International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 6, June 2017)

Taxonomy Learning to Improve Overall Associative Strength among Concept-A Graph Based Approach

Dr. D. Bujji Babu[1], N. Sravanthi[2]
[1]QIS College of Engg & Technology, Ongole, AP, India.

***Abstract*— In this Paper we propose an automatic and unsupervised methodology to obtain taxonomy. Taxonomies are the key to developing successful applications in an domain such as information retrieval, knowledge searching and classification. The manual construction of the domain taxonomies is a time consuming task. To reduce the time we will build an new taxonomy learning approach named as TaxoFinder. TaxoFinder takes three steps to automatically build a Taxonomy. First it identifies the concepts from the domain corpus using concept extractor. second it builds a graph representing how such concepts are associated together based on their co-occurrences. As the key method in TaxoFinder we propose a method for measuring associative strength among concepts which quantify how strongly associated in the graph, using similarities between sentences and spatial distances between the sentences. Lastly taxofinder induces a taxonomy from the graph using graph analytic algorithm.**

***Keywords*— Taxo Finder, Associative Strength, Concept Graph, Similarity,Taxonomy Learning.**

I. INTRODUCTION

A taxonomy is the result of a classification task where categories are ordered in a hierarchical subclass structure. In recent years the extraction and the execution of a domain specific taxonomies has become increasingly relevant. This is due to two main facts. First it is a critical process in information science[2], since it is usually a part of information extraction. Second The manual construction of a taxonomy is a time consuming tasks.

Taxonomy or an ontology provides a shared conceptualization of a domain recently there is a lot of interest using structured data to empower search or other applications. A general purpose taxonomy about wordly facts is indispensable understanding user intent and many efforts are being devoted to composing and managing such taxonomies[9].

Taxonomies are the key to developing successful applications in an domain such as information retrieval, knowledge searching and classification. In particular considering the text ever growing amount of text digital data per year taxonomy learning from text is primarily research area for developing applications now a days.

The basic goal of corpus based approach is to collect large amount of textual data as input for semantic processing[5].starting off from a rather simple data model tailored for large amounts of data and efficient processing of a relational database system as a storage level. An important nature of taxonomy is that it enables us to representing highly related concepts together and the path between the two concepts that reflects how these are semantically in the domain. Manually building a taxonomy posses a great challenge that requires huge amount of time and effort of humans. Taxonomy learning uses methods to develop the fields of natural language processing, information retrieval, machine learning in an attempt to reduce the human effort and build a high quality of taxonomy.

This paper proposes a graph based approach unsupervised approach named as taxofinder for taxonomy learning that automatically builds a taxonomy from semantic graph named cgraph of concepts modeled from target corpus. First we extract the concepts from the domain corpus using concept extractor[6].second based on the co-occurences of the concepts in a sliding window which is a set of consecutive sentences in each document from the corpus. We will build an undirected graph cgraph. In the cgraph a node is a concept and an edge is created if two concepts are co-occur in a sliding window thus making an association between them. from the cgraph we measure an associative strength between two concepts by leveraging the sentence information that concepts appear in the corpus. Lastly we induce the taxonomy from the cgraph by applying maximum spanning tree(MST)[4] algorithm.

II. LITERATURE SURVEY

Domain terms are the building blocks of a taxonomy. while relevant terms for the domain could be selected. Manually in this work we aim at fully automatizing the taxonomy induction process. Thus we start from a text corpus for the domain of interest and extract domain terms from the corpus by means of terminology extraction algorithm. To this end we used our term extracting tool term extractor. Note that any equally valid term extracting tool can be applied in this step.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 6, June 2017)

As a result a domain terminology is produced which includes both single word and multi word expression we add to our graph g one node for each term.

In order to construct a domain taxonomy First a hierarchy based on the hierarchical clustering algorithm is constructed this algorithm of constructing hierarchal clusters is applied separately for each similarity measure presented. To compute the distance between two clusters the technique of average linkage clustering is employed. The other two techniques the single or the complete linkage method is that are not used.

The main draw back of the single linkage method is that clusters are not very similar. The complete linkage method is that outliers have a high influence on the clustering process. We choose the average linkage clustering as it has shown to provide a good balance between these two extremes.In this paper we present a graph based approach aimed at learning a lexical taxonomy automatically starting from a domain corpus and the web. Unlike many taxonomy learning approaches in the literature our novel algorithm returns both concepts and relations.

Similarity measure are used in similarity based retrieval to approximate the usefulness of cases with respect to the target problem. Similarity Based reasoning has been widely used in the different case based reasoning application. such as medical diagnosis, IT Service management, product recommendation and personal rostering decision to predict the similar cases having the appropriate solution for the target problem.

Document Collection browsing has been studied as an alternative to the ranked list representation for search results by the information retrieval community. The popular IR approaches include clustering and monthetic concept hierarchies. Clustering approaches are hierarchically cluster documents in a collection and label the clusters. Monthetic approach oraganise the concepts into hierarchies and link documents to related concepts. Both approaches are mainly based on pure statistics such as document frequency and confidential probability. The major drawback of these pure statistical approaches is their neglect of semantics among concepts.

III. METHODOLOGY

Given a domain corpus concept extraction is the first step for taxonomy learning[9].if extracted concepts are irrelevant a taxonomy may not correctly represent domain knowledge as such irrelevant concepts can also lead to generating irrelevant taxonomic relations. We build a cgraph where a node represents each of such concepts and an edge represents an association between nodes.

Each edge has a weight indicating the associative strength between two nodes.

A key challenge is to constructing the cgraph from concepts in c is the calculation of an associative strength between two concepts. This strength quantifies how semantically close these two concepts are. The associative strength among all extracted concepts will be used as the key for building a taxonomy from the cgraph.

$$w(c_1,c_2) = \frac{1}{k}\sum_{j=1}^{k} w_j(c_1,c_2)$$

where k is the number of documents in D. and $w(c_1,c_2)$ represent the associative strength between $c_1$and c_2 with respect to the document $D_j \in D$ thus $w(c_1,c_2)$ is calculated as the mean of associative strength between c_1 and c_2 across all documents in D. the value of $w(c_1,c_2)$ is normalized between 0 and 1,where 1 means that the associative strength between two concepts is highest and 0 indicates the strength is lowest.

$$w_j(c_1,c_2)=\frac{1}{m*n}\sum_{p,q} as(s_{jp},s_{jq})$$

Where

- $as(s_{jp},s_{jq})$ represents the function that calculates tha associative strength between two sentences s_{jp} and s_{jq} in $D_j \in D$ where $s_{jp} \in s_j(c_1)$ and $s_{jq} \in s_j(c_2)$
- p and q are the sentences sequential indices belonging to $I_j(c_1)$ and $I_j(c_2)$ in D_j
- m and n are the number of elements in Ij(c1)and $I_j(c_2)$ respectively i.e., $m=|I_j(c_1)|$,$n=|I_j(c_2)|$

It calculates the associative strength between c_1 and c_2 using the associative strength among their concepts more specifically the mean of associative strength between all pairs of two sentences $s_j(c_1)$ and $s_j(c_2)$ that contain c_1 and c_2 in D_j.

$$as(s_{jp},s_{jq})=sim(s_{jp},s_{jq})^{p-q}$$

Where the sentence similatiry $sim(s_{jp},s_{jq})$ between two sentences s_{jp} and s_{jq} is based on the approach proposed which showed a high performance. to illustrate how to calculate the associative strength between two concepts.

To illustrate how to calculate the associative strength between two concepts, let us consider an example.

Suppose that there is a corpus D which has a document D1 consisting of five sentences, D1 ={s1; s2; s3; s4; s5}. Suppose that ss1, ss2 and ss3 are three sentence sets, each having k-sequential sentences that appear in D. Assuming that a sliding window size is 3 (i.e. k =3), we set ss1={s1, s2,s3},ss1={s2, s3, s4}, and ss3={s3, s4, s5}.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 6, June 2017)

In Fig. 3a, each directed edge from one a to the other b means a belongs to b. For example, looking at ss1, s1 and c1, we see that s1 belongs to ss1, and c1 belongs to (i.e. appears in) s1.

Suppose that we also extracted three concepts c1, c2, and c3 from D1, i.e., C={c1; c2; c3}, where c1 appears in s1 and s3, c2 appears in s1 and s5, and c3 in s5. Thus, s11 is the sentence represented as s11={c1; c2}, since these two concepts appear in the sentence s1. Also, s13={c1}, s15={c2; c3}. In addition, s1(c1) denotes the set of sentences that contain the concept c1 appears in D1, i.e. s1(c1)={s1; s3}. Also,s1(c2)={s1; s5}, and s1(c3)= {s5}. Also, I1(c1) indicates the set of sequential indices of sentences s1(c1), thus I1(c1)={1; 3} Also, I1(c2)={1; 5}, and I1(c3)={5}.

$W(c_1, c_2)=w1(c_1, c_2)=$

$$\frac{as(s11;\ s11) + as(s11;\ s15) + as(s13;\ s11) + as(s13;\ s15)}{4}$$

$$\frac{1^0 + 0{:}5^4 + 0{:}6^2 + 0.7^2}{4}$$

0.48

Following the above calculation, we can also obtain W(c2,c3) as 0.28 and w(c1; c3) as 0.53.

Deriving CGraph

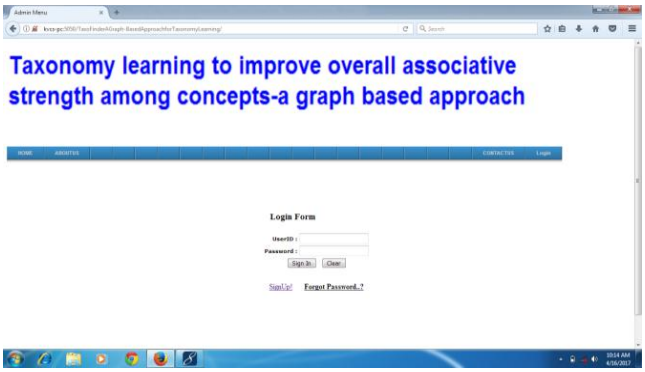
Once we build a CGraph, the third step is to derive a taxonomy from it. Our eventual goal is to build a taxonomy in such a way that it maximizes the overall associative strengths among all concepts in CGraph to find a good taxonomy. This is aligned with the notion of a good taxonomy used in the prior work [6]. The learned taxonomy guarantees that highly associated concepts are closely positioned.

The more concepts a CGraph has, the more complicated theCGraph tends to be, as the number of edges in the Cgraph could be substantially increased as a result. Note that the maximum number of edges in a CGraph From a CGraph, one possible way for deriving a taxonomy might be to reduce the overall number of edges in the graph by adjusting the size of the sliding window.

IV. Experimental Results

We developed the work using java, My Ecllipse 8.6, Oracle 10g Express Edition.

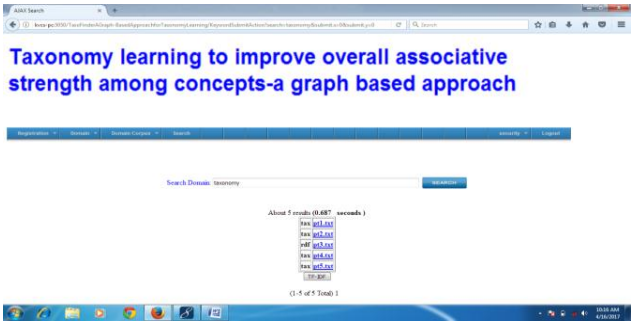
Home Page



Scrren 1:Home Page

Description: The above screen is homepage. It contains the home, registration, login pages. When user is new to this application user goes to the registration page and register to the details, if user is already register goes to login.

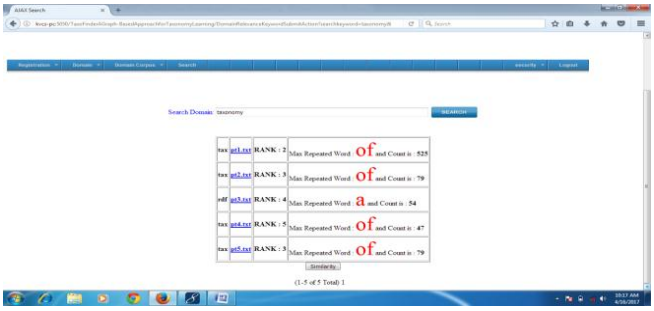
Search Domain:



Scrren 2: Search Domain

Description: In the above screen searching the domain name on this screen. This screen contains the search button.when we click on search button it will display all the related files based on the domain name.

Similarity:

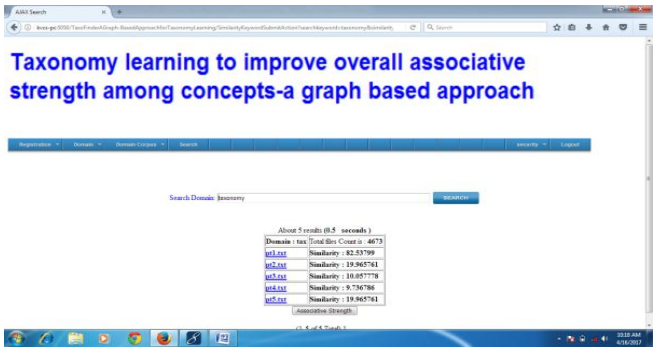


Scrren 3: Similarity

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 6, June 2017)

Description: In the above screen the domain relevance will be calculated and give the ranks based on the repeated text.and then we calculate the similarity of the given documents.

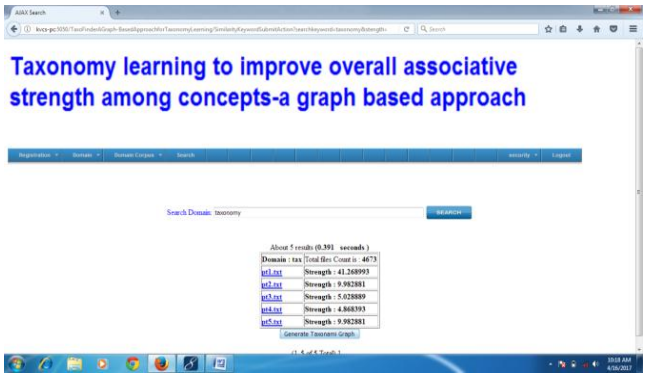
Associative Strength:



Scrren 4:Associative Strength

Description: In the above screen we calculate the similarity between the concepts and also display the repeated words in the chosen documents.then we calculate the associative strength.

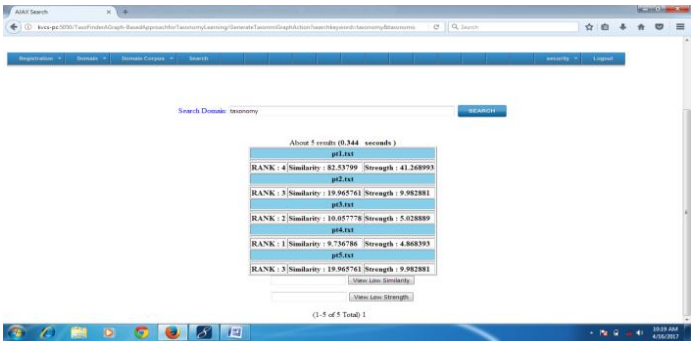
Generate Taxonomy Graph:



Screen 5:Generate Taxonomy Graph

Description: In the above screen we calculate the associative strength and it will displayed.and then we will generate the graph using similarities and strengths.

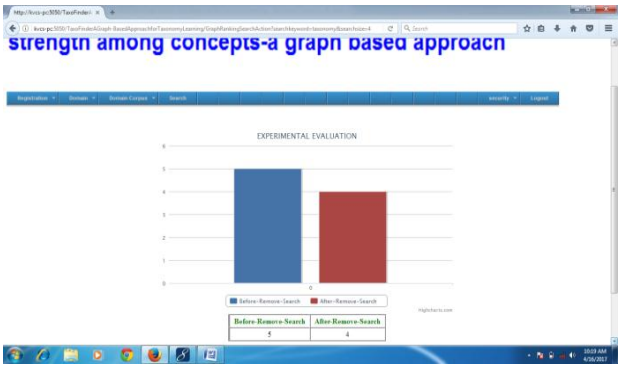
Building Cgraph:



Screen 6: Building CGraph

Description: In the above screen it will displays all similarities and associative strengths and also their ranks.And the two buttons also will be displayed In the first button will be named as View low similarity And the next button will be named as view low strength.we will give the low similarity either strength and then it will generates the graph based on the given value.

Taxonomy Graph:



Screen 7:Graph

Description: In the above screen it will displays the graph.The graph will be displayed based on the Removing search values.the graph will contains both before removing the search values and after removing the search Value.

V. CONCLUSION

In this paper we propose a taxolearn, a corpus based semantic taxonomy learning framework. For the implementation of taxolearn we aggregate and adapt steps from existing approaches. Taxofinder aims to build a cgraph representing concepts extracted from a domain corpus and their associative strengths. to measure such strengths we propose a formula for combining (1) the co-occurrence frequency of concepts with in a sliding window i.e., the set of consecutive sentences and (2) the distance and similarity sentences where such concepts are co- occur together. from the cgraph we used a graph analytic algorithm to induce a taxonomy aiming to maximize the overall associative strength among concepts to find a good taxonomy.

Acknowledgement

We acknowledge our sincere thanks and deep sense of gratitude to Mr.N.S.Kalyan Chakravarthy, Executive Chairman of QIS Educational Institutions, ongole, Andhrapradesh, for providing an excellent computational facilities, a very good learning environment in the campus with numerous journals and the digital library. We also thank Mr.N.Nageswara rao, President, SNES for his leadership and inspirational talks on importance of ethics and quality education.

REFERENCES

[1] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in Proc. 14th Conf. Comput. Linguistics, 1992, vol. 2, pp. 539–545.

[2] E.-A. Dietz, D. Vandic, and F. Frasincar, "TaxoLearn: A semantic approach to domain taxonomy learning," in Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intell. Agent Technol., 2012, pp. 58–65.

[3] Z. Kozareva and E. Hovy, "A semi-supervised method to learn and construct taxonomies using the web," in Proc. Conf. Empirical Methods Natural Language Process., 2010, pp. 1110–1118.

[4] P. Velardi, S. Faralli, and R. Navigli, "OntoLearn Reloaded: A graph-based algorithm for taxonomy induction," Comput. Linguistics, vol. 39, no. 3, pp. 665–707, 2013.

[5] G. Heyer, M. Luter, U. Quasthoff, T. Wittig, and C. Wolff, "Learning relations using collocations," in Proc. Workshop Ontol Learning, 2001, vol. 38.

[6] J. Seo, G.-M. Park, S.-H. Kim, and H.-G. Cho, "Characteristic analysis of social network constructed from literary fiction," in Proc. Int. Conf. Cyberworlds, Oct. 2013, pp. 147–150.

[7] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in Proc. 32nd Annu. Meeting Assoc. Comput. Linguistics, 1994, pp. 133–138.

[8] H. Yang, "Constructing task-specific taxonomies for document collection browsing," in Proc. Joint Conf. Empirical Methods Natural Language Process. Comput. Natural Language Learn., 2012, pp. 1278–1289.

[9] D. S_anchez and A. Moreno, "Web-scale taxonomy learning," in Proc. Workshop Extending Learn. Lexical Ontologies Using Machine Learn., 2005.

[10] Y.-B. Kang, S. Krishnaswamy, and A. Zaslavsky, "A retrieval strategy for case-based reasoning using similarity and association knowledge," IEEE Trans. Cybern., vol. 44, no. 4, pp. 473–487, Apr. 2014.

[11] Yong-Bin Kang, Pari Delir Haghigh, and Frada Burstein"Taxo Finder:A Graph Based Approach For Taxonomy Learning",in proc.