Ontology Based Approach For Instance Matching

M.Preethi, R.Madhumitha

Abstract ___One of the important barrier that hinders achieving semantic interoperability is ontology matching. Instance-based ontology matching (IBOM) or concept based ontology matching(CBOM) uses the extension of concepts, the instances directly associated with a concept, to determine whether a pair of concepts is related or not. Practically, instances are often associated with concepts of a single ontology only, rendering IBOM rarely applicable. This is achieved by enriching instances of each dataset with the conceptual annotations of the most similar instances from the other dataset, creating artificially dually annotated instances. We call this technique concept based ontology matching by concept enrichment (CBOMbCE). We are using the instance matching process with web crawlers mediating three world's leading publishers such as Oxford, ScienceDirect and Springer. We are obtaining keywords from the articles of these four journals which acts as the instances. We are collecting all possible journals available in these three websites since the access permission of these three journals can be restricted to some constraints within it. After searching and finding keywords those instances are matched with their ontology creation and further enrichment of instances. Through this technique we will obtain instances that are uncommon among two datasets.

Keywords __Ontology Matching, Web crawler, Concept Enrichment

I.INTRODUCTION

Semantic interoperability is a requirement to enable machine computable logic, inferencing, knowledge discovery, and data federation between information systems.__This is accomplished by linking each data element to a controlled, shared vocabulary. The Semantic web is nothing but a web with a meaning. It is a group of methods and technologies. It is the total formula of searching, aggregating and combining the web information. It is a logical method of accessing meaningful and accurate information. Data are interlinked. The Semantic Web is an idea of World Wide Web that the Web as a whole can be made more intelligent and perhaps even intuitive about how to serve a user's needs. The goal of Semantic Web Services is to enable dynamic, execution-time discovery, composition, and invocation of Web Services. Ontology matching has taken a critical place for helping heterogeneous resources to interoperate. Ontology alignment tools find classes of data that are semantically equivalent.

The new proposed system works on application of journals extracting the concept of concept enrichment.

2 LITERATURE REVIEW

2.1 A VECTOR SPACE MODEL FOR AUTOMATIC INDEXING

In document retrieval or other pattern matching environment where stored entities(documents) are compared with each other or with incoming patterns(search requests)it appears that the best indexing (property) space is one where each lies far away from the others as possible. An approach based on space density computations is used to choose an optimum indexing vocabulary for a collection of documents.

2.2 DOCUMENT SPACE CONFIGURATION

Clustered centroid is a typical clustered space where the various document groups are represented by circles and the centroids by black dots located more or less at the center of the respective clusters.The main centroid represented by a small rectangle in the centre may then be obtained from the individual documents. The main centroid of the complete space is simply the weighted average of the various cluster centroids. The average similarity between document pairs is smallest, thus guaranteeing that each given document may be retrieved when located sufficiently close to a user query without also retrieving its neighbors. This insures a high precision search output, since a given relevant item is then retrievable without also retrieving a number of non relevant items in its vicinity.Overlapping occurs only when similarity between terms occurs. Term frequency (TF) is the ratio number of times the word has occurred in a document by its document size. Inverse document frequency (IDF)is the ratio of logarithm of size of dataset to the total number of documents.

Schema matching aims at identifying semantic correspondences between metadata structures or models, such as database schemas, XML message formats, and ontologies. Solving such match problems is a key task in numerous application fields, in particular to support data exchange, schema evolution and virtually all kinds of data integration. Unfortunately, the typically high degree of semantic heterogeneity reflected in different schemas makes schema matching an inherently complex task. Hence, most current systems still require the manual specification of semantic correspondences, e.g. with the help of a GUI. While such an approach is appropriate for matching a few small schemas, it is enormously time-consuming and error prone for dealing with large schemas encompassing thousands of elements or to match many schemas. Matching large XML schemas, e.g. e-business standards and message formats. Matching large life science ontologies describing and categorizing biomedical objects or facts such as genes, the anatomy of different species, diseases, etc. Matching large web directories or product catalogs. Matching many web forms of deep web data sources to create a mediated search interface, e.g. for travel reservation or shopping of certain products.

3.WEB CRAWLER

Web Crawler is a meta search engine that blends the top search results from Google Search and Yahoo Search. WebCrawler also provides users the option to search for images, audio, video, news, yellow pages and white pages. A web crawler (also known as a web spider or web robot) is a program or automated script which browses the World Wide Web in a methodical, automated manner. There are several uses for the program, perhaps the most popular being search engines using it to provide webs surfers with relevant websites. Other users include linguists and market researchers, or anyone trying to search information from the Internet in an organized manner.

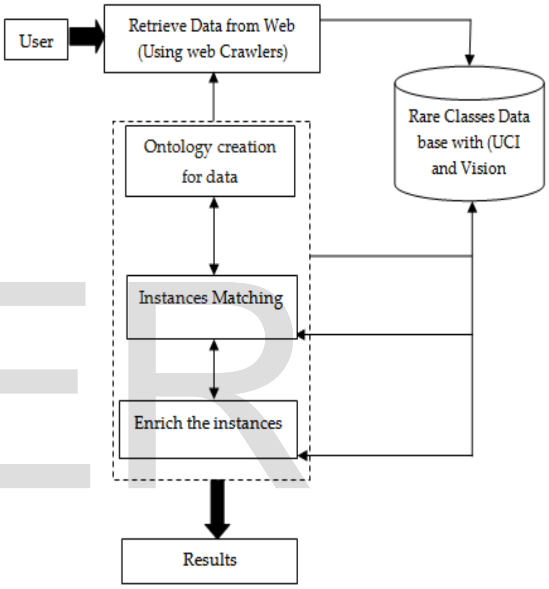

Figure 1 Systematic Data Flow Diagram

4. DATA PREPROCESSING

Data Preprocessing is a Computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. In this process we use web crawlers to retrieve online data from web. A Web crawler is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing. A Web crawler starts with a list of URLs to visit, called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the crawl frontier. URLs from the frontier are recursively visited according to a set of policies. Here, we use crawler to search the titles of journals from Oxford,Springer, ScienceDirect.

5.ONTOLOGY CREATION

We have already created dataset. That contains information about the journals and articles. We want to create ontology for every data by using following steps. Organizing and Scoping. The organizing and scoping activity establishes the purpose, viewpoint, and context for the ontology development project, and assigns roles to the team members. In the paper the ontology design of paper title, author, volume and publication are taken into account. During data collection, raw data needed for ontology development is acquired. Data analysis involves analyzing the data to facilitate ontology extraction. The initial ontology development activity develops a preliminary ontology from the data gathered. Ontology Refinement and Validation is done at the final stage. Here validation and comparison of final titles of the articles are considered .The ontology is refined and validated the ontology to complete the development process.

6. INSTANCES MATCHING

In this process we need to determine which instance(s) actually is (are) most similar. Instances Matching (IM) algorithms are required that use features to predict similarity between objects. The Vector Space model provides an abstract model, where journals are represented as vectors of features (in our case words) in a vector space. The similarity between two journals is quantified by the cosine similarity:

Algorithm of concept matching with lexical similarity

```
String labels S,S' representing concept names in ontology $O_1$,$O_2$

$\text{Similarity}_{\text{lexical}}$(S,S')->Sim(lex)// measure of accurateness

Function lexical similarity

Thesaurus->∑,lexical similarity measure->$M_0$

If similarity threshold value matches->1, unmatches->0

For each string query word, DO

For each string query word, DO

If w=w' THEN

Total threshold value->1

else

if w=w' Then

total threshold value->0

ENDIF
```

Algorithm of concept matching with semantic similarity

```
$\text{Similarity}_{\text{semantic}}$(S,S')->Sim(sem)// measure of semantic relatedness

Function semantic similarity

Thesaurus->∑, semantic similarity measure->$M_0$

If similarity threshold value matches->1, unmatches->0,relativeness->0.1…..0.9

For each string query word w, DO

For each string query word w', DO

If w=w' THEN

Total threshold value->1

else

If w%w'(related with w'semantically) THEN
```

```
Total threshold value->0.1 to 0.9 e,lse
if w=w' Then
total threshold value->0
ENDIF
ENDIF
END
```

7. ENRICH THE INSTANCES

The enrichment of instances depends on replacing the concepts of both concepts that are to be compared. During the instance enrichment process the same top N results and similarity threshold parameters are used so that the results are to be found effectively even the concepts are unrelated between them. Top N gives us the top results and similarity threshold shows us the similarity threshold value starting from one 0.1 to 0.9. hence the results predict the availability of journals in all the sets of data to be enabled.

calculations.

String 1	String 2	Lexical similarity measure algorithm	Semantic similarity measure algorithm
Royal	Royal-	1	1.0
Royal	monarchal	0	-0.25
Royal	kingly	0	0.0003046
Royal	princely	0	0.0555
Royal	aristocratic	0	0.000029733
Royal	lordly	0	0.000029733
Royal	noble	0	0.0020833
Investment	Investment	1	0
Investment	Envelope	0	0.05
Investment	Investiture	0	0.001322
Investment	Blockade	0	0.0002506
Investment	Siege	0	0.00471

Graphical results of tabulation

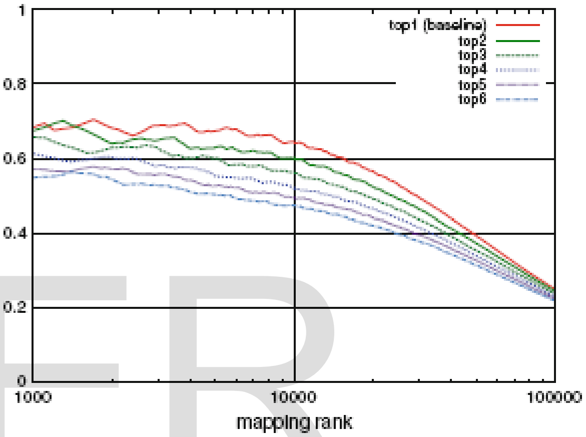

Fig 2: results of graph based on rankings

VIII.CONCLUSION

Thus the paper gives us good results because of getting information from the all four journals and hence matching takes place between uncommon dataset ie) articles. Enrichment of instances is new to the topic of journal data retrieval and the efficiency of matching will get increased. Term frequency calculates the occurrences of keywords which shows its importance or weightage of the instance in the paper. Thus user get results while typing the need or the particular word the results will be displayed like Google Instant and the user can select from the choices. The choices are the papers that the user wishes to view. This can be achieved by finding the relationship among all the aspects about a particular title. Overall performance of the retrieving results will get increased by using CBOMbCE algorithm

REFERENCES

1. Isaac A, vann derMeij L, Schlobach S,Wang S (2007) An empirical study of instance-based ontology matching. In: ISWC/ASWC,pp 253–266. Rahm E (2011) Towards large-scale schema and ontology matching. ReCALL 5:1–26. http://www.springerlink.com/index/ M5055K8721752228.pdf.

2. Rahm E, Bernstein PA (2001) Asurvey of approaches to automatic schema matching. VLDB J 10(4):334–350 25. Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. CommunACM18(11):613–620. doi:10.1145/ 361219.361220

3. Wang S, Englebienne G, Schlobach S (2008) Learning concept mappings from instance similarity. In: International semantic web conference, pp 339–355

4. Leme LAPP, Casanova MA, Breitman KK, Furtado AL (2009) Instance-based owl schema matching. In: Filipe J, Cordeiro J (eds) Enterprise information systems, Proceedings of 11th international conference, ICEIS 2009, Milan, May 6–10. Lecture notes in business information processing, vol 24. Springer, Berlin, pp 14–26. doi:10.1007/978-3-642-01347-8_2

5. Wang S, Isaac A Schopman B, Schlobach S, van der Meij L (2009) Matching multi-lingual subject vocabularies. In: Proceedings of the 13th European Conference on Digital Libraries (ECDL2009)

6. Wartena C, Brussee R (2008) Instanced-based mapping between thesauri and folksonomies. In: ISWC'08

7. Zaiss KS (2010) Instance-based ontology matching and the evaluation of matching systems. Ph.D. thesis, Heinrich Heine UniversitätDüsseldorf

8. Stumme G, Maedche A (2001) Fca-merge: bottom-up merging of ontologies. In: Proceedings of the 17th international conference onartificial intelligence (IJCAI '01), Seattle, pp 225–230

9. Thor A, Kirsten T, Rahm E (2007) Instance-based matching of hierarchical ontologies. In: Kemper A, Schning H, Rose T, Jarke M, Seidl T, Quix C, Brochhaus C (eds) BTW, LNI, GI, vol 103,pp 436–448. http://dblp.unitrier.de/db/conf/btw/btw2007.html#ThorKR07

10. Todorov K, Geibel P (2009)Variable selection as an instance-based ontology mapping strategy. In: ArabniaHR, MarshA(eds)SWWS.CSREA Press, USA, pp 3–9

11. Todorov K, Geibel P, Kuhnberger KU (2010) Mining concept similarities for heterogeneous ontologies. In: Perner P (ed) ICDM.Lecture notes