International Journal of Hybrid Information Technology
Vol.9, No.4 (2016), pp. 25-40
http://dx.doi.org/10.14257/ijhit.2016.9.4.03

Novel Approach of Semantic Annotation by Fuzzy Ontology based on Variable Precision Rough Set and Concept Lattice

Hongsheng Xu, Ruiling Zhang*, Chunjie Lin, Youzhong Ma

College of Information Technology, Luoyang Normal University
Henan Luoyang, 471022, China
Corresponding author,e-mail: zhangruilingls@163.com

Abstract

The paper firstly describes some typical manual and automatic semantic annotation, then analyses strategy of semantic annotation based on ontology. Secondly, this paper proposes semantic hierarchical structure of fuzzy ontology based on Variable Precision Rough Set and concept lattice in order to make up the shortage of traditional semantic annotation methods. Novel algorithm of semantic annotation based on fuzzy ontology for domain Webpage is presented in the paper by comparing ontology data of extractive webpage. Finally, experiments show that this presented algorithm is better than the traditional method in semantic annotation rate of support and recall.

Keywords*: Variable precision rough set; Concept lattice; Fuzzy ontology; Semantic annotation; Domain Webpage.*

1. Introduction

The conception of semantic Web aims to improve the existing Web; in order to overcome the present can not solve the Web information automatic processing and precise search difficulties. On the improvement of this series of measures, the most important is the introduction of semantic knowledge representation in Web, that is, the semantic Web will not only confined to a page from the content and form of expression, but the emphasis is on the increase with semantic information, so as to ensure that the Web page can be machine understanding and automatic processing to a certain extent [1]. Therefore, how to represent the semantic information for the semantic Web is very critical, need an effective semantic information model to support. The ontology is the description of the modeling means of semantic knowledge in the semantic Web, it formally defines the common in the domain of knowledge, is the core of semantics in Web system.

Two well-known annotation frameworks exist in the semantic Web: a project the Annotea annotation is supported by W3C; the other is a CREAM annotation framework. Annotea is a W3C project, it specifies the annotation schema for Web documents, and emphasize the application of collaborative tagging information. For all W3C projects, the use of open standards is to enhance the interoperability and scalability is a very important principle. The main tagging Annotea file format is RDF; the label is limited to HTML or XML document. It uses XPointer to realize the marked location information within the file.

In this paper is including the establishment of rule matching between here and named entity in the ontology concepts and instances of classes, semantic annotation. The key technology is the choice of similarity algorithm, this project uses edit distance similarity calculation combined with the method of sharing information, in order to improve the efficiency of the algorithm and the syntax and semantics of the named entities are taken into account, first on the named entity similarity based on edit distance calculation, if greater than the threshold set by the grammar angle dimension, whereas below the threshold sharing information content similarity calculation method based on dimension

ISSN: 1738-9968 IJHIT

from the semantic point of view. Semantic tagging results are stored; the project uses the non embedded semantic annotation in the form of preservation of semantic annotation results, in order to reduce the maintenance difficulty.

In the aspect of semantic annotation, supervised system mostly requires learning from user labeled sample, but collect enough training examples is a very difficult task, unsupervised system uses a multi strategy learning how to mark, but its accuracy is limited.

The key problems are: efficient use of knowledge acquisition tools; large-scale ontology merging speed, accuracy and efficiency; Deep Web tagging accuracy and speed, how too automatically on the Deep Web data annotation, in order to be able to automatically generate pages and Webpage corresponding labeled information; automatic semantic tagging accuracy, with a minimum of user intervention to maximize.

This paper presents semantic annotation based on fuzzy rough concept lattice, and it is with as the classification model, first extract the document, to form lexical properties represent the document, and then form the formal context, the form background lattice construction operation using the algorithm of concept lattice structure improvement, the formation of concept lattice, using concept lattice specific classification effect the classification of the article, and then you can learn both feature annotation information itself distribution using the classification model, also can study the distribution characteristics of the annotation information and context boundary (usually including: start and end boundary), so as to realize the automatic recognition of the marking information. This paper also proposes a best in the field Webpage semantic annotation ontology learning method based on mining semantic information in social tagging implication, and put forward a kind of latent consumption hierarchy to describe the structure underlying the label space, and based on this model is derived from the ontology learning algorithm. Through the experiment is to verify the effectiveness of the method.

2. Strategy Analysis of Semantic Annotation based Ontology

Automatic semantic annotation based on a specific algorithm, let the automatic annotation process of manual annotation tool, check the tagging results. Its goal is based on the training set (to be marked Webpage and given ontology) automatically tagging process. The automatic semantic annotation system tagging accuracy generally is not high, tagging granularity are relatively coarse, automatic semantic annotation by many related technical constraints in the actual operation process, really difficult to achieve.

From the point of view of ontology, knowledge representation can be seen as formed by logic, ontology and the calculation of the combination of three parts. Logic provides a new interpretation of a description logic functions from existing knowledge through logic operation; computing refers to the process of determining whether a description can describe the deduction from the given [2]. The ontology is related to the knowledge sharing concept model and explicit formal specification; it provides the basic terminology knowledge model (knowledge atom) and relations, and the use of these terms and the relationship between the extension rules and the complex definition form. Ontology is a new form of knowledge representation, it can be the domain knowledge expressed as knowledge discovery algorithm can understand form. Knowledge modeling ontology into knowledge space, and it is the construction of domain knowledge ontology.

2.1. Manual and Automatic Semantic Annotation

Research on the traditional semantic search of previously defined a unified body, various defects of the reuse of the ontology annotation relationship among all kinds of resources and resource network exists, using "semantic ontology" train of thought, research on social tagging data users, the relationship between the cyber source and label three, the establishment of a no statistics model to guide learning, dig out the social

annotation semantic information, solve social tagging of polysemous words and multi word meaning problem, the explicit semantic information in social tagging.

Definition 1: class of C 1 expression in the definition of ontology for the predicate form C (P1, P2,..., Pn), where P1, P2,..., N data type attribute Pn is a class C (such as DAMI, +OIL daml: DatatypeProperty). The semantic rules are a query can use the following representation:

$$Q(X): \{C_1(X_1), \ C_2(X_2) \}, C_n(X_n) \quad (1)$$

The Q (X) called the Datalog rule head, C1 (X1), C2 (X2)}... (Xn, Cn) is called the rule body, Ci (Xi) (I = 1 ~ n) is called the sub goals, X is the attribute vector, to query Xl, X2,..., Xn is C1, C2,..., attribute vector in Cn, where Ci(Xi) ($1 \leq i \leq n$) can be viewed as a relation in the relational model.

Justice is never really asserted, is defined in the "concept" and "property" of the rules. In the field of application of a specific, have its specific meaning of each concept, relationship, functions, axioms, examples, there are also some inherent relationship between elements and constraints, axioms can be used to describe and explain the relevance and constraints between the elements and it.

Definition 2: let Q be a finite set of domain knowledge of the problem, K is spatial knowledge on field Q. For any state of knowledge of K in K, a collection of N knowledge state (K) ={k', =k' = K, d = 1} (k, k') is called the knowledge state according to another; which, between the knowledge state of K and K' distance D (k, K) are equal to the size of the symmetric difference set, D (k, k') = kOk' =I (k\k}U (k1k) I. The set F (k) =[UN (k) \nN (k)] is called the state of knowledge of the convenience of K, namely the state of knowledge of K and its neighbor set different subject.

User friendly manual is annotation tool case. The system is designed to encourage users to use some produced in semantic services such as department mail list or schedule such data in the labeling of their HTML file [3]. The annotation tool with an intuitive graphical user interface, the user can choose the number of tag from the tag set in the interface, and put these labels associated to highlight text on them. At present, the Mangrove system has been integrated into a semantic mail service center. This semantic email services support the semantic message preliminary treatment, such as through the text form of meeting scheduling.

Memory user operation flow (similar to user defined rules), and it is the realization of semantic annotation. It is mainly in the XML and HTML document annotation. AeroDAML uses information extraction module AeroText for semantic annotation. AeroText relationship between proper nouns recognition is by predefined rules in the document and the noun, which is then mapped to corresponding concepts and properties in the ontology on. AeroDAML only supports the ontology representation language DARPA Agent Markup Language (DAML).

Property: provides a mechanism for describing the relationship between the abstract class, can be two yuan relationship as between classes, attributes and the relationship between sub attributes can be described by rdfs:subPropertyOf will form a hierarchical relationship attribute. OWL defines two kinds of attributes: object attribute and value type attribute, object attribute that represents individuals to two yuan in the relationship between individuals, value type attribute that represents individuals to two yuan relationship among data values.

The COHSE labeling machine generated is compatible with Annotea standard, the annotation information in links stored in the distributed link service (Distributed Links service), but the current system only realize a vocabulary to highlight those exists in text ontology term of service matching [4]. This tagger is provided in the form of plug-ins, supports both Mozilla and Internet Explorer. The user can freely choose their own like

browser environment. COHSE has been applied to many fields of applications, including semantic for visually impaired users label generation and provides modified Java tutorial site.

Definition 3: C is a class, if a member of the C fixed, then C is called closed classes (CloseClass), otherwise the Chen C for classification (OpenClass).

$$C_1 \coloneqq (A_1, B_1) \in \underline{B}(K_1), \quad C_2 \coloneqq (A_2, B_2) \in \underline{B}(K_2) \tag{2}$$

Definition 4: (super class and sub class) for any kind of C1 and C2, if i:i (isInstanceOfx) C1, I (isInstanceOf) C2, then C1 is called a subclass of C2, denoted as C1 (isSubclassOf) C2. Accordingly, C2 C1 called the super class, denoted as C2 (isSuperclassOf) C1.

$$f(x) \supseteq D_1 \Rightarrow f(x) \supseteq D_2$$

$$K = (U_1 \cap U_2, A_1 \cap A_2, I) \tag{3}$$

Among them, X C represents a class of domain ontology; the base Croot class represents the definition of domain ontology, representing the top class in the field. If Pi is an instance of class X or x object, PI associated with X. According to the gamma (PI) mapping between Webs pages to the domain ontology definition, classification can be achieved Web page.

The semantic Web service Web Service markup language OWL-S annotation framework based on WSAnnotator, it can support atomic Web Service and semantic annotation simple compound Web Service by atomic service through the "order" execution mode is formed by the combination of it.

Definition 5: (ontology classes closure). Ontology in O class, C class closure is the set of all C C and O subclasses, denoted C+. Made for: a query: Q C} (Xl), C2 (X2),... , Co (Xn) (C (1; sisn) belongs to O). The query should be in accordance with the expanded form of solving the following:

$$Q(X): C_1(X_1)^+, C_2(X_2)^+, \cdots, C_n(X_n)^+ \tag{4}$$

Note: for a query Q the user, wherein the rule body using the middle layer in the ontology class representation.

Definition 6: (member relationship, 6asMember). Member relation refers to the overall collection contains a specific sub element, said members between the element and the whole relationship. The parent concept and sub concept are membership. For example, A is a member of the group of team, is expressed as a teamhasMemberA.

$$\hat{u} = \left(\hat{a}, \hat{b}\right)^T = \left(B^T B\right)^{-1} B^T Y_1 \tag{5}$$

In the rule learning before, this paper uses GATE to preprocess the text; then the context information annotation information using Amilcare (including a variety of natural language information after pretreatment) to establish the initial rule set; then using a coverage algorithm for rule induction, the induction process is iterative in a bottom-up process, each iteration from the initial rules focus on two randomly selected rules were summarized and concluded the rules to rejoin the rule set, and then start the next iteration is applied to multiple semantic tagging system.

2.2. Ontology based Semantic Annotation Method

In the automatic annotation of this group, we also consider the automation tool of two types: one kind is the tool to the user (including knowledge workers) to provide automation component annotation information advice, but still user intervention. Another

kind is to automatically obtain the marking information of large-scale system. These tools some can only be used by some experts, while others are suitable for knowledge workers. Automation system is designed to support the relevant knowledge workers need to take into account the user interface design problems, in order to accuracy with minimal user intervention to maximize.

The realization of semantic annotation tools, an instance of it using the SHOE Ontology (Instance) and (Claim) a statement to mark a Web page, you can open multiple ontologies to mark a document, suitable for manual annotation occasions. For the MS Internet Explorer browser plugin for Markup (Plug-In) and SemanticWord was developed in the TeKnowledge project, which is characterized by the function of marking in the form of plug-in added has tools support semantic annotation; this thought is helpful to extend semantic annotation function to the existing text or web page editor.

Definition 7 [5]: study ability of the model is defined as two tuple (KE, LS), where KE is the user master the knowledge module, LS users to master the learning skills. The user knowledge space (UKS:userknowledgespace) is defined as three tuple (UI, C, P), where UI is the basic user information, C for the user's learning ability, the learning performance for P users.

$$w_{i+1}^{1}\left(t+1\right)=\left(1-wd_{i}^{1}\left(t\right)\right)x_{i}^{1}\left(t\right)-rs_{i}\alpha N^{1}\left(t\right) \tag{6}$$

Definition 8: (relationship inconsistency). If the body view P1 and P2, and it is the concept of C1 and C2, asymmetric relationship R. If the containing type Pl and P2 makes the following is false, we say that the existence of the relationship between the inconsistency between Pl and P2.

$$\begin{aligned}&(\exists c_1 \in P_1 \cdot C)(\exists c_2 \in P_1 \cdot C)(((\mathrm{r}, c_1, c_2) \in P_1 \cdot R) \wedge (c_1 \in P_2 \cdot \mathrm{C}) \wedge (c_2 \in P_2 \cdot \mathrm{C}) \rightarrow \\ &((\mathrm{r}_1, c_1, c_2) \in P_2 \cdot \mathrm{R}) \wedge (\mathrm{r} \neq \mathrm{r}_1))\end{aligned} \tag{7}$$

PowerPoint environment based on semantic annotation tools, it will mark stored in PowerPoint document, but the label is not visible. OntoMat-Annotize is a user friendly Web page annotation tool, the tool is generating annotated Web pages, the content for the semantic Web Agents reasoning, it allows users to create and maintain DAML+OIL markers based on ontology, which is user oriented individual tagging, namely the user using DAML metadata to enrich his Webpage.

Web content data Web content data is presented to the user's Web page content. Web content data consists of text, image, audio, video, metadata and hyperlinks and other data types. The main purpose of Web content mining is to improve the information search and filtering. Web data structure data structure includes two parts: (L) the page structure, mainly including Web page within the HTMLScript or XML code; (2) link structure, namely the link between actual Web pages.

This paper finally proposes a learning semantic annotation method based on semantic information in ontology mining social annotations in the implication, put forward a kind of latent subsumption hierarchy to describe the structure underlying the label space, and based on this model is derived from the ontology learning algorithm.

In order to solve this problem, in the absence of supervised, unsupervised systems using a variety of strategies to learn how to dimension, but its accuracy is still limited. Another method is the application of natural language processing technology, through the analysis of word, sentence to complete the tagging task. In addition, according to whether the use of ontology information extraction process, also divides the information extraction technology using mode is the traditional information extraction method based on ontology and information extraction method.

Based on the rules and not marking method of classification model representation based on annotated dependency relationships between information, which makes them in the strong dependence of the annotation application effects are usually not very good.

Sequence model using annotated dependency relationships between information, the input sequence unified mark [6]. In the label text, sequence can be sequences of all lines of text in a document can also be a word sequence tagging task, is the annotation results for all unit are given in this sequence a globally optimal.

Rule 1: (vertical classification relationship without circulation principle) in the body, vertical classification of all relations cannot exist cycle. Let R be a body in O Hc two yuan between, does not exist the class Cl, C2, C3, meet:

$$R(C_1, C_2) \wedge R(C_1, C_3) \wedge R(C_2, C_3)$$
$$r_3(r_2'(c_1, P_1), = r_2'(c_2, P_2)) \leftrightarrow r_1(c_1, c_2) \quad (8)$$

Definition 9: If there is a set of concepts in the ontology in each view, wherein the body view belongs to each concept in the concept of connected relationships and concepts are the same, we call this set of concepts for homomorphisms. For homomorphisms, if their attributes set intersection is not empty, then we hope to use a common conceptual representation in the middle layer of body, at the same time the public concept and homomorphisms to maintain the relationship between father and son.

This ontology semantic annotation for automatic annotation tools using OWL ontology. The system comprises a client server (C/S Architecture) version and a Web based on the demo version. The user can input a URL in the Web demo, the system will automatically on another page returns an annotation information file. If you want to see the annotation results in Webpage context, users need to store the RDF files to label information on the server, and then use a good annotation information browser can display.

Finally, on the analysis of the relationship between compositions of natural statement dependent results are semantic annotation. It first uses ontology dependency parser based on natural statement of Web page analysis, get the dependency relations between words in a sentence; and then the different words are mapped to the corresponding ontology concepts, and then between words by relational mapping relations between the concepts of the ontology, the machine learning method to the realization of dependency to the mapping between RDF three tuple.

3. Analysis of Semantic Hierarchical Structure on Fuzzy Ontology based on Variable Precision Rough Set and Concept Lattice

Based on the construction of domain ontology, using product ontology meta model for UNSPSC core ontology, the use of unstructured and structured data (Electronic Commerce log, tables, a relational database) combined with experts in the field of e-commerce knowledge, through the method of core ontology semi automatically extended extraction and establish the product ontology model.

In the construction of ontology research field, although there have been a series of ontology construction methodology and development tools, but ontology building automation still is a difficult technical problem in the field of ontology research. Although the construction of ontology is inseparable from the human expert's intervention, but many methods used in pattern recognition, artificial intelligence and statistics technology has been used to assist ontology construction, these methods can greatly reduce the workload of ontology developers, construction of semantic integrity ontology assisted ontology developers.

Ontology is to describe the concepts in a domain. On a shared property description called a concept definition. Concepts are organized into a superclass subclass relationship through a classification hierarchy, similar to the information retrieval in the classification hierarchy and object-oriented classes in an inheritance hierarchy. A class of objects and another class objects often have a certain relationship, the relationship between objects to illustrate the structure [7]. It is not enough only with words of ontology to describe, but there is no formal definition of ontology unified, here is now generally accepted ontology seven tuple definition: O= (C, AC, R, AR, H, I, X), where C is the set of concepts, AC is attribute set; R is a collection of relationship; AR is a collection of attribute relationship; H said the concept hierarchy; I is an instance of the axiom set is set X.

This paper first extract the type annotation attribute types in the list of vocabulary words as the feature set, and then use the extraction algorithm of the whole sentence space cut features of sentence similarity based on strength formula, and utilize the feature sequence structure and feature generation algorithm sentence set features corresponding to the sequence database.

Here take method is: the probability and statistics method to obtain the key concept words can represent the text; in the conceptual vocabulary found form, vocabulary, file background, and then use the model of concept lattice generation unit body above: the use of the reduction of the form background variable precision rough set to reduce the redundant object, noise reduction. For after the reduction of the formal context, and it is formal concept analysis technique using tectonic unit body. Finally, through the experimental analysis of the variable precision rough set theory and formal concept analysis and combined with fuzzy set theory (Fuzzy) mapping method based on domain ontology.

From the source ontology and the target ontology semantic related entity semantic annotation Ontology (i.e. classes, relations, attributes), encapsulating all the necessary information from the source ontology, deformation case an entity instance to target ontology entities. The results are close to the ONION in the ontology definition, the following formula.

$$LF(I_i) = \frac{F_3(I_i)}{\sum_{j=1}^{n} F_3(I_j)} \tag{9}$$

Using the top-down development approach prototype iteration, emphasizing the first determine the core concepts and relationships at the top in domain ontology, constitute an available fuzzy ontology prototype, and then in accordance with the prototype theory of iteration, expansion and perfection of the body gradually. The core process in the above model in an iterative process of analysis, and it is from the prototype iterative idea. In the fuzzy ontology model design, process of iterative analysis for concrete:

Step1: Domain ontology building top core class hierarchy and core class attribute;

Step2: Here the concept lattice model according to the refinement of the extracted attribute information core of a class, a hierarchical structure of classes to improve;

Step3: Rough set an example to verify the core classes, class hierarchy and attribute;

Step4: Deductive (Reasoning) can be in the structure of knowledge in the ontology, relying on the interpretation of the way to obtain more useful information;

Step5: Object oriented (Object-oriented) this attribute is most often with the concept in the ontology, particularly the application in information processing field.

All the concepts in object sets form a concept lattice (concept lattices). As the data structure of FCA core, the concept lattice is essentially describes the relationship between objects and attributes, show that the generalization and specialization relations between concepts, and concept extraction process in essence has become a concept lattice

constructing process [8]. Structural efficiency concept extraction depends directly on the concept lattice.

Definition 10: (formal concepts). A formal concept J is an ordered pair (A, B), where A G, B M, and A'=B, B'=A. The A is called J (formal concept extension (extent), B is called the connotation of form concept of J ((intent). In a formal context, the connotation and extension are equivalent, may be used to express a concept.

To build a lot of ontology are established on the basis of a specific field, list this field relates to the entry (terms), so it is easy to acquire knowledge and body function description; summarized and modified in accordance with the inherent property of entries and exclusive characteristics of entry, establishment of class (class) and the level of classification model the (taxonomy), the method used here is to join the relationship (relation) of terms and taxonomies connection.

The RDF data model does not depend on XML, but based on XML syntax. RDF Schema based on the RDF, provides primitives to implement the hierarchical organization of the Web object. Key primitives include classes, properties etc.. In the semantic Web, RDF and RDFs are often collectively known as the data layer for the semantic Web.

Definition 11: approximate about called $RED^{\beta}(C,D)$, the definition of $\beta \in (0.5,1]$ is: for a given value of ontology, satisfy the following two conditions:.

(1) $\gamma^{\beta}(C,D) = \gamma^{\beta}(RED^{\beta}(C,D),D)$

(2) Remove any one attribute from $RED^{\beta}(C,D)$, (1) will make the untenable.

Definition 12: let (A profile) is a lattice, if one defines two operation of two yuan and on A, such that for any a, b∈A, a∨b equal to the minimum upper bound of a and B, a is equal to a∧b the greatest lower bound of B, a and B so called (A, ∨, ∧) by lattice (A, ≤) by algebraic system guide. Operation of two yuan and respectively is called Union and intersection operations.

Usually we use a B instead of sup ({a, b}), a B instead of inf ({a, b}). Similar to B and B were used to instead of sup (B) and inf (B).

The core idea of fuzzy ontology: firstly, based on the similarity analysis (revealed in similarity stage), the first step is to select pairs of entities; they may be concepts, relations and attributes, to set up the corresponding relationship between them using the concept of bridge. The relationship between MAFRA allows different cardinality relationships between source entity and target between the entities of the. Therefore, source ontology or the target ontology can belong to one or more semantic bridge.

The second example of determining the core classes, to verify the class hierarchy and attribute; once again determine the limit condition attributes, function relation is established between the objects; iterative development finally proceed detailed, complete contact established between class [9]. In addition, because of the existence of the domain ontology classes are not isolated, a variety of semantic relationships exist between classes, the following formula:

$$\frac{|f(M)-f(x)|}{M-x} = |f'(\xi)| < \frac{\varepsilon}{2} \tag{10}$$

Definition 13: (class between incomplete) for an arbitrary ontology classes in O C, and C non root class, if the C does not exist in the super class, it prompts the kind of relationship is not complete. (Relationship instances incomplete) on Ontology in O individual I, if there is no C, meet I (isInstanceOf) C, it prompts the instance relation is not complete.

Definition 14: (K concept and K attribute set). For formal concept (A, B), if |B|=k, we call B for a K item concept connotation. For any X M and |X|=k, we call X the K attribute set.

An integrated knowledge management platform, support, knowledge management, semantic annotation and semantic retrieval etc.. The semantic annotation module using GATE implementation, GATE is integrated with the Natural Language Processing module series, such as word segmentation, POS tagging, named entity recognition, rules matcher and refer to recognition etc.. KIM was extended to realize the automatic semantic annotation on news data through the GATE existing rules.

In the context of the semantic Web, ontology has undertaken the underlying machine processable semantics defined task. One also need to note is, the above four kinds of semantics can be used for the machine, the difference is: for the first three semantic, need to understand the semantics based on, by hard coding manner, will be its ability to write programs to give the machine understand these semantics and machine processable semantics; it can be handled automatically by the machine.

Definition 15: That $S = (U, A)$ $P \subseteq A$, $Y = \{Y_1, Y_2, ..., Y_n\}$ is a U classification or partitioning, this classification is independent of knowledge of P. A subset of is a classification of Y. Division of Y S on the P of the lower and upper approximation are defined as $\underline{PY} = \{\underline{PY}_1, \underline{PY}_2, ..., \underline{PY}_n\}$ and $\overline{PY} = \{\overline{PY}_1, \overline{PY}_2, ..., \overline{PY}_n\}$.

Integrated the above semantic tagging and two aspects, we can know: the broad sense, all annotations are endowed with semantics can be regarded as a kind of semantic annotation; the narrow sense, semantic annotation should be machine processable, the label should be explicit and formal.

4. Research on the Semantic Annotation of Fuzzy Ontology for Domain Webpage

This paper studies and analyzes the user session, said that define the semantics of user preference and discovery algorithm. The model of the domain ontology integration into the Web mining and personalized are recommendation process, so that can make full use of semantic knowledge in the preprocessing, pattern discovery and online recommendation phase [10]. Then it puts forward recommendation algorithm based on semantic clustering personalized domain ontology: first use of Web data preprocessing, and the pre processed transactions by cluster analysis. To characterize each cluster to generate user use each cluster vector access preferences, and then compare the user similarity clustering to generate user preferences for the current session and Web access preference.

Many methods and tools for matching and mapping, also need the source ontology in the form of expression is shown to determine. Although, from the expression of any body language to the specific expression of the translation will symbolically exist, but whether the semantic preserved but not guaranteed. Matching now review the above language level of the error, and describe the existing in the existing system and the technology conditions: (1) since the body both use the same syntax, then the grammar will pass any different according to the internal representation of the compiler has been solved. (2) in the grammar in different situations, differences between the logical expression occurs, but the presentation logic equivalent will be used to express the same thing. An example of this is "disjoint (disjointness)" in the expression of OWL Lite in style, and expressed in OWL DL in style.

Modeling of OntoAES ontology construction method is applicable to the field of education knowledge ontology, the system specification will each discipline domain knowledge with the ontology modeling method and the framework of organization, the

various existing learning resources with domain knowledge and reasonable link of expansion, higher level application and in this grave foundation. The method also can be extended to other areas of the building. For example: 1) the user Ontology: a variety of property is mainly described the use of various learning resources subject object; 2) provider of educational services, ontology, introduces the related attributes provide various education service development party.

RDF is a semantic data model for two yuan, the relationship between the expressions of Web resources. For any complex relations after decomposition and it can be used in a plurality of simple two element relation to express. Therefore, any complicated relation model can be RDF data model as the basis. The basic concept of RDF including resources (Resource) and it is property (Property) and presentation (Statement). In RDF, all need to describe objects or things are called resource [11]. Resources can represent an author, a book, people, places, or in Webpage Webpage fragment. Each resource with a uniform resource identifier (Universal Resource Identifier, abbreviated as URL) to identify.

Tagging object for all tools to static content, such as: Email, Image, HTML page, Word, Powerpoint and PlainText, but Web contains a large number of multimedia objects, the e-commerce application service data and other dynamic content, at present, only a handful of annotation tools to support the content writing and semantic annotation synchronized manner, the vast majority of annotation tool used to create content, after the first dimension way. Semantic annotation automation aided reasoning support and metadata ontology query, process is not enough, these deficiencies affect the enthusiasm and the possibility of a large number of Web users in general use tools to create semantic Web content.

The importance of ontology evolution has been fully recognized, but the current ontology edit tools for ontology evolution support in the reason of evolution on ontology information retention and treatment has not been. That paper compared some typical ontology editing tools, evaluate them for ontology evolution support; and then, in order to strengthen the treatment of historical information, mining evolutionary reasons, proposes an ontology versioning and logging technology combined with the evolutionary framework, and in which the current OWL language extension for the support of ontology evolution, retention and analysis processing evolution causes information to effectively support, as is shown by equation(11).

$$E = \sum_{j=1}^{k} \sum_{p \in Cj} \left| p - M_j \right|$$

$$L(K_1) \cap L(K_2) = L(K) \tag{11}$$

If AC ={a1C, a2C,... , anC} represents a collection of C class attribute defined in ontology, if object P with attribute Ap ={a1p, a2p,... Anp} BP, or the presence of attributes in Ap, BP AC, but BP is aiC (I = 1 ~ n an equivalence attribute, then the P object is an instance of the C class. In the semantic Web environment, each user cluster is interested in the Web page might be distribution corresponding to different classes in ontology in the field. According to the content of the Web page, a formal definition of class and property and the individual body, can be directly put up a Web page to the mapping relationship between the ontology classes.

The concept can be used is a relationship between organization in a hierarchical fashion, can also be non hierarchical (related links) between. So the CS-Domain domain ontology can be used to describe the SKOS. Each concept in CS-Domain is represented as an instance of the skos:Concept class, the conceptual framework of the field described by skos:ConceptScheme. Help the concept hierarchy we define cs-domain in various semantic attribute spread from the skosaemanticRelation property in the SKOS.

The Web form is currently a Web page using data storage, more extensive presentation format [12]. The Web table data is usually through the backstage script generation or from the database directly extracted, so it's very valuable data. In Web form usually use table

column name to locate the concept of ontology, but for Chinese Web form, the expression of the concept of ontology is the name of a column is usually a Chinese phrases, and because of the complexity of Chinese information, Chinese phrase to express the same concept in ontology may vary in different Web table..

OWL with DAML and OIL as a starting point to develop and come is a revision of DAML+OIL is the latest achievements of semantic Web research field on the ontology language for semantic Web applications. The syntax of OWL and describe the structure similar to RDFs, but increased the amount of primitive description logic based on semantics to describe and construct a variety of body. The introduction of description logic ensures that the OWL have the ability to reason, as is shown by figure1.

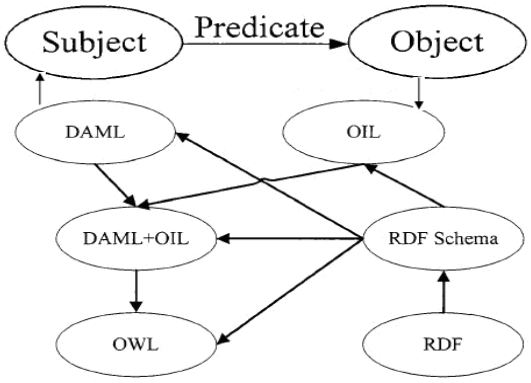

Figure 1. The Relation between the Label of the Semantic OWL and PDF

Given data sequence of random variables (X1, X2, x,... , Xn), C defines the annotation results of sequence of random variables (Yl, Y2, Y,... The conditional probability distribution), and it is Yn P (Y|X), through training methods to make the conditional probability P (Y|X) maximum. It is a description of such an undirected graph structure of G= (V, E), where V is the node set, E is the edge set, X={Xv|v V} says all node values in G, Y={Yv| V in V} said the annotation results on X.

A domain ontology to describe the set of classes for C={x1, x2,... , xn}, semantic user profile can be expressed as spri={< x1, mi1>, X2, mi2>,... A, B: XK, mik>}. Among them, m ij expressed class XJ in spri weight, M I in [0,1]. According to the user cluster access preference correlation matrix M, mapping and the relationship between domain ontology and Web pages, access preference relation matrix can be established corresponding to the semantic layer.

$$M_{t\times k}'=\begin{bmatrix} m_{11} & m_{12} & \ldots & m_{1k} \\ m_{21} & m_{22} & \ldots & m_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ m_{t1} & m_{t2} & \ldots & m_{tk} \end{bmatrix} \quad (12)$$

A Web Services semantic annotation method new to Web, the existing Services' description language (WSDL) document based, full use of similarity in WSDL XML schema semantic information and ontology entities, the extraction of semantic information from WSDL file, and semantic annotation by name similarity and structure similarity between entities, generation Web Service semantic description based on OWL-S.

User session can be expressed as As=<uldA, {P, hits, time}n>, the representation for the user to the Web page preference representation of prAs={<p1, wil>, P2, wil2>,... , < PS, wis>}, and convert it into a semantic preference form SprAs={< gamma (P1), mil >,

< gamma (P2), mi2 >,... , < known (PK), mik >}. The user session As, and cosine similarity coefficient between Bs expressed as the following formula.

$$\lambda p_1 = 2up_2, p_2 = \left(\frac{\lambda}{u}\right)^2 \frac{p_0}{2!} = \frac{\rho^2}{2!} p_0 \tag{13}$$

At these sites, some of the content is in the form of "providing information directly, and some of the content is in the Web database (Deep Web), need to submit a query term in a query form, by the background program from the database retrieval results page. Web extraction ontology semantic annotation algorithm steps are as follows:

Step1: get the need for extracting target Webpage;

Step2: information extraction of target Webpage extraction rules in application;

Step3: Based on the body structure of Hj concept lattice derived Oi, Oi ontology so as to get the B's (Oi). For the two to be merged ontology in the system of sub ontology mapping method based on the attribute, the joint distribution of multi strategy presented in this project;

Step4: the instance data is merged into the repository;

Step5: Webpage pretreatment in duplicate removal, denoising, Webpage classifier selection problem, in order to achieve high spatial efficiency and error rate two better balance, the project intends to use the block hash function method for URLs duplicate removal, SVM classification method to construct classifiers using multi classification;

Step6: establishes the mapping between Web pages and the ontology class;

Step7: is used in this paper to put forward a method for calculating the ontology based on semantic tagging, similarity computation between ontology concepts are attributes, attribute mapping unit body set, then according to the fields with similar unit body relation;

Step8: through the network packet capture tools Wire Shark interception of network data packets and then submitted by AJAX, simulation of AJAX code in the extraction procedure to generate a query request. After extraction the generated RDF fragments;

Step9: in the sequence labeling task, given unlabeled data sequence X (X1, X2,... The annotation results, Xn), Y (Y1, Y2 sequence data,... Yn), tree like conditional random field model is the purpose of the training is to learn the most suitable parameters by a given training corpus, makes the following certain rules tagging P (Y|X) value of the largest, by contrast we can get the final result of semantic annotation;

Step10: will be described by OWL ontology into RDF three tuples, according to three different tuple predicate concept classes and instances, attributes and relations in a database.

Step11: training required memory space is larger, longer running time. A method of kernel function with distance performance based on reduction, through the kernel function selection, denoising, reduce, and eliminate noise reduction steps, the final completion of the task of semantic annotation.

There is a need to explain, on the Web content of the extract by submitting a query word needs to obtain the target Webpage, this step into the Ontology library.

5. Experiments and Analysis

Collaborative filtering recommendation algorithm based on domain ontology and user preferences based on change, using the semantic similarity measure to calculate the semantic similarity between items, and then according to user rating project predict the Unrated item ratings, filled user item rating matrix in a, after pretreatment of the user item rating matrix for processing.

Sequence S: S is an ordered sequence of symbols of a sequence set, denoted as S[0..N], where S[N] is stored in the end symbol sequences (sequence of symbols is less than the Arts Office marks the end of a symbol, a part, as expressed sequence of the following

contents mentioned in the sequence includes the end marker, no longer one one instructions) [13]. S[i] represents the position of I in S symbols, 0 i<N. The length of S is N, denoted by |S|=N. The sequence of S[i... j] called S sequence, which 0 i< j< N.

Set this option to C, C1, C2,... , Cn as ontology classes in O, C1, Cn (isSubclassOf) C, for C1, C2,... And any instances of Cn, is an instance of C, then {C1 is called,... C, Cn} are disjoint decomposition.

Theorem 1: C, C1, C2,... , Cn as ontology classes in O, C1, Cn (isSubclassOf) C, if I, I (isInstance0f) C, there exists a unique J, 1<j<n, such that I (isInstanceOf) Cj, then {C1 is called,... , Cn} is a partition of C (Partrition) or a classification (Classification). This complete disjoint division called complete sub class decomposition.

Given the definition of knowledge base, the provisions of concept of value recorded as: AC={(I, c)}, I in LC (c) I * CKB; the relationship assignment denoted as: AR={(I, R, J), (I, J) in lR (R)}; attribute assignment denoted as AA={(i, a, V), (i, V) in IA (a) I * AKB * STRING. On the contrary, at a given AC, and the R and AA, lC can be obtained by the lA, lR and. By instantiating concepts and relations, can be extended body of concepts and relations, defined as follows.

$$\xi_{ij}(k)=\left[1+\left|\frac{\Delta x_i(k)}{\sigma_i}-\frac{\Delta x_j(k)}{\sigma_j}\right|\right]^{-1} \tag{14}$$

We take in the ontology concept lattice classes as the experimental object, about agricultural information page selection in e-commerce website WRBT as experimental data and Chinese lexical analysis system ICTCLAS as each attribute segmentation tool to run the EPFDS algorithm for generating Emp_ info dialect mode. The experiment for each attribute, we first select 100 samples, support and then each added 10 new samples to calculate dialect and dialect mode mode. The experimental programming environment for the Microsoft VS.NET, about the property of Position experimental results are shown in figure 2. The experimental hardware environment: Pentium (R) 4CPU 2.8oGHz, 1GB memory; experimental software environment: operating system by Windows XP, the algorithm using Java language, running in the JDKI.5 environment, development platform using IBM's open source software Eclipse 3.1, the server uses Bea company's WebLogie10, the database is SQL Server 2005.

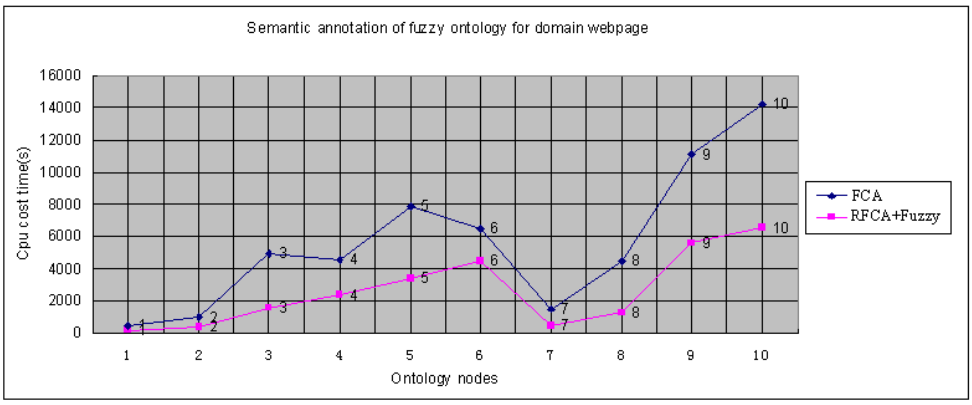

Figure 2. Experimental Results

Focus on two groups of experimental data, the RFCA+Fuzzy method of the Precision average value is the highest, but FCA is 92.6%, much higher than the other two methods, attributes this mainly due to the variable precision rough set reduction effect of generated rules. Because the rule is based on the manual mode of writing, and can generate implicit in the instance model, the semantic relations among the features included fully embody

the mode. The FCA method of the average value of Recall in the three methods is highest, for 95.39%, but considering the mining process characteristic sentence in the document in the sentence set accounts for the proportion and the mode in the feature sequence database minimum support (respectively 2.56 and 3.88), the FCA ontology method to obtain good support recall rate is satisfactory.

According to the Chinese Web form, the method can automatically generate the regular expression used to position the concept of ontology based on concept lattice (dialect mode), and provides the fuzzy ontology to solve the strategy to ensure the precision of information extraction. The experimental results show that this method has a certain practical value. User preferences obtained their FCA and VPRS arranged in descending order. From this two curve can be seen, tagging algorithm proposed ontology concept lattice based hierarchical agglomerative clustering algorithm the acquired semantic annotation quality is superior to the traditional. The test corpus which contains many verb predicate sentences and it is but also contains a large number of noun predicate sentences, close to the real language environment. The experiment in order to obtain a good labeling effect, selection in the manual syntax analysis tree, and it is statistical learning method using the maximum entropy model training speed. The results of the tests taken CoNNL2, 3 evaluation indicators evaluation procedures commonly used Eval08.pl 8 offers, namely accuracy rate (P), the recall rate (R) and comprehensive index F value (Fl) to evaluate the results of semantic role labeling.

6. Summary

Domain ontology based semantic annotation is the process of adding in the domain ontology under the guidance of the standardization of knowledge representation for the document based on that definition, in a written by OWL language ontology, text knowledge document using RDF language to describe them.

This paper introduces the ontology based semantic annotation theory; as the theoretical basis of semantic annotation, ontology plays an important role in the semantic tagging. Ontology semantic annotation is defined as the use of existing ontology based on Webpage insertion mark, or through the annotation document media have a knowledge base, which is a concept of the keywords and ontologies in the document in the process of mapping.

On the existing semantic annotation method, annotation tools were analyzed and summarized, and compares their advantages and disadvantages. Because of the existence of billions of Web pages on the Internet, fully manual annotation method has no practical significance. Finally this paper presents intelligent semantic annotation method of fuzzy ontology based on rough concept lattice. For tagging method proposed in this paper, just do some tentative experiment, the results from the practical application is still some distance. Future work needs to be carried out the following research: how to realize multi ontology semantic annotation. At present, semantic annotation are single domain ontology based on the Web data, but may involve a number of areas, a single ontology annotation objects can not meet all the. Simple to solve the cross domain semantic annotation is to the realization of ontology integration and expansion of costly. Under guidance of the ontology annotation of query form will enable data integration easier, so that each Web database interoperability between improved.

Acknowledgement

This paper is supported by the National Natural Science Funds of China (61272015), and also is supported by the science and technology research major project of Henan province Education Department (13B520155) and Henan Province basic and frontier technology research project (142300410303).

References

[1] H. Hassanzadeh and M. R. Keyvanpour, "A Machine Learning Based Analytical Frame for Semantic Annotation Requirements", International journal of Web & Semantic Technology, vol. 2, no. 2, **(2011)**, pp. 27-38.
[2] Shamsfard M and Barforoush AA, "Learning ontologies from natural language texts", Int'l Journal Human-Computer Studies, vol.60, no.l, **(2004)**, pp. 17-63.
[3] Milos Kudelka,Vaclav Snasel etal , "Semantic annotation of web pages using web patterns", Lecture Notes in Computer Science, vol. 5, no.39,**(2009)** , pp. 280-291.
[4] Handschuh S, Volz R and Staab S, "Annotation for the deep Web", Intelligent Systems IEEE, vol. 18, no.5, **(2005)**, pp.42-48.
[5] Heflin J and Hendler J, "A portrait of the Semantic Web in action", Intelligent Systems IEEE,vol.16, no.2, **(2005)**, pp.54-59.
[6] Kalyanpur A, Hendler J, Parsia B,et al, "SMORE-semantic markup, ontology, and RDF", editor: Maryland University,**(2006).**
[7] X.-p. Kang, D.-y. Li and S.-g. Wang, "Rough set model based on formal concept analysis", Information Sciences, vol. 222, **(2013)**, pp.611-625.
[8] H.-s. Xu, X.-j. Shen and Z.-t. Liu, "Construction and Presentation of Ontology on Semantic Web Based on Formal Concept", Journal of Computer science, vol.34, no.2, **(2007)**, pp.171-174.
[9] L. Wei, J.-j. Qi and W.-x. Zhang, "Attribute Reduction on Concept Lattice for Decision Formal Contexts", Science in China Ser.E Information Sciences, vol.38, no.2, **(2008)** ,pp. 195-208.
[10] J. Li, J.-Y. Song and H. Zhong, "Ontology-Based query division and reformulation for heterogeneous information integration", Journal of Software, vol.18, no.10, **(2007)**, pp.2495−2506.
[11] A.A. Estaji, M.R. Hooshmandasl and B. Davvaz, "Rough set theory applied to lattice theory", Inf. Sci. vol. 200, **(2012)** ,pp.108-122.
[12] R.-l. ZHANG and H.-s. XU, "Building and mapping ontology of e-business based on fuzzy rough concept lattices", Journal of Convergence Information Technology, vol.6, no.9, **(2011)**, pp. 81-88.
[13] L. Wei and J.-j. Qi, "Relation between concept lattice reduct and rough set reduct", Knowledge based Systems, vol.23, no.8, **(2010)**, pp.934-938.

Authors

Hongsheng Xu, He was born on December 28,1979.

Educational background: master, Henan University,Kaifeng, China, 2007;

Major field of study: data mining, Knowledge discovery, artificial intelligence, the Semantic Web.

Ruiling Zhang Professor, She was born on December 23,1964.

Educational background: master, Northwestern Polytechnical University,Xian, China, 2007;

Major field of study: data mining, Knowledge discovery, artificial intelligence, the Semantic Web.

