ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 International Conference on Innovations in Engineering and Technology (ICIET'14)

On 21st & 22nd March Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

An OTR Driven Semiautomatic- Method for Annotations of Web Data

T.Rajkumar[1],T.Chellatamilan[2]

[1]Department of Computer Science and Engineering, Arunai Engineering college, Tiruvanamalai,Tamilnadu,India

[2]Department of Computer Science and Engineering, Arunai Engineering college, Tiruvanamalai,Tamilnadu,India

ABSTRACT-The data warehouse is composed of data tables extracted from the web documents and it has been supplement to existing local data sources. The semiautomatic method is to annotate data tables by an OTR. XML/RDF data warehouse is composed of XML documents representing data tables with their fuzzy RDF annotations. This system also has flexible querying method which is to query the local data sources and data warehouses uniformly. SPARQL is used to retrieve the answers. Ontological and Terminological Resource (OTR) is made of a generic set of concepts dedicated to the data integration task an specific set of concepts and terminology dedicated to a given domain. RDF provides mechanisms for describing groups of related resources and the relationships between these resources. RDF Schema is used to define the vocabulary descriptions that are written in RDF using the terms described in this document.Web data table is both formed and also obtained from the web to form data table with which the annotation is formed.The main function is to form the data table to make the annotation and with the use of RDF relation is to be defined.Correct and relevant information obtaining to the query is an important task. This task is complex if the query terms have many meanings and occur in different varieties of domain. 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.

I.INTRODUCTION

ONDINE is used for loading and the querying of a data warehouse opened on the Web which is guided by an Ontological and Terminological Resource (OTR).ONDINE is the ontology based data integration that uses semantic framework and language recommendations such as xml, RDF,OWL to implement the entire management system. They represent a very interesting potential external source for loading the data warehouse of a company dedicated to a given domain of application and they can be used to enrich local data sources. The ontology based data integration is closely tied to the consistency and expressivity of the ontology used in the integration process. The Resource Description Framework (RDF) is a general-purpose language for representing information in the Web. RDF describes how to use RDF to describe RDF vocabularies. RDF defines a vocabulary for this purpose and defines other built-in RDF vocabulary initially specified in the RDF Model and Syntax Specification. SPARQL provides a standard way to query RDF data. The SPARQL

update language allows the user to modify RDF graphs in an RDF dataset at various levels of granularity, including individual RDF statements. The SPARQL is used to provide a minimal set of uniform, HTTP operations for managing a semantic web of network-manipulable RDF at a large level of granularity.

The process of acquiring, maintaining and enriching the domain ontologies is referred to as ontology engineering.It is also defined as the explicit and shared conceptualization of the domain. For small Web sites with only static Web pages, it is feasible to construct a domain knowledge base manually or semi-manually. In Loh, Wives, & de Oliveira (2000) a semi-manual approach is adopted for defining each domain concept as a vector of terms with the help of existing vocabulary and natural language processing tools. The entire management system, presented in supplement existing local data sources with data tables which have been extracted from Web documents.

Data are extracted from data tables in Web documents.Web documents are very heterogeneous in nature is a key issue in this workflow ability to assess the reliability of retrieved data.The idea is to provide systems using the annotation power of a user community to collect information about reliability. This method do not rely on users but rather on information about the Web data table origins to compute a reliability estimations. RDF is used to annotate Web tables and SPARQL to query annotated Web tables.

The problem is obtaining correct and relevant information from the web as it consists of huge amount of technical and scientific documents.Web consists of documents which is represented in the form of web data tables. User searches the information in the web where large amount of information is available and related ,other related information is presented to the user where the exact information is not obtained. Since web consists lot of similar information as it searches for and those information may also be represented in the form of paragraph. Data tables can be seen as small relational databases even if they lack the explicit metadata associated with a database. Information on the web presented in the form of data tables and paragraph where the user finds its difficulty in obtaining their information from the web. They represent a very interesting potential external source for loading the data warehouse of a company dedicated to a given domain of application.

II.RELATED WORKS

The focus of this research 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 an 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. Retrieval, clustering, and classification, annotated data lends itself to tasks such as automatic data annotation and retrieval of un-annotated data from annotation type queries.

III.BACKGROUND

The ontological and terminology resource is used for data integration. ONDINE system is used to allow local data sources to be supplemented with data tables which have been extracted from Web documents.The domain is specific part of the OTR was manually built by ontologists .The vocabulary used in the preexisting local databases in order to index the data and the domain information available within the databases schema.

The Component of the OTR has two main parts Core ontology and Domain ontology.Core ontology contains the structuring concepts of the data table semantic annotation task.Domain ontology contains the concepts specific to the domain of interest. A data table is composed of columns, themselves composed of cells. The cells of a data table may contain terms or numerical values often followed by a measure unit. Cells content are semantically annotated in order to identify the symbolic concepts or quantities represented by its columns. The core ontology is composed of three kinds of generic concepts: simple concepts which contain the symbolic concepts and the quantities. Unit concepts which contain the units used to characterize the quantities. Relations which allow n-ary relationships to be represented between simple concepts. The entire information is to be maintained and its relationship is defined with it where the user makes authentication to retrieve the information. SPARQL query is used to retrieve the information.

A.Unit And Symbolic Concepts

Unit concepts allow the meaning of units to be represented. Symbolic concepts allow the meaning of terms to be represented. The meaning of numerical values to be represented by the quantities. Quantity is defined as the set of units and it is sub concepts of unit concept to form a numerical range.

B.Relationship

Relations are defined by the several simple input concepts and followed by which it forms a single output concept. Relations allow the meaning of n-ary relationships between simple concepts to be represented.The input simple concepts represent the domain of the relation. There may be several input concepts in single relation.The Range of relation is formed as the output of the simple concept.Results are to be represented in the data tables and it consists of several result columns which represent the as many relation as it has in the results.

C.Terms

Sequence of words represented in a language is called as a term and it has been divided based on their source language. A term denotes a concept that it must denote at least one concept and it can denote several concepts. The semantic annotation of a data table is composed of two steps: Relation identification in the OTR are made and it is represented in the data table and making instantiating the identified relations that is used to consists of associating a set of fuzzy RDF annotation graphs with each row of the data table.These terms are used to define the relationship between the concepts present to it which also represents the unit associated with it.

D.Web Data Table

Web data table is both extracted from the web and also formed.Data table are extracted from the web when the related information is available.It can also be formed from the paragraph contents available from the web.As a result is to be formed in order to provide valid information to the user.The information obtained has to be annotated for making the description of the data.The data table retrieved for managing the user requests.hence it has to be filtered to remove the unnecessary contents and make annotation.Annotated data provide a valid information to the user requests.Numeric and symbolic contents are to be separated for providing user needs based on the relations defined in it.Further the semantic annotation has to be made for defining the meaning of the data present in it.

E.RDF

RDF is defined as the Resource definition framework which is used as a framework for describing a Web metadata . The information about the information on the site which provides interoperability between applications that exchange machine-understandable information on the Web. For example RDF is used to provide information such as a site's sitemap the dates of when updates were made and keywords that search engines look for and the Web page's intellectual property rights. In this RDF is used to clearly define the relation between the concepts and its terms used in it. It is used to clearly define the relations that are present in the data tables.RDF schema is used to provide the vocabulary description of the data that is to be defined in the data table.

F.SPARQL

SPARQL is an RDF query language, that is, a query language for databases, able to retrieve and manipulate data stored in Resource Description Framework

format. SPARQL is data-oriented in that it only queries the information held in the models; there is no inference in the query language itself. The Jena model is to provide the impression that certain triples exist by creating them on-demand, including OWL reasoning. SPARQL does not do anything other than take the description of what the application wants, in the form of a query, and returns that information in the form of a set of bindings or an RDF graph.

G.System Architecture

This architecture is to define the entire system with OTR. Data table is extracted from the web, then it is annotated and validated.XML/RDF data warehouse is used from where using SPARQL is used to retrieved.

The semantic annotation function for the web subsystem is to allow the data tables extracted from the web documents to be fuzzy annotated using the OTR. In this process user requests information to the server which searches the related documents from the web. Initially it searches for the web data table and paragraph contents related to that information and then filter it to remove the unnecessary contents presented in to it.If the information is available in the form of data table, then the data table extraction is made to it and if the related information is available in the form of paragraph it will be retrieved to form the data table.

Ontology and terminological resource is used to annotate the information obtained during that search related contents and the validation process is to be made to make the updating the contents with that available in the web.The information is to be made available in the form of RDF file.User extracts information from the domain using sparql which allows the user to obtain the exact information.

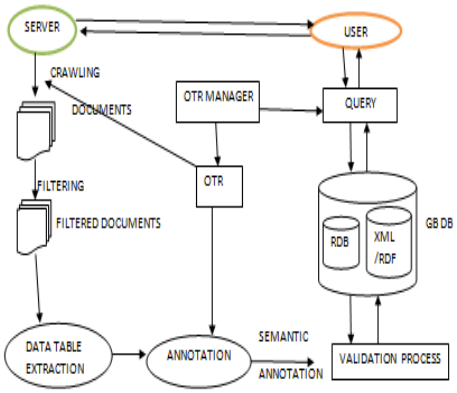

Fig1:EntireProcess

IV.MODULE DESIGN

1.Data Table Retreiving

Information in the web is presented in the form of data table.Data table is retrieved to present the information which is to be annotated .Data table which is retrieved consists of information as it is represented in the web.This has to be filtered to remove unnecessary information and to present the clear table content .Thus data table retrieved is created to form the clear content and it is maintained.

2.Annotation Of Data Table

A.Annotate Symbolic and numeric content:

The data tables extracted are semantically annotated with the use of OTR.Fuzzy annotations for the fuzzy extension of RDF are generated that is to be associated with the data tables. As a matter of fact, the main objective of the semantic annotation method is to identify which relations of the OTR are represented in a data table: those simple concepts are called in the following simple target concepts. It is to identify the simple concept of the OTR that corresponds to a column classified as a symbolic column and as a numerical one.The only simple concepts that appear in the signatures of the relations belonging to the OTR are considered.

.

$$score_{unit}(c,u) = \begin{cases} \frac{1}{\mid C_u \mid} & \text{if } c \in C_u \\ 0 & \text{otherwise} \end{cases}$$

All columns associated with the data table have been annotated with a simple concept of the OTR and we want to identify the relations of the OTR that are represented in the data table. It is used to compute the final score of each relation of the OTR for the data table that are identified. The title score of a relation for a data table is computed using, where title is the title of the data table. The content score of a relation r for the data table tab is the proportion of simple concepts in the signature of r which were represented by columns of tab. Let Sign be the set of simple concepts in the signature of r.

B.Validation:

The validation of the Fuzzy RDF semantic annotations associated with data tables are made by the end user before loading them in the XML/RDF data warehouse.This annotation makes the clear definition of the resources present in the data table which makes the user to obtain the clear information that is to be needed.

C.Find Relation:

The data table column represents the instance of a relation that relies on instances of target symbolic concepts and quantities of its signature.

3.Query RDF Data Sources

The end-user asks his/her query to 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 and a SPARQL query in the XML/RDF data warehouse .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 tables, represented in XML documents, by means of SPARQL queries. We remind the notions of view and query we then present the construction of an answer retrieved from the XML/RDF data warehouse. We conclude this section with experimental results.

.

V.CONCLUSION

This 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 taking into account the vocabulary used in the preexisting local databases in order to index the data and the domain information available within the databases schema. We define the system by completing the cosine similarity measure used to compare terms with other syntactical and semantic techniques, completing the semantic annotation of data tables in Web documents with the annotation of the text using the OTR and managing OTR evolution by taking into account annotation results.

REFERENCES

[1] Olufade, F. W. ONIFADE, Oladeji, P. AKOMOLAFE," A Fuzzy Search Model for Dealing with Retrieval Issues in Some Classes of Dirty Data"*IEEE transactions* on VOL. 2, NO. 11, October 2011.

[2] Comfort T.Akinribido, Babajide S. Afolabi, Bernard I. Akhigbe ,Ifiok J. Udo," A Fuzzy-Ontology Based Information Retrieval System for Relevant Feedback", *IEEE TRANSACTIONS ON IJCSI International Journal of Computer Science Issues*, Vol. 8, Issue 1, January 2011.

[3]RallouThomopoulos,PatriceBuche,OllivierHaemmerle,"*Representation of weakly structured imprecise data for fuzzy querying*", 2003Published by Elsevier Science B.V.

[4] Shawn Bowers and Bertram Lud¨ascher," *An Ontology-Driven Framework for Data Transformation in Scientific Workflows*", Springer-Verlag Berlin Heidelberg 2004.

[5] David M. Blei Michael I. Jordan," *Modeling Annotated Data*", 2003 ACM.

[6] Yang W. Lee, Diane M. Strong, Beverly K. Kahn, Richard Y. Wang," *AIMQ: a methodology for information quality assessment*", 2002 Elsevier Science B.V..

[7] S´ebastien Destercke and Patrice Buche and Brigitte Charnomordic,"Data reliability assessment in a data warehouse opened on the Web",*IEEE Transactons* on 2011.

[8] Patrice Buche, Juliette Dibie-Barthelemy, Liliana Ibanescu, and Lydie Soler "Fuzzy Web Data Tables Integration Guided by an Ontological and Terminological Resource" *IEEE transactions on knowledge and data engineering*, vol. 25, no. 4, april 2013.