Journal of Applied Sciences, 2010
ISSN 1812-5654

Integration of Heterogeneous Data Sources

[1]N.S. Ougouti, [1]H. Belbachir, [2]Y. Amghar and [2]N.A. Benharkat
[1]System, Signal and Data Laboratory, Departement of Computer Science, Faculty of Science,
University of Science and Technology, Oran, Mohamed Boudiaf USTO,
P.O. Box 1505, El MNaouer, Oran, Algeria
[2]LIRIS, UMR 5205, Insa of Lyon, 7 Avenue Jean Capelle F-69621 Villeurbanne CEDEX, France

Abstract: Access to distributed, heterogeneous and autonomous information sources, becomes possible with the Internet. These information sources are distinguished by the nature of information, namely, the ontological domain to which they belong but also by the type of media they are issues, such as image, text, video, etc. With the advent of semantic web, new opportunities in multi-sources integration are emerging and many approaches are revisited with taking into account the new requirements. Also, there is the use or reuse of datawarehouses, mediators and peer-to-peer systems. Our project aim to propose a distributed and open system for indexing and searching multimedia content (DIOSYS) and especially an integration system based on the Peer-to-peer paradigm. In this study, we propose a state of the art of the integration problem by examining the most representative approaches of the three currents and we will try to summarize this study with use of tables after having presented and justified a set of criteria.

Key words: Semantic web, mediation, peer-to-peer, ontologies, wrappers

INTRODUCTION

Reach a wide quantity of various, heterogeneous and distributed information on several sites and in particular on the Web, introduces new problems to the users and pushes the community of researchers to double efforts in order to try to integrate the corresponding sources with the aim of carrying out increasingly pointed applications which often contribute to an effective and fast decision-making.

With the advent of the semantic Web, new possibilities are offered and many traditional approaches are revisited with these new requirements. Also, we observe the use or the re use of datawarehouses, mediators and especially of peer-to-peer systems.

The datawarehouse approach (Vodislav, 2007). consists in carrying out integration by building real databases gathering relevant information of considered applications. The user will work directly on the data stored in the warehouse.

The mediator approach is a method where the data are accessible only from the information sources, the user in this case will work on abstract views built with the aim of describing the various data sources. Searching information from these sources requires the construction of execution plans to obtain the whole results from information sources, the most important mediators are in (Chawathe *et al*., 1994; Haas *et al*., 1997; Tomasic *et al*., 1995; Rousset *et al*., 2002; Mena *et al*., 1996).

The (Peer-to-peer or P2P) approach is a recent paradigm, it can be seen like a generalization of médiators/datawarehouses architectures. These integration systems follow a decentralized approach for integration of autonomous and distributed peers containing data which can be shared. The principal objective of such systems is to provide a semantic interoperability between several sources with the absence of global schema, several P2P systems exist like Edutella (Nejdl *et al*., 2002), PeerDB (Ng *et al*., 2003).

In this study, we propose to examine the most representative approaches of the two most recent currents: the mediation and the P2P, we will try after to summarize this study after having presented and justified a set of criteria.

THE DATAWAREHOUSE APPROACH

The purpose of datawarehouse systems are to recover, organize, integrate and store data of multiples heterogeneous and distributed sources into a same site,

Corresponding Author: N.S. Ougouti, System, Signal and Data Laboratory, Departement of Computer Science, Faculty of Science, University of Science and Technology, Oran, Mohamed Boudiaf USTO, P.O. Box 1505, El Mnaouer, Oran, Algeria

in order to allow a centralized interrogation and a global view of data. Adding data into datawarehouse is generally done in a batch mode through a module allowing the extraction, the transformation and the loading of data (ETL). The query done on the schema of the warehouse is then carried out directly on its contents.

The advantage of these systems is the query treatment simplicity, on the other hand a problem lies in the refreshing of data: the warehouse must be updated regularly in order to keep a faithful image of the data which it represents.

THE MEDIATORS APPROACH

The mediators approach consists in making an interface between the user and the various accessible data sources through the Web. This interface gives the impression of a centralized and homogeneous system, it is made up of a global schema and often with the use of an ontology which help the user in the formulation of his request thanks to a structured vocabulary. We distinguish two types of mediation systems: a global schema as view on local schema as GAV (Global As View) or local views as view of the global schema (Local as View). The effective interrogation of the data is done by the use of wrappers which translate the query rewritten into terms of view into a specific query language accepted by each source.

In what follows we describe some integration of information systems using mediators.

TSIMMIS: Chawathe *et al*. (1994) (The stanford-IBM Manager of Multiple Information Sources) have an architecture based on mediators and wrappers generated automatically, exploiting a canonical object model for the structured and semi structured data exchange in dynamic and heterogeneous environments. The wrappers have as a role to convert the objects corresponding into a common model named OEM (Object Exchange Model), this last allows a simple construction of the objects. A specific interrogation language (OEM-QL) was developed to query these objects. On top of translators, there are the mediators, they have as role to encapsulate necessary knowledge to treat a specific type of information, to direct the request towards the most adapted data sources and to arrange the result who must be turned over to the user.

GARLIC (Haas *et al*., 1997; Carey and Haas, 1995): It is a project whose objective is to integrate various multimedia sources by providing an integrated view of the local data sources schemas. These schemas are fused in a total schema expressed in ODMG. Access to GARLIC objects can be done in two ways: via a graphic interface or by using the query language of GARLIC. This Last is named GQL, it is an extension of SQL supporting way's expressions, overlapping collections and methods. The queries formulated by users are sent to the queries processor, who develops execution plans decomposed for the multiple data sources before being sent to the wrappers, which are associated to each information source playing thus the role of an interface to reach the local data.

The DISCO project (Distributed Information Search COmponent) (Tomasic *et al*., 1995), aimed to deal with the problems due to the great expansion of the Web. These problems are due to the fact that it is difficult in information search domain to locate the relevant data, to reach these data then to integrate them if they are heterogeneous into a global network. It is a data model which represents an extension of the object data model ODMG and its query language OQL. Its architecture consists of three levels: (Patrick, 1995) the Mediators who encapsulate the representation of the data sources, the Translators which convert the requests on the local data sources and the Catalogue which index all the components of the system and their localization.

The mediators manage connections to the data sources. Adding other sources is facilitated by modelling them with objects. These mediators manage also repertories of metadata and indices which have as role to optimize the access to data sources. They return the queries to the wrappers after having reformulated them and optimized them. Once these queries carried out, they recover the results, recompose them and return them to the application.

The information Manifold (Kirk *et al*., 1999; Levy *et al*., 1996) is a system for the extraction and the organization of information starting from various structured or not structured information sources based on the Web. The architecture of this system uses a knowledge bases which contains a rich field model, for describing the properties of the information sources and giving to the user the possibility of formulating high level requests. The language used permit to describe the semantics of structured sources contents. It is a combination of Horn rules and concepts resulting from the traditional logic of description, which permits to determine the information source the most appropriate to a given request. Its data model is relational, increased with the hierarchies of classes. The user is responsible for the cleaning of redundant information after the query execution.

PICSEL: Rousset *et al*. (2002) offered an environment of mediators construction, it has the possibility of expressing the mediator schema in CARIN language, combining by this, the expression capacity of a formalism containing rules and an other one containing classes (the logic of description ALN). The global schema gathers the whole predicates modelling the domain application of the system. It plays the role of domain ontology which provides the structured vocabulary being used as support for the expression of the query. Knowledge bases are also connected to the mediator to describe the contents of the information sources and to determine which are those which can provide results to a user query. The mediator has only abstract views on the sources data, the adapters are thus introduced to query the data sources by translating the views requests in the specific query language to each source. Research is done in the terms of the global mediator schema, or in terms of the ontology.

A query language was defined, it handles the terms of the domain ontology. A rewriting process allow to identify the relevant sources to answer the query, the required data and the way in which they should be combined to give a precise answer. The CARIN-ALN formalism was adopted as support of domain ontology description and as query language.

OBSERVER (Mena *et al.*, 1996): It is a system which allows interoperability between various sources, by using multiple ontologies to describe the data sources. These ontologies are described by using the logic of description, more precisely the CLASSIC language. There is no total ontology in OBSERVER; the mapping between multiple ontologies is carried out by using correspondence tables. However, the relations between ontologies are limited to basic lexical relations such as synonyms.

InfoSleuth: (Bayardo *et al.*, 1997): It is a Carnot project product, its goal is to integrate heterogeneous information sources. This method tried to introduce new technologies such as the use of agents, domain ontologies, data and services interoperability and internet programming in an open and dynamic environment. Thus, it can be regarded as a network of agents co-operating and communicating by the means of a high level agents query language (KQML). The users formulate their requests by using ontologies via user interfaces based on applets. The knowledge representation language KIF and the interrogation language SQL are used internally to describe the queries on ontologies. These requests are directed by means of brokers towards the suitable agents in order to search for and to integrate the data coming from the various heterogeneous sources.

Mind (Nottelmann and Fuhr, 2003): is a system which integrates heterogeneous multimedia and no cooperative database management systems and gives to the user the impression to work with only one coherent system. Its architecture consists of only one component charged of mediation and several components charged to encapsulate the local data sources (wrappers or proxies) with a proxy for each data source.

The mediator communicates with the co-operating proxies. On its top, exists a data fusion component whose role is to combine the results together. The queries and the documents in Mind are modelled in a formalism using DAML+OIL.

MIROWeb (Luc *et al.*, 1999): It is a project which developed a single technology to integrate multiple data sources under the relational-object model with semi structured data type. It deals with integration problem of irregular sources of the Web and regular relational databases through a mediator architecture based on a hybrid model which supports relational-objects and semi structured types.

In this architecture, the sources are transformed into tables with the possibility of having semi structured attributes. Instances of these attributes are modelled in directed graphs. The atomic objects are stored in relational-objects tables.

MIROWeb has three levels: users, mediator and sources. The user level provides an interrogation interface and a JAVA API. The mediator is based on Oracle 8, Other units exist like the decomposer of queries and the semi-structured unit.

The interrogation interface is a graphic interface to surf through virtual XML documents starting from any repertory chosen by the user. To avoid errors, a list of names is provided by the dictionary of mediator, who contains all the domain metadata. The user can choose a root, develop the tree structure, formulate the predicates of joint and selection, choose the projection nodes and validate resulting request XMLQL.

Xylème (Vodislav, 2007): It is the product of a project which combines the two approaches datawarehouse and mediator, It builds a dynamic warehouse gathering the whole of XML documents of the WEB. On top of this warehouse, a mediator exists, it plays the role of interface of requests between the user and XML documents relating to the same subject.

A semi-automatic tool for acquisition of mappings was built, it uses WordNet and dictionaries of synonymies specific to the application in order to propose correspondences between ways of tags coming from

DTDs of XML documents. In Xylème, the global schema is a set of terms trees. The mappings generated are then presented for validation or rejection at a human expert (Libourel, 2003).

PEER-TO-PEER SYSTEMS (P2P)

The vertiginous increase in information sources on the Web obliges to re-examine the way of building information search systems. A new idea consists in using a peer-to-peer architecture, inspired by popular file sharing systems on Internet like Gnutella and Kazaa. This architecture allows a very great number of connected sources and a network dynamicity.

These peer-to-peer integration systems follow an decentralized approach for the integration of autonomous and distributed peers containing data which can be shared. The principal objective of such systems is to provide a semantic interoperability between several sources in the absence of a global schema.

Senpeer (Faye *et al.*, 2006): It is a peer-to-peer system for data sharing having various data models. It is organized under a super-peer type with a regrouping of peers by semantic fields. Each peer publishes data described by a model in conformity with the relational, object or XML data model and has its own interrogation language. With an aim of a flexible mediation, the data are exported in a pivot model which has a structure of enriched semantically graph called sGraph (semantic Graph), with key words resulting from the schemas and intended to guide the discovery of the semantic correspondences. The requests are exchanged in an internal common format, rewritten and directed towards the relevant peers thanks to the semantic correspondences.

When a peer Formulate a query with its interrogation language (SQL, XQuery, etc.) this one is initially carried out locally then translated into the queries exchange formalism SQUEL(SenPeer Query Exchange Language) and finally sent to its super-peer. The result is the list of the relevant peers accompanied by the semantic rewriting of the request. The query can now be directed towards these various peers. Lastly, the communication is ensured by JXTA platform from Sun.

Edutella (Nejdl *et al.*, 2002): It is an integral part of the peer-to-peer and open source project JXTA, it is a system which provides an access to distributed collections of numerical resources through a network P2P. These resources are described by using metadata and RDF. To extract information from the Edutella network, language RDFQEL is used, it is a language based on Datalog semantics, it is thus compatible with all the query languages existing. The common data model is described thanks to Datalog in the form of JAVA classes and the queries transmitted between peers are represented by RDF.

PEPSINT (Cruz *et al.*, 2004): It is a peer-to-peer system of data management which combines traditional techniques of schemas integration with an P2P infrastructure. It permits to integrate semantically heterogeneous XML and RDF data sources, by using an hybrid peer-to-peer architecture and a GAV mediation approach.

This system contains two types of peers: a super-peer which contains a global RDF ontology and the peers which contain the local schemas and the local data sources. Each peer represents an autonomous information system and is connected with the super-peer by establishing several mappings. A XML-RDF adapter is used to transform an XML schema into an RDF one. It offers two modes of interrogation: The Data-integration mode where the global ontology is queried and P2P hybrid mode where the user can query the local source, this request can be directed towards other peers by using transitive mappings.

PeerDB (Ng *et al.*, 2003): It is a peer-to-peer data management system based on the agents where each peer contains a relational database. The metadata of relations which are shared with other peers are specified in a local export dictionary. There is no mapping between the peers. The reformulation of requests is assisted by agents through a matching of relations strategy i.e. matching of metadata between relations of different peers.

SUMMARY

Mediation integration systems: The various mediators approaches are characterized by the quoted properties below:

- The relation between the local sources schemas compared to the global unified schema (GAV or LAV).
- The common model
- Global schema Query languages
- Formats or types of the data sources used
- Formats of the turned over results

Table 1 contains the different mediator systems, showing for each system the following properties:

- Model Type: LAV or GAV

Table 1: A summary table of mediation integration systems

Name	Type	Common model	Formats of the data sources	Query language of the common model
TSIMMIS	GAV	OEM objects XML	Structured, Semi-structured	Oem-QL,
Garlic	GAV	ODMG	Relation. objects	GQL
Disco	GAV	Extension of ODMG	objects	OQL
The information manifold	LAV	Rules hierarchy of classes	Structured relation. not structured	Request on the view
PICSEL	LAV	Rules classes Carin-ALN	Databases HTML XML	Carin/core-classic
OBSERVER	LAV	Logic description	Not expressed	logic descript. Lang. ontology
Infosleuth	LAV	Based on agents domain ontology	Structured Semi-structured	KQML
Mind	LAV	DAML+OIL Schema	Textes images relational databases objects	DAML+OIL
Xylème	LAV+GAV	tree Based	Documents XML	Queries on trees
MIROWeb	GAV	Relation. object enriched by semi structured data	Relationel Objets semi-structured	XMLQL

Table 2: A summary table of peer-to-peer integration systems

Name	P2P Archite-cture	Common model	Formats of peers	Common model Query Lang.
Eduttella	hybrid Super-peer	Datalog	Relation. XML	RDF-QEL
Sinpeer	Super-peer hybrid	SGraph (Sémantic Graph)	Relation. Objets XML	SQUEL
Pepsint	Hybrid P2P GAV approach	Global Ontologie RDF	XML RDF	RDF QL
Hyperion	pure	Without global schema	relation.DB	Without com. Lang

- Common Model: Object, Relational, XML, Logical or a mixed model
- The format of the data sources used : Relational object, Semi-Structured or web Format
- The language used to query The common model: KQML, DAML+OIL, XMLQL, OQL

Peer-to-peer integration systems: A set of comparison criterions seems to us relevant for the peer-to-peer integration systems, we can summarize them in what follows:

- P2P Architecture of the system (pure or hybrid)
- The common model in case where there would exist
- The format of the data sources used
- The query language of the common model

Table 2 contains the different Peer-To-Peer systems, showing for each system the following properties:

- **P2P architecture:** Hybrid or Pure
- **Common model:** Datalog, Ontology, Object, Relational, ….
- **Peers format:** Relational, object, RDF, Semi-Structured or web Format
- **The language used to query The common model:** SQUEL, RDF-QEL, ….

CONCLUSION

The current tendency is to revisit the integration approaches based on mediation and datawarehouses or to suggest other peer-to-peer systems with the new possibilities offered by the semantic Web.

The use of ontologies proved very effective in semantic integration in the mediators approaches. Several recent works used this concept in a single way such as in Picsel where each source to be integrated is related to only one global domain ontology, in a multiple ways such as in Observer where each source to be integrated is described by its own ontology with the possibility to find correspondences between these ontologies, or in a hybrid way, by having an ontology for each local source connected to a global one. Dealing with ontologies create a new problem in this field, it acts of the definition of semantic correspondences between ontologies in an automatic way.

But these mediation integration systems are not very flexible and the global schema could be a bottleneck. The need for new decentralized and dynamic tools is felt. The peer-to-peer systems are regarded as a good solution for the Web scale passage. They have the advantage of not needing a single schema, to be able to add data and information on the schema in each peer and to query each peer with its own query language but they do not hands with data semantics.

For datawarehouse approaches, it would be interesting to take into account the knowledge on data and to introduce domain ontologies in their function. Moreover, it would be necessary that data integration and data analyzes are made in real time, because it is inconceivable nowadays and especially for Internet applications to make decisions based on relatively old data.

In all the systems reviewed, little of them take into account multimedia data sources like Image or Video types. It would be interesting to propose a mediation or peer-to-peer architecture which would include in their

local data sources these formats combined with the other formats, with taking account of the innovations brought by the semantic Web in the semantic description of these data. The major problems in this case will be the interrogation, the integration and the indexing of these data.

REFERENCES

Bayardo, R.J.Jr., W. Bohrer, R. Brice, A. Cichocki and J. Fowler *et al*., 1997. InfoSleuth: Agent-Based Semantic Integration of Information in Open and Dynamic Environments. MCC, Austin, Texas, pp: 1-12.

Carey, M.J. and L.M. Haas, 1995. Multi-media towards heterogeneous systems information: The garlic approach. Proceedings of the International Workshop on Research Issues in Data Engineering: Distributed Object Management, March 6-7, Taipei, Taiwan, pp: 124-131.

Chawathe S., H. Garcia-Molina, J. Hammer, K. Ire- land, Y. Yiapakonstantinou, J. Ullman and. J. Widom, 1994. The TSIMMIS Project: integration of heterogeneous information sources. Proceedings of the IPSJ Conference, Oct. 1994, Tokyo, Japan, pp: 7-18.

Cruz I. F., H. Xiao and F. Hsu, 2004. Peer-to-peer semantic integration of XML and RDF data sources. Internal Repport, Department of Computer Science University of Illinois at Chicago, USA. http://www.cs.uic.edu/~advis/publications/dataint/ap2pc04.pdf.

Faye, D., G. Nachouki and P.Valduriez, 2006. Integration of heterogeneous data in SenPeer. ARIMA., 5: 1-8.

Haas L.M, D. Kossmann, E.L. Wimmers and J. Yang, 1997. Various optimising query across dated sources. Procedings of the 23th International Conference on Very Large Data Bases, (VLDB'97), Athens, Greece, pp: 276-285.

Kirk, T., A.Y. Levy, Y. Sagiv and D. Srivastava, 1999. The information manifold. Proceeding of the ACM SIGMOD International Conference on Management of Data, Oct. 22, Philadelphia, pp: 299-310.

Levy, A.Y., A. Rajaraman and J. Ordille, 1996. Querying heterogeneous information sources using source descriptions. Proceedings of the 22th International Conference on very Large Data Bases, Sept. 3-6, Bombay, India, pp: 251-261.

Libourel, T., 2003. Mediation via metadata. Specific action 97, STIC, CNRS. http://www.lirmm.fr/~libourel/MM/MetaMedia.htm.

Luc, B., T. Chan-Sine-Ying, T.T. Dang-Ngoc, J.L. Darroux, G. Gardarin and F. Sha, 1999. MIROWeb: Integrating multiple data sources through semistructured data types. Procedings of the 25th VLDB Conference, Sept. 7-10, Edinburg, Scotland, pp: 750-753.

Mena E., A. Illarramendi, V. Kashyap and A.P. Sheth, 1996. OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-Existing Ontologies. Kluwer Academic Publishers, Boston, pp: 1-49.

Nejdl, W., B. Wolf, C. Qu, S. Decker and M. Sintek *et al*., 2002. EDUTELLA: A P2P networking infrastructure based on RDF. Proceedings of the 11th International World Wide Web Conference, May 7-11, Honolulu, Hawaii, pp: 604-615.

Ng, W.S., B.C. Ooi, K. Tan and A. Zhou, 2003. Peer DB: A P2P-based system for distributed data sharing. Proceedings of the 19th International Conference on Data Engineering, March 5-8, Bangalore, India, pp: 633-644.

Nottelmann, H. and N. Fuhr, 2003. The mind architecture for heterogeneous multimedia federated digital libraries. Lecture Notes Comput. Sci., 2924: 112-125.

Patrick, V., 1995. Bases de donnees et Web: Enjeux, problemes et directions de recherche. Internal Report, Institut De Recherche En Informatique Et En Automatique. http://horizon.documentation.ird.fr/exl-doc/pleins_textes/pleins_textes_6/colloques2/010008711.pdf.

Rousset, M.C., A. Bidault, C. Froidevaux, H. Gagliardi, F. Goasdoue, C. Reynaud and B. Safar, 2002. Construction of mediators to integrate multiple and heterogeneous data sources: PICSEL Project. Information-Interaction-Intelligence Journal, Volume 2.

Tomasic, A., L. Raschid and P. Valduriez, 1995. Scaling heterogeneous databases and the design of disco. Internal Repport No. 2704, Institut de Recherche en Informatique et en Automatique. http://www.cs.cmu.edu/~tomasic/doc/1995/RR-2704.pdf.

Vodislav, D., 2007. Integration, Partage et diffusion de donnees sur le Web. Habilitation a Diriger des recherches. http://cedric.cnam.fr/PUBLIS/RC1317.pdf