A Product Search Engine Supporting "Best Product" Queries

Furkan GÖZÜKARA*[1], Selma Ayşe ÖZEL[1]

[1]*Çukurova University, Faculty of Engineering and Architecture, Computer Engineering Department, Adana*

Abstract

In this study, a novel product search engine system which supports "find the best products for a given category" type queries is proposed. The product search engine system consists of a focused crawler, a record linkage system, a sentiment analyzer, and a query engine system. The focused crawler is used to crawl product information from various e-commerce sites; the record linkage system determines the identical products that are crawled from different e-commerce sites; the sentiment analyzer classifies users' reviews about the products as positive or negative so that our product search engine can decide which product is the best for a given category; and the query engine takes the user queries and displays the result. All implementations are done by using C# programming language in .NET 4.5 framework, and MS-SQL Server 2014 database management system is employed for data storage. The core of our system is the record linkage part which is based on a modified incremental Hierarchical Agglomerative Clustering algorithm. To improve the success of record linkage process we also develop a product code matching system such that if the two products from different e-commerce sites have the same product code they are considered as the same. In our experimental analysis we observe 96.25% F-measure in record linkage of E-commerce products and 100% precision in most related products search. Our system can successfully offer best products for a given category. The proposed system achieves to provide better user experience than the existing systems.

"En İyi Ürün" Sorgularını Destekleyen Bir Ürün Arama Motoru

Öz

Bu çalışmada, "verilen bir kategori için en iyi ürünleri bul" tarzındaki sorguları destekleyen özgün bir ürün arama motoru sistemi önerilmektedir. Geliştirilen ürün arama motoru sistemi, bir odaklı tarayıcı, bir kayıt eşleştirme sistemi, bir duygu analizi sistemi ve bir sorgu motoru sisteminden oluşmaktadır. Odaklı tarayıcı sistemi çeşitli e-ticaret sitelerindeki ürün bilgilerini elde etmek için kullanılmaktadır; kayıt eşleştirme sistemi farklı e-ticaret sitelerinde satılan aynı ürünleri tespit etmektedir; duygu analizi sistemi ürünlere yapılan kullanıcı yorumlarını olumlu veya olumsuz olarak sınıflandırmakta ve bu sınıflandırma hangi ürünlerin aranan kategori için en iyi ürün olduğunu belirlemek için kullanılmaktadır ve sorgu motoru ise kullanıcıların sorgusunu alıp kullanıcılara sonuçları göstermektedir. Bütün sistem C#

---

* Sorumlu yazar (Corresponding author): Furkan GÖZÜKARA, *furkangozukara@gmail.com*

programlama dilinde .NET 4.5 alt yapısı kullanılarak geliştirilmiştir ve veri depolamak için MS-SQL 2014 veri tabanı yönetim sistemi kullanılmıştır. Önerilen sistemin temelini artımsal olacak şekilde değiştirilmiş Hiyerarşik Aglomeratif Kümeleme algoritmasına dayanan kayıt eşleştirme sistemi oluşturmaktadır. Kayıt eşleştirme işleminin başarısını artırmak için ürün kodu eşleştirme sistemi geliştirilmiştir. Bu sistem farklı e-ticaret sitelerinde satılan ismi farklı olarak yazılmış ancak aynı ürün koduna sahip ürünleri tespit edebilmektedir. Deneysel analizlerimiz sonucunda, e-ticaret ürünlerinin kayıt eşleştirmesinde %96,25 F-ölçeği ve en alakalı ürünler aramasında %100 kesinlik elde edilmiştir. Geliştirilen sistem verilen kategori içinde en iyi ürünleri başarılı bir şekilde kullanıcıya sunabilmektedir. Önerilen sistem mevcut sistemlere göre daha başarılı bir kullanıcı deneyimi sunabilmektedir.

Anahtar Kelimeler: Duygu analizi, Odaklı tarayıcı, Kayıt eşleştirme, Karşılaştırmalı alışveriş ajanı, Ürün sıralama

1. INTRODUCTION

For the last 2 decades, the rapid spread of usage of the Web and the advances in Internet technology brought many commercial opportunities along with it, both in Turkey and in the world. Interactive product selling can be said as one of the biggest opportunity. The volume of interactive buy/sell is exponentially increasing and it has become a common part of the trade business. The interactive buy/sell of products over electronic systems such as Internet is described as E-commerce (Electronic Commerce) [1]. The E-commerce shopping is becoming an irrevocable part of our lives as each year passes. With the emerging new technologies, and the increase in the technological education level of the consumers, it is now much easier to do E-shopping than before. According to the Turkey's official BKM (Interbank Card Center) statistics [2], the electronic shopping volume in Turkey has increased to 50.682.910.000 TL in 2015 from 16.553.790.000 TL in 2011. This massive increase in transaction volume attracts many businesses to start operating in E-commerce area. Thus, many new E-commerce websites are starting to operate continuously. Due to the stiff competition among E-commerce businesses, many companies also have to halt their operations. Additionally, because of the competition, the E-commerce websites constantly update their prices and make discounts or campaigns. Therefore, it is quite not possible for average consumers to keep track of numerous E-commerce websites when going to make an online purchase. To provide best prices for an online product and the other features of the product to the consumers, CSA (Comparison Shopping Agent) systems are developed [3]. These systems collect the necessary information from E-commerce websites and provide the processed structured information to the consumers. Because of these stated reasons, in this study we present a full scale CSA system for Turkish E-commerce websites. Our proposed system crawls 50 different E-commerce sites in Turkey to collect product information. After that our developed record linkage system clusters the same products from different sites, then we extract title, price, category, and user reviews for each unique product. Our sentiment analysis system classifies users' reviews for each product and according to positive and negative reviews we rank the products for each product category. Finally the query engine of our system takes user query and returns the "most related" and "best" products with their price and reviews for the given category. To develop our product search engine, we design our own focused crawler, sentiment analyzer, record linkage, and product ranking metric, and we observe performance improvement in the developed systems with respect to state of the art systems.

The rest of the paper is organized as follows: in the next section a summary for the related work about CSA systems, focused crawlers, sentiment analysis, and record linkage are presented. In the third section we briefly explain the methods used in our system. The fourth section includes experimental evaluation results, and finally section five concludes the study.

2. RELATED WORK

2.1. CSA Systems

The first widely recognized comparison agent is Bargain Finder [4] which is a focused CSA system that scans only online CD (Compact Disc) stores. It uses internal search system of the E-commerce websites in order to display results to the users. When a user submits a query, it transforms query into the internal searching structure of the online CD stores and then submits query to the registered online stores in its system. When the results return from the online CD stores, it integrates results and then displays to the users. It does not utilize any record linkage system. There exist numerous CSA studies in the literature such that [5,6] conducted by Guttman et al. and Sadeddin et al. Moreover, Pathak has published a detailed survey about the CSA systems [3].

2.2. Focused Web Crawlers

Web crawlers are software systems that are used to collect information from websites. Their main task is fetching websites, processing fetched source code and extracting new target hyperlinks to crawl. General web crawlers start with root URLs and continue until crawling all of the hyperlinks that they can find. However, this task is tedious and not doable for small or even average scale applications. Only a few large-scale commercial general search engines (e.g., Google, Bing, Yahoo, Yandex, and so on) can cope with the challenges and the massiveness of the entire Web and keep their index fresh. Therefore, developing focused Web crawlers are much more feasible and commonly practiced. Focused Web crawlers are specialized versions of general Web crawlers that crawl only certain topics or certain Web sites [7]. Even though they are much smaller scale than general Web crawlers, still many challenges and tough tasks await the developers who are going to build focused Web crawlers [8]. Studies of Heydon and Najork [9], Shkapenyuk and Suel [10], Boldi et al. [11], and Gomes and Silva [12] include how to design a general web crawler; studies of Shkapenyuk and Suel [10], Yohanes et al. [13], and Liu and Milios [14] discuss how to decide

more relevant pages for a focused web crawler to reduce resource requirements such as space, bandwidth, computation power, etc., in the crawling task.

2.3. Sentiment Analysis

With the emerging new technologies, today, electronic platforms such as internet blogs, E-commerce Web sites, digital newspapers, Facebook, Twitter, online forums, etc. become crucial for our lives. People not only read these platforms, but also leave their opinions about the related subjects. The analysis of these opinions has become a major interest recently in the NLP (Natural Language Processing) field. Sentiment analysis or opinion mining is the computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes [15]. There has been much research in the opinion mining and sentiment analysis area recently. There are three recent major surveys made by Pang and Lee [16], Liu and Zhang [15], and Vinodhini and Chandrasekaran [17].

2.4. Record Linkage

Grouping same entities from different sources into the same cluster is known as record linkage [18] in the literature and has been widely researched. As the record linkage task is heavily domain dependent, there exist numerous different algorithms proposed by Tejada et al. [19], Jin et al. [20], Yan et al. [21], Köpcke and Rahm [22], and Christen [23] for this task. Supervised (i.e., classification), unsupervised (i.e., clustering), and mixture of supervised and unsupervised methods are generally used for record linkage tasks. Many of the previous works are done on single category entities; for example, only on restaurants, or only on authors, or only on electronic devices, etc. However, in our study we have infinitely many number of categories such as video cameras, led TVs, smartphones, hard disks, printers, fans, grass mowers, perfumes, etc. and we try to develop a record linkage system to deal with all these categories.

3. ARCHITECTURE OF THE PRODUCT SEARCH ENGINE

The proposed product search engine consists of four main components that are i) focused crawler, ii) sentiment analyzer, iii) record linkage system, and iv) query processor. The details for each components are explained in the below subsections.

3.1. The Focused Crawler

The proposed focused web crawler is used to crawl and extract information from 50 Turkish e-commerce web sites including hepsiburada, vatanbilgisayar, gittigidiyor, n11, trendyol, markafoni, etc. The crawler system starts crawling process from the seed URLs of the e-commerce sites, downloads and processes the Web pages,

extracts product information from product pages, and generates new URLs to be crawled. This process continues until no new pages from the e-commerce sites are found. Estimated numbers of Web pages in some e-commerce sites are presented in Table 1. As can be seen from the Table 1, there are thousands of Web pages to be crawled and processed. To achieve this goal our focused Web crawler is implemented in multi-threaded fashion and for each e-commerce site up to 50 concurrent connections have been made to provide both politeness for the e-commerce servers and fast crawling of the pages. Moreover, we develop some rules to distinguish product pages from other pages such that if a Web page satisfies the rules, the page is a product page and we extract product title, price, category, and users' reviews for the product and store these information in a database for further processing.

Table 1. Estimated number of Web pages to be crawled for some e-commerce sites

E-commerce Site	Number of Web pages	E-commerce Site	Number of Web pages
Gittigidiyor	788,000	Teknosa	1,070,000
N11	781,000	Markafoni	473,000
Hepsiburada	778,000	Sanalpazar	552,000
Vatanbilgisayar	3,230,000	Kitapyurdu	627,000
Trendyol	570,000	Tozlu	1,110,000

3.2. The Sentiment Analyzer

We use sentiment analysis to determine the polarity of the users' reviews for each product, therefore try to rank products for a given category and be able to answer "best product" queries. The crawled user reviews are classified by using an SVM classifier. To train the classifier, we manually labeled some of the users' reviews and obtain 600 positive and 600 negative comments. Then we apply lowercase conversion, transform diacritics, remove punctuations and non-ASCII characters. We do not remove stopwords and apply stemming as they decrease classification performance. After these preprocessing steps, the user reviews are converted to bag-of-words format and vectors are formed. Classifier is trained by using the labeled reviews then other reviews can be labelled by the trained model.

3.3. The Product Record Linkage System

As we crawl product information from different e-commerce Web sites, we need to identify the same products from different sources. As each e-commerce site can give different names to the same products this identification process is not easy. For example, while "Sony Xperia Z2 Cep Telefonu" and "Telefon Sony Z2 Xperia" refer to the same products, "Sony Xperia Z2 Cep Telefonu" and "Sony Xperia M2 Dual Cep Telefonu" are completely different products and in this study we need to successfully distinguish these two cases. To identify the same products from different sources, we propose a record linkage system which is the core part of this study.

The proposed record linkage system is based on a hierarchical agglomerative clustering algorithm and works in incremental fashion. Our system uses

ÖS 262

Ç.Ü. Müh. Mim. Fak. Dergisi, 31(ÖS 2), Ekim 2016

some domain expert defined rules while clustering the products. First of all prices of the products are compared, if the ratio of the prices of the two products are greater than a threshold value (e.g., 1.5) the two products can not be the same. If the products pass the price test, then their brands are compared. If both of the two products have brand information and the brands are different, they can not be the same product. If the products also pass the brand test then word conflict check is made. For example "*pc*" and "*ps3*" are conflicting words. The first one is used to define computer games while the second one defines PlayStation games. As an example, products having titles "PC DRAGON AGE INQUISITION" and "PS3 DRAGON AGE INQUISITION" are different although they seem very similar. If the two product titles have conflicting words, then they are considered as different. The last test that is applied is product code test. If the two product titles have product codes and these codes are different, then these two products can not be the same. We also propose an algorithm for product code detection and our algorithm can successfully detect product codes as in titles "18.5 PHILIPS **193V5LSB262** 5MS SİYAH LED MONİTÖR" and "19.5 PHILIPS **203V5LSB26** 5MS SİYAH LED MONİTÖR" and our record linkage system can decide that these two products are different.

3.4. The Query Engine and User Interface

The user interface allows user to enter his query and shows the results. Two different query types are supported by our system. The first query type is used to search for the most related products according to the user input as shown in Figure 1. As an example when user enters the search term "photograph (fotoğraf)" and clicks on the "Search a Product" button, the search terms are queried from the database and the products having the titles which includes the search term are displayed. The other products that are at the same cluster with the found ones are also listed. As shown in the Figure 1, minimum and maximum price information, and number of positive and negative comments for the products are also displayed. User is allowed to see the details of the products by clicking on the product names as shown in Figure 2.
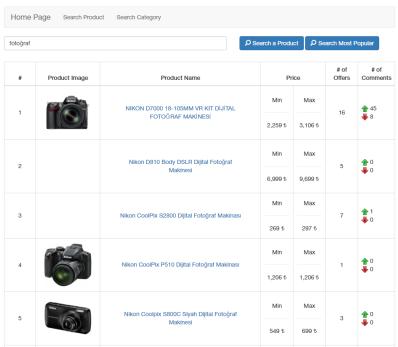

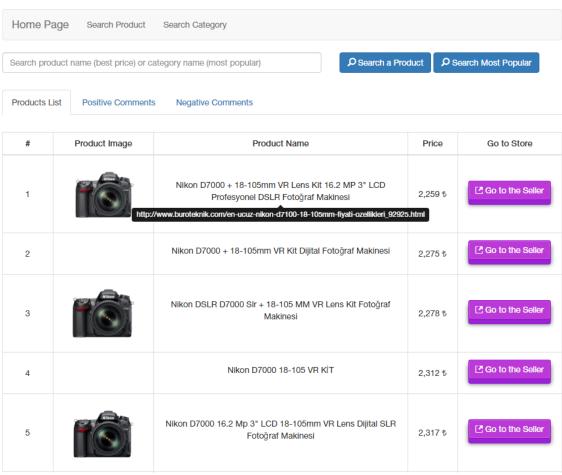**Figure 1.** Most related products search result in our proposed CSA system

#	Product Image	Product Name	Price	Go to Store
	Home Page Search Product Search Category			
	Search product name (best price) or category name (most popular)		🔎 Search a Product	🔎 Search Most Popular
	Products List Positive Comments Negative Comments			
1		Nikon D7000 + 18-105mm VR Lens Kit 16.2 MP 3" LCD Profesyonel DSLR Fotoğraf Makinesi http://www.buroteknik.com/en-ucuz-nikon-d7100-18-105mm-fiyati-ozellikleri_92925.html	2,259 ₺	↗ Go to the Seller
2		Nikon D7000 + 18-105mm VR Kit Dijital Fotoğraf Makinesi	2,275 ₺	↗ Go to the Seller
3		Nikon DSLR D7000 Slr + 18-105 MM VR Lens Kit Fotoğraf Makinesi	2,278 ₺	↗ Go to the Seller
4		Nikon D7000 18-105 VR KİT	2,312 ₺	↗ Go to the Seller
5		Nikon D7000 16.2 Mp 3" LCD 18-105mm VR Lens Dijital SLR Fotoğraf Makinesi	2,317 ₺	↗ Go to the Seller

Figure 2. Details of a product

The second query type is to find the "best product". As shown in Figure 3, when user enters search terms "graphic card (ekran kartı)" and clicks on the "Search Most Popular" button, the product categories that matches with the user query are fetched. Then, the sub-categories of the previously found categories are also retrieved. After that, all products that belong to the found categories and their clusters are searched. All products that belong to these clusters are also found and displayed. Then, all comments for these products are retrieved, popularity scores of all clusters are calculated, and the clusters are presented to the user according to the popularity score in descending order.

The popularity score is computed according to equation 1, where $scoreP_i$ is the popularity score of the $i\text{-}th$ cluster, C_{pi} is the total number of positive comments in the $i\text{-}th$ cluster, C_{ni} is the total number of negative comments in the $i\text{-}th$ cluster, S_i is the number of different E-commerce stores in the $i\text{-}th$ cluster.

$$scoreP_i = (C_{pi} - C_{ni}) \times S_i \qquad (1)$$

User can also see positive and negative comments made for that product as shown in Figure 4. These comments are cumulatively collected from all of the E-commerce stores that particular product is sold.
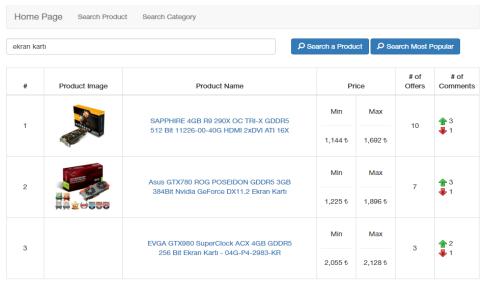
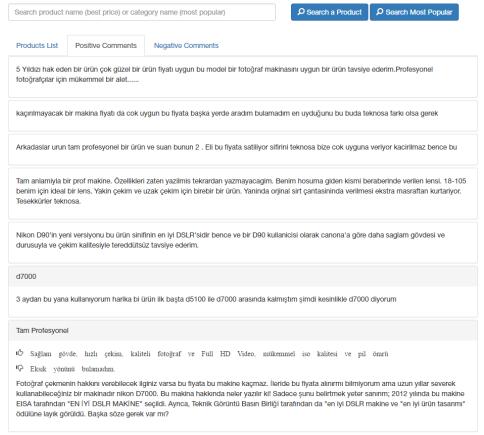**Figure 3.** Screenshot of search best products results

Figure 4. Screenshot of how positive comments of a product are displayed

4. EXPERIMENTAL EVALUATION

The success of our system highly depends on the performance of the record linkage system. Therefore we perform several experiments to evaluate the performance of our record linkage algorithm. We collect a dataset from our crawled Web pages for performance evaluation experiments and this dataset is available on https://github.com/FurkanGozukara/Record-Linkage. Our system has 96.25% F-Measure value while the best state-of-the-art algorithm has 89.12% F-Measure for the dataset. For more information we refer to the study [24].

The performance of the proposed search engine is evaluated by making some queries and computing the top-n precision of the results. We compare our system with most popular 7 commercial CSA websites in Turkey by submitting 5 different queries. For calculating the precision, we check whether the top 10 returned results are related to the query or not. The results of this experiment are presented in Table 2.

Table 2. Top-10 precision of the CSA systems for the given queries

Query	Top-10 precision of the CSA systems							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Nikon D810	100%	100%	100%	N/A	N/A	N/A	50%	100%
Dishonored Pc Oyun	100%	N/A	100%	N/A	N/A	N/A	0%	100%
Canon CLI-551	100%	100%	100%	N/A	N/A	N/A	100%	100%
Asus Xonar DSX 7.1	100%	N/A	100%	N/A	N/A	N/A	33%	100%
Fakir Buharlı Ütü	100%	100%	N/A	N/A	N/A	N/A	0%	100%

In Table 2, the query column lists the queries made on the CSA systems. Column numbers from 1 to 8 are used to give short names to the CSA systems. Names of the CSA systems that are used in the experiment are listed as follows: **1** refers to www.akakce.com, **2** is www.cimri.com, **3** means www.teknofiyat.com, **4** is www.iyi.net, **5** is used for www.uygunfiyat.com, **6** represents www.efiyat.com, **7** refers to www.fiyatarat.com, and finally **8** is used for our system. As shown in Table 2, akakce.com and our system have performed 100% precision for all queries. Teknofiyat.com has performed 100% precision for four queries, cimri.com has performed 100% precision for 3 queries; iyi.net, uygunfiyat.com, and efiyat.com can not give relevant results to any of the queries; and fiyatarat.com has performed with 36.6% precision on the average for all queries.

For analyzing performance of "best products in a category" query of our system, we compare our system with Google as other CSA systems do not have support for this query type. For the query "best graphic card (en iyi ekran kartı)", the top result returned by Google is the sponsored results as shown in Figure 5. Basically they are the ads promoted by the E-commerce Web sites. Moreover, two results of the displayed six results are not even graphic cards as shown in Figure 6. The two results of the displayed four graphic cards are professional graphic cards and most of the users who look for graphic cards would not be interested in them. CSA systems cannot be queried with such "best product" queries. However, CSA systems like akakce.com support categorical listing and in these listings, they sort the results by their internal ranking mechanisms such as in our search engine as shown in Figure 3. The 2nd, 10th and 11th results of Google lead to category listing of the E-commerce stores in Turkey. The 3rd, 6th and 8th results are journal posts about best graphic cards however, they are heavily outdated. The 4th result is a video about best graphic cards and again it is outdated as well. The 5th result is pictures of graphic cards. The 7th and 9th results of Google are about what consumers should consider when purchasing a graphic card. However, when a user query our system to find out best products in a category, our system always provide the most updated information and rank them according to our proposed ranking mechanism as shown in Figure 3.

Figure 5. The top result of Google "best graphic card (en iyi ekran kartı)" search

Figure 6. Google best graphic cards query

Figure 6. Google best graphic cards query (Continued)

5. CONCLUSION

In this study we present the main components of a product search engine which allows finding products on a given category, listing the prices from different vendors, and ranking the products according to user reviews. The core of the system is the record linkage part and the proposed record linkage algorithm clearly outperforms the similar algorithms that are previously designed. The product code detection along with the price matching, brand matching, and conflict words matching rules that we have proposed significantly improve the success of the record linkage task. We also show that the proposed system is successful by submitting queries to the system and other well-known CSA systems.

6. ACKNOWLEDGEMENTS

7. REFERENCES

1. Chen, S.C., Dhillon, G.S., 2003. Interpreting Dimensions of Consumer Trust in E-Commerce, Information Technology and Management, 4:2-3, p. 303-318.
2. http://bkm.com.tr/en/internetten-yapilan-kartli-odeme-islemleri. Accessed, 23.05.2016.
3. Pathak, B., 2010. A Survey of the Comparison Shopping Agent-Based Decision Support

Systems, Journal of Electronic Commerce Research, 11:3, p. 178-192.

4. Krulwich, B., 1996. Information Integration Agents: Bargainfinder and Newsfinder, In Internet-Based Information Systems: Papers From The 1996 AAAI Workshop, Vol. 79.

5. Guttman, R., Moukas, A., Maes, P., 1999. Agents as Mediators in Electronic Commerce, In Intelligent Information Agents, 131-152.

6. Sadeddin, K.W., Serenko, A., Hayes, J., 2007. Online Shopping Bots for Electronic Commerce: The Comparison of Functionality and Performance, International Journal of Electronic Business, 5:6, p. 576-589.

7. Chakrabarti, S., Van Den Berg, M., Dom, B., 1999. Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery, Computer Networks, 31:11, p. 1623-1640.

8. Gupta, S.B., 2012. The Issues and Challenges with the Web Crawlers, International Journal of Information Technology & Systems, 1:1, p. 1-10.

9. Heydon, A., Najork, M., 1999. Mercator: A Scalable, Extensible Web Crawler, World Wide Web, 2:4, p. 219-229.

10. Shkapenyuk, V., Suel, T., 2002. Design and Implementation of a High-Performance Distributed Web Crawler, In Data Engineering, 2002. Proceedings. 18th International Conference, 357-368.

11. Boldi, P., Codenotti, B., Santini, M., Vigna, S., 2004. Ubicrawler: A Scalable Fully Distributed Web Crawler, Software: Practice and Experience, 34:8, p. 711-726.

12. Gomes, D., Silva, M.J., 2008. The Viúva Negra Crawler: An Experience Report, Software: Practice & Experience, 38:2, p. 161-188.

13. Yohanes, B.W., Handoko, H., Wardana, H.K., 2013. Focused Crawler Optimization Using Genetic Algorithm. TELKOMNIKA (Telecommunication Computing Electronics and Control), 9:3, p. 403-410.

14. Liu, H., Milios, E., 2012. Probabilistic Models for Focused Web Crawling, Computational Intelligence, 28:3, p. 289-328.

15. Liu, B., Zhang, L., 2012. A Survey of Opinion Mining and Sentiment Analysis, In Mining Text Data, 415-463.

16. Pang, B., Lee, L., 2008. Opinion Mining and Sentiment Analysis, Foundations and Trends In Information Retrieval, 2:1-2, p. 1-135.

17. Vinodhini, G., Chandrasekaran, R.M. 2012. Sentiment Analysis and Opinion Mining: A Survey, International Journal, 2:6, p. 282-292.

18. Fellegi, I.P., Sunter, A.B. 1969. A Theory for Record Linkage, Journal of the American Statistical Association, 64:328, p. 1183-1210.

19. Tejada, S., Knoblock, C.A., Minton, S., 2002. Learning Domain-Independent String Transformation Weights for High Accuracy Object Identification, In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 350-359.

20. Jin, L., Li, C., Mehrotra, S., 2003. Efficient Record Linkage in Large Data Sets, In Database Systems for Advanced Applications, 2003. (DASFAA 2003). Proceedings. Eighth International Conference, 137-146.

21. Yan, S., Lee, D., Kan, M.Y., Giles, L.C., 2007. Adaptive Sorted Neighborhood Methods for Efficient Record Linkage, In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, 185-194.

22. Köpcke, H., Rahm, E., 2010. Frameworks for Entity Matching: A Comparison, Data & Knowledge Engineering, 69:2, p. 197-210.

23. Christen, P., 2012. A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication, IEEE Transactions on Knowledge and Data Engineering, 24:9, p. 1537-1555.

24. Gözükara, F., Özel, S.A., An Incremental Hierarchical Clustering System for Record Linkage in E-Commerce Domain, In The Publishing Process.