AUTOMATED EXTRACTION OF TYPICAL EXPRESSIONS DESCRIBING PRODUCT FEATURES FROM CUSTOMER REVIEWS

Karel Barák[1], František Dařena[1], Jan Žižka[1]

[1] *Mendel University in Brno*

EUROPEAN JOURNAL OF BUSINESS SCIENCE AND TECHNOLOGY

Volume 1 Issue 2
ISSN 2336-6494
www.ejobsat.com

ABSTRACT

The paper presents a procedure that helps in revealing topics hidden in large collections of textual documents (such as customer reviews) related to a certain group of products or services. Together with identification of the groups containing the topics the lists of important expressions is presented which helps in understanding what characterizes these aspects most typically from the semantic point of view. The procedure includes determining an appropriate number of groups representing the prevailing topics, partitioning the documents into a desired number of groups using clustering, extracting significant typical features of documents from each group with application of feature selection methods, and evaluating the outcomes with the assistance of a human expert. The results show that the presented approach, consisting mostly of automated steps, is able to separate and characterize the aspects of a certain product as discussed by the customers and be later useful, e.g., for handling customer complaints, designing promotional campaigns, or improving the products.

KEY WORDS

product aspects identification, text mining, cluster analysis, feature selection

JEL CODES

C38, C89

1 INTRODUCTION

Understanding how customers perceive and evaluate products and services is an important element in improving business processes and increasing competitive advantage. Besides using customer feedback to enrich marketing strategies of companies, reviews and ratings contributed by customers provide information for other consumers, thereby reducing their uncertainty about the product or service and affecting sales in various contexts (Engler et al.,

BARÁK, Karel, DAŘENA, František, and ŽIŽKA, Jan. 2015. Automated Extraction of Typical Expressions Describing Product Features from Customer Reviews. *European Journal of Business Science and Technology*, 1 (2): 83–92. ISSN 2336-6494.

2015). In order to exploit the information contained in customers' messages relevant aspects and their importance need to be revealed. The aspects might be known in advance (Guyon and Elisseeff, 2003) or determined automatically, for example, according to their relatedness to sentiment bearing words (Liu, 2012). Then, the aspects' characteristics might be analyzed with respect to subjectivity according to the sentiment polarity of respective expressions; alternatively, just objective facts or facts without considering their sentiment might be examined. Both approaches require understanding the content of the relevant messages and an ability of deriving useful knowledge from them.

Traditional methods often rely on surveys where customers answer to a set of predefined questions. These answers are then analyzed by application of different statistical or other techniques and then interpreted with relation to a given objective. There are many difficulties related to these traditional approaches. The responses of the customers might be influenced by method bias when the respondents cannot provide accurate responses and/or when they are unwilling to try to provide accurate responses (MacKenzie and Podsakoff, 2012). When studying how customers perceive some product features, the order of them might play an important role (Ares and Jaeger, 2013). This approach also requires a precise specification of the properties of the interviewed subject in order to ensure some representativeness and there exists a risk of some important aspect omission (Bell and Bryman, 2015). The number of processed responses is usually relatively low, typically a few hundreds (Alpu, 2015), even when using automated machine learning methods (Bafna and Toshniwal, 2013).

With the growth of volumes of electronic data, especially thanks to massive use of various on-line channels and platforms, such as social networks, discussion boards, or on-line review sites, huge collections of documents containing customers' opinions useful for decision making are available. However, manual analysis of the data by linguistic and domain experts within a reasonable time and budget is not feasible.

Mining knowledge from textual data, known as text mining (Feldman and Sanger, 2007), is a domain, which therefore has gained a lot of attention in the last decade. The Internet is indeed a good source of user generated textual data in these days while a lot of new data originates every day. Analyzing these document collections is certainly helpful and leads to interesting and sometimes unexpected findings.

This paper presents a procedure that can be used in order to reveal important aspects of a product or service and the ways of expressing these aspects. A situation when a large amount of documents containing opinions related to a product or service is available is emphasized. A significant portion of the steps that need to be performed is automated which enables to achieve desired results in a reasonable amount of time with acceptable effort.

2 FINDING TOPICS AND THEIR RELEVANT CHARACTERISTICS

Having domain knowledge, including aspects related to a product or service and the ways of evaluating these aspects, a set of relevant attributes might be constructed directly (Guyon and Elisseeff, 2003). In the opposite case, relevant features need to be extracted using a method based on some defined principles or rules. When a collection of labeled data (the labels express the membership of data elements in some groups) is available, feature selection algorithms might be applied in order to extract attributes that are relevant for particular classes. Filter methods that are based on correlation between features and target, or wrapper methods that use a learning machine in order to assess a set of features with respect to a classification algorithm might be used (Kohavi and John, 1997).

When the class labels are related to the topics the task of revealing characteristic features of topics might be realized by the application of feature selection methods. When information about the topics (their number and subject matter) is not available a different procedure needs to be performed. There exist methods for feature selection also for unsupervised learning, i.e., when no labeling for the processed data is available. Their goal is to find the smallest feature subset that best uncovers interesting groupings from data according to the chosen criterion. What is interesting and what is the criterion needs to be specified. However, no single true answer exists here (Dy and Brodley, 2004). Even when a set of important features for an unsupervised task (without known labels) is found, only the process of partitioning the data into some groups is simplified. However, the relation between these features and the topics is not obvious. Thus, a procedure combining two steps – topic separation, and extraction of relevant attributes must be performed (Žižka and Dařena, 2013).

The assumption related to every document collection is that it consists of some more or less independent topics. A topic is a probability distribution on the universe of terms; it is typically concentrated on terms that might be used when discussing a particular subject (Bingham et al., 2003). This means, that documents related to the same topic share some common words or expressions and are therefore somehow similar. This similarity might be used to cluster the documents according to their similarity using some of the clustering algorithms. Separated groups of documents, representing the topics, might be then used as classes (labels) employed by a feature selection method.

In a supervised learning task, the quality of the extracted document subsets might be easily evaluated by examining the values of standard classification performance measures. When classification is not the main goal, validation is more complicated. There doesn't exist one objective criterion measuring the result, unlike in a classification task where the outcome might be evaluated according to the correctness of label assignment. Here, the quality of results is related to the number of identified topics and topic granularity, the way the topics are characterized, and who evaluates the representative characteristics (for example, features relevant for classification don't necessarily need to have a clear semantic meaning related to a certain topic). Without detailed knowledge of the data, contained topics, and their characteristic features, such an evaluation is always subjective and might be evaluated only qualitatively in terms of usefulness.

3 FINDING A STRUCTURE IN DATA

Clustering algorithms partition a set of documents into subsets called clusters. The goal is to create the clusters that are coherent internally, but clearly different from each other. In other words, the documents within a cluster should be as similar as possible; and documents in one cluster should be as distinct as possible from documents in other clusters. When using vector document representation two basic types of clustering algorithms might be used: hierarchical, and flat (Kaufmann and Rousseeuw, 2005).

Hierarchical clustering constructs a tree like, nested structure partition of the document set where the clusters are hierarchically arranged (Xu and Wunsch, 2009). Partitioning clustering methods do not consider any explicit structure between the clusters. Their result is a set of k clusters, where k is given or automatically determined. It has been found that partitioning clustering algorithms are well suited for clustering large document data sets due to their relatively low computational requirements (Zhao and Karypis, 2001).

3.1 Evaluating the revealed clusters

Using an unsupervised approach, the perfectness of the output is usually expected to be much lower that desired. The reason is the fact that the missing labels must be assigned automatically without having any prior knowledge of the data. Thus, the labels might be finally assigned differently than a human expert would assign them because only he or she has a clear objective and can use some additional, external information (Weiss et al., 2010).

Sufficiently high quality (acceptable for a user) of clusters is essential for the success of the entire process. It is obvious that having only one cluster is unacceptable because there is no structure visible in the data. On the other hand, having the same number of clusters and instances (i.e., each cluster contains only one object) lacks any generalization although the clusters are perfect in terms of all measures of cluster quality. The task of determining the right number of clusters is thus not easy and a compromise has to be found.

There exist many approaches how to set an optimal number of clusters, see, for example Tibshirani and Walther (2005). The elbow methodology (Meyer zu Eissen and Stein, 2002) is often employed as a rule to determine the number of clusters in data set. A number of clusters is chosen such that adding another cluster doesn't give much better modeling of the data set (Morozkov et al., 2012). The quality of modeling is measured using some of the clustering evaluation measures for different numbers of clusters; when the value of these measures doesn't change significantly a good number of clusters has been found.

Because the absence of the ground truth (as opposed to a supervised learning task where class labels are known) external evaluation measures (Zhao and Karypis, 2001) couldn't be used. Instead, internal measures evaluating the clusters according to the characteristics derived from the data itself or expert-based procedures need to be applied.

Internal measures are usually based on the criteria of compactness and separation. Compactness measures how much are the objects in a cluster related to each other. Lower variance measured, e.g., in terms of pairwise or center-based distances in the cluster, signifies higher compactness. Separation evaluates how a cluster is separated from other clusters. Measures using distances between cluster centers, pairwise distances between objects from different clusters, or measures based on density might be applied (Liu et al., 2010).

Evaluation of clustering results by experts may reveal new insight into the data, but is generally very expensive and demanding. The results that are subjectively influenced are also not very well comparable (Färber et al., 2010). In order to prevent demanding analysis of the clusters and the documents in them, a procedure applying some machine learning methods might be used in order to reveal typical characteristics of the clusters (Žižka and Dařena, 2013). It is also possible to examine not all of the documents in every cluster but only some of them. There exist several approaches of how to choose the representative documents – an average document, the least typical element, or the most typical document (Gelbukh et al., 2003).

4 DATA USED IN THE EXPERIMENTS

The data was obtained from Julian McAuley, who collected reviews from Amazon (McAuley et al., 2015). The total number of all product reviews in this dataset is 143.7 million. The reviews might be classified into several categories according to the product categories in the famous e-shop. In this paper, the category of cell phones was used in order to have a sufficient number of documents available and to avoid extensive heterogeneity in the data. Products from different categories would be evaluated from different perspectives (for example, performance parameters like data processing speed or memory size are relevant for computers, while flavor or nutrition facts are important for grocery products). Processing such diverse collections would be thus more complicated, in many cases also unreasonable, and a high number of included topics hardly interpretable. When people search for certain information their effort is usually constrained in a more detailed scope rather than unbounded in a global range.

The data set from the cell phone category contained 3,447,275 reviews from which subsets consisting of 25,000 and 50,000 were randomly selected. Our previous work (Žižka and Dařena, 2012) demonstrated that these amounts of data are relatively stable in terms of distribution of terms across included topics. The selection process and all experiments were repeated 10 times in order to confirm the usefulness of the method for different data. The longest review contained 32,384 characters, the shortest just one character, the average length was 320 characters. The smaller data set contained approximately 24K, the bigger about 30K unique words on average.

The conversion of the documents into a structured format – the vector space model (Salton and McGill, 1983) – included removing unwanted characters (e.g., digits, punctuation, and other special symbols) and breaking each document down into individual tokens (useful units for processing). The tokens were stemmed, converted to lower case, and rare terms were removed. The filtered or derived features, referred to as terms, later formed the base of structured representation of the documents. In order to prevent excessive importance of common words, known as stop words, they were removed before further analysis (the list provided by Kevin Bougé at `https://sites.google.com/site/kevinbouge/stopwords-lists` was used).

The terms derived from the documents were represented numerically using the popular method known as tf-idf (term frequency times-inverse document frequency), which is a numerical statistic that is intended to reflect how frequent a term is in a document and how rare is in the entire collection (Salton and Buckley, 1988).

5 EXPERIMENTS

In order to identify the aspects (demonstrating themselves as topics) that are relevant for a specific group of products a separation of these aspects needed to be performed. Clustering was used as the method for topics separation; the topics were expected to be isolated in the created clusters.

CLUTO software package was used to identify groups of similar documents. As the clustering algorithm, CLUTO's implementation of k-means algorithm (Manning et al., 2008) denoted here as the direct method, was used. K-means is the most widely used flat clustering algorithm. In the first step, k randomly selected cluster centers are selected (very often randomly). Then, all objects are assigned to a cluster which is the closest to the centroid. In the following step, the cluster centroids are re-computed according to the positions of the objects in the clusters. The steps of assignment of the objects to the clusters and re-computation of cluster centroids are repeated until a stopping (a fixed number of iterations or, most commonly, when the cluster centroid positions do not change between iterations) criterion has been met. As the similarity measure, cosine similarity was used (Duda et al., 2001).

To evaluate a clustering solution, internal evaluation measures were used. These measures are represented by so called criterion functions that are optimized during clustering. Internal criterion functions try to maximize the similarity of documents in individual clusters while not considering the documents in different clusters. External criterion functions focus on optimization of dissimilarity of individual clusters. Hybrid criterion functions combine both internal and external criteria, i.e., they do not focus only on intra-cluster similarity but also take similarity with documents in different clusters into account (Zhao and Karypis, 2001).

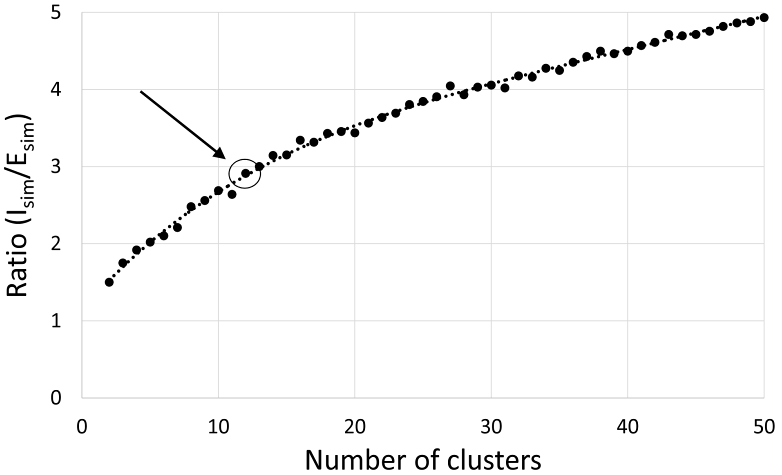

Fig. 1: Ratios of internal and external similarities for different numbers of clusters created from 25,000 reviews. The arrow shows the “elbow” of the curve approximating these points.

For each data set size, ten random selections of the desired quantity of documents were made. These data sets were clustered into $2, 3, \ldots, 50$ clusters. For each solution the values of internal and external similarity were calculated and a function approximating the ratio of internal and external similarities created. Using the elbow method (see Fig. 1) an optimal number of clusters was determined. The values were then averaged and these numbers were used later in relevant attributes extraction process.

For the data set consisting of 25,000 reviews the average cluster number was 17 (standard deviation 11) and for the larger data set with 50,000 reviews the number was 18 (standard deviation 9). The documents were thus clustered into the desired number of clusters to be prepared for relevant features extraction.

In order to identify significant attributes characterizing groups of documents, Žižka and Dařena (2013) used the C5.0 decision tree generator (Quinlan, 2015). This approach enabled extraction of the features that were important in the classification problem the solution of which was represented by an induced decision tree. The significant features were present in the tests in tree nodes and their importance was proportional to the position of the nodes in the tree (the most important feature was in the tree root, towards the leafs the importance decreased). This approach supported the reliability by providing the decision tree classification error estimates; on the other hand, this process was very demanding since the computation complexity was exponentially dependent on the number of attributes. This number is generally high in text mining tasks (Joachims, 2002).

In this paper, the chi-square (χ^2) method was used for the feature selection process. The method measures independence between a feature and a category. When a term and a category are completely independent the value of this measure is zero. The features most important with respect to a given class have thus the highest value. Computational times achieved by this method are significantly shorter than in the case of using decision trees. The identified important features are, however, very similar as demonstrated by Krupník (2014).

6 RESULTS AND DISCUSSION

The above mentioned procedure was applied to the data collection representing customer reviews of cell phones. After determining a desired number of clusters the data was clustered and a feature selection procedure was used in order to reveal the attributes characterizing these groups. For simplicity, the first ten most significant attributes for the clusters are presented in this paper. A more sophisticated approach could employ some thresholding (i.e., presenting features with their importance higher than a specified or calculated threshold); alternatively, numbers according to the requirements of a human expert might be used (even different numbers for different clusters).

The lists of important attributes for the separated groups of reviews are presented in Tab. 1. Because of relatively high number of examined clusters, only eight groups of words with derived topics are presented. The topics were determined according to reasoning of a human expert. Because of a clear relation of the words usually to one theme such decisions were often not too complicated.

Because the lists of important words are not always perfectly related to a single, clearly identifiable topic they might be combined with some representative reviews from the corresponding clusters. As representative documents, the ones residing close to cluster centers were selected, see Tab. 2 for some examples. A certain number of them might be used in order to support the process of deriving a suitable topic.

When not removing stopwords from the original documents some other interesting perspectives of the examined products emerged. For example, a group described by the words *she, her, daughter, wife, cute, mom, love, sister, gift,* and *mother* pointed to reviews that were somehow related to females (the review *I bought this cover for my daughter's blackberry phone. It fit perfectly and she was very pleased with the product.* was one of the reviews close to the cluster centroid).

Processing the data sets consisting of 25,000 and 50,000 reviews provided almost identical results in terms of identified clusters, their number, and the lists of significant words describing their semantic content. This demonstrates the fact that the amount of 25,000 documents is representative enough; with more documents some expressions are rather repeated and no (or very few) new topics and their characteristic features appear. Thus, only the results for the smaller data set are presented in this paper.

In order to support the process of determining an optimal number of clusters, the clustering solutions consisting of more and less clusters (7 and 27 for the data set consisting of 25,000 reviews) were analyzed. Having more groups, some of the topics naturally appeared more than once, like *Mobile accessory (screen protection)*. Some of the aspects spread across more groups and were more specialized compared to the situation with lower number of clusters, like *Mobile accessory (protection bumper, protection case), Mobile accessory (protection cases from silicon)*, and *Mobile accessory (protection cases)*. When processing the data partitioned into lower number of groups some topics were obviously mixed and not so particularized. For example, a group described by the words *case, color, protect, drop, cover, work, snap, rubber, cute, bumper* discussed more aspects of *mobile accessories* (cases, colors, types).

It seems that the elbow method was able to provide a reasonable number of document groups in terms of their relatedness to hidden topics (or aspects). After examination of different groupings, it can be concluded that it was generally better to partition the documents into slightly higher number of clusters in order to not loose some of the semantic information.

Because an entire review might typically address more than one aspect of a product the assignment of the review into one group will not be completely perfect (the review should in fact belong to more groups). Thus, smaller portions of the documents, such as paragraphs or sentences might be considered as meaningful elements. A few experiments with documents primitively partitioned into sentences were conducted. A significant change in the identified clusters, their important features, and derived

Tab. 1: Important attributes (stemmed) and the derived topics for the clustered data set consisting of 25,000 reviews

Important attributes	**Derived topic**
batteri, charg, life, mah, evo, hour, origin, oem, stock, die	Mobile accessory (battery)
signal, antenna, bar, hous, roof, booster, boost, unit, cell, feet	Mobile accessory (signal booster, antenna)
charger, cord, car, retract, plug, charg, work, wall, usb, transmitt	Mobile accessory (charger)
sound, ear, headset, hear, bluetooth, nois, comfort, music, pair, listen	Mobile accessory (headset, sound)
cabl, lg, usb, comput, transfer, data, tracfon, charg, pc, micro	Mobile accessory (cables, connectivity)
money, wast, worth, save, spend, junk, dont, buy, don, total	Customers discussing price
glare, film, anti, matt, finish, screen, mirror, protector, fingerprint, retina	Mobile accessory (screen film and protection)
color, pink, white, love, pictur, purpl, case, black, yellow, green	Customers discussing colors

Tab. 2: Examples of reviews close to the centers of the identified clusters

Mobile accessory (batteries)
After replacing my battery my phone no longer worked and I had to buy a new one so I would not recommend this.
This is a good battery. it works 90% like the original battery.If you need a replacement battery this will make a good one.
Mobile accessory (chargers)
This is a perfect charger for my car. It works great, and the price is right. I recommend this product to all.
Works well, charges phone quickly, easy to use. Would recommend to others who need a charger. I would buy again.

topics did not occur unlike in (Dařena et al., 2014). On the other hand, some of the representative documents were very short (like *best*, *awesome*, or *I love it*) and thus bringing no additional semantic insight into the data.

7 CONCLUSIONS AND FUTURE WORK

The paper demonstrated a procedure that helps in revealing topics (aspects, as perceived by customers) hidden in large collections of textual documents (customer reviews) related to a certain group of products or services. Together with identification of the groups containing the topics the lists of important expressions (here words and the entire reviews) were discovered which facilitated understanding what characterized these aspects most typically from the semantic point of view.

This procedure did not require a specific domain knowledge that could be used in feature identification process. It was also not based on linguistic information, like in (Bafna and Toshniwal, 2013) where as the features frequently appearing nouns were used, or in (Hu and Liu, 2004) where adjectives helped in identification of opinions. In this paper, no additional knowledge, like a sentiment lexicon (Maks and Vossen, 2012) was needed which made the entire proces straightforward and self-contained.

Future research will concentrate on deeper analysis of parameters of the used methods and on alternative approaches to individual steps of the proposed method. For example, a hierarchical clustering algorithm might be used instead of a partitioning one (this might support a hypothesis of hierarchical topics arrangement), different feature selection methods or their combination might be applied, a more sophisticated method of selecting representative documents and their combination with representative attributes might be employed (for example, retrieving documents containing the identified significant attributes). The major problem or deficiency of the presented procedure still lies in the absence of clear quantitative evaluating criterion; thus more attention might be paid to this direction. However, even when including more human experts, a clear uniform conclusion doesn't have to be reached (Saratlija et al., 2011).

8 ACKNOWLEDGEMENTS

Realization of the paper was supported by Mendel University in Brno and by its Internal Grant Agency, project number PEF_DP_2015_026.

9 REFERENCES

Alpu, O. 2015. A methodology for evaluating satisfaction with high-speed train services: A case study in Turkey. *Transport Policy*, 44, 151–157.

Ares, G. and Jaeger, S. R. 2013. Check-all-that-apply questions: Influence of attribute order on sensory product characterization. *Food Quality and Preference*, 28 (1), 141–153.

Bafna, K. and Toshniwal, D. 2013. Feature Based Summarization of Customers' Reviews of Online Products. In: 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems – KES2013. *Procedia Computer Science*, 22, 142–151.

Bell, E. and Bryman, A. 2015. *Business research methods*. Oxford: Oxford University Press.

Bingham, E., Kabán, A. and Girolami, M. 2003. Topic identification in dynamical text by complexity pursuit. *Neural Processing Letters*, 17, 69–83.

Dařena, F., Žižka, J. and Přichystal, J. 2014. Clients' freely written assessment as the source of automatically mined opinions. In: *17th International Conference Enterprise And Competitive Environment*. Amsterdam, Netherlands: Elsevier Science Bv, 103–110.

Duda, R. O., Hart, P. E. and Stork, D. G. 2001. *Pattern Classification*. New York, NY: Wiley.

Dy, J. G. and Brodley, C. E. 2004. Feature Selection for Unsupervised Learning. *Journal of Machine Learning Research*, 5, 845–889.

Engler, T. H., Winter, P. and Schulz, M. 2015. Understanding online product ratings: A customer satisfaction model. *Journal of Retailing and Consumer Services*, 27, 113–120.

Färber, I., Günnemann, S., Kriegel, H. et al. 2010. On Using Class-Labels in Evaluation of Clusterings. In: *Proceedings of the 1st International Workshop on Discovering, Summarizing and Using Multiple Clusterings (MultiClust 2010) in conjunction with 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Washington.

Feldman, R. and Sanger, J. 2007. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge University Press.

Gelbukh, A. F., Alexandrov, M., Bourek, A. and Makagonov, P. 2003. Selection of Representative Documents for Clusters in a Document Collection. In: *Proceedings of Natural Language Processing and Information Systems, 8th International Conference on Applications of Natural Language to Information Systems*, 120–126.

Guyon, I. and Elisseeff, A. 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3, 1157–1182.

Hu, M. and Liu, B. 2004. Mining and summarizing customer reviews. In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. Seattle, USA.

Joachims, T. 2002. *Learning to classify text using support vector machines*. Norwell: Kluwer Academic Publishers.

Kaufmann, L. and Rousseeuw, P. J. 2005. *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: Wiley.

Kohavi, R. and John, G. 1997. Wrappers for feature selection. *Artificial Intelligence*, 97 (1–2), 273–324.

Krupník, J. 2014. Stopwords removal influence on text mining task results. In: *PEFnet 2014*. Brno: Mendel University.

Liu, B. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Liu, Y., Li, Z., Xiong, H., Gao, X. and Wu, J. 2010. Understanding of Internal Clustering Validation Measures. In: *Proceedings of ICDM 2010, The 10th IEEE International Conference on Data Mining*, 911–916.

Maks, I. and Vossen, P. 2012. A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems*, 53 (4), 680–688.

Manning, C. D., Raghavan, P. and Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

McAuley, J. J., Targett, C., Shi, Q. and Van Den Hengel, A. 2015. Image-Based Recommendations on Styles and Substitutes. In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM, 43–52.

MacKenzie, S. B. and Podsakoff, P. M. 2012. Common Method Bias in Marketing: Causes, Mechanisms, and Procedural Remedies. *Journal of Retailing*, 88 (4), 542–555.

Meyer zu Eissen, S. and Stein, B. 2002. Analysis of Clustering Algorithms for Web-Based Search. In: *4th International Conference, PAKM 2002 Vienna, Austria, December 2–3*. Berlin: Springer, 168–178.

Morozkov, M., Granichin, O., Volkovich, Z. and Zhang, X. 2012. Fast algorithm for finding true number of clusters. applications to control systems. In: *Control and Decision Conference (CCDC)*, 2001–2006.

Quinlan, J. R. 2015. Data mining tools See5 and C5.0. RuleQuest Research. [online]. Available at: `https://www.rulequest.com/see5-info.html`. [Accessed 2015, October 14].

Salton, G. and Buckley, C. 1988. Term-weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24 (5), 513–523.

Salton, G. and McGill, M. J. 1983. *Introduction to Modern Information Retrieval*. New York: McGraw Hill.

Saratlija, J., Šnajder, J. and Dalbelo Bašić, B. 2011. Unsupervised Topic-Oriented Keyphrase Extraction and its Application to Croatian. In: *14th International Conference on Text, Speech and Dialogue*, 340–347.

Tibshirani, R. and Walther, G. 2005. Cluster Validation by Prediction Strength. *Journal of Computational and Graphical Statistics*, 14 (3), 511–528.

Weiss, S. M., Indurkhya, N., Zhang, T., Damerau, F. 2010. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. New York, Springer.

Xu, R. and Wunsch, D. C. 2009. *Clustering*. Hoboken, NJ: Wiley.

Zhao, Y. and Karypis, G. 2001. *Criterion Functions for Document Clustering: Experiments and Analysis*. University of Minnesota, Technical Report.

Žižka, J. and Dařena, F. 2012. Parallel Processing of Very Many Textual Customers' Reviews Freely Written Down in Natural Languages. In: *IMMM 2012: The Second International Conference on Advances in Information Mining and Management*. Venice, Italy, October 21–26. IARIA, 147–153.

Žižka, J. and Dařena, F. 2013. Revealing Prevailing Semantic Contents of Clusters Generated from Untagged Freely Written Text Documents in Natural Languages. In: *Text, Speech, and Dialogue*. Heidelberg: Springer, 434–441.

AUTHOR'S ADDRESS

Karel Barák, Department of Informatics, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czech Republic, e-mail: info@karelbarak.cz

František Dařena, Department of Informatics, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czech Republic, e-mail: frantisek.darena@mendelu.cz

Jan Žižka, Department of Informatics, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czech Republic, e-mail: jan.zizka@mendelu.cz