Ekonomiczne Problemy Usług nr 1/2017 (126), t. 2

ISSN: 1896-382X | www.wnus.edu.pl/epu
DOI: 10.18276/epu.2017.126/2-16 | strony: 153–159

Julian Vasilev
University of Economics Varna
Department of Informatics
e-mail: vasilev@ue-varna.bg

Nataliya Marinova
D. Tsenov Academy of Economics Svishtov
Department of Business Informatics
e-mail: n.marinova@uni-svishtov.bg

Text mining of articles in an issue of the journal „Economics and Computer Science" dedicated on the DIMBI project

JEL codes: C45, C88

Keywords: ontology, text mining, Rapid Miner, the DIMBI project

Summary. The purpose of this article is to use business intelligence techniques to analyse articles in an issue (Volume 2, Issue 5) of the journal „Economics and Computer Science". Since business intelligence methods are many, the research is limited to text mining methods. The research aim is to find terminology which is common for all articles in one issue of the journal. Since the journal has published several thematic issues, it is a research questions to find ontologies in each thematic issue. Rapid Miner is used as a software tool to conduct the text mining techniques. The most frequently used terms are found by Rapid Miner. A manual thematic classification of terms is done. The main groups are: educational, research and software. The proposed methodology may be used by other authors for other surveys in different thematic content.

Introduction

During the last decade the volume of data, generated from various information sources in a world has arisen significantly. A research work by IDC (Gantz, Reinsel, 2012) performed in the end of 2012 predicts that in the 2012–2020 period of time the volume of digital data in a world scale will arise 40 times. Retrieving of relevant contents within the vast quantity of text information became a real challenge, requiring detailed search in the different thematic areas. The process may significantly be facilitated via using appropriate methods for automated processing and analysis of the differently formatted sets of data, such as for instance, text mining.

Feldman and Sanger (2007) define text mining or text analytics as a general term, describing a set of techniques, which are aimed toward retrieving useful information from a collection of documents through identification and exploration of repeating patterns within non-structured text data in various documents – books, web pages, e-mails, papers or product descriptions. The authors cited above expand more formal definition of Hearst (1999), according to whom text mining is related only to creation of new non-obvious information (models, trends or links) from a given collection of text documents.

Nevertheless the idea of Text mining was known in the 60s of the 20th century, it became popular in 90s, when it was recognized as a main applied area in the field of the information systems. Creating algorithms of machine learning which can perform text mining tasks drastically reduces the efforts and time of hand processing of the text contents.

That is why the first tasks for retrieving text are connected to document’s classification and indexation. Ontology, which is another frequently used technique in text mining, contains the domain knowledge and these are in the form of relationships between the entities i.e. concepts and their level order (Kumar, Ravi, 2016).

Knowledge about data or text mining from important and relatively larger database has been recognized by numerous scholars and researchers. Hashimi, Hafez, and Mathkour (2015) have reviewed a lot of scientific literature and applications to underline primary Text mining domains.

The main areas of text mining are: text clustering, topic tracking and categorization. The result of clustering or categorization is usually done through procedures of data preparation, data filtering and tokenization.

Usual text mining tasks include activities of search engines, such as assigning texts to one or more categories (text categorization), grouping similar texts together (text clustering), finding the subject of discussions (concept/entity extraction), finding the tone of a text (sentiment analysis), summarizing documents, and learning relations between entities described in a text (entity relation modelling).

Text mining applications transform unstructured textual data into a structured format for further analysis. This process can be summarized in three steps, as pointed in Figure 1. Firstly, a data source is selected (Step 1). Secondly, this data is preprocessed (Step 2) and thirdly, analyzed (Step 3). Finally, the results are interpreted. For this study, firstly we have used articles in an online journal, published in a thematic issue. We assume that similar topics are discussed. This assumption may be accepted or rejected through the empirical study (given in the second part of this paper – “material and method” section and “results and discussion” section). Secondly, text pre-processing procedures are done (tokenization and stop-word removal). Thirdly, data analysis (in this study) is based on term frequencies. Future research may focus on other techniques for data analysis, such as cluster analysis, network analysis and association analysis.

The increased usage of text mining applications in late years is justified by the concept development “Big Data” and the tools of data mining. Text mining has something in common with better known data mining area, but possesses some differences, using text processing methods, borrowed from other scientific areas as for instance, general statistics, machine learning, database management, artificial intelligence and computational linguistics. Data mining instruments work well with structured data. Often, the data that has not been well structured yet still contains a lot of hidden information. Text mining tools automatically analyze the text document body, and discover that invisible information displaying it as text of a new contents or a diagram.

The goal of text mining tools usage is to retrieve (mine) meaningful information and knowledge from large document repositories, to convert the text into data and to analyze it using various data mining techniques. In most cases the data, collected from different types of information sources is in vast quantities and it is not possible to process it manually. The identification, separation and clustering of any specific type of information within a text document requires to apply text mining techniques or methods. In the field of education, the text mining techniques (statistical, linguistic and machine learning) support the data study and the analyses, coming from new scientific discoveries and research as well as the control and checking of complex data applying specific criteria.

For an organizational point of view the text mining applications are more practical and effective from the traditional techniques for working with databases, because it is assumed that the larger part of a company information is stored in non-structural format on different computers, not on a single common repository. The technology of text mining could be applied in: automated processing of open questions in electronic studies, to provide automated feedback answering the on-line questions of the customers asked also on-line, matching frequently asked customer’s questions, automated check of the CV’s of potential staff, following the customer’s opinions related to the company products in social networks, checking of the large databases related to patents, in order to avoid eventual violations, etc.

1. Material and method

The research aim is to find terminology which is common for all articles in one issue (volume 5, issue 2) of the journal “Economics and Computer Science.”[1] This research work uses as input data four articles in volume 5, issue 2 of the scientific journal. All articles are dedicated on the DIMBI project.[2] These four articles are written by four different authors. The titles are different. But all of them are dedicated on the DIMBI project and on teaching business informatics. That is why it is assumed that there may be common terms used in all articles.

Finding such terms manually is a difficult task. That is why text mining techniques are used. Rapid Miner[3] is used as a software tool to make ontologies from these four published articles.

Rapid Miner has a lot of instruments for text mining. A new process is created in Rapid Miner. The following operators are used:

1. Process documents for files – for reading the text information from articles.
2. Tokenize – to separate words by comma, dot, interval, semi-column.
3. Filter Tokens (by length) – to skip some of the words by length (min 4, max 25).
4. Filter Stopwords (English) – removes English stopwords.

2. Results and discussion

The output of Rapid Miner contains a word list with columns: word, total occurrences, document occurrences and other columns. The word list is sorted by the column “total occurrences” in descending order (table 1).

[1] http://eknigibg.net/index.php?route=information/information&information_id=8.

[2] http://dimbi.paragonweb.eu.

[3] https://rapidminer.com/.

Table 1. A part of the word list, generated by Rapid Miner, with words with the greatest total occurrences

Word	Total occurrences	Document occurrences
business	154	4
students	87	3
methods	69	4
software	65	4
skills	58	4
data	50	4
knowledge	48	4
universities	47	4
intelligence	44	4
tools	44	4
DIMBI	33	3
project	33	4
innovative	27	4
Informatics	25	3
products	25	4
analysis	24	3
research	23	4
practical	22	4
conclusion	18	3
requirements	18	3

Source: own elaboration. Word list, generated by Rapid Miner.

Grouping of mostly mentioned terms may be made by their thematic content. Some of them are oriented to:

1. Education: students, universities, teaching, skills, knowledge.
2. Research work: research, innovative, methods, practical, requirements, analysis.
3. Software: informatics, intelligence, business, software, data, products, tools.

Other methods of grouping are k-means clustering, decision trees, CHAID trees, CRT trees exist and they are well-known, but they are not applicable in our example.

Conclusions

Text mining techniques are well known in theory. This paper is an empirical study on the application of text mining techniques for counting total occurrences of terms in several documents. The research aim – to find terminology which is common for all articles in one issue of the journal – is fulfilled. Rapid Miner is used. The operator "Process Documents from Files" to create a word list with "total occurrences" and "documents occurrences". The results and discussion section of this paper shows that a lot of terms are used in all articles of one issue of the Economics and computer science journal. These terms are classified manually in three directions: education research and software. Future research may focus on the application of other text mining techniques using the same dataset or the application of the same text mining techniques with other text documents.

Bibliography

Economics and Computer Science Journal. Retrieved from: http://eknigibg.net/index.php?route=information/information&information_id=8 (7.02.2017).

Feldman, R., Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. New York: Cambridge University Press.

Gantz, J., Reinsel, D. (2012). *The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the Far East*. Retrieved from: http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf (3.06.2016).

Hashimi, H., Hafez, A., Mathkour, H. (2015). Selection criteria for text mining approaches. *Computers in Human Behavior*, *51*, 729–733.

Hearst, M. (1999). *Untangling Text Data Mining*. University of Maryland: Proceedings of the ACL '99.

Kumar, B. S., Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, *114*, 128–147.

The DIMBI project. Retrieved from: http://dimbi.paragonweb.eu/ (7.02.2017).

Rapid Miner. Retrieved from: https://rapidminer.com/ (7.02.2017).

EKSPLORACJA TEKSTÓW ARTYKUŁÓW W JEDNYM Z WYDAŃ CZASOPISMA „ECONOMICS AND COMPUTER SCIENCE" DEDYKOWANEMU PROJEKTOWI DIMBI

Słowa kluczowe: ontologia, eksploracja tekstu, Rapid Miner, projekt DIMBI

Streszczenie. Celem artykułu jest wykorzystanie technik analizy danych biznesowych w celu przeanalizowania artykułów w jednym z numerów (tom 2, wydanie 5) czasopisma „Economics and Computer Science". Ze względu na różnorodne metody analizy danych biznesowych badanie ogranicza się do metod eksploracji tekstu. Celem badania jest znalezienie terminologii, która jest wspólna dla wszystkich artykułów w jednym, wybranym numerze czasopisma. Ponieważ w tym dzienniku poruszono wiele zagadnień tematycznych, to zadaniem badawczym jest to, aby znaleźć ontologie w każdym wydaniu tematycznym. Stosowanym narzędziem i oprogramowaniem do wydobywania tekstu jest Rapid Miner. To oprogramowanie używane jest przede wszystkim do wynajdowania najczęściej używanych słów kluczowych. Nastepnie, ręcznie został stworzony słowniczek pojęć. Głównymi grupami okazały się być: edukacyja, badania i oprogramowanie. Proponowana metoda może być wykorzystywana przez innych autorów dla innych badań o różnej tematyce.

Tłumaczenie Maciej Czaplewski

Cytowanie

Vasilev, J., Marinova, N. (2017). Text mining of articles in an issue of the journal „Economics and Computer Science" dedicated on the DIMBI project. *Ekonomiczne Problemy Usług*, *1* (126/2), 153–159. DOI: 10.18276/epu.2017.126/2-16.