Enhancing Online Similar Web Pages Advisor with Support of Text Processing

Metin Turan

***Abstract*—The Web is a lifestyle of this era. User searches information on Web data by daily usage. The problem is that when user browsing a Web page and interested in similar pages, then an application is needed to find out related information locations (web pages) called similar Web page advisor. It is obvious that this task requires more than a Web search engine.**

In this study, a simple text processing technique for English is devised in order to rearrange the output of the Web search engine. In other words, the HTML content of the Web pages on the links suggested by Web search engine are further processed and evaluated so that enhanced ranking of the top ten links is presented to the user.

The output of the System is compared with the well-known similar tool Chrome "similar Web pages" add-on application. The average Cosine similarity of the original Web page and suggested ten Web pages is considered. Our System overwhelms Chrome "similar Web pages" add-on. Moreover, it is more stable if different types of Web pages are considered.

***Index Terms*—Social recommendation, content analysis and feature selection, text processing.**

I. Introduction

Similarity is an interesting research area for any kind of objects. It is a hard problem to find out similar objects in big data especially [1]. As the Internet is a big data store, it is getting more importance searching in the bulk of documents residing on the Internet called Web mining [2] or document management [3].

Web search engines do a simple search in a hypertext [4]. These engines have ranked the links of hypertexts in a degree similar to searched keywords. However, keywords generally can't be determined correctly or even unrelated for an inexperienced user [5]. Moreover tagging (clustering) of hypertexts is generally done manually instead of an automated procedure [6], [7]. Scalability is also an important factor in the case of struggling with such a big data [8]. All these things have been resulted in lower quality links list outcomes from Web search engines. Nowadays, modern Web search engines use some ranking factors such like that classification, localization and linguistic features (entities, citations).

The strategies to find out similar Web pages using Web search engines that can be applied by the user and which one is the best also discussed [9]. Web search engines can be categorized so that it is useful when a specialized Web search is intended by user [10]. Web search engines need more clever algorithms [11], [12] in order to present more scalable, categorized [13] and accurately ranked links [14], [15] to the user. Moreover, it would be more useful, suggesting Web page links similar to browsed Web page (or user preference) [16], [17] instead of searching with keywords. Some researchers have got ahead and focused on guessing the next movement of the user lately [18].

The hyperlinks referred in the Web pages are the most commonly used technique by researchers in order to obtain similar Web pages [19], [20]. Moreover, these links are clustered [21] so that some kind of categorizing is supported to the Web search engine. This leads to an assumption that user surfing on the Web for similar Web pages (using some of these links) [22]. Nowadays, they tend to process Web pages in HTML format [23], [24] or using Web search engine parameters and textual content for structural similarity [25], [26]. Both require text processing (formatted as HTML or natural language) and text similarity measures via Information Retrieval.

Text similarity research requires text mining techniques [27]. It includes extracting the features in text and comparing with others. Some researchers tend to use unsupervised techniques obtain structured features from hypertext of Web pages [28], [29]. The others believe that text is written in a natural language, so it would be wise, including some way of natural language processing [30]. However Web pages are shorter than a classical text, by the way text processing techniques require more attention [31]. Practical applications for textual similarity can be given, such as clustering [32], plagiarism [33] and summarization [34], [35].

The problem considered in this article is actually a kind of clustering problem. In other words, the Internet documents may be categorized into different clusters. However, a huge amount of documents restricts making such a classification instantly. What is the number of clusters? It is unknown and it totally depends on user preference. On the other hand, it is obvious a Web search engine can be used to get similar links as an answer of user queries. If it is possible to describe a Web page with some keywords using text processing and execute a query, then a list of possible similar Web pages would be suggested in a rank by Web search engines. Moreover, for the reasons mentioned above, it could be further processed and evaluated for document cosine similarity (using document vectors) to get a better ranking. The System compared with the best similar application (Chrome "similar Web pages" add-on). It produces better similar links.

Section II is a part about the similar live tools developed in the problem area. Section III explains the approach. The Section IV discusses the theory behind the system. The Section V explains experiments and results. The Section VI and VII discuss the findings (conclusion) and further work respectively.

Manuscript received October 9, 2016; revised March 20, 2017.

Metin Turan is with the İstanbul Commerce University, Küçükyalı, İstanbul, Turkey (e-mail: mturan@ ticaret.edu.tr).

doi: 10.18178/ijfcc.2017.6.1.481

II. SIMILAR TOOLS

There is a list of tools developed in order to suggest similar Web pages (or sites). Some of them are listed and compared in Table I.

TABLE I: TOOLS FOR SIMILAR WEB PAGE ADVISING

Tool Name	Add-on	Similar Sites	Similar Web Pages
www.similarsitesearch.com	no	yes	no
www.similarsites.com	yes	yes	no
www.similarweb.com	no	yes	no
www.similarpages.com	no	yes	no
Chrome similar web pages	yes	no	yes
www.siteslike.com	no	yes	no
www.sitesimilarto.com	no	yes	no
www.moreofit.com	no	yes	no

The problem handled in this article is to suggest similar Web pages to the user online (when browsing a Web page), the only tool support this idea is the Chrome "similar Web pages" add-on application. The others are applications, giving whether similar Web sites using categorized (indexed information on the Web) or statistical information about the Web site usage. They use offline information about the Web sites. On the other hand the problem is to suggest similar Web pages (not sites) dynamically (when user browsing a Web page). By the way, the System (our application) is compared with Chrome "similar Web pages" add-on application in success.

The opportunities of tools are summarized at Table I briefly. The first one is www.similarsitesearch gives a rating and topics about the searched Web site. It also supports filtering results by languages and/or country. www.similarsites.com is the most useful application in the similar Web site category. It lets user to select similar sites by category and presents lots of statistical information related to the Web site (similar sites traffic/ visits together / searches together/ topics). www.similarweb.com presents some statistical information such like that ranking (in global/country/category level), total user visits, traffic by countries and subdomain information (traffic distribution). It also has a professional version called "similarwebpro". www.similarpages.com is another application for similar Web sites searching for the indexed Web sites. It doesn't support Web pages similarity. www.siteslike.com,www.sitesimilarto.com and www.moreofit.com are similar applications as www.similarpages.com.

On the other hand, Chrome "similar Web pages" add-on is an online application using the active Web page and advises up to ten similar Web pages. It is easy to use with one click on the button added to the Chrome toolbar.

III. APROACH

The System schema of similar Web page advisor is given in Fig. 1. The first phase of the System is composed of determining the keywords represent the current Web page. In order to obtain keywords, text processing is applied to the all Web content (text). The inner text between the <P>, <a>, <li> and <td> tags in the HTML content of the Web page is considered. Tokenizer is applied to obtain the words in the text by discarding the stop words.

Porter Stemmer is applied to all words found, so that the standardization is supported by obtaining the root of a word (term). Then frequencies of the terms are calculated and the most frequent 10 terms (representative terms) are selected to compose representative vector of the Web page.

In the second phase, Web search engine is queried using the combinations of top 3 terms in the representative vector of the Web page. For example, if a Web page is represented by the following 3 top terms.

"diet", "food", "health"

The queries are composed of combinations of these top three terms are given at Table II. When different combinations of words are queried in Web search engine, it could return back links in different order or even with new links attached.

The first three hyperlinks returned by each query combination are evaluated in the third phase. These 12 hyperlinks (not similar) are processed further one by one to compose Web page representative vectors separately as described in the first phase.

TABLE II: TOP THREE TERMS QUERY COMBINATIONS

Query Combinations
diet food health
diet food
food health
diet health

Cosine similarity between the hyperlinks representative vectors and current page representative vector are calculated in the last phase. The cosine similarity value determines the similarity between Web pages, so that giving the ranking order of hyperlinks.

Finally, the top 10 links presented to the user (in order to compare with the Chrome similar Web page tool).

The System is an experimental application. It displays all the information needed by the researcher.

IV. THEORY

Web site advisors use indexing of the Web sites. However Web pages are dynamic and may contain detailed information on one of the topics of the Web site. Indexing doesn't work for categorizing the Web pages. Anyway the content of the Web page must be analyzed (possibly a dynamic page and content may change on time) to determine the specific details on the Web page. This could be only possible by text processing.

Processing the Web page online and searching for similar pages requires significant time. One of the smartest solutions for that problem could be searching in pre-filtered Web pages which are a result of the search using a Web search engine.

The problem is now how to pre-filter Web to get limited more similar Web pages from uncountable Web resources. The simple answer to this question is based on the content (text) located on the Web page. The Web page is an HTML formatted structure. It contains tags and tags have inner text. If these inner texts joined together to compose unstructured text then the rest of the application is only text mining. The well-known technique in text mining in order to analyze and describe the current Web page is finding the term frequencies (TF). Term is the root of a word in different form (e.g., plural, derivational affix). This standardizes words and let count terms correctly.

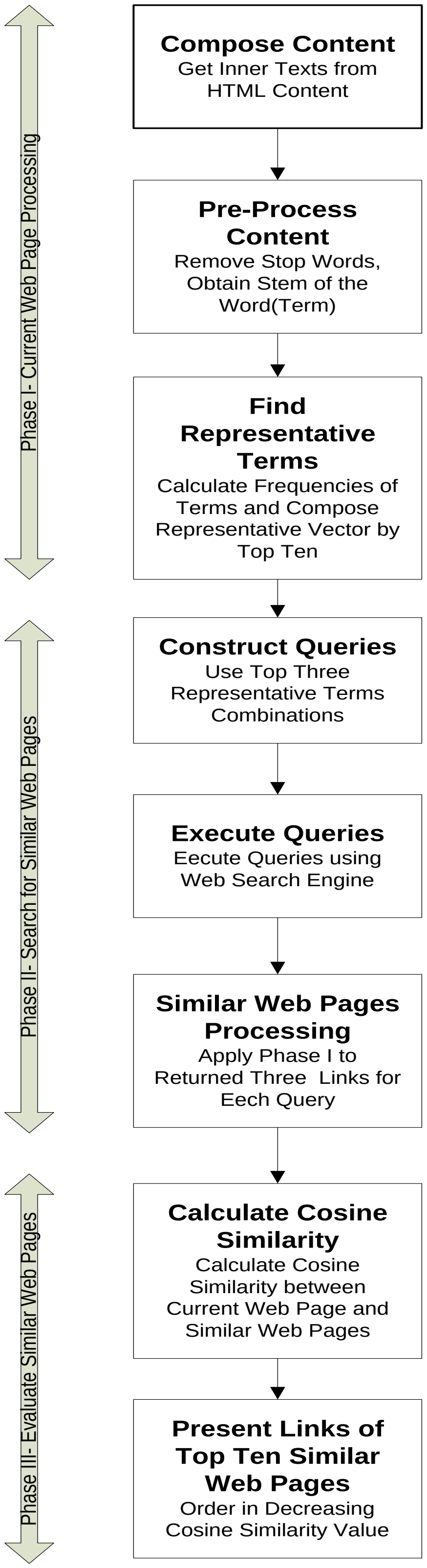
Phase I- Current Web Page Processing
Compose Content
Get Inner Texts from HTML Content
Pre-Process Content
Remove Stop Words, Obtain Stem of the Word(Term)
Find Representative Terms
Calculate Frequencies of Terms and Compose Representative Vector by Top Ten
Phase II- Search for Similar Web Pages
Construct Queries
Use Top Three Representative Terms Combinations
Execute Queries
Eecute Queries using Web Search Engine
Similar Web Pages Processing
Apply Phase I to Returned Three Links for Eech Query
Phase III- Evaluate Similar Web Pages
Calculate Cosine Similarity
Calculate Cosine Similarity between Current Web Page and Similar Web Pages
Present Links of Top Ten Similar Web Pages
Order in Decreasing Cosine Similarity Value

Fig. 1. General system schema of similar web page advisor.

Top 10 terms (most frequent terms) are used to represent current Web page as a vector. However Web search engine must be focused on the more valuable terms in the Web page representation vector. Experiments show that the top three terms are just enough generally (assuming terms frequencies is normally distributed) for searching.

The similarity between two documents can be computed using Cosine similarity function in (1).

$$\mathrm{Sim}\left(\overrightarrow{\mathrm{d}_1}, \overrightarrow{\mathrm{d}_2}\right) = \frac{\overrightarrow{d_1}.\ \overrightarrow{d_2}}{\|d_1\|\ \|d_2\|} \quad (1)$$

$\overrightarrow{d_j}$ is term representation vector of the jth Web page ($d_j(t_1, t_2, \ldots, t_{10})$, where t_i is the ith term sorted in decreasing frequency). Current page, top ten terms used in the Cosine similarity function. Since top three terms are selected for searching similar Web pages, then the similarity of the left seven terms (their frequencies are lower) determines the value of the Cosine similarity. If Cosine similarity reaches one, then this points out the highest similarity.

V. Experiments and Results

Chrome "similar Web pages" add-on is the most similar tool for the System developed. The Web page language is English. Some Web pages are selected and used in experiments.

Experiments include the following activities.

1) Current Web page content is composed.
2) Chrome "similar Web pages" add-on advised links are registered and their contents are composed.
3) System advised links are registered and their contents are composed.
4) Cosine text similarities are calculated between the current Web page content and Web pages suggested by the System and Chrome "similar Web pages" add-on.

The averages of similarities (for 10 pages) are compared. Some extreme examples and evaluation are given in Table III.

System overwhelms the Chrome "similar Web pages" add-on suggestions in general.

The most important observation, System suggests similar links in any case (Chrome "similar Web pages" add-on couldn't suggest at experiment 2 and only one link at experiment 4). The System is a more stable tool if compared with Chrome "similar Web pages" add-on.

VI. Conclusions

If the current Web page subject is specific (for example experiment 3) or it produces text is long enough (for example experiment 1) to determine the subject of the Web page, then the suggested pages are getting more similar to the current Web page. However, if the current Web page subject is general (not specific, for example experiment 4) or it produces a short text (for example experiment 2) then suggested pages similarities decrease rapidly. Chrome "similar Web pages" add-on is unsuccessful on latter cases. It may not even suggest a similar page (experiment 2).

If the frequencies of the left seven terms getting closer to zero, top three terms dominates the Cosine similarity. However, experiments indicate that the vital term (for example, in experiment 2, term Washington is in order 8) representing the Web page sometimes beyond the first three terms in order. This results in poorly advised similar Web pages.

VII. Further Works

If more than top three words are used, then Web pages could be more similar (but response time would be longer).

The Web page link address sometimes contains important cues (terms) as in example 2 (Washighton). However, this term is resolved to be eight representative term in the current Web page and it is not used in Web search engine. A method to pick up such terms from the Web link and evaluate separately would be useful.

Moreover, inner texts in the Web page HTML content may contain special characters or invaluable text (e.g. operational or commands) which drops down the Cosine similarity unexpectedly. If they could be discarded from the text, then representative terms could be determined with higher precision.

TABLE III: Experiments

Experiment Number: 1 **Current Page:** https://en.wikipedia.org/wiki/Health

System Representative Words for Current Web Page

health^{1}	*organ*^{2}	*world*^{3}	*healthi*^{4}	*diseas*^{5}
sleep^{6}	*mental*^{7}	*public*^{8}	*doi*^{9}	*social*^{10}

System Suggested Similar Pages	**Cosine Similarity**
http://health.howstuffworks.com/medicine/healthcare/who.htm	0.2427147
http://www.paho.org/hq/	0.2529031
http://www.cdc.gov/globalhealth/organization.htm	0.1713016
http://www.healthworldeducation.org/	0.1071393
http://www.who.int/about/en/	0.2204852
http://healthworldoutreach.org/default2.asp	0.1019565
https://humanhealth.org/	0.03965585
http://www.healthworld.com.au/index.html	0.1470681
https://www.chathamhouse.org/publication/what%E2%80%99s-world-health-organization	0.2732412
http://www.nytimes.com/topic/organization/world-health-organization	0.1620316
Average	**0.1718497**

Chrome "similar Web pages" ad-on Suggested Similar Pages	**Cosine Similarity**
http://www.businessdictionary.com/definition/health.html	0.2427147
https://en.wikibooks.org/wiki/Introduction_to_Sociology/Health_and_Medicine	0.2529031
http://www.merriam-webster.com/dictionary/health	0.1713016
https://www.nih.gov/health-information	0.1071393
http://health.usgs.gov/	0.2204852
http://nca2014.globalchange.gov/report/sectors/human-health	0.1019565
http://www.medicalnewstoday.com/articles/150999.php	0.03965585
http://health-and-medicine.wikia.com/wiki/Health_and_Medicine_Wiki	0.1470681
http://consumerwiki.dca.ca.gov/wiki/index.php/Health_and_Medicine	0.2732412
http://www.medicinenet.com/health_and_living/focus.htm	0.1620316
Average	**0.21639788**

Experiment Number: 2 **Current Page:** https://washington.org

System Representative Words for Current Web Page

free^{1}	*think*^{2}	*hotel*^{3}	*place*^{4}	*street*^{5}
attract^{6}	*bar*^{7}	*washington*^{8}	*museum*^{9}	*capitol*^{10}

System Suggested Similar Pages	**Cosine Similarity**
http://www.merriam-webster.com/dictionary/thing	0.1169229
http://www.thesaurus.com/browse/thing	0.07267261
http://www.discoverlosangeles.com/blog/100-free-things-do-los-angeles-free-activities	0.2097096
http://www.thefreesite.com/	0.08982144
http://travel.nationalgeographic.com/travel/city-guides/free-chicago-traveler/	0.1905304
https://www.timeout.com/los-angeles/free-things-to-do-in-LA	0.1473063
http://www.ebay.com/sch/i.html?_nkw=free+things	0.05466633
http://www.exploregeorgia.org/article/20-free-things-to-do-in-metro-atlanta	0.08708263
http://www.discoverlosangeles.com/blog/100-free-things-do-los-angeles-free-activities	0.2097188
http://www.inetgiant.com/	0.08053833
Average	**0.1258969**

Chrome "similar Web pages" ad-on Suggested Similar Pages	**Cosine Similarity**
No Suggestion	
Average	**undetermined**

Experiment Number: 3 **Current Page:** http://www.asp.net

System Representative Words for Current Web Page

net^{1}	*asp*^{2}	*commun*^{3}	*web*^{4}	*core*^{5}
2016^{6}	*microsoft*^{7}	*privaci*^{8}	*api*^{9}	*mvc*^{10}

System Suggested Similar Pages	**Cosine Similarity**
https://msdn.microsoft.com/en-us/library/aa286485.aspx	0.2969063
http://www.w3schools.com/aspnet/	0.2187691
https://aspnet.codeplex.com/	0.347537
https://docs.asp.net/en/latest/intro.html	0.3585593
https://github.com/aspnet/Home	0.2186927
https://msdn.microsoft.com/en-us/library/4w3ex9c2.aspx	0.4533583
https://www.lynda.com/ASP-NET-tutorials/ASP-NET-Essential-Training/784-2.html	0.1772252
http://weblogs.asp.net/	0.1475988

https://docs.asp.net/	0.3849156
http://www.w3schools.com/asp/default.asp	0.2267737
Average	**0,2830336**
Chrome "similar Web pages" ad-on Suggested Similar Pages	**Cosine Similarity**
http://www.4guysfromrolla.com/	0.2210709
https://www.aspfree.com/	0.1533101
http://www.devx.com/	0.05344735
http://www.wrox.com/WileyCDA/	0.09793852
https://www.mysql.com/	0.04211563
http://dotnetslackers.com/	0.1630901
http://www.aspmessageboard.com/	0.0843255
https://bytes.com/	0.07697085
https://www.devexpress.com/	0.1508854
http://aspalliance.com/	0.2510862
Average	**0,12942406**

Experiment Number: 4 **Current Page:** https://programming.com

System Representative Words for Current Web Page

sql[1]	*share*[2]	*manag*[3]	*html*[4]	*mongodb*[5]
codeandy[6]	*forum*[7]	*php*[8]	*java*[9]	*web*[10]

System Suggested Similar Pages	**Cosine Similarity**
https://msdn.microsoft.com/en-us/library/ms140203.aspx	0,1445813
http://www.sqlmanager.net/en/products/manager	0,0657774
http://www.windowsnetworking.com/articles-tutorials/windows-server-2008/Windows-2008-Share-Storage-Management-Tool.html	0,1280211
http://samsung-pc-share-manager.en.lo4d.com/	0,0604795
http://www.w3schools.com/SQl/default.asp	0,0795934
http://www.sqlcourse.com/intro.html	0,0474377
http://www.netapp.com/us/products/management-software/snapmanager-sql.aspx	0,0654655
https://msdn.microsoft.com/en-us/library/hh759341.aspx	0,1340521
http://stackoverflow.com/questions/12881455/sql-server-database-on-network-share	0,1112169
https://technet.microsoft.com/en-us/library/ms365247(v=sql.105).aspx	0,1120148
Average	**0,094864**
Chrome "similar Web pages" ad-on Suggested Similar Pages	**Cosine Similarity**
http://www.techxtend.com/content.aspx?name=solutions-programmers-paradise	0.0200839
Average	**0.0200839**

REFERENCES

[1] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of Massive Data Sets*, Cambridge, England: Cambridge University Press, 2014, ch. 3, pp. 73-128.

[2] T. Srivastava, P. Desikan, and V. Kumar, *Foundations and Advances in Data Mining*, Berlin Heidelberg, Germany: Springer, 2005, ch. 10, pp. 275-307.

[3] F. Meziane and Y. Rezgui, "A document management methodology based *on* similarity contents," *Information Sciences*, vol. 158, pp. 15-36, Jan. 2004.

[4] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search *engine*," in *Proc. the Seventh International Conference on World Wide Web 7*, 1998, pp. 107-117.

[5] J. Huang and E. N. Efthimiadis, "Analyzing and evaluating query *reformulation* strategies in Web search logs," in *Proc. the 18th ACM Conference on Information and Knowledge Management*, 2009, pp. 77-86.

[6] R. L. *Cecchini*, C. M. Lorenzetti, A. G. Maguitman, and F. Menczer, "A semantic framework for evaluating topical search methods," *CLEI Electronic Journal*, vol. 14, 2011.

[7] A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on Web-page clustering," *Workshop of Artificial Intelligence for Web Search*, 2000, pp. 58-64.

[8] B. B. Cambazoglu and R. B. Yates, "Scalability challenges in web search engines," *Advanced Topics in Information Retrieval*, vol. 33, pp. 27-50, 2011.

[9] M. Hagen and C. Glimm, "Supporting more-like-this information needs: finding similar Web content in different scenarios, information access evaluation. Multilinguality, multimodality, and interaction," *Lecture Notes on Computer Science*, vol. 8685, pp. 50-61, 2014.

[10] K. S. Esmaili and H. Abolhassani, "A categorization schema for semantic web search enginees," *IEEE International Conference on Computer Systems and Applications*, 2006, pp. 171-178.

[11] M. R. Henzinger, R. Motwani and C. Silverstein, "Challenges in web *search* engines," *Newsletter ACM SIGIR Forum*, vol. 36, pp. 11-22, 2002.

[12] R. B.-Yates, "Algorithmic challenges in Web search enginees, experimental algorithms," *Lecture Notes on Computer Science*, vol. 4007, pp. 277-278, 2006.

[13] G. Attardi, A. Gull ì, and F. Sebastiani, "Automatic web page *categorization* by link and context analysis," in *Proc. THAI-99, 1st European Symposium on Telematics, Hypermedia and Artificial Intelligence*, 1999, pp. 105-119.

[14] K. Rehman and M. N. A. Khan, "The foremost guidelines for achieving higher ranking in search results through search engine optimization," *International Journal of Advanced Science and Technology*, vol. 52, pp. 101-110, 2013.

[15] Nisha and P. Singh, "A review paper on SEO based ranking of web documents," *International Journal of Advanced Search in Computer Science and Software Engineering*, vol. 4, pp. 1136-1140, 2014.

[16] S. Courtenage and S. Williams, "Finding relevant web pages through equivalent hyperlinks," in *Proc. 3rd International Workshop on Web Dynamics (WWW2004)*, 2004, pp. 24-31.

[17] Z. Ma, G. Pant, and O. R. L. Sheng, "Interest-based personalized search," *ACM Transactions on Information Systems*, vol. 25, article 5, 2007.

[18] P. Thwe, "Using markov model and popularity and similarity-based page rank algorithm for Web page access," in *Proc. International Conference on Advances in Engineering and Technology (ICAET'2014)*, pp. 197-201, 2014.

[19] J. Hou, Y. Zhang, and J. Chao, "Web page clustering: A hyperlink-based similarity and matrix-based hierarchical algorithms," *APWeb 2003*, vol. 2642, pp. 201-212, 2003.

[20] D. Donato, S. Leonardi, and P. Tsaparas, "Stability and similarity of link analysis ranking algorithms," *Internet Mathematics*, vol. 3, pp. 479-507, 2006.

[21] J. Hou and Y. Zhang, "Utilizing hyperlink transitivity to improve web page clustering," in *Proc. the 14th Australasian Database Conference*, 2003, pp. 49-57.

[22] A. Kritikopoulos, M. Sideri, and I. Varlamis, "WORDRANK: A method for ranking Web pages based on content similarity," in *Proc. 24th British National Conference on Databases (BNCOD'07)*, 2007, pp. 92-100.

[23] M. *Alpuente* and D. Romero, "A visual technique for web page comparison," *Electronic Notes in Theoretical Computer Science (ENTCS)*, vol. 235, pp. 3-18, 2009.

[24] A. H. Kulkarni and B. M. Patil, "Template extraction from *heterogeneous* web pages with cosine similarity," *International Journal of Computer Applications*, vol. 87, pp. 4-8, 2014.

[25] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring semantic similarity between words using Web search enginees," in *Proc. International World Wide Web Conference Committee*, 2007, pp. 757-766.

[26] S. Rani and U. Garg, "A ranking of web documents using semantic *similarity* and artificial intelligence based search engine," *International Journal of Science, Engineering and Technology Research*, vol. 3, pp. 3354-3357, 2014.

[27] M. Radovanovic and M. Ivanovic, "Text mining: Approaches and applications," *Novi Sad J. Mat*., vol. 38, pp. 228-234, 2008.

[28] T. *Grigalis* and A. Čenys, "Unsupervised structured data extraction from template-generated Web pages," *Journal of Universal Computer Science*, vol. 20, pp. 169-192, 2014.

[29] D. A. Popescu and D. Radulescu, "Approximately similarity measurement of web sites," *Neural Information Processing*, vol. 9492, pp. 624-630, 2015.

[30] T. H. Haveliwala, A. Gionis, D. Klein, and P. Indyk, "Evaluating *strategies* for similarity search on the Web," in *Proc. the 11th international conference on World Wide Web*, 2002, pp. 432-442.

[31] W.-T. Yih and C. Meek, "Improving similarity measures for short segments of text," in *Proc. AAAI 2007*, 2007, pp. 1489-1494.

[32] H. H. Duan, V. G. Pestov, and V. Singla, "Text categorization via *similarity* search: An efficient and effective novel algorithm, similarity search and applications," *Lecture Notes in Computer Science*, vol. 8199, pp. 182-193, 2013.

[33] H. Y. Zhang, "CrossCheck: An effective tool for detecting plagiarism," *Learned Publishing*, vol. 23, pp. 9-14, 2010.

[34] M. K. Dalal and M. A. Zaveri, "Heuristics based automatic text summarization of unstructured text," in *Proc. International Conference and Workshop on Emerging Trends in Technology*, 2011, pp. 690-693.

[35] F. Ren, S. Li, and K. Kita, "Automatic abstracting important sentences of web articles," in *Proc. 2001 IEEE International Conference on Systems, Man, and Cybernetics*, vol. 3, pp. 1705-1710, 2001.

Metin Turan was born in İstanbul. He graduated from the Computer Science Department of the Hacettepe University in Ankara. He worked as a research assistant when he completed an MSc at the same department. He got his PhD degree from Computer Engineering Department of the Yıldız Technical University in 2015. The major research areas of author are artificial intelligence, NLP, programming languages, game programming, data mining, software engineering and image processing.

He has been experienced in analyzing, designing, programming and project management more than 15 years of sectoral work. He worked nine years as lecturer at computer engineering department of the İstanbul Kültür University. He was department head of computer engineering at the İstanbul Nişantaşı University. He is currently a member of computer engineering department of the Istanbul Commerce University. He published 6 international and 3 national research articles.