I.J. Modern Education and Computer Science, **2015, 10, 12-19**
Published Online October 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2015.10.02

Cost Estimation of the Homogeneous Websites Using Hyper-links Analysis

Mohammed Abdullah Hassan Al-Hagery
Department of Computer Science, Qassim University, Buraydah, Qassim, KSA
E-mail: dr_alhagery@yahoo.com

Abstract—Websites hide thousands of links and sub-links. Websites' links contain a huge amount of information and knowledge. This research concentrates mainly on the difficulty of early prediction of web structure size at the beginning of Website Development Life Cycle (WDLC), especially during planning and gathering requirements. There is a lack of finding an appropriate mechanism to assist developers in these steps. The objective of this research is to measure the logical size of a website in order to predict the development time and cost earlier before the development process based on based on the website contents and its internal structure. This objective includes three sub-objectives. First, analysis of seven classes of websites to collect real data sets. Second, extracting a set of relations from the gathered data and use these relations to establish a proposed model. Third, apply the gathered data in the proposed model to predict the development time and cost of a website. This research provides strong and important results that would help developers before the development process to predict total development time and cost which in turn used directly to specify development tools, draw project plan, formulate contract conditions and determine project duration and final price.

Index Terms—Homogenous Websites, Hyper-links Analysis, Data Repositories, Web Structure Mining, Development Time, Websites Cost Estimation.

I. Introduction

Websites development becomes the main contributor to the production of data, information and as well knowledge to support decision-makers in several domains such as economic, social networks, education, etc. Currently, it has become of paramount importance with the growing need to build hundreds of thousands, as well as millions of websites. On this basis websites' developers are facing problem to find an appropriate mechanism to calculate the website logical size and determine the development time and cost earlier at the beginning of WDLC. So this research contributes to solve this problem. The objective is to estimate website size and predict the development time and cost. The research achieves this task based on an extraction of basic attributes and relations from 154 websites as real data sets. The extracted attributes provide important metrics such as link, sub-links, pages, media files, images, documents, etc.

The relations employed to create the proposed model, which is used to estimate the logical size of websites. The logical size is an essential parameter to determine the development time, cost, price, and development staff. It can be noted that this research is indirectly relevant to the topic of data mining in data stores. This research is an extension of a previous research [1].

II. Literature Review

This section covers three main parts; methods of web structure analysis, web structure mining, and web structure measures. The traditional methods used to analyze structure and components of big and complex websites need experts or professionals who know what requirements are, what size it will be to handle all components and contents. Many researchers are highlighting on the web structure analysis and structure mining. Hou and Zhang presented two hyper-link analysis-based algorithms to find relevant pages for a given web page URL. The first algorithm comes from the extended co-citation analysis of the web pages. It was intuitive and easy to implement. The second takes advantage of linear algebra theories to reveal deeper relationships among the web pages to identify relevant pages more precisely and effectively [2].

There are no researches relevant to the size estimation of websites' structure, although a number of research works have been undertaken to discuss limited subjects such as websites usability measures and interestingness measures. For example, Jalali-Heravi and Za ̈ane applied 53 different statistical interestingness measures to associate classification rules, compare it based on the number of rules and the accuracy. It was aimed to reduce the number of rules generated while not jeopardizing the accuracy of the classifier [3]. In addition, Lee and Fu, proposed a method to incorporate hierarchical characteristics of the web usage mining process. They supposed the hierarchical properties of the website are incorporated into the hierarchical folder structure of web pages. The method decreases the size of the candidate set, increasing the effectiveness of the mining process [4].

Cherifi and Santucci presented and investigated the interactions between semantic web services models from the complex network perspective. The results show parameter and operation networks exhibit core features of typical real-world complex networks such as the 'small-

world' property and an inhomogeneous degree distribution. The generated results produce valuable insight in order to develop composition search algorithms [5]. Birla and Patel proposed a method used to recognize the connection between web pages linked by information or direct link connection. This organization of data is discoverable by the condition of web structure schema through database techniques for web pages. This relationship allows a search engine to pull data concerning a search query directly to connecting web page from the website the content rests upon. The websites are an important subject for researchers as repositories of information and data, where a large amount of data is available in different formats and structures. Finding a useful data from the web is a complex task [6]. The volumes of data and information available on the internet are increasing continuously. People search for useful information on the mass are important for search engines to provide useful information to users [20]. Useful data and information can be extracted from websites by search engines. The important part of a search engine is the crawler, which is used for gathering web pages indexed by the different search engines [21].

Gupta proposed a model reflects the topology of the hyper-links underlying the website. The objective is to generate information on the similarity or the difference between websites. According to web structural data, the web structure mining can be divided into two classes: extracting patterns from hyper-links of the web and mining the document structure. In this research, we are focusing on the first class [7].

Johannes gave an overview of web mining, with a special focus on techniques that aim at exploiting the graph structure of the web without statistical details. This research identified three main areas of research within the web mining community. First, the web content mining, which is an application of data mining techniques, usually organizes HTML documents. Second, web structure mining which use the hyper-links structure of a website as an information source. Finally, the web usage mining that concentrates on the analysis of user interactions with a web server (e.g., click-stream analysis) [8].

Wookey offered the WWW as a hierarchy of web objects that can be viewed as a set of websites. Each website is a set of web pages with arcs and content elements. The research focused on the website modelled as a directed graph with web nodes and web arcs, where the web nodes correspond to HTML files with page contents, and the web arcs correspond to hyper-links interconnecting the web pages. The researcher also focused on the extraction of spanning tree for a website with the objective of connecting web contents closely that are relatively important to each other. The Page-Rank algorithm was used to measure a relative importance of web contents to other web contents [9].

Miguel and Zhiguo, concentrated on the web structure mining/link mining within this type, they introduced link mining and reviewed two popular methods applied in web structure mining and Page-Rank[10] that were also used in [11,12]. Web structure mining focuses on the hyper-links structure of each website. The different objects are linked in some way. Simply applying the traditional processes and assuming that the events are independent can lead to wrong conclusions. However, the appropriate handling of the links could lead to potential correlations, and then improve the predictive accuracy of the applied models [13].

Alqurashi and Wang introduced a graph-based methodology for web structure mining. Firstly, they mapped the structure of a website onto a graph with its nodes (web pages) and links between nodes (hyper-links between pages). The properties of the web graph, such as the degree of each node, density, connectivity, the closeness centralization, and the node clusters analyzed quantitatively. The methodology was tested on the web structural data collected from 110 UK's university websites. Based on the evaluation of the properties, some guidelines and criteria were devised for quantifying the structural quality of the webs into five categories from very poor to very good [14].

Li and Kit developed an Adaptive Window Algorithm (AWA) for discovering the navigational structure in a website. They performed several usability experiments to correlate the usability and the structural design of websites. They concentrated on the study of building a static optimal structure of a website [15].

Mishra et al., established a method to generate the structural summary about the website and web page to discover the link structure of the hyper-links at the inter-document level [16]. On the other hand, Zhou and Leung suggested a novel navigability measure MNav. The proposed measure and the experimental results show that MNav can be efficiently computed. It provides an active and convenient measurement of website navigability [17].

Weigang et al., developed a new measurement model, web-entropy based on information theory to study the influence of individuals according to different social networks. The model was tested using data from Facebook™, Twitter™, YouTube™, and Google™ search. The proposed model can be extended to other platforms [18].

In 2014, Al-Hagery suggested an algorithm for homogeneous websites analysis. The objective is to extract new attributes and to establish a data set from any website. He provided a number of suggestions for others researchers. One of the suggestions is to establish big data sets based on the proposed algorithm to discover hidden relations and to build a standard model relevant to the estimation methods of websites development process [1].

III. Problem Statement

Cost estimation is one of more challenging requirements of project management procedures. Basically, it is a prediction methodology towards fine tuning the cost estimates for a successful conclusion of a project [22]. There are many difficulties are facing

websites' developer at the beginning of WDLC, especially during the early steps, which focus on requirements gathering process. These difficulties appear clearly for websites that have complex structure, huge amount of contents, and big size, exceeding thousands of pages or more. In this case, websites development process costs a lot because websites' analyst spends more time and money to accomplish analysis task based on several meetings and discussions with stakeholders directly or indirectly, based on several methods of requirements gathering, then developers can determine web requirements and start to design the structure. This method needs an expert who knows requirements, he can guess the size, the number of working days and the project cost required for the development. So, the developer will do these tasks manually according to their experience rather than using an estimation model.

IV. Objectives

The objective of this research is to propose an estimation model to predict the logical size of websites earlier at the beginning of the development processes, based on its contents and internal structure. It includes three sub-objectives. First, analysis of different classes of websites to extract real data sets (seven samples, each sample contains 22 websites). Second, analyse the collected data then discover main relations amongst the attributes of the extracted data sets. Use these relations to build the proposed model. Third, apply the extracted data in this model.

For a new website, which is similar to any element included in these samples, the developers can use the outputs of this model to estimate the total development time of this website at the early stage of development. The developers can get many indicators to help estimate the project duration, its cost, price, help to write project contract conditions, etc.

V. Research Methodology

This approach involves many steps; first, apply the website analysis tool. Second, analyse of seven samples of homogeneous websites.

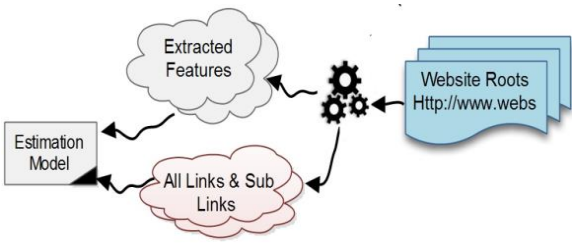

Fig.1. Methodology Headlines

Third, discover the hyper-links connections and components with all attribute values. Fourth, establish data sets from these samples to extract a set of important relations. Fifth, construct and apply the proposed model. Finally, apply these data in the proposed model, analyse the results, and estimate the required values to help websites developers as decision-makers to reduce the development cost. Fig.1 illustrates general layouts of the methodology, whereas Fig.2 presents steps details of the methodology.

It presents the general steps of following-up website links by the analysis tool, which analyse websites structure and contents. It illustrates processes of data sets extraction, relations construction, establishment, application of the proposed model, and generation of the final results.

A. Creation of the Estimation Model

A number of attributes extracted as a result of the analysis process. These attributes include; Total Links (TL), Number of Other Attributes (NOF), Number of Active Links (NAL), Total External Links (TEL), and Number of Pages (NOP). These attributes are incorporated in the formulas from (1 to 5). These formulas employed in this research as basic components to establish Websites Structures Size Measure (WSSM) as a proposed model to measure the logical size of websites' structures, as shown in formula (6).

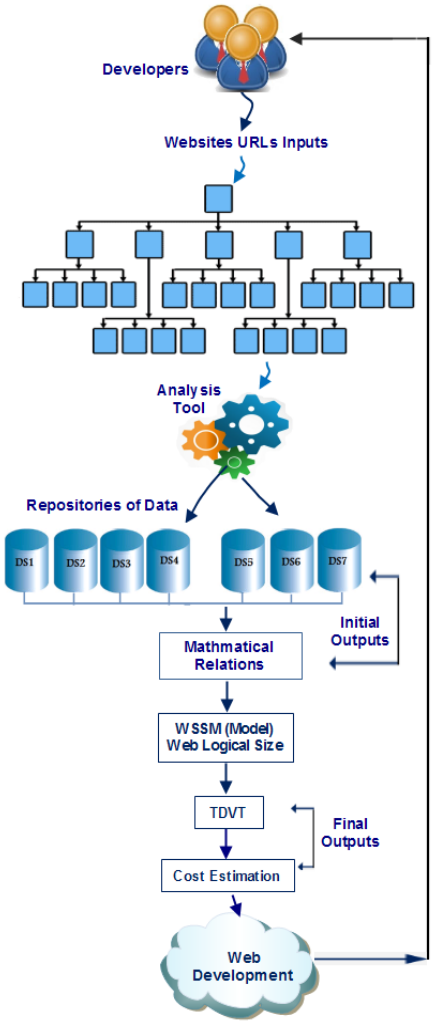

Fig.2. Methodology Details

Total Links (TL) = Number of Active Links (NAL) + Total External Links (TEL) (1)

$$NOF=\sum_{a=1}^{an} Audio_a +\sum_{v=1}^{vn} Video_v +\sum_{m=1}^{mn} Movies_m + \sum_{d=1}^{dn} Doc\_Files_d +\sum_{i=1}^{in} Images_i \quad (2)$$

Where, An, Vn, Mn, Dn, and In ≥ 1.

$$NAL = NOP + NOF \quad (3)$$

$$TEL = TL - NAL \quad (4)$$

$$NOP = NAL - NOF \quad (5)$$

$$WSSM = TL + (TEL \times \frac{1}{2} \times \sqrt[2]{TS}) - (NOF + \sqrt[2]{\frac{NOP + T\_N\_Analz\_Links}{NAL}}) \quad (6)$$

The model output is measured in pages and its inputs include TL, TEL, Total Time Analysis (TTA) in seconds, NOF, NOP, Total Number of Analysed Links (T_N_Analz_Links), and NAL. The model calculates the results based on these inputs.

Formula (7) is used to calculate the Total Development Time (TDVT), it includes two components; the first is the result of WSSM model and the second, is the Average Development Time of a Page (ADTP). The value of ADTP is taken from 1. It equals 20.8 working hours, according to the real results of several categories of web pages calculated in [19]. It is used as a factor in this formula to calculate the TDVT.

$$Where, TDVT = WSSM \times ADTP \quad (7)$$

Table 1. Web Page Development Time Ranges in Days and Hours [19]

Page	Category	Normal/day	High/day	AV/H
About Us	Static	0.25	0.50	3
Login	Easy	1	1	8
Home Page	Medium	3	4	28
Detailed Usage Report	Hard	5	6	44
Average	Normal	2.3	2.9	20.8

VI. Research Results

The research results comprise two types; initial results and the final results. These types are discussed with more details in the following two subsections.

A. Initial Results

The initial results are the outputs of analysis processes, it includes two classes; the first is five mathematical formulas. The second class includes seven data sets, which organized as small repositories of data; DS1, DS2, …, and DS7. The seven samples belonging to different organizations such as Hospitals, Academic Colleges, Institutes, Schools, Hotels, Tourism Companies (TC), and News Agencies (NA). Each sample contains 22 websites. The total number of websites analysed in this research was 7 × 22 = 154 websites. The total analysed links in this research for all samples was 588434 links. All samples were selected to cover the whole community contents of each sample. Each sample includes the following attributes; Site Root, T_Analinks, TL, TEL, NAL, NOP, Images, Document Files (Doc_Files), OthF, etc. Table 2 presents the average values of all attributes in the seven samples, including the average analytical time of each sample estimated in seconds.

Table 2. The Average Values of Analysis of the 7 Samples

Data sets	Sample Type	OthF	TEL	A_Ana Links	TL	Doc_ Files	Images	NOP	NAL	TTA/S
DS1	Hospitals	8.6	27	7736	657	62	73	563	707	73.2
DS2	Colleges	3	19	5567	533	30	84	382	499.1	41
DS3	Institutes	2.42	9	4770	477	31	42	358	433	39
DS4	Schools	2	11	4110	497	27	37	352	418	34.4
DS5	Hotels	1.9	24	2790	302	21	33	280	336	24
DS6	TC	1.3	9	1020	250	7.3	76	163	248	12
DS7	NA	3	8	754	153	3.2	31	110	145	4.3

After websites, analysis, what was found was that the number of NAL is higher than the number of NOP for each website because the number of NAL encompasses many other attributes including the number of NOP, Doc_Files, all types of media, etc. Fig.3 presents a comparison between the two attributes as a part of the

results of sample 2. Fig. 4 shows five attributes of websites in each sample. These attributes include the number of NAL, NOP, TL, TEL, and the time analysis. B. Final Results.

The final results are the outputs of the WSSM model which estimates the total number of pages of a website (logical size estimated in pages). The estimated results compared with the real results, which collected in [19]. The outputs of WSSM used as primary inputs to calculate the TDVT of a website estimated in working hours, assuming that the number of working hours per person is eight hour per day. Table 3 presents the results of sample 2 and Table 4 clarify the average values of the estimation results and the real results for all samples.

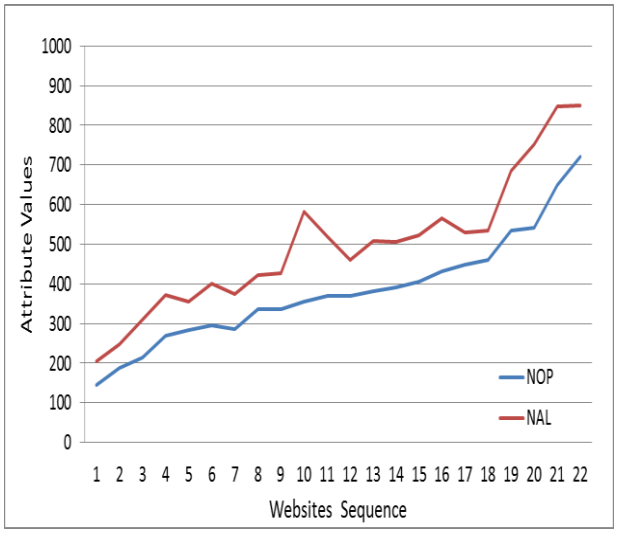

Fig.3. Comparison of NAL & NOP in Sample 2

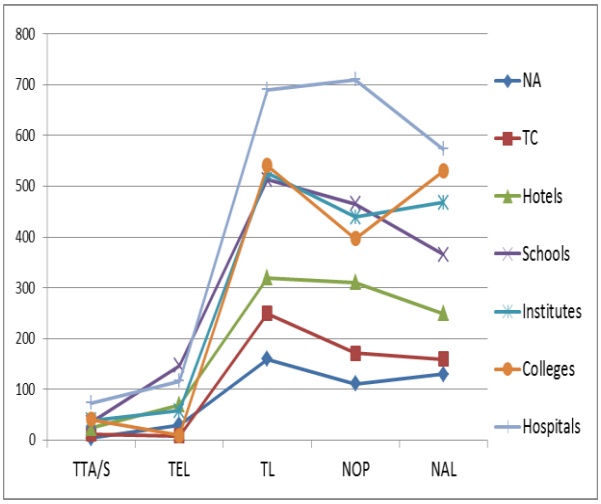

Fig.4. Five Attributes in All Samples

VII. Results Discussion and Interpretation

As shown in the results above, the Hospitals and Colleges have the highest values, specifically for the TL, NOP and NOL.This in turn gives an initial view of their general structure, in addition. The time analysis values for instance in sample 2, were located between 10 and 72 seconds, and the average value of time analysis of this sample was 41 seconds.

Whenever this value of a specific website is high, it means this website is bigger in size than the other websites which have the lower value. Results of all samples are ranging from the largest to the smallest, and most of them show strong indications and significant differences among all attributes which used as metrics in the estimation process. The results reflect a positive difference between all values of the data sets. Each sample has results appear in a determined level. This distribution is attributable to the presence of real differences in the architecture and the logical size of each sample when compared with the other samples based on the development time and other requirements.

When interpreting the results in Table 4, it found that these results are supporting the outputs of the proposed model. For instance, "Hospitals" sample got an average of real size = 563 pages and average value of real development Time = 11710.4 hours, whereas the average of estimated size by the proposed model = 650 pages with average value of estimation of the development time = 13509 hours. The percentage of error equal to 15% hours and the prediction accuracy = 85% for this sample. This sample has the highest values, the biggest size and the highest number of working hours if compared with other six samples. The generalisation of results covers the other websites that are similar and not included in these samples that relevant to the whole community of the same type (homogeneous websites).

The result of the first sample shows that any website is similar or found in this sample requires nearly an average of 13509 hours for development. For instance, if developers plan to develop a Hospital website as a project and they have a development team consists of five persons, for example, two analysts, one designer and two programmers. Based on the research results, the number of months required to development this project can be calculated as follows:

$$\text{Project Development Time} = (\text{TDVTES}/(\text{Hours\_Per\_Day} \times \text{No\_of Persons}))/\text{Dyes\_Per-Month}.$$
$$= (13509 / (8 \times 5)) / 30$$
$$= 11.25 \text{ months}.$$

On these bases, it is easy to estimate the total development cost of this project early. In the same method, the average values of real sizes of the next six samples are ranging in the following sequence; 382, 358, 352, 280, 163, and 110 pages, versus the values of 476, 426, 460, 302, 179, and 122 pages as an average of estimated sizes, respectively.

Fig. 5, illustrates the real values of TDVT versus the estimated values for the sample no.2. This Figure presents the results of one sample (Colleges) while Fig. 6 illustrates the results average of the seven samples, including the real values and the estimated values of the total development time. The error rate of the estimation process is low as shown in Fig. 5 and Fig. 6.

Table 3. Estimation Results of Sample 2

I	NOP (Real)	WSSM (Estimated)	TDVT R_Values/h	TDVT ES_Values/h	Error Rate/h
1	145	145.9	3016	3034.4	18
2	188	231.2	3910.4	4808.9	899
3	215	259.4	4472	5396.3	924
4	270	290.3	5616	6037.5	422
5	284	295.5	5907.2	6146.5	239
6	296	339.5	6156.8	7060.8	904
7	286	262.2	5948.8	5453.5	495
8	335	438.7	6968	9124.1	2156
9	337	403.2	7009.6	8387.1	1378
10	355	368.2	7384	7657.9	274
11	370	378.2	7696	7867.3	171
12	370	395.1	7696	8218.6	523
13	381	473.4	7924.8	9846.9	1922
14	391	472.2	8132.8	9821.6	1689
15	405	467	8424	9713.1	1289
16	431	652.1	8964.8	13563.2	4598
17	448	603.7	9318.4	12556.7	3238
18	461	670.3	9588.8	13942.5	4354
19	534	700.5	11107.2	14570.6	3463
20	541	783	11252.8	16285.5	5033
21	649	759.3	13499.2	15793.1	2294
22	722	1073.4	15017.6	22326.4	7309
AV	382	476	7955	9891	1981

Table 4. The Results of the Proposed Model

I	Sample	NOP (Real)	WSSEM/p (Estimated)	TDVT R_Values	TDVT ES_Values	Error Rate%/h	Prediction Accuracy%
1	Hospitals	563	650	11710.4	13509	0.15	0.85
2	Colleges	382	476	7955	9892	0.24	0.76
3	Institutes	358	426.241	7446.4	8866	0.19	0.81
4	Schools	352	459.991	7321.6	9568	0.31	0.69
5	Hotels	280	301.865	5824	6279	0.08	0.92
6	TC	163	178.804	3390.4	3719	0.1	0.9
7	NA	110	121.654	2288	2530	0.11	0.89
Average						0.16	0.84

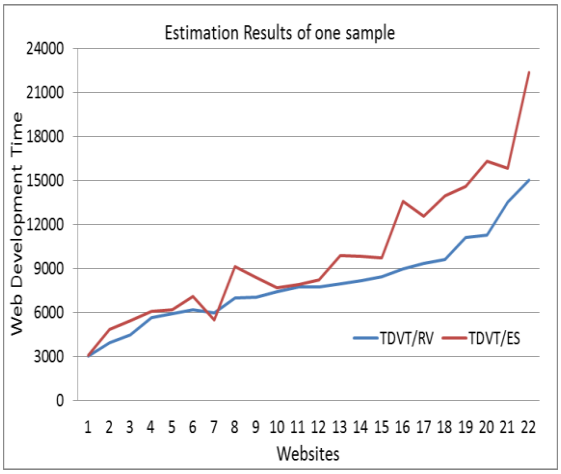

Fig.5. Final Results for Sample 2

Based on the example given above for the first sample "Hospitals", the results can be generalized for the Community of Hospitals and also by the same method it can be generalized for other samples.

The model results are very strong and support developers as decision-makers to apply the WSSEM as an estimation model early at the development process for any website.

The estimation results which calculated based on this model are accurate when compared with the real results. The error and the prediction accuracy rate for all samples are shown in Table 4, and the average rate of the prediction accuracy for all samples reached to 84% versus 16% as an average of error.

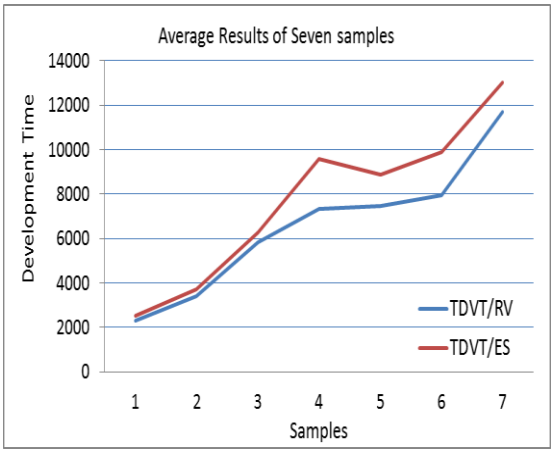

Fig.6. Final Results of WSSM against the Real Values

VIII. Conclusions

The research achieved the planned objectives and provides strong indications with a high degree of importance based on the final results, where it found that the error rate of the estimation is low. So, the developer can use the model and its results with high confidence because, all applied samples are representing their communities.

For example, results of Schools sample can be generalized to all School websites, with a limited percentage of error, as well a the results of other groups, are interpreted and generalized by the same method. If another website is required to be developed as a different type does not belong to any type in the seven samples analysed in this research (for instance, website of a bank), in this case developers can't use these results, but the solution is to apply the research steps for that sample. The results of each sample can be generalized to support websites' developers. It provides a clear view of website components, its structure and predicts important requirements which are necessary for a website development tasks.

As discussed above, the results give a good percentage of prediction accuracy for the basic requirements, such as development time estimated in working hours, the total cost, which can be used to identify the required tools, help to formulate the project plan, and to find exact view of the web architectures, and other restrictions emergent from the cost that should be incorporated into the project contract. However, all results of the proposed model located above the real level line, and from this perspective, the proposed model needs further improvement and adjustment to reduce the error rate and to increase the results accuracy, then the results will be better than current results.

IX. Future Work

After the completion of this research, many ideas and perspectives have emerged to enable researchers and people interested in this area to continue development. Some of these perspectives include the following points:

- Establishment of big repositories of data sets to be employed in the exploration process to discover additional features and hidden patterns related to websites behaviour, structure, and contents.
- The proposed mechanism can be employed to study the relation between universities, in other words to analyse the websites of the top 100 university in the world to discover whether if there is a relationship between this rank and the logical size of this website.
- Based on the methodology of this research, researchers can accomplish and enhance more solutions associated with data mining and knowledge discovery in websites' structure, behaviour, nature of design, design optimization, discovery of redundant components, use classification and/or clustering based on the structure similarity or web behaviour, etc.
- The proposed model of this research needs further adjustment to reduce the error rate and to increase the estimation accuracy.
- Additional attributes can be extracted to include more than the current, this helps to modify the current analysis tool.
- Finally, the applied tool can be compared to other analysis tools/crawlers in the analysis of websites based on novel criteria.

References

[1] M.A. Al-Hagery, “Data and knowledge extraction based on structure analysis of homogeneous websites”, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 5, No. 11, 2014, pp.56-62. http://thesai.org/Publications/ViewPaper?Volume=5&Issue=11&Code=IJACSA&SerialNo=10, (DOI): 10.14569/IJACSA.2014.051110.

[2] J. Hou, and Y. Zhang, “Effectively finding relevant web pages from linkage information”, IEEE Transactions on Knowledge and Data Engineering, 2003, Vol. 15 No. 4, pp.940-951. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1209010,DOI:10.1109/TKDE.2003.1209010

[3] M. Jalali-Heravi, and O.R. Za ̈ane, “A study on interestingness measures for associative classifiers”, Proceedings of the ACM Symposium on Applied Computing, 2010, pp.1039-1046. http://dl.acm.org/citation.cfm?id=1774306,DOI:10.1145/1774088.1774306

[4] C. Lee, Y. Lo, and Y. Fu, “A novel prediction model based on hierarchical characteristic of website”, International Journal of Expert Systems with Applications, Elsevier, Vol. 38 No. 4, 2011, pp. 3422-3430. http://www.sciencedirect.com/science/article/pii/S095741741000936X, DOI:10.1016/j.eswa.2010.08.128

[5] C. Cherifi, and J. Santucci, “On topological structure of web services networks for composition”, International Journal of Web Engineering and Technology, Vol. 8 No. 3, 2013, pp.291-321. http://arxiv.org/ftp/arxiv/papers/1305/1305.0467.pdf

[6] B. Birla, and S. Patel, “An implementation on web log mining”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4 No. 2, 2014, pp. 68-73.

http://www.ijarcsse.com/docs/papers/Volume_4/2_February2014/V4I2-0138.pdf.

[7] R. Gupta, "Web mining using artificial ant colonies: a survey", International Journal of Computer Trends and Technology (IJCTT), Vol. 10. No 1, 2014, pp. 12-17. http://arxiv.org/ftp/arxiv/papers/1404/1404.4139.pdf, DOI:10.14445/22312803/IJCTT-V10P103

[8] J. Fürnkranz, "Web structure mining exploiting the graph structure of the world-wide web", Austrian Research Institute for Artificial Intelligence, 2002, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.9645.

[9] L. Wookey, "Hierarchical web structure mining", Proceedings of Data Engineering, World Wide Web Conference (DEWS), Sungkyul University, Korea, 2006. http://www.ieice.org/~de/DEWS/DEWS2006/doc/2A-v1.pdf

[10] G.D.C. Miguel, and G. Zhiguo, "Web structure mining: an introduction", Proceedings of the IEEE, International Conference on Information Acquisition, June 27-July 3, Hong Kong and Macau, China, 2005, pp.590-595. http://fumblog.um.ac.ir/gallery/429/01635156.pdf,DOI:10.1109/ICIA.2005.1635156

[11] J.M. Kleinberg, "Authoritative sources in a hyper-linked environment", Journal of the ACM, Vol. 46 No. 5, 1999, pp. 604-632, New York, USA. http://dl.acm.org/citation.cfm?id=324140

[12] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: bring order to the web", Technical report, Stanford University, 1999. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.1768

[13] L. Getoor, "Link Mining: A new data mining challenge", ACM SIGKDD, New York, USA, Vol. 5 No 1, 2003, pp. 84-89. http://dl.acm.org/citation.cfm?id=959242.959253

[14] T. Alqurashi, and W, Wang, "A Graph-based methodology for web structure mining - with a case study on the webs of UK universities", Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14), ACM, New York, USA, 2014. http://dx.doi.org/10.1145/2611040.2611058

[15] C. Li, and C. Kit, "Web structure mining for usability analysis", Proceedings of the IEEE/WIC/ACM, International Conference on Web Intelligence, 2005. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1517862 .

[16] S.N. Mishra, A. Jaiswal, and A. Ambhaikar, "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, 2012, pp. 278–282, http://www.ijarcsse.com/docs/papers/April2012/Volume_2_issue_4/V2I400148.pdf

[17] Y. Zhou, and H. Leung, "MNav: a markov model-based website navigability measure", IEEE Transactions on Software Engineering, 2007, Vol. 33 No. 12, pp.869-890. http://dl.acm.org/citation.cfm?id=1314044

[18] L. Weigang, Z. Jianya, and G. Liu, "W-entropy method to measure the influence of the members from social networks", International Journal of Web Engineering and Technology, Vol. 8 No. 4, 2013, pp.369-394. http://inderscience.metapress.com/content/l0h78t73j4x054u8/

[19] J. Bicer, "Software estimation: estimating the duration of web-based software development", Technical report, 2009. http://www.septium.com/Estimation-Tips.pdf

[20] M. S. Iraji, H. Maghamnia, and M. Iraji, "Web Pages Retrieval with Adaptive Neuro Fuzzy System based on Content and Structure", International Journal of Modern Education and Computer Science, 2015, Vol. 8, pp.69-84, http://www.mecs-press.org/),DOI: 10.5815/ijmecs.2015.08.08

[21] M. A. Kausar, V. S. Dhaka, and S. K. Singh, "Implementation of Parallel Web Crawler through .NET Technology", International Journal of Modern Education and Computer Science, 2014, Vol. 8, pp.59-65, 2014. http://www.mecs-press.org/), DOI: 10.5815/ijmecs.2014.08.07

[22] M. M. Al_Qmase, and M. R. J. Qureshi, "Evaluation of the Cost Estimation Models: Case Study of Task Manager Application", International Journal of Modern Education and Computer Science, Vol. 8, pp.1-7, 2014, (http://www.mecs-press.org/), DOI: 10.5815/ijmecs.2013.08.01

Authors' Profiles

Mohammed A. H. Al-Hagery received his B.Sc in Computer Science from the University of Technology in Baghdad Iraq-1994. He got his MSc in Computer Science from the University of Science and Technology Yemen-1998. Al-Hagery finished his Ph.D. In Computer Science, (Software Engineering) from the Faculty of Computer Science and IT, University of Putra Malaysia (UPM)- 2004. He was a pointed a head of Computer Science Department at the Faculty of Science and Engineering, USTY, Sana'a from 2004 to 2007. From 2007 to this date, he is a staff member at the Faculty of Computer, Qassim University in KSA. He published more than 13 papers in international journals. Dr. Al-Hagery was appointed a head for Research Centre at the Computer College, Qassim University, KSA from September 2012 to this date.