Decision making with textual and spatial information

HANA KOPACKOVA, JITKA KOMARKOVA, PAVEL SEDLAK
Faculty of Economics and Administration, Institute of System Engineering and Informatics
University of Pardubice
Studentska 95, Pardubice, 53210
CZECH REPUBLIC
hana.kopackova@upce.cz, jitka.komarkova@upce.cz, pavel.sedlak@upce.cz

Abstract: - The aim of the paper is to show the way how textual and spatial information can be used in decision making process. Structured information represents only 10 % of available information. Despite of this fact managers mostly rely on this type of information. Due to their marginalization in decision making practice, we focused on textual and spatial information in this article. Four case studies clearly illustrate usage of these types of information on practical examples.

Key-Words: - Decision Making, Information, Text Categorization, Clustering, GIS.

1 Introduction

According to Porter [24] organisations need information to support decision making in various levels of management in order to remain or become truly globally competitive.

Peter Drucker [7] wrote in his article that what we really need is to be information-literate. „Computer people still are concerned with greater speed and bigger memories. But the challenges increasingly will be not technical, but to convert data into usable information that is actually being used".

Process of decision making represents complex and difficult task which is, among others, highly dependent on the accessibility of high quality information. In spite of classical presumption of bounded rationality theory [26], which supposes scarcity of information, today situation brings different problem. The amount of information is too huge.

As far as the amount of produced information is growing faster than the ability of information consumers to find, retrieve and use information, decision makers must somehow face this situation.

It was found [6] that all decision makers repeatedly revealed the preference to collect more information than needed. This finding indicates the idea of voluntary information overload of managers. Also work of Feldman and March [9] describes this situation. Their main claim is that organizations systematically collect more information than they use. A lot of the information gathered has a little decision relevance and plenty of information is collected after the decision has been made.

So, there is a conflict in managers' needs. Managers do not want to make decision without all possible information gathered so they postpone the decision and wait for additional information. On the other side, number of collected information can be so vast that its processing is very demanding and sometimes even impossible.

A possible solution is based on the pre-processing of data and information, and transformation of them into an actionable knowledge. The knowledge or intelligence producer (no mater if person or software) is supposed to make analyses and prepare clear outputs. Outputs must be in the best perceivable form for management. Thus, knowing how management accepts intelligence reports is crucial.

People prefer to take in information visually and verbally, using multimedia support, and adequately in brief. Various visualization methods are used to ease the interpretation of results and create a kind of interface between mathematical results and users.

In addition to traditional histograms and other charts, evolution of visualisation techniques provides also other methods to represent knowledge. In this article we will focus on three different types of methods. Their utilization is demonstrated on four case studies. Two methods which are from the branch of soft computing cover text categorization and text clustering tasks. In the third and second case we explain how to transform spatial data into actionable knowledge.

2 Information and Communication Technologies in the Process of Decision Making

The influence of information and communication technologies (ICT) can be viewed in two different perspectives. The first one takes into consideration development of ICT as the primary reason for information overload. The second point of view focuses on functionalities of modern ICT that have the potential to improve information retrieval and processing, e.g. by its structuring and visualisation in graphical form, and thus improving information perception.

From the common point of view most computer systems support decision making because all software programs involve automating decision steps that people would take.

The definition of DSS, which has evolved since the 1970´s was described in *Building Effective Decision Support Systems* by Sprague and Carlson [27] define it as a computer-based system, which contains data and analysis models and allows direct interaction. This definition can be taken from the narrow or broad point of view. The narrow view shows the DSS as a system that essentially solves or gives options for solving a given problem. The decision process is structured in a hierarchical manner, user inputs various parameters, and DSS essentially evaluates the relative impact of doing x instead of y. The broader definition incorporates the above narrow definition but also includes other technologies that support decision making such as knowledge or information discovery systems, database systems and geographic information systems (GIS) [1], [13].

Power [25] suggested the following broad categories:

- Data driven DSS - includes file drawer and management reporting systems, data warehousing and analysis systems, Executive Information Systems (EIS) and GIS. Data-Driven DSS emphasizes access to and manipulation with large databases of structured data and especially a time-series of internal company data and possibly external data.
- Model driven DSS - includes systems that use accounting and financial models, representational models, and optimization models. Model-Driven DSS emphasizes access to and manipulation with a model. Simple statistical and analytical tools provide the most elementary level of functionality. Some OLAP systems that allow complex analysis of data may be classified as hybrid DSS systems providing modelling, data retrieval and data summarization functionality. Model-Driven DSS use data and parameters provided by decision makers to aid them in analyzing a situation but they are not usually data intensive.
- Knowledge driven DSS - can suggest or recommend actions to managers. This DSS is a person-computer system with specialized problem-solving expertise. The "expertise" consists of knowledge about a particular domain, understanding problems within that domain, and "skills" at solving some of these problems. A related concept is Data Mining which is used to analyse large amounts of data in databases in order to find hidden patterns. Data mining tools can be used to create hybrid Data-Driven and Knowledge-Driven DSS.
- Document driven DSS - integrates a variety of storage and processing technologies to provide complete document retrieval and analysis. The Web provides access to large document databases including databases of hypertext documents, images, sounds and video. A search engine is a powerful decision-aiding tool associated with this type of DSS.
- Communication driven and group DSS - communication driven DSS includes communication, collaboration and coordination and group DSS focuses on supporting groups of decision makers to analyse problem situations and perform group decision making tasks.

In this article the broader concept of DSS is used as a framework. Only textual and geographic information are concerned within the article due to their marginalization in decision making practice.

3 Textual Information in Decision Making

Significance of the problem of textual information is considered by Tucker [28]: ”The ratio of unstructured to structured information in most organizations is easily 9 to 1, yet many of us spent most of our time worrying about – indeed, dedicating our careers to – managing the most familiar 10 percent of the problem: structured information…”. Forest Research [10] has predicted that unstructured data (such as text) will become the predominant data type stored online. Also Gartner group predicted that the amount of textual information double in every three months. Unlike the tabular information typically stored in databases today, documents have only limited internal structure, if any.

So, managerial decision making process can be highly dependent on hidden information in text documents. However, careful reading and sorting of documents is time consuming work. This type of activity wastes working time of managers and in the end it can even cause wrong decision. Text mining can be used for pre-processing of textual information in order to find hidden knowledge and ease the process of decision making.

Some examples of possible usage of text mining are: building personalised Netnews filter which learns about

preferences of a user [18], classification of news stories [15] or guidance of a user's search on the Web [1], [21], [22].
A growing number of statistical classification methods have been applied to text mining, such as Naive Bayesian [10], and Support Vector Machines [5], [16]. A comprehensive comparative evaluation of a wide-range of text categorization methods is in [18], [13].

4 Geographic Information in Decision Making

Activities of mankind are closely connected to the surface or near the surface of the Earth. Today, 70-80% of the tasks solved by local government are geographically related. In many situations knowledge of the place where something happens can be critically important. Data which contain spatial information are special – they allow to link place, time and attributes. People mostly use data connected to the Earth, so called geographic data/information but spatial data can be connected to any space. GIS are usually used as a software tool to process and analyze geographic data [20].
Importance of solving spatially oriented problems and making spatially influenced decisions have been recognized for several years. Many spatial decision support systems (SDSS) have been proposed and an influence of utilization of GIS as a SDSS on decision-maker performance was studied [3]. So, interest of managers and users in utilization of spatial information and services increases rapidly.
Environment impact assessment [31], environment protection [11, 23] and route planning belong to significant branches of GIS and SDSS utilization [17, 23]. Utilization of GIS in crisis management is a very important issue, e.g. a Multi-Criteria SDSS was proposed to help manage flooding [19] or another GIS Based Multicriteria Decision Making Tool was proposed to help select proper remediation technology [8]. Solid waste management and time and cost minimization can be given as an example of utilization GIS for solving a practical problem [23].
Visualization of geographic data by maps can quickly provide required information. If text or tabular information is used it is necessary to spend long time by its reading. When a map is used only a short look can be sufficient for understanding geographic information [30]. But cartography which deals with visualisation of spatial data has its own rules and principles, e.g. a correct interval of classification and cartographic method has to be selected. Otherwise, information can be perceived in a wrong way.
Unfortunately, classic GIS software packages (desktop or professional GIS) are too sophisticated so they limit users at least because of the complicated user interface and necessity of using the given computer. High price should be mentioned too [17]. This has resulted into rapid spreading of Web-based solutions along with an increasing demand for easy access of end-users to geographic data [29]. Factors which influence success of web-based SDSS have been already studied [14]. Web presentation is quite difficult task because various experience, skills and equipment of users must be taken into account during application design. Heterogeneity of data from various sources can cause serious problems too but shared databases can improve effectiveness of information production and utilization [8, 11].

5 Case Study A

Manager in this situation needs to find all available and accessible information about waste management in the Czech Republic. His company wants to introduce a new product to the market, however manufacturing technology results into waste production too. Desired information then covers legislation, dump locations, possible courses about this topic, case studies and so on. Text categorization is done with the effort just to support managerial decision by providing only relevant documents.
In the testing environment information retrieval was conducted. In advance prepared datasets containing documents about environment protection were used; one contained 25 documents focused directly on the branch of waste management and the second one contained 25 documents covering various environment protection topics.
Next, 36 experiments were conducted; this number is given by combination of six methods of feature selection, two methods of term weighting and three text categorization methods.
The process of text categorization started with parsing - bag of words representation [21] was used in this stage along with stop list usage (600 words). For term selection six different methods were used: term frequency (terms with only one occurrence in the particular document were omitted) – TF; document frequency (terms present only in one document were omitted) – DF; term frequency combined with document frequency (terms present only in one document with only one occurrence in this document were omitted) – TFDF; Chi-square; Mutual information; and Information gain. Methods for term weighting were selected by TF method and TF combined with inverse document frequency method (TFIDF). After pre-processing stage, the database was filled with 10571 words. On the average there is one word present in two documents but the reality was different in this case. The most frequent situation is that particular word is present only in one document (6529 words in our case). The fact that so

many words are infrequent can lead to the hypothesis that document frequency can be used for term selection.
Then, text categorization was done by means of Naive Bayes (NB), K-nearest neighbour (K-NN) and SVM-SMO algorithms. In all cases correctly classified instances (CCI-xxx) and Kappa statistics (K-xxx) were used for measuring accuracy of the text categorization. All results are given in Table 1.

Table 1. Results of text categorization. Percentage share of correctly classified documents.

Method of Term Weighting – Method of Term Selection	Method of Text Categorization					
	Naive Bayes		K-NN		SVM-SMO	
	CCI-NB [%]	K-NB [%]	CCI-KNN [%]	K-KNN [%]	CCI-SVM [%]	K-SVM [%]
TF - chi-square	96,00	92,00	62,00	24,00	92,00	84,00
TF - mutual information	98,00	96,00	62,00	24,00	92,00	84,00
TF - information gain	96,00	92,00	50,00	0,00	92,00	84,00
TF - TF	88,00	76,00	56,00	12,00	78,00	56,00
TF - DF	90,00	80,00	56,00	12,00	82,00	64,00
TF - TFDF	88,00	76,00	56,00	12,00	90,00	80,00
TFIDF - chi-square	96,00	92,00	66,00	32,00	96,00	92,00
TFIDF - mutual information	96,00	92,00	68,00	36,00	90,00	80,00
TFIDF - information gain	92,00	84,00	50,00	0,00	94,00	88,00
TFIDF - TF	90,00	80,00	56,00	12,00	84,00	68,00
TFIDF - DF	92,00	84,00	58,00	16,00	82,00	64,00
TFIDF - TFDF	90,00	80,00	54,00	8,00	92,00	84,00

All algorithms except for K-NN algorithm proved to be very good classifiers, applicable to selection of relevant information. SVM and NB can serve as a very helpful tool even though they are slightly complicated. Usage of DF method for the term selection provided very good results which are comparable with results of the other methods so the above stated hypothesis was confirmed. Differences between TF and TFIDF methods are not so significant to prove that one of them is better.
This case study was focused on testing text categorization methods being usable for selection of documents containing relevant information. The tested methods are not 100% precise but even the simplest one can decrease number of documents which have to be read by manager so they can significantly decrease the time demandingness of decision making process. In this particular case manager has to go through only approx. 30 documents instead of 50 documents available in the dataset. Precise number of documents to read depends on the classifier accuracy. Disadvantage of this approach is that some documents may stay hidden for manager.

6 Case Study B

University teacher is a manager in this case. He wants to run a quick scan of thematic seminar papers to see if some topics do not overlap. Too much similar papers seems to be copied or prepared in cooperation. Especially effective is this analysis for making year-to-year comparison.

The dataset contains 110 documents of very different length. Shortest took only two pages, longest covered 13 pages. Papers were gathered from three different subjects – data mining, system engineering and geographic information systems. These subjects are sufficiently distinct from each other so as the result of clustering we would obtain three clusters with inner structure.
These steps were used in the preprocessing stage:

- exclusion of stopwords,
- exclusion of words covered in more then 90 % of documents,
- exclusion of words covered in less then 2 documents,
- exclusion of words with term frequency less then 2,
- exclusion of words shorter then 3 letters.

The quick scan was conducted using hierarchical cluster analysis that gives visual output. Result of this experiment (see Fig. 1) shows that only two documents were classified into incorrect group – one paper from data mining and one from geographic information systems were classified as system engineering. Division into three groups approved separability of studied papers but expected results were focused on different aim - to find some abnormality.

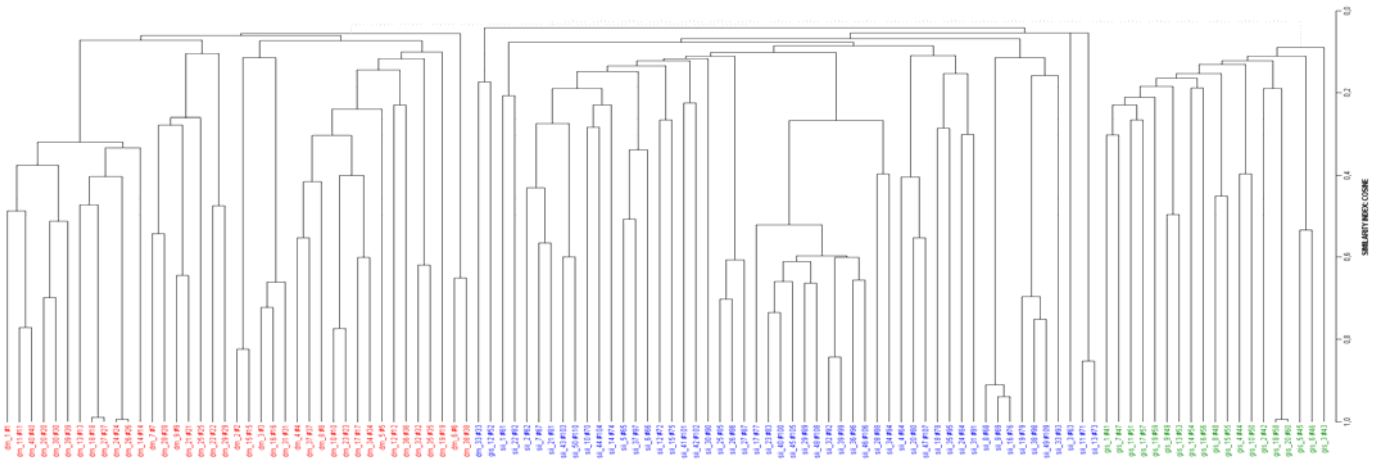

Fig. 1. Dendrogram as a result from hierarchical clustering.

In the Fig.1. we can find three overlaps. One overlap in subject geographic information systems and two in data mining. After careful reading, we found that similarity in data mining papers is given by copying. For geographic information system is the situation different. One paper was saved twice, first as draft and in second case as final paper. Explained example proved that visualization of outputs can ease the decision making process but unfortunately this kind of text pre-processing can not replace human work completely.
In this case the teacher had to read suspicious papers to decide if they abnormality is given by copying or if there is some other reason.

7 Case Study C

A manager wants to find an optimum location for construction of a recreation resort that have to meet many given requirements. Case study solves selection of an optimum location that have to be placed in Olomouc district at maximum 15 kilometres from the city Olomouc and 2 kilometres away from railway-station for good accessibility. Possible localities have to lie in distance greater than 1.5 kilometres away from watercourse because study area is very flat and it was affected by floods. Manager requires forest area only because of relaxation. Result of spatial analyses (distance measuring, buffering and topological overlay) is given on the map on the Fig. 2. It is obvious that man is not able to obtain this information by one look at a topographic map; some analyses have to be done. GIS is the proper tool to support decision making process in this case.

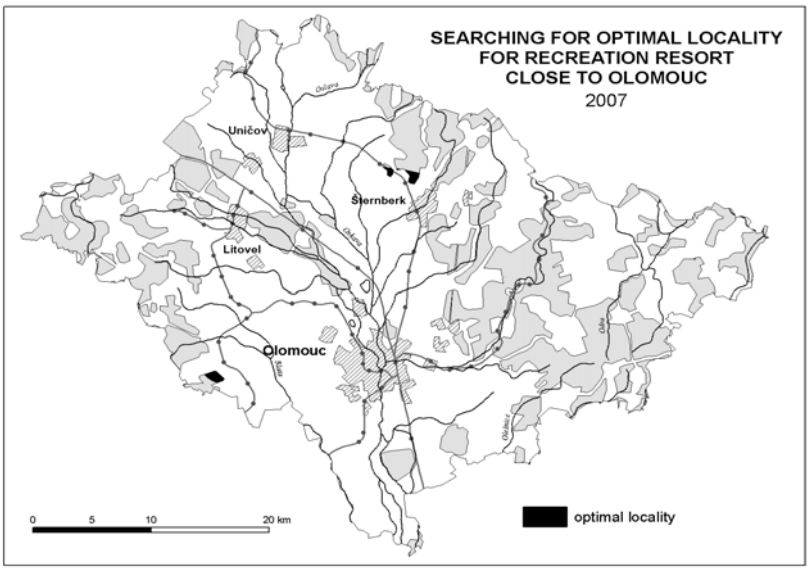

Fig. 2. Solving spatially oriented problem. Result of spatial analyses displayed on the map.

8 Case Study D

In the forth case, regional development agency is in the role of decision maker. This agency is supposed to prepare study about financial situation of municipalities of Pardubice region. Essential aims of the study were defined as follows:

I. Find municipalities having financial problems based on the given data. Financial problems mean that expenditures exceed incomes in two defined years (2004, and 2006) and a municipality is in debt.

II. Find municipalities having the deficit bigger than 10 000 CZK per capita in these years and being in debt. Such municipalities could have great problems and some state intervention may be necessary.

III. Find out if there is any relation between microregions and problematic municipalities. The presumption is that microregions were formed in order to make larger investment projects possible so majority of municipalities involved in microregions will be marked as problematic municipalities from the previously given point of view.

IV. Find out if the problematic municipalities form clusters. This aim is very similar to the preceding one but it covers also municipalities that are not engaged in any microregion.

V. Find out if the problematic municipalities are located only in distant (border) parts of region. To be classified as an outlying municipality, the municipality can not neighbour with former district towns (Pardubice, Chrudim, Svitavy, and Ústí nad Orlicí).

To realize these entire analyses manager can use different methods. Some of them can be done using only MS Excel, but parts three, four and five are especially suitable for processing in geographic information systems.

Given data were in XLS format and covered:

- name of municipality,
- income per capita in the year 2004 and 2006,
- expenditure per capita in the year 2004 and 2006,
- debt per capita in 2004.

For the point I we found 36 municipalities meeting given criteria. Figure 3 shows cartographic output representing these municipalities.

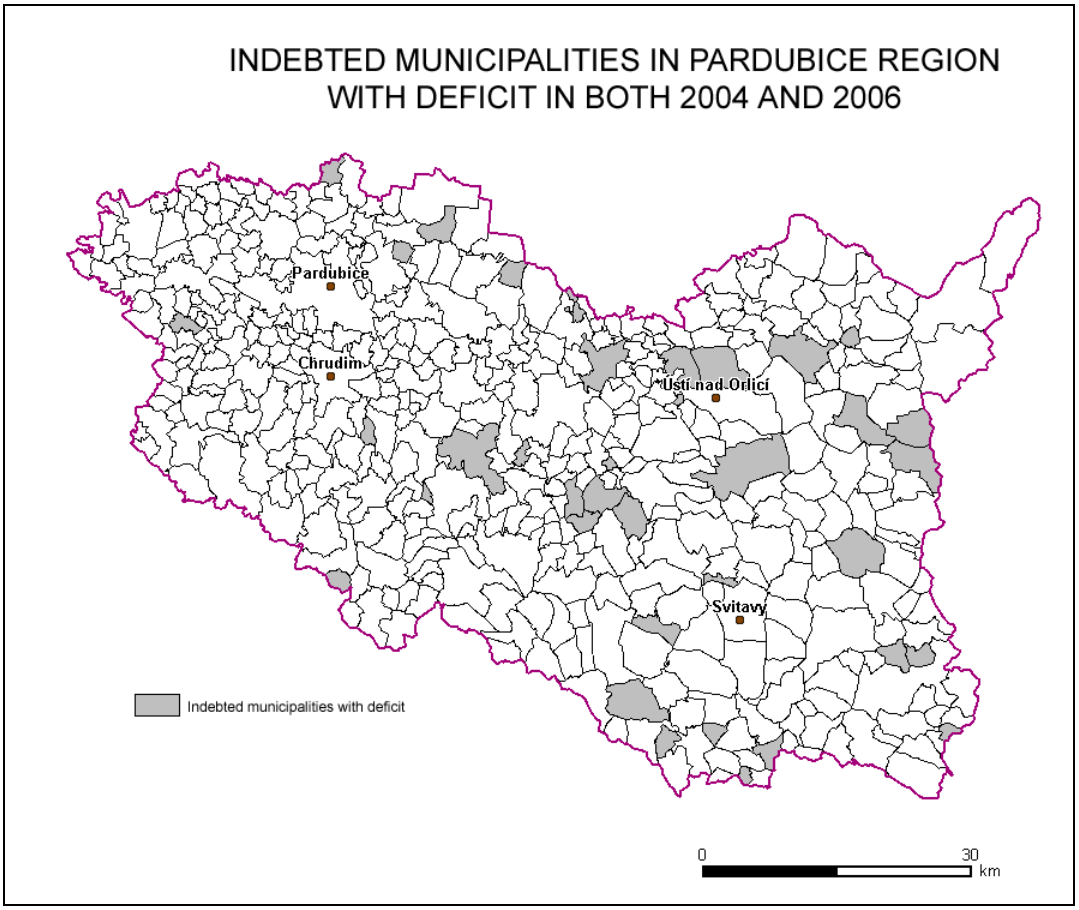

Fig. 3. Problematic municipalities in Pardubice region

As a result of point II, no municipality meeting given criteria was found. It means that no municipality has real problems with solvency.
Possible relation between microregions and problematic municipalities were looked for within the point III. Figure 4 shows microregions in Pardubice region along with the problematic municipalities.
We can see that there is no relation between problematic municipalities and microregions. One half of municipalities (18) is located in some microregion and the second half not. There is no microregion containing only problematic municipalities (see figure 4).
As it can be easily seen from figures 3 and 4, there are not real clusters of problematic regions (point IV). Only eleven of them neighbour with another problematic municipality, the rest of them form isolated isles.
Presumption of point V was mostly confirmed. Only three municipalities neighbour with former district towns. Two of them are near Ústí nad Orlicí (Libchavy, Orlické Podhůří) and one is near Svitavy (Kukle).

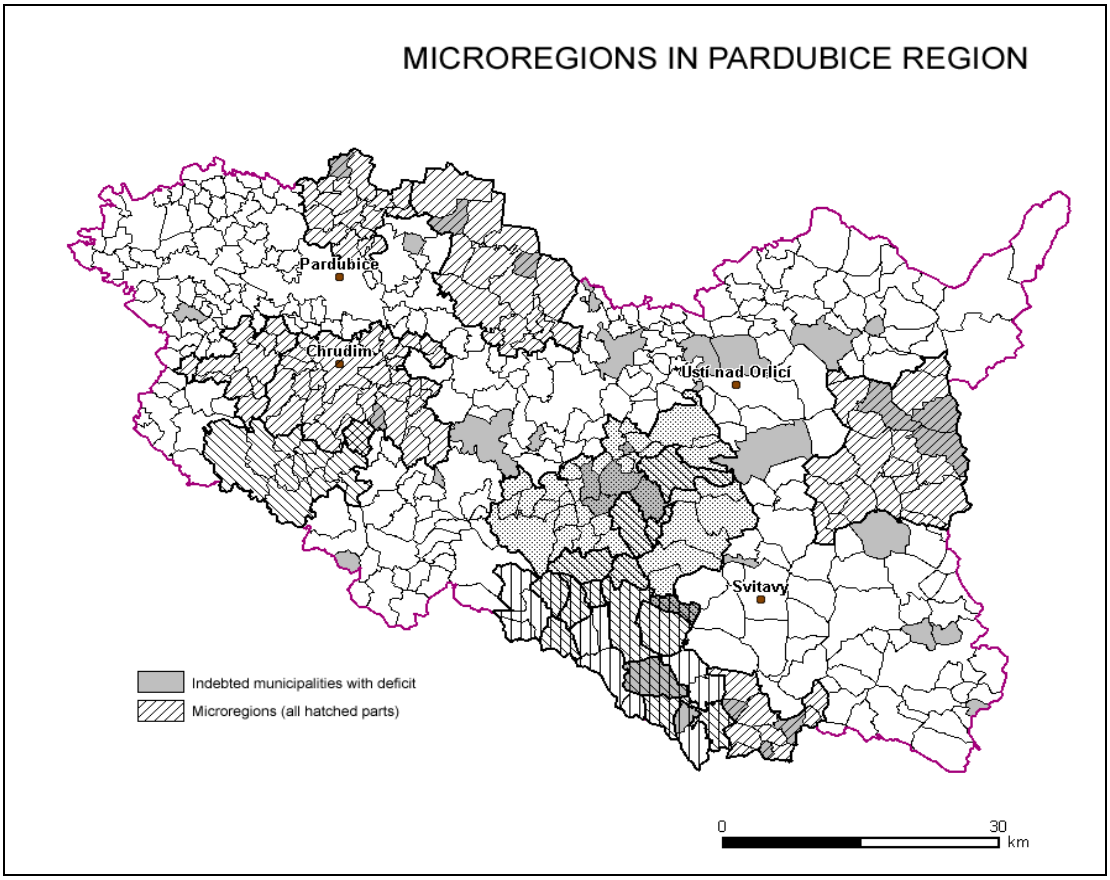

Fig. 4: Microregions in Pardubice region

9 Conclusions and future work

Pre-processing of data and information and transformation of them into actionable knowledge is one possible way how to decrease manager workload and improve the quality of final decision. In this article we concern on textual and spatial information because unlike numerical tabular data, processing of this kind of information is much more complicated but it is becoming much more important for decision making.
Textual documents today contain a lot of buried information which can be useful for decision making process. As Sid Banerjee mentioned in [2], corporations should be able "to collect all of the online product reviews, call-center notes, survey verbatims, customer-relationship management (CRM) text fields, press releases, and other textual information available across and beyond an enterprise and quickly convert them into useful business intelligence". Text mining techniques have been developed to help people to solve this problem.
In the paper, there are two case studies showing advantages of utilization of selected methods for retrieving relevant information from text documents. Specifically, Naive Bayes and SVM-SMO algorithms provided very good results in adequate documents selection. Also hierarchical clustering method can be used as a supporting method for managerial decision making.

Disadvantage of this approach is that some documents may stay hidden for manager so selection of the proper method is a very important issue.

Geographic information is today very important in decision making process because almost everything what happens is connected to some location. But only very basic information can be obtained just by looking at the maps. Usually, sophisticated spatial analyses like spatial querying, topological overlay, network analyses, spatial statistics, and modelling, etc. have to be done in order to study spatial relationships and patterns and obtain useful actionable knowledge. Anyway, utilization of spatial DSS can significantly improve quality and speed of decision making process. Case studies C and D provide a very brief example of possible benefits of utilization of geographic information for supporting decision making.

In this article we wanted to draw manager's attention to the tools processing textual and geographic information. Recently are these tools taken as something special, not generally applicable. Reasons are mostly two: high price of software (in the case of GIS also data) and high skill demandingness [4]. To overcome these barriers, managers need many successful stories from their branch.

Future work will be focused on searching for methods suitable for mining of geographic data from textual documents and then searching for methods of easy processing of retrieved geographic data by means of spatial analyses into actionable knowledge and its easily perceptible presentation.

Acknowledgments. This research and paper was created with a kind support of the Grant Agency of the Czech Republic, grant number GACR 402/05/P155.

References:

[1] Alter, S.: Decision support systems: Current practices and continuing challenges. Addison-Wesley, Reading, MA (1980)

[2] Banerjee, S. 'Text Mining': Shortening the Distance Between You and Your Customers. Marketing Profs. [online] [cit. 03-03-2008] URL: <https://www.marketingprofs.com>

[3] Crossland, M.D., Wynne, B.E., Perkins, W.C.: Spatial decision support systems: An overview of technology and a test of efficacy. Decision Support Systems 14, (1995) 219-235

[4] Drew, R. Text mining tools take on unstructured data. Computerworld, June 21, 2004.

[5] Dumais S., et. al.: Inductive learning algorithms and representations for text categorization. In Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM 98). (1998)

[6] Driver, M.J., Mock, T.J.: Human information processing, decision style theory, and accounting information systems. The Accounting Review 50 (1975) 490-508

[7] Drucker, P. Be Data Literate -- Know What to Know. *Wall Street Journal*, December 3., 1992.

[8] Dudukovic, J., Stanojevic, M., Vranes, S.: GIS Based Decision Support Tool for Remediation Technology Selection. In Proceedings of the 5th IASME/WSEAS Int. Conference on Heat Transfer, Thermal Engineering and Environment, Athens, Greece, August 25-27, 2007, p. 232-237

[9] Feldman M.S., March J.G.: Information in organizations as signal and symbol. Administrative Science Quarterly 26 (1981) 171-186

[10] Forrester Research. Coping with Complex Data. The Forrester Report (1995)

[11] Grabaum, R., et al: Use of GIS and Field Site Network for Assessing Changes in Biodiversity. In Proceedings of the 5th WSEAS International Conference on Environment, Ecosystems and Development, Venice, Italy, November 20-22, 2006, p. 89-93

[12] Hayes, P., et al.: A news story categorization system. In Second Conference on Applied Natural Language Processing (1988)

[13] Holsapple, C.W., Whinston, A.B.: Decision support systems: A knowledge-based approach. West Publishing Company, Minneapolis/St. Paul (1996)

[14] Jarupathirun, S., Zahedi, F.M.: Exploring the influence of perceptual factors in the success of web-based spatial DSS. Decision Support Systems 43 (2007) 933-951

[15] Joachims, T.: A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In Proceedings of ICML-97, 14th International Conference on Machine Learning (1997)

[16] Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In Proceedings of ECML-98, 10th European Conference on Machine Learning (1998)

[17] Keenan, P.B.: Spatial decision support systems for vehicle routing. Decision Support Systems 22 (1998) 65-71

[18] Lang, K.: NewsWeeder: Learning to Filter Netnews. In International Conference on Machine Learning (1995)

[19] Levy, J.K., et al.: Multi-criteria decision support systems for flood hazard mitigation and emergency response in urban watersheds. Journal of the American Water Resources Association 43 (2007) 346-358

[20] Longley, P. A.: Geographic information systems and science. John Wiley & Sons, Chichester (2001)

[21] Mitchell, T., et al.: WebWatcher: A Learning Apprentice for the World Wide Web. In AAAI Sprig Symposium on Information Gathering from Heterogenous, Distributed Environments (1995)
[20] Mladenic, D.: Personal WebWatcher: Implementation and Design. Tech. Report IJS-DP-7472, J. Stefan Inst. (1996)
[22] Mladenic, D.: Text-Learning and Related Intelligent Agents: A Survey. IEEE Intelligent Systems 14 (1999) 44-54
[23] Ozkan, A., et al: Optimization of Solid Waste Collection and Transportation Routes by using GIS. WSEAS TRANSACTIONS on ENVIRONMENT and DEVELOPMENT, Vol. 2 (2006) 1322-1327
[24] Porter, M. E.: On competition. Harvard Business School Publishing, Boston (1998)
[25] Power, D.J.: Decision support systems: Concepts and resources for managers. Quorum Books, Westport, Connecticut (2002)
[26] Simon, H. A. *Models of Bounded Rationality*., Cambridge, M.A.: Harper and Row, 1983.
[27] Sprague, R.H., Carlson, E. *Building Effective Decision Support Systems*. Prentice Hall, Upper Saddle River, NJ, 1982.
[28] Tucker, M.: Dark Matter of Decision Making. Intelligent Enterprise Magazine 2 (1999)
[29] Vatsavai, R.R., Shekhar, S., Burk, T.E., Lime, S.: UMN-MapServer: A High-Performance, Interoperable, and Open Source Web Mapping and Geo-Spatial Analysis System. Lecture Notes in Computer Science, Vol. 4197, Springer-Verlag, Berlin Heidelberg New York (2006)
[30] Vozenilek, V.: Cartography for GIS. Vydavatelství UP, Olomouc (2005)
[31] Zhao, Y., Li, D.: A GIS-Based Support System for Environmental Impact Assessment of Rehabilitation of Coal Mine Dump. In Proceedings of the 3rd IASME/WSEAS International Conference on ENERGY, ENVIRONMENT, ECOSYSTEMS and SUSTAINABLE DEVELOPMENT (EEESD '07) (2007) p. 480-485