A CONCEPTUAL MODEL FOR A MULTIAGENT KNOWLEDGE BUILDING SYSTEM

Antonina Kloptchenko, Tomas Eklund, Adrian Costea, Barbro Back
Turku Centre for Computer Science and IAMSR / Åbo Akademi University, Lemminkäisenkatu 14 B, 20520 Turku, Finland
Email: akloptch@abo.fi, toeklund@abo.fi, acostea@abo.fi, bback@abo.fi

Keywords: Software agents and multiagent systems, financial analysis, data mining, text mining

Abstract: Financial decision makers are challenged by the access to massive amounts of both numeric and textual financial information made achievable by the Internet. They are in need of a tool that makes possible rapid and accurate analysis of both quantitative and qualitative information, in order to extract knowledge for decision making. In this paper we propose a conceptual model of a knowledge-building system for decision support based on a society of software agents, and data and text mining methods.

1 INTRODUCTION

A huge amount of electronic information concerning different companies' financial performance and market situation is available in various databases and on the Internet today. This information can potentially be very valuable to companies' decision makers, their partners, competitors, investors, analysts, and stakeholders. These individuals want to extract relevant information for decision-making purposes from the widely available data storages on time and, preferably, by the click of a mouse button. The enormous supply of data available often exceeds our capacity to analyze it, leading to information overload. Users need to transform new data into valuable knowledge very quickly in order to react to rapidly changing conditions and make crucial decisions in time.

Although there are a number of methods and technologies available for creating, storing, and monitoring new data, there are not very many comprehensive and popular techniques for transforming all data into valuable information and knowledge. The fields of *knowledge discovery in databases (KDD), data mining (DM), and text mining (TM)* have provided a number of new approaches for analysis of large databases of financial data. KDD is the entire process of discovering interesting knowledge, such as patterns, associations, changes and anomalies, and significant structures from large amounts of stored data, while DM refers to the actual use of data mining tools for identifying patterns in the data (Fayyad et al. 1996). Most data mining techniques for financial applications deal with quantitative data. The analysis of qualitative information (company strategy, economic market outlook, i.e. the textual parts of financial statements, as well as information from outside sources) is very important and can be done using text mining approaches. TM refers to the nontrivial extraction of implicit, previously unknown, and potentially useful information from large textual datasets (Dorre et al., 1999). Unlike numeric data, textual statements contain not only the factual event but also the explanation for why it happens (Wuthrich et al. 1998).

The individuals are fortunate if the valuable data that they need are already stored in one available database on the web. More often the data are located on a number of different sites. An emerging problem is how to find and collect these data and process them so that they provide additional valuable knowledge. The majority of data mining techniques are meant for extracting meaningful patterns from numeric, well-structured databases. At the same time, ambiguously structured text databases grow large in size and significance, and require effective text mining techniques. A multi-agent software system consisting of a collection of individual software agents, each of which provides a certain task (Lesser 1995) and/or uses different data mining techniques, can be a possible solution for accomplishing this task.

In this paper we create a conceptual model of a knowledge building system based on a society of software agents, and data and text mining methods. Each agent exhibits intelligence by using different data and text mining methods. We believe that software agents, which are able to execute tasks on behalf of a business process, computer application, or an individual, are well suited to dealing with collecting, processing, and compiling vast volumes of dynamic data from distributed sources. The system could monitor new financial updates from a variety of sources, and calculate financial ratios for different companies. These data could be used for various tasks, for example, financial benchmarking and assessing creditworthiness of different companies.

Our model suggests the integration of several computing techniques, namely self-organizing maps for clustering quantitative information, decision trees and/or multinomial logistic regression for classifying new cases into previously obtained clusters, prototype-matching for semantic clustering qualitative information, and various techniques for text summarization. We have previously tested some of the techniques in certain modules of the conceptual model.

The paper is organized as follows: In Section 2 we describe the problem area and the approaches used in financial data analysis for solving the discussed problems and provide an overview of literature and related work in multiagent system design. We describe the conceptual model of our multiagent decision support system in Section 3. We explain the methodological issues of the different computational techniques we propose in Section 4. We discuss the possible limitations and difficulties associated with building and using the proposed system in Section 5. Section 6 contains our conclusions and the directions of future work.

2. DESCRIPTION OF PROBLEM AREA AND RELATED WORK

Financial analysis is very important in today's global economy. Access to more information should be beneficial to any investor or financial stakeholder. Financial benchmarking is an important and valuable tool for assessing the actual financial performance of a company. Financial benchmarking is the process of comparing a number of competitors according to, most commonly, a number of financial ratios, chosen based on the motive for the benchmarking (for example, to compare profitability, efficiency, etc). This type of benchmarking is often external, and does not require the participation of the benchmarked companies. Indeed, financial benchmarking is often performed by consulting companies, or business or industry-specific journals (such as *Pulp and Paper International*). Financial benchmarking can also be used by individual investors seeking to evaluate the actual financial performance or state of an investment object in comparison to competing investment opportunities.

An assessment of the creditworthiness of debt-issuing companies is based on the financial statements of the issuer and on expectations of future economic development using a combination of qualitative and quantitative analysis (Tan et al. 2002). Credit rating agencies (e.g. Moody's Investor Services, Standard & Poor Corp., FLIP) are commercial firms that receive payment for publishing an evaluation of the creditworthiness of their clients. Creditworthiness information is especially useful when borrowing takes place through the issue of securities, rather than by bank loans, since buyers of securities do not know the issuers as well as banks usually know their customers.

The idea of a society of software agents was introduced in Wang et al. (2002) for monitoring and detection of financial risk. In a society of software agents each agent carries out different functions autonomously. We use a multiagent approach for building our knowledge creating system.

There have been a number of attempts to use multiagent systems to support business processes and deal with business environment. Liu (1998) suggested a software agent approach in environmental scanning activities for senior managers. An agent system developed by PriceWaterhouseCoopers, called EdgarScan, scans the financial reports in the Securities and Exchange Commission's database (EDGAR). The agent works by scanning the document for tags that indicate certain financial data. The system also includes a basic graphical benchmarking system, which only allows the user to compare companies by one ratio or value at a time. The agent can be found at http://www.pwcglobal.com/gx/eng/ins-sol/online-sol/edgarscan. Nelson et al. (2000) have proposed an auditing system (FRAANK) based on an agent that retrieves financial information from the EDGAR database.

One popular data mining technique for quantitative data analysis is the *self-organizing map (SOM)* (Kohonen 1997). The SOM has been used for a variety of tasks relating to financial analysis, for example, credit analysis (Martín del-Brío and Serrano-Cinca 1993; Back et al. 1995; Serrano-Cinca 1996; Kiviluoto 1998; Tan et al. 2002), financial benchmarking (Back et al. 1998; Karlsson et al. 2001; Eklund et al. 2002), and macro level economic environment analysis (Kaski and Kohonen 1996). Tan et al. (2002) studied the rating process using Self-organizing maps for clustering and visualizing the financial ratios. Lavrenko et al.

(2000), Back et al. (2001), and Kloptchenko et al. (2002) have combined quantitative and qualitative financial data using quantitative and qualitative clustering techniques for knowledge discovery.

3. THE CONCEPTUAL MODEL

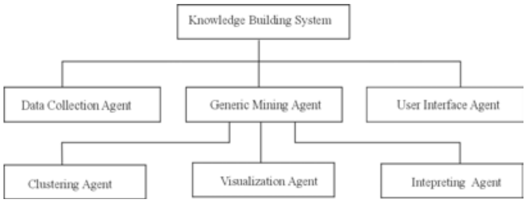

Figure 1: Architecture of the Knowledge Building System.

The proposed conceptual model of the knowledge-building system is depicted in Figure 1. It consists of six agents, i.e. *the Data Collection Agent, the Generic Mining Agent, the User Interface Agent, the Clustering Agent, Visualization Agent* and *the Interpreting Agent*. Each agent carries out its own functions and uses information provided by other agents connected to it. These agents handle three main activities (that are provided by three autonomous agents): data collection and storage (Data Collection Agent), searching for hidden patterns (Generic Mining Agent), and user-interface design (User Interface Agent).

The *Data Collection Agent* is intended to collect, assemble, and sort the quantitative and qualitative data from various Internet resources, such as Bloomberg, Reuters, Wall Street Journal, MSNBC, and individual companies' web sites. These data consist of, for example, market updates, quotes, financial reports, market reports, etc.

The *User Interface Agent* is intended to be responsible for providing the communication channel between the system and the human user that chooses the goal for the system. It should offer the choice of a number of tasks defined by the user in their setup of the system. For example, two possible applications are financial benchmarking and credit rating. These tasks are defined by the data (numeric and textual) included, as well as by the importance placed on each piece of data (for example, the importance of a particular financial ratio). In short, the agent should present the system options, receive the user input commands, and show the final results after it has interacted with the other agents.

The *Generic Mining Agent* is intended to include at least three activities in data processing (see Figure 1.): clustering of the data, visualization of the intermediary results of the previous process, and interpretation of the final results. The clustering techniques are instance dependent, in the sense that we can apply different clustering algorithms when performing data and text mining. We have three agents for the three distinct steps in data processing: the *Clustering Agent, the Visualization Agent*, and the *Interpretation Agent*.

Depending on what mining techniques and data are used, there are two main instances of the Generic Mining Agent: *Data Mining Agent* (Figure 2.) and *Text Mining Agent* (Figure 3.). We see the Generic Mining Agent as a generic class (in programming language understanding), which does not exist physically, but rather is an abstract class that is implemented via its instances. A distinction between the two instances of the Generic Mining Agent is based on the types of data they mine: Data Mining Agent (for processing numeric data) and Text Mining Agent (for processing text data).

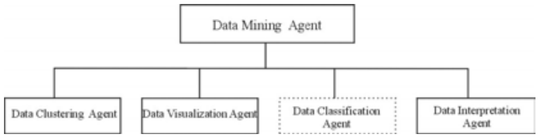

Figure 2: Data Mining Instance of the Generic Mining Agent.

In addition to the activities that are common for both the Data and Text Mining Agents, there are other activities that can be implemented, for example, constructing classification models in the case of the Data Mining Agent and information summarization for the Text Mining Agent. Two new agents can perform these two different activities: the *Data Classification Agent* (see Figure 2, dot-line rectangle) and the *Summarization Agent* (see Figure 3, dot-line rectangle).

The *Knowledge Building System* aims at creating new knowledge by consolidating the obtained new information from the Data Mining and Text Mining Agents. The Knowledge Building System will behave reactively to the goal of the system.

The *Data Mining Agent* would be responsible for numeric data processing and pattern discovery. The Data Mining Agent should provide the Knowledge Building System with the cluster that a company (or other data, depending upon the intended goal) belongs to, as well as the characteristics of the clusters (high profitability, low solvency, etc.), i.e. the results of the entire clustering. The *Data Clustering Agent* should calculate the chosen financial ratios for the chosen companies, standardize the data, and cluster them using self-organizing maps. Finally, the *Data Visualization Agent* visualizes the results.

After we visualize the map clusters provided by the Data Clustering Agent we could use the *Data Classification Agent* that creates a decision tree and/or a multinomial logistic regression model for classifying new financial data (Costea and Eklund 2003). The Data Classification Agent might also use other classifiers. Among these, the agent should use the model that achieves the highest accuracy in training and the best prediction performance.

Then, using all the information from the previous agents, combined with knowledge from other agents in the system, the *Data Interpretation Agent* would attempt to explain the findings. For example, in quantitative clustering, it is important to find explanations for a particular event, such as decreased profitability. This type of information can be found in the textual part of the annual report.

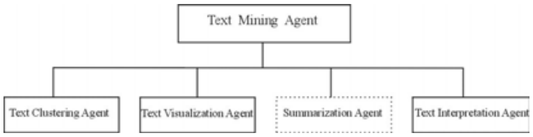

Figure 3: Text Mining Instance of the Generic Mining Agent.

The *Text Mining Agent* is intended to be responsible for processing textual information, and choosing the essential indications in it. It could use the *Summarization Agent* that deals with domain information, creating news summaries for any chosen company, or general market information for any chosen time period, and reports it to the user. Then, the *Text Clustering Agent* would perform financial statement clustering by using the prototype-matching methodology (Visa et al. 2002; Back et al. 2001) and reports which financial reports are close in meaning to each other. The *Text Visualization Agent* would present a visual U-matrix map with cluster representation and labels of the companies, which are clustered according to the similarity of their financial statements. The *Text Interpretation Agent* would have the same functionality as the Data Interpretation Agent, the difference being the type of data that is processed.

The Knowledge Building System combines the information from the Data and Text Mining Agents, i.e. it reports to the user how well the chosen company is performing in light of the chosen task, what level of performance the company displays in comparison with other companies in the analysis (clusters), and explains why (text summaries and clusters). The outputs of the two "instance" agents (Data and Text Mining Agents) can be validated one against the other, and the Knowledge Building System can do this automatically, alarming the user if the results are not convergent.

4. METHODS USED BY THE AGENTS

Our agents use several specific data mining techniques for clustering, visualization, and classification of quantitative and qualitative data.

We have used the SOM for clustering the quantitative data. The SOM is an unsupervised neural network for exploratory data analysis. The SOM takes multidimensional numeric data and clusters them on a two-dimensional topological map. Kiang and Kumar (2001) made a comparison between self-organizing maps and factor analysis and K-means clustering. The authors compared the tool's performances on simulated data, with known underlying factor and cluster structures. The results of the study indicate that self-organizing maps can be a robust alternative to traditional clustering methods.

Once trained SOM models are created, the problem of dealing with new data arises. Instead of time consuming retraining, a different method was proposed in Costea and Eklund (2003). The authors suggest a two-level methodology including initial clustering using SOM, and decision tree or multinomial logistic regression classification models trained on the original SOM model. This way the user is able to deal with new data without retraining maps. We have compared the two classification techniques in terms of their accuracy rates and class predictions and reached the conclusion that choosing among possible classifiers is problem dependent. We can extend the number of variables used for training the SOM maps, since the algorithm does not have restrictions from this point of view. Conversely, this methodology can be used as an alternative way of assessing the creditworthiness of companies as opposed to that provided by, say, Standard & Poor's (Tan et al. 2002).

We have tested the use of the prototype-matching approach for text clustering. This method is based on textual collection processing on word and sentence level processing (Visa et al. 2002; Toivonen et al. 2001). The prototype is a document, or a specific part of it, which is of interest to a particular user. A prototype is matched with an existing text collection to obtain a cluster of semantically similar documents. The methodology is based on text preprocessing, and word and sentence level text encoding and histogram creation.

The text summarization algorithm should extract the most relevant sentences from one or multiple documents with regard to a query. Therefore, we propose the use of a text clustering algorithm (e.g. prototype-matching or bisect k-means) for organizing one or more relevant documents into a tight cluster, and a feature extraction algorithm (e.g. occurrence of cue words, frequent words and proper nouns, position of the sentence with them in the text, sentence length, etc.) and classification algorithm (e.g. Naïve-Bayes classifier, C4.5) for extracting relevant sentences in the relevant documents. The combination of the mentioned techniques requires thorough study for successful summarization. We realize that straightforward word matching is not enough for effective detection of similarity between text pieces.

5. LIMITATIONS AND DIFFICULTIES

There are, of course, a number of problems associated with building a system of this complexity based on data that are freely presented on the Internet. We can divide system limitations in, at least, two categories: limitations that are specific for each individual agent and limitations regarding the integration of different agents. The data collection agent's ability to automatically retrieve financial data from Internet resources is severely hampered by a lack of standard for online financial reporting. A possible future solution to this problem is XBRL (eXtensible Business Reporting Language). XBRL is an XML (eXtensible Markup Language) standard created specifically to address the problem of online business reporting. Currently, there is no way for collection agents to automatically retrieve financial data from diverse web sites without specifically coding the agent for a specific page. (Debreceny and Gray 2001)

Another type of limitation of the system is due to the limitations of the deployed DM and TM techniques (Data and Text Mining Agents). For example, with all its advantages over standard clustering techniques, the SOM has one major drawback: verification of the achieved clustering results. This issue is addressed in Wang (2001), in which the author proposes a number of techniques for verifying clustering results. Similar techniques will have to be used in the system we are proposing.

Text mining techniques have a number of disadvantages due to the highly dimensional structure of text. Two textual pieces can often be nearest neighbors in terms of using similar vocabulary, without actually belonging to the same semantic class. Prototype-matching clustering is an exploratory technique that possesses some difficulties with determination of the clusters, and with their comparison with quantitative clustering. Although, theoretically, text implies richer information about an event than a numerical snapshot of the fact does, this is difficult to verify. Even having excellent text mining techniques on hand that could mine the indications of future financial performances of the company, those indications can be easily concealed by smart word choice and sentence construction.

Also, as was illustrated by the Enron and WorldCom scandals, the financial information presented in annual reports is not always reliable. Of course, if this incorrect information is inserted into our system, the results will also be incorrect. Moreover, there might be unintentional mistakes in the data. Therefore, some kind of error detection and handling capabilities should be built into the system. This is also required by the actual definition of KDD, which includes data cleaning and error detection (Fayyad et al. 1996).

The integration limitations are closely related to the individual agents limitations, e.g.: because of the lack of standard of financial information available on the Internet, the Data Collection Agent might not be able to provide the data that we need to address a specific problem, which makes its integration with the Knowledge Building System extremely difficult.

6. CONCLUSIONS AND FUTURE WORK

In the current research paper we introduced a conceptual model of a system based on different data/text mining methods for knowledge building from freely available data distributed on the web. The system aims to automatically perform different tasks such as data collection, financial benchmarking, assessing creditworthiness of companies, and finding hidden patterns in unordered and unstructured text data. The system uses two types of data (numeric and textual) and data processing techniques (data and text mining techniques) to support and explain the phenomena.

In this paper we discussed the operational facilities of the proposed system that will be accomplished by text and data mining methods. The system knowledge base, system external interface and limitations should be researched further.

As further research problems we could investigate new methods for collecting the input information for the Data and Text Mining Agents (that is improve the Data Collection Agent), extend the conceptual model to include subagents that perform tasks for their "parent" agents: Data Cleaning Agent, Data Aggregator Agent (aggregates information find on different web sources and presents this information further to Data Collection Agent).

REFERENCES:

Back, B., G. Oosterom, K. Sere, m. van Wezel, 1995. Intelligent Information Systems within Business:Bankruptcy Predictions Using Neural Networks. In *The 3rd European Conference on Information Systems (ECIS'95)*, Athens, Greece.

Back, B., K. Sere, H. Vanharanta, 1998. Managing complexity in large data bases using self-orginizing maps. In *Accounting Management and Information Technologies* 8(4): 191-210.

Back, B., J. Toivonen, H. Vanharanta, A. Visa, 2001. Comparing numerical data and text information from annual reports using self-orginizing maps. In *International Journal of Accounting Information Systems* 2: 249-269.

Costea, A. and T. Eklund, 2003. A Two-Level Approach to Making Class Predictions. In *The 36th Hawaii International Conference on Systems Sciences (HICSS-36)*, Hawaii, USA, IEEE.

Debreceny, R. and G. L. Gray, 2001. The production and use of semantically rich accounting reports on the Internet: XML and XBRL. In *International Journal of Accounting Information Systems* 2(1): 47-74.

Dorre, J., Gerstl, P., and R. Seiffert, 1999, Text Mining: Finding Nuggets in Mountains of Textual Data, In *Proceedings of th 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, California, USA

Eklund, T., B. Back, H. Vanharanta, A. Visa, 2002. Assessing the Feasibility of Self-Organizing Maps for Data Mining Financial Information. In *The Xth European Conference on Information Systems (ECIS 2002)*, Gdansk, Poland.

Fayyad, U., G. Piatetsky-Shapiro, P. Smythe, 1996. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *The Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, Portland, Oregon, AAAI Press.

Karlsson, J., B. Back, H. Vanharanta, A. Visa, 2001. *Financial Benchmarking of Telecommunications Companies*. TUCS Technical Report No. 395, Turku Centre for Computer Science. Turku.

Kaski, S. and T. Kohonen, 1996. Exploratory Data Analysis by the Self-Organizing Map: Structures of Welfare and Poverty in the World. In *The Third International Conference on Neural Networks in the Capital Markets*, World Scientific.

Kiang, M. and A. Kumar, 2001. An Evaluation of Self-Organizing Map Networks as a Robust Alternative to Factor Analysis in Data Mining Applications. In *Information Systems Research* 12(2): 34-41.

Kiviluoto, K., 1998. Predicting bankruptcies with the self-organizing map. In *Neurocomputing* 21(1-3): 191-201.

Kloptchenko A., T. Eklund., B. Back, J. Karlsson, H. Vanharanta, A. Visa, 2002. Combining Data and Text Mining Techniques for Analyzing Financial Reports. In *The 8th Americas Conference on Information Systems (AMCIS2002)*, Dallas, USA.

Kohonen, T., 1997. *Self-Organizing Maps*, Springer-Verlag. Leipzig, 2nd edition.

Lavrenko, V., M. Schmill, D. Lawrie, P. Ogilvie, 2000. Mining of Concurrent Text and Time Series. In *Text Mining Workshop of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, MA, USA. ACM.

Lesser, V., 1995. Multiagent Systems: An emerging subdiscipline of AI. In *ACM Computing Surveys* 27(3): 340-342.

Liu, S., 1998. Business Environment Scanner for Senior Managers: Towards Active Executive Support with Intelligent Agents. In *Expert Systems with Applications* 15: 111-121.

Martín-del-Brío, B. and. C. Serrano-Cinca, 1993. Self-organizing Neural Networks for the Analysis and Representation of Data: Some Financial Cases. In *Neural Computing and Applications* 1: 193-206.

Nelson, K. M., A. Kogan, R. P. Srivastava, M. A. Vasarhelyi, H. Lu, 2000. Virtual auditing agents: the EDGAR Agent challenge. In *Decision Support Systems* 28(3): 241-253.

Serrano-Cinca, C., 1996. Self organizing neural networks for financial diagnosis. In *Decision Support Systems* 17(3): 227-238.

Tan, R., J. den Berg, W. den Bergh, 2002. Credit Rating Classification Using Self-Organizing Maps. In *Neural Networks in Business: Techniques and Applications*, ed. by K. Smith and J. Gupta, Idea Group Publishing. Hershey.

Toivonen, J., A. Visa, H. Vanharanta, B. Back, 2001. Validation of Text Clustering Based on Document Contents. In *Machine Learning and Data Mining in Pattern Recognition (MLDM 2001)*, Leipzig, Germany. Springer-Verlag.

Visa, A., J. Toivonen, B. Back, H. Vanharanta, 2002. Contents Matching Defined by Prototypes: Methodology Verification with Books of the Bible. In *Journal of Management Information Systems* 18(4): 87-100.

Wang, S., 2001. Cluster Analysis Using a Validated Self-Organizing Method: Cases of Problem Identification. In *International Journal of Intelligent Systems in Accounting, Finance and Management* 10(2): 127-138.

Wang, H., J. Mylopoulos, S. Liao, 2002. Intelligent Agents and Financial Risk Monitoring Systems. In *Communications of the ACM* 45(3): 83-88.

Wuthrich, B., D. Permunetilleke, S. Leung, V. Cho, J. Zhang, W. Lam, 1998. Daily Prediction of Major Stock Indices from textual WWW data. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98),* New York, NY, USA, AAAI Press.