IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 2, Issue 2, Apr-May, 2014
ISSN: 2320 – 8791 (Impact Factor: 1.479)
www.ijreat.org

Ontology Based Data Integration with Flexible Querying System

Ilakkiya.P[1], Amudha.S[2]

[1]M.E. II Year, Department of Computer Science and Engineering, Sriram Engineering College, Perumalpattu – 602 024

[2]Assistant Professor, Department of Computer Science and Engineering, Sriram Engineering College, Perumalpattu – 602 024

Abstract

The data warehouse, composed of data tables extracted from Web documents, has been built to supplement existing local data sources. First, we present the main steps of our semiautomatic method to annotate data tables driven by an OTR. The output of this method is an XML/RDF data warehouse composed of XML documents representing data tables with their fuzzy RDF annotations. We then present our flexible querying system which allows the local data sources and the data warehouse to be simultaneously and uniformly queried, using the OTR. This system relies on SPARQL and allows approximate answers to be retrieved.

Index Terms—Knowledge and data engineering tools and techniques, XML/XSL/RDF, uncertainty, "fuzzy," and probabilistic reasoning, representations, data structures, and transforms and knowledge modeling.

1. INTRODUCTION

Today's Web is not only a set of semi structured documents interconnected via hyperlinks. A huge amount of technical and scientific documents, available on the Web or the hidden Web (digital libraries,. . .), include data tables. Those data tables can be seen as small relational databases even if they lack the explicit metadata associated with a database. They represent a very interesting potential external source for loading the data warehouse of a company dedicated to a given domain of application. They can be used to enrich local data sources. In order to integrate data, a preliminary step consists in harmonizing external data with local ones, i.e., external data must be expressed with the same vocabulary as the one used to index the local data. We have designed a software called ONtology-based Data INtEgration (ONDINE), using the semantic Web framework1 and language recommendations (XML, RDF, OWL, SPARQL),which implements the entire management system, presented ONDINE system relies on an Ontological and Terminological Resource (OTR) which is composed of two parts: on the one hand, a generic set of concepts dedicated to the data integration task and, on the other hand, a specific set of concepts and a terminology, dedicated to a given domain of application. ONDINE system is composed of two subsystems: 1) Web subsystem designed to load an XML/RDF data warehouse with data tables which have been extracted from Web documents and semantically annotated using concepts from the OTR; 2) MIEL++ subsystem designed to query simultaneously and uniformly the local data sources and the XML/RDF data warehouse using the OTR in order to retrieve approximate answers in a homogeneous way. Web subsystem has four steps as detailed in Fig. 1. In the first step, relevant documents for the application domain described in the OTR are retrieved from the Web and filtered by a human expert. In the second step, data tables are semi automatically extracted from the documents. In the third step, the extracted data tables are semantically annotated using the OTR. This step generates fuzzy annotations, represented in a fuzzy extension of RDF, which are associated with data tables represented in XML. In the fourth and last step, the end user has to validate the fuzzy RDF semantic annotations associated with data tables before loading them in the XML/RDF data warehouse. Let us notice that Web subsystem does not pretend to annotate

all data tables extracted from any Web documents, but to annotate accurately target data tables extracted from documents identified as relevant for a given domain. The human intervention at each of its step is therefore required to guarantee the accuracy of the approach. In this paper, we focus on the third step, that is the semantic annotation method, of Web subsystem. Its main originality is to produce fuzzy RDF annotations which allow: 1) the recognition and the representation of imprecise numerical data appearing in the cells of a data table; 2) the computation and explicit representation of the semantic distance between terms in the cells of a data table and terms of the OTR. MIEL++ subsystem allows the fuzzy RDF annotations to be queried using SPARQL2 which is recommended by W3C to query RDF data sources. This subsystem is an extension of the MIEL flexible querying

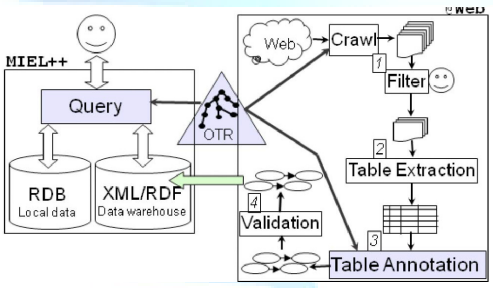

Fig. 1.ONDINE system.

system proposed in [1] and [2]. The main originalities of our new flexible querying subsystem are: 1) to retrieve not only exact answers compared with the selection criteria but also semantically close answers; 2) to compare the selection criteria expressed as fuzzy sets representing preferences with the fuzzy annotations of data tables. Some preliminary studies of this work have already been published in [3], [4], and [5]. This paper provides a synthetic overview of ONDINE system which relies on a new modeling of the OTR dedicated to the data integration task. The definition of this OTR, central in ONDINE system, was essential to consolidate the approach and ensure its sustainability and its future evolutions. @Web subsystem (previously presented in [3] and [4]) and MIEL++ subsystem (previously presented in [5] have been revised to take into account this new OTR. In Section 2, we present the new model of the OTR. The new Web and MIEL++ subsystems are then presented in the three next sections. The semantic annotation method of Web subsystem, which allows data tables, extracted from Web documents, to be fuzzy annotated using the OTR, is presented in two sections. In Section 3, we present the method which allows one to identify which concepts of the OTR are represented in a data table. The instantiation of these concepts for each row of the annotated data table, relying on fuzzy RDF annotations, is presented in Section 4. In Section 5, MIEL++ subsystem which allows a flexible querying of the fuzzy annotated data tables, stored in the XML/RDF data warehouse, using SPARQL is presented. Experimental results are given all along Sections 3, 4, and 5. Our approach is compared with the state of the art in Section 6. We conclude and present the perspectives of this work.

2. The Ontological and Terminological Resource

In [6], [7], [8], [9], [10] ontologies are associated with terminological and/or linguistic objects. In [6], Cimiano et al. motivate why it is crucial to associate linguistic information (part-of-speech, inflection, decomposition, etc.) with ontology elements (concepts, relations, individuals, etc.) and they introduce LexInfo, an ontology lexicon model, implemented as an OWL3 ontology. Adapting LexInfo, [7] presents a model called Lexicon Model for Ontologies (lemon) that supports the sharing of terminological and lexicon resources on the Semantic Web as well as their linking to the existing semantic representations provided by ontologies. The CTL model from [8] is a model for the integration of conceptual, terminological and linguistic objects in ontologies. In [9] a meta model for ontological and terminological resources in OWL DL is presented, called an Ontological and Terminological Resource, extended afterward in [11] in order to be used for ontology-based information retrieval applied to automotive diagnosis. The ontology we used in our previous works [3], [4], [5] was not designed to allow one to define the terminology and its variations (multilingual, synonyms, abbreviations,.) denoting the concepts. We therefore propose to use an ontological and Terminological Resource [9] allowing joint representation of an ontology and its associated terminology. According to [9], three factors influence the OTR structuring: the task to realize, the domain of interest and the

application. The OTR used in ONDINE system has been designed for the data table integration (annotation and querying) task. In this paper, the domain of interest is food safety but the OTR structure we propose is generic enough to be applied to many other domains. For example, in this paper, experimental results in aeronautics will be also presented. The application is the construction of a data warehouse opened on the Web. Since ONDINE system allows local data sources to be supplemented with data tables which have been extracted from Web documents, the domain specific part of the OTR was manually built by ontologists taking into account 1) the vocabulary used in the preexisting local databases in order to index the data and 2) the domain information available within the databases schema. Examples given in this paper concern the microbial risk domain. We present first, the conceptual component of the OTR and second, its terminological component, using the
OWL2-DL model.4

2.1 The Conceptual Component of the OTR

The conceptual component is the ontology of the OTR. It is composed of two main parts: a generic part, commonly called core ontology, which contains the structuring concepts of the data table semantic annotation task, and a specific part, commonly called domain ontology, which contains the concepts specific to the domain of interest.In order to understand the structure of the core ontology, let us detail the data table semantic annotation task. A data table is composed of columns, themselves composed of
cells. A data table must be structured in a standardized way, otherwise preliminary transformations are applied on it using state-of-the-art tools like spreadsheets (which is included in the table extraction step in Fig. 1). The cells of a data table may contain terms5 or numerical values often followed by a measure unit. During the semantic annotation of a data table, cells content are semantically annotated in order to identify the symbolic concepts or quantities represented by its columns and finally the semantic n-ary relationships linking its columns. For instance, in Fig. 2, the cell content "E.coli" is associated with the symbolic concept Escherichia coli by our annotation method (detailled in Sections 3 and 4), the content of the three cells 4.9, 41.1, and 45.8 are associated with the quantity Temperature and the entire content of the second row of the data table is
.

Table 1: Approximate temperature values for growth of selected pathogens in food

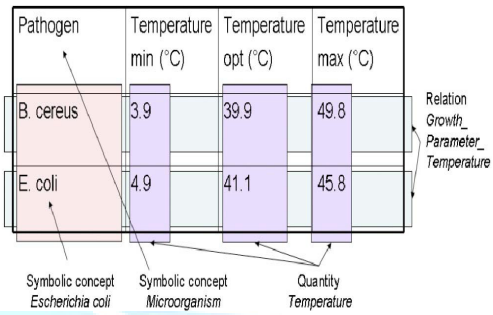

Fig. 2. Annotation of a table according to concepts defined in the OTR

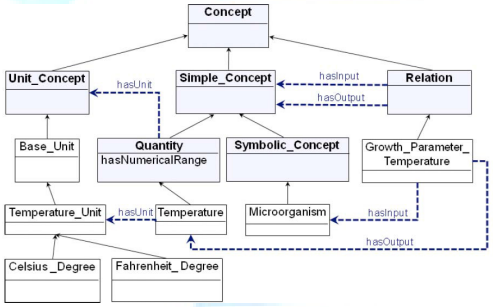

Fig. 3. An excerpt of the conceptual component of the OTR in microbial risk domain.

considered as an instance of the n-ary relation Growth Parameter Temperature which associates a given microorganism (like Escherichia coli) with its temperature growing conditions in a food product. The core ontology is therefore composed of three kinds of generic concepts: 1) simple concepts which contain the symbolic concepts and the quantities, 2) unit concepts which contain the units used to characterize the quantities, and 3) relations which allow n-ary relationships to be represented between simple concepts. The concepts belonging to the domain ontology, called specific concepts, appear in the OTR as sub concepts of the generic concepts. Fig. 3 presents an excerpt of the conceptual component of the OTR in microbial risk domain. In OWL, all concepts are represented by classes which are pair wise disjoints and are hierarchically organized by the sub Class Of relationship. The nodes represent the OWL classes, the solid arrows the "is-a" relationship between classes and the dashed arrows properties between classes. For instance, the

property has Unit links a quantity (e.g., a Temperature) with its units of measurement (e.g.Celsius_-Degree and Fahrenheit_Degree). We detail below the three kinds of generic concepts and their sub specific concepts in microbial risk domain.

2.1.1 The Unit Concepts

Unit concepts allow the meaning of units to be represented. Our classification relies on the international system of units. which decomposes the units into base units and derived units. There exist several ontologies dedicated to quantities and associated units (OM,7 QUDT,8 QUOMOS, OBOE,9 . . .).

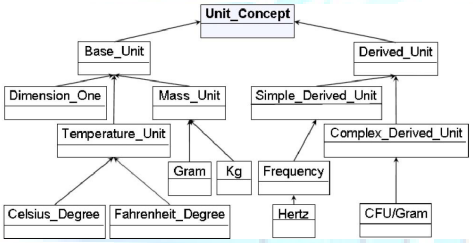

Fig. 4. An excerpt of the unit concepts in microbial risk domain.

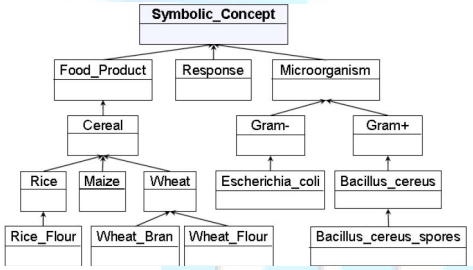

Fig. 5. An excerpt of the symbolic concepts in microbial risk domain.

We learn from these ontologies to build ours, but they cannot contain all the required specific units for a given domain. For instance, in microbial risk domain, the ontologist has added some units such as ppm10 or CFU=g.11 Fig. 4 presents an excerpt of the unit concepts in microbial risk domain.

2.1.2 The Simple Concepts

Symbolic concepts allow the meaning of terms to be represented. Symbolic concepts are hierarchically organized by the "is-a" relationship. Fig. 5 presents an excerpt of the specific symbolic concepts in microbial risk domain. The microbial risk domain OTR contains three distinct sub hierarchies of specific symbolic concepts: the specific symbolic concept Food_Product with more than 400 sub concepts, the specific symbolic concept Microorganism with more than 150 sub concepts and the specific symbolic concept Response with three sub concepts: growth, absence of growth, and death, which represent the possible responses of a microorganism to a treatment. These sub hierarchies have been defined by ontologists. We could not reuse preexisting terminologies for food products such as AGROVOC12 (from FAO—Food and Agriculture Organisation of the United Nations) or Gems-Food13 (from WHO—World Health Organisation) because those terminologies are not specific enough compared with the one built from our corpus in microbial risk (only 20 and 34 percent of common words, respectively). Quantities allow the meaning of numerical values to be represented. A quantity is described by a set of units, which

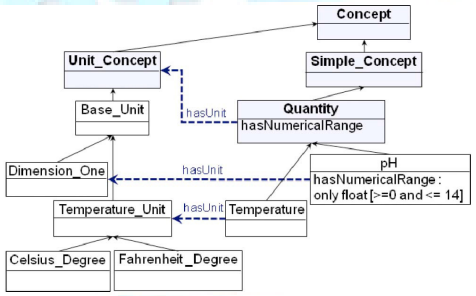

Fig. 6. An excerpt of the quantities associated with unit concepts in microbial risk domain

are sub concepts of the unit concept, and eventually a numerical range. Two properties has Unit and has Numerical Range, belonging to the core ontology, link, respectively, quantities to their associated units and numerical range. The OWL object property has Unit allows a quantity to be described by one or several unit concepts. OWL2-DL data type restrictions using facet spaces allow the numerical range of a quantity to be represented in the OWL data type property has Numerical Range. Fig. 6 presents an excerpt of the quantities in microbial risk domain. Eighteen specific quantities have been defined for the microbial risk domain. The specific quantity Temperature can be expressed using the unit _C (represented by the concept Celsius_Degree) or _F (represented by the concept Fahrenheit_Degree) and has no numerical range. The specific quantity pH is associated with the unit Dimension_One (i.e.,

with no unit) and is restricted to the numerical range [0,14].

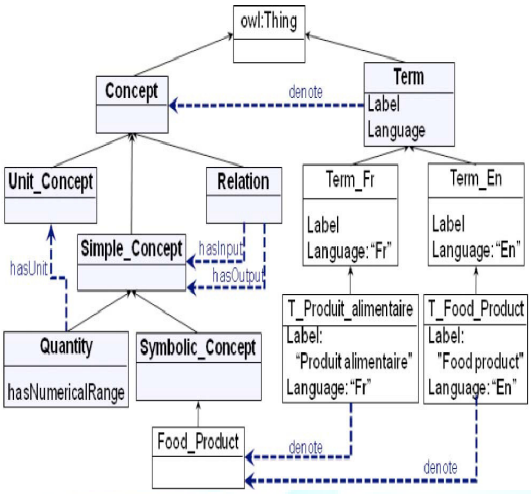

Fig. 7. An excerpt of the OTR in microbial risk domain.

3 . The Relations Identification in A Data Table

Given the OTR, described above, and given a data table extracted from a document found on the Web, we want to find which relations of the OTR are represented in this data table. An aggregation approach is used for that purpose, looking first at the contents of the cells, then identifying the simple concepts of the OTR represented in the columns and finally comparing the signature of the data table (the column concepts) with the signatures of the relations in the OTR. The main steps of the relations identification method are presented in Fig. 8: first, symbolic and numerical columns are distinguished, using some of the knowledge described in the OTR (mainly the unit concepts; for better description of this step, please refer to [3] which is a preliminary version of this work); then, the simple concepts represented by the symbolic columns and by the numerical columns are identified;

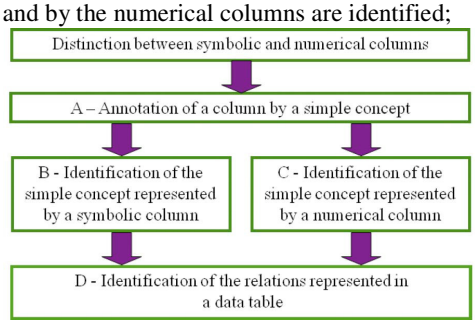

Fig. 8. The main steps of the relations identification method

finally, the relations represented in the data table are identified. We detail below the steps A, B, C, and D from Fig. 8. Each of these steps was experimented on three domains: microbial risk, chemical risk, and aeronautics. Three OTR were build, their domain specific part being manually built by ontologists. The size of each set of concepts belonging to the three OTR is given in Table 1, where U is the set of unit concepts, Q is the set of quantities, SC is the set of symbolic concepts, R is the set of relations and T is the set of terms. OTR and corpora are available on the Web .

4. The Fuzzy Querying Method

We present in this section the querying subsystem, called MIEL++, of ONDINE system. MIEL++ querying subsystem allows a uniform querying of two kinds of data sources: the local data sources and the XML/RDF data warehouse, which has been loaded with the data tables extracted from Web documents and semantically annotated. It relies on the OTR used to index the local data sources and to annotate the data tables. MIEL++ querying subsystem allows the end-user to express preferences in his/her query and to retrieve the nearest data stored in the two kinds of data sources corresponding to his/her selection criteria: the OTR—more precisely the hierarchical set of symbolic concepts—is used in order to assess which data can be considered as near to the selection criteria. The end-user asks his/her query to MIEL++ subsystem through a single graphical user interface (GUI), which relies on the OTR. The query is translated into a query comprehensible by each kind of data source, using two subsystems wrappers: an SQL query in the relational source (see [2] for more details about the SQL subsystem wrapper) and a SPARQL query in the XML/RDF data warehouse (see [5] for a complete description of the SPARQL subsystem wrapper). The final answer to the query is the union of the local results retrieved from the two kinds of data sources, which are ordered according to their relevance to the query selection criteria. In this section, we present the extension of MIEL++ subsystem which allows the end user to query fuzzy RDF annotations of data tables, represented in XML documents, by means of SPARQL queries. We remind the notions of view and MIEL++ query (see [2] for more details). We then present the construction of a MIEL++ answer retrieved from the XML/RDF data warehouse. We

conclude this section with experimental results.

4.1 MIEL++ Query

A MIEL++ query is asked in a view which corresponds to a given relation of the OTR. A view is characterized by its set of queryable attributes and by its actual definition. Each queryable attribute corresponds to a simple concept of the relation represented by the view. The notion of view must be understood with the meaning of the relational database model. It allows the complexity of the querying into different data sources to be hidden to the end user. A MIEL++ query is an instantiation of a given view by the end user, by specifying, among the set of queryable attributes of the view, which are the selection attributes and their corresponding searched values, and which are the projection attributes. An important feature of a MIEL++ query is that searched values may be expressed as continuous or discrete fuzzy sets. A fuzzy set allows the end user to express his/her preferences which will be taken into account to retrieve not only exact answers (corresponding to values associated with the kernel of the fuzzy set) but also answers which are semantically close. When a MIEL++ query is asked by the end user into the XML/RDF data warehouse which contains fuzzy RDF graphs generated by our annotation method to annotate XML data tables, the query processing has to deal with fuzzy values. More precisely, it has 1) to take into account the certainty score associated with the relations represented in the data tables and 2) to compare a fuzzy set expressing querying preferences to a fuzzy set, generated by our annotation method, having a semantic of similarity or imprecisionThe SPARQL query is automatically generated 1) from the signature of the relation represented by the view and associated with the MIEL++ query and 2) from the sets of projection and selection attributes of the MIEL++ query.

.

5. CONCLUSION

We have presented in this paper a complete system, called ONDINE, built, using the recommendations of the W3C, on a generic OTR expressed in OWL. ONDINE system allows XML data tables, which have been extracted from Web documents, to be annotated with fuzzy RDF descriptions and to be flexibly queried using SPARQL. Fuzzy RDF annotations are used to represent (1) the set of most similar symbolic concepts of the OTR which are automatically associated with the content of a cell belonging to a symbolic column, (2) imprecise values associated with a quantity expressed in one or several numerical columns, (3) a degree of certainty associated with each n-ary relation recognized in a data table. ONDINE system has been implemented through the development of @Web software on the one hand and the development of MIEL++ software on the other hand .To the best of our knowledge, ONDINE is the only software which allows one to simultaneously 1) annotate accurately a data table with an OTR and 2) perform approximate reasoning during the flexible querying process, comparing preferences expressed by the end-user with fuzzy annotations. In the very next future, we want to explore four new ideas to extend our approach. The first one consists in associating the data tables, which have been extracted from Web documents, with a reliability degree which takes into account several criteria to qualify the trust in the data source as for example the type or the reputation of the data source. The other perspectives concern the improvement of ONDINE system by 1) completing the cosine similarity measure used to compare terms with other syntactical and semantic techniques, 2) completing the semantic annotation of data tables in Web documents with the annotation of the text using the OTR, and 3) managing OTR evolution by taking into account annotation results and other ontologies. associated with a given quantity. Those perspectives will allow us to test the genericity of our OTR, which we pretend to be dedicated to the data integration task.

REFERENCES

[1] P. Buche and O. Haemmerle´, “Towards a Unified Querying System of Both Structured and Semi-Structured Imprecise Data Using Fuzzy Views,” Proc. Linguistic on Conceptual Structures: Logical Linguistic, and Computational Issues (ICCS), pp. 207-220, 2000.

[2] P. Buche, C. Dervin, O. Haemmerle´, and R. Thomopoulos, “Fuzzy Querying of Incomplete, Imprecise, and Heterogeneously Structured Data in the Relational Model Using Ontologies and Rules,” IEEE Trans. Fuzzy Systems, vol. 13, no. 3, pp. 373-383, June 2005.

,[3] G. Hignette, P. Buche, J. Dibie-Barthe´lemy, and O. Haemmerle´, “An Ontology-Driven Annotation of Data Tables,” Proc. WISE Workshops Web Data Integration and Management for Life Sciences, pp. 29-40, 2007.

[4] G. Hignette, P. Buche, J. Dibie-Barthe´lemy, and O. Haemmerle´, “Fuzzy Annotation of Web

Data Tables Driven by a Domain Ontology," Proc. Sixth European Semantic Web Conf. The Semantic Web: Research and Applications (ESWC), pp. 638-653, 2009.

[5] P. Buche, J. Dibie-Barthe´lemy, and H. Chebil, "Flexible Sparql Querying of Web Data Tables Driven by an Ontology," Proc. Eight Int'l Conf. Flexible Query Answering Systems (FQAS), pp. 345-357, 2009.