Managing Knowledge for Business Analysts: The Executive Information Portal[1]

John Mylopoulos, Attila Barta, Raoul Jarvis,
Patricia Rodriguez-Gianolli, and Shun Zhou[2]

Strategic business analysts keep track of trends that are relevant to their organization and its strategic objectives. To accomplish their mission, they monitor news stories and other reports as they become available, looking for evidence that these objectives remain on track, or have encountered obstacles. The paper presents a prototype enterprise information portal intended to support this type of knowledge work. The system supports three key functions. Firstly, it offers tools for building and analyzing semantic models of strategic objectives, the relationships among them, as well as events and actors who can play a role, positive or negative, in their fulfillment. Secondly, the system uses a powerful query language and the semantic model to assist analysts as they search external sources for relevant material, also provides semi-automatic classification and clustering of documents. Thirdly, documents are placed in an XML format and stored in an XML server. In addition, analysts can annotate or summarize documents and relate them to nodes of the semantic model.

Overview

An important ingredient of the Information Revolution is that individuals and organizations alike have access to orders of magnitude more information than ever before. Moreover, most of this information is available in computer-based forms, including electronic documents, databases and websites. The objective of our research is to develop technologies which help a group of knowledge workers search, find, retrieve, and organize information that is useful to their daily work.

---

[1] An extended description of this work can be found at http://www.cs.toronto.edu/~jm/exip.ps.

[2] Authors' address: Department of Computer Science, University of Toronto, Toronto, CANADA; {jm,atibarta,prg,jarvis,szhou}@cs.utoronto.ca.

These technologies include data retrieval and analysis for structured data (i.e., database tuples and/or records), semi-structured data (such as hypertext and source code), and unstructured data (such as documents, digitized photos). In general, such technologies are bundled into a solution through an integration architecture often called an *enterprise information portal* (EIP). EIPs are "*...applications that enable companies to unlock internally and externally stored information, and provide users a single gateway to personalized information needed to make informed business decisions"* [Shilakes98]. EIPs are not off-the-shelf software. Rather, they are integrated solutions built out of components. The majority of EIP providers develop their EIP solutions from product suits offered by several cooperating partner companies. Typically, EIPs offer facilities for search, retrieval, analysis and organization of structured data and documents, along with facilities for network connectivity and computing platform interoperation.

Our approach to enterprise information portals is founded on the use of a semantic model that captures knowledge shared by a group of collaborating knowledge workers. This model is used to drive information retrieval, classification, and analysis. For example, consider a group of software engineers who own a legacy software system. An EIP solution for this group would include a semantic model which represents concepts such as (software) global architectures, call graph structures, module inter-dependencies, versions, maintenance histories, and more. Reverse engineering tools can then extract information from the legacy system and associate it with the relevant components of the semantic model. When a software engineer is working on a particular task, say fixing a bug, she can rely on the EIP (and the semantic model) to find relevant information, such as global variable declarations, who calls the procedure that is being debugged and more. This is precisely the approach explored in [Finnigan97].

We are currently developing a prototype toolset, called the Executive Information Portal (or EXIP), intended to help a group of strategic business analysts working for a large corporation. Their task is to keep track of current events as they unfold, and make sure that their company's strategic objectives remain on track. To work on this task, business analysts scan news stories (Reuter's, CNN, etc.), analysts' reports (Yankee Group, Forrester Research, and the like) and other document sources, looking for relevant material. Once they have decided that a particular document is useful, they add it to their own library, write annotations and prepare memos to be circulated to their colleagues. This work is currently done without any tool support, or vanilla computer tools (e.g., a web browser and search engine). Our toolset aims to support the semi-automatic search and classification of documents, also the analysis of collected information with respect to a given set of strategic objectives.

The development of our toolset was guided by a number of quality requirements. Firstly, we wanted a system that evolves semi-automatically in the sense that its model and classification scheme may be modified by users, or automatically, thanks to the application of Machine Learning techniques. In addition, the classification of downloaded information is also assumed to evolve, along with the model and the classification scheme. Finally, we expect that our system will be used with minimal feedback from users, who are busy people and can't afford to spend much time training the system's classifiers and other functions.

System Architecture

The EXIP global architecture is shown in figure 1. The outermost layer of the architecture (bottom part of the figure) includes external information sources, such as CNNfn (cnnfn.cnn.com/news/technology/), CBC business news (cbc.ca/business/), the Globe and Mail (globeandmail.com/hubs/rob.html) and Forrester Research, whose reports we assume that the analysts download manually. This layer also includes wrappers for each source, which specify expected outputs for input queries. The information sources may have formats that are structured or semi-structured, proprietary or public domain (e.g., plain text.) In the prototype implementation, we wrap semi-structured (HTML) sources and plain text sources only. Structured sources (such as relational data) are easy to wrap and access, while documents in proprietary formats (PDF, MS-Word, RTF, etc.) can be translated to a common HTML or XML format using commercial tools such as Document Navigator from Verity [Verity].

It is important to note that documents transformed to HTML from a proprietary format are different from "native" HTML documents in that the former contain only useful information, while the latter contain an abundance of noise (advertising, site navigational menus, etc.) It is the job of the wrapper to locate the information of interest in a page for a given query, and separate it from such noise.

Another significant difference between the two types of documents is that those in a proprietary format are very likely to be local, easy to access files. In contrast, "native" HTML documents are often buried deep in a website and require considerable navigation in order to be retrieved. To specify the navigational path that will lead to a document to be retrieved, we use WebOQL, a second-generation web query language [Arocena98].

Wrappers export data in XML format. One important reason for selecting XML over HTML is that XML is fast becoming the lingua franca for information exchange, and its adoption allows us to benefit from a wealth of research in this area. Moreover, XML tags carry more useful information than HTML ones. This information can be used to better index and otherwise process retrieved documents. All exported data are translated into a common document schema, essential for the processing of documents after they have been downloaded.

The Document Clustering and Summarization component attempts to classify all downloaded documents by relating them to one or more elements of the semantic model. The proposed classification may be approved or overruled by the users of the system, or it may be accepted "as is" when the system is in "automatic" mode.

The semantic model provides a description of the strategic objectives of an organization in terms of goals and subgoals, also the events and actors that can influence any of these goals (positively or negatively). The model can be thought as a network of relationships between goals, actors, events and documents. Thanks to the rich modeling framework, the model supports various forms of analysis. For example, the analyst can visualize if and how the organization is advancing towards achieving its goals, what are the obstacles, critical events to look out for, and what are the dependencies to external organizations.

The shared knowledge of a group of analysts, represented by a semantic model, can be quite large and complex. Moreover, users of the EXIP (that is, business analysts and executives) are usually concerned with strategic planning from different perspectives. To aid the understanding and acquisition of knowledge at different levels of granularity, the user interface of EXIP supports a variety of visualization techniques for browsing/exploring the model as well as the contents of the EXIP.

The Document Management component, DocMan, provides support for document viewing, annotation and manipulation. DocMan manages collections of thousands of documents that have multiple links among them. Furthermore, DocMan supports complex search operations on these documents, involving both full-text and metadata search. Because the documents the their components (annotations, sumaries, links) are in an XML format, DocMan uses an XML Data Server for its operations.

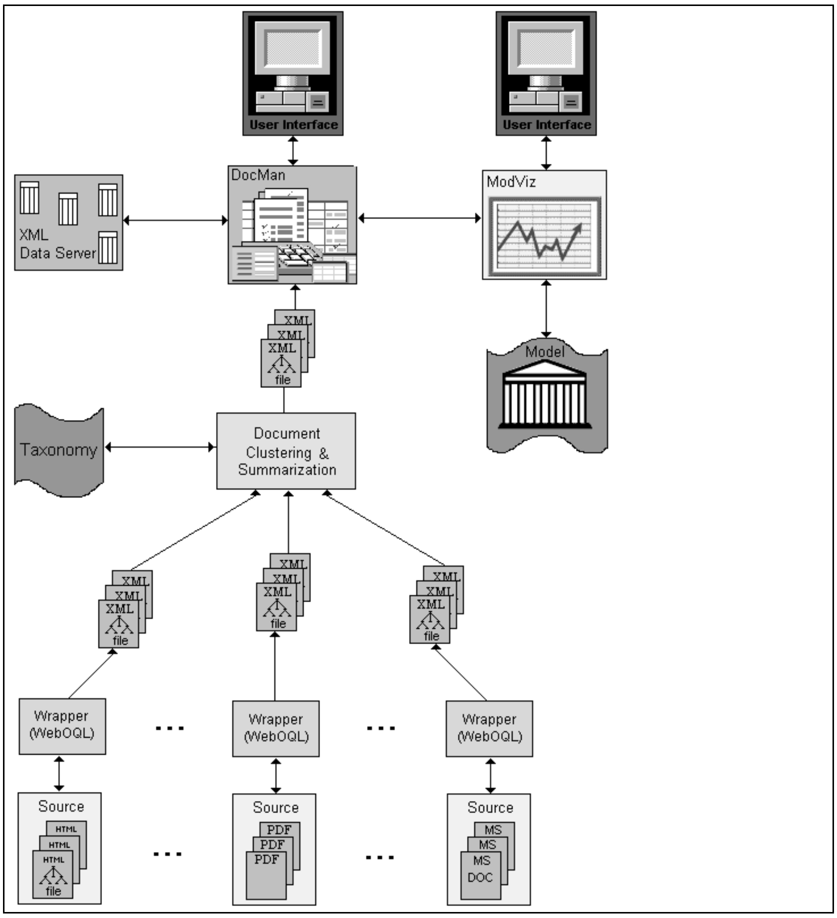

Figure 1: The global architecture of the EXIP.

Conclusions

EIP is intended to support a group of knowledge workers in retrieving, classifying and using information downloaded from a variety of information sources. Our approach differs from state-of-practice EIPs in that it treats all information managed by the system as semi-structured data, thereby exploiting tools such as declarative query languages and XML data servers. Moreover, our approach adopts a Machine Learning framework for quick learning and continuous evolution of its classifiers and clustering algorithms. Finally, our framework supports lightweight semantic models of relevant domain knowledge which is used for classification, retrieval and analysis.

We are currently completing the implementation and evaluation of the prototype. In addition, we are working on the definition of goal analysis techniques, such as evaluating the satisfiability of top-level goals, given a set of event instances about the application domain. Another form of analysis under review involves the identification of critical goals and events. These are goals or events which have become key to the fulfillment of top-level goals and about which additional information is needed.

cknowledgements

The research reported in this paper has been funded by the Bell University Laboratories (BUL), the Natural Sciences and Engineering Research Council (NSERC) of Canada, and the University of Toronto. We are grateful to Geoff Riggs (Bell Canada) for introducing us to the world of strategic business analysis, also to Michael Milton and Adele Newton (BUL) for helpful direction and moral support.

References

[Arocena98] Arocena, G., Mendelzon, A., "WebOQL: Restructuring Documents, Databases and Webs", Proceedings ", Inernational Conference on Data Engineering (ICDE'98), Florida, 1998.

[Barta98] Barta, A., *WebCat Information Integration for Web Catalogs*, Master's Thesis, Dept. of Computer Science, University of Toronto, 1998.

[Finnigan97] P. Finnigan, I. Kalas, K. Kontogiannis, H. Müller, J. Mylopoulos, S. Perelgut, M. Stanley, K. Wong, "The Software Bookshelf", *IBM Systems Journal,* November 1997.

[Goldman97] Goldman, R., Widom, J., "Dataguides: Enabling Query Formulation and Optimization in Semistructured Databases", Proceedings of the 23th International Conference on Very Large Databases, 1997.

[Iwayama95] Iwayama, M., Tokunaga, T., "Cluster-Based Text Categorization: A Comparison of Category Search Strategies," ACM International SIGIR Conference on Research and Development in Information Retrieval, 273-281, 1995.

[Manolescu00] Manolescu, I., Florescu, D., Kossmann, D., Xhumari, F., Olteanu, D., "Agora: Living with XML and Relational", Proceedings of the 26th International Conference on Very Large Databases, 2000.

[Masand92] Masand, B., Linoff, G., and Waltz, D., "*Classifying News Stories Using Memory Based Reasoning*," In Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 59-64, 1992.

[McCallum98] McCallum, A., and Nigam, K., "A Comparison of Event Models for Naive Bayes Text Classification," AAAI Workshop on Learning for Text Categorization, 1998.

[Mylopoulos98] Mylopoulos, J., "Information Modeling in the Time of the Revolution", *Information Systems 23(3-4)*, June 1998, 127-156.

[Mylopoulos01] Mylopoulos, J., Chung, L., Liao, S. S. Y., Wang, H., Yu, E., "Extending Requirements Analysis to Explore Alternatives", *IEEE Software 18(1),* January/February 2001, 92-96.

[Sahuquet00] Sahuquet, A., "Kweelt, The Making Of: The Mistakes Made and The Lessons Learned", Technical Report Department of Information and Computer Science, University of Pennsylvania, 2000.

[Schmidt00] Schmidt, A., Kersten, M., Windhouwer, M., Wass, F., "Efficient Relational Storage and Retrieval of XML Documents", Proceedings of the Third International Workshop on the Web and Databases, 2000.

[Shilakes98] Shilakes, C., and Tylman, J., "Enterprise Information Portals," Merrill Lynch, November 16, 1998.

[Verity] http://www.verity.com/products/docnav/index.html.

[Yang99] Yang, Y., and Liu, X., "A Re-Examination of Text Categorization Methods," Proceedings ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, 42--49.

[Yu93] Yu, E., "Modeling Organizations for Information Systems Requirements Engineering," Proceedings First IEEE International Symposium on Requirements Engineering, San Jose, January 1993, 34-41.

[Yu96] Yu, E., and Mylopoulos, J., "Using Goals, Rules, and Methods to Support Reasoning in Business Process Reengineering", *International Journal of Intelligent Systems in Accounting, Finance and Management 5(1),* January 1996.