Supporting Information Fusion with Federated Database Technologies

— Position Paper —

Kai-Uwe Sattler and Gunter Saake

Department of Computer Science, University of Magdeburg
P.O. Box 4120, D-39016 Magdeburg, Germany
{kus|saake}@iti.cs.uni-magdeburg.de

Abstract. A common problem facing many users today is to extract and combine information from multiple, heterogeneous sources and to derive information of a new quality or abstraction level. Though essential parts of this information fusion process can be supported by techniques developed in the field of federated databases, new approaches for managing consistency, uncertainty or quality of data and enabling efficient analysis of distributed, heterogenous sources are still required. This paper presents a brief survey of requirements and applications of information fusion from the database view, discusses the usage of federated database technologies in fusion systems and points out possible research directions.

1 Introduction

Today's state of the art in database technology enables storage and management of large volumes of data. Modern communication networks like the Internet promote the access to world-wide distributed databases. However, the expanding number of available information sources leads to what users perceive as information overload. In addition, information providers often represent their data in a heterogeneous fashion regarding structure and semantics. Therefore, the problem of information discovery and integration has become a significant challenge. This affects questions like ensuring actuality and credibility of data, efficient querying distributed sources, converting insufficient structured data as well as combining data from different sources.

In addition to a consistent and uniform access to heterogeneous data sources there is a further value of integration. The integrated data can contain information, which are not represented explicitly but in the form of dependencies, relationships or patterns over the various sources. Classical query techniques from the database field fail for the search and extraction of this implicit or hidden kind of information. Therefore, an important requirement for integrated information systems is the (semi-)automatic and intelligent transformation of data into useful information. The notion of transformation includes various aspects like integration, filtering, analysis and preparation of data aimed to discover and represent the hidden knowledge.

Besides data management and integration the fields of data mining and data fusion contribute solutions to these tasks. Data mining – as core of the process of knowledge discovery in databases – deals with searching and discovering pattern and dependencies in data [5]. Data fusion addresses the combination and interpretation of data from different sources [2]. Applying these techniques in information systems opens new potentials regarding the analysis and extraction of data from large heterogenous sources. The overall process of integration and interpretation of data from different sources as well as the further construction of models for a given domain in order to derive information of a new quality is often called *information fusion*.

The objective of information fusion results in important requirements to methods and techniques of database management. Some of these are addressed already by current research in federated databases. Others, like treatment of inconsistent or uncertain information, have still to be considered.

In this position paper we discuss these requirements from the point of view of federated databases. We point out important research directions supporting information fusion and present potential applications.

2 Requirements

Though various potential fusion applications result in different requirements we are able to identify several tasks which are similar for a wide range of fusion systems. In particular, these tasks are:

– *Data access:* At first, we have to support an uniform access to different sources. This involves the usage of database gateways in order to hide the heterogeneity, accessing Web sources via the appropriated protocols and extracting semistructured data from these sources as well as query translation, optimization and processing.
– *Data integration:* An integrated view should represent data from the different sources in a homogeneous model. This involves repairing conflicts at schema or instance level and dealing with aspects of data quality. In addition, inter-source associations have to be represented and managed at the global layer.
– *Analysis and abstraction:* Filtering or condensing data and extracting dependencies or abstractions offers the opportunity to yield information of a new quality. The notion of new quality depends on the concrete application. Possible representations are generalized aggregations and associations [1], clusters [6] and classes.
– *Presentation and processing:* The discovered information have to be presented according the problem domain or to be prepared for further processing [13].
– *Representation of meta-information:* An important prerequisite for fusion is the existence of information about the sources, the fusion objects and problem domain. These meta-information should be managed by the system and updated or extended during the fusion process, e.g. for optimization purposes.

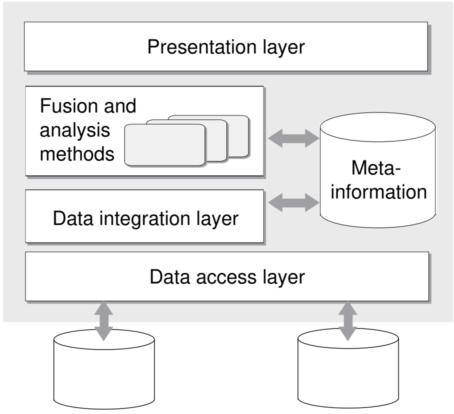

Fig. 1. Infrastructure for Information Fusion

These services should be provided by an infrastructure (Figure 1), which acts as an implementation base for various applications of information fusion.

3 Federated Database Technology in Information Fusion

Technologies from the federated database field could fulfill an essential part of the discussed requirements. Federated database design concerns with schema integration [3, 15, 4] and integrity constraints; FDBS components support transparent data access and manipulation [7] as well as translation and optimization of queries [14]. For fusion applications the most important tasks are as follows:

– *Intelligent support of integration:* In many cases the integration of the individual schemas is a complex process that cannot be done automatically. The database designer has to eliminate structural conflicts and to integrate different class hierarchies. This process has to be supported by tools taking into consideration semantical aspects as well as the quality of data [8, 9].
– *Efficient data access:* Analyzing huge amount of data from several sources requires efficient access mechanisms. Especially for distributed heterogeneous sources building and using suitable indexes, caching or replication can improve performance of processing significantly. Moreover, the requirements of the analysis methods regarding the access interfaces of sources have to be considered.
– *Integration of semistructured data:* Because of the rapid growing of the Web many sources provide access to information that may be not conform to a rigid schema but is stored in HTML documents or semistructured repositories. Therefore, integration of semistructured data and processing queries over the data are further tasks of a fusion system [16].
– *Extraction of meta-information:* Information about semantics and quality of data form an important base for fusion. The appropriated information has to be extracted from the data or provided by the user.

One critical feature of information fusion systems for large datasets is an efficient processing of queries. In particular, these results in the following requirements:

- The query processor of the DBMS has to contain an open optimizer in order to optimize fusion methods together with database operations.
- The DBMS has to provide functionality for integrating data more sophisticated than simple export/import procedures.
- Query processing should support ranking or quality valuation, e.g. information retrieval techniques.
- The DBMS should base on an open software architecture and contain an extensible repository for meta-information in order to embed fusion methods.
- Query processing needs to be supported by advanced techniques, for example building indexes on the fly or integration of source-specific indexes.
- Many statistical methods use a random set of data for initialization before the whole database is analyzed. This sampling is currently not supported by commercial databases.

Methods and techniques from the federated database field form an essential part of information fusion systems. Anyway, important tasks regarding adaptable, efficient query processing and optimization for heterogeneous sources have still to be considered.

4 Application Examples

Possible applications of information fusion exist there, where data from different sources have to be combined in order to derive new information and support decision processes. This overlaps at first sight with the topic of data warehousing. However, information fusion deals with (semi-)automatic and intelligent transformation of heterogenous sources, whereas data warehousing is aimed at the interactive exploration of integrated and materialized data. In the following, we will outline two application scenarios which are currently considered by our research group in cooperation with domain experts.

Bioinformatics. As part of worldwide efforts in the area of biotechnologies many research groups collect molecular-genetic data and make them available via the Internet. Commonly these sources are focused on a specific application of data only (e.g. genes, metabolic pathways, diseases etc.). Often the data are stored in flat files using an own proprietary file format. An integration of these sources enables complex fusion queries and offers opportunities for discovering new associations [11]. However, these tasks are complicated by insufficient or unknown structured, redundant or incorrect data, the huge amount of data and the continual updates and extensions. Therefore, integration and analysis of bioinformatics data is an important application of information fusion.

Telecommunication. Providers in the telecommunication market manage a large volume of data in their various business units. These data contain information about products, customers and their connections as well as network-specific information. Integrating and combining these data sources is an crucial task not only for management or marketing units, but also for system and network management. For example, one function of fault management is alarm correlation [12]: identifying network faults and their causes. Defining the necessary correlation model is a complex task, especially for large-scale networks. By analyzing historical alarm sequences the network operator can identify frequent pattern and derive rules for generalization and compression of alarms [10]. Combined with information from the configuration database these rules form the foundation of the correlation model.

5 Conclusion

A significant challenge for database technology has been the extraction and combination of data from different heterogeneous sources followed by filtering, condensation and derivation of information at multiple abstraction levels. As a result of the increasing information load there is a high demand for intelligent information fusion. This brief survey has discussed requirements, which have to be fulfilled by technologies from the federated database field in order to support efficient fusion techniques and has outlined directions for future work. As part of several planned research projects at our university we will investigate important aspects of these requirements.

References

1. R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *Proc. of the 20th Int. Conf. on Very Large Data Bases (VLDB)*, pages 478–499. Santiago, Chile, September 1994.
2. H. Arabnia and D. Zhu, editors. *Proc. of the Int. Conf. on Multisource-Multisensor Information Fusion - FUSION '98*. CSREA Press, Las Vegas, NV, 1998.
3. C. Batini, M. Lenzerini, and S. Navathe. A Comparative Analysis of Methodologies for Database Schema Integration. *ACM Computing Surveys*, 18(4):323–364, 1986.
4. S. Conrad. *Föderierte Datenbanksysteme: Konzepte der Datenintegration.* Springer-Verlag, Berlin/Heidelberg, 1997.
5. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From Data Mining to Knowledge Discovery: An Overview. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusuamy, editors, *Advances in Knowlede Discovery and Data Mining*, chapter 1, pages 1–34, AAAI/MIT Press, Cambridge, MA, 1996.
6. D. Fisher. Optimization and simplification of hierarchical clustering. In *Proc. of 1st Int. Conf. on Knowledge Discovery and Data Mining (KDD-95)*, pages 118–123. Montreal, Canada, August 1995.
7. G. Gardarin, S. Gannouni, B. Finance, P. Fankhauser, W. Klas, D. Pastre, R. Legoff, and A. Ramfos. IRO-DB — A Distributed System Federating Object and Relational Databases. In *Object-Oriented Multidatabase Systems — A Solution for Advanced Applications*, chapter 20, pages 684–712, Prentice Hall, Eaglewoods Cliffs, NJ, 1996.
8. M. Gertz. Managing Data Quality and Integrity in Federated Databases. In *2nd Annual IFIP TC-11 WG 11.5 Working Conf. on Integrity and Internal Control in Information Systems.* Warrenton, Virginia, November 1998.

9. M. Gertz and I. Schmitt. Data Integration Techniques based on Data Quality Aspects. In I. Schmitt, C. Türker, E. Hildebrandt, and M. Höding, editors, *Proceedings 3. Workshop "Föderierte Datenbanken", Magdeburg, 10./11. Dezember 1998*, pages 1–19, Shaker Verlag, Aachen, 1998.
10. K. Hätönen, M. Klemettinen, H. Mannila, P. Ronkainen, and H. Toivonen. Knowledge Discovery from Telecommunication Network Alarm Databases. In *Proc. of 12th Int. Conf. on Data Engineering (ICDE'96)*, pages 115–122. New Orleans, 1996.
11. M. Höding, R. Hofestädt, G. Saake, and U. Scholz. Schema Derivation for WWW Information Sources and their Integration with Databases in Bioinformatics. In *Advances in Databases and Information Systems – ADBIS'98, Poznań, Poland, September 1998*, LNCS 1475, pages 296–304. Springer-Verlag, Berlin, 1998.
12. G. Jakobson and M.D. Weissman. Alarm Correlation. *IEEE Network*, 7(6):52–59, November 1993.
13. D. Keim and H.-P. Kriegel. Visualization Techniques for Mining Large Databases: A Comparison. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):923–938, December 1996.
14. W. Meng and C. Yu. Query Processing in Multidatabase Systems. In W. Kim, editor, *Modern Database Systems*, pages 551–572. ACM Press, New York, NJ, 1995.
15. E. Pitoura, O. Bukhres, and A. K. Elmagarmid. Object Orientation in Multidatabase Systems. *ACM Computing Surveys*, 27(2):141–195, 1995.
16. R. Yerneni, Y. Papakonstantinou, S. Abiteboul, and H. Garcia-Molina. Fusion queries over internet databases. In *Advances in Database Technology - EDBT'98*, LNCS 1377, pages 57–71, Springer-Verlag, 1998.