An e-Knowledge application for local government and small and medium businesses

M. Castellano[1], N. Pastore[2], F. Arcieri [2], V. Summo[2]
& G. Bellone de Grecis[2]
[1]*Dipartimento di Elettrotecnica ed Elettronica, Politecnico di Bari, Italy*
[2]*Global Value, ACG srl, an IBM and FIAT Company, Italy*

Abstract

This paper presents an e-Knowledge application able to support decision makers in strategic planning for the growth of a territory. The application integrates a decision support system and provides e-Knowledge services for the analysis of territorial dynamics by processing and modelling huge amounts of data through mining techniques, in order to discover rules and patterns in a distributed and heterogeneous content environment. The application is build on a technologic infrastructure of Business Integration, based on Service Oriented Architecture, and a virtual private network for the cooperation between Local Government and Small and Medium Business applications. Techniques of crawling and Web Mining are respectively used for the extraction and the analysis of unstructured and semi-structured data. For the analysis of structured data, the application covers the whole Knowledge Discovery in Database processes. Finally, a case study of Employment Analysis is presented.
Keywords: e-Knowledge, data mining, web mining, Local Government, Small & Medium Business, cooperative information systems.

1 Introduction

In recent years, Information and Communication Technologies have profoundly transformed the ways in which the society produces and exchanges information as well as the time required to do so. This innovation is taking on a primary role in the economic development of more advanced countries and is creating new social behaviors.



Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

Today, the analysis of huge amounts of data coming from geographically distributed sites can be problematic for both private and public organizations. The aim is to produce added value knowledge in order to orient and support strategic decision-making. This need for knowledge is becoming an important feature also in economic fields. In these sectors, decision makers are called on to analyze increasingly large banks of information in short time spans. For example, the public sector's knowledge about the availability and characteristics of its unemployed human resources may be of great value to the private sector. For the private sector, access to this information can be crucial in determining the viability of developing and addressing investment strategies in a given geographic area. Hence, the efforts to search for relationships among information coming from different sources are incremental and continuous since this process can produce new knowledge through the collection, organization and furnishing of information.

In this paper we propose the Knowledge Center, an e-Knowledge application that represents a flexible decision support tool for a strategic planning both for Local Government (LG) in the growth of the territory and for Small and Medium Business (SMB) in marketing policies, increasing the ability of the decision maker in planning more flexible and rapid resolutions, with the awareness of taking the most appropriate ones. The purpose of the paper is not to present new algorithms or other techniques, but to show how to implement them in a flexible mining architecture in order to provide added value e-Knowledge services. At first the state of art is described, then the Knowledge Center model is presented and lastly a case study and conclusion are reported.

2 Backgrounds

2.1 Mining techniques

A large variety of data and web mining techniques have been developed by researchers in the fields of artificial intelligence and statistics.

Data Mining and Knowledge Discovery in Databases: Knowledge Discovery in Databases is a nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from large collections of data [1]. Data Mining is only a single step of the process, mainly concerned with means by which patterns are extracted and enumerated from the data. However, both terms are concerned with extracting useful and new information from data, by combining fields of databases and data warehousing with algorithms from machine learning and methods from statistics to gain insight in hidden structures within the data. A generic method of knowledge discovery follows these tasks [3]:

- **Data selection:** data relevant to the analysis task are retrieved from the database.
- **Data pre-processing:** noise and inconsistent data are removed (data cleaning) and multiple data source are combined (data integration).
- **Data transformation:** data are transformed or consolidated into forms appropriate for mining summary or aggregation operations.



Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

- **Data mining:** in this essential process intelligent methods are applied in order to extract data patterns.
- **Pattern evaluation:** interesting patterns representing knowledge based on interestingness measures are identified.
- **Knowledge presentation:** the mined knowledge is presented to the user through visualization and knowledge representation techniques.

Web Mining: Recent research [12] focuses on utilizing the Web as a knowledge base for decision making. With the huge amount of information available in Internet, application of data mining techniques to the World Wide Web, referred to as Web mining, has been the focus of several recent research projects and papers. The main processes of knowledge discovery in data coming from the Web are the Web content mining and the Web usage mining [14]. The latter is the process of mining for user browsing and access patterns [8]. Depending on the location of the source, the type of collected data differs. This makes the techniques to be used for a particular task in web mining widely varying. Anyway, a generic method of knowledge discovery of data coming from the web follows these tasks [7]:

- **Information Retrieval:** It deals with automatic retrieval of all relevant documents, while at the same time ensuring that the non-relevant ones are fetched as few as possible. The IR process mainly includes document representation, indexing, and searching for documents.
- **Information Extraction:** Once the documents have been retrieved the challenge is to automatically extract knowledge and other required information without human interaction. IE has the task to identifying specific fragments of a single document that constitute its core semantic content.
- **Generalization:** Characterized by pattern recognition and machine learning techniques, which are usually used on the extracted information.
- **Analysis:** Once the patterns have been discovered, they need to be understood, visualized and interpreted.

2.2 Service oriented architecture

Service-oriented architecture (SOA) is a concept specifying that an application can be made up from a set of independent but cooperating subsystems or services. The SOA model isolates each service and exposes only the necessary declared interfaces to other services. In this way, as technology changes, services can be independently updated, limiting the impact of changes and updates to a manageable scope. Managing change is an important benefit of leveraging component architectures and models. Today there are many collaboration techniques but they vary from one case to another and are often owner solutions or systems that collaborate without any vision or architecture. A shift towards a *service-oriented* approach will not only standardize interaction, but also allows for more flexibility in the process. The complete value chain within an organization is divided into small modular functional units, or services. A service-oriented architecture [5,6] thus has to focus on how services are described and organized to support their dynamic, automated discovery and use.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

3 The Knowledge Center application

The purpose of the Knowledge Center application is generating and making available to decision makers information and e-Knowledge services to support them in the decision-making process.

The application works as an increasing and continuous process in which the existing knowledge comes collected, organized and so made available.

The real aim is to integrate heterogeneous sources of information, like databases, unstructured documentation, web-sites, capturing the context and giving it more value through the possible new relations with other information already present in the organizations. By integrating Data and Web Mining techniques, the framework is able to extract, to select, to process and to model huge amount of data in a distributed and heterogeneous content environment of informative resources, coming from both the World Wide Web [10] and the LGs and SMBs, in order to discover rules and patterns unknown a priori.

In fig. 1 we show the Knowledge Center model and all the actors with which it interacts:

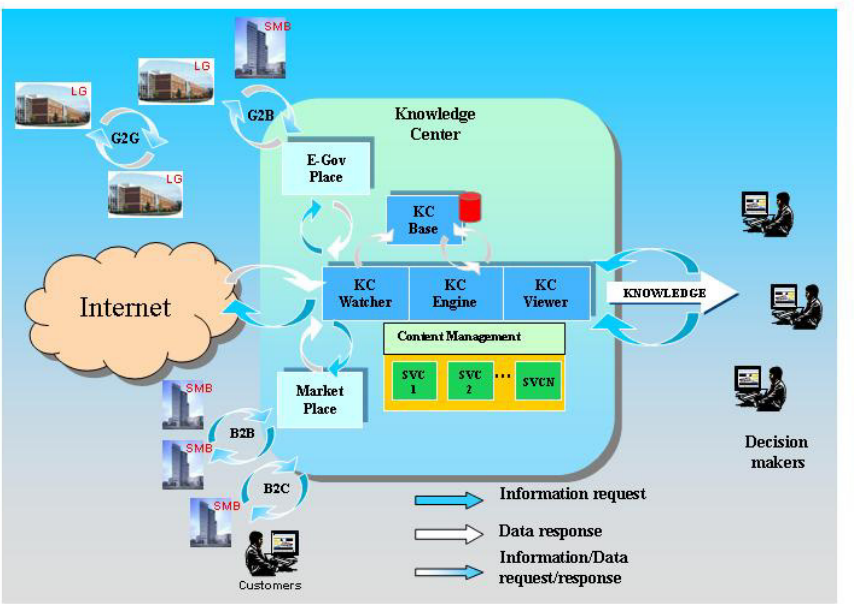


Figure 1: The Knowledge Center.

The main components have been developed with J2EE technology. They are:

KC-Watcher: This component has the task to extract structured and unstructured data coming from the World Wide Web, LGs and SMBs. It is able to discover information about the content of web pages and the navigation behaviour of the user by implementing Web Content Mining e Web Usage

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

Mining techniques [13]. To collect structured data from LGs and SMBs Information System it makes use of Web Services technology.

KC-Base: This component represents the knowledge repository of the system. It contains raw data coming from the Watcher and knowledge data elaborated by the Engine. For a standard representation of data, metadata and mining models the Base uses the Common Warehouse Metamodel (CWM) of the Object Management Group [15]. The CWM is designed to maximize the reuse of Object Model (a subset of UML) and the sharing of common modelling constructs where possible.

KC-Engine: This component is the core of the framework. It works on raw data stored in the Base and provides the e-Knowledge services by covering the whole process of the Knowledge Discovery in Databases (KDD) and in Text (KDT). Each e-Knowledge service realizes a choreography of complex Data and Web Mining operations built on libraries of well-known algorithms.

KC-Viewer: This component represents the single access point for Knowledge Sharing and has the task to show the knowledge depending on the characteristics of the decision maker. It provides the e-Knowledge services through human and machine interfaces by using Portlet and Web services technologies.

The Knowledge Center includes also e-Government and e-Business market places, which consist of unique access points, such as portals, respectively oriented to services for Local Governments and Small and Medium Business.

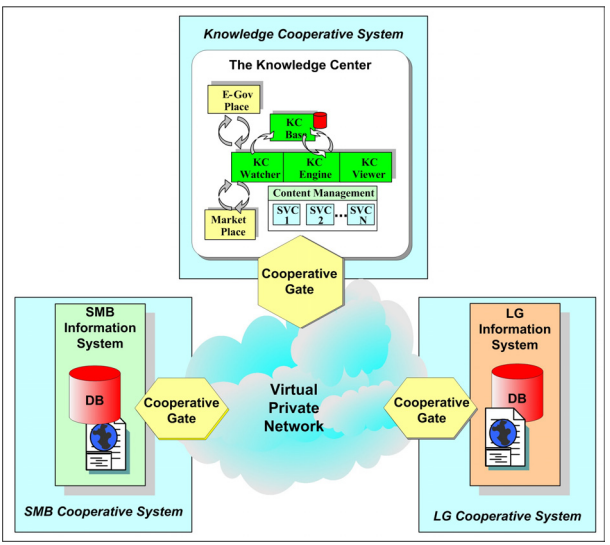


Figure 2: The Knowledge Center cooperative architecture.

3.1 The Knowledge Center cooperative architecture

To integrate sources of heterogeneous information, such as databases, documentation and web sites, the Knowledge Center implements an

© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

infrastructure of Business Integration based on Service Oriented Architecture [17] and built on a virtual private network constituted by LGs and SMBs that use "Cooperative Gates". These interfaces, based on Web services, transform LGs e SMBs legacy Information Systems in Cooperative Systems able to cooperate among them. A Cooperative Architecture's Information System consists of numerous components distributed over large, complex computer and communication networks that work together, requesting and sharing information, constraints, and goals [16]. Through the Cooperative Gates, the framework knows each change in each Information System. In fig. 2, the Knowledge Center Cooperative Architecture is showed.

3.2 The engine for mining

The engine is a distributed data and web-mining environment, where a set of services are managed and made available through a controller. The main components of the KC-Engine, showed in fig. 3, are the Controller, the Miners and the Kernel:

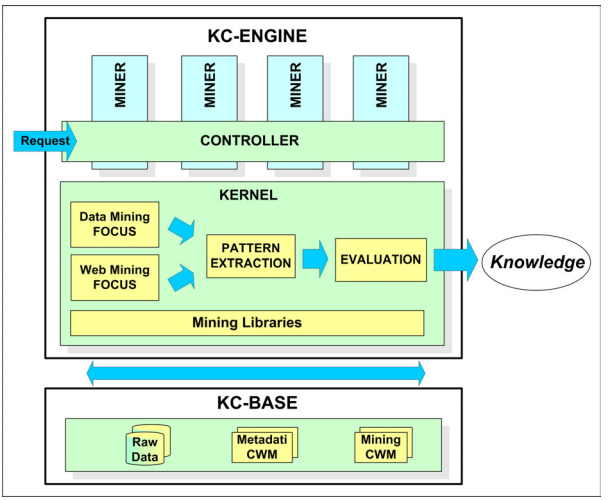


Figure 3: The KC-engine.

Controller. When a service request comes to the system, this is received by the Controller, analyzed and then forwarded to an action. Each action corresponds to an e-Knowledge service and it may call one or more Miners to provide the results.

Miners. The Miners are building blocks that can be used to build a complex application. In this case, the application is represented by an e-Knowledge service and one or more Miners that represent the business logic of the service may form it. A Miner can work in different ways:

- Load the mining models associated to a required e-Knowledge service.
- Call the KC-Watcher if it needs data from the Web for a specific service.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

- Activate a training process in the Kernel, by activating the process of the Knowledge Discovery in Databases or in Text [11] in order to rebuild mining models according to the presence of new data in the training sets.

Kernel. The Kernel follows the process of discovering knowledge from raw data, involving iterations of three main stages: data preparation, data and web mining, and results analysis. It is composed by two Focuses for Data Mining and Web Mining, a Pattern Extraction, an Evaluation and a set of Mining Libraries. The Focuses are useful for data preparation, in order to determine which data are necessary to provide the service. The Data Mining Focus holds the phases of selecting, pre-processing and transforming, while the Web Mining Focus holds the phase of the Information Extraction. The results of the Focuses are passed to a unique Pattern Extraction step, where a set of Web Mining techniques, collected together with other Data Mining algorithms into Mining Libraries, are used for the analysis. Finally, an Evaluation step interprets the utility and the carefulness of the extracted patterns and models [2], which will be stored in a CWM standard into the KC-Base.

3.3 The Crawlers for the web

In fig. 4, the component of the KC-Watcher is designed:

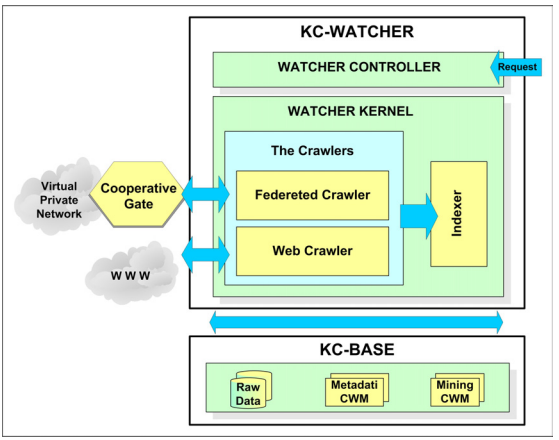


Figure 4: The KC-watcher.

A Controller and a Kernel form the component. The first receives from the KC-Engine requests of service that need web data. The second has the task to effectively crawl from the Web structured and unstructured data and to store them into the KC-Base. The Kernel includes two main components:

Federated Crawler: its scope is to acquire structured data both from e-Government and e-Business market places and from LGs and SMBs. This is made possible through the Cooperative Gate that connects the Knowledge Center to the virtual private network.

Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

Web Crawler: its task is to crawl unstructured information starting from a given set of URLs [4] submitted by the administrator of the system. Crawled data are then processed through Information Retrieval techniques to extract keywords that, suitably combined, represent the input for a search engine [9], such as Google, in order to get new URLs to crawl.

Indexer: its task is to index data crawled by the Web Crawler and the Federated Crawler. Then, it organizes them into clusters according to the extracted keywords.

4 A case study

The case study we introduce is an Employment Analysis service able to provide possible relationships between employment and commercial trends in a given job sector. LGs and SMBs are both highly sensitive to this type of service; LGs need knowledge tools to improve their planning of social and employment policies, while SMBs need to be able to improve their allocation of investments in the territory.

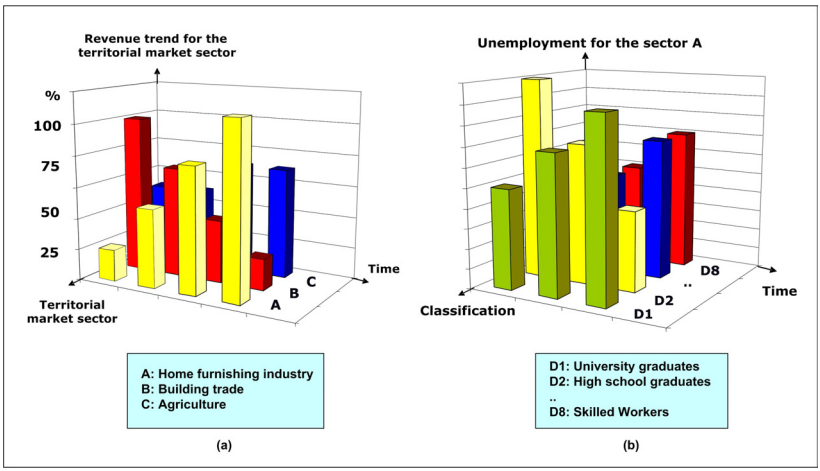


Figure 5: Experimental results.

The kinds of questions that decision makers try to answer are:

- In order to diversify, what sectors would be the best investment choices for an SMB in a given territory? What sector is growing?
- How can an LG help SMBs to invest in a territory?
- What are the characteristics of the available human resources (not employed) in a territory and how are the resources distributed?

In fig. 5, samples of experimental results are illustrated. The graphs show the revenue trend for the territorial market sector and the unemployment for a specific sector. We considered the sectors of Home furnishing industry, Building trade and Agriculture.

© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

These results were obtained through the analysis of about 500,000 database records, including curricula and electronic commercial transactions. They furnish the following responses to decision makers:

- The analysis in fig. 5a highlights that sector A is expanding and thus could be considered a target for new investments in the territory by SMBs.
- The analysis in fig. 5b illustrates the distribution of available human resources according to their specified characteristics. This information eased the ability of the LG to better interpret the needs of the territory and to plan the objectives of developmental policies.

5 Conclusion

This paper has presented a web-based decision support system able to provide added-value e-Knowledge services to Local Governments and Small and Medium Businesses. As remarkable results, the framework provides a single access point for Knowledge Sharing and is able to collect, integrate, elaborate and store data in a distributed and heterogeneous environment of informative resources. The system realizes a decoupling between the e-Knowledge services and the knowledge discovery processes and this is made possible by adopting the concept of reusable components as miners, which let's the creating of new services by using the same building blocks. Furthermore, to promote a clearer understanding of the contents and a more efficient sharing of knowledge, the use of the CWM standard has been considered.

Acknowledgements

The authors would like to thank Mario Logic and Vito Elvis for the important roles that they played during the development of this work.

References

[1] Fayyad, U.M., Piatetsky-Shapiro, G. & Smyth, P. & Uthurusamy, R., *Advances in Knowledge Discovery and Data Mining,* MIT Press, Cambridge, Massachusetts, 1996.

[2] Hand, D., Mannila, H. & Smyth, P., *Principles of Data Mining*, A Bradford Book, The MIT Press, 2001.

[3] Han, J. & Kamber, M., *Data Mining: Concepts and Technique,* Morgan Kaufmann Publishers, 2001.

[4] Chakrabarti, S., *Mining the Web*, Morgan Kaufman Publ., pp.19-21, 2003.

[5] Krishnaswamy, S., Zaslavsky, A. & Loke, S.W., An Architecture to Support Distributed Data mining Services in E-Commerce Environments. *Workshop on Adv. Issues of E-Commerce and Web/based Information Systems*, 2000.

[6] Chattratichat, J., Darlington, J., Guo, Y., Hedvall, S., Kohler, M., & Syed, J., An Architecture for Distributed Enterprise Data Mining. *7th*

© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9

International Conference on High Performance Computing and Networking Europe, 1999.

[7] Pal, S.K. & Mitra, P., Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions. *IEEE Transactions on Neural Networks*, Vol.13, No 5, 2002.

[8] Srivastava, J., Cooley, R., Deshpande, M. & Tan, P.N., Web Usage Mining: Discovery and applications of usage patterns from Web, *ACM SIGKDD*, 2000.

[9] Lee, C.H. & Yang, H.C., Developing an Adaptive Search Engine for E-Commerce Using a Web Mining Approach. *Proc. Int. Conf. Inform. Technol.: Coding and Computing*, pp. 604-608, 2001.

[10] Kumar, R., Raghavan, P., Rajagopalan, S. & Tomkins, A., Extracting large-scale bases from the Web, *Proceedings of the 25th VLDB Conference,* Edinburgh, Scotland, 1999.

[11] Nasukawa T. & Nagano, T., Text Analysis and Knowledge Mining System. *IBM Systems Journal*, Vol. 40, No. 4, 2001.

[12] Madria, S.K., Bhowmick, S.S., Ng, W.K. & Lim, E.P., Research issues in web data mining. *Proceeding of Data Warehousing and Knowledge Discovery, First International Conference,* DaWaK '99, pp.303-312, 1999.

[13] Cooley, R, Mobasher, B. & Srivastava, J., Web mining: information and pattern discovery on the World Wide Web. *Proceeding of IEEE International Conference Tools with AI.* Newport Beach, California, pp. 558-567, 1997.

[14] Mobascher, B., Dai, H., Luo, T., Sun, Y. & Zhu, J., Combining Web usage and content mining for more effective personalization. *Proceedings of the International Conference on E-Commerce and Web Technologies.* UK, 2000.

[15] Object Management Group, *Common Warehouse Metamodel (CWM) Specification.* OGM Press, Needham, Massachusetts, 2001.

[16] Mylopoulos, J. & Papazoglou, M., Cooperative Information Systems. *IEEE Expert Intelligent Systems & Their Applications*, Vol. 12, No. 5, 1997.

[17] Channabasavaiah, K., Holley, K. & Tuggle, E.M. Jr., Migrating to a service-oriented architecture, IBM, 2003.



Data Mining V, A. Zanasi, N. F. F. Ebecken & C. A. Brebbia (Editors)
© 2004 WIT Press, www.witpress.com, ISBN 1-85312-729-9