Research on Text Mining Based on Domain Ontology

Jiang Li-hua[1,2] , Xie Neng-fu[1,2] , and Zhang Hong-bin[3,*]

[1] Agricultural Information Institute of Chinese Academy of Agricultural Sciences, Beijing, 100081
[2] Key Lab of Agricultural Information Service Technology of Ministry of Agriculture, Beijing, 100081
[3] Institute of Agriculture Resources and Regional Planning of Chinese Academy of Agricultural Sciences, Beijing, 100081

Abstract. This paper improves the traditional text mining technology which cannot understand the text semantics. The author discusses the text mining methods based on ontology and puts forward text mining model based on domain ontology. Ontology structure is built firstly and the "concept-concept" similarity matrix is introduced, then a conception vector space model based on domain ontology is used to take the place of traditional vector space model to represent the documents in order to realize text mining. Finally, the author does a case and draws some conclusions.

Keywords: Ontology, text mining, domain ontology, vector space model.

1 Introduction

Natural language is the main communication and expression thought tool in today's economic society. Although it has been studied for a long time, the understanding and using ability is still limited. The data mining technology based on statistics had matured and applied successfully in large scale relational database in the early nineteenth century. Naturally scholars had the idea of applying the technology of data mining to analyze the text block described by natural language and called it text mining or knowledge discovery in text. Different from the traditional natural language processing's focusing on understanding the words and sentences, the main goal of text mining is to find out the unknown and valuable knowledge or their relationship in large scale text sets. However, I found that most text mining lack of semantic considerations in application, only analyze grammatically, but not the content, so the results are always barely satisfied.

* Corresponding author.

D. Li and Y. Chen (Eds.): CCTA 2013, Part II, IFIP AICT 420, pp. 361–369, 2014.

2 Text Mining Based on Ontology

Text mining, or knowledge discovery in text database, is the process of finding unknown, useful and understandable knowledge in large scale text database. The objects of text mining are semi-structured or unstructured. And they always contains multi-layer ambiguity, so a lot of difficulties of text mining are caused.

The traditional text mining method based on vector space model converts the text to word frequency vectors. The major defect of this method is neglecting the importance of semantic role leading to text mining results are unsatisfied. Therefore, semantic analysis and processing technology should be combine with text mining technology in order to develop more effective mining method to realize deep semantic level mining. Appling ontology to text mining provides theoretical support and a feasible approach to solve above problems.

At present, the representative semantic dictionaries applied common ontology are English WordNet and Chinese HowNet. There are many text mining methods based on WordNet and HowNet. But the text mining methods based on common ontology can not get a very good effect in partial field. Therefore a lot of research in recent years began to carry out the research of text Mining based on domain ontology. Bloehdom etc. put forward OnTology Based Text mining frame wOrk; The bag of words representation used for these clustering methods is often unsatisfactory as it ignores relationships between important terms that do not cooccur literally. In order to deal with the problem, Hotho etc. integrate core ontologies as background knowledge into the process of clustering text documents. Song etc. suggests an automated method for document classification using an ontology, which represses terminology information and vocabulary contained in Web documents by way of a hierarchical structure.

In our country, Knowledge Engineering Research Laboratory of computer science department in Tsinghua University has developed ontology data mining test platform based on semantic Web. And also there are some researchers who discussed applications of semantic processing technology in text mining. Xuling Zheng etc.proposed a corpus based method to automatically acquire semantic collocation rules from a Chinese phrase corpus, which was annotated with semantic knowledge according to HowNet（Zheng Xuling etc.， 2007）. By establishing domain ontology as the way of knowledge organization, Guobing Zhou etc. introduced a novel information search model based on domain ontology in semantic context（Zou Guobing etc., 2009）. An Intelligent search method based on domain ontology for the global web information was proposed to solve the problem of low efficiency typical in traditional search engineers based on word matched technology by Hengmin Zhu（Zhu Hengmin etc., 2010）. In order to improve the depth and accuracy of text mining, a semantic text mining model based on domain ontology was proposed by Yufeng zhang etc.. And in this model, semantic role labeling was applied to semantic analysis so that the semantic relations can be extracted accurately (Zhang Yufeng etc., 2011).

Taken together, the research of semantic text mining based on domain ontology is still in research theory spread stage, but relative actively in foreign countries. But there is few whole text mining based on domain ontology solutions. And the research scope is only in foundation of shallow knowledge such as classification and clustering of text

(Bingham,2001; Montes-y-Gómez, 2001)but rarely in rich useful deep semantic knowledge such as semantic association foundation(Zelikovitz,2004)、topic tracking(Aurora, 2007) and trend analysis (Pui Cheong Fung, 2003)and so on.

3 Research on Key Technology of Text Mining Based on Domain Ontology

As an expert to guide the entire mining process, ontology is used to pre-process the text structure to realize semantic mining and improve mining effect.

3.1 Knowledge Representation Based on Domain Ontology

At present, most of the ontology system basic structure is similiar which are entity, conception, attribute and relation. Namely, the features and corresponding parameters of entities and conceptions are studied by certain rules. At the same time, the relationship of entity and cpnception is described. Agriculture ontology is choosen as the subject and domain knowledge organization in this paper. It can not only deal with the inner basic relation in agriculture subjection, but also more formal specific relationship. Formalized agriculture ontology is defined:

Agri_Onto=(Onto_Info,Agri_Concept, AgriCon_Relation, Axion)

Onto_Info is the basic information of ontology including name, creator, design time, midified time, aim, souce and so on; Agri_Concept is the set of agriculture knowledge conception; AgriCon_Relation is relationship set of conceptions including hierarchical relationship and non hierarchical relationship; Axion includes axiomatic set in ontology.

3.2 Conception Semantic Correlation

In domain ontology, there are words to present class and conception. And the words are not only serve as the bridge for class and conception, but also basic element. The relationship in ontology depend the words to connect with each other, so word set is the key of building agriculture ontology. In this paper, the class words and conception words in agriculture ontology are isolated to make up conception word set. T is conception sum in ontology. Matrix is used to construct conception correlation.

$$R=\begin{bmatrix} R(C_1,C1) & R(C_1,C_2) & \ldots & R(C_1,C_T) \\ R(C_2,C_1) & R(C_2,C_2) & \ldots & R(C_2,C_T) \\ .. & \ldots & \ldots & \ldots \\ R(C_T,C_1) & R(C_T,C_2) & \ldots & R(C_T,C_T) \end{bmatrix}$$

Formula(1)

In matrix, $R(C_1, C_T)$ is semantic correlation of C_1 and C_T . With respect to ontology conception correlation calculation, there are a lot of scholars to study the method. It is stated that there are always two methods: Information capacity and Conceptual distance. Conception semantic correlation method is adopted in this paper. Described as follows:

If conception C_i and C_j are synonymous relationship, the correlation of C_i and C_j is 1 and $R(C_i, C_j) = 1$; semantic correlation of no synonymous conceptions C_i and C_j is calculated by the following formula:

$$R(C_i, C_j) = \frac{(Dist(C_i, C_j) + \alpha) * \alpha * (d(C_i) + d(C_j))}{CE(C_i, C_j) * 2 * Dep * \max(|d(C_i) - d(C_j)|, 1)}$$

Formula(2)

Therein, $d(C_i)$ and $d(C_j)$ are their corresponding levels in the binary tree; $Dist(C_i, C_j)$ is the weight of all weighted edges in the shortest route from C_i to C_j;

; $CE(C_i, C_j)$ is the sun of edges in the shortest route from C_i to C_j; Dep is max depth of ontology tree; α is a controllable parameter, generally more than o or o. On this basis, semantic correlation matrix R can be built to represent all conceptions in agriculture ontology.

3.3 Calculation Documents Similarity

The key of automatic document clustering is to calculate similarity of documents. The most widely used method is cosine measure. Compare two documents

$d_i = (t_{i1}, w_{i1}, t_{i2}, w_{i2}, ...t_{iu}, w_{iu})$ and

$d_j = (t_{j1}, w_{j1}, t_{j2}, w_{j2}, ...t_{jv}, w_{jv})$

Included angle cosine is used to present the level of similarity of documents:

$$Sim(d_i, d_j) = \frac{(d_i, d_j)}{\|d_i\| * \|d_j\|} = \frac{(d_i, d_j)}{\sqrt{(d_i, d_i) * (d_j, d_j)}}$$

$$(d_i, d_j) = \sqrt{\sum_{m=1}^{u} \sum_{n=1}^{v} w_{im} * w_{jn} * Sim(t_{im}, t_{jn})}$$

$$\|d_i\| = \sqrt{(d_i, d_i)} = \sqrt{\sum_{m=1}^{u} \sum_{n=1}^{v} w_{im} * w_{jn} * Sim(t_{im}, t_{jn})} \qquad \text{Formula (3)}$$

Therein, t_{im} is conception feature word; w_{im} is corresponding weight; $Sim(t_{im}, t_{jn})$ is semantic similarity of conception feature words, which can be got from formula (1).

4 Text Mining System Design Based on Domain Ontology

Combined with text mining and domain ontology, the text mining model based on domain ontology is put forward. The basic processing route is that: at first, "Conception - conception" correlation matrix of ontology is built. And the documents in which feature words are extracted are represented to space vector model based on conception. Then similarity between documents are calculated according to "Conception - conception" correlation matrix. At last, clustering analysis method is used to mine the deep knowledge.

4.1 System Frame

The text mining system frame mades up of six modules: text mining pretreatment, text feature extraction, text mining, ontology management, ontology reasoning, evaluation and output the modes.

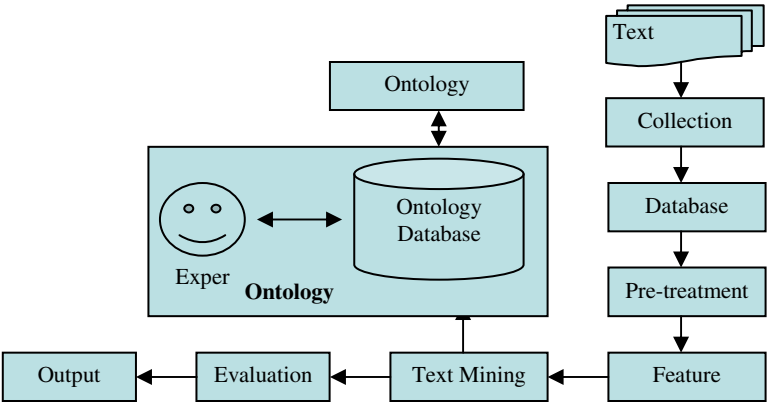

Fig. 1. System Structure

4.2 Structural Design

4.2.1 Text Data Pretreatment

The data source of text mining is unstructured text collections. They can are web pages, text documents, text files, word and excel documents, pdf files, E-mails and various forms of electronic documents. After the data resource acquired, they should be pre-treated. The process includes: data cleaning, such as denoising and incline

deduplication; data selection: appropriate and related to field text data are selected; text segmentation: conception set in ontology is regarded as references to realize professional vocabulary segmentation.

4.2.2 Text Feature Extraction

After data pretreatment, the text feature words must be extracted from the "clean" data. The process includes: (1) feature representation: The VSM is adopted to represent the documents. In the VSM, every document is presented to the following feature vector:

$$v(D)=(t_1,w_1(d);t_2,w_2(d);...;t_i,w_i(d);...;t_n,w_n(d))$$

Thereinto, t_i is key words; $w_i(d)$ is the weight of t_i in document d. $w_i(d)$ is often defined as the function of $tf_i(d)$ that is the occurrence frequency of t_i in documentd, namely $w_i(d)=\psi(tf_i(d))$. The frequently-used ψ is the function TF-IDF:

$$\psi = tf_i(d) * \lg \frac{N}{n_i}$$

Thereinto, N is the sum of documents. n_i is the sum documents of t_i.

(2)Feature extraction: the disable word list and threshold λ_d are used to delete the unimportant words to reduce dimensions of document space. According to the function of TF-IDF to calculate the weight of key word t_i. If the weight is more than λ_d set in advance, the key words are reserved, or deleted. Obviously, a much larger threshold will filter too many words. And the left words can not represent the document content well. Instead, a much lower threshold will affect feature selection too little. Through analysis of a large number of experimental data. when $\lambda_d = 0.3$, a much better filter effect will come into being.

4.2.3 Text Mining

Clustering analysis algorithm for mining is used and the aim is to divide the objects set into multiple Classes made up of similar objects. Firstly, every document is represented to feature vector based on key words $v(D)=(t_1,w_1(d);t_2,w_2(d);...;t_i,w_i(d);...;t_n,w_n(d))$. Then match t_i with domain ontology conceptions, if so t_i is taken place by domain ontology conceptions. Otherwise, t_i is regarded as unknown word. After matching, the key words are taken place by domain ontology conceptions in order that text document is represented to conception set. Then clustering analysis is used to cluster large amount of documents to few meaning cluster quickly so that the hidden knowledge or mode is acquired.

4.2.4 Ontology Management

Ontology management is the key of the text mining model and provides semantic support for model. The main work of ontology management is building, storing, maintaining and optimizing the domain ontology. It provides a platform for users easy to build and maintain ontology database conveniently; to manage the ontology database in order to build, add, delete and modify the classes, relations and constraint rules in ontology; to find new conceptions or cases to extend ontology structure on the basis of text mining clustering algorithm.

4.2.5 Ontology Reasoning

The function of ontology reasoning is semantic reasoning clustering results and deleting redundant or useless cluster by domain ontology as background knowledge or priori knowledge. It can refine and generalize related knowledge to enhance effectiveness and feasibility of clustering results.

4.2.6 Evaluation and Output

The knowledge acquired from text mining may be inconsistent non-intuitive and not easy to understand. It is necessary to post process text knowledge. The main two indexes often used to evaluate text clustering effect are recall rate and accuracy rate to reflect completeness and correctness. The evaluation method which combined the two indexes is F measure.

$$\mathrm{Re}(i,j)=\frac{n_{ij}}{n_i}$$

$$\mathrm{Pr}(i,j)=\frac{n_{ij}}{n_j}$$

$$F(i,j)=\frac{2*\mathrm{Pr}(i,j)*\mathrm{Re}(i,j)}{\mathrm{Pr}(i,j)+\mathrm{Re}(i,j)}$$

Therein, n_{ij} is the sum of class i in cluster j; n_i is the sum of documents of class i; n_j is the sum of documents in cluster j; n is the sum of documents.

4.3 Application Analysis

In order to verify the effectiveness of the system, 400 documents are selected from agricultural encyclopedia column of Chinese Agriculture Academy Science website as research objects which are 100 documents of farming, 100 documents of aquaculture, 100 documents of plant protection and 100 documents of veterinary medicine. After the 400 documents are pretreated, K-means based on space vector model and conception space vector model are used to realize cluster analysis. Clustering results are compared with accuracy, Recall and F measurement. The results are shown in the following table.

Table 1. Clustering results comparison

Arithmetic	Accuracy Rate	Recall Rate	F Measure
K-means Based on VSM	70.1	85.3	59.4
K-means Based on CVSM	84.2	93.8	89.9

As can be seen from the experimental results, space vector model based on domain ontology in which conceptions in domain ontology are instead of feature words so that correlation of feature words are reduced and dimensions of document vectors are decreased making better clustering results in accuracy, recall, F measure than K-means based on space vector model.

5 Conclusion

With the development of information technology and network resource, a flood of information is produced. To analyze the text content and potential valuable knowledge, this paper put forward text mining model based on domain ontology and introduced "conception-conception" correlation matrix and used space vector model based on domain ontology instead of space vector model to represent document and clustering algorithm to discovery knowledge.

References

1. Feldman, R., Dagan, I.: Knowledge discovery in textual databases (KDT). In: Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD 1995), pp. 112–117. AAAI press, Montreal (1995)
2. Rosso, P., Ferretti, E., Jimenez, D., Vidal, V.: Text Categorization and Information Retrieval Using WordNet Senses. In: Proceedings of the Second Global Wordnet Conference GWC 2004, pp. 299–304 (2004)
3. Sedding, J., Kazakov, D.: WordNet-based Text Document Clustering. In: Proceedings of the Third Workshop on Robust Methods in Analysis of Natural Language Data (ROMAND), Geneva, pp. 104–113 (2004)
4. Ino, Y., Matsui, T., Ohwada, H.: Extracting Common Concepts from WordNet to Classify Documents. Artificial Intelligence and Applications, 656–661 (2005)
5. Shehata, S.: A Wordnet-based Semantic Model for Enhancing Text Clustering. In: 2009 IEEE International Conference on Data Mining Workshop, pp. 477–482 (2009)
6. Bloedorn, S., Cimiano, P., Hothon, A., Staab, S.: An Ontology-based Framework for Text Mining. LDV-Forum 20(1) (2005)
7. Hotho, A., Staab, S., Stumme, G.: Ontologies improve text document clustering. In: Third IEEE International Conference on Data Mining, ICDM 2003, pp. 541–544 (2003)

8. Song, M.-H., Lim, S.-Y., Park, S.-B., Kang, D.-J., Lee, S.-J.: Ontology-based automatic Classification of Web Pages. International Journal of Lateral Computing 1(1) (2005)
9. Zelikovitz, S.: Transductive LSI for Short Text Classification Problems. In: Proceedings of the 17th International FLAIRS Conference. AAAI Press, Miami (2004)
10. Aurora, P.-P., Rafael, B.-L., José, R.-S.: Topic discovery based on text mining techniques. Information Proceeding and Management (43), 752–768 (2007)
11. Fung, G.P.C., Yu, J.X., Lam, W.: Stock prediction: Integrating text mining approach using real-time news. In: Proceedings of the 2003 IEEE International Conference on Computational Intelligence for Financial Engineering, pp. 395–402 (2003)