ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Concept Based Ontology Matching By Concept Enrichment

Preethi. M[1]

Post Graduate in Software Engineering, Sri Krishna College Of Engineering And Technology Coimbatore, Tamilnadu, India[1]

ABSTRACT: One of the important barrier that hinders achieving semantic interoperability is ontology matching. Instance-based ontology matching (IBOM) uses the extension of concepts, the instances directly associated with a concept, to determine whether a pair of concepts is related or not. In practice, however, 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 instance-based ontology matching by instance enrichment (IBOMbIE). We are using the instance matching process with web crawlers mediating four world's leading publishers such as Willey, Oxford, ScienceDirect and Springer. We are obtaining keywords from the articles of these four journals which acts as the instances. We are particularly considering ARTIFICIAL INTELLIGENCE and COMPUTER NETWORKS since these four journals consists of huge database regarding articles 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.

I. INTRODUCTION

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 inventor Tim Berners-Lee 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.

WEB CRAWLER

WebCrawler 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. Alternative names for a web crawler include web spider, web robot, robot, crawler, and automatic indexer.

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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

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.

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. During data collection, raw data needed for ontology development is acquired. Data analysis involves analyzing the data to facilitate ontology extraction. The intitial ontology development activity develops a preliminary ontology from the data gathered. Ontology Refinement and Validation is done at the final stage. The ontology is refined and validated the ontology to complete the development process.

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 documents are represented as vectors of features (in our case words) in a vector space. The similarity between two documents is quantified by the cosine similarity:

IDF_local use the local word distribution of w to calculate the IDF value of w, (i.e) when w is part of a document in dataset D1 we consider the word distribution of D1 to calculate IDF (w). In the IBOMbIEalgorithm there are always two word distributions: the word distributions of D1 and D2.

ENRICH THE INSTANCES

There are two crucial parameters of the IE process: the topN and the similarity threshold (ST) parameters. Tuning these two parameters may have a significant influence on the quality of the end result. The topN parameter defines from how many instances we add the associated concepts to the instance that will be enriched. A larger value of the topN parameter means that instances will be enriched with more concepts. Therefore, a larger N causes more concept associations to be created, resulting in a higher number of mappings generated by applying JCc and thus a final result with a potentially higher coverage.

II. RELATED WORKS

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.

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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

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 quaranteeing 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.

III. PROPOSED SYSTEM

This paper presents an extensional ontology matching method that works in the absence of dually annotated corpora, and assesses the viability of the method in a specific use-case, where we show that it can be a very useful extension of existing methods. Given problem-driven approach, driven by a real-world application in the library domain that started this line of research, we focus on technical aspects of the approach, rather than performing a broad, domain-cross comparison.

This paper extends previous work in two ways: we apply the method introduced in on a large-scale, we are using *web crawlers* to check the URL of *four international publishers*, and second, exhaustively evaluate the possible parameters of the algorithm using two different ways of evaluating the matching results. In addition to the evaluation results in we consider this sufficient proof for the power of IBOMbIE for Ontology Matching, especially when instances are available of which the similarity can be measured, as is the case in our application domain.

IV. 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 IBOMbIE algorithm.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

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 in computer science, vol 6171. Springer, Berlin, pp86–10.