Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 7, July 2015, pg.194 – 199

RESEARCH ARTICLE

EXTENDED ONDINE APPROACH FOR WEB DATA TABLES INTEGRATION

Urmila Bavkar[1]

Master of Engineering in Computer Engineering

Pd. Dr. D.Y.Patil Institute of Engineering and Technology, Pimpri, Pune-18, India

urmila.bavkar@gmail.com

Prof. Dr. Akhil R. Khare[2]

Professor in Computer Engineering

Pd. Dr .D.Y.Patil Institute of Engineering and Technology, Pimpri, Pune-18, India

Khare.akhil@gmail.com

***Abstract:** This dissertation report presents design of ONDINE system which loads and queries a data warehouse constructed from the Web documents, with the help of an Ontological and Terminological Resource (OTR) concept. The data warehouse, consisting data tables extracted from documents present on web, has been designed to support existing native databases. Initial semiautomatic technique is employed for annotating data tables acquired by an OTR (Ontological and Terminological Resource). The output of this technique is an XML/RDF data warehouse containing XML documents which produce data tables associated with their fuzzy RDF annotations. Then adjustable querying system is used, that permits the local databases and also the data warehouse that is obtained by extracting internet documents to be unvaryingly and at the same time queried, using the OTR. Approximate answers are retrieved with the help of SPARQL (Simple Protocol and RDF Query Language) querying language.*

Keywords — ONDINE, OTR, data table integration, XML/RDF, "fuzzy", data structures, and transforms.

I. INTRODUCTION

Most of the scientific as well as technical documents, present on the net having data tables, which could be observed as relational sources that were less important when they are not having metadata associated with them. They hold the attention for the possible external source for loading the data warehouse of a company related to a particular domain of application. Then that can be wont to increase efficiency of the local databases. To combine external obtained data with local ones, initial step is external data must have the same vocabulary as the one used to index the local one. For that Ontology-based Data INtEgration (ONDINE) software is meant, with exploitation of the semantic Web framework and different recommended languages such as XML, RDF, OWL, and SPARQL, to support previously available local data sources with data tables which have been obtained from Web documents.

ONDINE uses an Ontological and Terminological resource (OTR) having two parts: a common pair of theories related to data integration job and a particular group of theories and a language, to some specified domain. ONDINE program composed of two sub-systems:

1) Web subsystem

Builds a data warehouse with data tables obtained from web documents and annotated using theories in the OTR;

2) MIEL++ subsystem

Designed to query unvaryingly and at the same time both local database and data warehouse created from web using the OTR, to retrieve more accurate answers.

II. RELATED WORK

The focus of this is not in detection and removal or what is referred to as data cleaning, but improve search operation despite the level of dirtiness of the database. Fuzzy searching can also be used to locate individuals based on incomplete or partially inaccurate identifying information in an attempt to deal with dirty data. Fuzzy search is to be done by means of a fuzzy matching program that will make list of results based on likely relevance even though search argument words and spellings may not exactly match. Data available is represented in the form of unstructured format which will not be easy to make the data analysis. The task of annotating an un-annotated image can be viewed formally as a classification problem for each word in the vocabulary we must make a decision.

Comfort T.Akinribido implemented the fuzzy-ontology based information retrieval system that determine the semantic equivalence between terms in a query and terms in a document by relating the synonyms of query terms with those of document terms. . The query terms can be expanded through a database that contains Keywords and their synonyms. Fuzzy-Ontology allows the easy determination of the precise meaning of a word as it relates to a document collection. Fuzzy-Ontology could be used in IR to locate precise information, which may be contained in a document content collection.

Shawn Bowers and Bertram Lud¨ascher defined the data integration and transformation tools are used to discover new knowledge through analysis and modeling. A generic framework for transforming heterogeneous data within scientific workflows. Framework is to provide that exploits ontological information to support structural data transformation for scientific workflow composition. Structural type similar to a conventional programming language data type defines the allowable data values for an input or output, whereas a semantic type describes the high level conceptual information of an input or output, and is expressed in terms of the concepts and properties of ontology.

David M. Blei Michael I. Jordan defined the solution for the task of annotating an un-annotated image can be viewed formally as a classification problem for each word in the vocabulary we must make a decision. Information retrieval are organized around the representation and processing of a document in problems in which one data type can be viewed as an annotation of the other data type.

III.SYSTEM ARCHITECTURE

First, important documents related to the particular application domain as mentioned in the OTR are taken and selected by human specialist, which used in the next measure. In next measure, data tables which are extracted in previous step are annotated using OTR. Fuzzy annotations in a fuzzy extension related to data tables in XML (Extensible Markup Language) are generated by this. In the last measure, before moving them into database, consumer has to confirm the fuzzy RDF (Resource Description Framework) annotations. Internet sub-system only annotates useful or a relevant document to an application that is it doesn't annotate all data tables removed from any internet files. To get relevant documents to any application domain, individual involvement at each step is important to ensure the truth of the strategy. In this project, concentration is on semantic system of annotation.

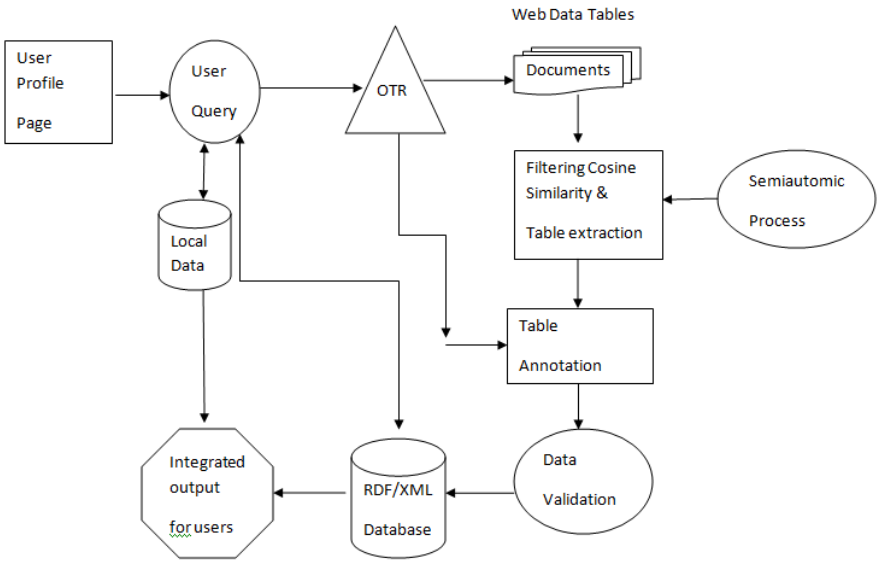


Fig: System Architecture

Its main function would be to create fuzzy RDF annotations which allow:
1) Acknowledgement of unknown numeric information in the cells of data table;
2) Calculating the semantic gap present between conditions in the cells of data table and conditions of the OTR and representing it.
MIEL++ subsystem query unclear annotations with the help SPARQL2, W3C recommends to issue RDF data resources. This subsystem is an expansion of the MIEL++ adaptive querying system.

IV. METHODOLOGY

This includes following steps
1. User Query
2. OTR Resource & Web Search
3. Filtering using cosine similarity & Table Extraction
4. Table Annotation based on OTR
5. Validation & Storing into RDF/XML Database
6. Users Integrated Output

1. User Query

In this a web application is designed in such a way to maintain the unique features of adapted querying system, which are:

a) Not only extracting exact answers but to extract semantically close answers to the given query.
b) Annotating results being founded on fuzzy logic and calculating similarity index.

The process is semi-automatic because user has to upload the related documents for executing query.

2. OTR Resource and Web Search

The OTR Recourse and web Search obscures the end user from the complexity of querying into different data sources. The set of query able attributes of the view and their corresponding searched values, are specified as selection attributes and projection attributes respectively. Initially an OTR resource has to be populated with the possible ontological relations. Unbroken or broken fuzzy sets can be used to search values in a MIEL++ query, which permit the user to indicate his/her preferences to extract related and exact answers.

3. Filtering using Cosine Similarity and Table Extraction

Colossal amount of data is present on the World Wide Web published either by experimenting or surveying by research organizations and various government bodies. The key is to find the related data among the entire pile of data.

Finding the similarity index will help as a good starting point for selecting the related documents of necessary information. We have used the cosine similarity as a measure of similarity. In Text mining and information retrieval, each term is assigned a different dimension and a document is characterized by a vector where the value of each dimension corresponds to the number of times that term appears in the document. Cosine similarity, which gives measure of how similar two documents are likely to be in terms of their subject matter [2].Cosine similarity ( ) is usually calculated by the underneath formula

$$x = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

4. Annotation of Tables

Tables are annotated based upon the OTR developed during one of the earlier modules. The efficiency of this module to recognize relations is heavily relied on the accuracy and range of the OTR resource developed.

5. Validation and Storing into RDF/XML Database

In the present module the query requested by the user is stored into the XML/RDF data warehouse after validation. The XML/RDF data warehouse consists of fuzzy RDF graphs which are employed in annotating the XML data tables.

The query processing has to deal with fuzzy values. Mostly, it has

a) To consider the exactness linked to the relations characterized in the data tables and
b) To evaluate a fuzzy set representing the querying preferences with respect to a fuzzy set having a semantic of similarity or imprecision.

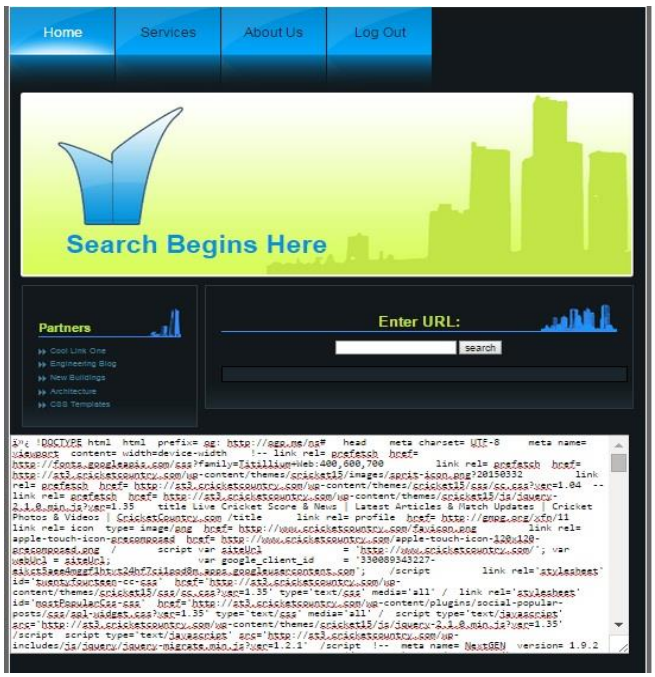
6. User's Integrated Output

Our approach in flexible SPARQL querying is a complete and integrated solution which allows one

a) To annotate the Web data tables which are stored in data ware house with the vocabulary defined in an OTR,
b) To complete the querying of the annotated tables using the same vocabulary and taking into account the fuzzy degrees generated by the annotation method according to their associated semantic.

V. RESULTS

Here, First admin have to login. Then next page is generated which is shown below where URL is entered from which tables are extracted and annotated. Then that web page's HTML format is displayed below.



In next figure, matrix is generated where precision recall result is displayed.

VI. CONCLUSION

This paper includes a whole system, called ONDINE, constructed, using guidelines of the W3C, on a universal OTR indicated in OWL.ONDINE system enables XML data tables extracted from internet documents, to be annotated with fuzzy RDF explanations and also to be flexibly queried using SPARQL. Fuzzy RDF annotations are used to signify (1) the set of most similar symbolic concepts of OTR which are

automatically associated with the contents of the a cell belonging to a symbolic column, (2) imprecise values associated with a quantity expresses in one or several numerical column, (3) a level of certainty related to each n-ary relationship identified in data table. Presented adaptive querying method allowing the data warehouse as well as the nearby data resources to be consistently and concurrently queried, utilizing the OTR. This method enables approximate responses by evaluating preferences indicated as fuzzy sets to be recovered and utilizes SPARQL.

ACKNOWLEDGEMENT

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, who has 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] Singhal, Amit (2001). "Modern Information Retrieval: A Brief Overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35–43.

[13] B.Gowthamipriya Darsini,B. Hanmanthu,"Ontological Semi Automic method for web table integration",International Journal of Computer applications, Volume 107 No-9,December 2014.

