2009 International Conference on Artificial Intelligence and Computational Intelligence

# Research on Ontology-Based Case Indexing in CBR

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

***Abstract*—This paper presents methods that support case retrieval in Case-Based Reasoning system. We used the Ontology to describe the relationship between terms in application fields. The similar cases are retrieval by calculating semantic similarity which we have defined. We evaluated traditional method of calculating the semantic similarity with lattice theory. We have constructed a decision support CBR prototype system of marketing strategy, based on this algorithm, which contains more than 600 cases. The evaluation shows that with the support of semantic, we can not only carry out data matching retrieval, but also perform semantic associated data access. CBR can quickly and accurately retrieve cases and improve efficiency of reasoning by semantic query.**

***Keywords*-Case Based Reasonin;, Case retrieval;, Ontolog;, Semantics similarity**

## I. BACKGROUND

Case-Based Reasoning (CBR) is a novel Artificial Intelligence(AI) method in dealing with questions and learning which is based on knowledge. It is the principal measure to solve new problems by consulting experiences and methods used in the same or similar problems. CBR supports decision simply by imitating this mechanism. Currently, as other DSS, CBR, is faced with many questions in the real-world application as well as in its own development, especially in terms of how to search for and extract cases efficiently.

As in the present circumstances that the operative mechanism of brain is not clear enough, it has become a research task to settle the problem of CBR using the existing theoretical basis and artificial intelligence (AI). Ontology has quickly become a hotspot of AI research, which covers knowledge engineering and knowledge representing etc., and has been used to establish knowledge bases, after its application in the computer field. Lorcan Coyle and his team in Trinity College Dublin Ireland published a case markup language named XML-based Case Mark-Up Language (CBML) and on this basis [1, 2] carried out researches on case retrieval, which has constructed a foundation for solving the problem of CBR system utilizing ontology theory.

The application effect of CBR system is evaluated by the accuracy of searches for similar cases. Though there are a number of relative investigations, the result is unsatisfactory. The present paper is to search after a method to realize higher reasoning efficiency and stronger solving ability and to achieve the goal of intelligent decision support by ontology applicaton in CBR system.

## II. RELEVANT CONCEPTS

The concept of ontology is originated from the philosophical field, which was put forward by the ancient Greek philosopher Aristotle (384-322b.c.). Its original definition in the philosophical field is the systemic description of objective existence in the world. It embodies the abstract essence of objective reality.

In the artificial intelligence field, the concept of ontology was initially forwarded by Neches etc. They have interpreted ontology as the definition of rules, which are composed by the basic terms and relations that are forming relative field words, to stipulate the extension of words. Neches considered that ontology defines the basic terms and relations of subject area vocabulary, and made the rules to define the extension of vocabulary based on them. In 1993, Gruber gave the most prevalent definition that ontology is the definite specification of conceptal model [3]. On this basis, in 1997 Borst gave another definition: ontology is the formal normalization of common concept model [4]. In 1998, Studer further researched the above two theories, he held that ontology is the definite formal normalization of common concept model [5].

Simply speaking, ontology is an entity and a result of analyzing and modeling a certain field through ontological method. In other words, ontology is to abstract a certain domain in the real world to a set of concepts and relations, which can be formalized as follows: $O = \{D, Cs, Rs\}$, ($D$: a certain domain, $Cs$: concept set, $Rs$: relation set).

Thus it can be seen that ontology offers a word set to describe the facts in a certain field, and all the ontology composes the knowledge of the domain. The goal of ontology is to capture the knowledge and provide common understandings of related fields, fix on common words in the fields and provide clear definition of these words and their relations from different levels of the formalization model. Now that ontology has become the descriptive language of knowledge, we can use it to express knowledge and set up

978-0-7695-3816-7/09 $26.00 © 2009 IEEE
DOI 10.1109/AICI.2009.449

knowledge base. Ontology enables communion and repetitive use of knowledge, improves the cross operation of heterogeneous systems, and promotes the sharing of information.

## III. ONTOLOGICAL CASE RETRIEVAL MODEL

Case retrieval is to search for one or more cases that are most similar to the present case from the case base. As asked, CBR finds out the most suitable cases from the case base according to the similarity degree and index. Quantity and quality of the retrieval result directly affect the problem-solving. Case retrieval includes three steps: feature discrimination, case matching and best selection. In conventional CBR, index is designed meticulously and similarity function is defined, but the variety of cases and semantic relations between the properties of cases are disregarded. The core task of ontology-based case retrieval is to let index machine analyze and understand characteristics of the problem, and retrieve the source of cases by semantic relations, through knowledge ontology. The model is shown in Fig. 1.

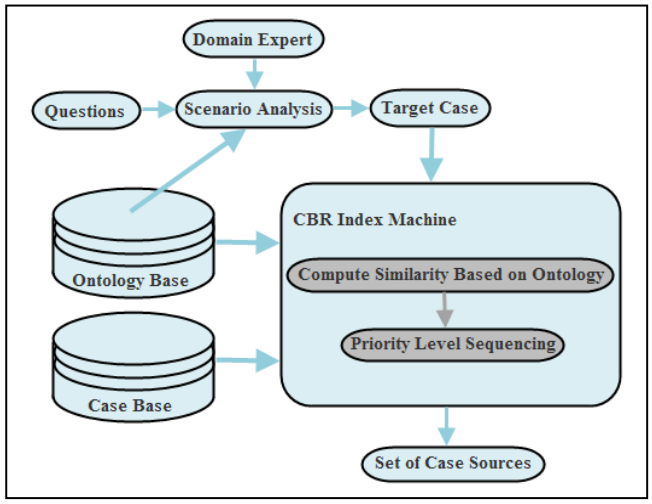


Figure 1. Schematic diagram of ontology-based case retrieva.

*1)* Scenario analysis: according to the context or requirement of the knowledge model, domain expert analyzes the question using the ontology technology, refines the scenario description, and extracts the appropriate characteristic information.

*2)* Generating target case: after the knowledge implicated in the question is marked in the ontology base, the question is conceptually modeled after the ontology case model and the target case is formalized. Then, the target case is filled with the knowledge elements set, scenario description of the case and reasoning target.

*3)* Similarity computing based on ontology: similar cases often have the same knowledge base, and the related knowledge can assist to solve the problem. Case retrieval based on the same or related knowledge base is more accurate than that based on property matching.

*4)* Priority level sequencing: select one or more cases that are most relative to the present question from the preliminary matching result. This step is closely related to the domain knowledge.

## IV. KEY ALGORITHMS IN ONTOLOGY-BASED CASE RETRIEVAL

The description of cases is composed of several properties and the case similarity is defined by the similarity degree of attributes. So, to compare two cases, we need to match their structure similarity of properties and the similarity of relevant attribute values, and organically combine them to measure the similarity of the cases. Actually, to match the comparability of properties and their values is to compute the relative knowledge similarity.

The computational method of semantic similarity of knowledge plays an important role in practical application of semantic with quantitative process concept. Ontological semantic similarity takes an important role in ontology integration and semantic information retrieval. At present, there are two main algorithms of semantic similarity: one is based on semantic distance; the other is on lattice theory. Here, we put forward a modified algorithm of semantic distance with principal advantage of both above.

The definition of ontological semantic distance is as follows:

Definition1. Semantic distance (SD): SD is a comprehensive measurement of the shortest length of relation chain in inheritance relationship of two different classes and common attribute relations of classes in the same ontology.

There are four points for attention in above definition:

*1)* SD only exists in different classes within the same cluster, and SD of different class cluster is infinity.

*2)* Make the best of characteristics of ontology. Relationship between ontological classes includes inheritance relation and binary relation. This dissertation is with a view to the method of calculating SD by comparing inheritance relation and common property.

*3)* Only the shortest length of relation chain can serve as the measure standard of SD. Because there may be several relation chains between classes, and the longest may be infinity.

*4)* The symmetry feature: calculation of SD should accord with the following equality: SD(A, B)=SD(B, A). The symmetry is beneficial to the comparison and conversion of similarity level of different concepts.

### *A. GCSM algorithm*

GCSM distance is defined as a dot product between two ontology concept nods of vectors, as the similarity of the two vectors is computed [6]. One characteristic of GCSM distance is that inheritance relation is calculated, whereas the other binary relation is not.

First of all, we introduce a concept named “Lowest Common Ancestor” (LCA).

Definition2. LCA refers to the ultimate one of the common ancestor of the two ontological concept nodes. Depth(node) is used to represent the depth of node, the depth of root vertex is 0, and by expanding one level, the depth of node adds 1. For example, depth(Organism)=0, depth(Animal)=1, depth(Fish)=2. Apparently, the LCA of "Plainwaterfish" and "Saltywaterfish" is "Fish", though "Animal" is also their "Common Ancestor", the depth of node of "Fish" is deeper. Similarly, the LCA of "Mammal" and "Fish" is "Animal" and the LCA of "Plant" and "Animal" is "Organism".

Definition3. GCSM distance, the calculation formula for GCSM distance is as follows:

$$\text{GCSM Distance} = \frac{\text{depth}(\text{uri}1) + \text{depth}(\text{uri}2)}{2 \times \text{depth}(\text{LCA}(\text{uri}1, \text{uri}2))}$$

Hence, the depth of concept node is deeper, while the GCSM distance is the shortest. Though the numerator is the sum of depth(uri1)and depth(uri2), the depth of their LCAs can counteract the enlargment effect of their own depths and the denominator is two-fold weighted for unitary quantification.

The limitation of GCSM distance is that it emphatically considers the global aspect of ontology conceptual network, but ignores the relationship of local feature of concept (property). Although, GCSM distance can embody the similarity level of the global structure of two concepts effectively, different values of common properties of classes which have the same ancestor in the same inheritance chain can also affect the similarity level of the classes to a large extent.

### *B. Lattice theory*

Lattice theory is a method to derive concept relations from attributed relationships. Lattice consists of partial ordering relation of a concept set, where the partial ordering relation is deduced from the order relation of property. For example, the properties of a color laser A4 printer are "color, laser, A4", that of a color inkjet A3 printer are "color, inkjet, A3". If color printer can replace monochrome printer and A3 printer can replace A4 printer, then we can draw a conclusion that "color> monochrome" and "A3>A4". Very few properties of them are different, so they are very much similar.

The limitation of lattice theory includes: (1) it emphasizes local but not global. For example, if there are concepts of "tiger" and "dog" in an ontology base, the properties of tiger and dog both consist of "four legs" and "zoophagous", then the erroneous deduction that "dog" can replace "tiger" will be drawn. (2) It is a qualitative metric form. It can only deal with the concepts which are similar in order relation.

### *C. Improved algorithm*

The new distance algorithm combines the advantage of GCSM and Lattice theory, which is thus enabled to comprehensively and accurately reflect the similarity between concepts from both a global and local view. The algorithm is as follows:

Firstly, compute GCSM as above.

Secondly, calculate $\Delta$A as follows:

The property of LCA is expressed as LCA_PropertySet (p1, p2, .., pn), just because uri1 and uri2 inherit LCA, thus also inherit LCA_PropertySet (p1, p2, .., pn). uri1 is (p'1,p'2,.., p'n) and uri2 is (p"1, p"2, .., p"n). p'i may be pi or a sub-property of pi. So

$$\left\{p'_i \,\middle|\, samas(p'_i, p_i) \| subproperty(p'_i, p_i)\right\}$$

$$\left\{p''_i \,\middle|\, samas(p''_i, p_i) \| subproperty(p''_i, p_i)\right\}$$

$$\left\{Rang(p'_i) \middle| Rang(p_i) \| SubClass(Rang(p_i))\right\}$$

$$\left\{Rang(p''_i) \middle| Rang(p_i) \| SubClass(Rang(p_i))\right\}$$

Then LCA(Rang(p'i), Range(p"i))=Range(pi)

Different properties are assigned with different weights, assuming the weight of (p1, p2, .., pn) is (w1, w2,…, wn).

$$\Delta A = \frac{\sum_{i=1}^{n} (w_i \times Sim)}{\sum_{i=1}^{n} w_i}$$

Refer to Tab.1for the detailed calculation flow.

There are four different relationships of property values in this paper:

*1)* Perfect matching, for example, user request="color", matching object="color", then, PropertyRelation((R'(i), R"(i))=1.

*2)* Exceeding matching, for example, user request="monochrome", matching object="color", then, PropertyRelation((R'(i), R"(i))=1.

*3)* Unsatisfactory matching, for example, user request="color", matching object="monochrome", then, PropertyRelation((R'(i), R"(i)) is in the range of 0-1, which is defined by the user.

*4)* Mismatching or no comparability, for example, user request="laser", matching object="inkjet", then, PropertyRelation((R'(i), R"(i))=0.

TABLE I. ALGORITHM OF GCSM-A

| Steps | Actions |
|---|---|
| (1) | Input uri1, uri2 |
| (2) | Execute the sentence of RDQL<br>Select class WHERE<br>(<uri1>, <#subClassOf>, ?class><br>(<uri2>, <#subClassOf>, ?class) |
| (3) | If H(uri1) include uri2, then, depth=depth+1 |
| (4) | Execute preorder traversal of all classes in the result of RDQL |
| (5) | Gain the class with $Depth_{max}$ : class[i], namely LCA, and its uri |
| (6) | GCSM=(depth(uri1)+depth(uri2))/(2*depth(uri)) |
| (7) | Gain property of LCA: (p1, p2, ..pn), and the corresponding uri1(p'1, p'2, ..p'n), uri2(p"1, p"2, ..p"n) |
| (8) | Calculate s(weight of property) |
| (9) | Calculate distance=GCSM+$^{F(\Delta A)}$ |
| (10) | Return distance |

## V. EXAMPLE

We have constructed a decision support CBR prototype system of marketing strategy, based on this algorithm, which contains more than 600 cases.

We compared two retrieval schemes in application. One scheme is based on ontological retrieval framework; semantic inference and extension of ontology model is created in compliance with domain semantic rules. The other is the conventional method without any reasoning processes, where retrieval is executed by characteristic words matching. The precondition of the experiment is that all resources can be exactly correlated to the relevant concepts.

As the result shown in Tab.2, with the support of semantic, we can not only carry out data matching retrieval, but also perform semantic associated data access. CBR can quickly and accurately retrieve cases and improve efficiency of reasoning by semantic query.

TABLE II. RETRIEVAL RESULT

| Retrieval Schemes / Retrieval Conceptions | Semantic Retrieval Result | Key words matching Retrieval Result |
|---|---|---|
| Channel | 95 | 79 |
| Quality | 36 | 27 |
| Customer | 127 | 105 |

## VI. SUMMARY

The ontological idea in CBR system theory and method research makes it possible to measure the comparability between cases based on the similarity of ontology in the case retrieval system. Pick out the most interrelated couple of properties as the structural comparability of cases and then compute ontological comparability of the property values. The structural and ontological comparability form the complex comparability of cases. The certificate mechanism of ontology uses the method of case retrieval by means of knowledge matching.


## 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, and the Project of Science and Technology Commission of Shanghai Municipality under Grant No. 08DZ1122300.


## REFERENCES


[1]Coyle L, Cunningham P, Hayes C, Representing Cases for CBR in XML[C], In Proceedings of 7th UKCBR Workshop, Peterhouse, Cambridge, UK:Springer Verlag, 2002, 212-220.

[2]Lorcan Coyle, Dónal Doyle, and Pádraig, Cunningham.Representing Similarity for CBR in XML [C], ECCBR 2004, LNAI 3155, pp119–127, 2004.Springer-Verlag Berlin Heidelberg 2004

[3]Gruber TR, A Translation Approach to Portable Ontology Specifications [J], Knowledge Acquisition, 1993, 199-220

[4]Brost W N, Construction of Engineering Ontology for Knowledge Sharing and Reuse, Ph D thesis, University of Twente, Enschede, 1997.

[5]Studer R, Benjamins VR, Fensel D, Knowledge Engineering Principles and Methods[J], Data and Knowledge Engineering, 1998, 25(1-2):161-19.

[6]Prasanna Ganesan.Hector Garcia-Molina.Jennifer Widom.Exploiting hierarchical domain structure to compute similarity, ACM Transactionson Information Systems, 2003, 21(1):64—93.