Improving Information-Carrying Data Capacity in Text Mining

Marcin Gibert(✉)

Department of Information Systems Engineering, Faculty of Computer Science,
West Pomeranian University of Technology in Szczecin, Szczecin, Poland
mgibert@wi.zut.edu.pl

Abstract. In this article the relation between the selection of textual data representation and text mining quality has been shown. Due to this, the information-carrying capacity of data has been formalized. Then the procedure of comparing information-carrying data capacity with different structures has been described. Moreover, the method of preparing the γ -gram representation of a text involving machine learning methods and ontology created by the domain expert, has been presented. This method integrates expert knowledge and automatic methods to develop the traditional text-mining technology, which cannot understand text semantics. Representation built in this way can improve the quality of text mining, what was shown in the test research.

Keywords: Text mining · Information-carrying data capacity · Vector space model · Text documents representation

1 Introduction

Data mining is the automatic or semi-automatic process of exploration and analysis of large quantities of data in order to discover meaningful patterns and rules [13]. Depending on the type of analyzed data, proper methods and techniques of data mining are used. In case of data such as text documents, which are usually unstructured or semi structured, the following methods are used, whose main tasks are:

- development of replacement representation (data structuring) of text documents, which gives the possibility of using different techniques of exploration;
- main exploration of textual data with a new structure.

In general, to achieve this goal, the Vector Space Model, in which every text document is represented by a vector of features, is used [2]. These features are the component elements of a text, e.g. single words, called terms. A set of terms is the transformation of an original form of textual data to a more useful text representation for the exploration task. Due to a properly selected structure of textual data, exploration is possible only when concentrating on the construction of a text or on semantic information, which the text carries. It can be a quantity of specified words from a text or precise extraction of semantic information, e.g. a fragment of stock company news,

M. Núñez et al. (Eds.): ICCCI 2015, Part II, LNCS 9330, pp. 648–657, 2015.
DOI: 10.1007/978-3-319-24306-1_63

which provides help for investors in purchasing stocks – *investor related to the board* [16].

Depending on the adopted strategy of textual data representation, except the structure of data, the quantity and quality of important information, which is carried by these data, are also changed. Improper selection of data structure can cause:

- redundancy of representation and subsequently information noise, both of which complicate identifying important information;
- difficulty in extracting semantic information;
- fewer interpretation possibilities of data in the exploration process.

Every element mentioned above can have a negative influence on the result of text mining. Therefore, the relations between the selection of textual data and quantity and quality of important information in the text mining has been characterized and the factor of the information-carrying capacity of data has been defined. Then, three different models of text representation have been described. One of them is γ –gram representation, which can be created by using the machine-learning methods and the information patterns based on the domain ontology. Information patterns are the formal model of factual information, which occurs in the text [10]. Therefore, this approach develops the traditional text-mining technology, which cannot understand text semantics. Based on this solution, the proper structure of data (terms) can be prepared by using ontology in the context of the exploration task. Representation built in this way can improve the information-carrying data capacity, what was shown in the test research.

2 General Rules of Text Mining

Text mining is based on a structure called the semiotic triangle [18]. This is a set of relations between the term, object and concept. The term is a form of linguistic expression, e.g. a single word. The object is part of the reality indicated by the term. The concept is the meaning, which constitutes an image (mapping) of the object in the mind. Due to above relations, text mining can include an analysis on the level of the text construction or semantic information. Depending on the exploration task, the new text representation, which, to a greater extent allows us to focus exploration activities on text construction or on semantic information, is prepared. The Vector Space Model is usually used for preparing this structured representation of a text. Data exploration based on this model is performed in two steps:

- Step I – exploring components of text documents to prepare the proper representation of textual data. This step precedes the main textual data exploration which is called the data preparation step.
- Step II – general exploration of new textual data representation, which is prepared in the previous step.

The role of data preparation in step I, in which methods of data exploration are also used, is eliminating redundant elements occurring in the text, i.e. information noise. This is because information noise could have a negative influence on the selection of proper representation of textual data and then on the exploration result. In this step, the text document, which is usually expressed in natural language form is mapped by replacement representation in the form of terms vector. Preparing a structured representation of textual data in step I usually consists of three main parts [6]:

- tokenization and segmentation;
- lemmatization and stemming;
- reduction and selection of data representation.

Segmentation is the transformation of a text from a linear form to a set of semantic lingual units, from graphemes, trough words, to whole sentences[9]. Sometimes segmentation is understood as tokenization, which is the process of dividing a text into a series of single tokens, usually single words [20].

Lemmatization is the process of reducing words from their original to basic form, called lemma. Stemming is a similar process, which relies on extracting identical part of words, called stems. Because of lemmatization or stemming processes, distinct grammatical forms are regarded as one word, which allows to identify the same word, located in different positions of the text. In practice, there is a possibility of using lemmatization in relation to recognized words in a text and stemming for unrecognized words [7].

In the next step, unimportant words are removed from a large set of words extracted from the text. This reduction is usually achieved by using the stop lists, containing words with little value to the exploration process. These words do not directly affect the meaning of a text, but they shape the course of the utterance. In general, these are words, which are most commonly used in texts of a given language, e.g. conjunctions, pronouns. Sometimes stop rules are used interchangeably, most commonly based on Zipf`s law [14]. These rules, using a given algorithm, specify a set of uninformative words for analyzing text sets. Sometimes, the so-called thesaurus e.g. WordNet is used to remove synonyms from a text [14].

Subsequently, text representation proper to the exploration task is created [21]. There are a few proposals of text representation, which have been described in the literature [5]:

- unigram model of representation – based on single words;
- n-gram model of representation – based on equal lengths of word sequences;
- γ -gram model of representation – based on different lengths of word sequences.

The unigram model of text representation is based on the notation of individual words occurring in a text. This notation builds a vector, which represents textual data. The elements of the vector r are calculated by formula [5]:

$$r = \begin{cases} 1\ when\ \exists j;\ w_j = v_i, v_i \in V \\ 0\ otherwise \end{cases},\ i \le n, j \le k \qquad (1)$$

where:
V – lexicon;
v_i – i^{th} word in lexicon;
n – number of word in lexicon;
w_j – j^{th} word extracted from the text t;
k – number of word extracted from the text.

The n-gram model of text representation is based on notation of the constant length combination of words, which occur successively in a text. This model of text representation is a matrix M created according to the formula [5]:

$$Mx,y = \begin{cases} 1\ when\ (w_j, w_{j+1}, \dots, w_{j+l-1}) = r_x \wedge w_{j+l} \in v_y \\ 0\ otherwise \end{cases} ;\ 1 \leq j + l - 1 \leq \mathrm{k} \qquad (2)$$

where:

$M_{x,y}$ – matrix representing text document, rows x of matrix corresponding to combination of words r_x;
r_x- variation including l words form lexicon V;
w_j – j^{th} word extracted from the text t;
v_y- words from lexicon V corresponding to column y of matrix.

The γ-gram model of the text representation is built by using function $\gamma(w_1,\dots,w_k)$, whose values respond to the usefulness of given word sequences $w_1,\dots,w_k$ in analyzing a text document [5]. Determining the most useful sequence of different lengths is done by using a proper algorithm.

After the step of data representation, a standard procedure of exploration is done. Appropriate weights are calculated for elements of text representations (terms), by using the selected function of importance e.g. tf, idf, tf-idf [15]. Optionally, Latent Semantic Indexing, which detects sematic relations between terms by using algebraic calculation, could be used [8]. Finally, by using the proper measure, e.g. the cosine measure, the similarity between text documents is calculated [22].

3 Preparing Text Representation Based on Machine Learning and on Domain Expert Knowledge

While preparing text representation two approaches should be used as text document analysis is carried out on text construction or semantic information. They are as follows:

- machine learning;
- expert knowledge.

The former, machine learning, is automatic and largely based on statistical-mathematical methods [12]. The latter, expert knowledge, concerns the techniques of knowledge management and is associated strongly with semantic analysis [24] [16].

This approach uses lexical and syntax rules of a given language and takes into account the meaning of analyzing words and phrases. In this solution it is important to have knowledge of analytical language, grammar and specification of utterance, which is related to the used vocabulary. Machine-learning is the most popular way of acquiring it in terms of simplicity of practical application. This is mainly because there is no expert at hand. Therefore, many commercial systems are based on automatic methods. However, it does not mean that using such methods is the most effective approach. P. Gawrysiak wrote [5]: *Rather, it appears that future NLP systems will benefit both expert knowledge of linguists, stored in the form of a knowledge base, and from the systems of automatic analysis too, as it will be able to modify and update this knowledge.* Therefore the optimal solution seems to be integrating machine learning methods and the base of knowledge defined by the expert e.g. in the form of domain ontology.

In the literature, few exemplary methods, which use the base of knowledge defined by the domain expert as a solution to this kind of approach, has been described [16] [19]. For example it is possible to select the proper terms in the Vector Space Model, which are important for exploration purposes, on the basis of domain knowledge [23]. In this case, terms are usually in the form of word sequences of different lengths, which have semantic relations with each other. In fact, this is the γ -gram model of representation, which extracts and effectively uses semantic information carried by the text. This method improves the traditional text-mining technology, which cannot understand text semantics.

The solution proposed by the author is selecting terms in the Vector Space Model based on the information patterns defined by the expert. Information patterns are a general model of semantic information, which are specified by the relation of proper words [24]. There are few proposals of formally defining this kind of information pattern [22][23][24]. In the test research, the Web Ontology Language - OWL - the standard of data description in ontology form, has been chosen for this task [17]. The definition of information patterns by using ontology is natural because of relations occurring between the object, the term and the concept , which are expressed by the semiotic triangle. Thus, every term corresponds to some concept ,, which is identical to the concept used in ontology creation. Moreover, the relations between concepts, included in the information patterns, are mapped in the ontology model.

Factual information, which can be extracted from a text based on a specific algorithm, which uses the information pattern defined by OWL, is for example *investor related to the board.* The definition of the information pattern based on OWL tags for the above factual information is:

```
<owl:Class rdf:ID="investor" />
<owl:Class rdf:ID="related" />
<owl:Class rdf:ID="board" />

<owl:ObjectProperty rdf:ID="whichOne">
<rdfs:domain rdf:resource="#investor" />
<rdfs:range rdf:resource="#related" />
</owl:ObjectProperty>
```

```
<owl:ObjectProperty rdf:ID="withWhom">
<rdfs:domain rdf:resource="#related" />
<rdfs:range rdf:resource="#board" />
</owl:ObjectProperty>
```

The pattern definition by using OWL facilitates the use of proper patterns for a specified task of data exploration. The set of patterns may form the basis of knowledge, which can be used as factual information, which is important for the exploration process. Furthermore, because of the use of an OWL construction it is easy to extend several patterns with new elements. This operation improves the construction process of the total model of knowledge, which is necessary to extract proper factual information for a given exploration task.

4 Rating Text Document Exploration Process with Use of Information-Carrying Capacity

The information-carrying capacity of data depends on all-important and interpretable information, which is carried by data D with structure s and concerns the considered task of exploration. Information-carrying capacity has been defined by the author's formula:

$$N_{inf}(\mathrm{D}, M) = f(\mathrm{W}, R, \zeta) \tag{3}$$

where:

N_{inf} – information capacity of data *D;*
D – textual data;
M– is the set of selected quality measures;
W– quality of data exploration result expressed by the adopted measures;
R – representation of text document;
ζ – is an unknown and unidentified set of information, which has an influence on information capacity of data *D* and on the result of the decision process;
f() – unknown function, whose value is specifying information capacity N_{inf}.

In the absence of knowledge about the function *f()* and the set ζ from formula (3) and in the absence of methods for their identification, the information-carrying capacity N_{inf} of data *D* with structure s_1 can be determined as the rank of information-carrying capacity of data with structure s_2 regarding structure s_2, based on the quality of result *W* of the data exploration process. This is a kind of categorical rating expressing the rank of values. Therefore, the information-carrying capacity of data *D* with structure s_1 is greater than the information-carrying capacity of data *D* with structure s_2 in case of achieving the result in the exploration process, which is characterized by higher quality according to the adopted measure *W*, e.g. precision, recall, accuracy, specificity, fall-out [1]. The adopted quality measure should be related to the defined aim of the exploration process [25].

The whole set of quality measures is also possible to take into account in the rating of information-carrying data capacity. Therefore, comparison of the information-carrying data capacity based on an adopted set of different quality measures, is performed by using one of the multi-criteria decision methods., e.g. Electre I [3].

There are many multi-criteria decision methods e.g. AHP, ANP, Electre I Promethee. Each of them works differently and is used for different tasks. The comparison of multi-criterion decision methods and a detailed description of them is presented in the literature[4]. For the decision problem, which includes selecting the best decision variant, the Electre I method has been adopted for the exemplary research. This method is used for problems of selecting the best decision variant. Electre I allows every decision variant to specify the weight of importance and the searching direction of the best decision. Subsequently, the relations between decision variants are appointed. The graph of superiority relations specifies their levels from the best to the worst variants. The highest value of the exploration quality measure combination could be obtained by using the best variant of data structure from this graph.

The quality rating of textual data exploration is made on selected levels of recall. Depending on the exploration task it can be either on one level or a whole set of selected levels of recall, e.g. above value 0.5. Therefore, the final rating of decision variants by using for example the Electre I method is made on the selected levels of recall. The final superiority factor of information-carrying capacity, is calculated, according to formula:

$$W_{N_{inf}} = \sum_{o=1}^{u} \begin{cases} 1\ when\ p_o = 1 \\ 0\ otherwise \end{cases} \tag{4}$$

where:

po- level of variant achieved in Electre I method, for recall o;
u – maximum level of recall;
po=1 – means that for recall o, the variant is on the first level of superiority graph.

5 Research on the Influence of Different Textual Data Structures on Quality of Exploration Process – Case Study

The test research was carried out on the corpus of 200 texts, derived from information for Polish stock company investors. The task of exploration was to identify company news, which gives the latter signals to buy stocks. For example this is news, in which information is given about the purchase of company stocks by people, who are in relation with members of the Board. In the first step, there is a set of 10 news in the analyzing corpus, which should be returned in ranking. This news determines 10 levels of recall. Next, those textual data structures were selected, which would be ranked in respect to the information-carrying capacity in the identical exploration process. The γ -gram model of representation has been selected on the basis of information patterns created by the expert by using the Web Ontology Language, according to the procedure described in chapter 3.

Finally, the set of quality measures, which was taken into account on the specified level of recall in rating, has been defined. For research purposes, three measures of quality have been selected, namely precision, negative prediction value and recall. Recall has values equal or higher than *0.5*. The example, taken into account in the selection of the most favorable ones, in respect of expert preferences, variant in Electre I method, has been presented in table 1.

Table 1. The data, which were taken into account in the rating of the most favourable decision variant for level of recall equal 0.8

For level of recall = 0,8			
criteria	precision	negative prediction value	Final rank (after normalization)
Direction of optimization	max	max	
weight	5	4	
Unigram model of representation	0,42	0.99	0,6
2-gram model of representation	0.12	0.98	0,4
γ -gram model of representation	0.62	0.99	1

Finally, according to the procedure, which has been described in chapter 4, the rank of data information capacity for individual structures has been calculated. All calculations in the test research have been made by using the author`s software, based on JAMA - basic linear algebra package for Java. On analyzing the results of the calculation it can be said that the most favorable data for the exemplary research has a γ-gram structure s_3, the second in order is a unigram structure s_1 and the last one is a 2-gram structure s_2.

6 Summary

In this article the importance of the correct choice of data structures in the text mining process has been described. The possibility of extracting more valuable knowledge was indicated, while taking into account the selected data structures in order to consider the problem of exploration. The relation between selecting terms and text mining quality has been shown. The method, which allows to select the proper term in the Vector Space Model has been presented. This method uses the information patterns defined by the Web Ontology Language to create better model of text representation for considering the exploration task. Moreover, by using the information-carrying data capacity factor, the procedure, which allows to compare the different data structure influence on a specified set of text mining quality measures, has been presented. Finally, in the article comparative research results of text mining quality has been presented with exemplary textual data of different structures. One of these structures, γ –gram representation, which is prepared by using machine-learning methods and ontology created by the domain expert, has proved to be capable of improving the information-carrying capacity of data, shown in this research paper.

References

1. Amiri, I.S., Akanbi, O.A., Fazeldehkordi, E.: A Machine-Learning Approach to Phishing Detection and Defense, p. 27 (2014)
2. Mbarek, R., Tmar, M., Hattab, H.: A New Relevance Feedback Algorith Based on Vector Space Basis Change. Computational Linguistics and Intelligent Text Processing **2**, 355–356 (2014)
3. Munier, N.: A Strategy for Using Multicriteria Analysis in Decision-Making: A Guide for Simple and Complex Environmental Projects, pp. 59–65 (2011)
4. Velasquez, M., Hester, P.T.: An Analysis of Multi-Criteria Decision Making Methods. International Journal of Operations Research **10**(2), 56–66 (2013)
5. Gawrysiak, P.: Automatyczna kategoryzacja dokumentów, pp. 36–45 (2001)
6. Ramasubramanian, C., Ramya, R.: Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm. International Journal of Advanced Research in Computer and Communication Engineering, Valume **2**(12), 4537 (2013)
7. Weiss, S.M., Indurkhya, N., Zhang, T.: Fundamentals of Predictive Text Mining, pp. 17–19 (2010)
8. Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W.: Handbook of Latent Semantic Analysis, p. 10 (2013)
9. Dale, R., Moisl, H., Somers, H.: Handbook of Natural Language Processing, p. 11 (2000)
10. Luna Dong, X., Gabrilovich, E., Murphy, K., Dang, V., Horn, W., Lugaresi, C., Sun, S., Zhang, W.: Knowledge_based Trust: Estimating the Trustworthiness of Web Sources. Computer Science Database (2015)
11. Gentile, A.L., Basile, P., Iaquinta, L., Semeraro, G.: Lexical and semantic resources for NLP: from words to meanings. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part III. LNCS (LNAI), vol. 5179, pp. 277–284. Springer, Heidelberg (2008)
12. Kononenko, I., Kukar, M.: Machine Learning and Data Mining, p. 17 (2007)
13. Berry, M., Linoff, G.: Mastering Data Mining: The Art and Science of Customer Relationship Management, p. 7 (2004)
14. Esposti, M.D.: Mathematical Models of Textual Data: A short Review, pp. 100–102 (2014)
15. Sabbah, T., Selemat, A.: Modified Frequency-Based Term Weighting Scheme for Accurate Dark Web Content Classification, pp. 185–187 (2014)
16. Jackson, P., Moulinier, I.: Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization, Amsterdam, vol. 5, pp. 125–126 (2007)
17. Bechhofer, S., Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F., Stein, L.A.: OWL Web Ontology Language (2015). http://www.w3.org/TR/owl-ref/
18. Merkelis, R.: Philosophy and Linguistics, p. 12 (2013). http://www.slideshare.net/robertasmerkelis/philosophy-and-linguistics-28940425
19. Jiang, L., Zhang, H.-b., Yang, X., Xie, N.: Research on semantic text mining based on domain ontology. In: Li, D., Chen, Y. (eds.) Computer and Computing Technologies in Agriculture VI, Part I. IFIP AICT, vol. 392, pp. 336–343. Springer, Heidelberg (2013)
20. Chakraborty, G., Pagolu, M., Satshi, G.: Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS, p. 70 (2014)
21. Sanders, T., Schilperoord, J., Spooren, W.: Text Representation: Linguistic and Psycholinguistic Aspects, pp. 1–19 (2001)
22. Śmiałkowska, B., Gibert, M.: The classification of text documents by using Latent Semantic Analysis for extracted information. Ekonomiczne Problemu Usług No. 106, Zeszyty Naukowe Uniwersytetu Szczecińskiego No. 781, pp. 345–358 (2013)

23. Smialkowska, B., Gibert, M.: The classification of text documents in Polish language by using Latent Semantic Analysis for extracted information. Theoretical and applied informatics **25**, 239–250 (2013)
24. Lubaszewski, W.: Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu (2009)
25. Wang, R.Y., Strong, D.M.: What data quality means to data consumers. Journal of Management Information Systems **12**(4), 7 (1996)