MSC2010: 68P20, 68T50

© *S. D. Sulova*

CREATING GROUPS FOR MARKETING PURPOSES FROM WEBSITE USAGE DATA

Customer grouping and knowledge extraction for these groups are important to online businesses because it allows purposeful application of marketing techniques. Individuals can be personally served with the groups, depending on the identified interests and preferences. In this article, we suggest a way to identify and create user groups by processing website usage data. We use the logs stored in the server log data for the visit to a selected website and then retrieve and process the text content of the visited web pages. The approach is based on the technology for natural language processing and uses the methods for clustering of text documents. The experimental testing of this method is done with the software product RapidMiner and data from visits to a Bulgarian e-shop.

Keywords: text clustering, group, text mining, Logfile, RapidMiner.

DOI: 10.20537/vm170314

Introduction

When working with large amounts of data, similar objects are often grouped in homogeneous sets. This allows the mass of the data to be reduced and the process of analysis to be made easier. The making of individual groups of data that represent different sets of similar objects is called clustering.

Frequently in the marketing analyzes the following segmentation is used — dividing the customers into groups based on a criterion or an indicator in order to later be able to apply certain marketing impacts on the groups. This is why in this research we offer the usage of clustering to identify groups of users, to which a differentiated way of servicing and special offers, based on their interests, can be used.

The approach that is suggested in this article is based on website usage data and the application of technologies for natural language processing on the text contents of the visited web pages.

§ 1. Related Work

The extraction of knowledge from web sources in literature is known as Web Mining [1]. It's the way of applying technologies for extracting knowledge on web resources — documents, hyperlinks, tags, server log files etc. Even though this process is generally based on Data Mining technologies, the specifics in it are connected to the fact that the data is different, initially it's in an unstructured form and they need to be processed and structured [2].

Depending on the general types of Internet resources, which are used in the process of WM there are three main types of extracting knowledge (Fig. 1) [3, 4].

- Web content mining (WCM) — extracting of useful data from the content of web documents. It is known that in the Internet space there is a lot of documents, most of which are unstructured or semi-structured. By processing their content, the most essential part of the text can be extracted, to separate key concepts and identify associative relations. For example, using automatic classification and grouping web pages by their subjects the following things can be explored: what the visitors are searching for on a certain e-shop and to find specific patterns for the set group of people; finding models in web pages on the basis of descriptions of the products; analyzing messages with opinions from forums etc.

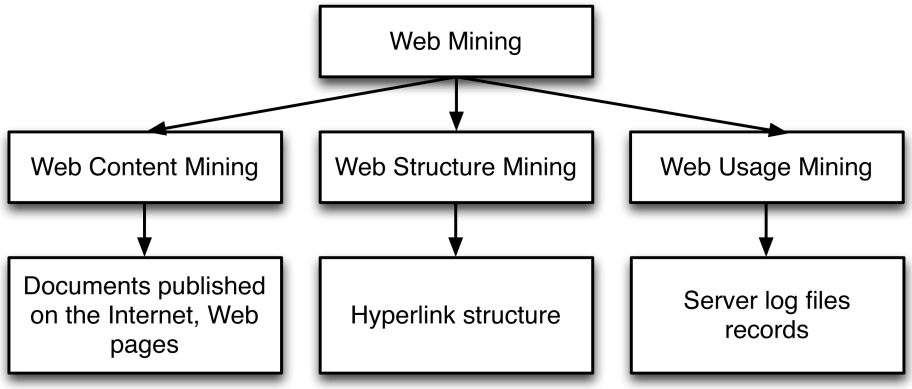

Fig 1. Web Mining types

- Web structured mining (WSM) — finding useful data in hyperlinks by using the structure of the incoming and outcoming connections (topology) in web. To do that the graph theory is applied to analyze a site and its link structure. In the last couple of years, the structure of web pages is widely used for analyzing significant information. The researches on web structure are influenced by the research for social networks and the analysis of quotations.
- Web usage mining (WUM) — discovering users' models based on the data for using Internet resources. It helps in finding templates in the flow of clicks and associated data, grouped or generated because of the users' interaction with websites. The records for the logging to servers are analyzed for the purposes of WUM and information about the sets of pages, objects or resources, which are often accessed by groups of users with common interests, is delivered.

For WCM, where data is retrieved from webpages, it is necessary to apply the natural language processing (NLP) technology to convert the text into a form suitable for analysis. The concept NLP 'is a field of computer science and linguistic concerned with the interactions between computers and human (natural) language' [5].

Different aspects of the processing of unstructured data and text are handled by many researchers. In general, the discovery of knowledge in unstructured text data in the literature is known as Text Mining (TM) [6]. The authors of [7] state that TM is mainly used for: 'information extraction, topic tracking, summarization, categorization, clustering, concept linkage, information visualization, and question answering'. Typical TM tasks are similar and Pena-Ayla describes them as: text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modelling [8].

Because our assignment is to identify groups for marketing purposes, we direct our attention to clustering, which allows searching for independent groups — clusters that are non-intermittent homogeneous sets of similar objects [4]. Grouping of data allows their size to be reduced and makes the analysis easier. For example, grouping can be applied to the clients. It's known that the users are different, they have different needs and requirements on the merchandize and like different ways of shopping. That is a premise for their segmentation and for the usage of differentiated methods when servicing them.

There are multiple algorithms for clustering, which in general can be divided into hierarchical and non-hierarchical [9]. Hierarchical clustering which includes agglomerative and divisive algorithms. In agglomeration methods, consecutive merging of units and clusters is performed. It starts with a few clusters representing the individual units and, after successive mergers, a cluster is brought together, integrating all units. In divisive methods, the approach is the opposite — it starts with a cluster that unites all units, and after successive divisions ends with many clusters, each unit forming a separate cluster. From existing cluster methods developed, hierarchical agglomeration methods are most commonly used. The results are represented by a tree diagram that graphically shows the

hierarchical structure received from the matrix of similarity and cluster merger rules. There are different strategies for merging objects into clusters and then to clusters themselves. The methods of the between-groups linkage and the nearest neighbor are used when creating clusters in the form of a 'chain'.

From the non-hierarchical algorithms the most distributed one is K-Means Cluster, by Hartigan and Wong [10]. In the beginning the chosen k random exit centers (points in space) and all objects are divided into k groups, depending on which center they are closest to. The distance proximity to centers is determined by one of the methods — Euclidean distance, squared Euclidean distance, Chebyshev distance, etc. After that the mew center of the cluster is calculated, based on the average values of the objects, resulting in the fact that some objects already fall into another cluster. The procedure is repeated recursively until the clusters' centers stop changing. This algorithm is used in a predefined number of clusters.

§ 2. Approach for identifying groups for marketing purposes

As already mentioned, for the purpose of our research we will focus our attention on analysis of already visited web pages and their grouping based on analysis of their textual content.

We suggest a model for identifying marketing groups, which is shown in Fig. 2.

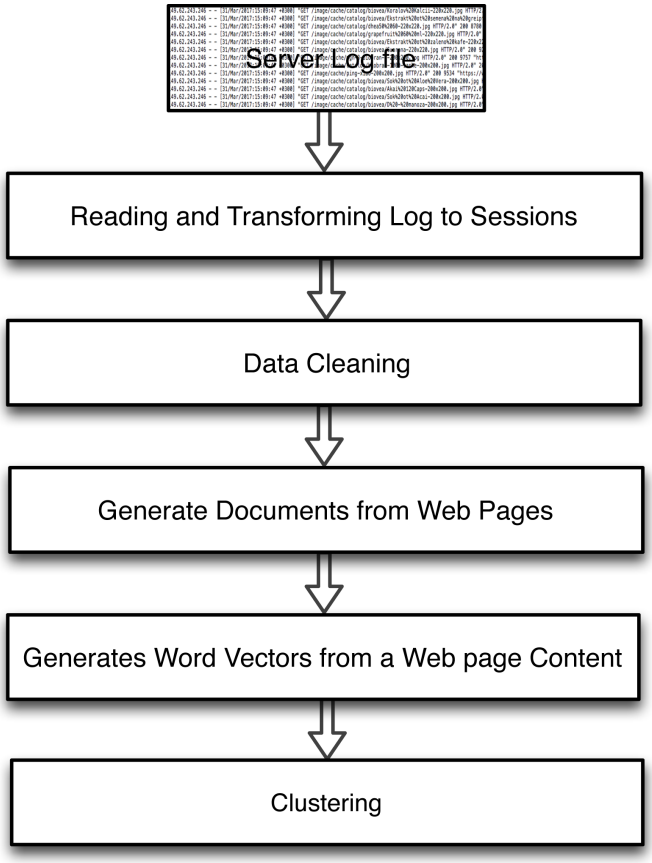

Fig 2. Model for identifying market groups by text clustering

The process of identifying market groups through a textual classification of the content of visited web pages includes the following steps:

1. Reading and transforming log to sessions. It's done by reading logfiles that are text files, also called journals or diaries, and in which data is stored for visits to a website. These files differ in their format, depending on the type of web server, but contain the following master data: the user's IP address; the time when the site is loaded, the address from which the user comes; the

type of browser and operating system used by the user, etc. Reading the logfiles and filing them in a tabular form is based on the knowledge of the format and type of text logfiles.

2. Data cleaning. Remove the irrelevant data from logfiles. Such are missing values and hyperlinks pointing to the following types of resources: gif, jpeg, video, audio, css, etc. That helps to significantly decrease the size of the logfiles [11].

3. Generating documents from web pages. At this stage, based on the saved web addresses of the visited webpages from each user session, textual documents with generated web text are generated.

4. Generating word vectors from a web page content. To prepare clustering data, a transformation from unstructured to structured format is needed. The so-called Vector space model (VSM) is used. In it each text document is represented as vectors [12].

To filter the multiple words from the unnecessary words, those who do not carry useful information at this stage the process text preprocessing is used, which often includes:

— tokenize — splits the text of a document into a sequence of tokens;
— filter tokens — filter tokens — based on their length;
— transform cases — transforms all characters in a document to lower ones;
— stem — replaces words with their basic forms;
— filter stop words — removes stopwords from a document.

In addition, some terms may be removed at this stage, the frequency of which is very small, and per researchers they are words that cannot be representative of the cluster they fall into [13].

5. Clustering. For clustering to take place, it is necessary to select a clustering algorithm to apply over the generated word vectors. Although there is virtually no 'best' clustering algorithm, based on multiple studies of text clustering [9,14,15] we can say that K-means clustering algorithm is an efficient algorithm for text clustering. The next step is to determine the number of clusters. There is no unambiguous rule for this task. Experiments with different number of groups are performed and the optimal option is chosen.

§ 3. Approbation and results

The approach presented in the previous section is tested using one of the most popular open source software for Data Mining — RapidMiner [16]. This software platform has multiple operators through which you can build and visualize processes. RapidMiner has a special add-on for text and web mining.

For experimental purposes, we use a log file by a Bulgarian e-shop which sells over 25 000 items — home goods, gifts and souvenirs, office supplies, school supplies, and more. The website's domain is in Bulgarian and so is the interface of the e-shop.

The identification of user groups is accomplished by following the steps proposed in paragraph 3. For the reading and transforming server log using RapidMiner we have built the model shown in Fig. 3.

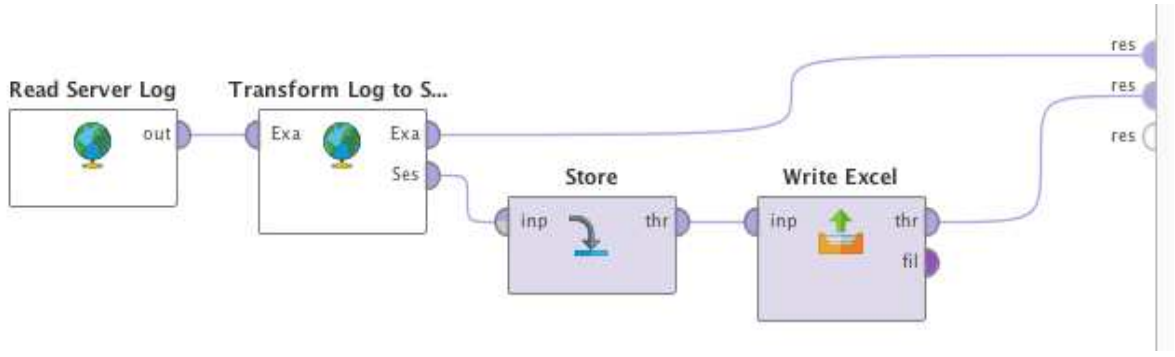

Fig 3. Model for reading and transforming server logfiles

To successfully recognize the logfile structure, we use a standard configuration file .xml file for the Apache web server as the parameter of the Read Server Log operator, in which the logfile structure is set. As a result, we get a structured file that is in the form of a two-dimensional table (Fig. 4).

Row No.	session	ip	agent	uri	referer	os_name	browser	language	time
2966	s233	213.226.6:	Mozilla/5.0	/index.php?route=pr	http://xn--80ace0ch.com/%D0%B0%	other	other	other	24050344
2967	s226	84.238.14:	Mozilla/5.0	/%D0%B7%D0%B0-%[	http://xn--80ace0ch.com/%D0%B7%	other	other	other	24050344
2968	s233	213.226.6:	Mozilla/5.0	/index.php?route=pr	http://xn--80ace0ch.com/%D0%B0%	other	other	other	24050344
2969	s233	213.226.6:	Mozilla/5.0	/index.php?route=pr	http://xn--80ace0ch.com/%D0%B0%	other	other	other	24050344
2970	s233	213.226.6:	Mozilla/5.0	/index.php?route=pr	http://xn--80ace0ch.com/%D0%B0%	other	other	other	24050344
2971	s233	213.226.6:	Mozilla/5.0	/index.php?route=pr	http://xn--80ace0ch.com/%D0%B0%	other	other	other	24050344
2972	s233	213.226.6:	Mozilla/5.0	/index.php?route=pr	http://xn--80ace0ch.com/%D0%B0%	other	other	other	24050344
2973	s144	188.165.1:	Mozilla/5.0	/%D0%B2%D1%8A%D(	?	other	other	other	24050344
2974	s233	213.226.6:	Mozilla/5.0	/index.php?route=pr	http://xn--80ace0ch.com/%D0%B0%	other	other	other	24050344
2975	s233	213.226.6:	Mozilla/5.0	/%D0%B0%D0%BA%D1	http://xn--80ace0ch.com/%D0%B0%	other	other	other	24050344
2976	s233	213.226.6:	Mozilla/5.0	/index.php?route=pr	http://xn--80ace0ch.com/%D0%B0%	other	other	other	24050344

Fig 4. Result of identifying visited web pages

We save the data in .xlsx format to make it easier to filter and remove the missing URL and address values pointing to gif, jpeg, png, and css resources. For each of the sessions, repeating IP addresses are also removed. After this Data Cleaning stage, the number of web pages from which the text content is to be extracted is significantly reduced. Virtually 1/3 of the originally identified and saved web pages remains. To retrieve, process, and group the text again, we use RapidMiner and the model built in Fig. 5.

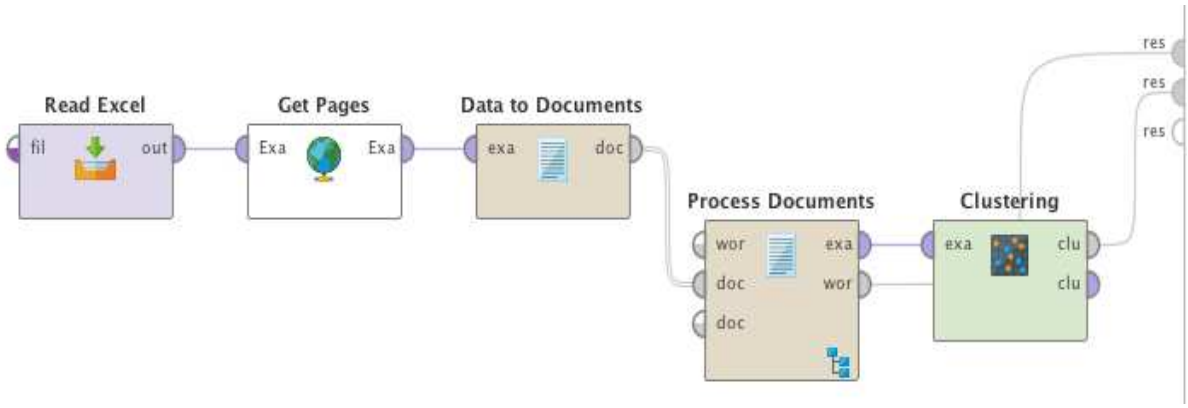

Fig 5. Model of the text extraction process and clustering

The Get Pages operator retrieves text from web pages, and Data to Documents is used to generate text documents from web pages. It should be noted that the received text documents contain as well as the text and the html code of the page. Because of the html tags, the attributes cannot extract essential information which can be useful for further text processing, it is necessary to filter the text and remove html business words. To do this, we use a vocabulary specially created by us with English words plus the words used by Hypertext Markup Language.

In the presentation of the text as word vectors (through Process Documents), we use word frequency calculation in the example case by TF–IDF (Term Frequency — Inverse Document Frequency) weighing. This indicator is statistics which shows how important the word for a collection of documents or corpus is. TF–IDF increases its value in proportion to the number of occurrences of the word, but also the frequency of the word in the body because some words as a rule appear more often in the texts.

Text processing includes the following sub-processes (Fig. 6):

- divides the text into words (Tokenize);
- removing words shorter with a length of less than or equal to 3 letters (Filter Tokens by Length);
- converting letters into small ones (Transform Cases);
- converting words into a normal form using Porter's algorithm (Stem Porter);
- deleting English terms, which are only html tags and other business words (Filter Stopwords with a specially created dictionary);
- erasing, etc. redundant words in Bulgarian (Filter Stopwords with a glossary of redundant words in Bulgarian).

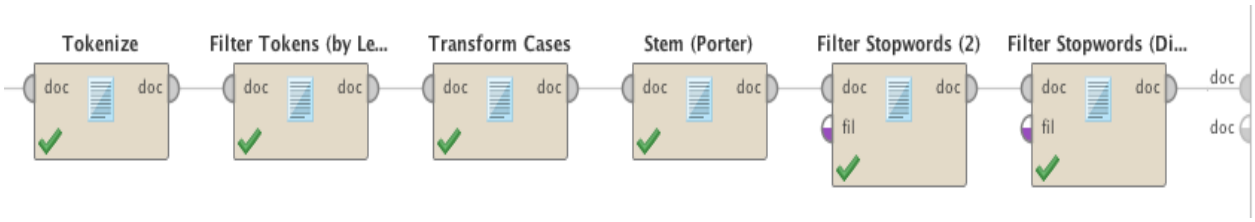

Fig 6. Text pre-processing

The result is a table with words, how many times they occur in the text, and in which document the word occurs. Table 1 shows part of the list of words.

Table 1. Result of text processing — list of words received

Word	Total Occurrences	Document Occurrences
Figurine	400	20
Decorative	389	18
Ceramics	382	20
Wooden	376	6
Fabric	359	20
Figurine	356	20
School	344	10
Flowers	334	4
Seasons	280	4
Notebook	277	19
Tiara	276	6
...	...	...

The final step is clustering using the K-means clustering method. The process of cluster grouping is as follows. First, cluster centers are defined, then clustering units are distributed using different algorithms to measure the distances of each unit to the centers. At the last stage, cluster centers are updated by re-measuring unit distances to new centers and reallocating them if necessary. The update process lasts until the cluster center changes exceed the convergence criteria or a predetermined maximum number of iterations is reached.

As we have already said, the similarity between objects is based on a measure of distance between them, and metrics for calculating similarity are extremely important because they lead to a different

distribution of documents by group. Based on studies on the application of metrics for calculating similarity, some authors suppose that K-means using Euclidean distance metric distance does not produce such good results, and it is advisable to use other algorithms to find similarity [17,18]. Other authors, however, who consider the clustering of textual documents in their studies, conclude that the Euclidean algorithm produces the best result [19]. Based on our experiments using the RapidMiner software and using the Davies–Bouldin index metric [20] for evaluating clustering algorithms, we found that the best results were obtained with the EuclideanDistance measure. The results of our experiments are shown in Table 2.

It is seen that 10 different metrics have been compared to calculate similarity to find the algorithm that creates clusters with low internal cluster distances and high cluster spacing distances and thus has the lowest Davies–Bouldin index. Tests are performed with 2, 3 and 4 cluster groups, and the table shows the absolute values of the results obtained. It should be noted that the results in RapidMiner are obtained in negative numbers, as the software product is designed to work in this way to maximize the efficiency of the process. The table shows that for $k = 2$ and $k = 4$ Euclidean has the best results, and for $k = 3$ it is in 2nd place. This causes us to conclude that in our case it is best to use this algorithm.

Table 2. A sample table for a sample file

	$k = 2$	$k = 3$	$k = 4$
EuclideanDistance	2.847	2.598	1.187
ChebychevDistance	3.209	2.630	2.556
CorrelationSimilarity	3.448	2.964	2.564
CosineSimilarity	3.066	2.702	2.210
DiceSimilarity	3.200	2.920	2.665
InnerProductSimilarity	3.200	2.964	2.665
JaccardSimilarity	3.200	2.920	2.665
ManhattanDistance	3.246	2.530	2.197
MaxProductSimilarity	3.268	2.804	2.190
OverlapSimilarity	3.475	2.717	2.578

Based on the experiments and verification of the data obtained for identifying groups for marketing purposes based on cluster analysis of the visited web pages, we suggest using the K-means clustering algorithm with the EuclideanDistance measure.

The resulting groupings and the basic terms that are closest to the centroid of the corresponding cluster help to create a profile of the identified groups and carry out targeted marketing activity with the users who fall into the respective groups. For example, when splitting the documents into 4 groups, the first cluster's words which are the closest to the centroid are: figurine, tiara, ceramics, souvenir (Table 3), which may be the reason for sending the users from this identified session to promotional offers for similar items.

Table 3. Attributes from cluster _0

Attribute	**Cluster 0**
Figurine	0.079
Tiara	0.069
Ceramics	0.066
Souvenir	0.064
Keychain	0.064
Statuette	0.060
Metal	0.059
...	...

COMPUTER SCIENCE 2017. Vol. 27. Issue 3

Conclusion

The present study proposes an approach for identifying groups that can be used for marketing purposes. The approach is based on processing the server logfiles and cluster analysis of the texts from the visited web pages. Approbation is made on the approach with the RapidMiner software product and data from the visits to a Bulgarian e-shop. The results show that the proposed approach can be successfully used to analyze webpage attendance. The new knowledge gained from the proposed analysis approach could help improve and refine customer relationship management processes and other marketing activities. Currently, extracting and successfully analyzing data from Internet resources could be an important competitive advantage for companies operating in online environments.

REFERENCES

1. Etzioni O. The World-Wide Web: quagmire or gold mine?, *Communications of the ACM*, 1996, vol. 39, issue 11, pp. 65–68. DOI: 10.1145/240455.240473
2. Sulova S. Application of web mining in customer relationship managament, *Izvestia, Journal of the Union of Scientists – Varna, Economic Sciences Section*, 2015, issue 1, pp. 105–110. https://ideas.repec.org/a/vra/journl/y2015i1p105-110.html
3. Cooley R., Mobasher B., Srivastava J. Web mining: information and pattern discovery on the World Wide Web, *Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, IEEE Computer Society, 1997, pp. 558–567. DOI: 10.1109/TAI.1997.632303
4. Markov Z., Larosed D.T. *Data mining the web: uncovering patterns in web content, structure, and usage*, New Jersey: John Wiley & Sons, 2007, 218 p.
5. Kumar E. *Natural language processing*, New Delhi: I. K. International Publishing House Pvt. Ltd., 2011, 224 p.
6. Fayyad U., Piatetsky-Shapiro G., Smyth P. From data mining to knowledge discovery in databases, *AI Magazine*, 1996, vol. 17, no. 3, pp. 37–54. DOI: 10.1609/aimag.v17i3.1230
7. Fan W., Wallace L., Rich S., Zhang Z. Tapping the power of text mining, *Communications of the ACM*, 2006, vol. 49, issue 9, pp. 76–82. DOI: 10.1145/1151030.1151032
8. Peña-Ayala A. *Educational data mining. Applications and trends*, Heidelberg: Springer International Publishing, 2014, xviii + 468 p. DOI: 10.1007/978-3-319-02738-8
9. Tarczynski T. Document clustering — concepts, metrics and algorithms, *International Journal of Electronics and Telecommunications*, 2011, vol. 57, issue 3, pp. 271–277. DOI: 10.2478/v10177-011-0036-5
10. Hartigan J.A., Wong M.A. Algorithm AS 136: a k-means clustering algorithm, *Journal of the Royal Statistical Society. Series C* (*Applied Statistics*), 1979, vol. 28, no. 1, pp. 100–108. DOI: 10.2307/2346830
11. Dixit D., Kiruthika M. Preprocessing of web logs, *International Journal on Computer Science and Engineering*, 2010, vol. 2, issue 7, pp. 2447–2452. http://www.enggjournals.com/ijcse/doc/IJCSE10-02-07-20.pdf
12. Wong S.K.M., Raghavan V.V. Vector space model of information retrieval: a reevaluation, *SIGIR '84 Proceedings of the 7th annual international ACM SIGIR conference on Research and development in information retrieval*, 1984, Cambridge, England, pp. 167–185. http://dl.acm.org/citation.cfm?id=636816
13. Jing L., Ng M.K., Yang X., Huang J.Z. A text clustering system based on k-means type subspace clustering and ontology, *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 2008, vol. 2, no. 4, pp. 1296–1308. http://waset.org/publications/2401
14. Steinbach M., Karypis G., Kumar V. A comparison of document clustering techniques, *KDD Workshop on Text Mining*, 2000. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.125.9225
15. Antony S., Wagh R. Study on text clustering for topic identification, *International Journal of Advanced Research in Computer Science*, 2017, vol. 8, no. 1, pp. 161–164. http://ijarcs.info/index.php/Ijarcs/article/view/2874
16. Linden A., Krensky P., Hare J., Idoine C.J., Sicular S., Vashisth S. Magic quadrant for data science platforms. https://www.gartner.com/doc/3606026/magic-quadrant-data-science-platforms
17. Huang A. Similarity measures for text document clustering, *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008)*, University of Canterbury, Christchurch, 2008, pp. 49–56. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.332.4480

18. Sandhya N., Lalitha Y.S., Govardhan A., Anuradha K. Analysis of similarity measures for text clustering, *International Journal of Data Engineering*, 2008, vol. 2, issue 4.
http://www.cscjournals.org/manuscript/Journals/IJDE/Volume2/Issue4/IJDE-63.pdf
19. Singh A., Yadav A., Rana A. K-means with three different distance metrics, *International Journal of Computer Applications*, 2013, vol. 67, no. 10, pp. 13–17. DOI: 10.5120/11430-6785
20. Davies D., Bouldin D.A. A cluster separation measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1979, vol. PAMI-1, issue 2, pp. 224–227. DOI: 10.1109/TPAMI.1979.4766909

Received 01.08.2017

Sulova Snezhana Dineva, PhD, Associate Professor, Department of Computer Science, University of Economics – Varna, 77 Knyaz Boris I Blvd, Varna, 9002, Bulgaria.
E-mail: ssulova@ue-varna.bg

С. Д. Сылова

Создание групп для маркетинговых целей из данных использования веб-сайта

Цитата: Вестник Удмуртского университета. Математика. Механика. Компьютерные науки. 2017. Т. 27. Вып. 3. С. 470–478.

Ключевые слова: кластеризация текстов, группы, анализ текстов, log-файл, RapidMiner.

УДК 519.688

DOI: 10.20537/vm170314

Формирование клиентских групп и извлечение информации для этих групп являются важными задачами онлайн-бизнеса, так как это позволяет наиболее полно применить методики маркетинга. Частные лица могут быть лично обслужены группами, в соответствии с выявленными интересами и предпочтениями. В данной статье мы предлагаем способ определения и создания пользовательских групп путем обработки данных использования сайтов. Используя данные журнала веб-сервера, мы заходим на выбранный сайт, просматриваем и обрабатываем текстовый контент страниц сайта. Данный подход базируется на технологии обработки естественного языка и использует методы кластеризации текстовых документов. Экспериментальное тестирование данного метода было проведено с помощью программного продукта RapidMiner и данных посещения сайта болгарского Интернет-магазина.

Поступила в редакцию 01.08.2017

Сылова Снежана Динева, PhD, доцент, факультет компьютерных наук, Экономический университет – Варна, 9002, Болгария, г. Варна, бульвар Княз Борис I, 77.
E-mail: ssulova@ue-varna.bg