EVALUATION OF TEXT CLASSIFICATION ALGORITHMS

for a Web-based Market Data Warehouse

Carsten Felden, Peter Chamoni

Institute of Logistics and Information Management,University of Duisburg-Essen, Lotharstraße 65, 47057 Duisburg, Germany

Keywords: Active Data Warehouse, Classification, Text Mining

Abstract: Decision makers in enterprises cannot handle information flooding without serious problems. A market data information system (MAIS), which is the foundation of a decision support system for German energy trading, uses search and filter components to provide decision-relevant information from Web-documents for enterprises. The already implemented filter component in form of a Multilayer Perceptron has to be benchmarked against different existing algorithms to enhance the classification of text documents. An evaluation environment with appropriate algorithms is developed for this purpose. Also a set of test data is provided and a tool selection is shown which implement different text mining algorithms for classification. The benchmark results will be shown in the paper.

1 INTRODUCTION

The information supply for decision makers is characterized by the problem of information overload (Kamphusmann, 2002). The World Wide Web (WWW) offers the opportunity to integrate textual information from external sources (Colomb, 2002). Decision-relevant unstructured data cannot be determined and extracted from these resources without problems (Hackathorn, 1998). A central component of an already implemented business intelligence (BI) system for energy trading is a filter component, which can classify relevant Web-based documents to ensure the information supply for decision makers. Empirical studies show the urgent need of automated and integrated information supply in almost all companies.

Paper documents and local data files still dominate the information infrastructure in most German enterprises. This is the result of a *COI* study on a sample of 50 companies, which was presented in the weekly German newspaper *Computer Zeitung* in 2004. Roughly a third of all documents is accessible from enterprise-wide file servers. Just 14 percent are scanned into digital archives and 29 percent of the requested enterprises have a knowledge management system for the administration of this important resource. One consequence is that users must search a stored file or paper document on local discs, change media or filing cabinets etc. on average 25 times a week. Approximately five minutes of work time are spent in each search case. The storage place of nearly a third (29 percent) of the documents is simply unknown. Every fourth document is stored at a completely wrong place. Therefore decision makers demand a faster search and a better access to the collected enterprise knowledge (Computer Zeitung, 2004). The ability to classify already stored documents and assign them to specific subjects can improve the search and access of unstructured data in a database. There is an implicit need to be able to classify unstructured documents according to specific subjects like time series or stored reports.

The goal of this paper is the development of an evaluation environment for the comparison and selection of classification algorithms as a filter component in an existing market data information system (MAIS) for energy trading. First we present the information system to give an impression about the already implemented solution. Subsequently, the structure of a test environment and the pre-processing for the evaluation of classification algorithms is shown.

Felden C. and Chamoni P. (2005).
EVALUATION OF TEXT CLASSIFICATION ALGORITHMS - for a Web-based Market Data Warehouse.
In *Proceedings of the First International Conference on Web Information Systems and Technologies*, pages 359-362
DOI: 10.5220/0001233903590362

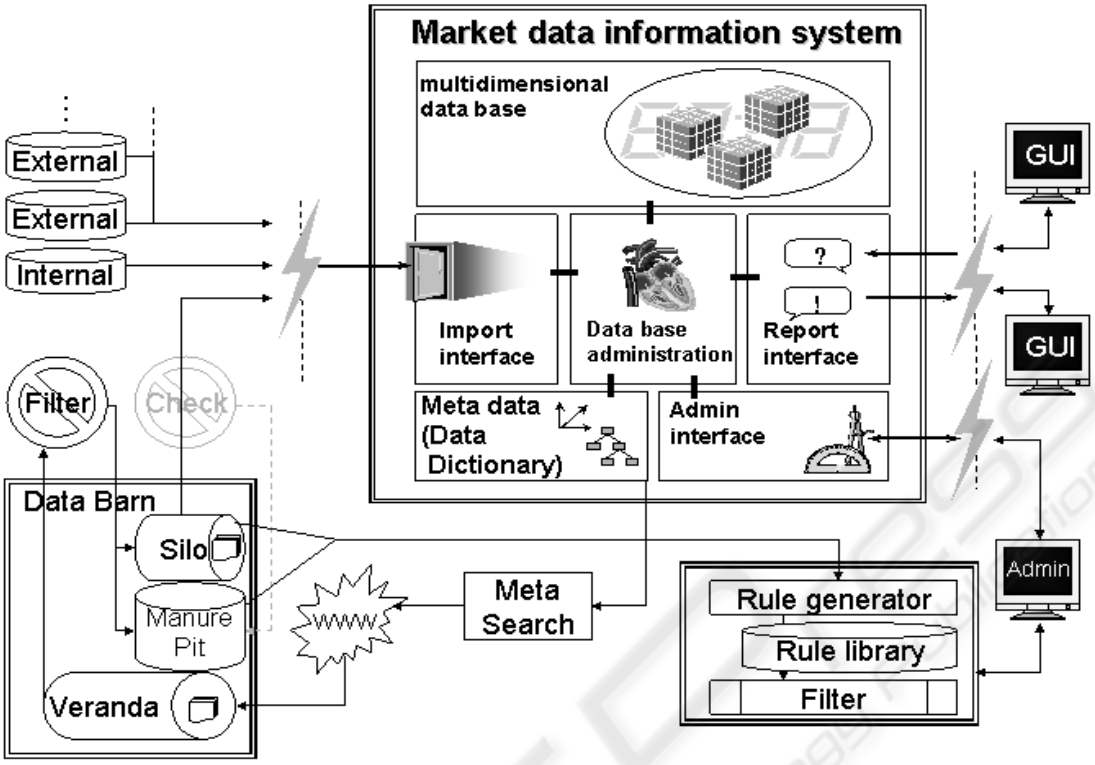

Figure 1: Architecture of the Market Data Information System

2 CONCEPTUAL FRAMEWORK AND CLASSIFICATION ALGORITHMS

There is an implicit need for fast decision-relevant information support. This leads to information systems, which can recognize situations automatically and enforce user actions. Thus data warehouses nowadays change into active data warehouses which provide a high availability of business-critical information (Inmon, 2002). To gain some knowledge about the classification done by the algorithm we present the architecture of MAIS first.

2.1 Basic Concept

The basic idea is to integrate company internal and external Web-based data for decision support. There already exist approaches to integrate structured and unstructured data from these sources in a data warehouse. This forms a base for further research to develop an active data warehouse triggered by text mining on Web-documents. The underlying information system was realized in a conjoint project with VEW Energie GmbH.

There are several approaches for the distinction of interesting and uninteresting internet documents to be mapped into a data warehouse. For this reason disjoint classes of documents are used in order to make a clear separation. The data integration process is divided in four different steps and illustrated in figure 1.

First, a set of dimensions (meta data) of the implemented multidimensional data model is the input for the retrieval process to find external Web-documents. The meta data are e. g. named information sources enriched with attribute values like stored regions. The main advantage is that the unstructured data can be mapped to the respective OLAP-slice (Codd/Codd/Sally, 1993) according to the meta data used for the internet search later on.

All internet pages, which are found, will be stored in the so-called data barn in a second step. The data barn makes a transient as well as a persistent storage for the subsequent processing of these pages available. The processing means to find out, whether the identified Web-pages are relevant for the appropriate multi-dimensional structure. Those documents are finally stored in the data warehouse. The main advantage of this architecture is flexibility concerning the maintenance of such a system. If another algorithm receives better results, only the filter component has to be changed.

2.2 Algorithms for Text Classification

The task of text classification is to group documents into two disjoint classes, interesting and uninteresting documents. It has to be distinguished between the examination of documents and the setting up of classification criteria. Both can be done manually or automatically. Alternative automatic procedures are the Multilayer Perceptron (Bishop, 1995), Rocchio algorithm, k-Next-Neighbour-Algorithm (kNN), Support Vector Machine (SVM), Naïve Bayes, Simple Logistics, Voted Perceptron, and HyperPipes which are already described in several publications, e. g. Sebastiani, 2002, Freund; Schapire, 1999; Hosmer; Lemeshow, 2000, Pampel, 2000, Rosenblatt, 1958, Sheng, 2005. These algorithms are chosen, because studies of several other evaluations have shown that they are the ones most examined (e. g. Collins, 2002, or Tveit, 2003). Additionally, we are able to compare our results with other studies. All these approaches are based on a vector space representation of textual documents (Kobayashi/Aono, 2004; Kamphusmann, 2002, Colomb, 2002).

There is not a single software suite with implementations of all algorithms. Therefore the authors decided to use several software packages. The Java Tool Weka (version 3.4.2) implements the procedures kNN, Naïve Bayes and decision trees via J48, based on C4.5 (Witten/Frank, 2000). SVM is provided by the C-program application SVMlight (Joachims, 1998). The Rocchio algorithm is processed by a self-developed java application. The $F_ß$ – measure of *van Rijsbergen* is used for the evaluation of the classifications. It is calculated from recall and precision.

3 EVALUATION PROCESS

The following chapter describes the structure of the test environment and the evaluation of algorithms for text classification. This description represents the basis of further evaluations in order to achieve an intersubjective and intertemporal comparability. This is important, because the majority of text classification studies use the Reuters textual data set which is highly normalized and standardized. Comparable pre-processing steps, as they are implemented in the context of MAIS, are insignificant of those evaluations. These aspects affect the result. Therefore the results of such studies are judged as insufficient.

The integration process is simulated in order to get a set of training data. The results of these queries are stored as text documents and define a set of training data. Finally, the documents are classified manually. The test data set has been collected during a period of two months and covers 1,304 documents. This amount is small compared to other test collections. But it seems appropriate because just a binary classification is desired and the documents are much larger compared to the Reuters Corpus. Another reason is that the original problem domain (energy trading) needs just a small number of documents which have to be classified. The necessary split into training, test, and validating data is to be made in the relative relationship 50: 20: 30.

The data set developed contains more than 67,000 different words. Capitalized and non-capitalized words were accumulated to a single value. Variations of pre-processing steps to develop different input files for the tools are created. The variations are as follows: An implementation of a stop-word-list is developed in order to delete meaningless words. Further on all hyphens between words have to be deleted and the remaining word parts are regarded as individuals. It is remarkable that during the conversion from HTML to TXT documents many special characters are produced automatically. The permissible terms are limited to a-z, A-Z, äüö, ÄÜÖ. Double ‘s’ and ‘ß’ are treated equally. ‘.’ is determined as decimal separator. A further step is the performing of stemming to improve the quality of the word list. Finally, the remaining words are integrated into a total list. The words which occur only once in a text document are deleted. Other words which only build the upper five percent in frequency per document are also deleted. Finally the word vectors are generated, which represent the input of the selected classification tools. The variations used obtain nine different input files with a number of terms between 10,343 and 33,776. Table 1 shows the final results of the text classification.

Table 1: Classification Results

Test-No.	Term #	Algorithm	$F_ß$
1	10,511	Naïve Bayes	0.7950
2	10,343	SVM	0.7868
3	15,676	Naïve Bayes	0.7810
4	31,602	Naïve Bayes	0.7870
5	33,247	Naïve Bayes	0.7870
6	33,392	SVM	0.7973
7	32,854	SVM	0.7844
8	33,602	Naïve Bayes	0.7865
9	33,776	Naïve Bayes	0.7865

Obviously, the Support Vector Machine of test No. 6 has highest $F_ß$-value. Most of the algorithms have reached their best result in this test (e. g. Voted Perceptron and Simple Logistic with value 0.7615, Rocchio with value 0.7261, or k-NN with value 0.6395). HyperPipes (0.6865) and AdaBoost.M1 (0.7360) reached the best results in test No. 3. The MLP (0.5000) and J48 (0.7135) get their best results in test No. 1.

Looking at the classification including the pre-processing, the point of view has to be changed. Test No. 1 gives good results, but requires enormous additional pre-processing efforts. Looking at test No. 9, an $F_ß$-value of 0.7865 can be found. This means comparable results can be gained with less pre-processing for the same algorithm. These results are comparable to other studies, where SVM is often the best classificator. As the MLP cannot reach the former value of 0.818 in the first implementation of MAIS, it is replaced by the Naïve-Bayse-algorithm which now builds the filter component with less pre-processing effort.

4 CONCLUSION

The recent development of analytical information systems shows that the necessary integration of structured and unstructured data sources in data warehousing is possible. The implementation of MAIS has proved this. Only documents of decision relevance should be delivered to the management. The ROI of data warehouse projects can be increased, if event-based and accepted information improves the decision quality significantly. The information flow alignment in MAIS is equivalent to a classification problem. The quality of classification algorithms must be examined in regular time intervals to guarantee best results. Therefore it is necessary to optimize the structure of the test environment. This environment has to support an intersubjective and intertemporal comparability of the test results. Classification evaluations are often accomplished; however these results are only important in the context of the selected data set and evaluation environment. In order to get concrete statements for MAIS, such an evaluation environment and the results are described in this paper. The most relevant documents are to be found so not just the classification itself has to be optimized, but the internet retrieval as well in order to find the *perfect* search terms.

REFERENCES

Bishop, C. M., 1995. Neural Networks for Pattern Recognition. Clarendon Press, Oxford 1995.

Codd, E.; Codd, S.; Salley, C., 1993. Providing OLAP (On-line Analytical Processing) to User-Analysts. An IT Mandate. White Paper. Arbor Software Corporation.

Collins, M., 2002. Ranking Algorithms for Named-Entity-Extraction: Boosting and the Voted-Perceptron. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 489-496.

Colomb, R. M., 2002. Information Retrieval – The Architecture of Cyberspace, London.

Computer Zeitung, 2004 (no author). Wildwuchs in der Ablage, in: Computer Zeitung, 35. Jahrgang, Nr. 50, 6. Dezember 2004, p. 17.

Freund Y.; Schapire R., 1999. Large Margin Classification Using the Perceptron Algorithm, Machine Learning 37, Dordrecht, pp. 277–296.

Hackathorn, R. D., 1998. Web Farming for the Data Warehouse, San Francisco.

Hosmer, D. W.; Lemeshow, S., 2000. Applied logistic regression, 2. edition. New York.

Inmon, W. H., 2002. Building the Data Warehouse, 3rd Edition. Wiley, New York.

Joachims, T., 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Forschungsbericht des Lehrstuhls VIII (KI), Fachbereich Informatik, Universität Dortmund.

Kamphusmann, T., 2002. Text-Mining. Eine praktische Marktübersicht. Symposium, Düsseldorf.

Kobayashi, M.; Aono, M., 2004. Vector Space Models for Search and Cluster Mining. In (Berry, M., Ed.): Survey of Text Mining. Clustering, Classification, and Retrieval. ACM, New York et al.; pp. 103 - 122.

Pampel, F. C., 2000. Logistic Regression. A primer. Thousand Oaks: Sage.

Rosenblatt, F., 1958. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65, 1958, pp. 386 - 408. (reprint in: *Neurocomputing* (MIT Press, 1998).)

Sebastiani, F., 2002. Machine Learning in Automated Text Categorization. In: ACM Computing Surveys, Vol. 34, No. 1, March 2002, pp. 1 - 47.

Sheng, J., 2005. A Study of AdaBoost in 3 D Gesture Recognition, http://www.dgp.toronto.edu/~jsheng/doc/CSC2515/Report.pdf, last call 2005-02-03.

Tveit, A., 2002. Empirical Comparison of Accuracy and Performance for the MIPSVM classifier with Existing Classifiers. http://www.idi.ntnu.no/~amundt/publications/2003/MIPSVMClassificationComparison.pdf, last call at 2005-02-02.

Witten, I. H.; Frank, E., 2000. Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, San Francisco.