Lin S., "A Fast PageRank Convergence Method based on the Cluster Prediction," IEEE/WIC/ACM International Conference on Web Intelligence, IEEE, Computer society, 2007.

[15] Kleinberg J., "Authoritative sources in a hyperlinked environment," Journal of ACM, vol.46, pp.604-632, 1999.

[16] Kumar R., Raghavan P., Rajagopalan S., and Tomkins T., "Trawling the web for emerging cyber-communities," IBM Almaden Research Center K53, 650 Harry Road, San Jose, CA 95120, USA, 1999.

[17] Kumar R. and Singh K., "Web structure mining: Exploring hyperlinks and Algorithms for Information Retrieval," Science Publications, American Journal of Applied Sciences, vol. 7, no. 6, pp.840-845, 2010.

[18] Mishra N., Jaiswal A., and Ambhaikar A., "An effective algorithm for web mining based on topic sensitive link analysis," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, no. 4, April 2012.

[19] Taherizadeh S. and Moghadam N., "Integrating web content mining into web usage mining for finding patterns and predicting users' Behaviors," International Journal of Information Science and Management, vol. 7, no. 1, January 2009.

[20] Xing W. and Ghorbani A., "Weighted PageRank algorithm," Proceeding of the 2nd Annual Conference on Communication Networks and Services Research, May 19-21, IEEE Computer Society, Washington DC, USA, pp.305-314, 2004.