International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Review on Extended Online Approach for Web Data Tables Integration

Urmila Bavkar[1], Dr. Akhil Khare[2]

[1]Department of Computer Engineering, Padmashree Dr.D.Y. Patil Institute of Engineering & Technology, University of Pune, India

[2]Professor, Department of Computer Engineering, Padmashree Dr.D.Y. Patil Institute of Engineering & Technology, University of Pune, India

Abstract: ***This review paper of ONDINE system which allows the loading and the querying of a data warehouse obtained from the Web, using an Ontological and Terminological Resource (OTR) concept. The data warehouse, composed of data tables extracted from Web documents, has been built to support existing local databases. First semiautomatic method is used to annotate data tables obtained by an OTR (Ontological and Terminological Resource). The output of this method is an XML/RDF data warehouse consists of XML documents which represent data tables with their fuzzy RDF annotations. Then used flexible querying system which allows the local databases and the data warehouse which is obtained by extracting web documents to be simultaneously and uniformly queried, using the OTR. Using SPARQL (Simple Protocol and RDF Query Language) querying language approximate answers are retrieved.***

Keywords: ONDINE, OTR, data table integration, XML/RDF, "fuzzy," and probabilistic reasoning, representations, data structures, and transforms, knowledge modeling.\

1. Introduction

A lot of scientific and technical documents, on the net or the concealed internet (e-libraries,), contain data tables. These data tables could be viewed as relational sources that were little actually when they lack the explicit metadata of 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 databases. To combine data, the initial step is external data must be expressed with the same vocabulary as the one used to index the local one. We have designed software called Ontology-based Data INtEgration (ONDINE), using the semantic Web framework and language recommendations (XML, RDF, OWL, and SPARQL), to supplement existing local data sources with data tables which have been extracted from Web documents. ONDINE program comprises two sub-systems:

1) Web subsystem made to fill an data warehouse with data tables which were pulled from internet documents and annotated utilizing theories in the OTR;
2) MIEL++ subsystem designed to query uniformly and simultaneously local database and data warehouse driven from web using the OTR in order to retrieve approximate answers in a homogeneous way.

The ONDINE method utilizes an Ontological and Terminological source (OTR) which comprises two components: on one hand, a common pair of theories focused on the data integration job and a particular group of theories as well as a language, devoted to some specified domain.

Ontological and Terminological Resource (OTR):

Since ONDINE program enables local data sources to be compounded with data tables which were removed from internet documents, the domain particular portion of the OTR was by hand constructed by ontologists using consideration:

1) The terminology used within the pre-existing local databases so that you can catalog the information
2) The domain information that can be found within the databases schema. Illustrations provided in this paper issue the microbial risk domain name. First, OTR's conceptual component then its terminological part utilizing the OWL2 DL version is presented.

(a) The Conceptual Component of the OTR
Since the conceptual component is the ontology of the OTR. It consists of two main parts: a generic part, also called core ontology, which structures data tables using semantic annotation method, and a specific part, called as domain ontology, concepts particular to the domain of interest. Core ontology is thus made up of three types of common theories: 1) symbolic concept includes symbolic concept & the quantities 2) unit concept that have the models employed to qualify the measures, and 3) connections which enable n-ary associations to be represented between straightforward theories. The notions of the domain ontology, called theories that are particular, come in the OTR as theories that are common.

(b) The Terminological Component of the OTR
The terminological component represents the terminology of the OTR: it contains the terms of the domain of interest. A term is defined as a sequence of words, in a language, and has a label. Terms are divided based on their source language. A term denotes a concept; which must denote at least one concept and it can denote several concepts.

2. Literature Survey

ONDINE [1] is most recently presented technique for data table integration in which mainly two tasks performs annotation as well as querying. The method presented in [2] was based on three factors that influence OTR structuring such as determining the domain of interest of the application. The ontology used in [3]-[5] was not designed to allow one to define the terminology and its variations (synonyms, multilingual, abbreviations,) denoting the concepts. In [6]-[10] ontologies are associated with terminological and/or linguistic objects. In [6], Cimiano et al. motivate why it is important to associate linguistic information to ontology elements (concepts, relations, individuals, etc.) and also introduce LexInfo, an ontology lexicon model, implemented as 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 objects, terminological objects and linguistic objects in ontologies. An Ontological and Terminological Resource [9] allowing joint representation of ontology and its associated terminology. According to [9], three factors influence the OTR structuring: finding the domain of interest to the application. ONDINE system uses OTR which has been designed for the data table integration task. Ontological and Terminological Resource, extended afterward in [11] in order to be used for ontology-based information retrieval applied to automotive diagnosis.

In recent time there are some other methods those are presented by different authors; however those are not well designed with same objectives as ONDINE. Another problem with existing methods is that recognition of n-ary relations not included. Some methods included the recognition of n-ary relations but not clearly defined, and inaccurate. The application is the construction of a data warehouse opened on the web. In ONDINE [1], an author overcomes all this limitations and presented the efficient method. However, this method still further requires extending by considering following points which are not considered in ONDINE system:

- Associating the data tables those are extracted from the 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.
- Need of using the cosine similarity measure which is used to compare terms with other syntactical and semantic techniques.

3. Basic System Architecture

In the initial step, important documents relevant to the application domain as described in the OTR are retrieved and selected by a manual process done by human specialist, is used in the next measure. In next measure, extracted data tables are annotated utilizing the OTR. Fuzzy annotations, symbolized in a fuzzy extension which is related to data tables represented in XML (Extensible Markup Language) are generated by this. In the last and next measure, the consumer needs to confirm the fuzzy RDF (Resource Description Framework) annotations before packing them into database. Internet sub-system doesn't annotate all data tables removed from any internet files, however, to annotate precisely goal data tables removed from records identified as important to get certain domain name. The individual involvement at each step is thus needed to ensure the truth of the strategy. In this document, we concentrate on the next measure which is the Web sub-system that is semantic system, of annotation.

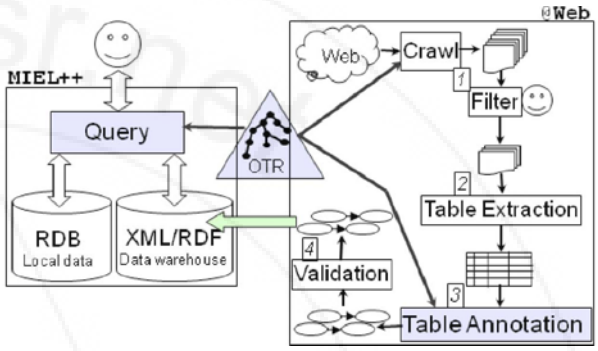

Figure 1: ONDINE System

Its main creativity would be to create fuzzy RDF annotations which permit:

1) The acknowledgement as well as the portrayal of unknown numeric information showing in the cells of data table;
2) The calculation and specific representation of the semantic space between conditions in the cells of data table and conditions of the otr.

MIEL++ sub-system allows the annotations that are unclear to be queried using SPARQL2 which W3C recommends to issue RDF data resources. This sub-system is an expansion of the MIEL++ adaptive querying system. Our adaptive querying subsystem's primary originalities are:

1) To recover not just responses that are precise compared with all semantically near responses;
2) (The collection policy) to evaluate the collection policy indicated as fuzzy models representing preferences using data tables' fuzzy annotations.

4. Conclusion

In the recent time, different efficient methods are present for web data table's integration those are extracted from the web documents using the ontological and terminological resource. This system is called as ONDINE system. This system allows extracted XML data tables from Web, to be annotated with fuzzy RDF descriptions and to be flexibly queried using SPARQL. Then investigated Fuzzy RDF annotations those 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) Estimate 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.

However, this ONDINE method still further needs to improve in different ways. So presented the extended version of this system called as Extended ONDINE. In this system, the use of cosine similarity measure for the comparative study as well as the efficient method for semantic annotation of web data tables in the web documents before extraction with aim of improving the efficiency of web data table's integration are added.

5. Acknowledgment

I would like to take this opportunity to acknowledge the contribution of certain people without which it would not have been possible to complete this paper work. I am thankful to the Principal Dr. R. K. Jain, Guide, Head, Coordinators, Colleagues of the Department of Computer Engineering, Dr. D. Y. Patil Institute of Engineering and Technology, Pimpri, Pune, Maharashtra, India, for their support, encouragement and suggestions. I would like to express my special appreciation and thanks to my guide Professor Dr. Akhil Khare, you have been a tremendous mentor for me.

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 Ontology," Proc. Eight Int'l Conf. Flexible Query Answering Systems (FQAS), pp. 345-357, 2009.

[6] P. Cimiano, P. Buitelaar, J. McCrae, and M. Sintek, "Lexinfo: A Declarative Model for the Lexicon-Ontology Interface," J. Web Semantics, vol. 9, no. 1, pp. 29-51, 2011.

[7] J. McCrae, D. Spohr, and P. Cimiano, "Linking Lexical Resources and Ontologies on the Semantic Web with Lemon," Proc. Eight Extended Semantic Web Conf. The Semantic Web: Research and Applications (ESWC), pp.245-259, 2011.

[8] T. Declerck and P. Lendvai, "Towards a Standardized Linguistic Annotation of the Textual Content of Labels in Knowledge Representation Systems," Proc. Seventh Int'l Conf. Language Resources and Evaluation (LREC '10),2010.

[9] A. Reymonet, J. Thomas, and N. Aussenac-Gilles, "Modelling Ontological and Terminological Resources in OWL DL," Proc. OntoLex 2007 - Workshop associated with ISWC '07, Sixth Int'l Semantic Web Conf. (ISWC '07),2007.

[10] C. Roche, M. Calberg-Challot, L. Damas, and P. Rouard, "Ontoterminology - A New Paradigm for Terminology," Proc. Int'l Conf. Knowledge Eng. and Ontology Development (KEOD), pp. 321-326. 2009.

[11] A. Reymonet, J. Thomas, and N. Aussenac-Gilles, "Ontology Based Information Retrieval: An Application to Automotive Diagnosis," Proc. Int'l Workshop Principles of Diagnosis, pp. 9-14, 2009.

[12] Yangarber, W. Lin, and R. Grishman, "Unsupervised Learning of Generalized Names," Proc. Int'l Conf. Computational Linguistics, 1-7, 2002.

[13] C.J. van Rijsbergen, Information Retrieval. Butterworth, 1979.

[14] J.C. Platt, Fast Training of Support Vector Machines Using Sequential Minimal Optimization, pp. 185-208. MIT Press, 1999.

[15] L. Zadeh, "Fuzzy Sets," Information and Control, vol. 8, pp. 338-353,1965.

[16] L. Zadeh, "Fuzzy Sets as a Basis for a Theory of Possibility," Fuzzy Sets and Systems, vol. 1, pp. 3-28, 1978.

[17] D. Dubois and H. Prade, "The Three Semantics of Fuzzy Sets," Fuzzy Sets and Systems, vol. 90, pp. 141-150, 1997.

[18] D. Dubois and H. Prade, Possibility Theory - An Approach toComputerized Processing of Uncertainty. Plenum Press, 1988.

[19] M. Baziz, M. Boughanem, H. Prade, and G. Pasi, "A Fuzzy Logic Approach to Information Retrieval Using a Ontology-Based Representation of Documents," Fuzzy Logic and the Semantic Web,vol. 1, pp. 363-377, 2006..J. van Rijsbergen, Information Retrieval. Butterworth, 1979.

[20] Y. Liu, K. Bai, P. Mitra, and C.L. Giles, "Tableseer: AutomaticTable Metadata Extraction and Searching in Digital Libraries," Proc. ACM/IEEE-CS Seventh Joint Conf. Digital Libraries (JCDL), pp. 91-100, 2007..J. van Rijsbergen, Information Retrieval. Butterworth, 1979.

Author Profile

Urmila V. Bavkar is Research Scholar Dr. D. Y. Patil Institute of Engineering & Technology, Pune, University of Pune, Maharashtra, India. She has received her Bachelor's Degree in Computer Science from SVERI's College of Engineering, Pandharpur, Shivaji University, Maharashtra. Currently she is pursuing her

Master's Degree in Computer Engineering from Dr. D. Y. Patil Institute of Engineering & Technology, Pune, University of Pune.

Prof. Akhil. Khare is presently working as a Professor in Department of Computer Engineering, Dr. D. Y. Patil Institute of Engineering and Technology, Pimpri, Pune, Maharstra, India. He has published more than 60 research papers in National/International Journals (Including IEEE, ACM, CIIT, IJCTEE-US, JICT-London.) Apart from university of Pune, also associated with Shivaji University, Bharati Vidyapith, University of Petrolium & Energy Studies, Dehradun, Gujarat Vidyapeeth, AMET University Chennai and ICFAI University Jharkhand at UG, PG and Ph. D. level. Also guided project for softKOASH Pune, Ministry of Defense, India and Infrarel, Indore and also working as Reviewer at IJCAT. Associated with various societies like ISTE, IETE, APACALL TIG, and are also senior member of the IACSIT, Singapore.