Indian Journal of Science and Technology, Vol 8(32), DOI: 10.17485/ijst/2015/v8i32/87210, November 2015
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Annotating Fuzzy Web Data Table Integration using ONDINE System

S. Selvapriya[1] and S. Thirunavukarasu[2*]

[1]Department of Information Technology, Jerusalem College of Engineering, Chennai - 600100, Tamil Nadu, India; selvapriya2005@gmail.com

[2]Department of Information Technology, Bharath University, Chennai - 600073, Tamil Nadu, India; thirunavukarasu.it@bharathuniv.ac.in

Abstract

To design ONDINE system for loading and querying of data from the data warehouse using Ontological and Terminological Resource (OTR). The data warehouse consists of data in the table format which is to be extracted from the web documents. The primary step is to annotate the data tables that are extracted. Then the querying system is presented for simultaneous responses to the user queries using OTR.

Keywords: Data Structures, Fuzzy, Knowledge and Data Engineering Tools and Techniques, Representations, Transforms and Knowledge Modeling, Uncertainty, XML/RDF

1. Introduction

The Ontology based Data Integration (ONDINE) system is designed for the correlation of external data with the local data, which is mainly based on the OTR, which consists of two kinds of concepts such as generic and specific (say to given domain). The ONDINE system comprised of two subsystems as

1. @ Web: loading of XML/RDF data extracted from the web and annotating using the OTR concepts.
2. MIEL++ for querying simultaneously and retrieving the equivalent answers to those queries of XML/RDF form of data from the data warehouse using the OTR.

@Web subsystem is involved of four stages namely:

- Collection of data(crawling) and filtering,
- Extraction of data tables from documents in semiautomatic manner.
- Semantic annotation of data tables using OTR, in which the association of RDF forms of annotated data with the XML data tables.
- Validation of semantic RDF annotations with data tables associate together, in prior of XML/RDF data loading.

Note that @web subsystem, annotation is not done to all data tables that are extracted from the web documents, only to the accurate data tables targeted which are related to the given domain. The RDF annotations allow representing of numerical data in data tables and also the semantic distance among the data in data tables and the OTR terms. But MIEL++ subsystem permits the querying of fuzzy RDF annotations by using the flexible querying system of MIEL++ (SPARQL). Also note that MIEL++ subsystem's genuine is that, this system helps to retrieve the logically related answers rather than the exact answers[2].

The OTR is the central part of ONDINE system, which combines the approach and assures its sustainability in future evolutions. In section 2, new model of OTR, @web and MIEL++ subsystems are presented in upcoming sections[1]. The semantic annotation method in @web subsystem allows the extracted data tables from the web documents to be as annotated fuzzy using the OTR, which is presented in three next sections[3]. In section 3, a proposed method that allows to identify the concepts of OTR which are represented in a data table. The instances of these concepts of the annotated data tables in each row mainly based on fuzzy RDF annotations, which is presented in section 4. In section 5, MIEL++ subsystem uses SPARQL to grant the flexible querying of fuzzy annotated data tables that are stored in XML/RDF data warehouse and the results of these experiments are given in respective sections 3, 4, 5.

* Author for correspondence

2. Related Work

The domain part of OTR was built by ontologists by taken into account of

- Vocabularies used in preexisting of local databases to index the data.
- Domain base information present within the database schema. The components that are presented are conceptual and terminological[4]. The conceptual components composed of two primary parts as: **generic** and **specific** parts[4]. The generic is also called as core ontology and specific is also called as domain ontology.

Thus the core ontology is of three generic concepts:

- Simple concepts for containing symbols and quantities
- Unit concepts to contain units and to characterize the quantities and
- Relations to allow n-ary relationships to be presented between simple concepts[5].

The concepts of domain ontology appear in OTR as sub-concepts of generic. concepts. The unit concepts permits the meaning of units to be represented and our classification is based upon international system of units, decomposes units into base units and derived units. There are several ontologies dedicated to quantities and associated units. From these kinds of ontologies, we learn that they do not contain all the required units for a given domain. The symbolic concepts allows meaning of terms to get represented and are hierarchically organized by 'is-a' relationship. The quantity permits the meaning of numerical values to get represent[6].

A relation is defined by signature, which consists of several input simple concepts and one output simple concept. The input represents the domain of relation. A relation may have several input simple concepts. The output concepts represent the range of the relation. A relation often represents semantic n-ary relationships between simple concepts with only one result in a data table[7]. If and data table contains several results then it is represented as many results. Two properties that belongs to core ontology is has Input and has Output, which links a relation to its domain and its range

3. Proposed System

To integrate data, a primary step consists of harmonized external data with local data, i.e., the external data must be demonstrated with same vocabulary as one used to index local data. For that users going to search data are through a single offline application in spite of going in for search at different web pages, also for getting of data accurately and with more frequently[8]. In this project, we designed a software called Ontology-based Data Integration (ONDINE), using semantic Web framework1 and language guidance (XML, RDF, OWL, and SPARQL), that implements entire management system, and to add-on existing local data sources with the data tables which are been extracted from the documents presented in web. User search data through offline application. ONDINE software is designed to supplement data tables extracted from web documents. Data is supplied to user with accuracy and the search time is reduced. Steps that are performed are:

- Data Table Creation.
- Annotate Data Tables.
- Querying the System.

3.1 Data Table Creation

In this, first step is to retrieve the relevant documents from the web for the chosen application domain. The documents retrieved from the web are then filtered to eliminate null data. The filtered document is extracted into data table's semi- automatically[9].

3.2 Annotate Data Tables

The semantic annotation of extracted data tables is performed[10]. Validation of fuzzy RDF semantic annotations is associated with data tables before loading into XML/RDF warehouse. Finding relations of instances with their column of data table is performed[11].

3.3 Querying the System

In this module, users query for RDF data sources either by SQL or SPARQL in XML/RDF warehouse[12]. The end-user query the fuzzy RDF annotations by SPARQL queries to subsystem using GUI relies on the OTR. The exact and related answers are supplied to user's queries from XML/RDF warehouse.

4. System Flow

The Figure 1 shows the steps of system flows are as follows:

- The crawling of data is performed by extracting the data tables from the web documents.
- The crawled contents are then filtered to get the appropriate data.

- Table is extracted and annotated to produce the data in table form.
- The data tables are stored at XML/RDF warehouse and retrieved in future.
- The end-users query the data and the exact and reliable data is supplied to the user within a short time[13].

The system demonstrate about data crawling, filtering and extracting into the table form, annotation is for separation of symbols and numbers in data and validation is for relating symbols with their numbers to the data[14]. Then data is supplied to the users query after checking the criteria of their access rights. The MIEL++ is used for the end users to query fuzzy RDF annotations of data tables, represented in XML documents, by means of SPARQL queries[15].

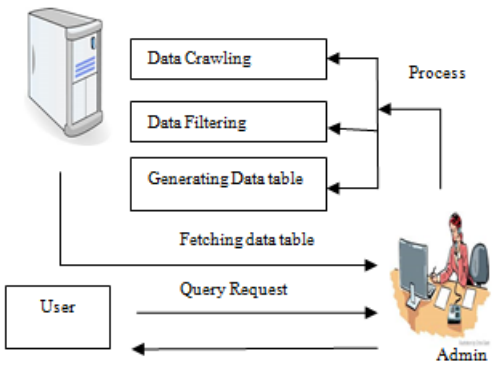

Figure 1. System Architecture.

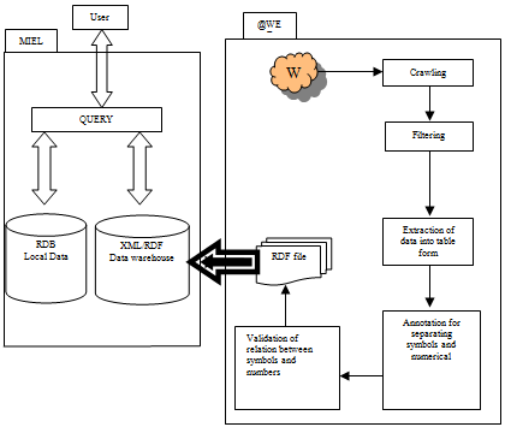

Figure 2. Functional Architecture of ONDINE Sytem.

5. Results and Discussion

Precision is the ability of a system to retrieve only relevant documents.

$$\text{Precision} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Retrieved}\}|}$$

Recall is the ability to retrieve all relevant documents.

Table 1. Precision, Recall, F1 Measure Experimental Results

Queried relation	p*	r*	F1
Original source	96%	95%	95%
Risky foods	78%	100%	87%
Time to develop	93%	100%	96%
Symptoms	95%	100%	97%

$$\text{Recall} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Relevant}\}|}$$

F1 Score is a combined measure that assesses precision/recall trade off. The greater F1 value indicates the better performance.

The proposed model is evaluated through precision, recall, and F1 measure values as follows.

True Positive (TP) =No. of relevant links retrieved.
True Negative (TN) =No. of irrelevant links not retrieved.
False Positive (FP) =No. of irrelevant links retrieved.
False Negative (FN) =No. of relevant links not retrieved.

$$\text{Precision} = \frac{\text{No. of relevant links retrieved(TP)}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{No. of relevant links retrieved(TP)}}{\text{TP} + \text{FN}}$$

The F1 measure is calculated by

$$\text{F1} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

*Where p is the precision and r is recall value.

6. Conclusion and Future Work

The ONDINE system is the software that allows one to perform simultaneously.

- Annotate accurately of data tables with an OTR and
- Performing approximate reasoning during querying process and comparison of preferences that are expressed by end-user with fuzzy annotations. In upcoming future, we must explore few ideas like associating data tables extracted from the web documents with a reliability degree, that take into account of sev-

eral criteria to qualify the trust in data source as like the type or reputation (opinion) of data source.

To the best of our knowledge, ONDINE is the only software that allows simultaneously to 1. Annotate accurately a data table with an OTR and 2. Performing rough type reasoning during querying process, comparison of preferences expressed by users with fuzzy annotations. ONDINE has been successfully tested on three different applications which demonstrates the generic potential of the proposal. In future, we want to explore four new ideas to extend the approaches. The primary one consists of associating the data tables, those are been extracted from Web documents, with a reliability that accounts to several criteria to qualify the trust in the data source as like type or opinion over the data source.

7. References

1. Buche P, Haemmerle´ O. 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); 2000. p. 207–20.
2. Latha RS, Vijayaraj R, Singam ERA, Chitra K, Subramanian V. 3D-QSAR and Docking Studies on the HEPT Derivatives of HIV-1 Reverse Transcriptase. Chemical Biology and Drug Design. 2011; 78(3):418–26. ISSN: 1747-0285.
3. Buche P, Dervin C, Haemmerle´ O, Thomopoulos R. Fuzzy Querying of Incomplete, Imprecise, and Heterogeneously Structured Data in the Relational Model Using Ontologies and Rules. IEEE Trans Fuzzy Systems. 2005 Jun; 13(3):373–83.
4. Declerck T, Lendvai P. 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.
5. Masthan KMK, Aravindha BN, Dash KC, Elumalai M. Advanced diagnostic aids in oral cancer. Asian Pacific Journal of Cancer Prevention. 2012; 13(8):3573–6. ISSN: 1513-7368.
6. Hignette G, Buche P, Dibie-Barthe´lemy J, Haemmerle´ O. An Ontology-Driven Annotation of Data Tables. Proc WISE Workshops Web Data Integration and Management for Life Sciences; 2007. p. 29–40.
7. Hignette G, Buche P, Dibie-Barthe´lemy J, Haemmerle´ O. Fuzzy Annotation of Web Data Tables Driven by a Domain Ontology. Proc Sixth European Semantic Web Conf The Semantic Web: Research and Applications (ESWC); 2009. p. 638–53.
8. Tamilselvi N, Dhamotharan R, Krishnamoorthy P, Shivakumar. Anatomical studies of Indigofera aspalathoides Vahl (Fabaceae). Journal of Chemical and Pharmaceutical Research. 2011; 3(2):738–46. ISSN: 0975-7384.
9. Buche P, Dibie-Barthelemy J, Chebil H. Flexible Sparql Querying of Web Data Tables Driven by an Ontology. Proc Eight Int'l Conf Flexible Query Answering Systems (FQAS); 2009. p. 345–57.
10. Cimiano P, Buitelaar P, McCrae J, Sintek M. Lexinfo: A Declarative Model for the Lexicon-Ontology Interface. J Web Semantics. 2011; 9(1):29–51.
11. Devi M, Rebecca LJ, Sumathy S. Bactericidal activity of the lactic acid bacteria Lactobacillus delbreukii. Journal of Chemical and Pharmaceutical Research. 2013; 5(2):176–80. ISSN: 0975-7384..
12. McCrae J, Spohr D, Cimiano P Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. Proc Eight Extended Semantic Web Conf the Semantic Web: Research and Applications (ESWC); 2011. p. 245–59.
13. Reymonet A, Thomas J, Aussenac-Gilles N. Modelling Ontological and Terminological Resources in OWLDL. Proc OntoLex 2007 - Workshop associated with ISWC '07; Sixth Int'l Semantic Web Conf (ISWC '07); 2007.
14. Roche C, Calberg-Challot M, Damas L, Rouard P. Ontoterminology - A New Paradigm for Terminology. Proc Int'l Conf Knowledge Eng and Ontology Development (KEOD); 2009. p. 321–326.
15. Reddy SV, Suchitra MM, Reddy YM, Reddy PE. Beneficial and detrimental actions of free radicals: A review. Journal of Global Pharma Technology. 2010; 2(5):3–11. ISSN: 0975-8542.
16. Kimio T, Natarajan G, Hideki A, Taichi K, Nanao K. Higher involvement of subtelomere regions for chromosome rearrangements in leukemia and lymphoma and in irradiated leukemic cell line. Indian Journal of Science and Technology. 2012 April, 5 (1):1801–11.
17. Cunningham CH. A laboratory guide in virology. 6th ed. Minnesota: Burgess Publication Company; 1973.
18. Sathish Kumar E, Varatharajan M. Microbiology of Indian desert. Ecology and vegetation of Indian desert. In: Sen DN, editor. India: Agro Botanical Publ.; 1990. p. 83–105.
19. Varatharajan M, Rao BS, Anjaria KB, Unny VKP, Thyagarajan S. Radiotoxicity of sulfur-35. Proceedings of 10th NSRP; India. 1993. p. 257–8.
20. 01 Jan 2015. Avaialable from: http://www.indjst.org/index.php/vision.