Knowledge Pre-Processing in Decision Making

HANA KOPACKOVA, JITKA KOMARKOVA, PAVEL SEDLAK
Faculty of Economics and Administration, Institute of System Engineering and Informatics
University of Pardubice
Studentska 84, Pardubice, 53210
CZECH REPUBLIC

Abstract: - Process of decision making represents so important part of the activity of managers that some people have dedicated their lives seeking for tools helpful in this process. Nowadays, importance of these tools has increased due to the increasing amount of available information. During last years many tools supporting decision making have arisen along with many new buzzwords. In this article we explain three different types of tools which can be used as pre-processing tools in decision making: text categorization, text clustering, and spatial analyses including results visualisation.

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

1 Introduction

The world's total yearly production of print, film, optical, and magnetic content would require roughly 1.5 billion gigabytes of storage. This is the equivalent of 250 megabytes per person for each man, woman and child on the Earth [18].

According to Porter [22] organisations need information to support decision making in various levels of management in order to remain or become truly globally competitive. But the necessity of having information can result in two problems: (1) shortage of information and (2) information overload. Process of decision making represents complex and difficult task which is, among others, highly dependent on the accessibility of high quality information. 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.

Nevertheless, 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 [7] 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 three case studies. Two of the methods which are from the branch of soft computing, cover text categorization and text clustering tasks. In the third 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 1970s 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], [10].

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 [23]: "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 [8] 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 [15], classification of news stories [9] or guidance of a user's search on the Web [1], [19], [20].

A growing number of statistical classification methods have been applied to text mining, such as Naive Bayesian [10], and Support Vector Machines [5], [13]. A comprehensive comparative evaluation of a wide-range of text categorization methods is in [12], [13].

4 Geographic Information in Decision Making

Almost everything what happens, happens somewhere. 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. Spatial data can be connected to any space, not only to the Earth but people mostly use data connected to the Earth which are very often called geographic data. GIS are usually used to process and analyze geographic data [17].

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 [4]. Route planning is a significant branch of GIS and SDSS utilization [14]. Planning facility location is another typical example of a solved problem, e.g. shopping mall location [3]. Utilization of GIS in crisis management is a very important issue, e.g. a Multi-Criteria SDSS was proposed to help manage flooding [16].

Web-based solutions are spreading rapidly along with an increasing demand for easy access of end-users to geographic data [24]. Web based spatial OLAP system has been proposed [2]. The factors which influence success of web-based SDSS have been studied [11].

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 [25]. 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.

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

Head of a department of a university is a manager in this case. He wants to run a quick scan of preliminary drafts of thesis to see if some topics do not overlap; if there are some relations among teachers' interests; or to make year-on-year comparison. The dataset contains 75 very short documents covering preliminary drafts of thesis of selected department.

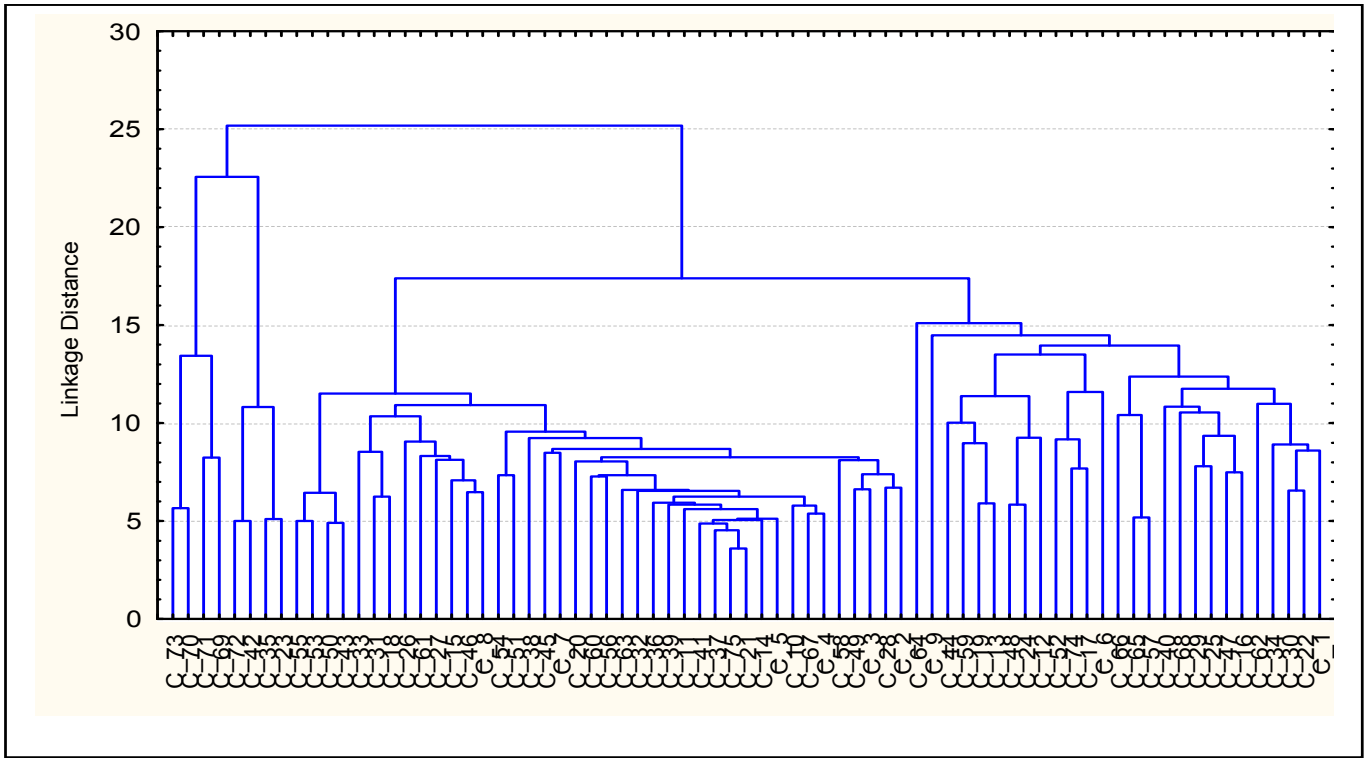

Fig. 1. Tree diagram of processed documents. Created by means of Ward's method and Euclidean distance.

The quick scan was conducted using hierarchical cluster analysis that gives visual output. Result of this experiment (see Fig. 1) shows that the documents are concentrated into two big clusters (No 3 and 4) and two minor clusters containing 8 documents (No 1 and 2). But no documents overlap. With regard to the number of department members (15) and the fact that each of them is a supervisor of more then one thesis, more different clusters could be assumed. Nevertheless, it is usual that several people work on similar problem or cooperate in the framework of one research project. In such cases overlapping of topics can happen. This particular example shows similarity of the drafts of the thesis of 4 people who are engaged in one research project. The less expected result is in the cluster 4. Drafts of thesis found in this cluster belong to the persons who do not cooperate. Moreover, they do not know about the similarity of the topics.

Unfortunately this kind of text pre-processing can not replace human work completely. In this case head of the department has to know engagement of the people in projects to interpret results, but visualization of outputs can ease the decision making process.

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.

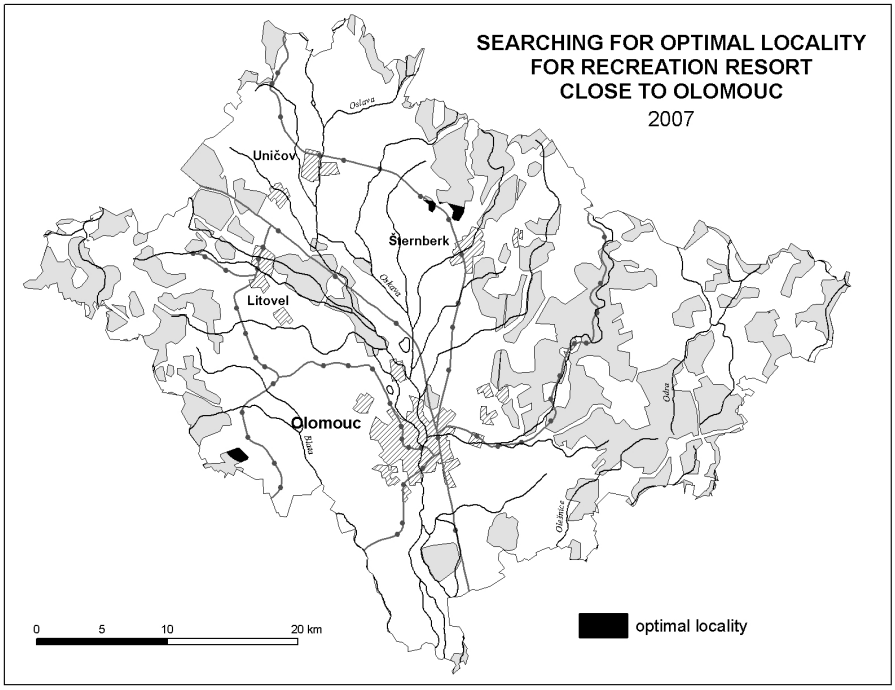

Fig. 2. Solving spatially oriented problem. Result of spatial analyses displayed on the map.

8 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 geographic 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 but it is usually very difficult to retrieve this information from the text. 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.

Study of spatial relationships and patterns is very often a significant part of preparatory phase of a project. But only very basic information can be obtained just by looking at the maps. Usually, sophisticated spatial analyses like topological overlay, calculating areas, network analyses, etc. have to be done in order to obtain useful actionable knowledge so utilization of spatial DSS can significantly improve quality and speed of decision making process. Case study C shows a potential of utilization of geographic information for supporting decision making.

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] Bimonte, S., Wehrle, P., Tchounikine, A., Miquel, M.: GeWOlap: A Web Based Spatial OLAP

Proposal. Lecture Notes in Computer Science, Vol. 4278, Springer-Verlag, Berlin Heidelberg New York (2006)
[3] Cheng, E.W.L., Li, H., Yu, L.: A GIS approach to shopping mall location selection. Building and Environment 42 (2007) 884-892
[4] 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
[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] Feldman M.S., March J.G.: Information in organizations as signal and symbol. Administrative Science Quarterly 26 (1981) 171-186
[8] Forrester Research. Coping with Complex Data. The Forrester Report (1995)
[9] Hayes, P., et al.: A news story categorization system. In Second Conference on Applied Natural Language Processing (1988)
[10] Holsapple, C.W., Whinston, A.B.: Decision support systems: A knowledge-based approach. West Publishing Company, Minneapolis/St. Paul (1996)
[11] 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
[12] 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)
[13] 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)
[14] Keenan, P.B.: Spatial decision support systems for vehicle routing. Decision Support Systems 22 (1998) 65-71
[15] Lang, K.: NewsWeeder: Learning to Filter Netnews. In International Conference on Machine Learning (1995)
[16] 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
[17] Longley, P. A.: Geographic information systems and science. John Wiley & Sons, Chichester (2001)
[18] Lyman, P., Hal R.V.: How Much Information [online]. [cit. 2007-08-08]. Available from <http://www.sims.berkeley.edu/how-much-info-2003>
[19] 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)
[21] Mladenic, D.: Text-Learning and Related Intelligent Agents: A Survey. IEEE Intelligent Systems 14 (1999) 44-54
[22] Porter, M. E.: On competition. Harvard Business School Publishing, Boston (1998)
[23] Tucker, M.: Dark Matter of Decision Making. Intelligent Enterprise Magazine 2 (1999)
[24] 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)
[25] Vozenilek, V.: Cartography for GIS. Vydavatelství UP, Olomouc (2005)