Ontological Semi Automatic Method for Web Data Tables Integration

B.Gowthamipriya Darsini
M.Tech(SWE),Dept.of CSE
Kakatiya Institute of Technology & Science
Warangal-15,Telangana,India

B.Hanmanthu
Assistant Professor, Dept.of CSE, Kakatiya Institute of Technology & Science
Warangal-15, Telangana, India

ABSTRACT
In this paper, we are presenting a model for implementation of ontology through invoking a semi- automatic process from web data tables. The singularity of this method is the implementation of cosine similarity for filtering the documents. We present the framework for the development of generic method concerning data integration. The data warehouse is composed of several data tables extracted from the web and it has been supplemented by the existing local data sources. Documents pertaining to different domains can be supplemented as input to the data sources which are stored in data ware house and then are converted into XML/RDF data. We have used cosine similarities to measure the similarity of two documents which are likely to be same in terms of their subjects. It is a semi automatic method to extract and integration of web documents by using ontology.

Keywords
Ontology, XML/RDF, SPARQL, Cosine similarity.

1. INTRODUCTION
In this paper, a model for an Ontological and Terminological Resource (OTR), dedicated to the task of n-array relations annotation in Web data tables is studied. This task relies on the identification of the symbolic concepts and quantities, defined in the OTR, which are represented in the tables' column. We propose to guide the annotation by an OTR because it allows a separation between the terminological and conceptual components and allows dealing with abbreviations and synonyms which could denote the same concept in a multilingual context. The OTR is composed of a generic part to represent the structure of the ontology dedicated to the task of n-array relations annotation in data tables for any application and of a specific part to represent a particular domain of interest. In this study the similarity is measured based upon Cosine similarity. We present the model of our OTR and its use in an existing method for semantic annotation and querying of Web tables. In this Model we use two subsystems. They are as follows

1. Web sub system.
Web sub system is used to load the XML/RDF data tables into the data warehouse which are retrieved from web documents and are annotated by using Ontological and terminological resources.

2. Querying sub system
It is used to retrieve the efficient results by using ontological and terminological resource uniformly. We can retrieve syntactically and semantically very close results by using this method.

2. RELATED WORK
Many methods are present in existing literature to annotate and query Web data tables by using domain ontology. Converting and Annotating Quantitative Data Tables for reuse is also proposed [5]. LexInfo allows us to associate linguistic information with respect to any level of linguistic description and expressivity to elements in ontology. And fuzzy web data tables are integrated and guided by using OTR. In these methods, symbolic concepts and numerical concepts are considered and the methods to instantiate concepts and relations, and querying the results are elucidated clearly [1][5][6][7][8]. In fuzzy web data tables integration we use ONDINE [1] system having web sub system and domain sub system. We also create ontological and terminological resource. In this method we use semi automatic ontology in order to retrieve the data from web. We propose an original method to assess Web data table reliability from a set of criteria by using evidence theory. Finally, we can understand the method to extend the workflow to integrate the reliability assessment step [9]LexInfo has been implemented as OWL ontology and is available together with an API. Our main work is the model and similarity itself, but importantly a clear motivation for more elaborate models for associating linguistic information with ontology [10].

3. PROPOSED WORK
The Proposed work 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

3.1 User Query
In this user query, a web application is designed to maintain the unique features of our querying system, which are:

1. To extract both exact answers and semantically related answers to the given query.
2. To annotate the results based upon fuzzy logic and to calculate the similarity index.

The process is semi-automatic because the user has to upload the related documents in which the querying has to be executed.

3.2 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.

3.3 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 (x) is usually calculated by the underneath formula

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

3.4 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.

3.5 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

1. To consider the exactness linked to the relations characterized in the data tables and

2. To evaluate a fuzzy set representing the querying preferences with respect to a fuzzy set having a semantic of similarity or imprecision.

3.6 User's Integrated Output

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

1. To annotate the Web data tables which are stored in data ware house with the vocabulary defined in an OTR,

2. 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.

4. DATA FLOW

The user is directed to a query page initially during which he is prompted to request for a query. The query may be a key word or a fuzzy set, based upon the desired set of results, expected by the user. The user participation in the process ends hereafter that the query is transferred to the OTR resource in which the symbolic concepts and semantically close terms are already enlisted. The queried keyword or fuzzy sets is/are related and identified to their respective symbolic concepts and then the modified query is searched in the uploaded web data tables. The similarity index for various web documents is calculated using cosine similarity and the tables are annotated by the described annotation method. RDF graphs are generated and are employed in annotating the XML data tables. Both the similarity indices and the xml data tables of the related web documents are presented to the user for further processing.

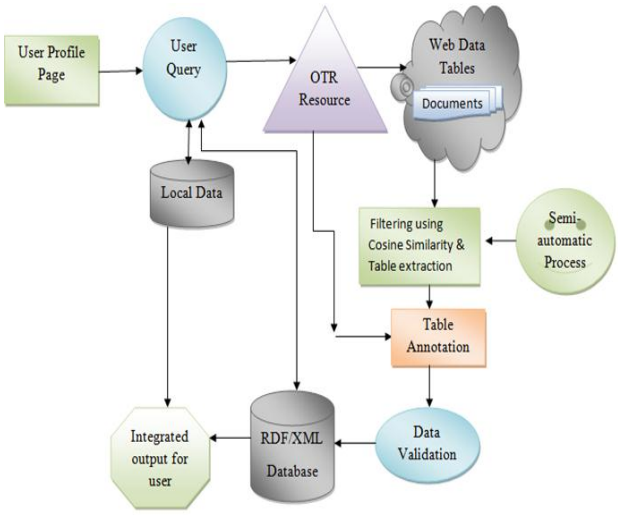

Fig1: System Architecture

From Fig 1: Shows that the user initially uploads the web documents which are having web data tables that is to be inputted to the process. And then the documents are stored in data ware house or local data sources. The web documents are converted into XML/RDF documents by using Fuzzy Ontology. By using semi automatic process, we extract web data tables and filtering can be done by using cosine similarity measure. Finally the valid documents in XML are retrieved by searching and similarity measure can be done. The results obtained indicate that the method is very relevant and it gives high accuracy and precision. In this process, we used Numerical concepts, symbolic concepts and terms, relations, SPARQL and web data tables in the background. The accuracy and precisions value graph is shown below. The accuracy is the degree of closeness of measurements of a value to that value's actual (true). The precision is related to repeatability and reproducibility, is the degree to which repeated measurements under unchanged conditions show the same results [3][4]. These methods are developed by using JAVA technology and Serve let pages, HTML and querying Language. These processes are done by using Tomcat Server.

From Fig 2 we identify the results are in high precision with best accuracy.

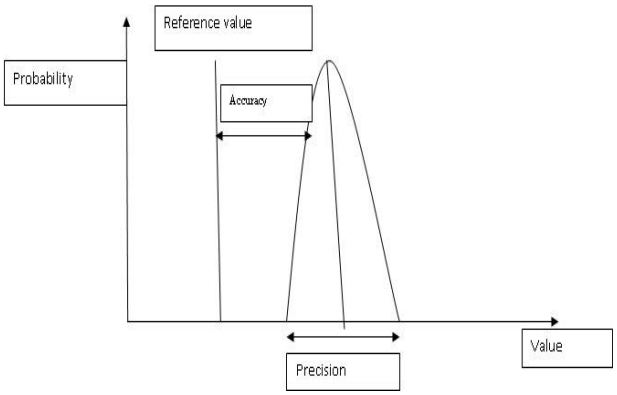

Fig 2: Precision and accuracy

5. ILLUSTRATION OF THE PROPOSED MODEL

As an illustration, web data tables pertaining to four different domains are considered. The four domains are Bacterial Reference table, Countries, Temperature conversion (named as Temperature), Temperature wiki.

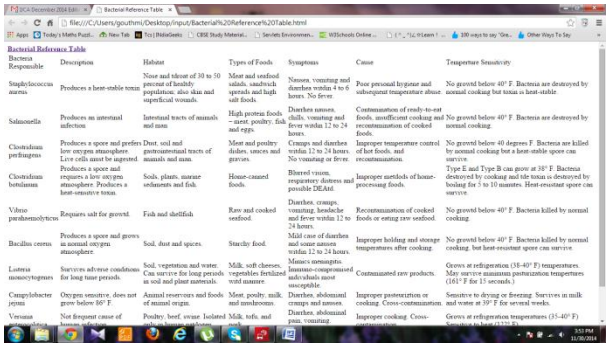

Fig. 3: Web document

Fig.3 web documents shows that document which is having web document that is converted into XML/RDF and stored in data ware house. Fig 3 is example of Bacterial reference table. First the word "Salmonella" is searched, which belongs to the bacterial domain. It can be seen from the corresponding results, that the similarity for Bacterial Reference table is displayed as the remaining documents do not contain the searched word.

Fig.4: Search salmonella

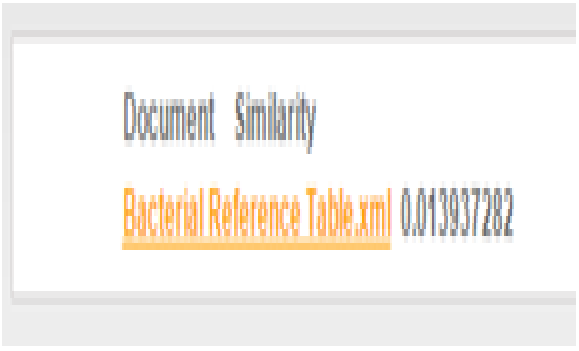

Fig.5: Result for salmonella

The word "Temperature" is searched now in the same web data tables. Now it can be seen, except Countries data table remaining tables are enlisted as they contain the word "Temperature". And the increasing value of similarity indicates the increase in the appearance of the word in the related documents.

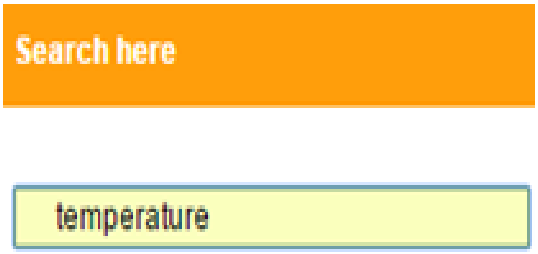

Fig. 6 Search temperature

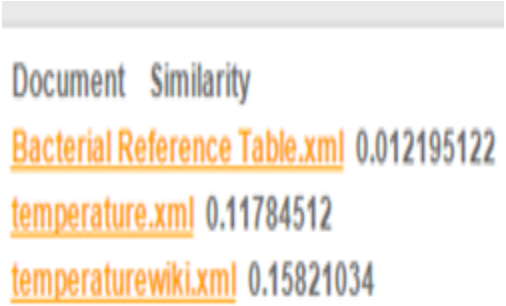

Fig.7: Result for temperature

Finally the word "temp" is searched in the same documents. The related snaps are shown below. As you can see the similarity value are exactly same as that of earlier case.

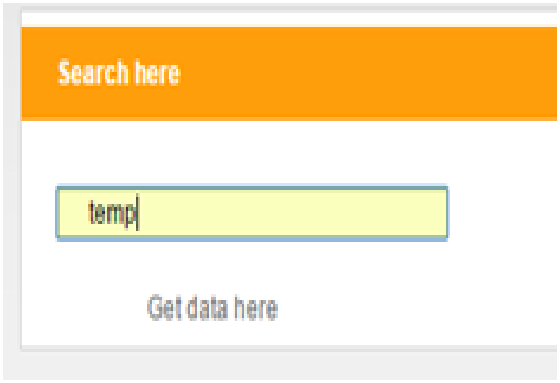

Fig.8: Search as temp

Fig.9: Result for temp

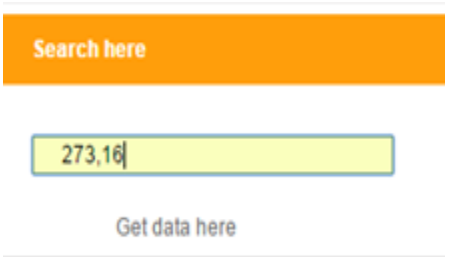

Fig.10: Numerical concepts

Fig.11: Result for numerical concepts

We can get the results for both symbolic concepts and numerical concepts. Examples for symbolic concepts are shown fig3, fig5 and fig 7. The results for Fig 3, 5, 7 are shown in Fig4, 6, 8. And the examples for numerical concepts are shown in fig9 and result for numerical concepts are fig10. So, we can get results with similarity. So the results are in expected accuracy. This is expected because when a keyword is queried; all the related phrases and symbolic constants are searched in the web documents. This shows that if a relation is present in the OTR, identical results are obtained and vice versa. This demonstrates that there is no need for searching all the different terminologies used for the queried keyword individually. Any one word in relation would result the same output as that of all keywords together which is the prime motive behind ontology.

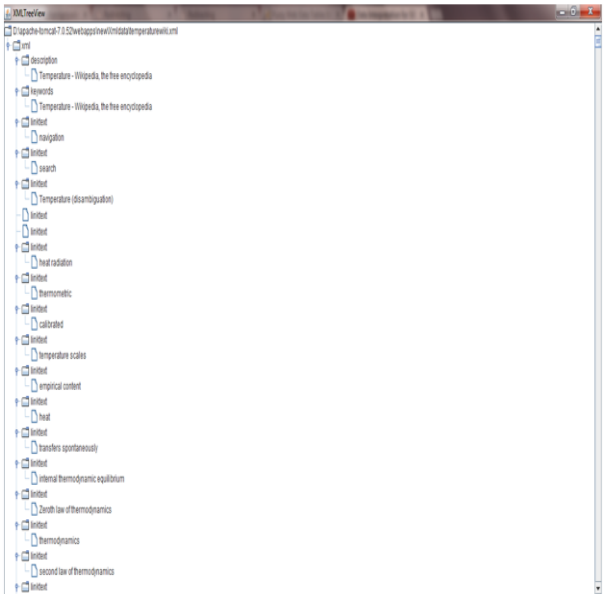

Fig12: XML tree view

The result of any search is a web document which is a XML document. If we click on the xml document view option, we get results in XML tree view. It contains the root value as XML which indicates the parent values and child values. By knowing the parent value we can search or extract the data very precisely.

6. PERFORMANCE GRAPH

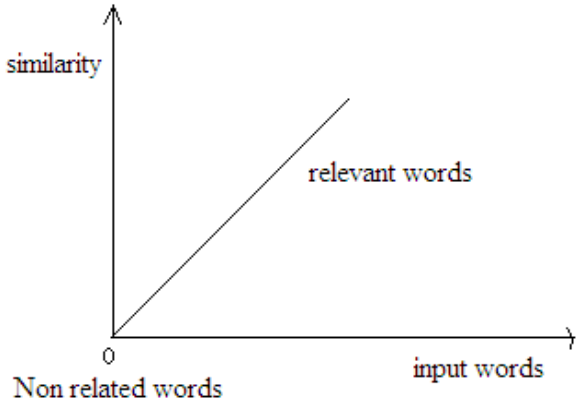

Fig .13 Performance graph

Figure represents how the word should be displayed according to data availability. If the input word does not exist within the database, then graph line points at origin, if any word exists in the data base, it will infer to a point other than origin in the graph. The more the slope of the line, the more is the similarity between the documents.

7. CONCLUSION

In this paper, we extract the documents using text. The texts in documents are in XML/RDF. By using ontology we load and query the web data tables and integration. We provide the cosine similarity measure for finding the value of similarity between two documents. We can also develop a database by using exact words from a piece of text, stem or lemmatize it, etc. and we can also annotate only relevant data. We can also integrate the attributes in web data tables. By grouping the attributes we can get results with high precision and accuracy. We integrate web data tables in web documents and store in data ware house. Converted documents are displayed in Xml format. Those XML documents are developed by using n array relationships in web data tables. A complete quasi automatic ontological method to integrate the web documents, which performs with high accuracy, is presented in this paper.

8. FUTURE SCOPE

Developing an efficient and robust database which contains all the possible synonyms or related symbols of various domains, and linking it to our proposed method will make the process automatic. Thus the performance can be increased when compared to our present system.

9. ACKNOWLEDGMENTS

Our sincere thanks to the Principal and management members of Kakatiya Institute of Technology and Science, Warangal, who have facilitated the resources to read, compute and develop the project. We like to convey our sincere thanks, to the Head of the department, Dr.P.Niranjan, who had encouraged and guided us towards research and publishing of this paper.

10. REFERENCES

[1] 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.

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

[3] JCGM 200:2008 International vocabulary of metrology — Basic and general concepts and associated terms (VIM)

[4] John Robert Taylor (1999). An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements. University Science Books. pp. 128–129. ISBN 0-935702-75-X.

[5] Mark van Assem, Hajo Rijgersberg, Mari Wigham and Jan Top; Converting and Annotating Quantitative Data Tables in 2009.

[6] G. Hignette, P. Buche, J. Dibie-Barthe´lemy, and O. Haemmerle´, "Fuzzy Annotation of Web Data Tables

[7] Driven by a Domain Ontology," Proc. Sixth European Semantic Web Conf. The Semantic Web: Research and Applications

[8] 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.

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

[10] P. Cimiano a;_, P. Buitelaar b, J. McCrae a, M. Sintek; LexInfo: A Declarative Model for the Lexicon-Ontology Interface in 2009.

[11] 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 ScienceIssues.

11. AUTHOR'S PROILE

B.Gowthamipriya Darsini is currently pursuing Master of Technology in Computer Science and engineering with specialization in Software Engineering.Computer Science and Engineering Department, Kakatiya Institute of Technology & Science (KITS), Kakatiya University-Warangal, Telangana, India.

B.Hanmanthu obtained his Bachelor's degree in Computer Science and Engineering from JNT University Hyderabad. Then he obtained his Master's degree in Computer Science and Engineering with specialization Software Engineering from JNT University Hyderabad, and he is also life member of ISTE. He is currently Assistant Professor of Computer Science and Engineering, Kakatiya Institute of Technology & Science (KITS), Kakatiya University-Warangal, Telangana, India. His specializations include Data mining and Data warehousing, Network security, Software Engineering.