INTERNATIONAL JOURNAL OF SCIENTIFIC & ENGINEERING RESEARCH, VOLUME 6, ISSUE 5, MAY-2015
ISSN 2229-5518

Ontology Based Approach for Semantic Similarity between Abstracts of Research Papers

Madhumitha.R, Assistant Professor/CSE, Shahanas Thasleem.A.G ,PG Scholar,CSE, Sri Krishna College of Engineering and Technology

Abstract — One of the challenging issues in paragraph similarity is ontology construction and ontology matching. An example taken for paragraph similarity is to find matching between abstracts of research papers. Generally researchers use keywords for searching in search engine for finding similar meaning for that particular keyword but in this project the abstract has been used for searching and finding the abstracts.
The main objective of this project is to find the semantic similarity between abstracts of research papers. Here ontology plays an increasingly important role in knowledge management and the Semantic Web. In this approach the domain ontology as well as abstract ontology is generated. Domain ontology is constructed based on the file which contains domains and terms. Abstract ontology is constructed based on the abstracts which are extracted from WebCrawler. After building the two ontology (domain ontology and abstract ontology), ontology matching is performed to match the concepts, then the similar abstracts will be obtained in that particular domain and thus users gives input (abstracts) in order to obtain the similar abstracts. This project will be helpful for the researcher who searches the similar abstracts existing in their domain.

Index Terms— Semantic web, Ontology, Ontology Matching, TF-IDF, WordNet, POS, K-Mean, WebCrawler.

1 INTRODUCTION

In order for the Semantic Web participants to share information, they must have some agreement on what elements in their shared domain of interest exist and how these elements can relate to one another. A formal specification/conceptualization of such an agreement is called ontology. Ontology for a domain gathers and gives semantic descriptions of concepts in the domain of discourse, defining domain-relevant attributes of concepts and various relationships among them. For example, an ontology describing a car will include such concepts as wheel, batteries, and so on. It will also include relations such as manufactured by, year, color and type of fuel.

1.1 Reasons to develop Ontology

(i).Based on people's common understanding of the structure of information or software agents will be shared - This is one of the more common goals in developing ontologies.

(ii).Domain knowledge can be reused – This was one of the driving forces behind recent surge in ontology research.

(iii).*Explicit domain assumptions can be built* -Underlying an implementation makes it possible to change these assumptions easily if our knowledge about the domain changes.

(iv*).From the operational knowledge*, domain knowledge can be separated- This is another common use of ontologies.

(v).*Domain knowledge should be analyzed* -This is possible once a declarative specification of the terms is available.

1.2 Problem Definition

Searching the web has become very important role in day to day life. A user either searches for specific data or just simply browses topics of their own interests. Typically, a user enters a query or it may be set of keywords, after a search engine answers with set of document which are relevant to that query. Then users have to go through the information that is relevant to the researchers. Only some part of query contains relevant information. Instead if a user has domain ontology and gives a paragraph or an abstract as input i.e., text document, the particular content will match with ontology and gives the result. Even though the ontology is used, the researchers will not get good result because ontology and text document will not a generate a proper result. If the particular research needs good performance, then the researches need to construct two ontologies and then matching should be performed.

1.3 Ontology Matching

By using concept similarity, related concepts from different ontologies are found. The phrase takes on a slightly different meaning, in computer science, philosophy or cognitive science. It aims at finding the semantic mappings between two given ontologies. The application that involves multiple ontologies must establish semantic mappings among them, in order to ensure interoperability. Examples of such applications arise in myriad domains, including e-commerce, e-learning, knowledge management, information extraction, tourism, web services, and bio-informatics.

As shown in Figure 1.a. Here the data is organized into a taxonomy that includes people and courses. Professors have attributes such as name, degree, and granting institution. Such marked-up information finds it easy for the softbot to find a professor with the last name Cook. Once the attribute "granting institution" was found, then the softbot quickly starts identifying the alma mater CS department in Australia. Here the work of softbot is to identify the data has been marked up using an ontology specific to Australian universities, then in Figure 1.b, there are many entities named Cook. However, knowing that "associate professor" is equivalent to "senior lecturer", obviously the bot will select the right subtree in the departmental taxonomy, and zoom in on the past homepage of your conference acquaintance.

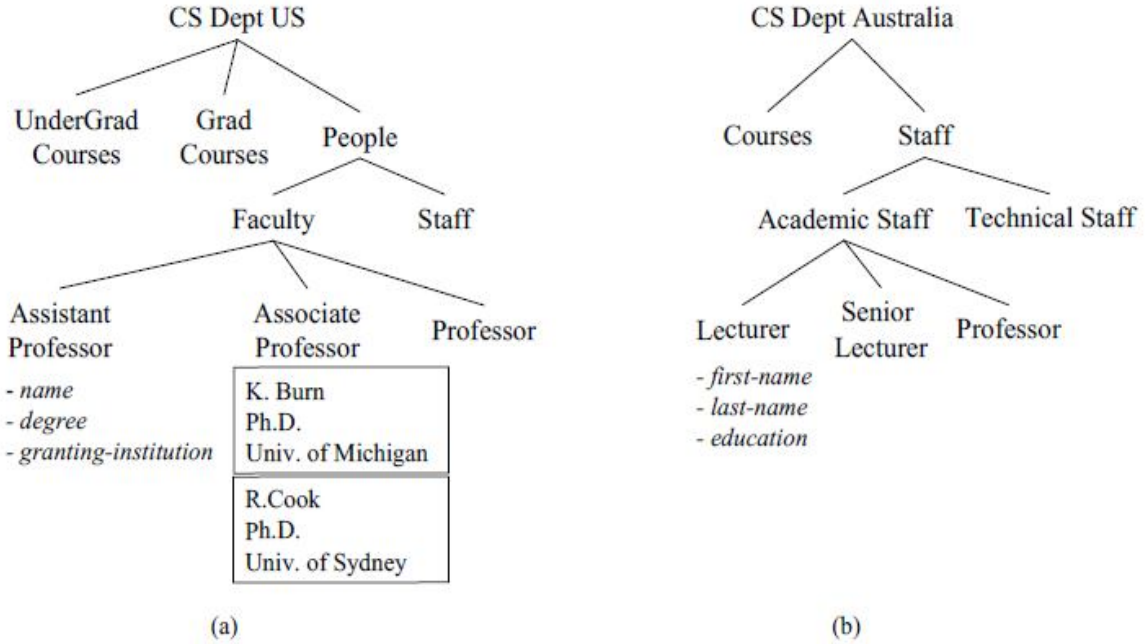

Figure 1: Ontology Matching

2 PROPOSED WORK

Searching the web has played an important role in human life in the past couple of years. A user either searches for specific information or just browses topics on their interest. Typically, a user enters a query in natural language, or as a set of keywords, and a search engine answers with a set of documents which are relevant to the query. Then, the user needs to go through the documents to find the information that interests researchers. However, usually just some parts of the documents contain query-relevant information. Here if the users gives a paragraph or abstract instead of a keyword then that particular abstract maps with domain ontology. Suppose similar concepts are available in that abstract then finally researchers will get similar abstracts.

Ontology is playing an increasingly important role in knowledge management and the Semantic Web. In this project, the domain ontology as well as abstract ontology will be generated. Domain ontology is constructed based domains and terms which are created by the researchers. Researchers can construct domain ontology based on their needs or requirement.

Abstract ontologies are constructed from the abstracts which are extracted from WebCrawler. The contents of proposals are usually unstructured. The research ontology is used to analyze, extract, and identify the keywords in the full text of the proposals. Finally, a further reduction in the vocabulary size can be achieved through the removal of all words that appeared only a few times in all proposal documents. TF-IDF encoding describes a weighted method based on inverse document frequency (IDF) combined with the term frequency (TF) to produce the feature. Thus, research papers can be represented by corresponding feature vectors. Based on the features in the text document, abstract ontology is constructed. Semantic meaning of the features in the text is calculated by using the WordNet tool. Based on the terms (keywords) and maximum count of terms, the domain ontology is matched with abstracts. Finally the abstract ontology is generated from the text.

After building the two ontology (domain ontology and abstract ontology), ontology matching are performed. The concept similarity measure, which we designed in the proposed method stands as the signature to this method [12]. The concept similarity measure is used to find related concepts from the different ontologies.

Abstract Matching is performed based on inputs given by the researchers. The matching is mainly performed based on terms and keywords matching are performed and hence similar Abstracts are obtained.

2.1 Extraction Of Url From WebCrawler

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 automated, methodical manner. This process is called spidering or Web crawling. Many legitimate sites, in particular search engines, up-to date data are provided by spidering.

In this project, WebCrawler is used to retrieve content of particular website[15] i.e., http://ieeexplore.ieee.org/. The content may be pdf files, documents, images etc i.e., entire data of that particular website. In order to retrieve the whole content of the any particular website, the depth of the search should be declared [14]. Only then the content can be retrieved up to the depth limit. Sometimes there may be some problem due to continue retrieval of content from the particular website. So here sleep time should be given. This sleep time is used mainly to give some break between one url to another url. The outcome of this module will be the content of specific url which has been given by users.

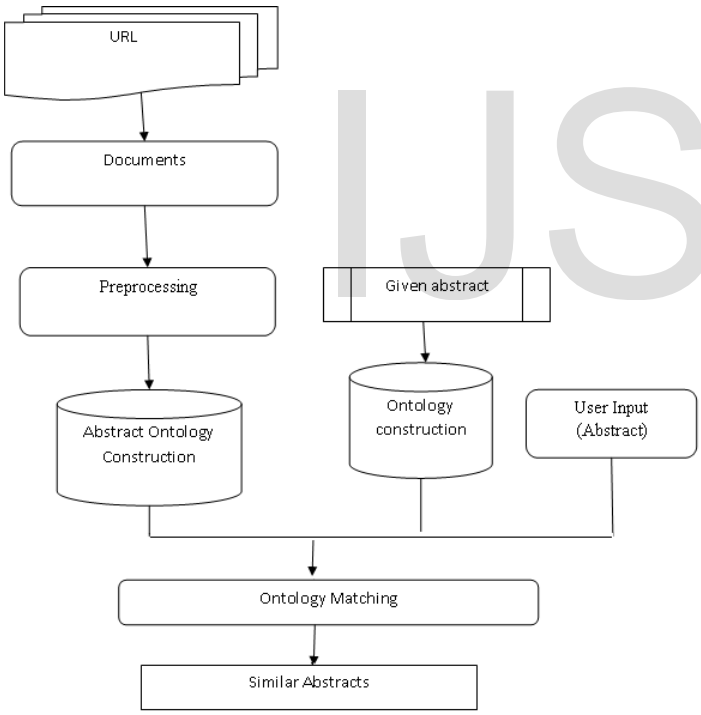

Figure 2: Block Diagram of Proposed Methodology

2.2 Preprocessing

For pre-processing, the keywords are taken from the respective urls [14] .After finding the keywords from the urls, then the content of that particular keywords are viewed. Once the content is obtained from the first step of preprocessing.

Next step the content should be processed i.e., only the abstracts are taken from that whole research papers. These abstract are made to be displayed on WebCrawler. Once the abstracts are obtained in WebCrawler and again this saved in notepad along with the title, author, publisher, copywriters.

2.3 Domain Ontology Construction

Domain Ontology is constructed based on the domains which are already saved in a file. For example, file contains domains and terms i.e., domain is Data mining and terms are cluster, classify, preprocessor, mining. In similar way, the Researchers can add domains and terms which are required to their research and save the file. Based on the file domain ontology is constructed.

2.4 Abstract Ontology Construction

In order to construct the abstract ontology the pre-processing and concept clustering have to be done.

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar (in some sense or another) to each other than to those in other groups (clusters).In this concept, centroid based clustering technique is implemented. In Centroid based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, *k*-means clustering gives a formal definition as an optimization problem: find the k-cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

Algorithmic steps for k-means clustering

Let $X = \{x_1, x_2, x_3, \ldots\ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots\ldots, v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, *'c_i'* represents the number of data points in *i^{th}* cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

Abstract ontology is constructed based the abstracts which are extracted by WebCrawler and placed in a separate folder in order to pre-process. The pre-process step, of this module is to separate some of the content related to the journal websites content. i.e. For example, Researchers download the abstract from ieee explorer website, then the copy writes of some content will be attached it and it is pre-

processed.

The contents of proposals are usually unstructured. The research ontology is used to analyze, extract, and identify the keywords in the full text of the proposals. Finally, a further reduction in the vocabulary size can be achieved through the removal of all words that appeared only a few times in all proposal documents

Based on the proposal documents which are extracted based on keywords (TF-IDF), the Domain ontology is matched with Abstracts (based on the terms which is already defined in the file with maximum count) and finally Abstract ontology is constructed by matching with domain ontology.

Example represents Ontology Matching of Domain and Abstracts. The matching is done based on the terms which already kept in a file. For example Let us take a Domain as Networks. The domain specified as Networks and terms are wireless, switch, topology etc , these contents are kept in a file. Based on these content Domain Ontology is constructed.

Next step, pre-process the abstracts which are obtained from the WebCrawler and thus perform clustering. After the clustering process, the POS (grammatical tagging) to be done and TF-IDF(maximum count or number of times word appears in a document) . Based upon these techniques, occurrence of terms is identified and its maximum count in that document. Hence matching is performed.

2.5 Abstract Matching

It is the process of determining correspondences between concepts. A set of correspondences is also called an alignment. It aims at finding correspondences between semantically related entities of different ontologies[11]. These correspondences may stand for equivalence as well as other relations, such as consequence, subsumption, or disjointness, between ontology entities.

In the same way objective, approach and result are matched, it will go and check the abstract (path of text document)[1]. The Researchers give the abstracts inorder to identify whether the similar concepts are published or for their literature survey.

Once the abstracts are given, this input is classified (documents are classified based on the terms which are already defined) and finally similar abstract with their domain is achieved.

The steps for computing semantic similarity between the abstracts:

- First, each abstract is partitioned into a list of tokens.
- Part-of-speech disambiguation (or tagging).
- Stemming words.
- Find the most appropriate sense for every word in a sentence (Word Sense Disambiguation).
- Finally, compute the similarity of the abstract based on the similarity of the pairs of words.

Tokenization: Each sentence is partitioned into a list of words, and removes the stop words. Stop words are frequently occurring, insignificant words that appear in a database record, article, or a web page, etc.

Tagging part of speech (+) : This task is to identify the correct part of speech (POS - like noun, verb, pronoun, adverb ...) of each word in the sentence. The algorithm takes a sentence as input and a specified tag set (a finite list of POS tags). The output is a single best POS tag for each word.

Stemming word (+) : The Porter stemming algorithm. Porter stemming is a process of removing the common morphological and inflexional endings of words. It can be thought of as a lexicon finite state transducer with the following steps: Surface form -> split word into possible morphemes -> getting intermediate form -> map stems to categories and affixes to meaning -> underlying form. I.e.: foxes -> fox + s -> fox.

Word sense disambiguation: Word sense disambiguation may be applied for identifying term meaning. This could help in identifying duplicate terms that represent the same concepts, or by distinguishing between multiple concepts that have the same lexical representation. Although such procedures are very useful, it should be noted that this only holds for noncompound terms, as compound terms usually have only one meaning. Next, we discuss four different WSD methods, of which one is an unsupervised method and the other three are supervised methods. The difference between supervised and unsupervised methods is that supervised methods use training data to train classifiers and subsequently disambiguate terms from a test set by using these classifiers, whereas unsupervised methods do not require this information.

In WordNet, each part of speech words (nouns/verbs...) are organized into taxonomies where each node is a set of synonyms (synset) represented in one sense. If a word has more than one sense, it will appear in multiple synsets at various locations in the taxonomy. WordNet defines relations between synsets and relations between word senses. A relation between synsets is a semantic relation, and a relation between word senses is a lexical relation. The difference is that lexical relations are relations between members of two different synsets, but semantic relations are relations between two whole synsets.

3 RESULTS

Experimental results on ontology matching gives better performanceThe matching of two ontology i.e., Abstracts gives

good output than the keyword matching. . This system has much advancement over the existing system and the following observations are confirmed.

- Automated ontology
- More than single ontology construction.
- WordNet
- K-Mean

4 CONCLUSION

A research ontology is constructed to categorize the concept terms in different discipline areas and to from relationship among them. It facilitates text-mining and optimization technique is used to research proposals based on their similarities and then to balance them according to the applicants characteristics. The proposed method can be use to expedite and improve the proposal grouping process. In the proposed ontology similarity measure, the similarities of the concept in the two ontologies are measured on the basis of its neighborhood set and the feature set. We define a function for the similarity measure that incorporates the neighborhood and the features of the selected query keyword. After performing the ontology matching, we can obtain the research proposal to the particular sub domains in the domain ontology. It also provides a formal procedure that enables similar proposals to be grouped together in a professional and ethical manner. The proposed method can also be used in other government research funding agencies that face information overload problems.

Future work is needed to cluster external reviewers based on their research areas and to assign grouped research proposals to reviewers systematically. Also, there is a need to empirically compare the results of manual classification to text-mining classification. Finally, the method can be expanded to help in finding a better match between proposals and reviewers.

REFERENCES

[1] Angela Locoro · Jérôme David · Jérôme Euzenat. Springer-Verlag Berlin Heidelberg 2013. “Context-Based Matching: Design of a Flexible Framework and Experiment”.

[2] K. Chen and N. Gorla, “Information system project selection using fuzzy logic,” IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 28, no. 6, pp. 849–855, Nov. 1998.

[3] K. Chen and N. Gorla, “Information system project selection using fuzzy logic,” IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 28, no. 6, pp. 849–855, Nov. 1998.

[4] W. D. Cook, B. Golany, M. Kress, M. Penn, and T. Raviv, “Optimal allocation of proposals to reviewers to facilitate effective ranking,” Manage. Sci., vol. 51, no. 4, pp. 655–661, Apr. 2005.

[5] D. E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning. Redwood City: Addison-Wesley, 1989.

[6.] Q. Liang, X. Wu, E. K. Park, T. M. Khoshgoftaar, and C. H. Chi, “Ontology-based business process customization for composite web services,” IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 41, no. 4, pp. 717–729, Jul. 2011.

[7] Y. Liu, C. Xu, Q. Zhang, and Y. Pan, “The smart architect: Scalable ontology-based modeling of ancient Chinese architectures,” IEEE Intell. Syst., vol. 23, no. 1, pp. 49–56, Jan./Feb. 2008.

[8] L. L. Machacha and P. Bhattacharya, “A fuzzy-logic-based approach to project selection,” IEEE Trans. Eng. Manag., vol. 47, no. 1, pp. 65–73, Feb. 2000..

[9] M. Nagy and M. Vargas-Vera, “Multiagent ontology mapping framework for the semantic web,” IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 41, no. 4, pp. 693–704, Jul. 2011.

[10] T. Ong, H. Chen, W. Sung, and B. Zhu, “Newsmap: A knowledge map for online news,” Decis. Support Syst., vol. 39, no. 4, pp. 583–597, Jun. 2005.

[11] Raghunadan P et al October 2012,“Fast Semi-Automatic generation of Ontologies and their exploitation”.

[12] Sharifullah Khan , Muhammad Safyan March 2014. “Semantic matching in hierarchical ontologies”.

[13] Sharifullah Khan , Jibran Mustafa. October 2013.” Effective semantic search using thematic similarity”.

[14] J. C. Trappey, C. V. Trappey, F. C. Hsu, and D. W. Hsiao, “A fuzzy ontological knowledge document clustering methodology,” IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 39, no. 3, pp. 806–814, Jun. 2009.

[15] YaJun Du*, YuFeng Hai, ChunZhi Xie, XiaoMing Wang, 2013 Elsevier. “An approach for selecting seed URLs of focused crawler based on user-interest ontology”.

INTERNATIONAL JOURNAL OF SCIENTIFIC & ENGINEERING RESEARCH, VOLUME 6, ISSUE 5, MAY-2015
ISSN 2229-5518