Building an Information and Knowledge Fusion System

Tamás Mészáros, Zsolt Barczikay, Ferenc Bodon,
Tadeusz P. Dobrowiecki, György Strausz,

Department of Measurement and Information Systems,
Budapest University of Technology and Economics (BUTE),
Műgyetem rkp. 9., Budapest, Hungary, H-1521
{meszaros, barczy, bodon, dobrowiecki, strausz}@mit.bme.hu

Abstract. In this paper authors present a system for information and knowledge fusion that provides an integrated management of information in a particular task domain. The proposed system uses structured (database and XML-based) and unstructured (information retrieval) data acquisition techniques, various knowledge representation schemes to integrate retrieved information, and customisable reports generated based on profiles supplied by the end user. This paper concentrates on modelling the problem domain, information and knowledge fusion methods, and technology fields used in the proposed system.

Introduction

Efficient knowledge retrieval is being investigated within the framework of the Information and Knowledge Fusion EUREKA Applied Research Project[1]. Our main goal is to analyse, design and implement a new Intelligent Knowledge Warehousing environment, which would allow advanced knowledge management in various application domains (e.g. banking, revenue service, insurance, legal information, training & education, health care, etc.) [1].

Conventional information sources in such application areas provide only limited and frequently unreliable information. It is also hard to evaluate such data in an optimal way. Despite the fact that nowadays quite a variety of information sources provide data for applications in these domains, the human evaluation is still needed to select the relevant information. This method has limited depth, scope and reliability. Only very few sources are considered for gathering information and the human resources are in all respects limited.

Successful applications of automatic extraction of information from data sources for the computer-based or human-based decision support show that even without considering the meaning of the stored data, useful information, e.g. trends or anomalies, can be determined. *Data mining* deals just with such extraction of implicit, previously unknown, and potentially useful information and with using it for crucial business decisions [2]. It uses machine learning, statistics and visualisation techniques for knowledge discovery and presentation in a form that is easily comprehensible to

[1] This research was also supported by the Hungarian NSF (OTKA) Grant Nr. F-030763 .

the humans. Data mining deals primarily with structured and well-defined data sources, especially with large relational databases.

Information retrieval means collecting information from unstructured text documents (like books, papers, and electronic documents) [3]. In the 90s the easiness experienced in the Internet publishing resulted in a greatly increased volume of documents stored in global computer networks. To find the right piece of information in this rapidly growing and unstructured data storage, new kinds of systems were developed, like e.g. Internet catalogues and search systems. Automatic indexing systems attempt to cope with vast amount of data and to build structured index schemes. Internet search engines use them then to find the required information.

In the today's information retrieval and data mining the meaning of the information, which is looked for, is not used. Any related knowledge, however, can vastly increase the effectiveness of the retrieval process.

The project formulated a number of objectives, namely:

(1) to determine suitable methods for describing semantic information, that can be used to enhance the retrieval efficiency in information retrieval;

(2) to evaluate the state of the art of knowledge representation and reasoning methods, and to integrate these techniques into the retrieval process; and

(3) to build knowledge-based information retrieval system.

Such system can be used as an electronic alternative to the conventional data collection and analysis tools. It can provide new services, greater depth of the analysis, and higher reliability in the data acquisition process. In particular it can **continuously monitor information sources**, and automatically update the related knowledge, and it can **select the relevant data and provide reliability information** as well**.**

Modelling the Problem Domain

At an abstracted level all of the applications mentioned before can be modelled as a formation and a purposeful interaction of three *information environments*. *Target Environment* is a fragment of the real world, where the targeted (monitored) objects (corporations, bank clients, business processes, etc.) do exist. *Information Cumulating Environment* comprises all forms and media, which cumulate information about the targets. In our case it is Internet, various Intranet resources, corporation databases, published resources, financial experts, etc. *Information Utilising Environment* represents the users of information (e.g. the staff of a bank), at various level of management. In the following, when convenient, we will address these three environments as the "Client", the "Web", and the "Bank" respectively, according to the financial application example.

Target Environment (TE, "Client") is the physical source of knowledge. It comprises objects, phenomena, relations, etc., whose particular properties (parameters, relations, and state variables) are of interest for both Information Cumulating ("Web") and Information Utilising Environments ("Bank"). Target objects are usually interrelated, i.e. part of the knowledge is common to whole subsets or structures of domain objects.

Information Cumulating Environment (ICE, "Web") is coupled to the Target Environment by knowledge acquisition process of various sorts. Knowledge embedded in the "Web" is thus to an extent a veritable model of the "Client". It is however important to note that such knowledge is heavily distributed within the "Web" and that the application (i.e. "Bank") has no control over the whole extent and timeliness of the acquisition process. Consequently the "Bank" has no control over what knowledge is stored and where.

Information Utilising Environment (IUE, "Bank") represents active and possibly intelligent entities that require particular knowledge about the objects from within the "Client" to achieve their specific goals within the "Bank". An interesting feature adding to the complexity is that the "Web" and the "Bank" can be related, even overlapping. This will be true if e.g. the "Web" involves knowledge cumulated in human (expert) resources.

Knowledge Retrieval and Fusion

Our primary problem to face is that the knowledge acquisition between "Client" and "Web", and its storage in the "Web" are not specific to the goals found in the "Bank". The character of the knowledge acquisition reflects rather the physical and technological possibilities of the "Web" or some unrelated goals pending within that environment. Consequently, the form, the quality, and the validity of information are questionable from the users' point of view.

Specific goals require more condensed information than that available in the "Web's" sources. On the other hand when the user goals vary, so does the character of the requested information, even about the same "Client's" objects. In consequence the "Web" and the "Bank" must be interfaced by a mediating system embedded in both, which can accept as input the "Bank's" goals and which to provide answers should be familiar with the "Web's" possibilities and limitations, see Fig. 1.

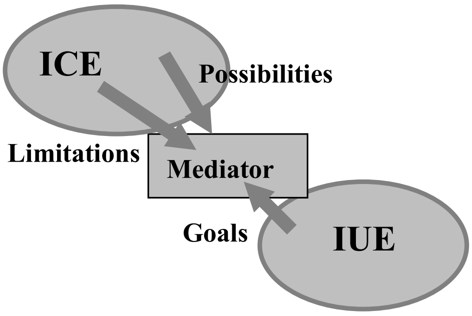

Fig. 1. Mediating system

The components of the required mediating system follow logically from its function. Such system must help to find the information essential for the "Bank's" goals, however, for typically recurrent and related goals producing the information by the repeated request to the "Web" would mean an unjustified and badly organised spending of system resources. Consequently a **Knowledge Repository (KR)** is needed, which reflects a portion of the "Client" and would puffer the information to avoid spending too much of the system resources on extracting from the "Web" knowledge needed for recurrent and related retrieval requests.

This repository must be up-dated with intensity and agility defined by the "hunger" for the information in the "Bank" and with the information accumulated in the meantime in the "Web". The question is therefore where from and in what form seek the information? To handle this problem an **Information Retrieval Subsystem (IRS)** is required, which will perform knowledge sensitive and knowledge intensive information retrieval. The IRS should know models of the "Client" (essentially the structure of the KR), should know the model of the "Web", and should perform search for information.

Similarly to the IRS a **Knowledge Fusion Subsystem (KFS)** interacts with the KR and the "Bank". The Repository stores more information than the particular needs of the particular users in order to serve the diversified goals of the whole "Bank". Therefore, the KFS component must be told what to fetch in the KR and in what form to present it to the ordering entity in the "Bank".

Managing the System

In the following let us shortly review various support functions provided by the system. From the design point of view the "Bank" is composed from end-users and managers. End-users ask simple queries and accept data. Managers, in addition, shape the overall flow of information.

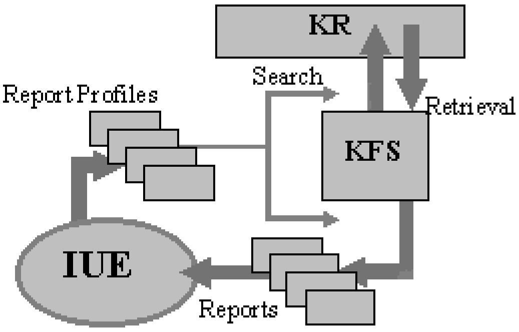

Fig. 2. Information flow related to the Knowledge Fusion Subsystem.

The information flow at the end-user level (i.e. the level of the individual goals in the "Bank") is governed by the Report Profile - a concise representation of the user's goals - stating which objects, parameters, etc. should be fetched from the KR and transformed into actual personalised reports (Fig. 2). The KFS finds the required information (continuously up-dated by the IRS), and reports it to the users in the

"Bank" with automatically generated periodic and one-shoot summaries. The KFS works with a number of report profiles.

Similarly to the user defined Report Profiles managers must define suitable Search Profiles describing what, how, and where to search. Search Profiles modify the searching strategy within the "Web" and the up-dating strategy within the KR (Fig. 3). Managers can also add Report Profiles to define reports to be distributed all over the "Bank" (e.g. corporation news, daily or hourly exchange rates, etc.).

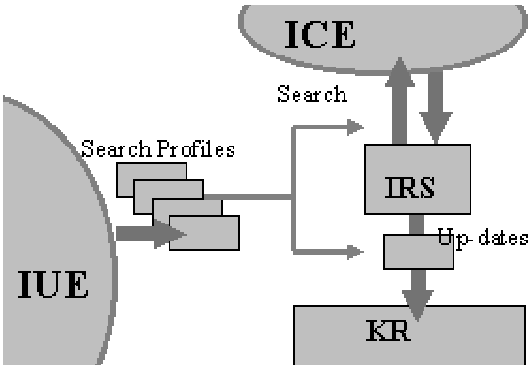

Fig. 3. Information flow related to the Information Retrieval Subsystem

The key part of the system is the internal **Knowledge Base (KB)** that models the application area and in particular the application task. The model building process is difficult and requires expertise in both the internal structure of the IKF system and in the specific application field. Therefore only managers can use its development and management services. These tools are hidden at the user level.

Main tasks of the management services are the followings:

(1) *Design and modification of the Knowledge Base.* The Knowledge Base contains general and application specific components. The main purpose of this decomposition is to allow the manager to design the domain specific information.

(2) *Defining and modifying search and evaluation strategies*. A set of information search and processing functions is available in the system. Managers can select methods appropriate to the tasks and can also modify them by setting the parameters of the methods. This customisation includes the specification of data sources, and retrieval and processing methods as well. The reliability of these sources can also be described.

(3) *System performance analysis*. In order to be able to accomplish the above mentioned functions managers should be able to analyse the behaviour of the system. A wide variety of analysing functions might be necessary to accomplish the general utilisation on the system.

(4) *Monitoring, and supervising*. Monitoring and supervising is necessary in every complex system, but this task is especially critical for systems with adaptive nature.

(5) *Input form specification*. One of the important tasks of manager is to create

application specific input forms for the user. Designing input forms means not only generating fields to fill-in, but also describing the relations between the fields and the internal representation.

(6) *Output report structure definition*. Manager can determine what part of the information should be recallable from the Knowledge Base. Manager can also define what part of the Knowledge Base is reachable for the end-users and what part should remain hidden for them. Pre-defined report forms can also be created and customised. The user can select these forms during the report creation.

User level services provide the interface for the end-users to define their requests and to read the results. The basic strategy applied here is that the end-user has only limited access to the system services. End-users can only reach information relevant to their working topics and are offered basically three types of services:

(1) *Generating queries using pre-defined input forms*. End-users can input information into the Knowledge Base and also pose queries to the system through pre-defined forms. End-users have limited possibility to configure the forms from the list of entities defined earlier by the manager.

(2) *Generating output reports by using and customising pre-defined report forms*. Users have similar possibilities for defining output report forms like in case of the input forms. This way end-users can produce dense reports on a given topic.

(3) *Asking for the information about the results of a query*. An important service is to provide explanation about the output data and also how the output data was generated.

Technologies in the IKF system

It is obvious that this system requires a wide variety of information acquisition, storage and access tools.

The "Web" contains a very diverse set of information repositories starting from relation databases, Internet resources up to the interfaces to various human data resources. In order to gather all these data the IRS should be equipped with several kinds of properly customised retrieval methods. We can categorise them into two main groups:

(1) *Structured retrieval*. **Data retrieval (DR)** from structured and well-defined data sources (like databases). It can also involve an automated **data mining (DM)** approach to exhaustively explore and determine complex relationships in very large databases.

(2) *Retrieving unstructured data*. **Information retrieval (IR)** deals with the retrieval process from unstructured data sources, e.g. the Internet resources.

We expect that many methods developed in data mining and information retrieval will be applicable to solve the knowledge acquisition problem.

Data Mining

Data mining investigates knowledge discovery techniques that obtain predefined structured knowledge solely from huge databases. Some data mining problems (like e.g. clustering [4] [5]) had already been known for a long time, and had been studied by many researchers in other fields of science (like statistics, machine learning, and visualization). Promising, possibly optimal, results were obtained (like K-L transform for dimensionality reducing), the size of databases, however, made it impossible to implement them in practice. The newly posed requirement, i.e. that the time and the memory demands of any algorithm in this field have to be linear in the size of the database, initiated further research.

There are several data mining approaches and application areas. *Association rules* try to find association between set of items in a given database of e.g. sales transactions (or market baskets), where each transaction contains some product [4] [6]. *Clustering* finds classes of closely related (similar) objects within a database of pattern vectors using a distance (similarity) function [5]. *Sequence matching* searches a database of event sequences, and it tries to find the closest to a query sequence [7]. In *Episode finding* a database about a long sequence of events with timestamps is given, and the task is to find those serial episodes that are present in at least a certain fraction of all windows [8]. *Classification* uses a database of training records with class labels for each, and it tries to build a concise model (generally a decision tree) to decide the classification for future, unlabeled records [9] [10]. *Web mining* is a data retrieval form in a linked environment that considers similar page/plagiarism finding, sophisticated query answering, page ranking etc [11].

Information Retrieval

Information retrieval is becoming a more and more important research topic as the publicly available information sources grow very fast. The World Wide Web, as an example, is a fast growing source of information. It introduced a set of technologies that allow the easy publication of electronic documents. However, its document formats, storage and access scheme do not help in finding the relevant information the user is looking for.

Information retrieval (IR), born as part of the library science in the 50's, deals with systems for indexing, searching, and recalling data, particularly text and other unstructured forms. With the growing amount of electronic data sources, IR received wide support from research organizations as well as from commercial companies in the 90's. Web indexing and search sites (like Altavista, Google, or Yahoo), text indexing and search options in databases (like Intelligent Miner for Text in DB2 from IBM, or Oracle interMedia), and client-based text indexing software (e.g. Autonomy's Kenjin) are examples of this.

Despite this growing activity in developing new IR methods and tools, the performance of these methods and tools is still far behind the similar tools for structured data retrieval. Web search engines overload their users with a waste amount of search hits (or they do not return any result at all). The basic problem behind this poor performance is that most systems are based on statistical indexing methods, and these methods cannot provide relevant results without understanding the meaning of the query and texts. Other systems based on natural language processing methods (where the aim is to "understand" the text and the query) have even worse performance due to missing proper techniques for automatically building semantic models from the text.

The Role of XML

The Extensible Markup Language (XML, standardized by the World Wide Web Consortium, W3C) defines text-based document in a structured way [12]. After successful attempts to standardise on the hardware and software, XML is an attempt to standardise on the information format, to make efficient search, storage and retrieval possible. XML is a so called mark-up language, i.e. a set of mark-up conventions to encode the meaning of the text. XML is extensible and various concrete languages can be created for particular domains.

XML is also a basis for many internationally accepted recommendations for electronic data exchange protocols. From the IKF point of view the most interesting developments are: XML-Based Ontology Exchange Language [13], Financial Products Mark-up Language [14], Artificial Intelligence Mark-up Language [15], DARPA Agent Mark-up Language [16], Extensible Financial Reporting Mark-up Language, Trading Partner Agreement Mark-up Language [17], Extensible Business Reporting Language, and the like [18].

Multi-agent System Architecture

The distributed nature of the problem suggests a generic multi-agent architecture, where the main components, and the parts of those are described as individual agents that co-operate or compete to fulfil their goals and thus overall system aims.

An intelligent agent is a software system that operates autonomously to fulfil locally specified goals [19]. The main characteristics of these systems are reactivity (it senses its environments and makes actions based on the perceptions), proactivity (it works to fulfil its goals), social ability (it is able to communicate with other agents and humans), and persistency (it continuously maintains an internal state). Other secondary characteristics may include being veritable, adaptive, robust, rational, and mobile.

There are emerging standards in this field, like the ARPA Knowledge Sharing Effort (KSE), the OMG Mobile Agent Facility (MAF), and the Foundation for Intelligent Physical Agents (FIPA) [20]. Multi-agent systems (MAS) contain autonomous agents that communicate with each other to solve common tasks. This requires a formalised communication language that can be implemented by the creators of the individual agents. The KQML (Knowledge Query and Manipulating Language) is and example of such a language [21].

Different agents can be designed and customised for different information sources. The multi-agent architecture can also be used to increase the reliability of the collected data. Co-operating agents will deal with the different information sources, and competing agents can be utilised to enhance the retrieval effectiveness. The agent methodology can be also applied during the interaction with the user. The complexity of the IKF system and the problem domain can be hidden from the user using specialised agents. Simple interface agents have tasks like helping the user in filling-in forms, or automatically correcting errors in user inputs. More sophisticated agents can be used as application helpers, explaining and teaching the usage of the system.

Knowledge Repository

Regarding the details of the Knowledge Repository two leading concepts are the knowledge organised around the suitable domain ontology and the XML-based document retrieval, mapping and storage. An ontology is an explicit (possibly formal) specification of the names for referring to the objects in the application and the (logical) statements that describe what these objects are, and how they are related or not to each other [22] [23].

Ontology therefore provides a vocabulary for representing and communicating knowledge that can exist for an agent or a community of agents, for the purpose of enabling knowledge sharing and reuse. The so-called ontological commitment means an agreement to use a vocabulary (i.e. for queries) in a way that is consistent with the theory. Research in ontology is one of the most far-reaching issues in the now-a-days artificial intelligence.

The crucial role of well-defined ontology has already been recognised. IEEE Computer Society pioneered a Standard Ontology Study Group [24]. A strong research group related to the IKF project is tackling the question of the meta-organisation of the ontological hierarchies [25], finally fairly recently a number of ontology based enterprise models has been developed [26] [27]. All these developments serve as a basis for the development of the suitable Hungarian enterprise ontology and knowledge base.

Summary

The collection of information forms a vast number of information sources available through the global computer networks and a knowledge intensive processing (reduction) of such information with the purpose to back up management decisions and evaluations opens new dimension in several applications fields.

Designing information systems that provide such support is already possible, if a number of emerging software, information processing, and system integrating technologies is used and further developed to meet the needs of the application. Designed approaches, tools and methods are portable to numerous domains, where the abundance of information and difficulty of making decisions made the introduction of the automated information system questionable until now.

In the framework of the Hungarian IKF project the authors have described a functional system architecture that integrates information acquisition, knowledge building, and profile-based report generation techniques to provide support for financial applications. In the current phase of the project technologies and software tools are investigated in the field of automated information retrieval, XML-based document processing, and data mining.

References

1. EUREKA PROJECT “IKF - Information and Knowledge Fusion”, March 2000.
2. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds), “Advances in Knowledge Discovery and Data Mining”, Mit Press, 1996
3. Korfhage, R., “Information storage and retrieval”, Wiley Computer Publishing, 1997
4. R. Agrawal, T. Imielinski, A. Swami: ”Mining Associations between Sets of Items in Massive Databases,” Proc. of the ACM SIGMOD Int'l Conference on Management of Data, Washington D.C., May 1993, pp. 207-216.
5. S. Guha, R. Rastogi, and K. Shim, “CURE: An Efficient Clustering Algorithm for Large Databases,” SIGMOD 1998.
6. H. Toivonen, “Sampling Large Databases for Association Rules,” VLDB 1996, pp. 134-145.
7. C. Faloutsos, M. Ranganathan and Y. Manolopoulos, “Fast subsequence matching in time-series databases,” SIGMOD, 1994, pp. 419-429.
8. H. Mannila, H. Toivonen, and A. I. Verkamo, “Discovering Frequent Episodes in Sequences,” In U. M. Fayyad and R. Uthurusamy, eds, Proc. of the 1st Int’l Conf. on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, Aug. 1995.
9. J.C. Shafer, R. Agrawal, M. Mehta, “SPRINT: A Scalable Parallel Classifier for Data Mining,” VLDB 1996, pp.544-555
10. J.E. Gehrke, V. Ganti, R. Ramakrishnan, and Wei-Yin Loh, “BOAT -- Optimistic Decision Tree Construction,” In Proceedings of the 1999 SIGMOD Conference, Philadelphia, Pennsylvania, 1999.
11. S. Chakrabarti, et.al., “Mining the Web's Link Structure,” IEEE Computer 32: 60-67 (1999)
12. P. Prescod and Ch. F. Goldfarb, “The XML Handbook - 2nd Edition”, Prentice Hall, 1999
13. XML-Based Ontology Exchange Language (XOL), http://www.oasis-open.org/cover/xol.html
14. Financial Products Mark-up Language (FpML), http://www.oasis-open.org/cover/fpml.html
15. Artificial Intelligence Mark-up Language (ALICE), http://www.oasis-open.org/cover/aiml-ALICE.html
16. DARPA Agent Mark-up Language (DAML), http://www.oasis-open.org/cover/daml.html
17. Trading Partner Agreement Markup Language (tpaML), http://www.oasis-open.org/cover/tpa.html
18. XML: Proposed Applications and Industry Initiatives, http://www.oasis-open.org/cover/ xml.html#applications
19. J. M. Broadshaw, “Software Agents”, The MIT Press, 1997
20. Y. Labrou, T. Finin, and Yun Peng, “Agent Communication Languages: The Current Landscape”, IEEE Intelligent Systems, March/April 1999, pp. 45-52
21. UMBC KQML Web, http://www.cs.umbc.edu/kqml/
22. John Sowa's web site devoted to knowlege representation and related topics of logic, ontology, and computer systems, http://www.bestweb.net/~sowa/direct/index.htm
23. N. Guarino and Ch. Welty, “Towards a methodology for ontology-based model engineering”, in Proc. of the ECOOP-2000 Workshop on Model Engineering. June, 2000.
24. Standard Upper Ontology, IEEE Study Group, IEEE Computer Society, Standards Activity Board, June 2000, wysiwyg://634/http://ltsc.ieee.org/suo/index.html
25. N. Guarino and Ch. Welty, “A Formal Ontology of Properties”, LADSEB/CNR Technical Report 01/2000, http://www.ladseb.pd.cnr.it/infor/ontology/Papers/OntologyPapers.html
26. M.S. Fox, J.F. Chionglo, and F.G. Fadel, “A Common-Sense Model of the Enterprise”, Proceedings of the 2nd Industrial Engineering Research Conference, 1993, pp. 425-429, Norcross GA: Institute for Industrial Engineers
27. US Taxonomies, US GAAP C&I Taxonomy 00-04-04, http://http://www.xbrl.org/US/default.htm