Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 18 (2013) 1198 – 1207

Procedia Computer Science

2013 International Conference on Computational Science

Data Pre-Processing Evaluation for Text Mining: Transaction/Sequence Model

Daša Munková[a], Michal Munk[a], Martin Vozár[a,]*

[a]Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, 949 74 Nitra, Slovakia

Abstract

Data pre-processing presents the most time consuming phase in the whole process of knowledge discovery. The complexity of data pre-processing depends on the data sources used. The aim of this work is to determine to what extent it is necessary to carry out the time consuming data pre-processing in the process of discovering sequential patterns in e-documents. We used the transaction/sequence model for text representation and sequence rule analysis as a method of modelling. We compare four datasets of different quality obtained from texts and pre-processed in different ways: data with identified the paragraph sequences, data with identified the sentence sequences, data with identified the paragraph sequences without stop words and data with identified the sentence sequences without stop words. We try to assess the impact of these advanced techniques of data pre-processing on the quantity and quality of the extracted rules. The results confirm some initial assumptions, but they also show that the stop words removal has a substantial impact on the quantity and quality of extracted rules in case of paragraph sequence identification. Contrary, in case of sentence sequence identification, removing the stop words has not any significant impact on the quantity and quality of extracted rules.

Selection and peer review under responsibility of the organizers of the 2013 International Conference on Computational Science

Keywords: Data pre-processing; stop words; sequence identification; transaction/sequence model; text mining; evaluation

1. Introduction

The present era is characterised by the amount of available electronic data on one hand, but often a lack of knowledge on the other hand [1]. A huge amount of data has a weak predictive value. The concept of Knowledge Discovery was created for this purpose [2]. We understand knowledge discovery as a process involving data collection, data pre-processing, data transformation, data analysis and results interpretation [3]. Knowledge discovery is characterised with a wide range of variables and data sources.

* Corresponding author. Tel.: +421-37-640-8678 ; fax: +421-37-640-8556 .
E-mail address: mvozar@ukf.sk .

1877-0509 © 2013 The Authors. Published by Elsevier B.V.
Selection and peer review under responsibility of the organizers of the 2013 International Conference on Computational Science
doi:10.1016/j.procs.2013.05.286

Knowledge Discovery in Databases (KDD) is the most common area of knowledge discovery. Fayyad defined this area as a non-trivial acquisition of hitherto unknown and potentially useful implicit information from data [3]. KDD has a methodological background in databases, statistics and machine learning. Production databases and data warehouses are data sources.

In this case, where data is obtained from text, this process is call text mining or knowledge discovery in texts, where the electronic documents are the data source. Knowledge discovery in texts is analogous with knowledge discovery in databases [4]. Similarly, Sullivan understands knowledge discovery in texts in concordance with the general definition of knowledge discovery [5]. The biggest differences are in data pre-processing itself, which means how to represent a text for use with analytical methods. Knowledge discovery in texts involves many scientific fields. Similarly like in KDD, statistical methods and methods of machine learning are used as the tools for data analysis in knowledge discovery in texts. However, knowledge discovery in texts builds on theoretical and computational linguistics by data pre-processing [6-9].

The biggest differences among areas of knowledge discovery are during the phase of data pre-processing in the process of managing CRISP-DM methodology [2]. Data pre-processing presents the most time consuming phase in the whole process of knowledge discovery. The complexity of data pre-processing depends on the data sources used. A data file of M variables and N cases is an input to an analytical procedure. Not only data transformation into analysis-ready format (by analytical tools into the required form) is the matter of data pre-processing, but also the data quality itself.

Relatively, the simplest data pre-processing is in the case of the use of data as a data source. Simply, the data pre-processing consists of data selection, cleaning, creating, integrating and formatting [10]. In the case of data warehousing, there is no data cleaning and integrating, which reduces the data pre-processing even more.

In electronic documents, the stem words would be the variables. The weights of individual stem words would be the cases in each text of documents, i.e. each text document represents a vector of weights primarily divided by frequency of incidences of stem words. Simply, data pre-processing consists of converting document into a plain text and then of stem word identification. More detail about e-document representation and data pre-processing for the purpose of data analysis is dealt with in books on the subject of text mining, content analysis etc. [6-7], [9], [11].

The aim of this work is to determine to what extent it is necessary to carry out the time consuming data pre-processing in the process of discovering sequential patterns in e-documents. For this purpose, an experiment was conducted focusing on data pre-processing in e-documents. We used the transaction/sequence model for text representation [12], and we were influenced by contributions [13-14] during the realisation of an experimental plan.

The rest of the paper is structured subsequently: in section 2 we summarize related work of other authors that deal with data pre-processing issues in the field of usage of text mining and text representation. We summarize transaction/sequence model in section 3. Subsequently, we particularize research methodology in section 4. This section describes how we prepared texts in different levels of data pre-processing. Section 5 provides a summary of the experiment results in detail. Finally, we discuss obtained results in section 6.

2. Related work

The disadvantage of electronic documents over databases consists of unstructured data. The solution is to create variables that will adequately represent the analysed text. Text representation is the essential step for text pre-processing.

Text (document) is a sequence of words [15] represented by an array of words. Most often a text of document is represented by a vector. Vector has as many components as stem words in dictionary or examined document collection. This results in disadvantages such as high dimension vector (a large number of variables) and sparse vectors (missing values). Each stem word for the document can be coded: world frequency – simple

frequency of word incidence; binary frequency – binary indicator of incidence; log frequency - logarithm function provides "damping" of word incidences, i.e. stabilization of variance [2] and inverse document frequency, which presents a very useful transformation taking into account the specificity of words and a total frequency of incidences [16-17].

Besides the most used vector space model (VSM) for text representation or document representation, a tensor space model (TSM) is used. TSM unlike the SSM models a text by multilinear algebraic high-order tensor instead of the traditional vector [18].

Another text representation - a matrix representation of document was designed by Xufei Wang et al. The document is a set of segments where rows represent distinct terms and columns represent cohesive segments [19].

The other methods based on different text representation as a vector are n-grams representation [20], Nature Language Processing [21-22], Bag-Of-Words [23-24] or Distributional Word Clusters [25]. But all these methods consider only term frequency of words incidences in texts and therefore ignoring the significance sequences in which they occur.

As in our case, there are other approaches taking into account the position of token in the text [26].

The aim of an experiment is to determine to what extent removing the stop words influences on sequence rules in short texts. There are many scientists [27] who state, that with removing the stop words we miss important information or lose, the text receives different meaning.

3. Transaction/sequence model

Text mining is analogous to KDD. Sometimes it is enough to slightly adapt the existing methods and procedures from other areas of knowledge discovery. In our case we chose a quite unusual representation of short texts, and we found the inspiration in area of KDD and web usage mining. We used the transaction/sequence model for text representation, which allows us to examine the relationships between the examined attributes and search for associations among the identified tokens (content words or stop words) in the texts. Similarly, like in market basket analysis, a transaction represents one purchase, or in web analysis it represents the set of user's visited pages during one session, in our case it is a set of words in short written texts.

The structure and data character predetermine the use of specific methods for analysis – data modelling. In case of the use of transaction/sequence model for text representation, it is mainly association rule analysis and sequence rule analysis. The difference between the association and the sequence rule analysis is that we do not analyse the sequences but the transactions in association rule analysis, which means, we do not include the sequence variable representing the order of the words in text into the analysis. The transaction represents a set of the words occurred in text, whereby the order of incidence of the identified words (content words or stop words) in the given text is not taken into account.

Association/sequence rule analysis has its application also in area of quantitative syntax analysis [28]. Specially, in our case, we focused on data from the above mentioned area by the evaluation of data pre-processing in process of knowledge discovery in texts.

Examined variables:

- Text ID,
- Paragraph ID,
- Sentence ID - within a paragraph,
- Transaction/Sequence ID - a set ID of tokens in text, it consists of previous two/three variables,
- Sequence - an order of words in text/paragraph/sentence,
- Content word - word that refers to object, action or property,

- Part of speech - words classification (nouns, verbs, adjectives, adverbs, articles, pronouns, prepositions, conjunctions and the rest-unclassified),
- Stop words - words which do not contain important significant information or occur so often that in text that they lose their usefulness,
- Snowball category - there is not one definite list of stop words, for our experiment we used a list defined in Snowball project.

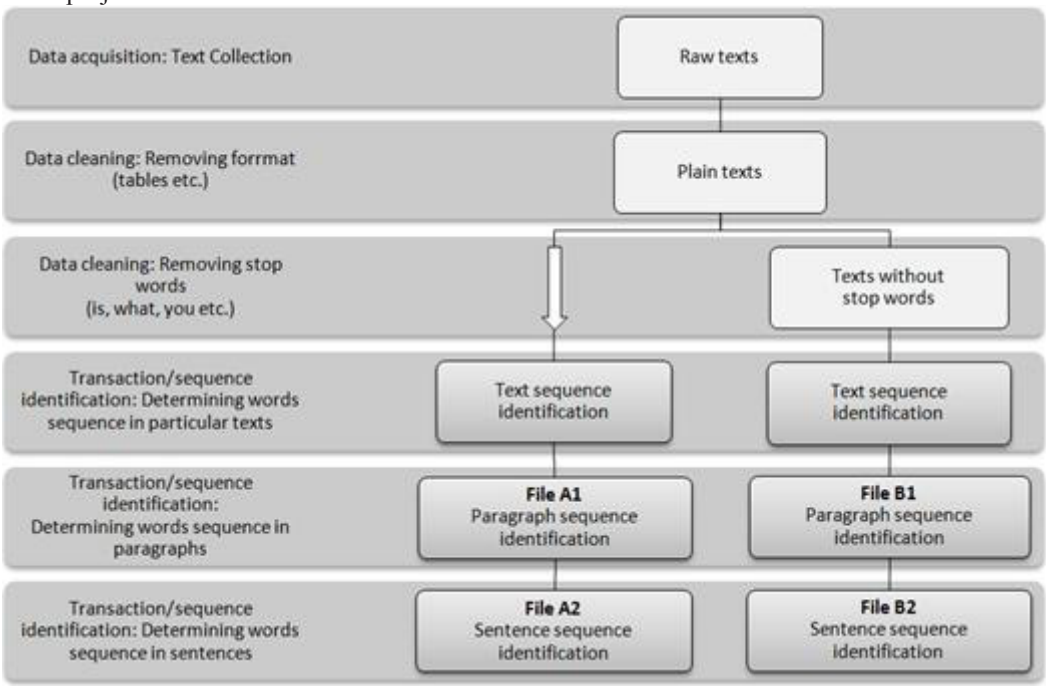

Fig. 1 Application of techniques of data pre-processing.

4. Research Methodology of Experiment

We aimed at specifying the inevitable steps that are required for gaining valid data from the texts represented by transaction/sequence model. Specially, we focused on the sequence identification and data cleaning (Fig. 1). We tried to assess the impact of this advanced techniques on the quantity and quality of the extracted rules which represent the sequential patterns.

Experiment was realized in several steps.

1. Text Collection.
2. Format removal.
3. Data pre-processing on various levels (Fig. 1):
   3.1. with an paragraph sequence identification (File A1),
   3.2. with an sentence sequence identification (File A2),
   3.3. with stop words removal and an paragraph sequence identification (File B1),
   3.4. with stop words removal and an sentence sequence identification (File B2).
4. Data analysis - searching for sequential patterns in individual files. We used *STATISTICA Sequence, Association and Link Analysis* for sequence rules extraction. It is an implementation of algorithm using the

powerful a-priori algorithm [29-32] together with a tree structured procedure that only requires one pass through data [33-34].
5. Understanding the output data - creation of data matrices from the outcomes of the analysis, defining assumptions.
6. Comparison of results of data analysis elaborated on various levels of data pre-processing from the point of view of quantity and quality of the found rules – sequential patterns.

We articulated the following assumptions:
1. we expect that removing stop words will have a significant impact on the quantity of extracted rules,
2. we expect that removing stop words will have a significant impact on the quantity of extracted rules in terms of decreasing the portion of inexplicable rules,
3. we expect that removing stop words will have a significant impact on the quality of extracted rules in the term of their basic measures of the quality.

5. Results

5.1. Data understanding

The analysed text documents consisted of short texts. After the conversion into the plain text, parsing and tokenisation, 57.41 % of content words were identified. Based on Snowball list of stop words 42.59 % of stop words were determined in our short texts. Among the most frequent content words belonged the nouns with portion higher than 22 %, the uncategorised (based on our simple rules in English morphology) and the verbs with portion higher than 15 %, then the pronouns and prepositions, each with portion approximately 10 % of tokens. The remaining content words: conjunctions, articles and adverbs achieved the portion less than 7 % from the point of view of incidence.

Based on snowball stop words list, the most occurred stop words (they are classified into several categories) the rest (overlap among prepositions, conjunctions, adverbs etc.) with portion higher than 17 % and the pronouns forms (demonstratives, interrogatives, reflexive and possessive adjectives or pronouns) with portion more than 8 %. The articles with portion approximately 7 % also belonged to the most frequent stop words. The commonest (the most common in English texts), verb forms (verb+negation, pronoun+verb and auxiliary+negation) and auxiliaries achieved the portion lower than 5 % from the point of view of incidence.

5.2. Comparison of the quantity of extracted rules in examined files

The analysis (Table 1) resulted in sequence rules, which we obtained from frequented sequences fulfilling their minimum support (in our case min s = 0.1). Frequented sequences were obtained from identified sequences based on the length of the text (paragraph/sentence).

There is a high coincidence between the results (Table 1) of sequence rule analysis in terms of the portion of the found rules in case of files with sentence sequence identification (File A2, File B2). The most rules were extracted from file with paragraph sequence identification without stop words; concretely 75 were extracted from the file (File B1), which represents over 69 % of the total number of found rules. Based on the results of Q test (Table 1), the zero hypothesis, which reasons that the incidence of rules does not depend on individual levels of data pre-processing for text mining is rejected at the 1 % significance level.

Kendall´s coefficient of concordance represents the degree of concordance in the number of the found rules among examined files. The value of coefficient (Table 2) is 0.25, while 1 means a perfect concordance and 0 represents discordance. Low value of coefficient confirms Q test results.

From the multiple comparison (Tukey HSD test) two homogenous groups (Table 2) consisting of files (File A2, File B2) and (File A1, File B1) were identified in term of the average incidence of the found rules.

Statistically significant differences on the level of significance 0.05 in the average incidence of found rules were proved between files with paragraph sequence identification (File A1, File B1) and files with sentence sequence identification (File A2, File B2).

Table 1. Incidence of discovered sequence rules in particular files.

Body	==>	Head	File A1	File B1	File A2	File B2
(verb)	==>	(definite article)	1	0	0	0
…		…				
(noun), (adjective)	==>	(verb)	0	1	0	0
…		…				
(adjective)	==>	(noun)	1	1	1	1
Count of derived sequence rules			59	75	23	22
Percent of derived sequence rules (Percent 1's)			54.63	69.44	21.30	20.37
Percent 0's			45.37	30.56	78.70	79.63
Cochran Q Test		Q = 79.3260; df = 3; p < 0.0000				

Table 2. Homogeneous groups for incidence of derived rules in examined files.

File	Mean	1	2
File B2	0.2037	****	
File A2	0.2130	****	
File A1	0.5463		****
File B1	0.6944		****
Kendall Coeff. of Concordance		0.2448	

Identification of sequence based on length (paragraph/sentence) has an important impact on the quantity of extracted rules. Naturally, we obtained higher number of frequented sequences and as well as rules from the longer sequences. On the contrary, removing the stop words has no significant impact on the quantity of extracted rules.

Now, we will look at the results of sequence analysis more closely, while taking into consideration the portion of each kind of the discovered rules [35]. We require from association rules that they be not only clear but also useful. Association analysis produces the three common types of rules [36]:

- the useful (utilizable, beneficial),
- the trivial,
- the inexplicable.

In our case upon sequence rules we will differentiate same types of rules.

The only requirement (validity assumption) of the use of chi-square test is high enough expected frequencies [37]. The condition is violated if the expected frequencies are lower than 5. The validity assumption of chi-square test in our tests is violated. This is the reason why we shall not prop ourselves only upon the results of Pearson chi-square test, but also upon the value of calculated contingency coefficient and graphic visualization of dependency.

Contingency coefficients (Coef. C, Cramér's V) represent the degree of dependency between two nominal variables. The values of coefficient (Table 3) are higher than 0.3. There is a medium dependency among the

portion of the useful, trivial and inexplicable rules and their incidence in the set of the discovered rules extracted from the data matrix File A1/File B1, the contingency coefficients are statistically significant.

Table 3. Crosstabulations - Incidence of rules x Types of rules: (a) File A1; (b) File B1.

Incidence\Types	File A1			File B1		
	useful	trivial	inexp.	useful	trivial	inexp.
0	6	9	33	9	15	8
	28.57%	27.27%	62.26%	39.13%	45.45%	15.09%
1	15	24	20	14	18	45
	71.43%	72.73%	37.74%	60.87%	54.55%	84.91%
∑	21	33	53	23	33	53
	100%	100%	100%	100%	100%	100%
Pearson Chi-square	12.8692; df = 2; p = 0.0016			10.3813; df = 2; p = 0.0056		
Contingency Coef. C	0.3277			0.2949		
Cramér's V	0.3468			0.3086		

Table 4. Crosstabulations - Incidence of rules x Types of rules: (a) File A2; (b) File B2.

Incidence\Types	File A2			File B2		
	useful	trivial	inexp.	useful	trivial	inexp.
0	2	5	5	3	8	2
	25.00%	29.41%	50.00%	37.50%	47.06%	20.00%
1	6	12	5	5	9	8
	75.00%	70.59%	50.00%	62.50%	52.94%	80.00%
∑	8	17	10	8	17	10
	100%	100%	100%	100%	100%	100%
Pearson Chi-square	1.5814; df = 2; p = 0.4535			1.9751; df = 2; p = 0.3725		
Contingency Coef. C	0.2079			0.2311		
Cramér's V	0.2126			0.2376		

The coefficient values (Table 4) are higher than 0.2, while 1 represents perfect dependency and 0 means independency. There is a small dependency among the portion of the useful, trivial and inexplicable rules and their incidence in the set of the discovered rules extracted from the data matrix File A2/File B2, and the contingency coefficients are not statistically significant.

The graphs (Fig. 2) visualize interaction frequencies – File A1/File B1 x Types of rules. Curves in this case are not copied too, they have different course – which only proves the results of the analysis.

The portion of the useful, trivial and inexplicable rules depends on paragraph sequence identification and does not depend on sentence sequence identification. Removing stop words has impact on increasing the portion of inexplicable rules, in case of paragraph identification, this impact is stronger. In addition, it has an impact on decreasing the portion of trivial and useful rules.

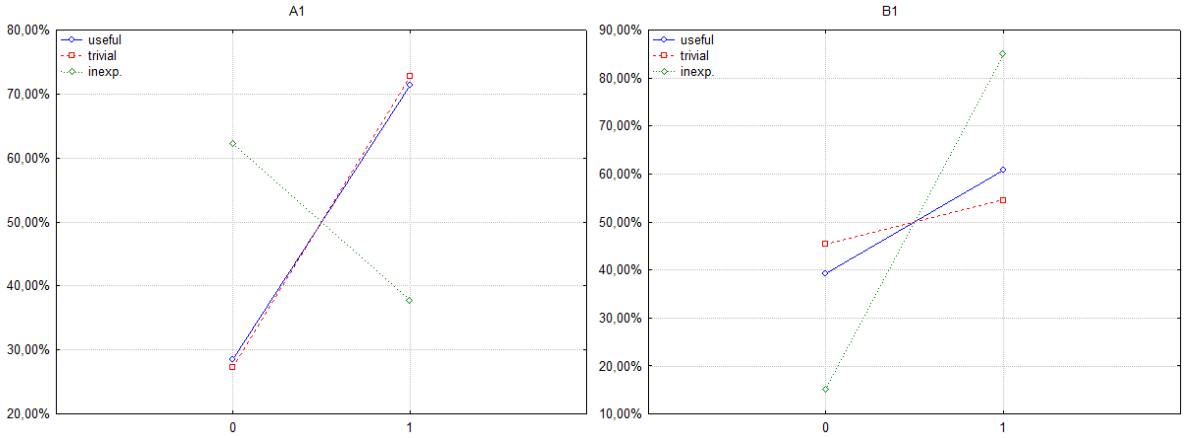

Fig. 2 (a) Interaction Plot - File A1 x Types of rules; (b) Interaction Plot - File B1 x Types of rules.

5.3. Comparison of the quality of extracted rules in examined files

Quality of sequence rules is assessed by means of two indicators [36]: support and confidence. Results of the sequence rule analysis showed differences not only in the quantity of the found rules, but also in the quality. Kendall´s coefficient of concordance represents the degree of concordance in the support of the found rules among examined files. The value of coefficient (Table 5a) is 0.5, while 1 means a perfect concordance and 0 represents discordancy.

From the multiple comparison (Tukey HSD test) a single homogenous group (Table 5a) consisting of files File A2, File B2 and File A1 was identified in term of the average support of the found rules. Statistically significant differences on the level of significance 0.05 in the average support of found rules were proved among File B1 and the remaining ones.

Table 5. Homogeneous groups for (a) support of derived rules; (b) confidence of derived rules.

Support	Mean	1	2
File A2	18.1373	****	
File B2	22.3232	****	
File A1	26.7797	****	
File B1	35.7627		****
Kendall Coeff. of Concordance		0.5000	

Confidence	Mean	1	2
File A2	33.4732	****	
File B2	41.0641	****	
File A1	42.1873	****	
File B1	53.4090		****
Kendall Coeff. of Concordance		0.5000	

There were demonstrated differences in the quality in terms of confidence characteristics values of the discovered rules among individual files. The coefficient of concordance values (Table 5b) is 0.5, while 1 means a perfect concordance and 0 represents discordancy.

From the multiple comparison (Tukey HSD test) a single homogenous group (Table 5b) consisting of files File A2, File B2 and File A1 was identified in term of the average confidence of the found rules. Statistically significant differences on the level of significance 0.05 in the average confidence of found rules were proved among File B1 and the remaining ones. Results (Table 5a, Table 5b) show that the largest degree of concordance in the support and confidence is among the rules found in the file with sentence sequence identification (File A2, File B2) and with paragraph sequence identification including stop words (File A1). On

the contrary, discordancy is among file with paragraph sequence identification without stop words (File B1) and the remaining files (File A2, File B2 and File A1). In case of paragraph sequence identification, removing stop words has a substantial impact on the quality of extracted rules. Contrary, in case of sentence sequence identification, removing stop words has not any significant impact on the quality of extracted rules.

6. Discussion and Conclusions

The experiment contribution in area of knowledge discovery in texts could be evaluated from the point of view of text representation and data pre-processing. In term of data pre-processing, it is a methodology design and recommendations for reliable data acquisition from e-documents in process of discovering the sequence patterns. The contribution, in term of text representation, consists in the design and description of transaction/sequence model.

The first assumption was not proved; removing stop words has no significant impact on the quantity of extracted rules. Identification of sequence based on length (paragraph/sentence) has an important impact on the quantity of extracted rules. Naturally, we obtained higher number of frequented sequences and as well as rules from the longer sequences.

The second assumption was also not proved. The impact of removing stop words for reducing the portion of inexplicable rules was not proved in case of paragraph sequence identification as a sentence sequence identification. On the contrary, removing the stop words caused the increase of inexplicable rules. We found that the portion of the useful, trivial and inexplicable rules depends on paragraph sequence identification and does not depend on sentence sequence identification. After removing stop words, the portion of inexplicable rules increased in case of sequences assigned by sentences as well as paragraphs. While in case of paragraph sequence, this portion was statistically significant. In this case, it may be the merging of unrelated sequences and the increase of inexplicable rules is stronger with removing the stop words. In addition, it has an impact on decreasing the portion of trivial and useful rules.

The third assumption was proved partially. Removing the stop words has an impact on quality of extracted rules only in case of paragraph sequence identification. It was showed removing stop words has a substantial impact on the quality of extracted rules in case of paragraph sequence identification. Contrary, in case of sentence sequence identification, removing stop words has not any significant impact on the quality of extracted rules.

We recommend sequences identification based on sentences when transaction/sequence model is used, assuming that the solution of particular problem does not require other approach to sequence identification. Naturally, this approach to sequence identification seems to be the most suitable in case of problems solving from the area of quantitative syntax analysis [28].

The question remains regards removing stop words which cause an increase of inexplicable rules. Their increase was identified in case of paragraph sequence identification as sentence sequence identification. In addition, stop words removal caused a decrease of useful and trivial rules. Therefore, in the further research we will focus on identifying an impact of various categories of stop words in knowledge extraction and suggest possible stop words reduction.

References

[1] Paralič, J. 2003. *Objavovanie znalostí v databázach*. Košice: Elfa; 2003. 80 p. ISBN 80-89066-60-7.

[2] Houškova Beranková, M., Houška, M. 2011. Data, information and knowledge in agricultural decision-making. In *Agris On-line Papers in Economics and Informatics*, 2011. 3(2): p. 74-82.

[3] Fayyad, U. et al. 1996. *Advances in Knowledge Discovery and Data mining*. AAAI Press/MIT Press; 1996.

[4] Hearst, M. A. 1999. Untangling text data mining. In: *ACL*, 1999. p. 3-10.

[5] Sullivan, D. 2001. *Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing and Sales*. John Willey & Sons; 2001.
[6] Neuendorf, K.A. 2002. *The Content Analysis Guidebook*. London: Sage; 2002. 320 p. ISBN 978-0-7619-1978-0.
[7] Titscher, S. et al. 2002. *Methods of Text and Discourse Analysis*. London: Sage; 2002. ISBN 978-0-7619-6483-4.
[8] Hajičová, E., Panevová, J., Sgall, P. 2003. *Úvod do teoretické a počítačové lingvistiky*. Praha: Karolinum; 2003. ISBN 80-246-0470-1.
[9] Weiss, S. M. et al. 2005. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Springer; 2005. ISBN 978-0-387-95433-2.
[10] Chapman, P. et al. 2000. *CRISP-DM 1.0 Step-by-step data mining guide.* SPSS; 2000.
[11] Paralič, J. et al. 2010. *Dolovanie znalostí z textov*. Košice: Equilibria; 2010. 184 p.
[12] Munková et al. 2012. Analysis of Social and Expressive Factors of Requests by Methods of Text Mining. In: *Pacific Asia Conference on Language, Information and Computation, PACLIC 26*, 2012. p. 515–524.
[13] Munk, M., Kapusta, J., Švec, P. 2010. Data Preprocessing Evaluation for Web Log Mining: Reconstruction of Activities of a Web Visitor. In: *International Conference on Computational Science, ICCS 2010*, *Procedia Computer Science*, 2010. 1(1): p. 2273-2280.
[14] Munk, M., Drlík, M. 2011. Impact of Different Pre-Processing Tasks on Effective Identification of Users' Behavioral Patterns in Web-based Educational System. In: *International Conference on Computational Science, ICCS 2011, Procedia Computer Science*, 2011. 4: p. 1640-1649.
[15] Leopold, E., Kindermann, J. 2002. Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?. In: *Machine Learning*, 2002. 46: p. 423-444.
[16] Salton, G. 1971. *The SMART Retrieval System: Experiments in Automatic Document Proccesing*. Prentice-Hall; 1971.
[17] Manning, C. D., Schütze, H. 2002. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press; 2002.
[18] Ning Liu, Benyu Zhang, Jun Yan, Zheng Chen, Wenyin Liu, Fengshan Bai, Leefeng Chien. 2005. Text Representation: From Vector to Tensor. In: *IEEE International Conference on Data Mining*, *ICDM*, 2005. p.725-728.
[19] Xufei Wang, Jiliang Tang, Huan Liu. 2011. Document Clustering via Matrix Representation. In: *IEEE International Conference on Data Mining*, *ICDM*, 2011. p. 804-813.
[20] Caropreso, M. F., Matwin, S., Sebastiani, F. 2001. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In: *Text Databases and Document Management: Theory and Practice*, 2001. p. 78-102.
[21] Basili, R., Moschitti, A., Pazienza, M. T. 2000. Language-sensitive text classification. In: *Proceedings of RIAO'00, 6th International Conference Recherche d'Information Assistee par Orinateur*, 2000. p. 331-343.
[22] Jacobs, P. S. 1992. Joining statistics with NLP for text categorization. In: *Proceedings of the Third conference on Applied Natural Language Processing*, 1992. p. 178-185.
[23] Dumais, S. T., Platt, J., Heckerman, D., Sahami, M. 1998. Inductive learning algorithms and representations for text categorization. In: *Proceedings of CIKM'98, 7th ACM International Conference on Information and Knowledge Management*, 1998. p. 148-155.
[24] Joachims, T. 1998. Text Categorization with support vector machines: learning with many relevant features. In: *Proceedings of ECML'98, 10th European Conference on Machine Learning*, 1998. p. 137-142.
[25] Bekkerman, R., El-Yaniv, R., Tishby, N., Winter, Y. 2002. Distributional Word Clusters vs. Words for Text Categorization. In: *Journal of Machine Learning Research*, 2002. 1: p. 1-48.
[26] Yih-kuen Tsay, Yu-fang Chen. 2008. *Introducing the Sequence Model for Text Retrieval*. 2008.
[27] Furdík, K., Bednár, P. 2009. Using Jbowl Library for Natural Language Processing (in Slovak). In: *Varia XVI, Proc. of 16th Colloquium of Young Linguists*, 2009. p. 122-131.
[28] Köhler, R. 2012. *Quantitative Syntax Analysis*. De Gruyter: Berlin; 2012. ISBN 978-3-11-027219-2.
[29] Agrawal, R., Imielinski, T., Swami, A. N. 1993. Mining association rules between sets of items in large databases. In: *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 1993.
[30] Agrawal, R., Srikant, R. 1994. Fast Algorithms for Mining Association Rules in Large Databases. In: *Proceedings of the 20th International Conference on Very Large Data Bases*, 1994.
[31] Han, J., Lakshmanan, L.V.S., Pei, J. 2001. Scalable frequent-pattern mining methods: an overview. In: *Tutorial notes of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2001.
[32] Witten, I. H., Frank, E. 2000. *Data Mining: Practical Machine Learning Tools and Techniques*. New York: Morgan Kaufmann; 2000.
[33] Electronic Statistics Textbook. 2010. Tulsa: *StatSoft*; 2010.
[34] Skorpil, V., Stastny, J. 2008. Comparison of Learning Algorithms. In: *24th Biennial Symposium on Communications*, Kingston, Canada, 2008. p. 231-234.
[35] Balogh, Z., Turcani, M. 2011. Possibilities of Modelling Web-Based Education Using IF-THEN Rules and Fuzzy Petri Nets in LMS. In: *Communications in Computer and Information Science*, 2011. 251: p. 93-106.
[36] Berry, M. J., Linoff, G.S. 2004. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Wiley Publishing; 2004.
[37] Hays, W. L. 1988. *Statistics*. New York: CBS College Publishing; 1988.