Information Retrieval for Domain Ontology Graph model using Clustering

Rupali.U.Patil

Abstract— Ontology deals with similar terms and relationship that can be used to describe and represent area of knowledge. In this paper, we propose to use knowledge represented in the form of ontology for categorizing documents. The novelty of this approach is we categorize text to a given ontology by using the Knowledge it contains. The information from the analyzed document is enriched with external ontological knowledge. The categorization process depends solely on entities, named relationships, and the taxonomy of classes represented in the ontology. We describe the processing phase of ontology graph generation system from input text documents of different domains. This paper focuses on processing of input text documents, processing is structured representation of the input text. Processing of ontology graph generation includes allowing tokenization of input text, POS tagging, removal of stop word and Stemming. Then class identification, matched the terms then generate tree, otherwise extract terms by matching input terms with dictionary.

Index Terms: Clustering method , Domain identification, Information management , Information Retieval , Knowledge representation ,Text analysis, , Text preprocessing.

1 Introduction

Ontology is a defined in a different ways. Ontology is defined as vocabulary, taxonomy or thesaurus. Ontology is a description of problem domain, where entities of the domain, its properties and its relationship are described.
Ontology is defined as “
“Ontology is a specification of conceptualization.”
Ontology is a set of definition of content –specification information represent primitives (constant, function, relation).ontology is very important aspect in different application to provide a semantic framework for knowledge management. Ontology represents the hierarchical structuring of knowledge about things by subcategorizing them according to their essential qualities. Ontology represents by UML [UML WEB], any OOL (object oriented language), and RDF [RDF W3C], DAML + OIL [DAML], OWL [OWL WEB] or any other representation. This defines object, properties and its relationship. Ontology engineering [1] is an association of ontology research for developing theories; method and software tools that help generate and preserve ontology. Text classification method is depends on knowledge representation in ontology. Text categorization is well research problem in computer science. The existing categorization method may be enhanced with the use of external knowledge that is not present or is virtually impossible to extract from the source document itself. One of commonly used source of external knowledge of text categorization is Word Net. Text categorization is also called as text classification, or topic identification. Each classification group is specified in term of ontology classes and may be defined as a single class, a subset of class taxonomy, a list of classes with a specific important, or perhaps, as a combination of taxonomy and a list of classes. Ontology learning [2] significantly facilitates the construction of ontology’s by the ontology engineer. Ontology learning aims the integration of multitude of disciplines in order to facilitate the construction of ontology’s, in particular ontology engineering. There are two text classification types i.e., term frequency-inverse document frequency (TF-IDF) [3], and term dependence [4]. Text classification is according to precision (to calculate accuracy of retrieval model) [5], recall [6] (from all data sets to calculate the capability of retrieval model to retrieve correct document), f-measure (to calculate vocal average of recall also precision) [7].

An ontology graph modeling process includes two phases: Pre-processing phase and Processing phase. The aim of this paper is to represent the pre-processing phase of domain ontology graph generation system for input text documents.

Representation of the pre-processing phase for domain ontology graph, to create system for input text documents

Overview of Clustering:

Clustering is a division of data into group of similar object. Each group, called cluster, consists of object that are similar amongst them and dissimilar compared to object of other groups.

Goal or objective of clustering:

To group data points that are close (or similar) to each other and identify such groupings (or clusters) in an unsupervised manner. Unsupervised means no information is provided to the algorithm on which data points belong to which clusters
A hierarchical clustering method works by grouping objects into a tree of clusters. Hierarchical clustering solutions, which are in the form of trees called dendrograms, are of great interest for a number of application domains. Hierarchical trees provide a view of the data at different levels of abstraction. Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a buttom-up (merging) or top-down (splitting) fashion. Agglomerative

algorithm build the three from bottom (i.e., its leaves) toward the top (i.e., root.)

2. MOTIVATION

Typically, classes in the ontology are structured into hierarchies. A class defines the type of properties common of to the individual objects within the class. The classes are interconnected by relationship defined some form of semantic interdependence (relationship are also regards as properties) class hierarchies and class relationship for schema level ontology. A comprehensive, well populated ontology with classes and relationship closely modeling a specific domain represent vast compendium knowledge in the domain. We belive that the knowledge represented in a compressive ontology can be leveraged to identify in a text document, provide the document is thematically related to the ontology domain. Furthermore the concept in the especially if they are organized into hierarchies of higher level categories, may be used to identify the category that best classifies the document based on its content.

3. LITERATURE REVIEW

The document clustering is a specialized data clustering problem. There are two main factors are as follow:

1) To characterized inherent semantic of document.

2) To define similarity measure based on semantic representation. It contains higher mathematical value to document pair which has higher semantic relation.

Different approaches have been proposed by many researchers to take care of semantic relation into a document clustering. In a latest approach to get better clustering result is recommended by applying background knowledge during preprocessing. In preprocessing ontology based heuristic for feature section featured aggregation is useful to make a no. of alternative text representation. It referred as COSA (concept selection and aggregation). COSA contain two stages as:

1) It maps terms onto concepts using shallow and reuseful NLP system.

2) It uses the concept hierarchy to propose good aggregation for subsequent clustering.

Result as comparable with sophisticated baseline preprocessing strategy on tourism domain dataset.

P. K. Bhowmick, D. Roy, S. Sarkar, and A. Basu [1], "A framework for manual ontology engineering for management of learning material repository, 2010 provide a structure of that enables the knowledge experts to build domain knowledge in their vernaculars. The residential ontology has been utilized to index learning materials into dissimilar ontological levels then the created index structure has been used for providing more accurate search in education domain. The framework also provides the ways to perform personalized search by consulting the user profiles and the created index structures.

E. H. Y. Lim, R. S. T. Lee, and J. N. K. Liu [4], "Knowledge Seeker—an ontological agent-based system for retrieving and analyzing Chinese Web articles," 2008, present the Knowledge Seeker, is an ontological agent based system. Knowledge is intended to help user find, retrieve, and analyze news article from internet. Then the present content in semantic web. Using the Chinese document corpus to evaluate knowledge seeker. For result testing to compared with other approaches. Knoeledgeseeker is able to identify the topic of Chinese web article with accuracy of almost 87% also the processing speed less than one second per article.

Knowledge is also able to organized content flexibity and understands knowledge more accuracy than method that use ontology definition.

George Forman [7], "An extensive empirical study of feature selection metrics for text classification" 2003 presented in general comparative study of feature selection metrics for the high dimensional domain text classification. It and 2 class problem classically with high class skew also it exposed the surprising performance of a new feature selection metric, Bi-Normal Separation. A further involvement is a novel evaluation methodology that considers the common problem of trying to select one or two metrics that have the best chances of obtaining the best performance for a given dataset. Somewhat surprisingly, selecting the two best performing metrics can be sub-optimal: when the best metric fails, the other may have correlated failures, as is the case for IG (Information Gain) and Chi for maximizing precision. The residual analysis determined that BNS (Bi-Normal Separation) paired with Odds Ratio yielded the best chances of attaining the best precision. For optimizing recall, BNS paired with F1 was consistently the best pair by a wide margin.

Dino Isa, Lam Hong Lee, V.P. Kallimani, and R. Raj Kumar [3] "Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine" , 2008, described the hybrid text document classification approach, using the naïve bayes method at front end for raw text data vactorization. In combination with a SVM classifier at the back end to classify the documents to the right category. Results show that to improved classification accuracy compared to the pure naive Bayes classification approach. Highest probability category to identify the correct class of document to use naïve bayes classifier. The SVM to the nai¨ve Bayes approach adds only one second to the train and test times.

4. PROBLEM STATEMENT

Document clustering is a specialized data clustering problem, where the object is in the type of document. Data mining is a technique has been profitably demoralized for this purpose. The set of documents, that to partition them into a predetermined or an automatically derived no of cluster such that document assigned to each cluster are more similar to each other that the document assign to different cluster.

5. DOCUMENT PREPROCESSING

In Natural Language Processing, text Preprocessing is the main part of the system. Given a document data set, initially to perform a preprocessing phase for to reduce the specific of terms that are used to compare with document D, by removing irrelevant data. This process is called information extraction or feature extraction.

1) Tokenization:

In tokenization process sentences split into individual tokens, typically words. Further, redefined method, drawn from the field of natural language processing, parse the grammatical structure of the text to choose important terms or chunks(Sequence of words),such as noun phrases.

2) POS tagging:

This process involves assigning a part-of-speech label such as adjective, noun, conjunction to the words in the corpus. This assignment of labels is known as part of speech tagging.

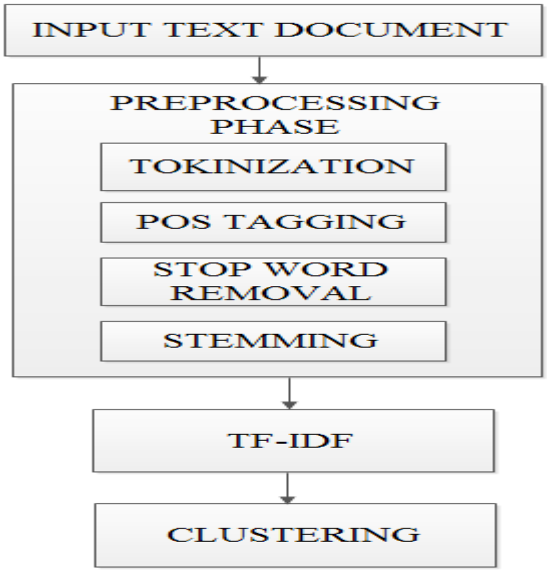
3) Stop word removal:

In any document, discarding these words (like: is, the, for, an) which greatly reduces the overhead of processing.

4) Stemming:

In this process all the misspelled words are removed. It involves stemming the term to its root word, such as a word 'specifically' is stemmed down to specific'. The porter's stemming algorithm is used for this process.

A System Architecture



B) TF- IDF weighting method:

TF-IDF is a weighting approach used for to evaluate the importance of a term in the corpus or identifies how relevant a term is to the corpus. It is mathematically expressed as:

$$\text{Tf-idf}_{(t,\,d)} = \text{tf}_{(t,\,d)} \times \text{idf}_{t}$$

From above expression, tf_i is the term frequency of term in a document, D is the number of documents in the corpus, and df_i is the document frequency or the number of documents containing term i. In other words, tf-idf$_{t,\,d}$ assigns to term *t* a weight in document *d* that is as:

- Highest when *t* occurs many times within a small number of documents (thus lending high discriminating power to those documents).
- Lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal).
- Lowest when the term occurs in virtually all documents.

C) Hierarchical Clustering Algorithm:

In clustering document, hierarchical clustering is one of the methods. It produces the hierarchical to the document. In this method, grouped of data object into tree of cluster. Then output is a structure that is the more informative than the unstructured set of cluster than the partition cluster. It is a used of hierarchical decomposition of a given set of data object. It is a better then partitioned clustering; its major job is to construct the hierarchical structure in a tree of clusters whose leaf node represent the subset of document collection. It can be divided into two methods:

A) Agglomerative hierarchical clustering.
B) Divisive hierarchical clustering.

A) Agglomerative hierarchical clustering:

This method is a bottom-up strategy; its start with each objet a separate clusters itself and merges the clusters according to distance measure. It merges until all object into a single cluster till the termination conditions are satisfied. It merges the cluster iteratively and it represent by genograms as a tree like structure that show the relation between object. In dendogram each merge is represented by horizontal line. The y co-coordinator horizontal lines are similarity of two clusters that where merged and document can be viewed in single clusters. Its similarity measure is calculated by:
1) Single-linkage clustering.
2) Complete-linkage clustering.
3) Group-average clustering.

1) Single link clustering:

Similarity between two clusters depends on most similar members of the clusters i.e. similarity between two clusters is measured by similarity of cloest pair of data point belonging o different clusters .The process of merging cluster repeat until that all object merged to form a single cluster. The minimum distance is calculated between document is:

$$Min \{d(x, y): x \square A, y \square B\}$$

2) Complete link clustering:

Similarity between the clusters is the similarity of their most dissimilar. A maximum distance calculated between the documents is:

$$Max \{d(x, y): x \square A, y \square B\}$$

3) average-linkage clustering:

Average calculates the cluster quality based on all similarity the document. Mean distance is calculated between documents is:

$$\frac{1}{[illegible]} d\{x, y\}$$

- *Algorithm steps for hierarchical clustering:*

Here the N is the set of items to be clustered and distance is N*N matrix (Similarity) .The process of hierarchical clustering:

Step 1: Assigning each item to its own clusters, so that if you have N items, you know have N clusters. Each clustering contains just one item. Distance between the clusters the same as the distance between the items they contain.

Step 2: calculate the cloest pair of the cluster then merges the pair into single clusters, so that now you have one cluster less.

Step 3: Calculate the distance between new cluster and each old clusters. (Single-linkage, complete-linkage and average- linkage)

Step 4: Repeat step 2 and 3 until all item are calculated in a single cluster of size N.

Advantages:

1) This algorithm can construct an ordering of objects, which may be informative for data display. Smaller clusters are generated, which may helpful in discovery.
2) It can be used any valid measure of distance.
3) The observations themselves are not required; all that is used is a matrix of distances.

Disadvantages:
1) They do not scale well; time complexity of at least O (n_2) where n is the total number of objects.
2) They can never undo what was done previously.

D) Extract Terms by matching input terms with dictionary Terms

Dictionary is used to identify the meaningful terms in input text after removing the stop words.

E. Mathematical Model:

Let the S be a system such that: S = {I, P, T, C,E}

Where,

S=proposed system.

I= Initial state at T <init> i.e., providing the input data base.

E= End state of concept cluster.

P= preprocessing phase

Where,

P= {P1, P2, p3, p4}

P1= Tokenization.

P2= POS tagging.

P3= Stop word removal.

P4= Stemming.

T= Text analysis and weighting (TF-IDF).

C= Clustering.

6. Result and Discussion

In text documents contain four different domains, which are Indian festival, Education, Politics, and Sports. These four domains are labeled as classes for the domain ontology learning process. These documents are stored in text files. First Select the location of the data sets then start process to perform a preprocessing phase.

Diwali is hindu festival			
Import Data	**Split**	**tokinize**	**Pos**

After pre-processing phase to calculate the weight with the help of TF-IDF. In TF-IDf and DOG-Onto approach to calculate the value of precision, recall and f-measure

Approach	**TP**	**FP**	**PRE**	**REC**	**F-M**
TI-IDF	**0**	**100**	**0**	**0**	**NAN**
DOG	**25**	**75**	**0.25**	**0.25**	**0.25**

7. Conclusion

In this paper we present the approach of concept clustering. In our approach, we are predicting the concept clustering by using hierarchical clustering. Experimental evaluation result shows that our proposed system improves result than existing system. In this system, hierarchical clustering algorithm is used which is readily helps to improve results. This algorithm helps to greater extend. In future, we can enhance our system with advanced decision making system. So, it can support to the target action depends on feature of discovered clustering

REFERENCES

[1] P. K. Bhowmick, D. Roy, S. Sarkar, and A. Basu, "A framework for manual ontology engineering for management of learning material repository," *Int. J. Comput. Sci. Appl.*, vol. 7, no. 2, pp. 30–51, 2010

[2] H. Alani, K. Sanghee, D. E. Millard, M. J. Weal, W. Hall, P. H. Lewis, and N. R. Shadbolt, "Automatic ontology-based knowledge extraction from Web documents," *IEEE Intell. Syst.*, vol. 18, no. 1, pp. 14–21, Jan./Feb. 2003

[3] D. Isa, L. H. Lee, V. P. Kallimani, and R. RajKumar, "Text document preprocessing with the Bayes formula for classification using the support vector machine," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 9, pp. 1264-1272, Sep. 2008

[4] E. H. Y. Lim, R. S. T. Lee, and J. N. K. Liu, "Knowledge Seeker—an ontological agent-based system for retrieving and analyzing Chinese Web articles," in *Proc. IEEE Int. Conf. Fuzz. Syst.*, Jun. 1–6, 2008, pp. 1034–1041.

[5] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006

[6] O. Etzioni, M. Cafarella, D. Downey, A. M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, "Unsupervised named-entity extraction from the Web: An experimental study," *Artif. Intell.* vol. 165, no. 1, pp. 91–134, Jun. 2005.

[7] G. Forman, "An Extensive empirical study of feature selection metrics for text classification," *Int. J. Mach. Learn. Res.*, vol. 3, no. 7–8, pp. 1289–1305, Mar. 2003