ISSN 1330-3651(Print), ISSN 1848-6339 (Online)
DOI: 10.17559/TV-20140512181430

A COMPARATIVE EVALUATION OF POPULAR SEARCH ENGINES ON FINDING TURKISH DOCUMENTS FOR A SPECIFIC TIME PERIOD

Yıltan Bitirim, Abdül Kadir Görür

Original scientific paper
This study evaluates the popular search engines, Google, Yahoo, Bing, and Ask, on finding Turkish documents by comparing their current performances with their performances measured six years ago. Furthermore, the study reveals the current information retrieval effectiveness of the search engines. First of all, the Turkish queries were run on the search engines separately. Each retrieved document was classified and precision ratios were calculated at various cut-off points for each query and engine pair. Afterwards, these ratios were compared with the six years ago ratios for the evaluations. Besides the descriptive statistics, Mann-Whitney U and Kruskal-Wallis H statistical tests were used in order to find out statistically significant differences. All search engines, except Google, have better performance today. Bing has the most increased performance compared to six years ago. Nowadays: Yahoo has the highest mean precision ratios at various cut-off points; all search engines have their highest mean precision ratios at cut-off point 5; dead links were encountered in Google, Bing, and Ask; and repeated documents were encountered in Google and Yahoo.

Keywords: *information retrieval; performance evaluation; search engine; Turkish*

Usporedna evaluacija popularnih mehanizama za pretraživanje u pronalaženju turskih dokumenata određenog vremenskog razdoblja

Izvorni znastveni članak
U ovom se istraživanju ocjenjuju popularni mehanizmi za pretraživanje, Google, Yahoo, Bing, i Ask, pri traženju turskih dokumenata usporedbom njihovog sadašnjeg rada s radom izmjerenim prije šest godina. Nadalje, istraživanje pokazuje sadašnju učinkovitost mehanizama u pronalaženju podataka. Najprije su učinjeni upiti za turske riječi odvojeno na svakom mehanizmu. Svaki pronađeni dokument je klasificiran, a izračunati su omjeri točnosti na raznim cut-off točkama za svaki upit i svaki mehanizam. Zatim su ti omjeri uspoređeni s onima od prije šest godina zbog procjene. Pored opisne statistike, korišteni su Mann-Whitney U i Kruskal-Wallis H statistički testovi kako bi se pronašle statistički značajne razlike. Svi mehanizmi za ispitivanje osim Google-a danas su učinkovitiji. Bing je njviše napredovao u odnosu na prije šest godina. Danas Yahoo ima najviše prosječne omjere točnosti u raznim cut-off točkama. Svi mehanizmi za pretraživanje imaju najviše prosječne omjere točnosti u cut-off točki 5; ugašene veze (dead links) su nađene u Google-u, Bingu i Asku, a ponovljeni dokumenti u Google-u i Yahoo-u.

Ključne riječi: *mehanizam za pretraživanje; ocjena uspješnosti; pronalaženje informacija; turski*

1 Introduction

Most of the people in the world use their native languages while accessing the Web [1]. There were more than 36 million Internet users in Turkey in the middle of 2012[2]. Most of the users use Turkish queries on international search engines to retrieve their information needs. Therefore, retrieval effectiveness of international search engines on finding Turkish documents is important.

Demirci et al. [3] investigated information retrieval effectiveness of the popular international search engines on Turkish document retrieval in terms of precision and normalized recall in 2007. In the same study, they also compared the international search engines with the local search engine.

Similar methodology and some results of the study done by Demirci et al. [3] are used in our study to investigate how the Turkish document retrieval effectiveness of the popular international search engines has been changed from six year ago to now.

There are a number of studies on evaluation of search engines such as follows: Li and Shang presented 2 different automatic scoring algorithms (term-based algorithm and a new three-level algorithm) and a manual method for evaluating search engine performance and carried out an experiment with 2 different query sets on the search engines, AltaVista, Google, and InfoSeek under the default search and exact-phrase search modes [4]. They mainly found out that three-level algorithm had consistent results with the results of manual method and Google was the best under both search modes; Bitirim et al. investigated retrieval performance of 4 Turkish search engines (Arabul, Arama, Netbul, and Superonline) with respect to various measures, e.g., precision and normalized recall ratios [5]. In the experiment, 2 different sets of queries (as 17 queries and 5 queries) were defined and used. One of the findings is that the search engine Arama was the best both in terms of mean precision ratio and in terms of mean normalized recall ratio; Tümer et al. selected 3 keyword-based search engines (Google, Yahoo, and MSN) and a semantic-search engine (Hakia) and evaluated their semantic search performance [6]. Although Yahoo was the best in terms of mean precision ratio and Google was the best in terms of mean normalized recall ratio, semantic search performances of 3 keyword-based and 1 semantic search engines were not good; Zhang and Fei studied to find out the effect of several search features of the search engines, Yahoo, Google, and Live Search, on information retrieval performance [7]. Some of their findings are that, for all the engines, the PDF file format restriction search had the best retrieval performance and the regular search had the best Web page ranking performance; Sadeghi introduced 2 new automatic performance evaluation methods for search engines and investigated the performance of three search engines (Ask, Bing, and Google) by using these methods [8]. Significant degrees of consistency were encountered between automatic and manual-based assessments. Among these search engines, Google was

the best engine; and Garoufallou evaluated the performances and characteristics of 4 international search engines (Google, AltaVista, Yahoo, and Exalead) and 4 Greek search engines (Google.gr, In.gr, Robby.gr, and Find.gr) in point of view of Greek librarians [9]. One finding is that the participants preferred to use international search engines rather than Greek ones in general.

The common inspiration of the aforementioned studies and our study is to motivate researchers and search engine providers towards improving the search engines for better information retrieval performance as well as to give an idea about search engine performances to the users.

This paper is organized as follows: the next section describes the methodology; Section 3 presents the experimental findings; and the last section concludes the paper.

2 Methodology

In order to be able to compare the results of this study with the results presented in Demirci et al. [3], 4 of the search engines, the same queries, and the same evaluation method of the retrieval outputs are used from [3].

The international search engines Google, Yahoo, Bing, and Ask are selected since they are the top 4 most popular search engines nowadays [10]. Note that the search engine Bing is the same search engine mentioned as MSN in [3]; therefore, it might also be called "New MSN".

The queries are ‘bahçe’ (garden), ‘pencere’ (window), ‘gazete’ (newspaper), ‘ağaç’ (tree), ‘renk’ (colour), ‘dolap’ (wardrobe), ‘burun’ (nose), ‘beyin’ (brain), ‘bilim’ (science), ‘kitap’ (book), ‘çiçek’ (flower), and ‘oyun’ (game).

Search engines may require settings for direct search of as many as possible Turkish documents among the world. Therefore, the following settings are done on selected search engines:

- For Google:
  - http://www.google.com.tr is used,
  - safe search filter is turned off,
  - location is not given since there may be Turkish documents located in countries other than Turkey,
  - search results language is selected as Turkish.
- For Bing:
  - http://www.bing.com.tr is used,
  - safe search filter is turned off,
  - worldwide search is selected since there may be Turkish documents located in countries other than Turkey,
  - wearch results language is selected as Turkish.
- For Yahoo:
  - Only http://tr.yahoo.com is used. There is no any other setting for Turkish language on this search engine.
- For Ask:
  - http://www.ask.com is used, since there is no setting for Turkish language available on this search engine.

After the aforementioned settings are done on selected search engines, the queries given above are run on the search engines separately and, at each run, the first 20 documents (it is worth to say that using the first 20 documents retrieved is meaningful since Spink and Jansen [11] found out that the users’ tendency is not to view the results pages beyond the first or second results pages and the selected search engines Google, Yahoo, Bing, and Ask displayed 10 documents in each of the first two results pages by default) in the retrieval output are evaluated based on human relevance judgment. In the evaluation, every document in the retrieval output is categorised as "relevant" or "non-relevant" with the authors’ consensus. During this categorisation, the following criteria, which are generally based on the criteria used in [5], are considered:

- If more than one document which have the same content but different Web addresses are retrieved in a retrieval output, they are considered as different ones.
- If more than one document which have the same content as well as the same Web address are retrieved in a retrieval output, the first one is considered while the other(s) (repeated document(s)) is/are categorised as non-relevant.
- If the document is not relevant but contains a link to a relevant document, it is categorised as relevant.
- If the document is not in Turkish, it is categorised as non-relevant.
- If the document is not accessible, it is categorised as non-relevant.

The precision measure can be defined as the ratio of the relevant documents retrieved to the retrieved documents. The 3 characteristics including the simplicity, fairness, and reliability make precision as one of the basic measures used in many search engine evaluations [12]. In addition to this, it is an important measure since the users are interested in the relevant documents in the retrieval output [4].

When every document is evaluated as relevant or non-relevant, precision ratios are calculated at various cut-off points (for first 5, 10, 15, and 20 documents retrieved) for each query and search engine pair. After that, the mean precision ratios at 4 cut-off points from the six years ago study [3] and the current study are examined to observe how the performance of the search engines has been changed.

Besides the descriptive statistics, Mann-Whitney U test [13] is carried out in order to find out whether there are statistically significant differences between six years ago (SYA[1]) , and present-day (PD[2]) precision ratios of the search engines for each cut-off point. Furthermore, Kruskal-Wallis H test [14] is used to find out if there is a statistically significant difference between PD precision ratios of the search engines for each cut-off point and when a difference is encountered, pair-wise Mann-Whitney U test is used to determine which search engines engendered it. Note that the 90 % confidence level is used for all statistics.

[1] Since these results were obtained and presented in 2007 by Demirci et al. [3], we will refer to them as six years ago (SYA).
[2] We will refer to the results of our study as present-day (PD).

3 Experimental results

For PD, total 960 documents are examined one by one with human relevance judgement. The number of relevant documents retrieved by the search engines in PD and SYA are given in table 1. Except for Google, the total number of relevant documents retrieved is increased for all search engines. In PD, while the ratio of relevant documents decreased 7 % for Google, it is increased 19 %, 13 %, and 10 % for Bing, Yahoo, and Ask, respectively. The ratio is mostly increased at Bing. However, the most relevant documents are retrieved by Yahoo with 78 %.

In SYA, Ask retrieved 0 relevant documents for two queries. However, none of the search engines retrieved 0 relevant documents in PD. When comparing the PD and SYA, the number of relevant documents retrieved is decreased by 50 % of the queries run on Google. Yet, it is increased by 83 %, 75 %, and 67 % of the queries run on Yahoo, Bing, and Ask, respectively.

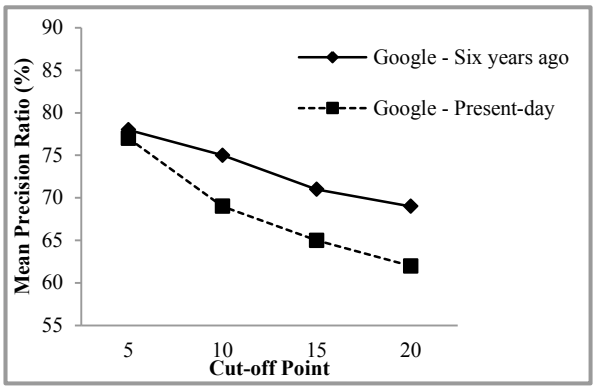

Figure 1 Mean precision ratios of Google for PD and SYA

Table 1 Number of relevant documents retrieved by each search engine

Query	Google		Yahoo		Bing		Ask	
	SYA	PD	SYA	PD	SYA	PD	SYA	PD
bahçe	11	11	9	10	7	9	6	5
pencere	11	4	7	11	9	18	6	3
gazete	19	19	15	19	10	20	10	18
ağaç	10	8	6	8	1	3	0	8
renk	12	4	12	9	1	7	0	1
dolap	11	8	9	17	5	16	3	5
burun	13	14	18	19	19	19	13	14
beyin	11	9	9	18	15	14	5	6
bilim	13	17	14	16	10	17	6	12
kitap	16	20	17	20	16	17	14	15
çiçek	19	15	20	20	17	14	11	10
oyun	20	20	19	20	18	19	20	20
Total:	**166**	**149**	**155**	**187**	**128**	**173**	**94**	**117**
Avg (%):	**69**	**62**	**65**	**78**	**53**	**72**	**39**	**49**

When the PD data is considered: Google displayed a dead link for each 2 retrieval outputs and a repeated document in 1 retrieval output; Bing and Ask displayed a dead link for 1 retrieval output and no repeated document; Yahoo displayed no dead link but a repeated document in 1 retrieval output; and finally, the dead link and document repetition are not encountered in the first 10 documents retrieved in general, although only the search engine Google retrieved a dead link in the first 5 documents retrieved for 1 retrieval output.

In Figure 1, mean precision ratios of Google are shown at various cut-off points (for first 5, 10, 15, and 20 documents retrieved) for PD and SYA. Even though mean precision ratios of PD and SYA are so close to each other at cut-off point 5, there are 6 or 7 % difference at cut-off points 10, 15, and 20. PD mean precision ratio is lower than SYA mean precision ratio at all cut-off points. Furthermore, for both PD and SYA, it is seen that mean precision ratios are decreased when the cut-off point increased. Statistically meaningful difference is not encountered between SYA and PD precision ratios of Google.

As it is seen in Figure 2, PD mean precision ratio of Yahoo is higher than its SYA mean precision ratio at all cut-off points. The difference of PD and SYA ratios is between 9 % and 14 % at all cut-off points. Mean precision ratios of PD are decreased 4 % from cut-off point 5 to 10 and from cut-off point 10 to 15, while the ratio is increased 1 % from cut-off point 15 to 20. However, mean precision ratios of SYA are decreased, when the cut-off point increased. Statistically meaningful difference is not encountered between SYA and PD precision ratios of Yahoo.

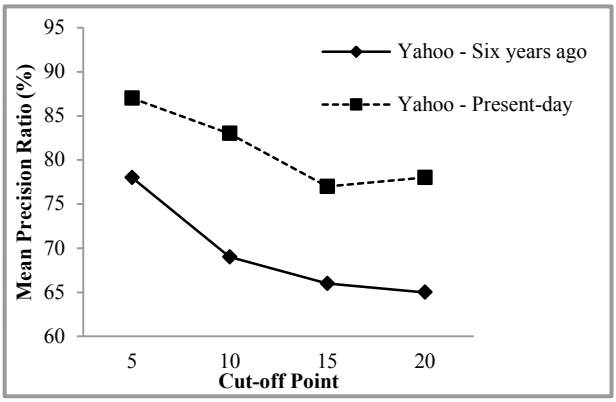

Figure 2 Mean precision ratios of Yahoo for PD and SYA

As shown in Fig. 3, the differences of Bing's PD and SYA mean precision ratios at cut-off points 5, 10, 15, and 20 are 31 %, 28 %, 20 %, and 19 %, respectively. Mean precision ratio of PD is remarkably higher than SYA ratio at all cut-off points. In PD, mean precision ratios are

decreased 2 % from cut-off point 5 to 10 and from cut-off point 15 to 20, while it is 7 % from cut-off point 10 to 15. The highest mean precision ratio of PD (83 %) is at cut-off point 5. For SYA, although mean precision ratios are close to each other at 4 cut-off points, the lowest mean precision ratio (52 %) is at cut-off point 5. There is statistically meaningful difference between SYA and PD precision ratios when the cut-off point is 5 (U = 34,000, p = 0,023, z = −2,269, r = −0,46), 10 (U = 36,500, p = 0,038, z = −2,071, r = −0,42), and 15 (U = 43,500, p = 0,097, z = −1,661, r = −0,34). This means that in PD, Bing retrieved more relevant documents than SYA at cut-off points 5, 10, and 15.

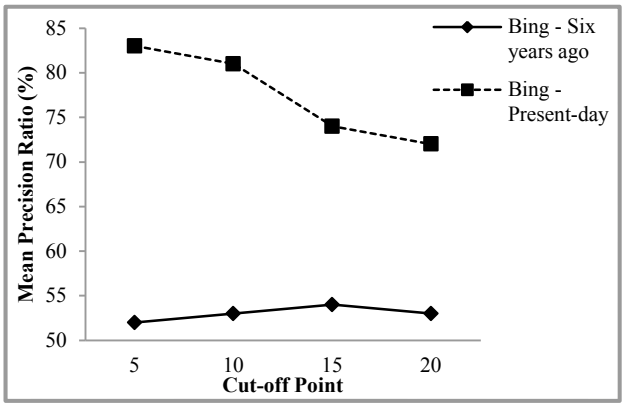

Figure 3 Mean precision ratios of Bing for PD and SYA

PD and SYA mean precision ratios of Ask for 4 cut-off points are shown in Figure 4. At all cut-off points, PD mean precision ratio is higher than SYA mean precision ratio.

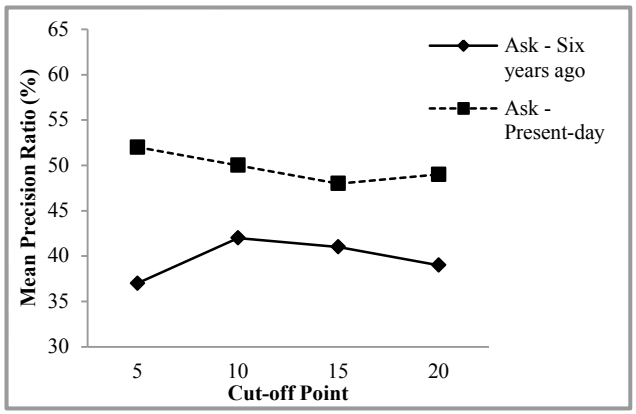

Figure 4 Mean precision ratios of Ask for PD and SYA

The highest difference of PD and SYA mean precision ratios is 15 % at cut-off 5. Besides, while the highest PD mean precision ratio (52 %) is at cut-off point 5, the lowest SYA mean precision ratio (37 %) is at the same cut-off point. From cut-off point 5 to10 and from cut-off point 10 to 15, mean precision ratios of PD are decreased whereas the mean precision ratio is increased from cut-off point 15 to 20. For SYA, mean precision ratio is increased 5 % from cut-off point 5 to 10. But, mean precision ratios are decreased for the rest. Statistically meaningful difference is not encountered between SYA and PD precision ratios of Ask.

Fig. 5 clearly shows that Yahoo is the best search engine at all cut-off points in PD. Then, Bing is the second, Google is the third, and Ask is the last. The highest mean precision ratio (87 %) belongs to Yahoo at cut-off point 5. Every search engine has the highest relevant document ratio, firstly, at cut-off point 5, and secondly, at cut-off point 10. The lowest mean precision ratios of Yahoo and Ask are 77 % and 48 %, respectively, at cut-off point 15. These ratios are 72 % and 62 % for Bing and Google, respectively, at cut-off point 20. Mean precision ratios are decreased between 2 % and 8 % from cut-off point 5 to 10 and between 2 % and 7 % from cut-off point 10 to 15 at all search engines. Average of mean precision ratios of Yahoo is 81 %, followed by Bing, Google, and Ask with 78 %, 68 %, and 50 %, respectively.

There is statistically meaningful difference between PD precision ratios of the search engines at cut-off point 5 (H = 10,195, 3 d.f., p = 0,017) and cut-off point 10 (H = 8,523, 3 d.f., p = 0,036). After Bonferroni correction is applied and significance level is accepted as 0,0167, Mann-Whitney U tests are used to reveal which search engine pair(s) engendered the difference. For cut-off point 5, there is statistically meaningful difference between Yahoo and Ask (U = 25,000, p = 0,005, r = −0,57) and Bing and Ask (U = 30,000, p = 0,012, r = −0,51). For cut-off point 10, there is statistically meaningful difference between Yahoo and Ask (U = 29,000, p = 0,012, r = −0,51) and Bing and Ask (U = 30,500, p = 0,016, r = −0,49). As a result, Yahoo and Bing have retrieved more relevant documents than Ask at cut-off point 5 as well as cut-off point 10.

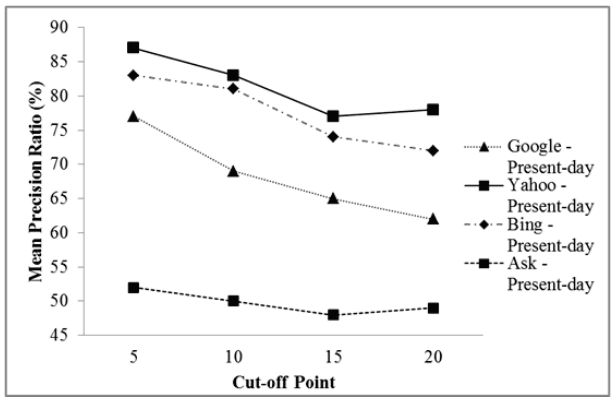

Figure 5 Mean precision ratios of the search engines for PD

4 Conclusion

When all documents in all retrieval outputs of each search engine are considered, the ratio of relevant documents is decreased 7 % from SYA to PD for Google. In contrast, this ratio is increased between 10 % and 19 % for the other search engines. Although the ratio is mostly increased at Bing (19 %), the most relevant documents are retrieved by Yahoo (78 %) in PD. In PD, none of the search engines retrieved 0 relevant documents, dead links are encountered in Google, Bing, and Ask, and moreover repeated documents are encountered in Google and Yahoo.

PD mean precision ratio is higher than SYA mean precision ratio at all cut-off points for all search engines

except Google. Bing has the most increased PD mean precision ratio at all cut-off points. In PD, the best search engine is Yahoo at all cut-off points and the highest mean precision ratio of Yahoo is 87 % at cut-off point 5. Average of PD mean precision ratio of Yahoo is 81 % and for Bing, Google, and Ask is 78 %, 68 %, and 50 %, respectively. In PD, all search engines have their highest mean precision ratios at cut-off point 5 and, then, at cut-off point 10.

Besides the descriptive statistics, some advanced statistical tests are conducted and it is encountered that in PD, Bing retrieved more relevant documents than SYA when considering the first 5, 10, and 15 documents retrieved and in PD, Yahoo and Bing more relevant documents have been retrieved than in Ask in both the first 5 and the first 10 documents retrieved.

Acknowledgements

We would like to express our sincere thanks to Prof. Dr. Yaşar Tonta from Hacettepe University for his invaluable guidance on the statistical work.

5 References

[1] Chung, W. Studying information seeking on the non-English Web: An experiment on a Spanish business Web portal. // Int. J. Hum-Comput Stud. 64, 9(2006), pp. 811-829. DOI: 10.1016/j.ijhcs.2006.04.009

[2] Miniwatts Marketing Group, Internet usage in Europe June 2012. URL:http://www.internetworldstats.com/stats4.htm. (19/02/2013).

[3] Demirci, R. G.; Kişmir, V.; Bitirim Y. An evaluation of popular search engines on finding Turkish documents. // Proceedings of the 2nd International Conference on Internet and Web Applications and Services (ICIW)//Morne, Mauritius, 13-19 May 2007, pp. 61. DOI: 10.1109/ICIW.2007.15

[4] Li, L.; Shang, Y. A new statistical method for performance evaluation of search engines. // Proceedings of the 12th International Conference on Tools with Artificial Intelligence (ICTAI)/ Vancouver, BC, Canada, November 2000, pp. 208-215.

[5] Bitirim, Y.; Tonta, Y.; Sever, H. Information retrieval effectiveness of Turkish search engines. // Proceedings of the Second International Conference on Advances in Information Systems (ADVIS '02), 2002, İzmir, Turkey, Tatyana M. Yakhno (Ed.). Springer-Verlag, London, UK, 2457, pp. 93-103. DOI: 10.1007/3-540-36077-8_9

[6] Tümer, D.; Shah, M. A.; Bitirim, Y. An empirical evaluation on semantic search performance of keyword-based and semantic search engines: Google, Yahoo, Msn and Hakia. // Proceedings of the 4th International Conference on Internet Monitoring and Protection (ICIMP)/Venice-Mestre, Italy, 24-28 May 2009, pp. 51-55. DOI: 10.1109/icimp.2009.16

[7] Zhang, J.; Fei, W. Search engines' responses to several search feature selections. // The International Information & Library Review. 43, 3(2010), pp. 212-225. DOI: 10.1080/10572317.2010.10762866

[8] Sadeghi, H. Automatic performance evaluation of search engines using judgments of metasearch engines. // Online Information Review. 35, 6(2011), pp. 957-971. DOI: 10.1108/14684521111193229

[9] Garoufallou, E. Evaluating search engines: A comparative study between international and Greek SE by Greek librarians. // Program-Electronic library and information systems. 46, 2(2012), pp. 182-198. DOI: 10.1108/00330331211221837

[10] eBizMBA Inc., Top 15 most popular search engines, February 2013. URL: http://www.ebizmba.com/articles/search-engines. (19/02/2013).

[11] Spink, A.; Jansen, B. J. Web search: public searching on the Web, volume 6 of Information science and knowledge management. Kluwer Academic Publishers Group, Norwell, MA, USA, and Dordrecht, the Netherlands, 2004.

[12] Li, K. F.; Wang, Y.; Yu, W. Chapter 7 Personalised search engine evaluation: Methodologies and metrics. // In: Lewandowski D, editor. Web Search Engine Research / Library and Information Science. Vol. 4, (2012), pp. 163-202. DOI: 10.1108/S1876-0562(2012)002012a009

[13] Mann, H. B.; Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the other. // The Annals of Mathematical Statistics. 18, 1(1947), pp. 50-60. DOI: 10.1214/aoms/1177730491

[14] Kruskal, W. H.; Wallis, W. A. Use of Ranks in One-Criterion Variance Analysis. // Journal of the American Statistical Association. 47, 260(1952), pp. 583- 621. DOI: 10.1080/01621459.1952.10483441

Authors' addresses

Yıltan Bitirim, Ph.D.
Department of Computer Engineering,
Faculty of Engineering,
Eastern Mediterranean University,
Famagusta/T.R.N.C., via Mersin 10, Turkey
yiltan.bitirim@emu.edu.tr

Abdül Kadir Görür, Ph. D.
Department of Computer Engineering,
Faculty of Engineering,
Çankaya University,
Yukarıyurtçu Mahallesi Mimar Sinan Caddesi No: 4 06810,
Etimesgut/ANKARA, Turkey
agorur@cankaya.edu.tr