Google pagerank algorithm: using efficient damping factor

Ali Ali Saber[1], Aso Kamaran Omer[2], Noor Kaylan Hamid[1]
[1]Department of Computer Engineering Techniques, College of Engineering Technology, AlKitab University, Kirkuk, Iraq
[2]Department of Administration and Accounting, Faculty of Humanities and Social Sciences, Koya University, Erbil, Iraq

Article Info

Article history:

Received Feb 11, 2022
Revised Aug 24, 2022
Accepted Sep 8, 2022

Keywords:

Damping factor
Dangling node
Link spam
PageRank
PageRank algorithm

ABSTRACT

A vital feature of modern web search engine is the ability to display relevant and reputable pages near the top of the list of query results. A well-used search engine nowadays is Google search engine, it is the world's most popular search engine, rely on PageRank technology to determine a website's ranking. We put our attention on important benefactions to improving the quality of rankings via the value which is called damping factor, commonly the original suggestion d=0.85 by Brin and Page is the most common choice. In this paper, we suggest a new value which plays an important role to rank web sites accurately, our work focuses on damping factor value which improves the efficiency of PageRank value for each website. Our results show that the suggested value can get greater performance. Finally, we will show satisfactory result without link spam and dangling node applying PageRank algorithm on graphs with over 5000 links.

Corresponding Author:

Ali Ali Saber
Department of Computer Engineering Techniques, College of Engineering Technology
Al-Kitab University
Altun Kopri-Kirkuk, Iraq
Email: Ali.a.saber@uoalkitab.edu.iq

1. INTRODUCTION

The age that we are living in, which is the age of informati on by many thinkers and scientists. This is for sure growing especially because now, almost everyone has the possibility to generate and share content with others [1]. Modern web search engines can display relevant, popular, and accurate pages towards the top of the query results list. The PageRank algorithm to web pages, it is a measure of the importance of these pages. PageRank technology is used to look at the full Internet's link structure and find the most essential pages. Created by Google co-founders Sergey Brin and Lary Page in the late 1990s, provides useful measure of relative importance of every web page in the network, called the PageRank. The PageRank algorithm is used by Google, the world's largest and most well-known search engine, to determine website ranking, because of Google is distinctive in its concentrate on developing the ideal search engine that knows directly and accurately what users mean or what they search for then gives them the most related page with desired information [2].

We will concentrate on damping factor in this study: we make an effort to shed light on the new value for damping factor while PageRank changes significantly when d is modified [3]. In this paper we suggest (d)=0.90 by appling algorithm to calculate PageRank value which gives more accurate PageRank for each webpage, gives final stable values and users will find this option to be quite natural and satisfying. The behavior of PageRank in relation to changes in d, known as the damping factor, has been proven to be beneficial in the detection of link spam [4]. Spamming is a way used to manipulate or affect search engine indexes. When using the PageRank algorithm by some web pages or sites, then they spamming so that to boost up the rank of their certain pages [5]. When a web page receives in- link from another web page which has higher PageRank value then the PageRank of current page will also increase [6]. So, they try to make use of such links to increase the PageRank value of their pages or make it important page as well. The selection of efficient value of d is important, and in most cases, the suggestion of d=0.85 by Brin and Page is used [7]. To the best of our knowledge, no study has been done to verify that d=0.85 is the best answer. Therefore sum of all page rank values should be ideally equal to one or real close to one [8].

This paper is structured: we have two aims firstly is calculating PageRank value of each page by taking into account dangling nodes. Finally, determining the twenty-five most popular pages. Much more of research on damping factor value d is motivated by the idea which calculates the value for webpages and most of researches that have done before on PageRank used the value d=0.85 because it is default value by Google's founder Sergey Brin and Larry Page [9].

In 1998, Google tested the PageRank algorithm on a network of 24 million sites.with default damping factor value 0.85 now twenty years later the size of the web has expanded in size, as a web continues to grow challenge arises in the computation of PageRank and selecting an accurate damping factor value, theses challenges caused to improving computer technology [10]. Our proposed approach using d=0.90 can be used to calculate the PageRank of all pages significantly and we will compare the results using the default value for d with our proposed value on dataset.

2. PROPOSED METHOD

While conducting random surfing on the internet, we will encounter two problems. First, we will discover keeping come back the same cycle or loop of node and get stuck in a loop. Next problem that we will face when arriving at a dangling node. Dangling node is a node without out links does not link to any other page, and this will get us to stuck [11]. Aiming to solve cyclic and dangling node problems, a new concept of damping factor, constant d as described by Page and Brin. The web-graph component of the process is obliterated when d is 0, resulting in a trivial uniform process. As d goes to 1, the web component becomes increasingly important [12].

In terms of our research, we used a six-step procedure to obtain the desired results:

a. List of Top 25 URLs along with their page rank values and in degree–out degree of each web page.
b. Sum of all page rank values (Ideally equal to one or close to one).
c. Number of iterations occurred when executing PageRank algorithm on the data.
d. From the input data, generate web pages for websites.
e. Calculate the in-link and out-link weights.
f. Iterate the processes until important information from the web sites is achieved.

In the suggested algorithm, we use the notion of traditional PageRank algorithm and apply it to a variety of web sites [13]. It assists users in gaining an understanding of the website's only significant and valuable web pages. As a result, the user may quickly choose only the most useful web sites [14].

3. METHOD

We propose the traditional PageRank algorithm to deal with our dataset for calculate PageRank value which depends on our suggested damping factor value d=0.90 to handling dangling nodes and link spam as shown in:

$$\mathrm{PRk} + 1(\mathrm{u}) = (1 - \mathrm{d})/\mathrm{n} \ + \mathrm{d} \ \sum\nolimits_{\mathrm{v} \in \mathrm{B(u)}} \frac{\mathrm{PRk(v)}}{\mathrm{L(v)}} \tag{1}$$

where we have used:

$$\mathrm{PR0(v)} = \sum\nolimits_{1-\mathrm{d}\ ,\mathrm{v} \in \mathrm{B(u)}=0}^{1\ ,\mathrm{v} \in \mathrm{B(u)} \neq 0}$$

in this algorithm the convergence speed of the method depends on a scalar parameter d, calculations are done iteratively until every PageRank values converge or until the specified or proposed number of iterations are executed [15]. At the start of the loop i=0, by sitting initial PageRank value PR0(v)=1 and PR0(v)=1/n where n is number of total web pages. If there are none in the node's links as in (1). The reason for setting PR0(v)=1 for any other nodes with in links is on assumption that the web pages has at most one in links [16]. The primary reason for setting PR0(v)=(1-d)/n for node without any in links, because the value will not change after the first iteration of the procedure, links will assist PageRank computation in meeting the convergence faster [17]. We decided to use the Page Rank algorithm that we have applied on input file mathworks 100 mat that contains 100 pages and more than 5000 edges along with the list of pages to which they are linked.

This data was generated in 2015 at mathwork site. We investigate which webpage is most important among the all web pages in the website using the page rank algorithm depending on the damping factor value is equal to 0.90. In particular, the significance of PageRank when d approaches 1 this effect has been illustrated in Figure 1.

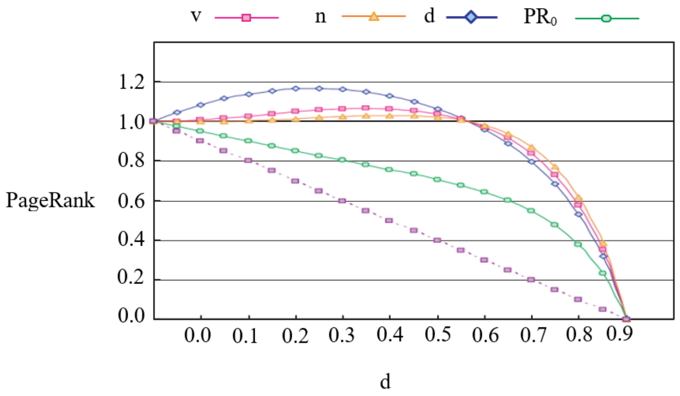

Figure 1. Ranking under traditional PageRank algorithm

Simultaneous as shown in (1) can be used to compute the PageRank of each page, Figure 1 shows the relationship between PageRank and damping factor d, indicating that lines v and d cross line n. The damping factor is clearly the most important component in changing the page ranking. Values of d=0.90 can be adopted to calculate the PageRank of all page [18].

4. RESULTS AND DISCUSSION

We will show the results when d=0.85 as indicated in our literature review. Next, we will show the results and compare with our proposed value d=0.90. We discovered throughout our literature study that the value d=0.85 as shown on Table 1 presents unsatisfied ranking of websites in mathworks 100 mat.

Table 1. PageRank values using d=0.85

Rank	URLs	In- degree	Out-degree	PageRank
1	www.mathworks.com	20	14	0.044342
2	ch.mathworks.com	20	14	0.043085
3	cn.mathworks.com	20	14	0.043085
4	jp.mathworks.com	20	14	0.043085
5	kr.mathworks.com	20	14	0.043085
6	uk.mathworks.com	20	14	0.043085
7	au.mathworks.com	20	14	0.043085
8	de.mathworks.com	20	14	0.043085
9	es.mathworks.com	20	14	0.043085
10	fr.mathworks.com	20	14	0.043085
11	in.mathworks.com	20	14	0.043085
12	it.mathworks.com	20	14	0.043085
13	nl.mathworks.com	20	14	0.043085
14	se.mathworks.com	20	14	0.043085
15	mathworks.com/index.html%3Fnocookie%3Dtrue	0	1	0.0015
16	mathworks.com/company/aboutus/policies_statements/patents.html	6	6	0.007714
17	mathworks.com/company/aboutus/policies_statements/trademarks.html	6	6	0.007714
18	mathworks.com/company/aboutus/policies_statements	5	6	0.006439
19	mathworks.com/company/ piracy.html	5	6	0.006439
20	mathworks.com /rss/index.html	5	6	0.006439
21	ch.mathworks.com/index.html%3Fnocookie%3Dtrue	0	1	0.0015
22	ch.mathworks.com/company /patents.html	5	6	0.0051817
23	ch.mathworks.com /trademarks.html	5	6	0.0051817
24	ch.mathworks.com/company/aboutus/policies_statements.html	5	6	0.0051817
25	ch.mathworks.com/company /piracy.html	5	6	0.0051817

Sum of PageRank values for 25 webpages calculated (0.6629188), it seems that the ranking obtained with this choice (d=0.85) are not satisfactory compared to our choice [19]. In our research following the data analysis for each website, we calculated page rank for each website. Table 2 presents the list of top 25 URLs along with their page rank value, in degree and out degree.

Table 2. PageRank values using d=0.90

Rank	URLs	In- degree	Out-degree	PageRank
1	www.mathworks.com	20	14	0.050561
2	ch.mathworks.com	20	14	0.049459
3	cn.mathworks.com	20	14	0.049459
4	jp.mathworks.com	20	14	0.049459
5	kr.mathworks.com	20	14	0.049459
6	uk.mathworks.com	20	14	0.049459
7	au.mathworks.com	20	14	0.049459
8	de.mathworks.com	20	14	0.049459
9	es.mathworks.com	20	14	0.049459
10	fr.mathworks.com	20	14	0.049459
11	in.mathworks.com	20	14	0.049459
12	it.mathworks.com	20	14	0.049459
13	nl.mathworks.com	20	14	0.049459
14	se.mathworks.com	20	14	0.049459
15	mathworks.com/index.html%3Fnocookie%3Dtrue	0	1	0.001
16	mathworks.com/company/aboutus/policies_statements/patents.html	6	6	0.0060468
17	mathworks.com/company/aboutus/policies_statements/trademarks.html	6	6	0.0060468
18	mathworks.com/company/aboutus/policies_statements	5	6	0.0051468
19	mathworks.com/company/ piracy.html	5	6	0.0051468
20	mathworks.com /rss/index.html	5	6	0.0051468
21	ch.mathworks.com/index.html%3Fnocookie%3Dtrue	0	1	0.001
22	ch.mathworks.com/company /patents.html	5	6	0.0040451
23	ch.mathworks.com /trademarks.html	5	6	0.0040451
24	ch.mathworks.com/company/aboutus/policies_statements.html	5	6	0.0040451
25	ch.mathworks.com/company /piracy.html	5	6	0.0040451

URL length: the URL is the website's address on the World Wide Web. PageRank values were calculated for only 219 numbers of iterations which we used maximum number until convergence occurred for each web page and stable values. Table 3 shows maximum number of iteration used in this work which it needs for each damping factor value, the higher damping factor value needs higher number of iterations [20].

Table 3. Effect of d on expected number of power iterations

d	Maximum number of iterations
0.90	219

a. Following graph shows websites linked to' https://mathworks.com 'with their PageRank values for each web page and number of in-degree and out-degree as shown in Figure 2.
b. In the Figure 2 there are 14 nodes in yellow color which they have the highest PageRank value (0.04 to 0.05) which means they are the most visited or the important web pages.
c. We ran this algorithm on input file of 25 URLs (N=25). Program ran successfully gives satisfying set of results. Sum of all page rank values evaluated (0.7392424) is so close to 1 while using d=0.85 the sum of all page was (0.6629188) which is less than our result and indeed it is not the optimal solution.
d. The importance of the page is determined by the proportion or ratio of time spent by the surfer on that page. A random surfer will visit 'http://www.mathworks.com' almost 5% of the time.
e. The ranking gained with this choice (d=0.90) are very natural and accurate.
f. The graph is column-stochastic because all the PageRank values are non-negative and the sum of all values is really close to 1.
g. For damping factor, the value of d =0.9, program converged in 200 maximum iterations.
h. The length of time a random surfer spends on a specific page, such as 'http://www.mathworks.com' is a measure of the relative or proportional value of that page; he or she spends a significant amount of time on it, thus it must be important. Because other significant pages must refer to them.
i. Some website having out links but no in links, has less weight (0.001) under PageRank algorithm. In reality, few websites can have out links without in links.

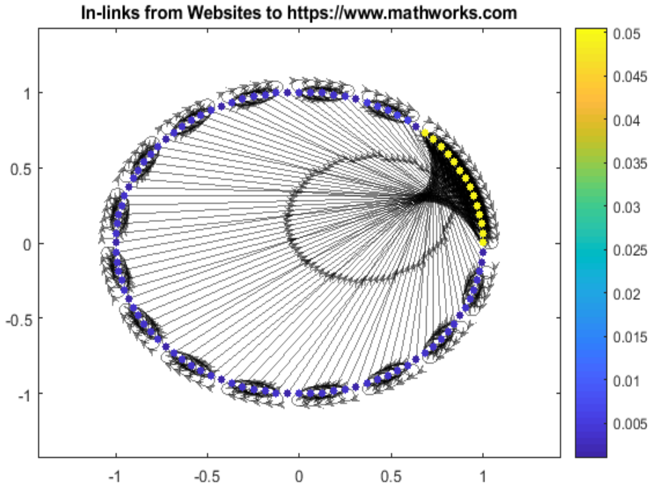

Figure 2. PageRank values

4.1. Comparision

In this section we discuss how d=0.85 and d=0.90 differ from each other by using traditional google PageRank algorithm. Table 4 illustrates the outcome of our research and all web pages with more in links having greater PageRank value when using d=0.90. In contrast to the findings of other studies [21]-[23], we discovered that the damping factor had an impact on a website's ranking, despite the fact that the differences in results are related to our study's strategy in comparison when d=0.85. More particularly, the elements that affect the results may have been influenced by changes in the damping factor's value [24]-[26].

Table 4. PageRank value comparision

Web	PageRank (d=0.85)	PageRank (d=0.90)
www.mathworks.com	0.044342	0.050561
ch.mathworks.com	0.043085	0.049459
cn.mathworks.com	0.043085	0.049459
jp.mathworks.com	0.043085	0.049459
kr.mathworks.com	0.043085	0.049459
uk.mathworks.com	0.043085	0.049459
au.mathworks.com	0.043085	0.049459
de.mathworks.com	0.043085	0.049459
es.mathworks.com	0.043085	0.049459
fr.mathworks.com	0.043085	0.049459
in.mathworks.com	0.043085	0.049459
it.mathworks.com	0.043085	0.049459
nl.mathworks.com	0.043085	0.049459
se.mathworks.com	0.043085	0.049459
mathworks.com/index.html%3Fnocookie%3Dtrue	0.0015	0.001
mathworks.com/company/aboutus/policies_statements/patents.html	0.007714	0.0060468
mathworks.com/company/aboutus/policies_statements/trademarks.html	0.007714	0.0060468
mathworks.com/company/aboutus/policies_statements	0.006439	0.0051468
mathworks.com/company/ piracy.html	0.006439	0.0051468
mathworks.com /rss/index.html	0.006439	0.0051468
ch.mathworks.com/index.html%3Fnocookie%3Dtrue	0.0015	0.001
ch.mathworks.com/company /patents.html	0.0051817	0.0040451
ch.mathworks.com /trademarks.html	0.0051817	0.0040451
ch.mathworks.com/company/aboutus/policies_statements.html	0.0051817	0.0040451
ch.mathworks.com/company /piracy.html	0.0051817	0.0040451

5. CONCLUSION

In conventional Google PageRank technology, the damping factor is a significant aspect in adjusting a website's ranking. The damping factor is a key component in the traditional Google PageRank algorithm that affects page ranking, and can be changed to any number between 0 and 1. After calculation has done with each probability d value, according to the findings we would like to pinpoint that the resuls of our work shows the best and accurate ranking when d=0.90 because the sum of all web page values is very close to ideal value (1) which it is (0.7392424) and gave better PageRank value for each web page in the given data. In our future work, we'll also talk about how to improve efficiency in order to discover important nodes and eradicate all link spam.

ACKNOWLEDGEMENTS

Thanks to computer technical engineering college's staff for providing research facilities to excute the work.

REFERENCES

[1] T. A.Jilani, U. Fatima, M. Mahmood Baig, and S. Mahmood, "A survey and comparative study of different PageRank algorithms," *International Journal of Computer Applications*, vol. 120, no. 24, pp. 24–30, Jun. 2015, doi: 10.5120/21410-4444.

[2] K. S. Patel and K. Sarvakar, "Search engine optimization PageRank algorithm," *International Journal of Computer Science Engineering and Information Technology Research*, vol. 8, no. 3, pp. 17–24, 2018, doi: 10.24247/ijcsetiraug20183.

[3] T. Liu, Y. Qian, X. Chen, and X. Sun, "Damping effect on pagerank distribution," in *2018 IEEE High Performance Extreme Computing Conference, HPEC 2018*, Sep. 2018, pp. 1–11, doi: 10.1109/HPEC.2018.8547555.

[4] Z. Bar-Yossef and L. T. Mashiach, "Local approximation of pagerank and reverse pagerank," in *International Conference on Information and Knowledge Management, Proceedings*, 2008, pp. 279–288, doi: 10.1145/1458082.1458122.

[5] M. Kaur and C. Singh, "A hybrid page rank algorithm: an efficient approach," *International Journal of Computer Applications*, vol. 100, no. 16, pp. 58–63, Aug. 2014, doi: 10.5120/17613-8420.

[6] C. Ziakis, M. Vlachopoulou, T. Kyrkoudis, and M. Karagkiozidou, "Important factors for improving Google search rank," *Future Internet*, vol. 11, no. 2, p. 32, Jan. 2019, doi: 10.3390/fi11020032.

[7] D. F. Gleich, "PageRank beyond the web," *SIAM Review*, vol. 57, no. 3, pp. 321–363, Jan. 2015, doi: 10.1137/140976649.

[8] C. Yuan, Y. Lu, K. Liu, G. Liu, R. Dai, and Z. Wang, "Exploration of Bi-Level PageRank algorithm for power flow analysis using graph database," in *2018 IEEE International Congress on Big Data (BigData Congress)*, Jul. 2018, pp. 143–149, doi: 10.1109/BigDataCongress.2018.00026.

[9] K. Kumar, "PageRank algorithm and its variations: a survey report," *IOSR Journal of Computer Engineering*, vol. 14, no. 1, pp. 38–45, 2013, doi: 10.9790/0661-1413845.

[10] M. Işık and H. Dağ, "An effective rocommender model for E-commerce platforms," *Mugla Journal of Science and Technology*, vol. 3, no. 2, pp. 143–149, Dec. 2017, doi: 10.22531/muglajsci.357313.

[11] J. Agrawal, N. Sharma, P. Kumar, V. Parshav, and R. H. Goudar, "Ranking of searched documents using semantic technology," *Procedia Engineering*, vol. 64, pp. 1–7, 2013, doi: 10.1016/j.proeng.2013.09.070.

[12] J. H. Kim, M. L. Li, K. Selçuk Candan, and M. L. Sapino, "Personalized pagerank in uncertain graphs with mutually exclusive edges," *SIGIR 2017 - Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 525–534, doi: 10.1145/3077136.3080794.

[13] S. Kumar, "Analyzing the Facebook workload," *Proceedings - 2012 IEEE International Symposium on Workload Characterization, IISWC 2012*, 2012, pp. 111–112, doi: 10.1109/IISWC.2012.6402911.

[14] K. Avrachenkov, N. Litvak, and K. S. Pham, "A singular perturbation approach for choosing the pagerank damping factor," *Internet Mathematics*, vol. 5, no. 1–2, pp. 47–69, 2008, doi: 10.1080/15427951.2008.10129300.

[15] J. Zhan, S. Gurung, and S. P. K. Parsa, "Identification of top-K nodes in large networks using Katz centrality," *Journal of Big Data*, vol. 4, no. 1, 2017, doi: 10.1186/s40537-017-0076-5.

[16] K. Sugihara, "Beyond Google's PageRank: complex number-based calculations for node ranking," *Global Journal of Computer Science and Technology*, vol. 19, no. 3, pp. 1–12, 2019, doi: 10.34257/gjcstevol19is3pg1.

[17] S. Singla, N. Duhan, and U. Kalkal, "A novel approach for document ranking in digital libraries using extractive summarization," *International Journal of Computer Applications*, vol. 74, no. 18, pp. 25–31, 2013, doi: 10.5120/12986-0179.

[18] K. Sagar and A. Saha, "Qualitative usability feature selection with ranking: a novel approach for ranking the identified usability problematic attributes for academic websites using data-mining techniques," *Human-centric Computing and Information Sciences*, vol. 7, no. 1, 2017, doi: 10.1186/s13673-017-0111-8.

[19] S. . Jayanthi and S. Sasikala, "Link spam detection based on DBSPAMCLUST with fuzzy c-means clustering," *International Journal of Next-Generation Networks*, vol. 2, no. 4, pp. 1–10, 2010, doi: 10.5121/ijngn.2010.2401.

[20] B. Jaganathan and K. Desikan, "Weighted page rank algorithm based on in-out weight of webpages," *Indian Journal of Science and Technology*, vol. 8, no. 34, pp. 0–5, 2015, doi: 10.17485/ijst/2015/v8i34/86120.

[21] H.-G. Jun, D.-H. Im, and H.-J. Kim, "An RDF metadata-based weighted semantic Pagerank algorithm," *International journal of Web & Semantic Technology*, vol. 7, no. 2, pp. 11–24, 2016, doi: 10.5121/ijwest.2016.7202.

[22] W. K. Li and G. Li, "Discussion," *Applied Stochastic Models in Business and Industry*, vol. 23, pp. 237–239, 2009, doi: 10.1002/asmb.

[23] N. NarayanDas, E. Kumar, and S. Sheetal, "Approaches of Page Ranking algorithms: review," *International Journal of Computer Applications*, vol. 82, no. 2, pp. 31–38, 2013, doi: 10.5120/14090-2094.

[24] A. Dode and S. Hasani, "PageRank algorithm," *IOSR Journal of Computer Engineering*, vol. 19, no. 01, pp. 01–07, 2017, doi: 10.9790/0661-1901030107.

[25] S. Ghiam, "A survey on web spam detection methods: taxonomy," *International Journal of Network Security & Its Applications*, vol. 4, no. 5, pp. 119–134, 2012, doi: 10.5121/ijnsa.2012.4510.

[26] A. Althaf Ali and R. M. Shafi, "Test-retrieval framework: performance profiling and testing web search engine on non factoid queries," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 14, no. 3, pp. 1373–1381, 2019, doi: 10.11591/ijeecs.v14.i3.pp1373-1381.

BIOGRAPHIES OF AUTHORS

Ali Ali Saber received the B.Sc. degree in Computer Science from Kirkuk University, Iraq, in 2018 and the M.S. degrees in computer software engineering from Tehran University, Tehran, Iran, More than three years of experience as a lecturer in the field of computer engineering including experience in software programs, over eight years of experience working in HR section, A key team member with strong leadership and ability to work under pressure .Trilingual with fluent verbal and written skills in Arabic, Turkish, English, Persian and Kurdish languages. Experienced in dealing with different cultures and nationality. Communicate effectively, thrives on responsibility and challenge. He can be contacted at email: Ali.a.saber@uoalkitab.edu.iq.

Aso Kamaran Omer holds a M.S. in business administration from University of Tehran, Iran. He is a Multi-task, efficient and reliable administrative professional with over ten years of experience supporting directors, chairperson and managers to improve internal departmental operations. Accustomed to working in fastpaced environments Excellent interpersonal skills, ability to work well with others, in both supervisory and support staff roles. Diversified skill sets covering administrative support, client relations, human resources, accounts payable and project management. He also Responsible for the preparation of all personnel and administrative documents and advises personnel on a variety of administrative issues reviews documents for accuracy. He can be contacted at email: aso.omer@koyauniversity.org.

Noor Kaylan Hamid received the B.Sc. degree in Computer Science from Kirkuk University Iraq in 2007-2008, and M.Sc. Degree in Computer science and information technology Accurate specialty (Mobile Network) Salahaddin University, Iraq in 2011-2014. More than two years of experience as a lecturer in the field of computer engineering including experience in software programs, written skills in Arabic, English and Kurdish languages. Experienced in dealing with different cultures and nationality. She can be contacted at email: noor.g.h@uoalkitab.edu.iq.