Research on Ontology-based Representation Method for Textual Cases

Dong Wang, Yang Xiang, Qian Chen, Guobing Zou
College of Electronics and Information Engineering,
Tongji University,
Shanghai China
E-mail: superwang1981@163.com

***Abstract*-How to map from texts to structured case representations and how to automatically generate representations have become the research hotspot in the fields of Textual Case-based Reasoning (TCBR). This paper presents methods that support automatically generation ontology-based representation for textual cases. We used the Ontology to describe the relationship between terms in application fields. First, we defined a new formal description method for Ontology. And then we propose a representations model for textual cases and describe the method of automatic generation method of this Model. Finally, based on the proposed model, a marketing case representation and management system is established. A real-world case of marketing is also applied in order to verify the feasibility of the proposed model. The verification results show that the system is efficient for keeping semantic information. As a whole, this research provides a knowledge representation and the automatic generation approach to facilitate knowledge management to efficiently and accurately describe the contents of text knowledge. The proposed model can be applied in TCBR to enhance reuse of domain knowledge.**

Keywords-Textual Cases; Textual Case-based Reasoning; Ontology; Representation formalism

I. INTRODUCTION

Case-based Reasoning (CBR) is a novel approach to problem solving and learning and has got a lot of attention over the last few years. CBR emulates human psychological mechanisms and takes successful experience as reference to help solve the new problems. The text-based CBR system (TCBR) is one of the most active sub fields in this domain. The cases in knowledge database of TCBR are presented in text form. The goal of TCBR is to extract similar text cases automatically or semi-automatically. Recently, there has been significant progress in theoretical support and guidance for TCBR. However, as other decision support system, it is facing many problems in the course of its application and development, especially the problem in how to represent text case and how to generate representations for TCBR automatically. [1, 2]

The most widely used method of text representation is "bag of words". 1997, the text representation method used in the SPIRE model proposed by Daniels and Rissland was bag of words, but the case retrieves in TCBR system dependent on information retrieval system and emphasized the extraction of feature word weight.[3] Lenz (1999) tried adding semantic information to text representation through WordNet. [4] Recent work in TCBR has considered other, more advanced representations. Wiratunga et al. (2004) introduced a fully automated method for extracting predictive features to represent textual cases. This approach included extracting feature words and analogizing semantic relations between words through association rule induction, additionally, finding logical combination of keywords in extended algorithm. [5] Cunningham et al. (2004) investigated automatic structural map to represent textual cases for TCBR. This approach retained some useful syntactic information by translating text into a network structure: the feature words of case serve as nods, the relationships of feature words are as edges. But, this method can not distinguish problems from solution, which go against re-utilization of case. [6] Guptą & Aha (2004) proposed a natural language understanding approach to extract deep semantic for TCBR that derives a first-order representation of the case texts. But, this method has been not yet perfect, and need to be improved in future. [7]

With the gradual maturing of ontology technology, it is gradually became a hotpot in the field of artificial intelligence including knowledge engineering and knowledge representation, etc. Ontology is the model about abstract description of the essence of things, which is a method for domain knowledge sharing and common understanding. This paper researched on ontology-based representation method for textual cases. This method supported automatically generation ontology-based representation for textual cases and solved the issue of lacking semantic information in the process of case representation in TCBR system. The experimental results showed that our method had higher accuracy and effectively decreased the latitude of text case representation, moreover, this method reserved semantic information of cases.

II. ONTOLOGY FORMAL MODEL

Domain ontology defines class, instance, property, relationship and axiom. It concludes and abstracted domain knowledge by describing in detail the conception, property of conception and relation among concepts. In application, the ontology that was constructed by ontology editing tools such as Protégé, Kaon, Ontolingua to make files corresponding to ontology description language (XML, RD, OWL, DAML+OIL, etc.). Formalization of domain ontology is to formalize the semantic information which are described by ontology language. In 1998, Guarino formally defined ontology as domain space structure: <D, W>, D: domain; W: the widest object state set among D. ρ^n denotes all n-element concept mappings from W to D in domain space. The conception of D can be represented as a ordered three-dimensional group: C=<D, W, R>, D: domain; W: the widest object state set among D; R is the set of the conception relationship (ρ^n) in <D, W>.[8] On the ontology based frame

978-0-7695-4225-6/10 $26.00 © 2010 IEEE
DOI 10.1109/AICI.2010.127

annotated by Semantic Web, the ontology is formalized as a set including six elements: {C, A^C, R, A^R, H, X}, C: concept set; A^C: attribute set of every conception; R: relationship set; A^R: attribute set of each relationship; H: concept hierarchy; X: axiom set.[9] Based on the above research, we defined ontology model as a five elements set including class or conception, property, relationship, axiom, instance.

Definition 1. Ontology model is represented as a five-tuple array {C, P, R, I, A }, C: conception or class set; P: property set; R: relationship set; I: instance set; A: axiom set.

It is far from enough to describe the formalized ontology model as a five-tuple array. To reserve various semantic information of ontology as possible, we defined and particularly described some of the five tuples of which the conception and axiom are one-tuple, therefore they are needless to be defined.

Definition 2. Property set P is a set of the same properties that are included in the corresponding conceptions. It can be expressed as two-tuple array: P={C_i, P_i}, of which C_i is the conception with property P_i.

Definition 3. Relationship set R is the hierarchical relation (father-son relationship) between conceptions, which can be represented as triple array: R={C_i, R_i, C_{i+1}}, C_i and C_{i+1} is the subject and object of the conception described by relationship R^i, respectively.

Definition 4. Instance set I can be a binary array: I={C_i, I_i}, C_i is conception of instance I_i.

After formalized, the ontology posses it own prolific semantic structure, and can describe more complex object.

III. ONTOLOGY-BASED REPRESENTATION FOR TEXTUAL CASES

In this chapter, a text representation model based on ontology was proposed and the method of automatic generation of text representation based on this model was introduced.

By using ontology , we extract the concept , relationships , instances , which could express the text's content. So we achieve the goal in text representation.

Definition 5. Ontology text model was defined as a five-tuple array:

$$Onto\_Case = \{C\_Case, P\_Case, R\_Case, I\_Case, A\_Case\}$$

C_Case : conception set of cases; P_Case : properties of conceptions; R_Case : subordinate relations of conceptions; I_Case : instances set of conceptions; A_Case : axiom set , it was used to define the relationship of two properties in P_Case such as functional relationship, inverse-function relation, transitive relation, etc.

The detailed definition of each tuple in $Onto\_Case$ model is same as that in ontology.

3.1 Generation of conception set

The generation of conception is divided into two parts. One is to extract dominant part of ontology conception set of case using keywords matching algorithm, with the aid of basic conception in domain ontology. The other is extract recessive part of ontology conception set of case with automated reasoning function of ontology. The algorithm is as follow:

Algorithm 1. Generation of conception set

Input: text *t*, domain ontology *O*

Output: conception set C_Case

Begin

Step 1. Change *t* to vector space model-*S*. *S* contained text segmentation and deleting stop words of *t*.

Step 2. Retrieve in *S* under the matching condition of conception C in ontology and relationship R, by means of keywords matching algorithm. The retrieval results- matching conceptions were reserved as dominant knowledge. However, the matching properties will be used in the subsequent ontology reasoning.

Step 3. Return dominant knowledge conceptions and matching relationship of properties to ontology to perform concept reasoning. The result of reasoning is the recessive part of conception set of ontology based case representation.

Step 4. Return text conception set C_Case that is composed by dominant and recessive conception.

End

According to Algorithm 1, we may draw such conclusion that text conception set includes not only its own conception but also the potential conception of semantic relatedness which is from ontology reasoning.

3.2 Generation of relationship set

Subordinate relation, as the most universal and the most important relation between the conceptions, composed the hierarchical structure of concepts. The concepts we proposed in above champer 3.1 generally distribute in every corner of the Ontology hierarchical structure. So subordinate relation of these conceptions should be reorganized to reconstruct the tree hierarchical structure of conceptions.

The hierarchical structure of conceptions of cases was reconstructed in the bottom up order by recursion method, in this paper. Algorithm is as follow.

Algorithm 2. Generation of relationship set

Input: text conception set V , domain ontology *O*

Output: conception set R_Case

Begin

Step 1. Scan concept hierarchy of domain ontology, reserve the conceptiones in leaf node of hierarchical structure to form a set , expressed as $C = (c_1, c_2, c_3, \cdots, c_n)$.

Step 2. Operate the intersection of sub-concept set C and text conception set V. Thus, a set of conception in leaf node of text hierarchical structure was gained and expressed as $T=(t_1,t_2,t_3,\cdots,t_l)$, $l \le n$.

Step 3. Search the parent nodes of each conception t_i in conception hierarchical structure of domain ontology. And add the parent nodes to text conception set V.

Step 4. Set all parent nodes in step3 as new leaf nodes. And then return to step3 to search their parent nodes, circulating until there is no parent node.

Step 5. Return all combinations of hierarchical relationship, that is R_Case.

End

3.3 Generation of property set

In domain ontology, the property of conceptions defined the relationship between existence and conception besides subordinate relation. On the Basis of definition of concept attribute, the property consisted of its name, domain and range. Of which, the domain and range of property is actually the set of conception of domain ontology. Therefore, if the domain and range of a certain property exists in text conception set, it should present in formalized text. The algorithm is as follow:

Algorithm 3. Generation of property set

Input: text conception set V, domain ontology O

Output: property set P_Case

Begin

Step 1. Scan property of domain ontology to gain set P.

Step 2. Operate the intersection of property set P and text conception set C_Case to gain property set P_Case related to text conception set.

Step 3. Return P_Case.

End

3.4 Generation of instance set

In domain ontology, the instance of conception is dependent on conception. One cenception posses several instances, moreover, one instance corresponds to multiple conceptions. Consequently, the instance set of conception was obtained by retrieveing enery conception in text conception set directly, according to the relationship between conception and instance in ontology definition. The algorithm was omitted.

Through the method introduced above, the conceptions, relationships, properties and instances in the case were obtained and the inner logic of text cases was reserved in the maximal degree, so semantic information were preserved. Moreover, this method reduced the dimension of the feature because concepts were only a small part in the case.

IV. SYSTEM DEVELOPMENT AND VERIFICATION

Based on the ontology-based representation model, a marketing ontology and marketing knowledge representation and management system (MKRMS) is established in this research. We first introduce the architecture of our MKRMS. Afterward, a demonstration case in the marketing field is applied in order to verify feasibility of the proposed system.

4.1 System Architecture and Functions

We exploited marketing domain ontology using Protégé. And it included more than 2000 vocabularies. Fragment of marketing ontology is shown in Fig. 1.

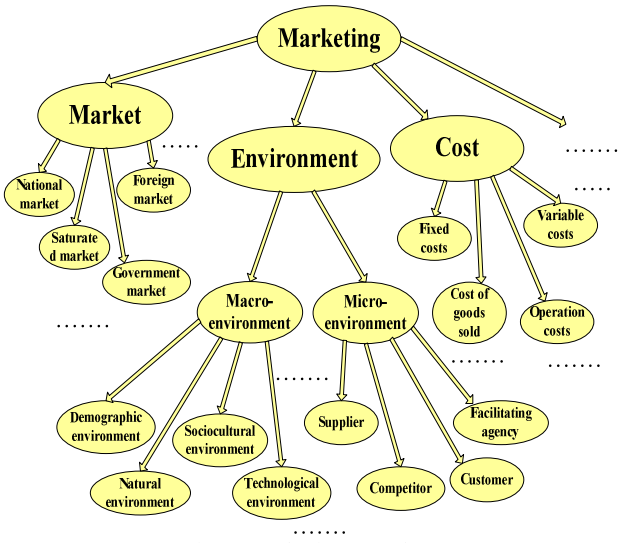

Fig. 1.Marketing Ontology

As shown in Fig. 2, the core framework of MKRMS shows that the main system processes include "Ontology Formalism," "Textual Cases Process" and "Matching".

We used Chinese lexical analyzer ICTCLAS for Chinese word segmentation. [10]

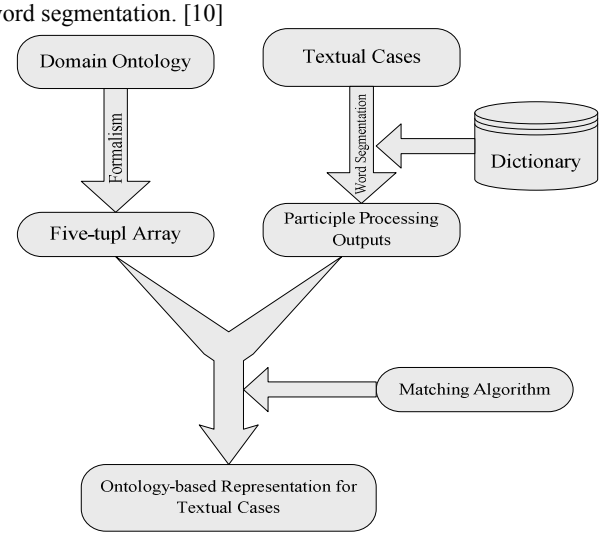

Fig. 2.The core framework of MKRMS

4.2 Case Study and System Verification

In order to verify the feasibility of the proposed MKRMS, a demonstration case in the market segmentation field is applied.

In the following, the experimental results are analyzed and interpreted.

Based on the ontology-based representation model, the case was expressed as follow:

{

{(market segmentation), (market), (segment market), (market subdivision strategy), (brand), (product), (profit), (market choice), (consumer), (advertisement), (advertisement expansion), (after service), (brand premium), (brand loyalty), (product quality), (process of market segmenting), (product segmentation)};

{(consumer, consumption habit), (consumer, buying motives), (product segmentation, linear type product segmentation), (market segmentation, process of market segmenting)};

{(market, hierarchical relation, market segmentation) (market segmentation, hierarchical relation, process of market segmenting) (market segmentation, hierarchical relation, market subdivision strategy) (brand, hierarchical relation, brand premium), (brand, hierarchical relation, brand loyalty), (product, hierarchical relation, product quality), (market, hierarchical relation, market choice), (advertisement, hierarchical relation, advertisement expansion)};

{};

{};

}

The example illustrated as follows： the automatically generated case representation can not only emphasize the feature concept of cases, but also embody the relationship between conceptions. The semantic logic of case is preserved effectively.

V. CONCLUSION

This paper introduced a method of case expression based on ontology in TCBR system. From this research, core information contained within large quantities of text-oriented documents can be automatically extracted, formatted, and represented in order that domain knowledge could be more concretely expressed and knowledge acquisition efficiency could be enhanced. Moreover, the proposed approach could also be applied in knowledge management, long-distance self-learning environments, and any other knowledge acquiring environment in order to lower knowledge management costs, improve knowledge acquisition efficiency, and enhance domain knowledge reuse.

In the future, the proposed information extraction method could be integrated with other NLP technology in order to improve the capability of text parsing, and the range of the marketing ontology knowledge could be expanded in order to enhance the usage and usability of the proposed system.

ACKNOWLEDGMENT

This work is funded by the National '863' High-Tech Research and Development Plan of China under Grant No. 2008AA04Z106, the NSFC under Grant No. 70771077, the Project of Science and Technology Commission of Shanghai Municipality under Grant No. 08DZ1122300 and the Special funds for the development of Shanghai Information Project No. 200901015.

REFERENCES

[1] S. Bruninghaus, & K. D. Ashley, "Reasoning with Textual Cases" in H. Muñoz, & F. Ricci, eds. Case-Based Reasoning Research and Development (LNAI 3620), Springer, Berlin, 2005, pp. 137-151.

[2]R. Weber, K. Ashley and S. Bruninghaus, Textual case-based reasoning. The Knowledge Engineering Review, Vol. 20:3, 255–260, 2006, Cambridge University Press.

[3]J. Daniels and E. Rissland, 2006, Finding Legally Relevant Passages in Case Opinions. [Proc.6th International Conference on Artificial Intelligence and Law]

[4]M. Lenz, Case Retrieval Nets as a Model for Building Flexible Information Systems. Ph.D. Humboldt University, Berlin, Germany.

[5]N. Wiratunga, I .Koychev, and S. Massie, 2004, Feature selection and generalization for retrieval of textual cases. In Funk, P and González Calero, PA eds. Advances in Case-Based Reasoning [Lecture Notes in Artificial Intelligence, 3155]. Berlin: Springer, pp. 806–820.

[6]C. Cunningham, R. Weber, J. M. Proctor, C. Fowler, and M. Murphy, 2004, Investigating Graphs in Textual Case-Based Reasoning. [Proc. 7th European Conference on Case-Based Reasoning]

[7] K. Gupta, and D. W. Aha, 2004, Towards Acquiring Case Indexing Taxonomies from Text. [roc. 6th International Florida Artificial Intelligence Research Society Conference]

[8] Guarino N, Formal Ontology in Information Systems. Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998. Amsterdam, IOS Press, pp. 3-15.

[9] Peng Wang, Bao-Wen Xu, Jian-Jiang Lu, Da-Zhou Kang, Yan-Hui Li. A novel approach to semantic annotation based on multi-ontologies. IEEE International Conference on Machine Learning and Cybernetics. vol. 3, 2004, PP 1452-1457

[10] Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, Qun Liu. 2nd SIGHAN workshop affiliated with 41th ACL. Sapporo Japan, July, 2003, p184-187