IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 3, March 2015.
www.ijiset.com
ISSN 2348 – 7968

Ontology Based Effective Text Mining Technique Using Singular Value Decomposition

[1]K.Anjugam, [2]D.Dinagaran.M.Tech.,
[1]Student, [2] Assistant Professor,
[1,2]Department of Computer Science and Engineering, IFET College of Engineering.

***Abstract*— Text mining is an effective method of obtaining potentially desired knowledge from text document. There is no accuracy in obtaining the features from the traditional techniques of text mining because the semantic information of the text is properly utilized. The semantic information needs technical support and theoretical basis, which can be provided by ontology. This paper introduces and analyzes text mining based on domain ontology using singular value decomposition technique in order to solve the high dimensional sparsity problem existing in D matrix computation. Finally this proposed method is implemented as a prototype tool and a benchmark database is used to validate it. The benchmark database consist of heart disease dataset that consist of diagnostic features. The system will also predict the condition of the patient by taking the input of diagnosis values.**

Keywords*— Domain ontology; D matrix; Singular value Decomposition.

INTRODUCTION

With the rapid development of Internet and information technology, network is no longer a simple platform for communication, and it becomes to information carrier of a set of digital and information resources. People get the information they need by querying search engines. However, due to the independence of the information in every field and the limitation of search engines, semantic information retrieval is unable to realize. In order to achieve information sharing and interaction better, we need professional ontology to guide the search.

Ontology is a word derived from the earliest philosophy. In recent years, there are more and more researches on ontology in computer science. Now ontology has become a very hot concept in computer field. It attracts many experts and scholars to study the ontology technology and plays a prominent role in many fields.

In recent years, experts from various countries are studying how to obtain a simple and efficient method of building ontology. Through introducing the technology of natural language processing and classification and data set mining to ontology construction, they have developed some methods and techniques by using some data set corpus of some fields to construct domain ontology, such as SOAT [1] developed by the Academia Sinica in Taiwan; DODDLE [2] developed by the research institution in Japan; SymOntos [3] developed by the research institution in Italy. All of these methods are based on the processing of data set corpus, including natural language processing, data set classification and domain keyword extraction. Then they will make use of rules or statistical methods to gain part of relationships between concepts. To complete or extend the core ontology, it will need artificial check. This semi-automatic ontology construction method is not very mature, so there are a lot of researches on space.

There are a lot of meanings to research ontology. Firstly, in terms of ontology's own form, it is a good show for all areas, all things and all concepts in the world [4]. From its structure, we can see that it includes the concept of things, the relationship between things, the nature of things, and more the mutual restriction between things, interrelated and interdependent. Ontology can reflect the nature of the objective things and its external performance. Secondly, in the field of artificial intelligence, ontology can help us to get the essential knowledge of things [4]. The main research purpose of artificial intelligence is to hope that computers can "think" like humans, and human wisdom embodies in the understanding of things and knowledge. Knowledge is a system structure that has hierarchy, logical relationship and reasoning rules, and all of these are the feature of ontology. Thirdly, ontology can solve the problem that results of general search engine are not accurate [4].

The main content of this article is the study on automatic ontology extension method, and raise an automatic ontology extension method based on data set clustering and supervised learning.

II. RELATED WORK

For the method of domain ontology construction, international scholars have already studied for many years. It has formed some relatively mature methods of ontology construction, such as Skeleton method, seven steps method and so on.

ISSN 2348 – 7968

Skeleton method [5], developed by the University of Edinburgh, was based on the experience of building enterprise ontology. The main methodology consists of the following steps: at first, be clear the purpose of the construction and application field. The process of building ontology is departed to ontology obtaining, ontology coding and ontology integration. Then establish the evaluation standard. At last, write the concepts and ontology into a document.

Seven Steps method [6], developed by the college of medicine, Stanford University, was mainly used in the domain ontology construction. The main steps are as follows: (1) Determine ontology scope; (2) Consider to reuse the existing ontology; (3) List the terms of the field; (4) Define the classification, including the hierarchical relationship between classed and subclasses; (5) Define properties, including the domain and range of properties; (6) Define the facet of properties, which means some special values or features; (7) Instance the classes to complete the ontology.

Seven Steps method is the most widely used and the most mature method to construct ontology. On the basis of Seven Steps method, Sun et al [4] proposed a semi-automatic ontology construction method, which included a new algorithm, ontology learning based on the data set statistics. Use the statistical information of data set as the data source of concept obtaining, apply data set classification and keyword extraction technology to deal with data set statistic table, extract the domain concepts and generate ontology concept candidate set. Extend ontology by reusing WordNet through calling Jena API. Do some specialized processing to complete the ontology.

Xu et al [7] proposed a kind of ontology construction and extension method based on relational database focusing on the relational database features. Through a detailed analysis of relational database schema, define a series of rules that extracts data in a relational database to ontology, and follow these rules to build a preliminary ontology. Then, use WordNet to expand the initial ontology. Finally, use ontology editing tools to modify and improve the ontology manually.

Wang et al [8] proposed a data set information retrieving method that employs domain ontology to extend users' query requirements, which include an extracting method for domain knowledge aimed at distributed and then domain ontology can be built by using these domain knowledge. Also, a query expansion method based on domain ontology was raised, which includes the analysis of user queries, the semantic vector representation of queries and the match between users' query and domain ontology.

III. AUTOMATIC ONTOLOGY EXTENSION METHOD BASED ON DATA SET CLUSTERING AND SUPERVISED LEARNING

A. Feasibility of the Method

Building domain ontology manually is a very complex and tedious process, and also it is unrealistic to complete a professional and highly specialized ontology automatically. In the process of ontology construction and extension, we often require the participation of domain experts. Semi-automatic ontology extension method becomes a research hotspot, and it is the most rational ontology extension method. The selection and extension of concepts will be done automatically by the program according some algorithm. In the process of the program, we will revise and improve the ontology through the guide of field experts. The algorithm will become more and more intelligent, and ultimately it will generate professional and complete domain ontology.

This paper studies the existing methods of semi-automatic ontology extension. On this basis, we proposed an automatic ontology extension method based on supervised learning and data set clustering. By doing some experiments, we verify the various steps in the process and prove the feasibility of this method.

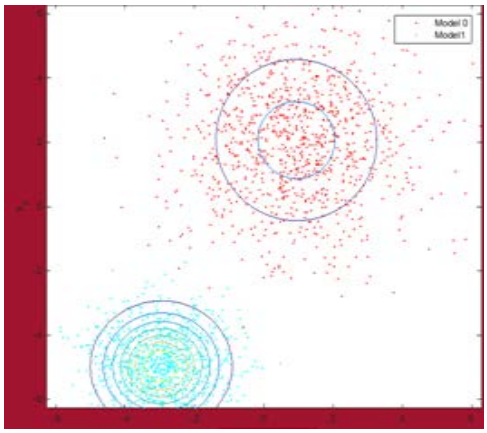

Figure 1:Default clustering

B. Basic Idea

The proposed method of automatic ontology extension in the paper employs data set clustering algorithm and K-means classification algorithm. The basic idea is as follows.

1) Extract distributed from the SVD database, take distributed as data sets and do the data set clustering algorithm;

2) Obtain classified clusters and select the top five words of frequency of occurrence among the cluster;

3) Select theme vocabulary for each cluster after manual intervention;

4) Update the ontology with the theme vocabulary, add corresponding training set to K-means classifier and update domain-independent vocabulary to SVD dictionary;

5) Extract domain candidate set of words after the distributed doing splitting and denoising by Chinese word splitting component;

ISSN 2348 – 7968

6) Classify the candidate set of words based on user-supplied training set by using K-means classifier, and set the threshold. If the score of word is lower than the threshold, we think this word is not related to the field, so the word will be added to SVD dictionary. On the contrary, if the score of word is higher than the threshold, we choose the classification which has the highest score and update the word to ontology;

7) Reload SVD dictionary, extract the content of distributed and do splitting and denoising;

8) Do data set clustering again, and the valuable data will be added to the training set of the category.

This algorithm is designed to have a feedback mechanism, in which "positive feedback" is to get the classifications of distributed, filter some effective keywords, add them to the training set and update the SVD dictionary with the words which are not related to the field. "Negative feedback" is a process by using K-means classifier to do supervised learning from candidate vocabulary set according to training set and classify them one by one. In addition, we set a threshold. If the score of word is lower than the threshold, the word will be added to SVD dictionary.

Therefore, the proposed method of automatic ontology extension in the paper based on data set clustering algorithm and K-means classification algorithm will become more and more intelligent. In the subsequent data processing, it will be more accurate.

C. Algorithm Description

The flow chart of the proposed method of automatic ontology extension based on data set clustering and supervised learning is as shown in Figure 1:

1) Description of Data set Clustering Algorithm

The data set clustering algorithms used in this paper mainly include K-means algorithm [9], getting the feature vector of each distributed, calculating TF-IDF weight, taking the highest frequency term from distributed, and doing recursive clustering. The TF-IDF (term frequency-inverted document frequency) is used to calculate the weights of terms, and its formula is defined by Salton [10] as follows:

$$tfidf_{t,d} = tf_{t,d} \cdot \log(N / df_t) \qquad (1)$$

Where $tf_{t,d}$ is the absolute frequency of term t in the document d, df_t is the document frequency of term t that counts how many documents in which term t appears, and N is the total number of documents.

The detailed algorithm description is as shown in Algorithm 1.

WP: Distributed;
P: Pages;
k: the number of clusters;
t: terms;
f: frequency of term;
S: DataSet, the list of page weight;
C: Centriole list;
C_{dis}: Centriole distance;
C_{class}: the content class of the cluster;
Extract *P* from *WP*s;
$P(t_1,f_1; t_2,f_2; ...; t_n,f_n)$ =
IKAnalyzer(*P*); foreach (*P* in *WP*s)
foreach (*t* in *P*)

$$_{t} p = (tdidf(t, p), ..., tdidf(t_n, p))$$

$$S = \{p_1, p_2, ..., p_n\}$$

K_means(DataSet *S*, int *k*)
new_C = Select_init_centriole(*S*, *k*);
do {
    C = *new_C*;
    foreach (*s* in *S*)
        foreach (*c* in *C*)
        distance = Calculate_distance(*s*, *c*);
        if(C_{dis}_*best* > distance)
            C_{dis}_*best* = distance;
            C_{class}_*best* = c.tag;
        category_list[Cclass_best].Add(s);
    new_C = Relocate_centriole(*category_list*, *C*, *k*);
} while (! Is_centrio_stable (*new_C*, *C*));
return *category_list*;

Algorithm 1: Data set Clustering Algorithm

2) Description of Supervised Learning Algorithm

The supervised learning algorithm used in this paper is the learning process of K-means classifier [11] using the given training set. We also use Chinese word splitting component, KAnalyzer, through positive and negative feedback mechanisms to expand vocabulary and SVD dictionary, to guide future learning process, which is as shown in Algorithm 2.

P: Pages;
DCS: Domain Candidate Set;
SD: SVD Dictionary;
O: Ontology;
t: terms;
TS: Training Set;
C: Classes;
DCS{t_1,t_2,...,t_n} = IKAnalyzer(P);
foreach (*t* in *DCS*)
  foreach (*c* in *C*)
    $Pro_{(t,class)}$ = Calculate_prior_probability(*t*,*class*);
  If ($Pro_{(t,class)}$ < threshold)
    Add (*SD*, *t*);
  ELSE
    c = max ($Pro_{(t,class)}$);
    Add (O, c, t);
return O;

Algorithm 2: Supervised Learning Algorithm

ISSN 2348 – 7968

IV. EXPERIMENT AND EVALUATION

In this section, we apply this method in the field of SVD. The design goal of SVD ontology is to capture the knowledge of the SVD industry, provide the understanding of SVD knowledge, and define the vocabulary which has common recognition in SVD field. Data Preparation

In order to evaluate the effectiveness of the proposed method, we download 139,568 distributed by using the SVD crawler named Cardinal . All the distributed are related to the SVD domain. The set of distributed are downloaded from some official SVD sites of port.

A. Ontology Construction

After learning the experimental data, this section is mainly to do experiments to validate the algorithm, to construct a SVD ontology automatically:

1) Prepare the original data: port.rdf, dump.txt, SVD.dic, Training Set.

2) K-means classifier extract data set clustering original set, extract the content of distributed, and do splitting and denoising. After that, written in the format of the clustering original set, one line represents one SVD content.

3) Do data set clustering and get hierarchical file library. Extract the most frequent five words occur in each clustered file and use them as the file name. Traverse these SVD contents again and make interactive clustering.

4) Update SVD dictionary, training set and ontology. The hierarchical file library obtained above tells us how to classify the field of SVD knowledge. We also know which words are not related to the field, so we need to update SVD dictionary, add new training set and add these new classification to the ontology manually.

5) Call k-means classifier and update SVD dictionary automatically. The classification result we obtained is displayed by ontology visualization tool.

6) After updating the ontology and SVD dictionary, we can do data set clustering again. At this time, the result has less “noise”.

B. Experimental Results and Analysis

After using the automatic ontology extension method based on data set clustering and supervised learning, it is obvious that in the process of ontology construction, there is less manual intervention, and accuracy has also been significantly improved.

V. CONCLUSION AND FUTURE WORK

This paper presents an automatic ontology extension method based on data set clustering and supervised learning. We apply the proposed method to a large set of distributed in the data set, and successfully construct an ontology for the port related concepts. In our future work, we will request domain experts to help us to identify relationships between concepts in the ontology, and that make ontology more reasonable.

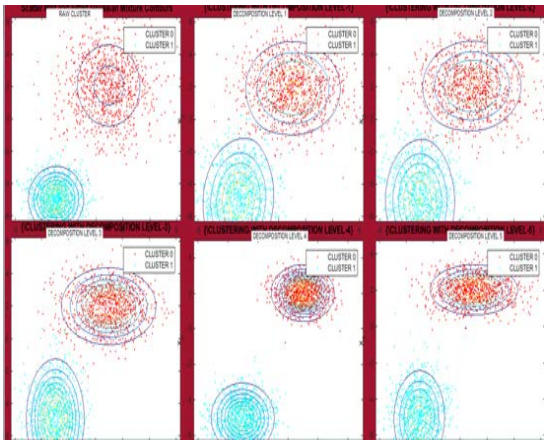

Figure 2:Clustering after decomposition using SVD

REFERENCES

[1] Shih-Huang Wu, Wen-Lina Hsu. SOAT : A Semi-Automatic Domain Ontology Acuqisition Tool form Chinese Coprus[J]. COLING 2002.

[2] N.Nakaya, M.Kurematsu and T.Yamagauchi. A Domain Ontology Development Enviroment Using a MRD and Data set Coprus[C]. Fifth Joint Conference on Knowledge-based Software Engineering. Frontiers in Artifical Intelligence and Application, Vo180, 105Press, 2002:242-251.

[3] M.Missikoff, P.Velardi, P.Fbarinai. Data set Mining Techniques to Automatically Enrich a Domain Ontology[J]. Applied Artifical Intelligence Journal. 2003:323-340.

[4] Y. Sun. Semi-automatic ontology construction methods[D]. Master's thesis. 2009.5.

[5] Uschold M. Knowledge Level Modelling:Concepts and Terminology[J]. The Knowledge Engineering Review, 1998, 13(1):5-29.

[6] S. Li, Q. Yin, Y. Hu. Review on Ontology Research[J]. Computer Research and Development. 2004, 41(7).

[7] G. Xu. Research on Ontology Construction Method based on Relational Database[D]. 2010.4

[8] J. Wang. Study on Data set Retrieval Based on Domain Ontology Extensions Query[D]. 2011.3

[9] Lizhang Zhan, Hong Xu, Xiuguo Chen. A Semi-Supervised Data set Clustering Approach Based on K-means Algorithm[C]. International Conference on Engineering and Business Management. 2011.

[10] G. Salton, M.J.MxGill, Introduction to Modern Information Retrieval, McGraw-Hill Inc., New York, NY, USA, 1986.

[11] C. Zong. Statistical natural language processing. Tsinghua University Press. 2008:340-349