Uniform management of heterogeneous semi-structured information sources

Giorgio Terracina
D.I.M.E.T. - Università di Reggio Calabria
Via Graziella, Località Feo di Vito
89100 Reggio Calabria, Italy
e-mail:terracina@ing.unirc.it

Abstract

Nowadays, data can be represented and stored by using different formats ranging from non structured data, typical of file systems, to semi-structured data, typical of Web sources, to highly structured data, typical of relational database systems. Therefore, the necessity arises to define new tools and models for uniformly handling all these heterogeneous information sources. In this paper we propose both a framework and a conceptual model which aim at uniformly managing information sources having different nature and structure for obtaining a global, integrated and uniform representation. We show also how the proposed framework and the conceptual model can be useful in many application contexts.

1 Research question

In the last years the information available electronically has dramatically increased; moreover, in the same period, a growing variety of models and languages for representing and manipulating this information has been proposed.

Indeed, nowadays, data can be represented and stored by using different formats ranging from non structured data, typical of file systems, to highly structured data, typical of relational database systems. In more detail, sometimes data do not have any structure (as an example images and sounds); sometimes they have an implicit structure which must be derived by analyzing them; sometimes a rigid and regular structure exists.

We call "semi-structured data" [1, 7, 17] all those data such that they have some structure but this is not regular such as that of databases. In the last years the development of the Internet led information system researchers to deeply study semi-structured information sources since data on the Web are generally of this nature.

Traditional systems for managing information sources appear to be inappropriate for: *(i)* handling the enormous quantity of data typical of the Web, *(ii)* managing information sources that do not have a precise structure and, finally, *(iii)* guaranteeing the cooperation and the uniform treatment of both structured and semi-structured information sources. Therefore, innovative tools, capable to face all the problems described above, appear to be compulsory.

All these tools should take into account the semantics of involved information sources; in order to derive it, the necessity arises to identify terminological and structural properties existing among object classes or subschemes belonging to different information sources (*interscheme properties*) [2, 4, 10]. Since the number and the dimension of information sources are great and since the information they store change quite frequently over time, manual extraction of interscheme properties is expensive and difficult; therefore, the necessity arises of designing semi-automatic tools capable to carry out this task. Finally, for uniformly managing information sources having different nature and structure, it appears mandatory the exploitation of a conceptual model capable of uniformly representing the various typologies of sources.

All these considerations led us to define a new framework for uniformly and semi-automatically handling information sources having great dimensions and different nature and structure. In more detail, the proposed framework consists of three steps:

- The representation of involved information sources through a new, graph-based, conceptual model, called SDR-Network, which can be exploited for defining a metrics allowing to derive and represent the semantics of each involved information source (*intrasource semantics*).
- The exploitation of both the conceptual model and the corresponding metrics for extracting interscheme properties which allow to define the relative semantics of involved information sources (*intersource semantics*).
- The exploitation of both intrasource and intersource semantics for obtaining an integrated and uniform representation of involved information sources.

The availability of both a conceptual model capable to uniformly representing heterogeneous information sources and an integrated representation of involved information sources can be useful in several application contexts such as data conversion, message exchange in E-commerce applications, semantic query processing and, finally, for designing advanced Web search engines.

To the best of our knowledge, SDR-Network is the first proposed conceptual model well-suited to *represent and manipulate* with intra- and inter-source semantics of information sources having different structures and natures. Lore's XML data model [7] is an interesting previous proposal of a model for *representing* XML documents, well suited for *representing* also OEM and E/R sources. However, this model has been conceived with different purposes than our model. In particular, it is supposed to serve querying tasks, rather than the task of representing and manipulating with intra- and inter-source semantics.

Analogously, also the model proposed by [6] for *representing* XML documents can be extended to represent also OEM and E/R sources. However, as for Lore's

XML data model, the model of [6] aims at supporting source querying and not to represent and handle the semantics of the information sources.

In the sequel of this paper we present the three steps of the framework and, finally, we give an overview of some possible applications. Due to space constraints we cannot describe all details of the various steps of the framework. However, the interested reader can find them in [12, 13, 15, 16, 18].

2 The conceptual model

In order to derive a conceptual model for uniformly representing information sources having different nature and structure we must face syntactic and semantic difficulties.

In particular, *syntactic difficulties* arise because models for representing semi-structured data are various and complex; moreover, they generally represent instances, i.e. extensional data, whereas the metrics we must define for deriving source semantics (and consequently the conceptual model which it is based on) refers to object classes, i.e. intensional data; therefore, the necessity arises to derive the object classes associated to instances represented by classical semi-structured data models.

Semantic difficulties arise because, in semi-structured information sources, the various objects of the same class can be described by different properties; in other words, a certain property could be present only in some objects belonging to a class.

The proposed conceptual model can be obtained by associating to each information source IS a network $Net(IS)$, called SDR-Network (Semantic Distance-Relevance Network). This can be represented as:

$$Net(IS) = \langle N(IS), A(IS) \rangle$$

where $N(IS)$ denotes the set of nodes and $A(IS)$ represents the set of arcs. In more detail, each node is associated to a concept and is characterized by a name[1]; each arc can be represented by a triplet $\langle x, y, l_{xy} \rangle$, where x is the "source" node, y is the "target" node and l_{xy} is a label associated to the arc. l_{xy} is composed, in its turn, by a pair $[d_{xy}, r_{xy}]$, where both d_{xy} and r_{xy} belong to the real interval $[0, 1]$. d_{xy} is the *semantic distance coefficient*; it indicates the capability of the concept associated to y to characterize the concept associated to x. r_{xy} is the *semantic relevance coefficient*; it indicates the participation degree of the concept expressed by y in defining the concept associated to x. Intuitively, this participation degree indicates the fraction of instances of the concept associated to x whose complete definition requires at least one instance of the concept denoted by y.

An example of an SDR-Network is shown in Figure 1; it is derived from the set of Italian Central Government Office Information Sources and represents the management process for projects supported by the European Social Funds. In the figure, in order to simplify the layout, a dotted node having name x is used to indicate that the arc incident onto x must be considered incident onto the corresponding solid

[1]Since a node represents a concept, i.e. an object class, from now on, we use the terms "node", "concept" and "object class" interchangeably.

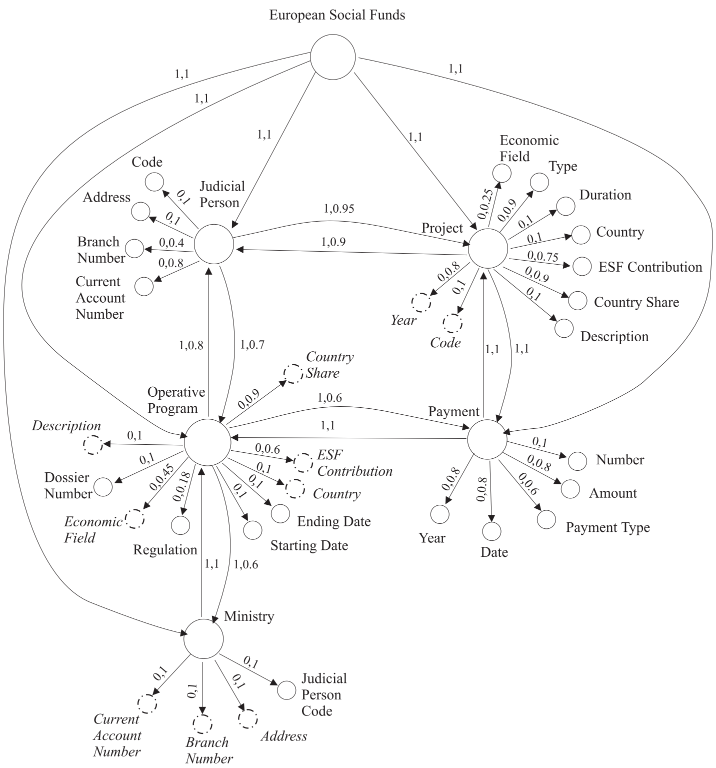

Figure 1: The SDR-Network of the European Social Funds information source

node having the same name. SDR-Network nodes such as *Judicial Person*, *Project*, *Payment*, etc. represent the involved concepts (i.e., they denote the involved object classes). The arc $\langle$*Judicial Person*, *Project*, $[1, 0.95]\rangle$ indicates that 95% of judicial persons are involved in some project. The other arcs have an analogous semantics.

Note that, basically, any information source can be represented as a set of (possibly structured) concepts and a set of relationships among concepts. Since SDR-Network nodes and arcs are well suited to represent such concepts and relationships, the SDR-Network can be used to uniformly model most existing information sources. In this respect, semantic preserving translations have been provided from some interesting source formats, such as XML, OEM and E/R to SDR-Network [13, 18].

Both the semantic distance and the semantic relevance define the *intrasource semantics* of an information source; in addition they can be exploited for deriving interscheme properties (which contribute to define the *intersource* semantics). A suitable metrics can be defined on the SDR-Network for measuring the semantic distances and relevances between object classes corresponding to nodes of the SDR-Network which are not directly connected by an arc. The metrics is based on the concepts of Path Semantic Distance, Path Semantic Relevance, D-Shortest-Path, CD-Shortest-Path, D_Path_n and neighborhood.

The *Path Semantic Distance* of a path P in an SDR-Network $Net(IS)$ is the sum of the semantic distance coefficients associated to the arcs of P. The *Path Semantic Relevance* of a path P in an SDR-Network $Net(IS)$ is the product of the semantic relevance coefficients associated to the arcs of P. The *D-Shortest-Path* between two nodes N and N' in $Net(IS)$ is the path having the minimum Path Semantic Distance among all paths connecting N and N'. If more then one path exists having the same minimum Path Semantic Distance we choose, as D-Shortest-Path, one among those having the maximum Path Semantic Relevance. The *CD-Shortest-Path* (Conditional D-Shortest Path) between two nodes N and N' in $Net(IS)$ comprising an arc A, is the path having the minimum Path Semantic Distance among all paths which connect N and N' and comprise A. As in the previous definition, if more than one path exists having the same minimum Path Semantic Distance, one among those having the maximum Path Semantic Relevance is chosen as the CD-Shortest-Path. The D_Path_n is a path P in $Net(IS)$ such that the Path Semantic Distance of P is greater than or equal to n and lesser than $n+1$. The *i-th neighborhood* of a class x (denoted by $nbh(x,i)$) consists of the set of arcs $\langle z, y, l_{zy} \rangle$ such that: *(a)* they do not belong to any neighborhood lesser then i; *(b)* a CD-Shortest-Path P between the node of $Net(IS)$ associated to x and y comprising $\langle z, y, l_{zy} \rangle$ exists such that it is a D_Path_i.

3 Extraction of interscheme properties

In this section we show how the conceptual model and the metrics defined above can be exploited for extracting interscheme properties. These can be classified into nominal properties, sub-source similarities and assertions between knowledge patterns.

Nominal properties are synonymies, hyponymies or homonymies. *A synonymy* between two object classes A and B indicates that they have the same meaning; a *hyponymy* from an object class A to an object class B denotes that A has a more specific meaning than B; B is the *hypernym* of A. An *homonymy* between two object classes A and B indicates that they have the same name but different meanings.

A *sub-source similarity* represents a similitude between fragments of different sources and corresponds, in the SDR-Network conceptual model, to a similarity between two sub-nets.

An *assertion between knowledge patterns* indicates either a subsumption or an equivalence between knowledge patterns. Roughly speaking knowledge patterns can be seen as views on involved information sources.

In the next three subsections, we describe the extraction of synonymies and homonymies, sub-source similarities and assertions between knowledge patterns.

3.1 Extraction of synonymies and homonymies

The technique for extracting synonymies and homonymies among concepts belonging to two semi-structured information sources IS_1 and IS_2 receives both the SDR-Networks associated to involved information sources and some lexical synonymies between names, stored in a Lexical Synonymy Property Dictionary $LSPD$[2]. It returns a Synonymy Dictionary SD and an Homonymy Dictionary HD. Derived

[2] All considerations regarding the construction of the LSPD and the associated problems can be found in [13, 18].

properties are fuzzy and are represented as triplets $\langle A, B, f\rangle$, where A and B are the involved concepts and f is a fuzzy coefficient, in the real interval $[0, 1]$, expressing the plausibility of the property.

The technique consists of two phases. The first one, for each pair of concepts $C_l \in IS_1$ and $C_m \in IS_2$, derives the so-called basic similarity between C_l and C_m. Basic similarities are rough properties taking into account only lexical similarities and the nearest neighborhoods of involved concepts; these properties are exploited as the starting point for deriving the real properties. Basic similarities are represented as triplets of the form $\langle C_l, C_m, f_{lm}\rangle$, where C_l and C_m are the concepts into consideration and f_{lm} is a coefficient, in the real interval $[0, 1]$, denoting the plausibility of the property; all basic similarities are stored in a *Basic Similarity Dictionary* BSD.

The second phase takes the BSD derived in the first phase as input and detects synonymies and homonymies between concepts of the information sources into consideration. First, the similarity degree associated to each tuple $\langle C_l, C_m, f_{lm}\rangle \in BSD$ is refined. Then, the set of significant synonymies (resp., homonymies) is constructed by selecting those pairs of concepts whose similarity degree is greater (resp., smaller) than a certain, dynamically computed threshold th_{Syn} (resp., th_{Hom}).

In order to refine the similarity coefficient associated to a tuple $\langle C_l, C_m, f_{lm}\rangle \in BSD$, we analyze both C_l and C_m and their neighborhoods $nbh(C_l, i)$ and $nbh(C_m, i)$, for each i such that $nbh(C_l, i) \neq \emptyset$ and $nbh(C_m, i) \neq \emptyset$. The influence of the similarity of neighborhoods of C_l and C_m on the similarity of C_l and C_m must be inversely proportional to their distance; in order to obtain this, a monotone decreasing weighting succession $\{p(i)\}$ is associated to the neighborhoods of C_l and C_m so that farthest neighborhoods have lightest weights.

Intuitively, the process of refining the similarity coefficient between C_l and C_m consists of the following steps: *(i)* Constructing, for each tuple $\langle C_l, C_m, f_{lm}\rangle \in BSD$, $nbh(C_l, i)$ and $nbh(C_m, i)$. *(ii)* Computing the similarity degree between $nbh(C_l, i)$ and $nbh(C_m, i)$ as an objective function associated to the maximum weight matching on a suitable bipartite weighted graph defined from nodes of both $nbh(C_l, i)$ and $nbh(C_m, i)$. *(iii)* Computing the overall similarity degree of C_l and C_m as a weighted mean of similarity degrees between the various neighborhoods of C_l and C_m; weights are the elements of the succession $\{p(i)\}$ described above.

3.2 Extraction of sub-source similarities

The proposed technique for deriving sub-source similarities takes in input a set of information sources represented by SDR-Networks and a Synonymy Dictionary SD and returns a Subscheme Similarity Dictionary SSD. As for synonymies and homonymies, derived properties are fuzzy and are represented by triplets $\langle SS_1, SS_2, f\rangle$, where SS_1 and SS_2 are the involved sub-sources and f is a fuzzy coefficient denoting the plausibility of the property.

Given an information source IS and the corresponding SDR-Network $Net(IS)$, the number of possible sub-sources that can be identified in IS is exponential in the number of nodes of $Net(IS)$. To avoid the burden of analyzing such a huge number of sub-sources, the proposed technique for deriving sub-source similarities first selects only the most *promising* ones according to the following rules:

- Given two information sources IS_1 and IS_2 and the corresponding SDR-Networks $Net(IS_1)$ and $Net(IS_2)$, it considers only those pairs of sub-sources $[SS_i, SS_j]$ such that $SS_i \in Net(IS_1)$ corresponds to a rooted sub-net having a node N_i as root, $SS_j \in Net(IS_2)$ corresponds to a rooted sub-net having a node N_j as root and N_i and N_j are synonyms.

- The selection of the most promising pairs of sub-sources, having the synonym nodes N_i and N_j as roots, is carried out by exploiting some information obtained during the application of the technique for computing the similarity degree between N_i and N_j described in the previous section. In particular, given a pair of synonym nodes N_i and N_j, the technique derives a *promising pair of sub-sources* $[SS_{i_k}, SS_{j_k}]$, for each k such that both $nbh(N_i, k)$ and $nbh(N_j, k)$ are not empty. SS_{i_k} and SS_{j_k} are constructed by determining the *promising pairs of arcs* $[A_{i_k}, A_{j_k}]$ such that $A_{i_k} \in nbh(N_i, l)$, $A_{j_k} \in nbh(N_j, l)$, for each l belonging to the integer interval $[0, k]$.

  A *pair of arcs* $[A_{i_k}, A_{j_k}]$ is considered *promising* if *(i)* the target nodes T_{i_k} of A_{i_k} and T_{j_k} of A_{j_k} are different from N_i and N_j; *(ii)* an edge between the target nodes T_{i_k} of A_{ik} and T_{j_k} of A_{jk} is present in the maximum weight matching computed on a suitable bipartite graph constructed from the target nodes of the arcs of $nbh(N_i, l)$ and $nbh(N_j, l)$, for some l belonging to the integer interval $[0, k]$; *(iii)* the similarity degree of T_{i_k} and T_{j_k} is greater than a certain given threshold.

The rationale underlying this approach is that of constructing promising pairs of sub-sources such that each pair is composed by the maximum possible number of pairs of concepts whose synonymy has been already stated. In this way it is probable that the overall similarity degree, resulting for each promising pair of sub-sources, will be high.

The second step of the technique for deriving sub-source similarities consists in deriving the similarity degree associated to each pair of promising sub-sources; this is determined by computing the objective function associated to the maximum weight matching defined on a suitable bipartite graph, constructed from the concepts composing the sub-sources of the pair. The exploitation of the maximum weight matching as the main step for the computation of the similarity between two sub-sources $SS_i \in IS_1$ and $SS_j \in IS_2$ is justified by observing that SS_i (resp., SS_j) can be considered similar to SS_j (resp., SS_i) only if it possible to determine a set of concepts belonging to SS_i (resp., SS_j), each of which being synonym with one of the concepts of SS_j (resp., SS_i). The maximum weight matching is exploited for selecting this set.

3.3 Extraction of complex assertions between knowledge patterns

The approach for extracting complex assertions between knowledge patterns takes in input a set of semi-structured information sources and the set of interscheme properties holding among them, derived by the techniques described in the previous sections.

For both representing complex information and inferring *complex* implicit correlations, a particular Description Logic, called DL_P, has been proposed [3, 5, 11]. Here we use the SDR-Network model and DL_P for inferring complex correlations among data encoded in information sources having different nature and structure.

The adoption of a formalism derived from Description Logics is motivated by two main reasons. First, Description Logics are well suited for representing complex semantic properties of represented domains. Second, DLs feature a precise formal inference system, of which we take advantage as the basis for deriving complex assertions.

Complex assertions derived by our approach are of the form $L_1 \dot{\leq}_{W_{\langle L_1, L_2 \rangle}} L_2$, where L_1 and L_2 are class expressions of DL_P and $W_{\langle L_1, L_2 \rangle}$ is a coefficient in the real interval [0,1]; in particular, $W_{\langle L_1, L_2 \rangle}$ represents the plausibility that the converse inclusion $L_2 \subseteq L_1$ (and, consequently, $L_2 = L_1$) holds as well. In other words $W_{\langle L_1, L_2 \rangle}$ represents the proportion of instances of L_2 which are also instances of L_1. In particular, if $L_1 \dot{\leq}_1 L_2$ then all instances of L_2 are instances of L_1.

The approach consists of two main phases: *(i) Pre-processing phase* (Phase 1), where information sources, represented by SDR-Networks, are analyzed for extracting an initial set of assertions. In particular, first SDR-Networks are translated into DL_P assertions, therefore obtaining a set of *intrasource assertions*; then the set of interscheme properties holding among the information sources is analyzed for obtaining basic assertions relating object classes belonging to different information sources (*intersource assertions*). In this phase human domain experts can provide other assertions which cannot be derived automatically from the structure of SDR-Networks and which encode knowledge the domain experts claim to hold for the given application. *(ii) Discovering complex assertions* (Phase 2), which is devoted to discover properties involving, in general, complex class expressions. This phase works by case analysis, considering various possible combinations of basic assertions.

4 Construction of the integrated and uniform representation of information sources

In order to obtain an integrated and uniform representation of involved information sources it is necessary to activate an integration procedure. This exploits the previously derived interscheme properties and the SDR-Networks associated to involved information sources and constructs a global SDR-Network. In particular, involved SDR-Networks are juxtaposed for obtaining a (temporarily redundant and, possibly, ambiguous) global SDR-Network SDR_G. SDR-Networks are rooted labeled graphs. Therefore, one operation to be done is to obtain only one root in SDR_G. In particular, if the roots of the two SDR-Networks in input are synonyms, they must be merged; otherwise, a new root node is created which is connected with the roots of the two SDR-Networks.

After this, the algorithm exploits derived synonymies, homonymies and sub-source similarities for normalizing SDR_G. In particular, it first determines which nodes must be assumed to coincide, to be completely distinct or to be renamed and carries out the corresponding transformations on nodes and arcs of SDR_G. Then, it determines which sub-sources must be merged.

In more detail, for each tuple $\langle N_x, N_y, f\rangle$ involving the nodes N_x and N_y and belonging to the Synonymy Dictionary SD, N_x and N_y must be considered coincident in the integrated SDR-Network and, therefore, must be merged into a new node N_{xy}. If two nodes are homonyms, they must be considered distinct in the integrated SDR-Network and, consequently, at least one of them must be renamed.

After that all nodes have been examined, the algorithm normalizes the set of arcs. In particular, for each pair of nodes $[N_s, N_t]$ such that N_s derives from a merge process, it must be checked if N_s is connected to N_t by two arcs (note that this situation could happen only if also N_t derives from a merge process) and, in the affirmative case, they must be merged into a unique one. If only one arc exists between N_s and N_t the corresponding coefficients must be updated.

Finally, the algorithm considers similarities between sub-sources; note that only similarities between sub-sources corresponding to rooted sub-nets are taken into account. The merge of rooted sub-nets can lead to the presence of two arcs between the same pair of nodes; if this happens, the two arcs must be merged.

The integration algorithm returns also some support information structures which can be exploited by all applications that use the global integrated SDR-Network. In more detail, these structures are: *(i)* a *Set of Mappings*, describing how each object belonging to output information sources has been obtained from one or more objects belonging to input information sources; *(ii)* a *Set of Views*, allowing to obtain the instances of objects of output information sources from instances of objects of input information sources.

5 Applications

In this section we show how both the proposed framework and the SDR-Network conceptual model can be useful in various application contexts.

E-commerce In this application case, trading partners frequently exchange messages that describe business transactions. Each trading partner uses its own message format. In order to enable systems to exchange messages, application developers must convert messages between the formats required by different trading partners.

In this application context the SDR-Network is extremely useful because: *(i)* it is able to uniformly representing information sources having different formats; *(ii)* it can support the extraction of intersource properties, such as synonymies, which the translation process is based on.

A possible approach for carrying out the message exchange is composed by the following steps:

- translation of the source and destination formats from their conceptual model to the SDR-Network;
- integration of the SDR-Networks associated to source and destination formats for obtaining a global SDR-Network;
- translation of the source data into the data of the global SDR-Network;
- translation of the data of the global SDR-Network into the destination data. It is worth pointing out that, in many cases, source and destination data are quite similar and that differences are mainly due to variations of their data models; in this case the technique overviewed here appears particularly suited and can be almost completely automated.

Semantic Query Processing In this case the user specifies a query and the system tries to answer it. A user query could involve many information sources, possibly having different formats and structures. A user may not know which information sources contain data he is interested in as well as he could not be able to handle a great number of information formats. For solving these problems, the mediator-based Cooperative Information Systems have been proposed in the literature [19]. In this architecture, the mediator has a global scheme expressed in a global conceptual model which could be different from models of the involved information sources. This global scheme can be constructed with the framework presented in this paper.

Data and Web Warehouses Data and Web Warehouses are decision support systems obtained from sets of information sources [14]. The process of data extraction must transform data from the source format into the warehouse format; in order to carry out its task it must exploit interscheme properties. When the involved information sources are numerous and large, it is quite difficult to derive data marts directly from transactional data; in this case a three level data warehouse architecture [8, 9] is well suited. In this architecture, transactional data are first reconciled, integrated and stored in a second level (called reconciled data level) and data marts are derived from reconciled data instead of from transactional data. We argue that insterscheme properties are extremely useful for reconciling data and that the SDR-Network is well suited as the conceptual model for representing reconciled data.

Advanced Web Search Engines Web search engines generally work on extensional data; indeed, they support users in the search of information sources containing some specific portion of data. However they are not able to associate data to concept. Therefore, when the same extensional data are associated to different concepts in different information sources, Web search engines are generally not able to distinguish the various sources on a "semantic" or "context-based" basis and, consequently, all found information sources are considered valid and returned to the user. As an example, suppose a user specifies the word *Stock* because he/she is interested in suppliers for his/her store; the Web search engine could return also sources about stock exchange.

As a consequence, the necessity arises of tools able to derive source semantics, on the one hand, and to describe the concepts of interest for the user, on the other hand. We have already seen that the SDR-Network can describe the semantics of information sources; the formal description of user interests can be obtained by exploiting user profiles. These can be represented by the SDR-Network conceptual model. The suitable cooperation of these two tools allows to realize an approach which exploits existing Web search engines but produces results which are semantically more precise and which best fits user needs w.r.t. those directly produced by the engines.

The basic idea is as follows: first the SDR-Network representation of the set of sites considered interesting for the user is obtained; then, the SDR-Network representation of the template of the document required by him/her is derived with the support of the user profile; finally, interscheme properties derived from involved information sources are exploited for improving the semantic characterization capabilities of the Web engine.

6 Conclusions

In this paper we have presented an approach for uniformly managing heterogeneous semi-structured information sources. The proposed approach provides the following novelties w.r.t. the existing ones: *(i)* the definition of a new conceptual model capable to represent information sources having different nature and structure; *(ii)* the presentation of a graph-based technique for automatically extracting interscheme properties and complex assertions between knowledge patterns holding among object classes of heterogeneous information sources; *(iii)* the design of an algorithm for the construction of a uniform representation of involved information sources.

At present we are working towards designing a tool capable to automatically obtain SDR-Networks associated to input information sources and to derive interscheme properties; moreover we are working on techniques that allow to extract other interesting interscheme properties, such as hyponymies and hypernymies.

In the future we are planning to exploit SDR-Networks to support the construction of advanced Web search engines.

References

[1] S. Abiteboul. Querying semi-structured data. In *Proc. of International Conference on Database Theory (ICDT'97)*, pages 1–18, Delphi, Greece, 1997. Lecture Notes in Computer Science, Springer-Verlag.

[2] C. Batini and M. Lenzerini. A methodology for data schema integration in the entity relationship model. *IEEE Transactions on Software Engineering*, 10(6):650–664, 1984.

[3] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati. Source integration in data warehousing. In *Proc. of Workshop on Data Warehouse Design and OLAP Technology*, pages 192–197, Wien, Austria, 1998. IEEE Computer Society.

[4] S. Castano and V. De Antonellis. Semantic dictionary design for database interoperability. In *Proc. of International Conference on Data Engineering (ICDE'97)*, pages 43–54, Birmingham, United Kingdom, 1997. IEEE Computer Society.

[5] T. Catarci and M. Lenzerini. Representing and using interschema knowledge in cooperative information systems. *Journal of Intelligent and Cooperative Information Systems*, 2(4):375–398, 1993.

[6] S. Ceri, S. Comai, E. Damiani, P. Fraternali, S. Paraboschi, and L. Tanca. XML-GL: A graphical language for querying and reshaping XML documents. *Computer Networks*, 31(11-16):1171–1187, 1999.

[7] R. Goldman, J. McHugh, and J. Widom. From semistructured data to XML: Migrating the lore data model and query languages. In *Proc. of International Workshop on the Web and Databases (WebDB'99)*, pages 25–30, Philadelphia, Pennsylvania, USA, 1999.

[8] IBM. *Information Warehouse Architecture I.* IBM Corporation, 1993.

[9] L. Palopoli, L. Pontieri, G. Terracina, and D. Ursino. Semi-automatic construction of a data warehouse from numerous large databases. In *Proc. of International Conference on Re-Technologies for Information Systems (ReTIS'00)*, pages 55–75, Zurich, Switzerland, 2000. Osterreichische Computer Gesellschaft 2000.

[10] L. Palopoli, D. Saccà, G. Terracina, and D. Ursino. A unified graph-based framework for deriving nominal interscheme properties, type conflicts and object cluster similarities. In *Proc. of Fourth IFCIS Conference on Cooperative Information Systems (CoopIS'99)*, pages 34–45, Edinburgh, United Kingdom, 1999. IEEE Computer Society.

[11] L. Palopoli, D. Saccà, and D. Ursino. DL_P: a description logic for extracting and managing complex terminological and structural properties from database schemes. *Information Systems*, 24(5):403–425, 1999.

[12] L. Palopoli, G. Terracina, and D. Ursino. Inferring complex intensional knowledge patterns from heterogeneous semi-structured information sources. Submitted for publication. Available from the authors.

[13] L. Palopoli, G. Terracina, and D. Ursino. A graph-based approach for extracting terminological properties of elements of XML documents. In *Proc. of International Conference on Data Engineering (ICDE 2001)*, pages 330–340, Heildeberg, Germany, 2001. IEEE Computer Society.

[14] E. Rahm and P.A. Bernstein. On mathing schemas automatically. In *Technical Report MSR-TR-2001-17*, http://www.research.microsoft.com/pubs/, 2001.

[15] D. Rosaci, G. Terracina, and D. Ursino. An algorithm for obtaining a global representation from information sources having different nature and structure. Submitted for publication. Available from the authors.

[16] D. Rosaci, G. Terracina, and D. Ursino. Deriving "sub-source" similarities for information sources having different structure and nature. Submitted for publication. Available from the authors.

[17] D. Suciu. Semistructured data and XML. In *Proc. of International Conference on Foundations of Data Organization (FODO'98)*, Kobe, Japan, 1998.

[18] G. Terracina and D. Ursino. Deriving synonymies and homonymies of object classes in semi-structured information sources. In *Proc. of International Conference on Management of Data (COMAD 2000)*, pages 21–32, Pune, India, 2000. McGraw Hill.

[19] G. Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, 25(3):38–49, 1992.