Revealing Prevailing Semantic Contents of Clusters Generated from Untagged Freely Written Text Documents in Natural Languages

Jan Žižka and František Dařena

Department of Informatics, FBE, Mendel University in Brno
Zemědělská 1, 613 00 Brno, Czech Republic
{zizka,darena}@mendelu.cz

Abstract. The presented work deals with automatic detection of semantic contents of groups of textual documents, which are freely written in various natural languages. The large original set of untagged documents is split between a requested number of clusters according to a user's needs. Each cluster is taken as a class and a classifier (decision tree) is induced. The words used by the tree represent significant terms that define semantics of individual clusters. The importance (weights) of the terms combined in individual tree branches are computed according to their particular meaning from the correct classification viewpoint – a certain word combined with other words may lead to different classes but a specific class can strongly prevail. The results are demonstrated using large data sets composed from many hotel-service customers' reviews written in six different natural languages.

Keywords: textual documents, semantics, natural language, clustering, tagging, classification, automatic disclosure of meaning, Cluto, c5/See5.

1 Introduction

Having a very large collection of untagged and unformatted textual documents freely written in a natural language, a question suggesting itself can be: Would it be possible to automatically find some document groups characterized by the same or very similar semantic contents? For thousands or millions of documents, it would be quite impractical or impossible to do it manually within an acceptable time and cost range. Computers are able to find such groups using appropriate clustering algorithms and procedures, however, a typical problem often is whether the individual clusters represent any *reasonable semantic meanings*.

The following sections describe a method that investigates a given set of document clusters from the semantic point of view. Inspired by [11], where the authors applied the significance of entropy-lowering words to classification of textual documents, the research presented here demonstrates how it is possible to specify the main sense of a short textual document, which does not have a particular structure and is freely written in a natural language to represent a certain opinion. Such documents are typical for various web-based applications as blogs, mind expression, opinion formulation, etc.

I. Habernal and V. Matousek (Eds.): TSD 2013, LNAI 8082, pp. 434–441, 2013.
© Springer-Verlag Berlin Heidelberg 2013

The accent was put on processing large number of documents without tags that could provide bias concerning their sense from the semantic point of view. As the data, the research used reviews of customers of hotel services. The reviews can be easily written using a common Internet web-browser. Providers of various services today collect meaning of customers as a certain feedback. Many reviews give typically a lot of information referring to a certain matter – here, the quality of hotel services and customers' (dis)satisfaction. Such reviews are usually ranked by their authors using a given scale like from five stars (complete satisfaction) to one star (complete dissatisfaction). This labeling is often too rough because the service provider may be interested in, e.g, what is very important, what can be ignored, or what is typical. In this case, the set of reviews should be additionally categorized. A typical method is an application of clustering, which generates groups of instances having similar characteristics representing certain contents.

Here, it is necessary to emphasize that 1) the research goal was not creating classes for classification, and 2) the text-mining problem belongs among the strongly *data-driven* tasks [2]. The primary aim was to find a method how to *reveal the main semantic contents of clusters* generated in different numbers. As almost each review described several aspects of evaluating the hotel service, the individual clusters expectedly did not represent just one topic. However, there typically was one prevailing theme accompanied by other minor ones. In some cases, certain clusters represented just a mixture of various aspects without any particular preference of some of them. Anyway, it corresponds to the different way how individual users look at the reviewed service: different people prefer different things (breakfast, transportation, price, etc.) while some service qualities are commonly shared (cleanness, food, quiet hours, staff friendliness, etc.).

2 Generating Clusters Using the Cluto System

Various clustering methods can be applied to untagged data sets. A report concerning the unsupervised approach can be found in [10]. Here, the research was based on a developed clustering system known as *Cluto* [6], which is very suitable for clustering high-dimensional data-sets like textual documents. A user has to specify a set of parameters as the number of clusters plus the method of particular clustering (details can be found in [6]), and Cluto divides the data into groups with minimized inter- and maximized intra-group group similarity. The essential question was how many clusters should be generated. It is necessary to avoid too high generality (just one/two cluster/s) or concreteness (as many clusters as instances). The generality is important as it can say what is common in a certain group of reviews. However, larger number of smaller and more concrete clusters may provide more detailed information. Thus, a user interested in what individual clusters represent can opt for different number of clusters and then investigate what is typically general and what are the details – in such a case, there is no optimal cluster number as it depends on a particular application.

2.1 Data Preparation

After the preliminary phase, the authors decided to experiment with three numbers of clusters: two (the highest generality), five (medium generality), and ten (higher

specificity). These numbers depended on the total number of available reviews (each cluster should contain sufficient amount of documents to avoid too low number of reviews per cluster). Another goal was finding out how the suggested method worked for different languages: DE (German), EN (English), ES (Spanish), FR (French), IT (Italian), and CS (Czech). The first five languages represented the 'big' ones as there were hundreds of thousands reviews available (for EN almost two millions) – due to too high computational complexity, each of those languages was finally represented by a repeated random selection of 50,000 reviews. The only exception was the CS-set because it contained only 17,103 reviews – it was used in one piece as a representative of 'smaller' languages.

The data source was prepared using the method *bag-of-words* [1]: each review was transformed into a vector where a word was replaced by its weight using the *tf-idf* formula ($term\ frequency \times inverted\ document\ frequency$) [9]. Alternative word representations like *n*-grams provided worse results due to increasing the original very high sparsity (av. 99.85%), which is one of difficult problems [7]. Because the experiments used data from six languages and there was no available *unified* tool for *uniform* stemming, stop-words removing, etc., only digits were removed (various tested tools were giving different results for the same data). The rest was left as it was because the intention was also to compare mutually the results for the six tested languages under the as same conditions as possible. However, when not thoroughly reducing irrelevant or grammatically incorrect terms, the result contains a lot of redundant terms that wrongly increase the dimensionality and introduce noise – for example, a certain word can be mistyped in several ways, which leads to an artificial increase of the dictionary size.

2.2 Clustering

The clustering procedure used the tool Cluto [6]. Except its many parameters that allow various experiments and looking for the best parameter combination, Cluto implementation is also very fast and uses the computer memory (RAM) very efficiently with relation to the possible sparsity of vectors. The primary parameter was the requested number of clusters, which was from 1 to 20. Because of the limited space, only the results for 10 are here demonstrated but they are quite representative and illustrative. As the method of clustering, the experiments employed the so-called *direct* one, which is Cluto's implementation of k-means [5]. The similarity between reviews was measured using the cosine of an angle between vectors (often used in text mining), and the criterion function for evaluating the clusters' quality was hybrid $H2$ based on combination of the internal criterion $I2$ (the intra-cluster similarity maximization based on a cluster's centroid vector) and external $E1$ (the inter-cluster similarity minimization); all the parameter details can be found in mentioned [6].

3 Searching for the Prevailing Semantic Contents

The semantic content of the generated clusters is – due to the applied *bag-of-words* representation – given by words (terms) that are significant for expressing the meanings revealed by the used data-mining techniques. Certain important terms relate to a specific

topics while other significant words to different ones. Now the question is how to find those significant words that would express the particular semantic meanings. In [11] and [4], the authors applied a generator of decision trees [8] that provided a rank of attributes the values of which were decisive for minimizing the entropy. The heterogeneous set of instances mixed from different classes was split between more homogeneous subsets representing instances belonging to more specific classes.

3.1 Looking for Significant Words

The main idea is to find such document elements that can say what a document is talking about – here, the *significant words*. The most significant word is in the root of the decision tree because the tree asks each time for the value representing the word. Other words in the rank get gradually lower significance according to their importance for decreasing entropy with respect to the classification accuracy. As it was shown in [12], those significant words (and phrases composed from them) corresponded very well to a reader's point of view. For the presented data type, such words represent the semantic contents of the clusters, see [12]. The most significant words (from the root and levels below the root) are the leading exponents, which provide the main meaning of the document.

The search for the prevailing semantic contents starts from creating a given number of clusters according to the need of generality or the level of details. Each of the clusters is taken as a class. Then, using the clusters, a decision tree is constructed (c5/See5 [3]) and the byproduct of the tree is the rank of significant words. The top-level words in the rank give the semantic meaning of the cluster: a group of reviews dealing with the same matter. The words appearing in the decision tree create a dictionary composed only from the significant words – their number is a small fraction of the all words used in the reviews, typically a couple of hundreds from tens of thousands (for EN it was just 198 significant words from 26,092, for CS 287 from 29,023, etc.). The classification using generated clusters as tags of the reviews worked with a relatively small accuracy error that was (applying 10-fold cross-validation testing) 8-13% and slightly higher for CS due to the smaller number (17,103 vs. 50,000) of review samples – 16.4%.

3.2 Weighting the Importance of Different Significant Words

The words contained in each branch of the classification tree present combinations of terms leading to a certain class. If a branch ending in a leaf (which represents a class) contains words that lead exclusively to that class, such words are typical just for that class. Branches leading to different classes may contain some identical words, for example, *always-bad-breakfasts*, or *always-almost-not-bad-breakfasts*, where only the word *bad* makes a difference while other words are the same. Another branch can contain *almost-not-friendly-personnel*, where the semantic meaning does not deal with breakfast even if there are also some identical words – in spite of certain identical words, the three branches lead to different classes (*good breakfast, bad breakfast, unfriendly personnel*). Anyway, the most significant word, which is in the tree root, is part of *every* branch.

Thus, there is a problem how to assign a degree of strength to words pointing correctly (in combinations with other words) via different branches to different classes, that is, to different semantic meanings. For example, a word W_1, *bad*, can be used 30 times for correct and 0 times for incorrect classification to a class B (bad breakfasts), and another word W_2, *always*, can be used 30 times for correct and 20 times for incorrect classification to that class B (50 times in total). Which of these two words contribute more to B? The word W_1 was used less times but in 100% correctly, while the word W_2 was used 5 times more but with only 60% correctness.

According to the method described in [4], the frequencies of the correct and incorrect directing by a given word in the tree were represented using a two-dimensional vector space to introduce a weight that balanced those two frequencies. The word weight w_w was given by the following formula:

$$w_w = \frac{N_{correct}}{N_{all}} \cdot \frac{ln\sqrt{N_{correct}^2 + N_{all}^2}}{ln(N_{max})}, \tag{1}$$

where N_{max} is the maximum of $N_{correct}$ (the number of a word usages for correct classifications) and N_{all} (the sum of a word usages for all classifications). The calculated weight then determines the importance of a word in relation to a given class – higher numbers mean greater relevancy.

4 Results of Experiments

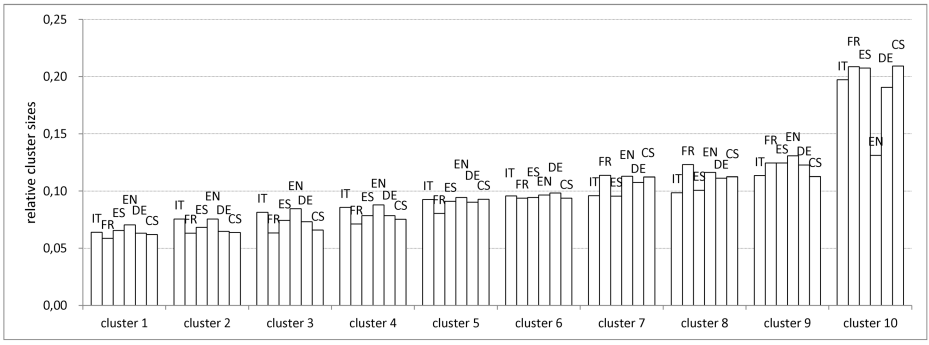

Fig. 1. The distribution of relative sizes of ten individual clusters for the six investigated natural languages. The last cluster 10 represents a mixture of several different topics where none of them prevails semantically others.

As described above, the first step included the clustering, which separated certain numbers of reviews between groups having high similarity inside and low outside. Fig. 1 illustrates how the original sets of reviews for individual languages were relatively split among requested 10 cluster groups (10 clusters for each investigated language). After looking at the generated significant words, the cluster 10 was semantically indifferent

because it contained various aspects while none of them prevailed. Other clusters represented prevailing topics. For other numbers of clusters, the situation was very similar; only the lower numbers (below 5) did not show big differences in the number of review distributions between clusters, especially for the minimum given by two clusters. The semantically relevant clusters had very close sizes within each group (cluster 1 to cluster 9).

The results are summarized in Table 1. The semantic contents (main topics) was assigned to individual clusters manually after looking at a list of significant words weighted in combinations with other ones by w_w's. It is interesting to see that for different languages (various nationality of hotel customers), there are often the identical topics and meanings, for example, *location, staff, breakfast*, and so like. As a brief illustration, the top terms representing English for *room negatives* according to their (gradually decreasing) weights: *too, small, noisy, little, old, dated, ...*; French: *peu, trop, manque, absence, odeur, insonorisation, bruit, ...*; Spanish: *poco, falta, caro, algo, antiguo, escaso*, ...; Czech: *hluk, malý, klimatizace, chybí*, ... Similarly, *room positives* for English: *comfortable, spacious, clean, modern, quiet, large, lovely, well*, ...; German: *schöne, schöner, wunderschöne, saubers, grosse, ...*; Italian: *camere, pulite, stanze, confortevoli*, ..., and so like. An example of an English *N/A* cluster: *I, nothing, my, have, like, didn't, that, would, they,* ..., and very similarly for other used languages.

5 Conclusions

The presented method of disclosing semantic contents from a very large volume of untagged textual documents written in natural languages demonstrates that it is possible to carry out it with a useful machine support. A user interested in the possible contents of textual documents has to decide how many groups the collected data set should be separated in. Such a separation can be realized by clustering, after which a user gets potential classes. Then, a classifier of the decision tree type is induced, which generates a small set of words significant for expressing individual semantic contents. Because the significant words are combined in each tree branch, it is necessary to give them weights that represent the words' importance for each possible semantic contents hidden in each cluster.

The experiments were carried out with large real-world data created in six different natural languages by customers reviewing used hotel services. The results demonstrated that – depending on the generality or specificity given by the requested number of created clusters – computers are able to reveal meaning of groups of textual documents and that such meanings are very often identical or similar between various languages. At the same time, the experiments shown that a user (which can be a hotel manager) can also reveal groups of meanings that are not very specific and, according to her or his needs, it is possible to more deeply study reduced number of reviews, which is without such a support impossible due to the extremely large volume of data.

Unfortunately, comparing and evaluating different similar systems is extremely difficult because of the different used data sets, sense inventories, and knowledge resources adopted. Text-mining belongs among strongly data-driven areas from the machine-learning viewpoint and the inductively obtained results often depend on particular data

Table 1. The revealed prevailing semantic contents (for 10 generated clusters) based on significant words in individual reviews for the six tested languages. N/A means that no specific topic could be derived from significant words and the cluster represented a mixture of several topics approximately balanced.

language	main topic of the clusters				
	1	2	3	4	5
CS	general positives	breakfast	N/A	positives, no diacritic	staff
DE	N/A	general positives	location	N/A	breakfast
EN	N/A	value	N/A	hotel facilities	room positives
ES	rooms	environment	location (no diacritic)	N/A	location (with diacritic)
FR	N/A	breakfast, facilities	environment	room negatives	location
IT	location	staff, facilities	rooms	N/A	convenience

language	main topic of the clusters				
	6	7	8	9	10
CS	room negatives	staff, cleanliness	location	surroundings	location
DE	N/A	general positives	room positives	quality/price	atmosphere
EN	room negatives	N/A	staff	room facilities	location
ES	general negatives	N/A	rooms	quality/price	location
FR	N/A	location	price, quality	N/A	comfort
IT	room facilities	location	N/A	good quality	room positives

[9]. In addition, comparing methods even on the same corpus is not eligible if there is different sense inventories. Primarily, the presented research aimed at particular large real-world data-sets with very sparse vectors, looking for an uncomplicated method applicable to not only one specific language.

The following research work aims at deeper analysis of such clusters, including more languages as well as more sophisticated data preparation (at least, removing stop-words and applying a kind of stemming). A big problem is subsequently (in bulk, after writing reviews) correcting mistyping of very large data volumes – it would be much better to apply this function simply during writing the reviews to get not so noisy data. It should be also investigated how (and in which) the various number of requested clusters differs and for what number of clusters the suggested method begins to be useless due to the loss of generality.

Acknowledgments. This work was supported by the research grants of the Czech Ministry of Education VZ MSM No. 6215648904 and IGA of the Mendel University in Brno No. 4/2013.

References

1. Berry, M.W., Kogan, J. (eds.): Text Mining: Applications and Theory. John Wiley & Sons, Chichester (2010)
2. Bloedhorn, S., Blohm, S., Cimiano, P., Giesbrecht, E., Hotho, A., Lösch, U., Mödche, A., Mönch, E., Sorg, P., Staab, S., Völker, J.: Combining Data-Driven and Semantic Approaches for Text Mining. In: Foundations for the Web of Information and Services: A Review of 20 Years of Semantic Web Research, pp. 115–142. Springer, Heidelberg (2011)
3. c5/See5 (June 2013), `http://www.rulequest.com/see5-info.html`
4. Dařena, F., Žižka, J.: Text Mining-Based Formation of Dictionaries Expressing Opinions in Natural Languages. In: Proceedings of the 17th International Conference on Soft Computing Mendel 2011, Brno, June 15-17, pp. 374–381 (2011)
5. Karypis, G., Zhao, Y.: Criterion Functions for Document Clustering: Experiments and Analysis. Technical Report 01-40, University of Minnesota, USA (2001)
6. Karypis, G.: Cluto: A Clustering Toolkit. Technical report 02-017, University of Minnesota, USA (2003)
7. Qu, L., Ifrim, G., Weikum, G.: The Bag-of-Opinions Method for Review Rating Prediction from Sparse Text Patterns. In: Proceedings of the 23rd Intl. Conference on Computational Linguistics, COLING 2010, Beijing, China, August 23-27, pp. 913–921 (2010)
8. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)
9. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys 1, 1–47 (2002)
10. Traupman, J., Wilensky, R.: Experiments in Improving Unsupervised Word Sense Disambiguation. Technical Report UCB/CSD-03-1227, February 2003, Computer Science Division (EECS), University of California, Berkeley (2003)
11. Žižka, J., Dařena, F.: Mining Significant Words from Customer Opinions Written in Different Natural Languages. In: Habernal, I., Matoušek, V. (eds.) TSD 2011. LNCS, vol. 6836, pp. 211–218. Springer, Heidelberg (2011)
12. Žižka, J., Dařena, F.: Mining Textual Significant Expressions Reflecting Opinions in Natural Languages. In: Proc. of the 11th Intl. Conf. on Intelligent Systems Design and Applications, ISDA 2011, Córdoba, Spain, November 22-24, pp. 136–141 (2011)