Chapter 5

Summarization and visualization

D. Mladenič & M. Grobelnik
Department of Knowledge Technologies,
Jozef Stefan Injstitute, Ljubljana, Slovenia.

Abstract

Both text summarization and visualization aim at providing some sort of general view of the text either giving a text summary in the required natural language or giving some visual representation of the text. In both cases the text can be either a single document or a set of documents written in some natural language(s). Here we present some basic methods for text summarization and text visualization and give an example on real-world data describing the research projects in information technology supported by European Commission.

1 Introduction

Most people are aware of the large amount of information available on-line and the potential value it contains. At the same time, in our everyday life we are faced with the impression of condensed time, where there are simply not enough hours in the day for all the desired activities. We try to cope with this in different ways. One of the important ways of using the available on-line information in a limited time is pre-processing it by means of summarization and visualization. Humans are good at making summaries such as abstracts of papers, reviews of books, news headlines, minutes of meetings, previews of movies. Different types of materials can be subject to summarization and visualization. It is worth mentioning that in addition to text, multimedia materials are becoming more popular and recognized as a valuable source of on-line information, resulting in a growing interest in the development and use of methods for handling multimedia materials including multimedia summarization. In this chapter, we concentrate on text materials as one of the most popular and investigated types of materials for summarization.

WIT Transactions on State of the Art in Science and Engineering, Vol 17, © 2005 WIT Press
www.witpress.com, ISSN 1755-8336 (on-line)
doi:10.2495/978-1-85312-995-7/05

There are different ways to distinguish between types of summaries. While most methods aim at generating generic summaries, recently user-focused summaries are gaining importance as a result of the growing importance of personalized information filtering. The other ways to characterize types of summaries according to their functions [13] are: indicative, informative and critical. Indicative summaries provide only a basic idea to alert the user to relevant sources. Informative summaries are used as substitutes for the sources mainly giving all the relevant information, while critical summaries in addition to relevant information substituting the source also incorporate opinion on the content. Based on the main characteristics of the underlying methods, we can talk about keyword summarization, sentence extraction and abstract generation, as described in Section 2.

The rest of this Chapter first describes the main methods used in text summarization (Section 2) and text visualization (Section 3). Section 4 gives an illustration of the methods on an application involving real-world data describing the information technology research projects supported by European Commission. We conclude with remarks on potentials for future directions of research and applications (Section 5).

2 Text summarization

Text summarization can be approached and observed from different perspectives. Similar as in document retrieval used by document search engines, in text summarization we are dealing with text documents. Moreover, summarization is often applied as a second stage of document retrieval, to help the user to get an idea about the content of the retrieved documents. Research in Information Retrieval (IR) 0 has a long tradition of addressing the problem of text summarization with the first reported attempts in the 1950s and 1960s, that were exploiting properties such as frequency of words in the text. When dealing with text, especially in different natural languages, properties of the language itself can be a valuable source of information. This brings in text summarization of the late 1970s, the methods from research in Human Language Technologies, especially Natural Language Processing (NLP) [24]. As humans are good at making summaries, we can consider using examples of human-generated summaries to find something about the underlying process by applying Machine Learning methods (ML) [27]. Dealing with data analysis in general is traditionally addressed by Statistics [6, 15] thus statistical methods are also importantly contributing to text summarization. Analysis of large datasets is in general a subject area of Data Mining (DM) [8] and when dealing with text data we often talk about Text Mining (TM) 0. All these and probably some other research areas contribute their methods and ideas to text summarization.

In general, text summarization can be applied on a single document written in some natural language or on a set of documents written in one or several languages. For instance, capturing on-line news from several countries and providing a summary of them involves multi-document and multi language summarization.

There are several ways to provide text summary. The simplest but also very effective way is providing **keywords** that help in capturing the main topic(s) of the text either for human understanding or for further processing such as indexing and grouping of documents, books, pictures, etc. As the text is usually composed of sentences we can talk about summarization by **highlighting or extracting the most important sentences**, and this is the way of summarization that is frequently found in human-generated summaries. A more sophisticated way of summarization is by **generating new sentences** based on the whole text, as for instance used by a human in writing book reviews. The following subsections describe basic methodology and give some examples of these three main approaches to text summarization.

2.1 Keywords

Using keywords to describe the topic(s) of the text is usually based on extracting some of the words from the given text 0. Availability of texts with manually assigned keywords (such as research papers with the keyword list) and document taxonomies equipped with keywords (such as Medline taxonomy of medical abstracts, or Yahoo taxonomy of Web documents) enables application of text/document categorization techniques [18, 32] to automatically generate a model for mapping text to a set of keywords.

2.1.1 Extracting keywords from text

Keywords used to characterize texts are often phrases of two or more words, thus some authors referred to them as keyphrases 0. Keyphrases can serve multiple goals, depending on a particular setting where they appear. For instance, keywords on the first page of an article provide summarization enabling the reader to quickly determine whether the given article is interesting for the reader. Cumulative index for a book or a journal gives an index rather than a summary and enables the reader to quickly find a part of texts related to a specific interest of the reader. Search engines also use keywords filed to help the user in making more precise search. Depending on the application goal different methods are used for generating a keyword list.

Automatic extraction of keyphrases from text, as proposed by Turney 0 defines the task as the automatic selection of important phrases from within the body of a document. The document is treated as a set of phrases, where each phrase is either keyphrase (positive example) or not (negative example). This leads to a supervised machine learning setting [27] in a classical way using labeled examples to generate a classification model. Once generated, the model is used to classify new examples, namely to classify each phrase from a document as keyphrase or non-keyphrase. The problem is difficult as only a small proportion of examples is positive (usually less than 1%), while typically machine learning is applied on more balanced class distribution. Some newer machine learning approaches investigate the issue, as it seems to be fairly common especially for the tasks that involve text data [3, 28].

In order to apply machine learning methods on document phrases, each phrase is described by a set of features (properties). This provides for so-called declarative domain knowledge and presents an important step in applying machine learning to different real-world problems. In 0, the features used to describe a phrase are for instance, the number of words in a phrase, the position in the text where the phrase first occurs, the frequency of the phrase etc. They have performed experimental evaluation of two machine learning approaches on five different collections of documents. The first set of experiments applies a decision tree induction algorithm to this task while the second set applies a specifically designed keyphrase extraction procedure where the parameters are tuned using genetic algorithms [27]. In decision tree induction, each phrase is evaluated based on the values of a subset of features and classified as positive or negative example of keyphrases. For a new document, all the phrases from the document are classified and the algorithm returns a list of keyphrases. In the specifically designed extraction procedure, each phrase is pre-processed (*e.g.* stemmed, converted to lowercase letters) and assigned a score depending on the feature values (*e.g.* more frequent phrases get higher score) and adopting a linear weighting model (the overall phrase score is a sum of weights of its features). Relative influence of each feature is determined by the feature weight that is set using machine learning with a genetic algorithm. The returns list of keyphrases contains a subset of highly scored phrases.

As expected, the custom-designed algorithm that incorporates more of so called procedural domain knowledge is more suitable for the task than a general machine learning algorithm. Statistical evaluation of the keyphrases generated by the proposed custom-designed algorithm shows that up to 30% of the returned keyphrases are among the desired keyphrases. Subjective human evaluation of the same results suggests that about 80% of the automatically extracted keyphrases are acceptable to human experts.

2.1.2 Keyword assignment using document categorization

Automatically assigning keywords to document based on a keywords-equipped document taxonomy was proposed in [29]. There, Yahoo taxonomy of Web documents is used to automatically obtain a model for categorizing documents by means of machine learning methods (as proposed for taxonomy of Web documents [25, 28] or Reuters news articles [20]). The assumption of a machine learning method is that we are given a set of documents each labeled by one or more categories from a predefined set of categories, such as “arts”, “business”, “entertainment”, “education”, etc. Applying the model on a new document assigns to the document one or more content categories. In the case of keywords-equipped document taxonomy such as Yahoo, in addition to a category each document is assigned to the associated set of keywords. Namely, in Yahoo taxonomy, each category is named by a sequence of keywords giving a path from the root of the taxonomy to the category node. For instance, a group of documents related to “Intelligent software agents” is in the taxonomy node named “Science - Computer Science - Artificial Intelligence - Machine Learning

- Intelligent Software Agents". The node name is actually formed by a sequence of keywords (actually keyphrases having mostly one or two words), ordered from more general to more specific. Document categorization in general assigns a category name to a new document and since here the category name is composed of keyword, it can be interpreted also as assigning a set of keywords to a document.

Mladenic and Grobelnik [30] propose document categorization into hierarchy of content categories based on the representation of each document as a word-vector of frequency for each word or word sequence from the document. In order to form the word-vectors, pre-processing is applied on the document removing stop-words (from the standard list of stop-words such as “a”, “the”, “from”), converting letters into lowercase, etc. The particular word sequences (phrase) to be included in the word-vector are generated using an efficient procedure as described in [30] based on the phrase frequency in the whole document collection. Difficulty of this document categorization problem comes from several factors. Similarly as in keyphrase extraction (described in Section 2.1.1), only a small proportion of examples are positive for a particular category. Namely, the whole problem of hierarchical document categorization is addressed as a set of binary problems, each considering if the document should be assigned to a particular category or not. As we are dealing with several thousands of categories, for each of the binary problems positive documents are only documents from the particular category while the negative documents are documents from all the other thousand(s) of categories. The other difficulty comes from a high number of features, representing different words and phrases. In [28] a new method for selecting a subset of features is proposed and experimentally evaluated on five different datasets. The results show that a combination of the proposed method with the Naïve Bayesian classifier on most tested datasets yields significantly better results than using the other methods in the same setting. Experiments using statistical testing with cross-validation show that about 80% of the desired keyphrases were successfully predicted (recall 0.8), while among all the predicted keyphrases about 40% were the same as provided by human (precision 0.4).

2.2 Sentence extraction

In early information retrieval efforts of summarization, sentence extraction was based purely on word frequency. This was followed by work on sentence extraction in late 1960s using not only word frequency [23] but also sentence position and some cue phrases [7] and this is still used by many approaches as a base for sentence extraction. Some newer approaches adopt machine learning methods for extracting sentences [21] and sentence segments [5] or for multi-document sentence extraction [10] and ordering [1]. Natural language processing is a base for approaches that exploit text structure, for instance, using a part of speech tagger and shallow parsing to grouping terms into lexical chains based on their relationships such as synonymy [2]. Lexical chains are scored based on their length, distribution in the text, density, graph topology etc. Extracted are full sentences containing the highly scored chains.

Sentence extraction using machine learning approaches defines the problem of selecting the sentences to be included in the summary as a statistical classification problem as shown in fig. 1. The assumption is availability of a document set with hand-selected document sentences. This set of documents is used to automatically obtain a model for selecting important sentences from a document. Each document is treated as a set of sentences and each sentence is represented by a set of features. The features are used to describe the importance of the sentence and contain information on the sentence position in the document (*e.g.* beginning of the document, beginning of a section), sentence length, presence of thematic words (*e.g.* most frequent words), presence of cue phrases (fixed-phrases such as “in conclusion” or “to summarize”). Each sentence is additionally labeled as important (positive example), if it was among the hand-selected summary sentences, or not important (negative example).

Based on a given set of documents with hand-selected summary sentences, a model is automatically obtained using machine learning methods. The approach described in [21] uses the Naïve Bayesian classifier to obtain a model that

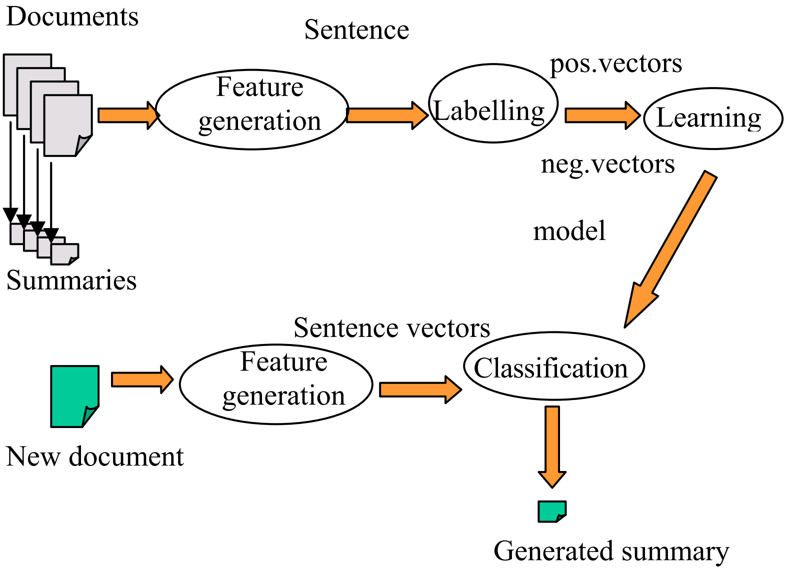

Figure 1: Process of sentence extraction using machine learning approach. It is given a set of documents with associated hand-selected summaries. Each document is treated as a set of sentences, each represented by features and a label. Machine learning is used to automatically generate the associated model that enables generating summaries for new documents.

automatically assigns importance weight to each feature. For a new document, all the sentences are classified by the model returning a probability that the sentence will be included in a summary (if the summary would be human-selected). The document summary is then generated by ranking all the document sentences according to this probability and selecting the highly ranked sentences. Usually, the user specifies the number of sentences to be selected. Experimental evaluation on 188 document-summary pairs sampled from 21 publications, reported in [21], has shown that the most important features are describing sentence position, presence of title keywords and presence of cue phrases. For summaries that are 25% of the size of the average testing document, the proposed method selects 84% of the sentences that are also chosen by human professionals.

Instead of extracting whole sentences, we can generate a document summary by extracting sentence segments. First, document is represented as a set of sentences and each sentence is broken into segments by a special cue markers [21]. Each segment is represented by a set of features such as position, presence of thematic words, etc. Machine learning methods are then used on the sentence segments instead of the whole sentence. In [5] two methods were experimentally compared: decision tree induction and the Naïve Bayesian classifier. The reported results show that decision tree outperforms Naïve Bayesian classifier finding the desired segments in 72% of cases (compared to 69% achieved by the Naïve Bayesian classifier).

2.3 Abstract generation

Generating new text instead of extracting parts of the existing text is a more demanding approach to summarization. One of the early approaches that was applied only to some simple examples is using semantic-based techniques of Artificial Intelligence transforming text into so-called Plot Units [22] and in some more recent work transforming text into frames or templates that are automatically filled by means of information extraction techniques [14, 26]. As abstract generation requires additional domain knowledge sources which are hard to provide and thus it is still dominated by the extraction approaches.

Natural language processing [24] is an important background technology for abstract generation. In [17] an approach is proposed that combines natural language processing with symbolic world knowledge embodied in the concept thesaurus WordNet, dictionaries and similar resources. Such a combined approach enables concept-level generalization (*e.g.* replacing a word by a set of its synonyms). Given a document, first the central topics of the document are found by simply extracting sentences based on phrase frequency, cue phrases and similar (see Section 2.2 for more details on sentence extraction). The next phase aims at removing redundancies, rephrasing sentences and merging related topics based on concept generalization (*e.g.* "orange, apples, figs" can be replaced by "fruits") and script identification ("she sat down, read the menu, ordered, ate and left" can be replaced by "she visited restaurant") [17]. In the last phase of abstract generation, the extracted text is reformulated and merged into a coherent, densely phrased, new text.

3 Text visualization

Visualization of data in general [9] and also of the textual contents of a document set is one of the important ways of how to deal with large amounts of textual data. The most frequent application of text visualization techniques is in particular cases when one needs to understand or to explain the structure and nature of large quantity of typically unlabeled and poorly structured textual data in the form of documents. The usual approach when dealing with text for visualization is first to transform the text data into some form of high dimensional data and in the second step to carry out some kind of dimensionality reduction down to two or three dimensions that allows to graphically visualize the data. There are several (but not too many) approaches and techniques offering different insights into the text data like: showing similarity structure of documents in the corpora (*e.g.* WebSOM [19], ThemeScape), showing time line or topic development through time in the corpora (*e.g.* ThemeRiver [16]), or showing frequent words and phrases and the relationships between them (Pajek 0).

One of the most important issues when dealing with visualization techniques is scalability of the approach to enable processing of very large amounts of the data.

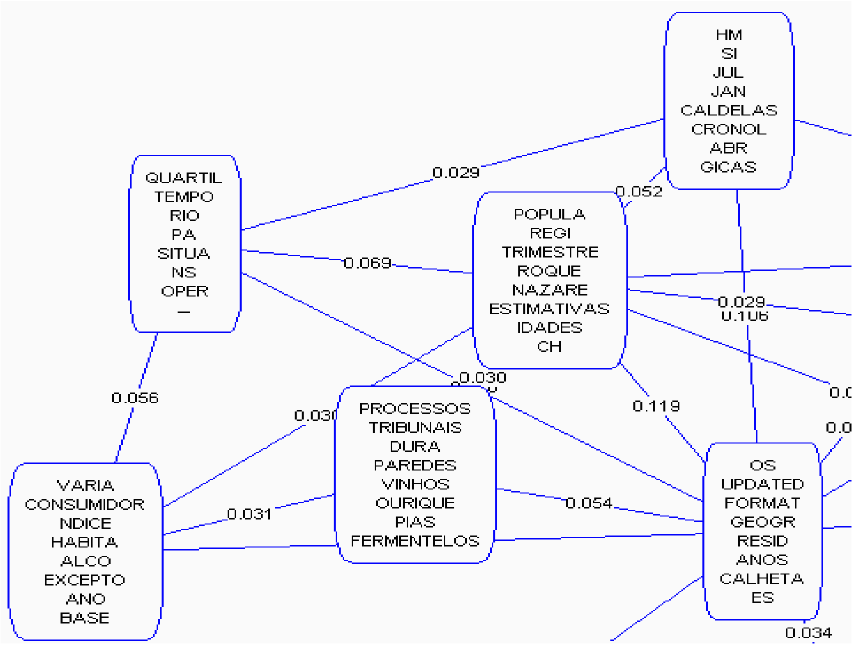

Figure 2: Visualization of a part of the Web site of Portuguese National Statistics Office. Nodes represent the clusters and are labelled by the characteristic keywords, while edges give similarity between them (the higher the more similar).

For instance, by applying document visualization we can get an overview of the content of documents at a certain Web-site or in some other document collection. One possible direction is to base text visualization on clustering of the documents, where the idea is [12] to represent the documents as word-vectors (using TFIDF weighting scheme) and to perform K-Means clustering [33b] on the set of word-vectors. The obtained clusters are then represented as nodes in a graph, where each node in the graph is described by a set of the most characteristic words in the cluster (top weighted words in the centroids of the cluster) and similar nodes (using standard cosine-similarity) are connected with a link. Finally, by applying a procedure for nice graph drawing we get visual representation of a document set. Figure 2 gives an illustration of the approach on a visualization of the Web site of the Portuguese National Statistics Office.

4 Example of summarization of a document set

Many applications today require dealing with a set of documents, where text is provided in some natural language. One such application is analysis of the data describing research projects in information technology that were supported by European Commission [Grobelnik and Mladenic 2002] in the 5th Framework Program (IST 5FP). That application required several approaches involving Text Mining and Link Analysis. Here we will describe just an example related to a specific kind of problem focusing on a specific kind of text summarization.

For the analysis of 2786 research projects, we had two sources of the data: (1) table of projects from internal database of European Commission and (2) project descriptions from the Web. The relevant information was automatically extracted out of the Web `HTML` files: project name, acronym, start and end dates, textual description of the project content, a list of participating organizations. The data were pre-processed mainly to remove inconsistencies.

One of the questions imposed by the customer required summarization of the document collection from a particular angle, namely finding a set of organizations relevant for a given topic description. To achieve that, each document was first divided into three parts: project acronym, content description of the project and a list of organizations participating in the projects. For the purpose of this application, the other available data were ignored. Next, a machine learning method is used on the documents so that the system returns a list of organizations sorted by their relevance for the given topic, based on the topic description (a set of keywords) provided on the fly by the user. Given a topic description, for each document its relevance score is calculated based on the project content and the set of 100 most relevant projects is selected. For each of the selected projects, the relevance score is used to assign weight to all the organizations participating in the project. If the same organization appears in several projects, the weights are added, resulting in higher weight of the organization. Finally, all the organizations are sorted according to their weight and the top highly weighted organizations are presented to the user. As an example in fig. 3 we give the highly weighted organizations relevant for Data Mining topic.

Organization name - [list of acronyms of relevant projects]

- GMD Forschungszentrum Informationstechnik - [SOL-EU-NET, SPIN!, XML-KM, CYCLADES, VESPER, COGITO]
- Universitaet Dortmund - [MINING MART, KDNET, DREAM, CYCLADES, APPOL II]
- Dialogis Software Services - [MINING MART, SOL-EU-NET, SPIN!]
- European Commission Joint Research Centre - [MINEO, KDNET, LINK3D, ETB, CTOSE, NOSE II]
- Universita Degli Studi Di Bari - [SPIN!, KDNET, LINK3D, COGITO]
- Fraunhofer Gesellschaft Zur Foerderung Der Angewandten Forschung - [KDNET, CERENA, VOSTER]
- Universita Del Piemonte Orientale Amedeo Avogadro - [MINING MART, KDNET]
- Schweizerische Lebensversicherungs Und Rentenanstalt Swiss Life - [MINING MART, KDNET]
- Perot Systems Nederland - [MINING MART, KDNET]
- Bureau De Recherches Geologiques Minieres - [MINEO]
- Katholieke Universiteit Leuven - [SOL-EU-NET, KDNET, VIBES, UP-ARIADNE]
- Institut National Des Sciences Appliquees De Lyon - [CINQ]
- University Bristol - [SOL-EU-NET, KDNET]
- Institut Jozef Stefan - [SOL-EU-NET, KDNET]
- Czech Technical University Prague - [SOL-EU-NET, KDNET]

Figure 3: The set of highly weighted organizations relevant for Data Mining topic summarizing the related content of the document set describing 5FP EU IST research projects.

The organizations in fig. 3 were obtained by providing to the developed machine learning system the following set of keywords: "*knowledge discovery text mining classification machine learning data mining data analysis personalization decision support*".

5 Future directions of research and applications

In the future we may expect that summarization will move towards more semantic representations (in comparison to the today's shallow representations) of the text. This will enable more complex manipulation with the content and therefore better summaries or extracts of the content tailored for particular needs. Visualization technology will expectedly move towards integration with other methods from the area of "intelligent user interfaces". One further important issue is development of more thorough evaluation of the visualization techniques with appropriate Human Computer Interaction methods to justify why and when the approach is better compared to some other ones, in what sense and in which application contexts.

References

[1] Barzilay, R., Elhadad, N., McKeown, & K.R., Inferring Strategies for Sentence Ordering in Multidocument News Summarization, *Journal of Artificial Intelligence Research*, **17**, pp. 35–55, 2002.

[2] Barzilay, R. & Elhadad, N., Using Lexical Chains for Text Summarization, In: *Advances In Automatic Text Summarization*, I. Mani & M.T. Maybury (editors), pp. 111–122, 1999.

[3] Brank, J., Grobelnik, M., Milić-Frayling & N., Mladenić, D. Training text classifiers with SVM on few positive examples, Technical Report, 2003.

[4] CORDIS: Public Web site of 5FP EU projects by European Commission <http://www.cordis.lu/>, 2002.

[5] Chuang, W.T. & Yang, J., Text Summarization by Sentence Segment Extraction Using Machine Learning Algorithms, *Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*, pp. 454–457, Lecture Notes In Computer Science, vol:1805, Springer-Verlag.

[6] Duda, R.O., Hart, P.E. & Stork, D.G. Pattern Classification, Wiley-Interscience, 2nd edition, 2000.

[7] Edmundson, H.P., New Methods in Automatic Extracting, In: *Advances In Automatic Text Summarization*, I. Mani & M.T. Maybury (eds.), pp. 23–42, 1999.

[8] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. & Uthurusamy, R. (eds.), Advances in Knowledge Discovery and Data Mining, MIT Press, February 1996.

[9] Fayyad, U.M., Grinstein, G.G. & Wierse, A. (eds.), Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann Publishers, September 2001.

[10] Goldstein, J., Mittal V.O., Carbonell, J.G. & Callan, J.P., Creating and evaluating multi-document sentence extract summaries. In: *Proc. of the International Conference on Information and Knowledge Management* CIKM, pp. 165–172, 2000.

[11] Grobelnik, M. & Mladenić, D., Approaching Analysis of EU IST Projects Database, *Proc. of the International Conference on Information and Intelligent Systems*, Varazdin, Croatia, 2002.

[12] Grobelnik, M. & Mladenić, D., Efficient Visualization of Large Text Corpora, *The 7th TELRI Seminar "Information in Corpora"*, 2002.

[13] Hahn, U. & Mani, I., The challenges of automatic summarization, IEEE Computer: Innovative Technology for Computer Professionals, November 2000.

[14] Hahn, U. & Reimer, U., Knowledge-based Text Summarization: Salience and Generalization Operators for Knowledge Based Abstraction, In: *Advances In Automatic Text Summarization*, I. Mani & M.T. Maybury (editors), pp.215–232, 1999.

[15] Hastie, T., Tibshirani, R. & Friedman, J.H., The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer Series in Statistics), Springer Verlag; October 2001.

[16] Havre, S., Hetzler, B. & Nowell, L., ThemeRiver™: In Search of Trends, Patterns, and Relationships. IEEE Symposium on Information Visualization, InfoVis-99, San Francisco, CA, 1999.

[17] Hovy, E. & Chin-Yew, L., Automated Text Summarization in SUMMARIST, In: *Advances In Automatic Text Summarization*, I. Mani & M.T. Maybury (editors), pp. 81–91, 1999.

[18] Jackson, P. & Moulinier, I., Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization (Natural Language Processing, 5), John Benjamins Publishing Co., July 2002.

[19] Kaski, S., Honkela, T., Lagus, K. & Kohonen, T., "WEBSOM Self Organizing maps of Documents Collections", *Neurocomputing*, Vol. **21**, pp. 101–117, 1998.

[20] Koller, D. & Sahami, M., Hierarchically classifying documents using very few words. *Proceedings of the 14th International Conference on Machine Learning* ICML97, pp. 170–178, 1997.

[21] Kupiec, J., Pedersen, J.O. & Chen, F., A trainable document summarizer. In SIGIR'95, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Seattle, Washington, USA. pp. 68–73, ACM Press, 1995.

[22] Lehnert, W.G., Plot Units: A Narrative Summarization Strategy, In: *Advances In Automatic Text Summarization*, I. Mani & M.T. Maybury (editors), pp. 177–214, 1999.

[23] Luhn, H.P., The Automatic Creation of Literature Abstracts, In: *Advances In Automatic Text Summarization*, I. Mani & M.T. Maybury (editors), pp. 15–22, 1999.

[24] Manning, C.D. & Schutze, H., *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, MA, 2001.

[25] McCallum, A., Rosenfeld, R., Mitchell, T. & Ng, A., Improving Text Classification by Shrinkage in a Hierarchy of Classes. *Proceedings of the 15th International Conference on Machine Learning*, 1998.

[26] McKeown, K., Robin, J. & Kukich, K., Generating Concise Natural Language Summaries, In: *Advances In Automatic Text Summarization*, I. Mani & M.T. Maybury (eds.), pp. 223–264, 1999.

[27] Mitchell, T.M., Machine Learning. The McGraw-Hill Companies, Inc., 1997.

[28] Mladenic, D. & Grobelnik, M., Feature selection for unbalanced class distribution and Naive Bayes, *Proceedings of the 16th International Conference on Machine Learning* ICML-99. Morgan Kaufmann Publishers, San Francisco, CA, pp. 258–267, 1999.

[29] Mladenic, D. & Grobelnik, M., Assigning keywords to documents using machine learning, *Proceedings of the 10th International Conference on Information and Intelligent Systems IIS-99*, Croatia, 1999.

[30] Mladenic, D. & Grobelnik, M., Feature selection on hierarchy of web documents. *Decision Support Systems* **35**, Elsevier Science B.V., pp. 45–87, 2003.

[31] Mladenić D., EU project: Data mining and decision support for business competitiveness: a European virtual enterprise (Sol-Eu-Net). *Proc. of OES-SEO 2001: Open enterprise solutions: Systems, experiences and organizations*. Rome, 14–15 September 2001. Rome: LUISS, pp. 172–173, 2000.

[32] Mladenić D., Text-learning and related intelligent agents: a survey. *IEEE Intelligent. Systems and their Applications*, **14**, pp. 44–54, 1999.

[33a] Rijsberg, C.J., van Information Retrieval. Butterworths, London, 1979.

[33b] Steinbach, M., Karypis, G., & Kumar, V. A comparison of document clustering techniques. *In Proceedings of KDD Workshop on Text Mining*, pp.109–110, 2000.

[34] Turney, P.D., Learning Algorithms for Keyphrase Extraction, *Information Retrieval Journal*, Vol. **2**, No. 4, Kluwer Academic Publishers, pp. 303–336, 2000.

[35] White, D.R., Batagelj, V. & Mrvar, A., Anthropology - analyzing large kinship and marriage networks with pgraph and Pajek. *Social science computer review*, **17**, pp. 145–274, 1999.

[36] Zanasi, A. (editor), Text Mining and Its Applications, WIT Press/ Computational Mechanics, 2005.