Research on Semantic Text Mining Based on Domain Ontology

Lihua Jiang [1,2], Hong-bin Zhang [3]， Xiaorong Yang [1,2]， Nengfu Xie [1,2]

1 Agricultural Information Institute of Chinese Academy of Agricultural Sciences, Beijing, 100081

2 Key Laboratory of Digital Agricultural Early-warning Technology, Ministry of Agriculture, P. R. China, Beijing, 100081

3 Institute of Agriculture Resources and Regional Planning of Chinese Academy of Agricultural Sciences, Beijing, 100081

{Jianglh@caas.net.cn}

Abstract: Text mining is an effective means of detecting potentially useful knowledge from large text documents. However conventional text mining technology cannot achieve high accuracy, because it cannot effectively make use of the semantic information of the text. Ontology provides theoretical basis and technical support for semantic information representation and organization. This paper improves the traditional text mining technology which cannot understand the text semantics. The author discusses the text mining methods based on domain ontology, and sets up domain ontology and database at first, then introduces the "concept-concept" correlation matrix and identifies the relationships of conceptions, and puts forward the text mining model based on domain ontology at last. Based on the semantic text mining model, the depth and accuracy of text mining is improved.

Key Words: ontology; domain ontology; text mining; semantic text mining

1 Introduction

Natural language is the main means to exchange and express thoughts and ideas in social-economic living for people. Though study of natural language for a long time,

the ability of interpretation is still limited. The existed technology has solved single sentence analysis, but it is hard to cover all language phenomenon, especially for the whole paragraph and chapter.

In the early nineteenth century, the data mining technology based on statistical technology had developed maturely, and application in the large-scale structural relational database achieved success. So people want to apply data mining technology to analyze natural language text and the method is called text mining or knowledge discovery in text. Difference with the traditional natural language processing words and sentences, the main goal of text mining is concentrated on founding hidden meaningful knowledge in mass text set, that is the understanding of text set and the relationship among texts. Now most of the text mining applications are lack of semantic level consideration, only in grammar level processing, so the obtained result is not pretty.

2 Text Mining Based on Domain Ontology

Text mining or knowledge discovery in text is the process of extracting unknown, useful and understandable pattern or knowledge in mass text data（Hearstma, 1997; Feldman R, 1995）. The research object is semi-structural or unstructured and natural language text contains multi-level ambiguities. So text mining has brought a lot of difficulties.

The traditional text mining method based on vector space model represented the text to lexical frequency vectors and mined the vectors (Mothe, 2001; Ghanem, 2005). The defect of this method is only dealing with document forms and ignoring the semantic role. This is the primary cause that traditional text mining can not live up to expectations. So text mining technology is needed to combine with semantic analysis to realize text mining in semantic level (Xin Xu, 2004; Chih-Ping Wei, 2008; Hui-Chuan Chu, 2009). Appling domain ontology to text mining provides theoretical support for semantic text mining and also provides a feasible technology way.

With the help of ontology to mine text sets, it amounts to a “domain expert’ is equipped to the process of text mining to guide the whole process. According to features of text mining applications, ontology is divided into common ontology and domain ontology. Common ontology usually starts with the epistemology of the philosophy and extracts the relationship of general objects. The typical semantic

dictionary established by common ontology is English WordNet and Chinese HowNet. At present, there are many text mining methods based on WorNet and HowNet (Rosso, 2004；Sedding J，2004；Raymond, 2000；Y. Ino, 2005；Shehata S, 2009). But the text mining methods based on common ontology are very difficult to achieve good results in some specific domains. Therefore, some researchers began to develop text mining study based on domain ontology. Bloehdom etc. put forward the OTTO frame（OnTology Based Text mining frame wOrk）(S. Bloehdorn, 2005).OTTO used text mining to learn the target ontology from text documents and used then the same target ontology in order to improve the effectiveness of both supervised and unsupervised text categorization approaches. The bag of words representation used for these clustering methods is often unsatisfied as it ignores relationships between important terms that do not co-occur literally. In order to deal with the problem, Hotho etc. integrate core ontology as background knowledge into the process of clustering text documents（Hotho A，2003）. Song etc. suggested an automated method for document classification using an ontology, which expressed terminology information and vocabulary contained in web documents by way of a hierarchical structure (Song Mh, 2005).

In our country, knowledge engineering lab of computer science department in Tsinghua university developed text mining platform based 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 the domestic research theory spread stage, and relative actively in foreign

countries. But there is few whole text mining based on domain ontology solutions results. 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 Key Technology of Text Mining Based on Domain Ontology

At present, Most of the ontology systems are almost the same in basic structure. Most of ontology is described for the entity, conception, generally properties and relationship. That is, by the way of some rules, the characteristic features and the corresponding parameters of the entities or conceptions are studied. At the same time, the relationship of entity and conception is described.

3.1 Knowledge Presentation Based on Domain Ontology

Knowledge presentation is the foundation of text mining. The quality of the knowledge representation is related to the efficiency of text mining. This paper uses WordNet and OWL to organize and present knowledge. This paper studies the following ideas: after data preprocessing, universal transformation format and conception extraction, domain ontology and ontology database are built by WordNet and OWL. Knowledge construction is defined and knowledge index is built in conceptual level.

In this paper, we take agriculture ontology as an example. Agriculture ontology is the system including agriculture terms, definition and standard relationship of terms. It is also formalize expression of conceptions and the relationship among conceptions. At the same time, it can not only deal with the inner relations among agriculture subject tale, but also more formal special relations. Formalize agriculture ontology is defined:

Agri_Onto=(Onto_Info, Agri_Concept, AgriCon_Relation, Axion)

Therein, Onto_Info is the basic information of ontology including name, creator, design time, modification time, aim and knowledge resource and so on; Agri_Concept is the set of agriculture conceptions; AgriCon_Relation is the conceptual relation set

including hierarchical relationship and the hierarchical relationships; Axion includes existing axiomatic set in ontology.

3.2 Construction conceptual semantic correlation matrix

Combined with the features of domain ontology and text mining, this paper proposes the text mining model based on domain ontology, and the main idea is as following: at first, "conception-conception" correlative matrix of domain ontology is constructed, and then key words in the collected documents are extracted in the light of domain ontology and the documents are represented to vector space model in order to calculate similarity between documents by reference to the "conception-conception" correlation matrix to realize document clustering. If new conception is found in the process of document clustering, the ontology database should be enlarged.

In domain ontology, there are vocabulary to represent classes and conceptions which are not only the bridge to communicate with classes and conceptions but also basic elements to represent classes. The relationships of domain ontology depend on the words to connect, so vocabulary is the key of construction domain ontology. In this study, the class vocabulary and conception vocabulary are extracted from text documents to make up vocabulary set $C=\{C_1,C_2,...,C_T\}$, in which T is the sum of conceptions in ontology. This paper uses matrix to structure representation method of conception relatedness, as follows:

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

In the matrix, $R(C_1,C_T)$ represents semantic relativity of C_1 and C_T.The result of $R(C_1,C_T)$ is semantic relativity of C_1 and C_T. A lot of scholars have discussed the conception relativity calculation. Generally speaking, there are two kinds of methods, information capacity method and the concept distance method. This paper adopts concept distance method, because in the ontology conceptions are arranged in tree. The method of concept distance can not only reduce complexity of algorithm but also easy to calculate. Before calculation the relativity, first of all to do the following description: ① if C_i and C_j is similar, the $R(C_i,C_j)=1$; ② if C_i and C_j is not similar, according document [25] this paper adopts the following formula to calculate semantic relativity:

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

In the formula, $d(C_i)$ and $d(C_j)$ represent the layers that C_i and C_j located. $Dist(C_i, C_j)$ is the total weights in the shortest root from C_i to C_j in domain ontology tree. Dep is the max depth of ontology tree. α is a controllable parameter, generally more than equal to zero.

3.3 Identification of relationship of conceptions

The aim of text mining is to find the inherent, useful knowledge and implicit relationship. This paper introduces the identification mode to judge weather there is relation $(C_i \cup C_j)$ between C_i and C_j.The method of this paper is to set a threshold to judge the relation between C_i and C_j. If $C_i \cup C_j$ accounts for high proportion that is C_i always accompany C_j, C_i and C_j inevitably has the relationship. If the value of $(C_i \cup C_j)$ is higher than threshold, there is correlation between C_i and C_j. At the same time, the implicit correlation is found.

4 Text Mining Model Based on Domain Ontology

4.1 System Frame

Design ideas in the following figure：

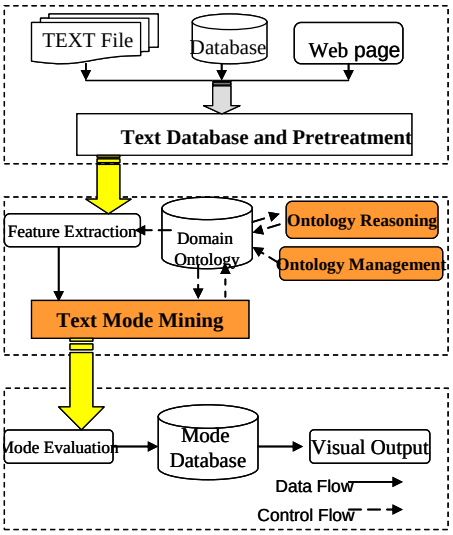

Fig1. Text Mining Model Based on Agriculture Ontology

4.2 Workflow of Text Mining Based on Domain Ontology

The text mining model based on domain ontology includes six parts.

⑴ Ontology management

Ontology management is the core of the whole text mining model and provides semantic support. It mainly takes charge of building, storage, maintenance and optimization domain ontology. After comparison several ontology building methods, this paper adopts Protégé developed by USA Stanford University to set up domain ontology. The building of ontology is a process of accumulation and updating domain knowledge.

⑵ Text database and text data pretreatment

The text data source of text mining is unorganized text documents, including Web pages, files, Words and Excels, PDF documents, E-mails and so on. Before acquisition text information, the text data must be pretreated including data cleaning such as noise reduction and duplication removal, data selection, text segmentation such as Chinese Word Segmentation and Paragraphs segmentation.

⑶ Text information extraction

After pretreatment, the text data must be clean and then feature information must be extracted including word segmentation, feature representation. After feature extraction, the text data will be changed to text information. And the text information can be stored in formal of structured or semi-structured.

⑷ Text mode mining

Text model mining is based on text information and needs the support of system resources including domain ontology database, ontology reasoning and ontology management. The module of text model mining use semantic model mining arithmetic to mine deep semantic knowledge.

⑸ Ontology reasoning

Ant colony optimization is introduced, combined with searching algorithm to realize intelligent semantic reasoning and provides technological support for mining deep semantic mode. The effect of ontology reasoning is to reason mining mode and obtain deeper level mode to avoid common sense knowledge.

⑹ Evaluation and output

The knowledge obtained from mining module may be inconsistent, non-intuitive and difficult to understand. So it is necessary to post process text knowledge including knowledge valuation and accept or reject, elimination knowledge inconsistency.

5 Conclusion

In order to improve the depth and accuracy of the text mining, a semantic text mining model based on domain ontology is proposed. In this model, conceptual semantic correlation matrix is applied to semantic analysis so that the semantic relations can be extracted accurately. The text mining model based on domain ontology in this paper can mine deep semantic knowledge from text documents. The pattern got has great potential applications.

Acknowledgment:

The work is supported by the special fund project for basic science research business Fee, AIIS “Tibet agricultural information personalized service system and demonstration” (No.2012-J-08) and The CAAS scientific and technological fund

project “Research on 3G information terminal-based rural multimedia information service” (No.201219).

References

1. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. J. Mol. Biol. 147, 195--197 (1981)
2. May, P., Ehrlich, H.C., Steinke, T.: ZIB Structure Prediction Pipeline: Composing a Complex Biological Workflow through Web Services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 1148--1158. Springer, Heidelberg (2006)
3. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
4. Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid Information Services for Distributed Resource Sharing. In: 10th IEEE International Symposium on High Performance Distributed Computing, pp. 181--184. IEEE Press, New York (2001)
5. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: an Open Grid Services Architecture for Distributed Systems Integration. Technical report, Global Grid Forum (2002)
6. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov

References

1. Hearstma. Text datamining Issues, techniques, and the relationship to information access [Z]. Presentation notes for UW/MS workshop on data mining, 1997.
2. Feldman R, Dagan I. Knowledge discovery in textual databases (KDT)[A]. Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD’95) [C]. Montreal, Canada, AAAI press, 1995: 112-117.
3. J. Mothe, C.Chrisment, T.Dkaki. Information mining – use of the document dimensions to analyse interactively a document set [Z]. European Colloquium on Information Retrieval Research, 2001: 6-20.

4. Ghanem, M.Chortaras, A.Guo, Y.Rowe, A.Ratcliffe, J.A grid infrastructure for mixed bioinformatics data and text mining [J]. In: Computer Systems and Applications, 2005, 34(1): 116-130.
5. Xin Xu, Gao Cong, Beng Chin Ooi, Kian-Lee Tan, Anthony K.H. Tung, Semantic Mining and Analysis of Gene Expression Data[A], In: Proceedings 2004 VLDB Conference[C], Morgan Kaufmann, St Louis, 2004:1261-1264.
6. Chih-Ping Wei, Christopher C. Yang, Chia-Min Lin, A Latent Semantic Indexing-based approach to multilingual document clustering [J], Decision Support Systems, 2008, 45(3):606-620.
7. Hui-Chuan Chu, Ming-Yen Chen, Yuh-Min Chen, A semantic-based approach to content abstraction and annotation for content management [J], Expert Systems with Applicaions, 2009, 36(2):2360-2376.
8. Rosso P, Ferrettle , Jmenez D, et al. Text Categorization and Information Retrieval Using WordNet Senses[Z].The Second Global Wordnet Conference GWC 2004[C].Czech Republic, 2004.
9. Sedding J, Kazakov D. Wordnet-based text document clustering [A]. Proceedings of the Third Workshop on Robust Methods in Analysis on Natural Language Data (ROMAND) [C]. Geneva, 2004: 104-113.
10. Raymond Kosala, Hendrik Blockeel. Web Mining Research: A Survey [A]. ACM SIGKDD[C]. 2000(2): 1-15
11. Y. Ino, T. Matsui, and H. Ohwada. Extracting Common Concepts from WordNet to Classify Documents [A]. Artificial Intelligence and Applications[C], 2005: 656-661.
12. Shehata S. A Wordnet-based Semantic Model for Enhancing Text Clustering[C]. 2009 IEEE International Conference on Data Mining Workshop, 2009: 477-482.
13. S. Bloehdorn, P. Cimiano, A.Hotho and S. Staab. An Ontology-based Framework for Text Mining [J], LDV-Forum, 2005, 20(1).
14. Hotho A, Staab S, Stumme C. Wordnet improves text clustering [A]. Proceedings of the Semantic Web Workshop at SIGIR-2003, 26th Annual International ACM SIGIR Conference[C], 2003.
15. SONG MH, LM SY, PARK SB, Ontology-based automatic Classification of Web Pages [J]. International Journal of Lateral Computing, 2005, 1(1).
16. Zheng Xuling, Zhou Changle, Li Tangqiu, Automatic Acquisition of Chinese Semantic Collocation Rules Based on Association Rule Mining Technique [J]. Journal of Xiamen University (Natural Science), 2007, 46(3): 331-336.

17. Zou Guobing, Xiang Yang. Information Search Model Based on Domain Ontology [J]. Journal of Tongji University (Natural Science), 2009, 37(4): 545-549.
18. Zhu Hengmin, Ma Jing, Huang Weidong, Fan Huangxi, Study on Method of the Global Web Intelligent Search Based on Domain Ontology [J]. Journal of the China Society for Scientific and Technical Information, 2010, 29(1): 9-15.
19. Zhang Yufeng, Hechao, Research on Semantic Text Mining Based on Domain Ontology [J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(8): 832-839,
20. E. Bingham. Topic identification in dynamical text by extracting minimum complexity time components [Z]. In: Proceedings of ICA2001. 2001:546-551.
21. M Montes-y-Gómez, A. Gelbukh, A.López-López. Discovering ephemeral associations among news topics [Z]. In: Proceedings of IJCAI- 2001 Workshop on Adaptive Text Extraction and Mining. 2001:216-230.
22. Zelikovitz S. Transductive LSI for Short Text Classification Problems[C]. Proceedings of the 17th International FLAIRS Conference, Miami: AAAI Press, 2004.
23. Aurora Pons-Porrata a, Rafael Berlanga-Llavori b, Jose' Ruiz-Shulcloper. Topic discovery based on text mining techniques [J]. Information Proceeding and Management. 2007(43):752-768.
24. Pui Cheong Fung, G.Xu Yu, J.Wai Lam. Stock prediction: Integrating text mining approach using real-time news [Z]. In: Computational Intelligence for Financial Engineering, 2003 IEEE International Conference, 2003: 395-402.
25. Wu Jian, Wu Zhaohui, Li Ying, Web Service Discovery Based on Ontology and Similarity of Words [J]. Chinese Journal of Computer, 2005, 28(4):595-602.