Data Integration Using Web Semantic for Searching of Persons at Heterogeneous Information Systems

Cesar Enrique Rose-Gomez[1], José A. Pacheco-Sánchez[2], Gilberto Gradías-Enriquez[2]

Instituto Tecnologico de Hermosillo[1]
Procuraduria General de Justicia del Estado de Sonora[2]
Contact:crose@ith.mx
Paper received on 11/08/08, accepted on 06/09/08.

Abstract. In this paper, a framework for data integration at heterogeneous information systems is presented. The framework is strongly based in the use of web semantic technology. A practical approach to creating mappings between a relational database schema and OWL ontology is presented. Also, the framework includes the use of semantic association and natural language processing to obtain the data integration for a criminality information domain.

1 Introduction

Today the main goal in data and information retrieval systems has been to provide efficient support for the querying and retrieval of data. However, one of the primary obstacles in the information integration applications is the heterogeneity of the distributed data sources. Multiple types of heterogeneity characterize data from multiple sources. The following hierarchy is often used: (i) syntactic heterogeneity; is a result of differences in representation format of data, (ii) schematic or structural heterogeneity; the native model or structure to store data differ in data sources leading to structural heterogeneity, (iii) semantic heterogeneity, differences in interpretation of the meaning of data are source of semantic heterogeneity, (v) system heterogeneity; use of different operating system, hardware platforms lead to system heterogeneity.

Traditionally, the solution for the information integration has been the following general approaches [1]: Federated databases [2, 3] where the sources are independent, there is some global view or schema of the federation of databases that is shared by the applications. Data warehousing, a collection of decision support technologies, a data warehouse is a collection of information as well as a supporting system. Data warehouses are optimized for data retrieval, not routine transaction processing. The data warehousing process includes possible cleaning and reformatting of data before it's warehousing. Mediation, a mediator is a software component that supports a virtual database; the user can consult as if outside materialized (just as a data warehouse). The mediator does not store data properly; in its place he transfers the query of the user in one or more queries to the sources. The mediator synthesizes the answer to the query of the user of the answers of sources and returns an answer to the user.

Advances in Intelligent and Information Technologies.
Research in Computing Science 38, 2008, pp.419-429.

Up to date, some of the current work in data integration research concerns the semantic integration problem. This problem is not about how to structure the architecture of the integration, but how to resolve semantic conflicts between heterogeneous data sources. A common strategy is the use of ontologies, this approach is called ontology based data integration. Ontology based data integration involves the use of ontologies to effectively combine data and/or information from multiple heterogeneous sources.

A sector with evident problems of information integration is the governmental. The explosive growth in the digital information stored in the data repositories in local, state and federal organizations, and the urgent necessity of interagency access to that information has caused problems or difficulty in its recovery and analysis. Specific examples are the police and justice information systems. We can find systems for the recognition and identification of people through their general data and diverse biometrics such as iris, facial patterns, voice, hand measurements or fingerprints. The sources of data to make such processes are own and in some cases external to the institution, these sources are distributed and heterogeneous, a concrete example is shown in Fig. 1, where we can see databases of different government agencies, for example; the Resides database contains the general data about persons, the Licencia database contains the data of persons that have a driver license, the SAP database contains data of persons with penal antecedents, the AFIS database contains biometric data of persons, the Padron Vehicular database contains the data of motor vehicles and the SIAMP database contains the data of the Ministerio Publico agencies.

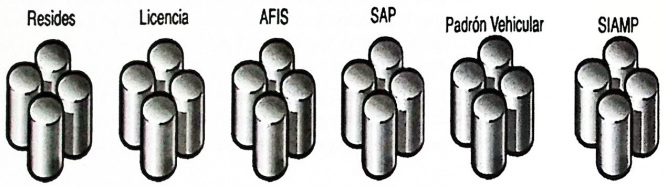

Fig. 1. Distributed and heterogeneous sources

Hence, the queries are made on each database as shown in Fig. 2, which does not allow having knowledge about the interrelation of the individuals in the other databases, causing that the identification process is inefficient. For instance, one person has different names in several databases, but their fingerprints are the same in the databases, therefore, the identification process is slower. Fig. 2 also shows the process for the identification of persons.

In [4] reviews some of the key research in information integration, theory and systems, describes the current state-of-the-art in commercial practice, and the challenges faced by chief information officers (CIOs) and application developers. Also in [5] we can found a tutorial that provides an overview of some of issues underlying the theory of data integration.

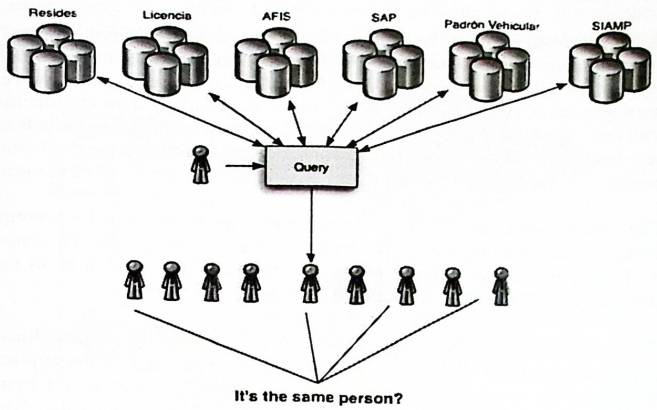

Fig. 2. Person identification process

The semantic integration of the information has acquired a great importance, due to the availability of an increasing number of data sources. Generally, these sources are distributed and heterogeneous, and in rare occasions, partially they are only translated to the new "franca lingua" of the information: XML [6]. This generates significant difficulties; nevertheless the potential applications are many. The works in the field of the Semantic Web [7,8] it is a representative example. The technologies developed around this vision of the Web can take advantage of the organizations with a need to combine diverse sources of information. Among the numerous projects that are being realized in the field of semantic integration [9,10] they observe some clear tendencies: on the one hand, the approach to data expressed in XML; on the other hand, the use of the ontologies like the global scheme on which the user realizes the queries and, generally, like the tool fundamental to express a scheme integrated conceptually.

The use of ontologies in the integration of data can be found in [11,12], [13] show as the ontologies expressed in RDFS, a semantically rich schema language, permit the bridge across syntactic, schematic, and semantic heterogeneities in data sources. [13] also present different study cases. In [14] reports the achievements on ontology-based heterogeneous data integration, and discuss the fundamental issues, including metadata representation, mapping process, and query processing, in several approaches to different applications of data integration.

2 Semantic data integration using ontologies

Data integration provides the ability to manipulate data transparently across multiple data sources. The two most important approaches for building a data integration system are Global-as-View (GaV) and Local-asView (LaV). Data sources can be heterogen ous in syntax, schema, or semantics, thus making data interoperation a difficult task. Syntactic heterogeneity is caused by the use of different models or languages. Schematic heterogeneity results from structural differences. Semantic heterogeneity is caused by different meanings or interpretations of data in various contexts.

The semantic data integration is the process of using a conceptual representation of the data of their relationships to eliminate possible heterogeneities. The important point of semantic data integration is the concept of ontology, which is an explicit specification of a shared conceptualization.

There are several definitions of ontology [15,16,17], from the philosophical perspective to the Artificial Intelligence (AI) perspective depending on the application. In philosophy, the word ontology means a theory about the nature of the being, or the types of existence; nevertheless, in AI we found definitions like Gruber [18] that establishes that ontology is a formal specification of a conceptualization for systems of AI, in which what "exists" is what can be represented. A more constructive definition is given by Noy and McGuinness [19] in which ontology is a formal explicit description of concepts in a domain of discourse (classes or concepts), with properties of each concept describing various features and attributes of the concept (slots, or roles, or properties) and with restrictions on slots (facets or role restriction). Uschold and Jasper establish in [20] that ontology can take several forms, but that it should necessarily have to include a vocabulary of terms and some specification of their meaning. This includes definitions and an indication of how the concepts are interrelated; this interrelation collectively imposes a structure in the domain and restricts the possible interpretation of terms. The global schema in a data integration system may be an ontology, which then acts as a mediator for reconciliating the heterogeneities among different sources. An approach in the use of ontology is that all source schemas are directly related to a shared global ontology that provides a uniform interface to the user. However, this approach requires that all sources have nearly the same view on a domain, with the same level of granularity. This approach is appropriate for GaV systems. Fig. 3 shows the architecture of our proposed semantic integration system to realize the identification of people using heterogeneous databases.

Ontology provides an explicit model obtained by consensus and it is described in a language that contains the concepts, properties and relations of a domain. There are diverse languages to define ontologies [21]. For example, SHOE, XML, RDF, DAML, OIL, DAML+OIL and OWL. Fig. 4 shows partially the ontology used in our proposal, the domain of discourse for the ontology is the criminal.
The ontology has been constructed and modeled using the Protégé tool. The language selected for its representation is OWL [22], since it allows the definition of

the concepts, relations and properties. OWL is a very expressive language that allows establishing the suitable semantics and is possible to use it with the Jena framework [23] that allows the use of the ontology and the accomplishment of the searches to create applications of the Semantic Web.

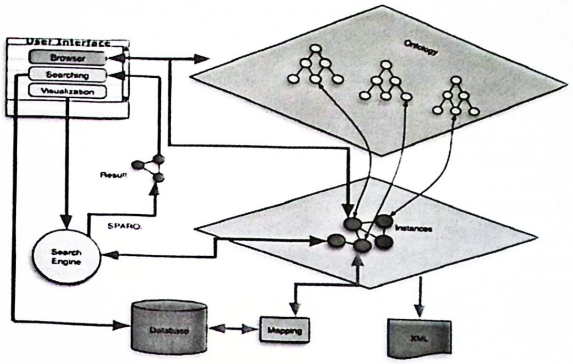

Fig. 3. Architecture of semantic integration system

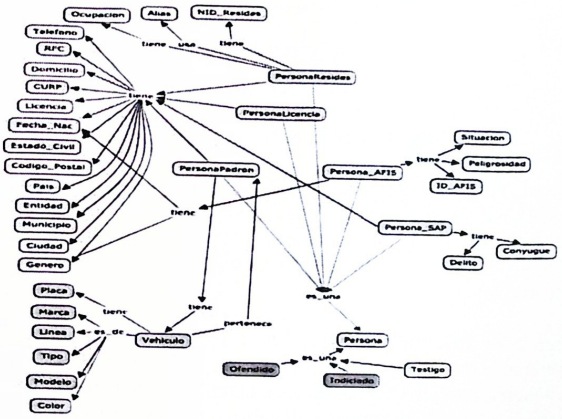

Fig. 4. Partial ontology

3 Mapping the data in relational databases to RDF graph

One of the most important aspects in the design of a data integration system is the specification of the correspondence between the data at the sources and those in the global schema. Such a correspondence is modeled through the notion of mapping. The user can formulate a query Q_{user} without specific knowledge of the different data sources; this query is carried out through a mediator to execute it in each of heterogeneous databases. This mediator basically makes a mapping of the schemes of the relational databases to a global scheme that is being supported by the ontology. Through this mapping we get the instances in a RDF graph [22]. Fig. 5 shows this mapping, where Answer = Mediator(Q_{user}), we can consider Q_{user} = < q_1, q_2, q_3,.., q_n >, where q_i is a query at the database i, we get a RDF graph for each q_i. Therefore, Answer is <$grdf_1$,$grdf_2$,$grdf_3$,..,$grdf_n$>, where $grdf_i$ is a RDF graph.

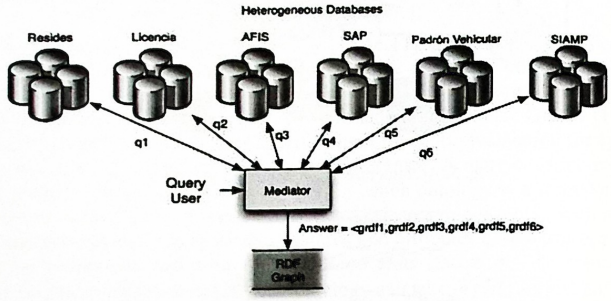

Fig. 5. Mapping databases and RDF

Definition 1. From [24], a relation r of the relation schema $R(A_1,A_2,\ldots,A_n)$, also denoted by r(R), is a set of n-tuples $r=\{t_1,t_2,..,t_m\}$. Each n-tuple t is an ordered list of n values $t=<v_1,v_2,\ldots v_n>$, where each value v_i, $1<=i<=n$, is an element of $dom(A_i)$ or is a special null value. The i^{th} value in tuple t, which corresponds to the attribute A_i, is referred to as $t[A_i]$.

Definition 2. An RDF triple contains three components: (1) Subject, which is an RDF URI reference or a blank node, (2) Predicate, which is an RDF URI reference, (3) Object, which is an RDF URI reference, a literal or a blank node.
An RDF graph is a set of RDF triples: <rdf:subject, rdf:predicate, rdf:object>

When the user do a query using some concept from the ontology O, for example the owl.name, we can get the values associated to the value of *name* using the query:
$\{ t \mid \exists u \in Resides, (t.name = u.name \wedge t.address = u.address \wedge t.RFC = u.RFC\}$
where t[name], t[address] and t[RFC] are instances for the classes owl.name, owl.address and owl.RFC of the ontology O.

Such that we get a RDF graph, where the triples of the graph is build using the following:
<uri:t[name], uri2:owl#name, t[name]>
<uri:t[name], uri2:owl#address, t[address]>
<uri:t[name], uri2:owl#RFC, t[rfc]>
With this mapping we are solving the syntactic heterogeneity problem that is caused by differences in data format representation. Also, we are solving the schematic heterogeneity, which results from structural differences.

4 Eliminating heterogeneities with OWL reasoning

In our study case we are looking for a person that could be found at different databases, but we can find diverse heterogeneities that do not permit us to know if it is the same person. For instance, the alias of a person could be *The Black Dog* in a database but in another one the alias is *Dog*. In a traditional search we cannot know if are the same values because the syntactic match not is success, however, semantically a *Black Dog* is a *Dog*.

In each RDF graph that we get from the queries we have the data of the persons that are closed to the person that is looking for. For the data integration we use the instances in the RDF graph and the ontology, also we have rules for reasoning as the following:

Persona_Licencia ∪ Persona_Padron ⇐ EquivalentPerson(Licencia_Person,Padron_Person).

EquivalentPerson(Licencia_Person.Padron_Person) ⇐
Homony(Licencia_Person, Padron_Person) ∧
(EquivalentAttribute(Licencia_Person.Licencia, Padron_Person.Licencia) ∨
EquivalentAttribute(Licencia_Person.RFC, Padron_Person.RFC) ∨
EquivalentAttribute(Licencia_Person.CURP, Padron_Person.CURP)).

Homony(X,Y) ⇐ O.Class(X) ∧ O.Class(Y) ∧
(Equivalent(X.Name, Y.Name) ∨
(X ⊂ O.Person ∧
Y ⊂ O.Person ∧ Equivalent(X.Name, Y.Name)) ∨
(Similar(X.Name,Y.Name) ∨
(X ⊂ O.Person ∧
Y ⊂ O.Person ∧ Similar(X.Name, Y.Name))).

EquivalentAttribute (X,Y) ⇐ Equivalent(X,Y) ∨ Similar(X,Y).

Equivalent(X,Y) ⇐ X = Y.
Similar(X,Y) ⇐ Certain_Factor(X,Y) >= 90.

The Fig. 6 shows the idea of the data integration for the identification of persons. The input data could be the person name, the car license, etc., or some biometrics such as a fingerprint, face image, etc.

When the input data is a string such that the person name, we use pattern matching to decide about similarity or string closed pattern matching. The similarity concept could be defined in several ways, so we use the Levenshtein distance. This distance between strings is the minimum number in insertion, delete and replacement of characters necessary to make them the same. Also we use regular expressions to find some pattern in the string. The above permit us find the similarities among the names, for example as: Joe Doe, Joe Doe II, Joe Does, together with others attributes in the pattern marching could be inferred that is the same person.

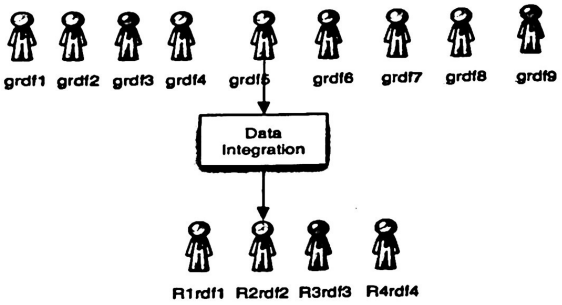

Fig. 6. Personal data integration

When another attributes are used such that the biometrics, the rule for the inference could be stronger because the pattern matching is safe, for instance, the fingerprints.

The data integration is obtained when the established rules are fulfilled in such a way that a resulting RDF graph is obtained that considers the union of the attributes for each of the persons who have been that they are the same. For example, it considers from the Fig. 6 that grdf1, grdf5 and grdf8 contains persons who are the same (person1), but grdf2 and grdf7 contain another person (person2), and so on, then:

$R1rdf = grdf1 \cup grdf5 \cup grdf8$
$R2rdf = grdf2 \cup grdf7$
$R3rdf = grdf3 \cup grdf9$
$R4rdf = grdf4 \cup grdf6$

Where R1rdf contains the union of the attributes of person1, only exists a resource for this person, R2rdf contains the union of the attributes of person2, R3rdf contains the union of the attributes of person3 and R4rdf contains the union of the attributes of person4.

Finally, the union with the rest of RDF graphs that do not correspond to person1 is obtained, in this case person2, person3 and person4.
resutFinalRDF = R1rdf ∪ R2rdf ∪ R3rdf ∪ R4rdf
where resultFinalRDF contains the data integrated of each person.

5 Experimental Results

The framework has been tested with the databases of the *Procuraduría General de Justicia del Estado de Sonora*, Jena was used in the implementation of the framework, for the test we used twenty cases of persons with characteristics previously know, for example, that the person had registers in several databases, that the person had several registers in the same database, persons with the same fingerprint but with different names, etc., we can say that the results obtained were correct, the merge was correct for the test with these cases, but we can not say that it is a result at 100%, because we need to use more cases in the test, unfortunately the obtained result couldn't be shown using real data, since the data are private and for security reasons it is not possible to publish them. We got the data integrated of the person and can be visualized as hyperbolic tree, or with several interfaces, and it is possible generate a XML file, at the Fig. 7 we show a XML structure, and a hyperbolic tree.

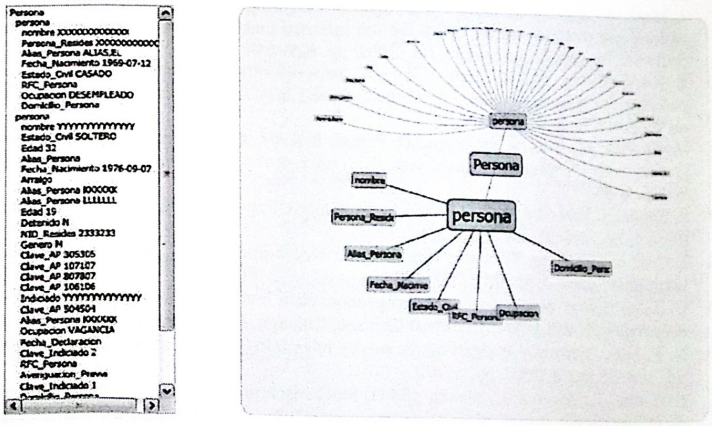

Fig. 7. Result of the integration

6 Conclusions

A framework for the integration of the semantic information for the identification of persons has been applied to the criminality domain using technology of Se-

mantic Web. The study case used has allowed to show the viability of the solution of the problem search when heterogeneous databases are maintained and to recover the integrated information of a person.

References

1. H. Doan and A. Y. Halevy, "Semantic integration research in the database community: A brief survey," AI Magazine, vol. 26, pp. 83 - 94, 2005.
2. H. Garcia-Molina, Database Systems: The Complete Book. New Jersey USA: Prentice Hall, 2002.
3. R. M. Colomb, "Impact of semantic heterogeneity and federating databases," Computer Journal, vol. 40, no. 5, pp. 235–244, 1997.
4. L. Haas, "Beauty and the beast: The theory and practice of information integration," I. 2007, Ed., 2007, pp. 28–43.
5. M. Lenzerini, "Data integration: A theoretical perspective," P. of the 21st ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS 2002), Ed., 2002, pp. 233–246.
6. R. Tous, "Data integration with xml and semantic web technologies," Ph.D. dissertation, Universitat Pompeu Fabra, Barcelona, España, 2006.
7. G. Antoniou and F. van Harmelen, A Semantic Web Primer. Cambridge, USA: The MIT Press, 2004.
8. M. C. Daconta, L. J. Obrst, and K. T. Smith, The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management. New Jersey, USA: Wiley, 2003.
9. J. Machado and J. Fernandes, "Heterogeneous information systems integration: Organizations and methodologies," P. of the 4th International Conference on Product Focused Software Process Improvement, Ed., 2002, pp. 629–644.
10. Z. Xu, S. Zhang, and Y. Dong, "Mapping between relational database schema and owl ontology for deep annotation," P. of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, Ed., 2006.
11. V. Alexiev, M. Breu, J. de Bruijn, D. Fensel, R. Lara, and H. Lausen, Information Integration with Ontologies: Experiences from an Industrial Showcase. New Jersey USA: Wiley, 2005.
12. P. Spyns, R. Meersman, and M. Jarrar, "Data modelling versus ontology engineering," S. Record, Ed., vol. 31, no. 4, 2002, pp. 12–17.
13. F. Cruz and H. Xiao, "The role of ontologies in data integration", Journal of Engineering Intelligent Systems, pp. 245–252, 2005.
14. H. Xiao, "Query processing for heterogeneous data integration using ontologies," Ph.D. dissertation, University of Illinois at Chicago, Chicago, Illinois, 2006.
15. N. F. Noy, "Semantic integration: A survey of ontology-based approaches," S. Record, Ed., vol. 33, no. 4, 2004, pp. 65–70.
16. H.Wache, T. Vogelea, U. Visser, and H. Stuckenschmidt, "Ontology- based integration of information-a survey of existing approaches," I. -. W. Ontologies and I. Sharing, Eds., 2001.
17. C. Calero, F. Ruiz, and M. Piatini, Ontologies in Software Engineering and Software Technology. Berlin, Germany: Springer, 2006.
18. T. R. Gruber, "1993," in A Translation Approach to Portable Ontology Specification, K. Acquisition, Ed., vol. 5, no. 2, 1993, pp. 199–220.
19. N. Noy and D. L. McGuinness, "Ontology development 101: A guide to creating your first ontology," University of Stanford, Technical Report SMI-2001-0880, 2001.

20. M. Uschold and R. Jasper, "A framework for understanding and classifying ontology applications," P. of the IJCAI99 Workshop on Ontologies and P.-S. M. (KRR5), Eds., Stockholm, Sweden, 1999.
21. A. Gomez, M. Fernandez, and O. Corcho, Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. Berlin, Germany: Springer, 2004.
22. W. Lacy, Owl: Representing Information Using the Web Ontology Language. BC, Canada: Trafford Publishing, 2005.
23. J. J. Carrol, I. Dickinson, C. Dollin, D. Reynolds, A. Seaborne, and K. Wilkinson, "Jena: Implementing the semantic web recommendations," P. of the 13th International World Wide Web, Ed., New York, USA, 2004, pp. 74–83.
24. R. Elmasri and S. Navathe, Fundamentals of Database Systems. USA: Addison-Wesley, 2000.