# Text categorization – potential tool for managerial decision-making

HANA KOPACKOVA
Faculty of Economics and Administration, Institute of System Engineering and Informatics
University of Pardubice
Studentska 84, Pardubice, 53210
CZECH REPUBLIC
hana.kopackova@upce.cz

*Abstract:* One of the most important manager activities is decision-making. Especially in these days, full of different information, it is necessary to distinguish between important and unimportant information. The aim of this paper is to find methods which fulfil criteria put on managerial decision-making process. Usage of text categorisation can significantly lower manager workload. Another aim can be defined as raise of objectivity in decision-making process. Automated processing of text documents can prevent simplifications and generalisation, which allow us to decide on the base of small amount of cases and widen this decision on all cases.

*Key-Words:* - decision-making, management, text categorization, DSS, Business Intelligence

## 1 Introduction

The work of managers, scientists and engineers is mainly focused on solving problems and making decisions. As the world is more and more globalized, this kind of work becomes much complex with demand on great amount of information. Management information is the information collected to determine how well the company is running. To manage some business effectively managers must decide what metrics (or key performance indicators) they will track to measure performance.

Necessity and volume of information forces us to improve our problem solving and decision-making capabilities using different tools and machines. Such tools can be divided into groups according to hierarchical level of managerial positions as each position needs different type of information.

Top Managers hold positions like chief executive officer or chief operating officer and are responsible for the overall direction of the organization. They are responsible for creating a context for change, just as they are responsible for developing employee's commitment to and ownership in the company's performance and for creating a positive organizational culture through language and action.

Middle Managers hold positions like plant manager, regional manager or divisional manager. This group of managers is responsible for setting objectives consistent with top management's goals and planning and implementing subunit strategies for achieving these objectives. Coordinating and linking departments, groups and divisions with in a company is another responsibility for middle managers.

First Line Managers hold positions life office manager, shift supervisor, or department manager. Their primary responsibility is to manage the performance of entry level employees, who are directly responsible for producing a company's goods and services. They also teach entry level employees how to do their jobs.[1]

Each level of management differ in information needed so possible tools used to support decision-making must be also different. Next chapter will describe some

## 2 Tools supporting decision-making

Classical supporting tools can be divided, according to levels of management, into 4 separate systems. These systems are vertically connected so that the lower level systems provide necessary inputs of the higher level. Description and figure can be found for example in [2].

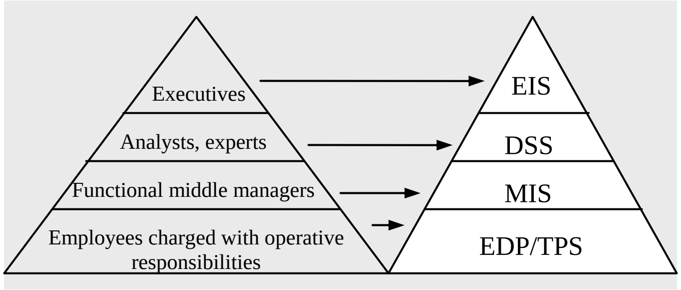
Executives
Analysts, experts
Functional middle managers
Employees charged with operative responsibilities
EIS
DSS
MIS
EDP/TPS

**Figure 1: Information system pyramid**

The association between strategic goals and performance indicators has been changing the way information systems have been being developed. Most major management software companies have been aggregating the so-called Business Intelligence (BI) to their systems.

According to [3, 4], Business Intelligence comes from the need to extract and make information available from the "pile" of data a company generates and to store it in its databases. However, the author claims that:

"This increasing information inundation makes the decision-making process more difficult, as high- and midmanagement feel impotent in the process of searching and recovering it. Legacy, and the emerging ERPEnterprise Resource Planning systems, integrated corporate systems, do not bring managerial information in its most palatable form. Quite the contrary, information that is vital for strategic decision-making is hidden in thousands of tables and files that are inaccessible to common mortals, connected by transnational relationships and correlations, in an autonomy that is inadequate for decision-makers."[3]

Business Intelligence involves integrating processes to provide a holistic view of a business that provides value (or intelligence) about the business going forward. Business Intelligence systems are designed to give executives information from their operational systems to help make better decisions for the business. Data Marts, Data Warehousing, OLAP and Web methodologies and technology are all tools to assist in providing this information.

All presented tools are in its basic form focused on structured numeric data, maintained in any kind of database. Some data mining tools cover also tools applicable on unstructured textual data, but such instruments are complicated and have too many functions, and due to this fact they are awfully expensive.

## 3 Do we need tools for unstructured data processing?

Mark Tucker very good answered this question in [5]. ”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… Business processes have always relied on unstructured information, and the volume and sources of this information are increasing, not decreasing. The World Wide Web, the corporate Intranet, email, and online discussion groups are just a few of the familiar examples… Information drives our business decisions, and it always has. What has changed dramatically is the kind of decisions we make. As the economy shifts from an industrial model to a knowledge-driven one, we need more and more information to support the decision-making process, and the dynamic nature of our business environment is such that less and less of this information fits the structured information model."

Forest Research [6] has predicted that unstructured data (such as text) will become the predominant data type stored online. This implies a great opportunity of possible more effective use of repositories of business communications, and other unstructured data, by using computer analysis. But the problem with text is that it is not designed for using by computers. Unlike the tabular information typically stored in databases today, documents have only limited internal structure if any. Furthermore, the important information they contain is not explicit but is implicit: buried in the text. [7]

Managerial decision-making process can be highly dependent upon hidden information in text documents nevertheless careful reading and sorting of documents is time consuming work. This type of activity waste working time of managers and at the end, it can cause wrong decision. Now we must ask how we can help managers to cope with numerous text documents.

The aim of this paper is finding of methods which can fasten decision-making process on all levels of management and make it much easier for managers using great amount of information in text form. Usage of text categorization methods, which are well known in the branch of information retrieval, but they are not often used to support decision-making process can significantly lower manager workload. Another aim can be defined as raise of objectivity in decision-making process. Automated processing of text documents can prevent simplifications and generalisation, which allow us to decide on the base of small amount of cases and widen this decision on all cases. Unfortunately, this approach is commonly used having too much text documents and only little time to read them.

## 4 Text categorization in managerial decision-making

Text categorisation is, the assignment of free text documents to one or more predefined categories based on their content, is an important component in many information management tasks; real-time sorting of e-mail or files into folder hierarchies, topic identification to support topic-specific processing operations, or structured search and browsing. The automated categorization (supervised learning) of texts into predefined categories has witnessed a booming interest in the last 10 years, due to the increased availability of documents in digital form and the real need to organize them.

Here are some examples of possible usage of text categorization; building of personalised Netnews filter which learns about the news-reading preferences of a user [8], classification of news stories [9] or guidance of a user‘s search on the Web [10, 11, 12, 13].

A growing number of statistical classification methods have been applied to text categorization, such as Naive Bayesian [14], Bayesian Network [15], Decision Tree [16] [17], Neural Network [18], Linear Regression [19], k-NN [20], Support Vector Machines [21, 22], Boosting [23] and Genetic Algorithms [24]. A comprehensive comparative evaluation of a wide-range of text categorization methods is reported in [21, 22].

Usually text categorisation methods are tested in ideal environment which cover:

- great amount of training documents (Reuters collection, newsgroups, TDT Pilot study corpus... – about 20000 documents),
- and approximately similar length of documents.

These conditions are almost impossible to guarantee in real managerial decision-making. The aim of this paper is to find such methods which can be applied directly by managers if they need them. This statement assumes that managers are the persons who will decide which documents will be used as training documents. No one can expect that he/she will have great number of training documents with similar length.

Methods suitable to support managerial decision-making process must fulfil these criteria:

- easy to implement,
- fast to process categorization,
- cheap,
- stable for differences in length of document,
- learning model build from small number of documents,
- high precision.

Three methods which are easy to apply will be tested for being used in managerial decision-making. These methods are K-nearest neighbour, Rocchio algorithm and Naive Bayes algorithm.

## 4.1 K-nearest neighbour

The *k*-nearest neighbour classifier labels an unknown document ***d*** with the label of the majority of the *k* nearest neighbours. A neighbour is deemed nearest if it has the smallest distance, in the Euclidian sense, in feature space. For *k* = 1, this is the label of its closest neighbour in the learning set. The *k*-nearest neighbour method is intuitively a very attractive method.

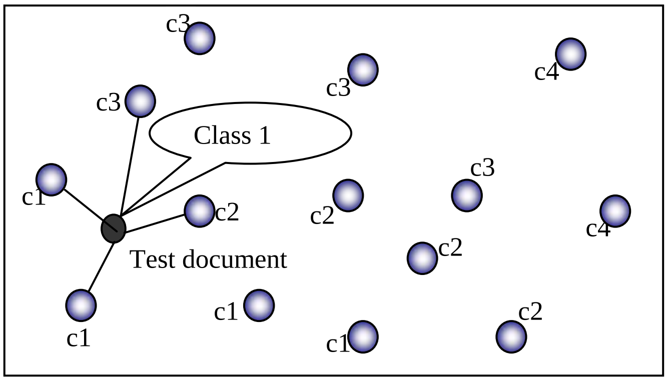

**Figure 2: K-NN classifier**

A disadvantage of this method is its large computing power requirement, since for classifying an object its distance to all the objects in the learning set has to be calculated.

## 4.2 Rocchio algorithm

The basic idea of the algorithm is to represent each document ***d*** as a vector $\vec{d}_i = \left(d_{i1},\ldots,d_{iN_T}\right)^T$ in a vector space so that documents with similar content have similar vectors. $d_{ik}$ is weight of the word $w_k$ in document, $\vec{d}_i$, $\vec{d}_i \in D$ where $D$ is set of all documents and $N_T$ is number of selected features.

The Rocchio algorithm learns a class model by combining document vectors into a prototype vector $\vec{c}_i$ for every class $C_i \in \boldsymbol{C}$. Prototype vectors are generated by adding the document vectors of all documents in the class.

$$\vec{c}_i = \sum_{\vec{d}_i \in C_i} \vec{d}_i \tag{1}$$

The resulting set of prototype vectors, one vector for each $C_i \in \boldsymbol{C}$, represents the learned model. This model can be used to classify a new document $\delta$. Again the document is represented as a vector $\vec{\delta}_i$ using the scheme described above. To classify $\delta$ the cosine of the prototype vector of each class with $\vec{\delta}_i$ is calculated. The new document is assigned to the class with which its document vector has the highest cosine.

$$H_{\cos}(\delta) = \arg\max_{C_i \in \mathrm{C}} \cos(\vec{\delta}_i, \vec{c}_i) \tag{2}$$

$H_{\cos}(\delta)$ is a hypothesis approximating function for assignment of document $\delta$ into wining category. The cosine measures the angle between the vector of the document being classified and the prototype vector of each of the classes. The smaller the angle, the larger will be the cosine. So $\delta$ is assigned to the class which has the smallest angle between its prototype vector and $\vec{\delta}_i$.

The algorithm can be summarized in the following decision rule:

$$H_{\cos}(\delta) = \arg\max_{C_i \in \mathrm{C}} \frac{\vec{\delta}_i \cdot \vec{c}_i}{\|\vec{\delta}_i\| \cdot \|\vec{c}_i\|} = \tag{3}$$

$$\arg\max_{C_i \in \mathrm{C}} \frac{\sum_{j=1}^{N_T} \delta_{ij} \cdot c_{ij}}{\sqrt{\sum_{j=1}^{N_T} (\delta_{ij})^2} \cdot \sqrt{\sum_{j=1}^{N_T} (c_{ij})^2}} \tag{4}$$

### 4.3 Naive Bayes algorithm

The Naive Bayes classifier is constructed by using the training data to estimate the probability of each class given the document feature values of a new instance. Although this model is a strong simplification of the true process by which text is generated, the hope is that it still captures most of the important characteristics.

Assumption: Documents are generated by drawing words from a probability distribution. Let's assume that we have $|\boldsymbol{C}|$ probability distributions, one for each category. All documents assigned to a particular class are generated from the probability distribution associated with this class in a number of indepen-dent trials. The i-th word of the document is generated by the i-th independent trial.

Probabilistic classifiers try to estimate $P(C_i|\delta)$, the posterior conditional probability that a document $\delta$ is in class $C_i$. Bayes' rule says that to achieve the highest classification accuracy, $\delta$ should be assigned to the class for which $P(C_i|\delta)$ is highest.

$$H_{BAYES}(\delta) = \arg\max_{C_i \in \mathrm{C}} P(C_i|\vec{\delta}_i) \tag{5}$$

Bayes' theorem can be used to split the estimation of $P(C_i|\delta)$ into two parts:

- *$P(C_i)$* is the prior probability that a document is in class *$C_i$*.
- $P(\vec{\delta}_i|C_i)$ is the likelihood of observing document $\delta$ in a given class.
-

$$P(C_i|\vec{\delta}_i) = \frac{P(\vec{\delta}_i|C_i) \cdot P(C_i)}{\sum_{C_j \in \mathrm{C}} P(\vec{\delta}_i|C_j) \cdot P(C_j)} \tag{6}$$

*$P'(C_i)$*, the estimation of *$P(C_i)$*, can be calculated from the fraction of the training documents that is assigned to this class. Easily this measure represents the rate of class $C_i$ volume to sum of volumes of all classes.

$$P'(C_i) = \frac{|C_i|}{\sum_{C_j \in \boldsymbol{C}} |C_j|} \tag{7}$$

The estimation of $P(\vec{\delta}_i|C_i)$ is more difficult. $P(\vec{\delta}_i|C_i)$ is the probability of observing a document like $\delta$ in class $C_i$. Since there are a huge number of different documents it is impossible to collect a sufficiently large number of training examples to estimate this probability without prior knowledge or further assumptions. Additional assumptions that denoted Bayes algorithm as naive are those:

- word's occurrence is dependent on the class the document comes from,
- it occurs independently of the other words in the document.
-

$$P(\vec{\delta}_i|C_i) = \prod_{k=1}^{N_T} P(w_k|C_i) \tag{8}$$

$w_k$ is particular word in document $\delta$ and conditional probability. $P(w_k|C_i)$ determine probability of occurence of word *$w_k$* in category *$C_i$*. As the result we obtain this decision rule:

$$H_{BAYES}(\delta) =$$

$$\arg\max_{C_i \in \mathrm{C}} \frac{P(C_i) \cdot \prod_{k=1}^{N_T} P(w_k|C_i)}{\sum_{C_j \in \mathrm{C}} \left( P(C_j) \cdot \prod_{k=1}^{N_T} P(w_k|C_j) \right)} \tag{9}$$

## 5 Experimental results

In the testing environment I used 50 documents in Czech language; 25 of them were focused directly on the branch of waste management and the rest 25 documents were not specialised – only covered problem of environment. The shortest document had only 98 words and the longest had 1400 words. For feature selection were used three different methods: Chi-square, Mutual information, and Information gain [25]. Two methods for term weighting were tested TF – term frequency and TFIDF. After pre-processing stage database was filled with 10571 words.

The result can be seen in figure 3 (CCI means correctly classified instances and K means Kappa statistics).

| | Naive Bayes | | K-NN | | Rocchio algorithm | |
|---|---|---|---|---|---|---|
| | CCI-NB [%] | K-NB [%] | CCI-KNN [%] | K-KNN [%] | CCI-VM [%] | K-VM [%] |
| TFchi-square | 96,00 | 92,00 | 62,00 | 24,00 | 100,00 | 100,00 |
| TFmutual information | 98,00 | 96,00 | 62,00 | 24,00 | 100,00 | 100,00 |
| TFinformation gain | 96,00 | 92,00 | 50,00 | 0,00 | 100,00 | 100,00 |
| TFIDFchi-square | 96,00 | 92,00 | 66,00 | 32,00 | 100,00 | 100,00 |
| TFIDFmutual information | 96,00 | 92,00 | 68,00 | 36,00 | 100,00 | 100,00 |
| TFIDFinformation gain | 92,00 | 84,00 | 50,00 | 0,00 | 100,00 | 100,00 |

**Figure 3: Experimental results**

Experiments that were published for example in [20] introduced K-NN method as very efficient, however these experiments used Reuters corpus filled with articles of similar legth. My experiments proved that K nearest neighbour algorithm is very sensitive to length differences (documents using same words have long Euclidean distance between them if they differ in length) so it is not suitable to support managerial decision-making as it was described here. On the other hand Rocchio algorithm and Naive Bayes can serve as very helpful tool.

Differences between TF and TFIDF methods are not so dominant to say that one of them is better.

## 6 Conclusion

Unstructured information in the form of textual documents are used in managerial-decision making very often, nevertheless support for this kind of action is inadequate. In this paper I shortly introduce tools that deal with textual information including proposal of methods that fulfil demands specially put on this kind of decision-making.

## 7 Acknowledgment

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] International Development Law Organization. Management. [Online] [cit. 04-06-05] URL: < http://www.idlo.int/texts/IDLO/mis6675.pdf>

[2] Drótos, G. Perspectives of Information Systems in Organisations. Notes to the CEMS course. Budapest, 2002.

[3] BARBIERI, C. BI-Business Inteligence – Modelagem e tecnologia. Rio de Janeiro: Axcel Books, 2001.

[4] Damiani, W. B., Galery, A. D., Madureira, R. T. Opinion Poll Regarding Business Report Forms. Hawaii International Conference on Business. Honolulu, June 18 -21, 2003.

[5] Tucker, M. Dark Matter of Decision Making. Intelligent Enterprise Magazine, vol. 2, num. 13. 1999

[6] Forrester Research. Coping with Complex Data. The Forrester Report, 1995

[7] Tkach, D. Text Mining Technology, Turning Information Into Knowledge. White paper from IBM. 1998

[8] Lang K., NewsWeeder: Learning to Filter Netnews, *International Conference on Machine Learning*, 1995.

[9] Hayes P. et al., A news story categorization system, *Second Conference on Applied Natural Language Processing*, 1988

[10] Mitchell T. et al., WebWatcher: A Learning Apprentice for the World Wide Web, *AAAI Sprig Symposium on Information Gathering from Heterogenous, Distributed Environments*, 1995

[11]Bollacker K., Lawrec S., Giles L., Citeseer: An Autonomous System for Processing and Organizing Scientific Literature on the Web, *Conference on Automated Learning and Discovery*, Pittsburgh, 1998

[12] Mladenic D.: Personal WebWatcher: Implementation and Design, *Tech. Report IJS-DP-7472*, J. Stefan Inst., 1996

[13] Sklenák V. a kol., *Data, informace, znalosti a internet*, 1. vyd. Praha, C. H. Beck, 2001

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

[15] Sahami M., Learning limited dependence Bayesian classifier. In KDD- 96: Proceedings of the second international *Conference on Knowledge Discovery and Data Mining*, AAAI press, 335-338, 1996.

[16] Quinlan J. R., C4*.5: Programs for machine learning*, Morgan Kaufmann, 1993.

[17] Weiss S. M et. al., *Maximizing text-mining performance*, IEEE Intelligent systems,1999.

[18] Wiener E., Pederson J. O., Weigend A. S., A neural network approach to topic spotting. Proceedings of *SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval*, 1995.

[19] Yang Y., Chute C. G., A linear least squares fit mapping method for information retrieval from natural language texts. Proceedings of the *14th International Conference on Computational Linguistics (COLING 92)*, 1992.

[20] Yang Y., An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, Vol.1, No.1/2, 1999.

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

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

[23] Schapire R. E., Singer Y., Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3), 2000.

[24] Prasanna K., Khemani D. Applying Set Covering Problem in Instance Set Reduction for Machine Learning Algorithms. WSEAS Multiconference: Software Engineering, Parallel & Distributed Systems (SEPADS 2004). Salzburg, Austria, 2004.

[25] Janakova, H. Text categorization with feature dictionary – problem of Czech language. WSEAS TRANSACTIONS on INFORMATION SCIENCE AND APPLICATIONS, Issue 1, Volume 1, July 2004, s. 368 - 372, ISSN 1790-0832