Mediators to Access Public Data Resources

Mário Flecha
Prodemge/Fundação João Pinheiro/Depto. de Ciência da Computação
Universidade Federal de Minas Gerais
R da Bahia, 2277 – Lourdes
Belo Horizonte, MG 30160012
Brasil
Flecha@prodemge.gov.br

José Luis Braga
Depto. de Informática
Universidade Federal de Viçosa
Viçosa, MG 36571000
Brasil
zeluis@gnv.ifas.ufl.edu

Abstract

This paper reports an experienced and evaluated three layer architectural solution to cooperate with user applications for database access in a large and legacy real environment. It looks for mediator software independence from idiosyncrasies of user application interfaces and specific database schemata. The cornerstone to implement this architecture is a set of relational databases and views manipulated by a Relational Inference Machine – RIM – which is based on the lexicon that comes from datafield content semantics. There is a variety of utilizations in real life, specially State's public services, which operate with different semantic domains applyied to individuals identification, localities and addresses naming, personnel data, bibliographic items location, amongst others. Some of those applications are over increasingly large and intensively accessed databases, especially in legacy and federated environments growing used over a network. Such several interactive public database systems are daily using these mediation services and tools over millions of data records which had posed some challenges along its use and evolution. Of course, lots of unexplored issues and possibilities still remain to be searched and solved. We are trying an approach based on semantic constellation for context hunting over non-structured data, in order to improve semantic searching.

1. Introduction

To co-operate with database consultant users over a network, mediator software plays the role of component in Co-operative Information Systems - CIS. One important mediators' goal is semantic support on information searching over heterogeneous and mixed legacy data dispersed in a medley of federated databases.

A general mediator which covers all topics of interest to any application is highly unwise. The concept of generic and modular components sounds as a better strategy to build flexible configurations with sets of mediators, each one focusing its own scope and service [Wiederhold and Genesereth, 1997].

Use of specialized mediator components permits to build flexible mediation layers and better co-operation with the users layer. Mediators which promote transparent and intelligent integration of information derived from heterogeneous sources of data may interact with others for data retrieval questions which include data resources stored in less recent technologies, originally not conceived for use over a network [Wiederhold and Genesereth, 1997; Wiederhold, 1996; Wiederhold, 1992]. A co-operative interface for database access is also helpful to the users applications, specially to publicize information and make it easier consultations and electronic commerce amongst other services over the infoways.

A conceptual support comes from semantic information processing research field. We have taken as basis the work of [Quillian, 1968] on a representative computational model about human associative memory and its use to create or obtain meaning from the context.

Pursuing a generic mediation architecture, independent from specific database schemata and user application peculiarities, we have developed a technique named Content Semantics - CS - [Flecha, 1994, Flecha, 1997]. It is intended to be a concise method to contextualize, extract and produce semantics infered from byte contents during data entry time or directly inserted by human specialists in specific domain ontology for late use in a variety of information recovery purposes.

CS main concern is to provide a practical way to model semantic domains based on lexical structure and contextual information. As defined in the Semantic Theory [Ullmann, 1977], the lightweight lexicon organization might be the lever to facilitate and speed automatic treatment of meaning for co-operative data access mediation, including poorly structured data like HTML or text files.

Our implementation has used mechanisms based on the lexicon for pragmatic and contextual approach on the treatment of disambiguation, polysemy, synonymy and homonimy [Ullmann, 1977]. The mechanism is based on lightness of lexical structure and its halfway position between systematized and non-systematized organization [Ullmann, 1977]. Lexical level holds more atomic and indivisible elements which are semantic cornerstones of language and do provide aid for co-operative information retrieval purposes. Natural Language Processing, differently, for its intrinsecal issues has to face syntactical structure which is heavier than the lexical one. Syntactical level implies order, function and all the natural language internal complexities and rules. Our work is not concerned in a mediator with such skill. We have striven for a mediator component able to offer semantic services and co-operate with the users to access data resources.

Though, lexical approach touches underlying questions related to ontology, semantic knowledge maintenance and re-usability over data resources in the same semantic domain, including terms from different idioms [Wiederhold, 1992; Wiederhold, 1996]. The terms may provide some meaning by themselves (autosemantic terms) but the context is the mechanism which sets up less ambiguously the meaning [Ullmann, 1977]. For structured databases the context is in general previously known from the semantic domain of datafield contents and also by the semantic derived from the processes of an application, which may create additional terms external to the datafield contents but completely related to the context of the database and its tuples.

The sets of terms and contexts are onto-components of one or more ontologies that may benefit from the situation but do not belong to an specific application. Beyond this automatic acquisition there is a non-automatic maintenance for human specialists in a domain which entails the complete process of ontology maintenance. The knowledge base is the point of contact between the mediator and a specific data resource. The ids of ontology terms are related to the ids of one or more data resources instances. Knowledge base maintenance is similar to ontology, except that it is particular to an application. Hence, the relationships between ontology terms and data resources instances are kept automatically by the applications or by human specialists, possibly other than the ontologies' ones.

Under CS model approach an ontology implements pragmatic concepts as contextualized lexical terms whose relationship with the instances of fact databases compose a semantic knowledge base. To use a CS-based knowledge base RIM was implemented in order to explore terms delivered by the users in their consultations. Using the ontology and the semantic knowledge base as components RIM searchs for relevant answers from data resources.

When searching for information a mediator based on CS model and RIM shall "understand" contents in datafield bytes with semantics derived mainly from domain-specific lexicon used in natural languages . In doing so it relieves users from the burdensome duty of learning a query language like SQL, even when simplified [Shneiderman, 1993; Wiederhold, 1992]. It reduces time consuming dialogue sessions, minimizes useless bytes and does not require the user or itself to know details of specific database schemata, something impossible over the Internet. RIM and CS approach permits a mediator to scale up from specific to generic use of semantic domains, with ontologies and knowledge bases instantiated at execution time without geting involved with DBMS model, database schema or user external interface.

2. An Architecture towards Independent Mediators

In a three layer architecture, to be a middle independent component a mediator cannot be restricted to application specifics (upper layer) or DBMS protocols (lower layer) [Wiederhold, 1992; Wiederhold, 1996; Wiederhold and Genesereth, 1997].

To attain this goal at the upper layer user application plays the role to interface externally with users and internally with mediation layer interface. This way a mediator gets itself free from any particular application detail, becoming widely reusable and useful to upper layer clients [Flecha, 1997].

A mediator must provide a downward interface able to separate itself from data resource lower layer yet providing a clear way to request services from underneath [Flecha, 1997] (see figure 1).

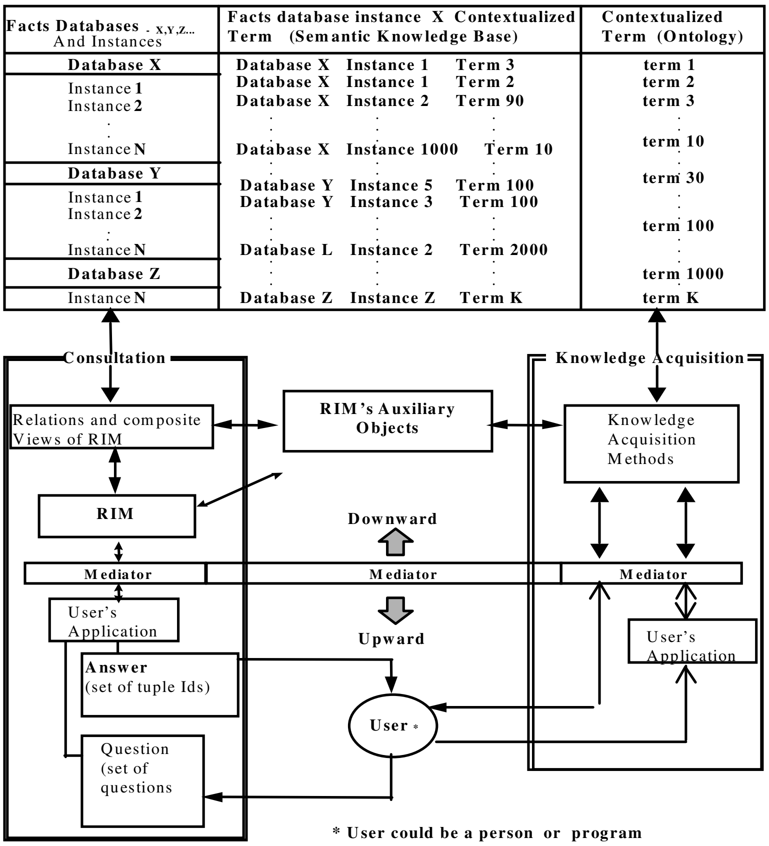
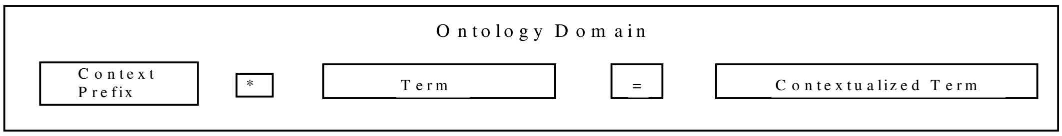


Figure 1 . Basic mediaton architecture

SQL alone does not provide user and user application with all flexibility and cognition needed to interface co-operatively and independently with data resources layer. Usual indexes are not enough [Wiederhold, 1992; Wiederhold, 1996; Wiederhold and Genesereth, 1997]. CS and RIM provide a way to implement indirect indexation [Flecha, 1994], [Flecha, 1997].

By its turn, RIM offers a co-operative way to implement user consultations, not merely *responding* reactively but properly *answering* with meaningful information [Webber, 1986]. RIM has a conceptual and logical schema that can be implemented in different ways. Architectural issues coped towards schema-independent mediators include layers openness with high degree of coupling between them [Rumbaugh, 1994; Fischer, 1994; Wiederhold, 1992]. Open layer-based architectures generate high coupling and compromise reusability and modularity [Rumbaugh, 1994; Fischer, 1994; Wiederhold, 1992]. Closed layer architectures can better achieve schema independence [Rumbaugh, 1994; Fischer, 1994; Wiederhold, 1992].



(a) Example: city*Seattle; state*WS; year*1998. The context database, beyond the prefix, keeps processing information for term treatment, like phonetization, words breaking etc

Figure2: CS Ontology's Ontocomponents

Contextualized terms may belong to one or more ontologies. They are created and captured early by semantic knowledge acquisition methods at the mediator upward interface.

CS and RIM increase mediator's fan-in due to easy-connection and recovering capacity over data resources whose universe of discourse belongs to the same semantic domain, in spite of the underlying implementation of them. Indeed, there is a database environment where RIM executes and resides, but it does not concern the execution environment of user applications or data resource layers.

An individual will make personal use of terms set to compose his speech habits [Ullmann, 1977]. This is the repertory that each one uses in situations where linguistic events happen, including database consultations and man-machine dialogs. Hence, an up-to-date dictionary of basic terms shall be kept as close as possible to completion, plus a context and contextualized term databases which are ontocomponents based on common sense pragmatic concepts expressed with terms (not exclusively words but also abbreviations, acronyms, expressions, compound words, etc.).

3. Evaluation & Measurements

There are mediator implementations specialized in bibliographical search, police criminal search and identification, State's institutional and departmental name search among others. There is use cases where information is merged from different DBMS models. In fact, as said before, there is no restrictions in this matter.

For example, the State Police Information System has three fact databases, which keep data about eleven million citizens, with a semantic knowledge base of sixty million tuples and ontology of twenty million terms. Another system keeps information about four hundred thousand State's public servants with a forty million tuples semantic knowledge base and a ten million terms ontology.

A comparison between two systems was set to evaluate and measure the degree of co-operation with database consultant users (tables 1 and 2). One of them uses a mediator component aid and the other doesn't. In this simple use case we fave focused measuring of semantical power of our mediator. It was not our concern to show the complete set of facilities which includes to search multiple heterogeneous data resources.

Systems compared are both specialized in localities from all Brazil and have the same original source data: the Post Office. Data ranges from states, cities, towns, quarters and all kinds of streets. Some mediators have bigger knowledge bases and ontologies than location mediator but it is the one with higher fan-in and concurrent use on the evaluated environment.

Non-mediated system, whose name is CepDigital, is a software product for personal use that executes in stand-alone equipments. It offers an external graphical interface and explores string-matching techniques for postal data recovery. Mediated system works in an interactive, concurrent, multi-user environment. It is supported by a mainframe that hosts localities fact database and mediators database components, which are intensively called in by corporate user applications.

Locality information found by the mediator is used in a variety of ways by user applications (e.g., to display postal information, to make data entry validation, to standardize or convert it into a location sequential number which hides or represents datafield like birthplace, residential address, commercial address and so on).

Compared systems execute in quite distinct scenarios, therefore the comparison is limited over the same data collection, using the same Consultation terms in different environments and media.

It was not our goal to answer questions related to external interface characteristics, hardware performance, RDBMS performance or environmental issues. Yet, RIM technical issues related to selectivity in large knowledge bases (from the order of dozens of millions of tuples) required performance solutions. The solution of such restrictions has produced drastic response time reduction.

Evaluated mediator's upward interface programming code was implemented in a 4GL under non-object oriented paradigm. Downward interface implementation uses solely SQL over a set of relations and views, which inter-operate together and are started from upward interface.

Location mediator is used for several corporative systems at the level of State or State's Secretariat of Minas Gerais. Some of them are intensively used 24 hours, seven days a week, under concurrence. However, the absence of OO paradigm has produced some undesirable effects on its implementation, programming code, ontology and knowledge maintenance. A project of an OO version is under specification.

City Search:

Table 1 presents data produced during test of co-operative recovering power of CepDigital software and Location's Mediator. Three cities were searched using different writings for their names: Belo Horizonte, São Thomé das Letras and Rio de Janeiro. Beyond the city name the State initials may or may not be informed. The table shows results carried out with each software. The test consisted in posing typical Consultation over location database using several variations in name writing, to check the semantics provided for each software to solve the problem co-operatively. In tables 1 and 2, "Y" stands for Yes, "N" stands for No, "BL" means Big List of items found for user choice (more than forty items), "SL" means Small List of items found for user choice (less than forty items).

City Names	Did CepDigital Find?		Did the Mediator Find?	
	Without State	With State	Without State	With State
BH	N	N	Y	Y
BHZ	N	N	Y	Y
B Horizonte	N	N	Y	Y
B Orizonte	N	N	Y	Y
Belo Horizonte	Y	Y	Y	Y
Belo	N	N	N	N
São Thomé das Letras	N	N	Y	Y
São Tomé das Letras	Y	Y	Y	Y
São Tomé Letras	N	N	Y	Y
São Tomé	BL	N	SL	SL
Rio	N	N	Y	Y
R de Janeiro	N	N	Y	Y
R Janeiro	N	N	Y	Y
R J	N	N	Y	Y
Rio de Janeiro	Y	Y	Y	Y
São Sebastião do Rio de Janeiro	N	N	Y	Y

Table 1: City's name Consultation and results returned from each application.

Location Search:
Table 2 shows test results with one street searched using different spelling for its original name in the fact database: Professor Aníbal de Matos. Beyond street's name, the State initials, city name, kind of street and quarter may or may not be informed. Test has consisted basically in posing typical Consultation over locality database using several variations in the name's spelling to check semantics provided for each software to solve the problem co-operatively.

State	City	Kind of Street	Street's Name	Quarter's Name	Did CepDigital find?	Did Mediator find?
—	—	—	Aníbal Matos*	São Pedro**	N	SL
—	—	St	Aníbal Matos	São Pedro	N	SL
—	—	Street	Aníbal Matos	São Pedro	N	SL
—	—	Street	Professor Aníbal Matos	São Pedro	N	Y
—	—	Street	Professor Aníbal de Matos	São Pedro	N	Y
—	—	Avenue	Prof.Aníbal Matos	São Pedro	N	Y
MG	Belo Horizonte	Street	Professor Aníbal de Matos	Santo Antônio	N	Y
MG	Belo Horizonte	Street	Prof Anïbal de Matos	Santo Antônio	N	Y
MG	Belo Horizonte	St	Professor Aníbal de Matos	Santo Antônio	Y	Y
MG	Belo Horizonte	Street	Professor Aníbal de Matos	— or S Antônio	N N	BL BL
MG	Belo Horizonte	Street	Professor	— or S Antônio	BL BL	BL SL
MG	Belo Horizonte	Street	Anïbal	— or S Antônio	N	SL

Table 2: Location name Consultation and results returned from each software under test

4. Conclusion

Context treatment, term synonyms, term range Consultations, multiple databases and users, multiple and simultaneous queries per user, approximate answers and schema abstraction are some aspects already implemented, tested and in use.

Several solutions for caching, selectiveness, concurrency, bottlenecks and combinatory explosion were added to the real scenario mentioned in section 4 ([Flecha, 1997] - shows details).

A high fan-in was achieved by location's mediator but due to non-object oriented programming used to implement it, the degree of encapsulation and re-use still can be better. This is the point where our research is being conducted. An improved co-operation might occur among datawarehouse, agents and mediation neighboring research areas. Standards like Corba may provide a more general basis for that. In the future, agent-mediators might be the integration element over the networks, increasing density and relevance of information through easy interfaces, able to hide complexity and sustain the decision making process to face changing speed.

References

[Bush, 1945] Vannevar Bush,. As We May Think. The Atlantic Monthly. July, 1945. p. 13. Ottawa, Canada: University of Ottawa. DUCHIER, Denys (ed). HTML version available in <www.isg.sfu.ca/~duchier/misc/vbush/vbush.all.shtml>

[Damaceno, 1997] Eduardo Teixeira Damaceno. *RIM Utilization Report.* Belo Horizonte, MG: Prodemge, DTP/STP/GPP, July 30th, 1997 (Performance Report, in Portuguese language).

[Fischer, 1994] Gerhard Fischer. Domain-Oriented Design Environments. *Automated Software Engineering,* p.177-203, 1994.

[Flecha, 1994] Mário L.O. Flecha. *Semantic Information Processing. Building a Relational Inference Machine -RIM - for Cognitive Systems' Development.* Belo Horizonte, MG: Prodemge, 1994, p.68. (Technical Report No. 25, in Portuguese).

[Flecha, 1997] Mário L.O. Flecha. *Semantic Processing on the Mediation of Public Database Querying.* MSc. Thesis. Belo Horizonte: Fundação João Pinheiro - Escola de Governo/Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, 1997. 223p. (in Portuguese).

[Genesereth *et al.*, 1995] Michael Genesereth *et al. Reference Architecture for the Intelligent Integration of Information Retrieval. Prepared by the Program on Intelligent Integration of Information (I^3).* Version 2.0 (draft). DARPA - Defense Advanced Research Projects Agency, August 2, 1995. 90p.

[Genesereth and Fikes *et al.*, 1992] Michael Genesereth and Richard E. Fikes *et al.*, *KIF - Knowledge Interchange Format. Version 3.0 Reference Manual.* Stanford, CA: Computer Science Department, Stanford University. (Report Logic-92-1) June 1992. 68p.

[Hart and Graham, 1997] Peter E. Hart and Jamey Graham. Query-free Information Retrieval. In: Co-operative Information Systems. *IEEE Expert.* p.32-37, Sep/Oct, 1997.

[Mylopoulos and Papazoglou, 1997] John Mylopoulos and Michael Papazoglou. Co-operative Information Systems. In: Guest Editors' Introduction. *IEEE Expert.* p.28-37, Sep/Oct, 1997.

[Quillian, 1968] Ross Quillian. Semantic Memory. In: *Semantic Information Processing*. Cambridge, Mass.: M.I.T. Minsky, Marvin (ed). Ed. The MIT, 1968, p.227-270.

[Rumbaugh, 1994] J. Rumbaugh. et al. *Modelagem e Projetos Baseados em Objetos*. Trad. Dalton Conde de Alencar, Rio de Janeiro, Rio de Janeiro: Ed. Campus, 1994, 654p. (Translation of: Object-oriented modeling and design. Englewood Cliffs, New Jersey: Ed. Prentice-Hall, 1991).

[Shneiderman, 1993] Ben Shneiderman – University of Maryland. *Designing the User Interface – Strategies for Effective Human-Computer Interaction.* 2nd ed. Massachusets, MA: Ed. Addison-Wesley, 1993, 573p.

[Ullmann, 1977] Stephen Ullmann. *Semântica - Uma Introdução à Ciência do Significado.* 4.ed. (Portuguese). Trad. J. A. Osório Mateus. Lisbon, Portugal: Fundação Calouste Gulbenkian, 1977. 578p. (Translation of: Semantics - An Introduction to the Science of Meaning. Oxford: Ed. Basil Blackwell, 1964).

[Webber, 1986] Bonnie L. Webber, Questions, Answers and Responses: Interacting with Knowledge Base Systems. In: *On Knowledge Base Management Systems. Integrating Artificial Intelligence and Database Technologies – Topics in Information Systems*. Michael L. Brodie/John Mylopoulos (ed). Ed. Springer-Verlag, 1986, p.365-402.

[Wiederhold and Genesereth, 1995] Gio Wiederhold and Michael Genesereth. Basis for Mediation. May 1995, In: PROC. INTERNATIONAL CONFERENCE ON CO-OPERATIVE INFORMATION SYSTEMS (CoopIS95). Vienna, Austria: available in <coopis@cs.toronto.edu>, May 1995. p.138-155.

[Wiederhold and Genesereth, 1997] Gio Wiederhold and Michael Genesereth. The Conceptual Basis for Mediation Services. In: Co-operative Information Systems. *IEEE Expert.* p.38-47, Sep/Oct, 1997.

[Wiederhold, 1996] Gio Wiederhold. Foreword and Glossary: Intelligent Integration of Information, in: *Intelligent Integration of Information.* Norwell, Mass.: Gio Wiederhold (ed). Ed. Kluwer Academic Publishers: p.5-9, p.193-201 May/June, 1996. (Journal of Intelligent Information Systems, Integrating Artificial Intelligence and Database Technologies, 6(2/3)).

[Wiederhold, 1992] Gio Wiederhold. Mediators in the Architecture of Future Information Systems. *Computer,* p.38-49, 1992.