Domain Ontology Mapping Based Semantic Web Mining Models Research

Jiaojie Cai[1, 2] Yufeng Zhang[1] Feng Hu[1] Jianfeng Dong[1]

[1]Center for Studies of Information Resources, Wuhan University, Wuhan, P. R. China
[2]Xiaogan University, Xiaogan, P. R. China

Abstract

This paper proposes two models about semantic web mining based on domain ontology mapping and gives domain ontology mapping rules. In the first model we semantic annotate web resources based on domain ontology mapping rules, and then implement web mining to get semantic knowledge rules; In the second model we implement fuzzy semantic mapping between initial knowledge rules and domain ontology mapping rules in order to get semantic knowledge rules. By experimenting, we prove the effects of two models which also can greatly improve the degree semantic results of web mining.

Keywords: semantic web mining, domain ontology mapping, semantic annotation, fuzzy mapping

1. Introduction

Web mining is a technology that is combined the traditional data mining technology with web, and is a process that extract the information and knowledge from web resources. As the web resources is massive, heterogeneous, semi-structured, dynamic and difficult to obtain, research of semantic based web mining has become hot spots which aims to improve the degree of semantic results of web mining. Till Plumbaum et al.[1]present a novel approach to track user interaction on a web page based on JavaScript-events combined with the semantic web standard micro-formats to obtain more fine-grained and meaningful user information. Yuefeng Li et al.[2]discuss ontology-based problem solving approaches for building a bridge between web mining and the effectiveness of using data mining. Ming Yi et al.[3]analysis the personalized recommendation method based on web semantic knowledge. Jie Tang et al.[4]propose an approach called RiMOM to automatically discover mapping between ontology based on Bayesian decision theory. It is important to note that all of the prior works adopt a shallow approach to discuss web mining based on domain ontology. Specifically, they only consider the importance of concepts but not that of relations in constructing the model of semantic web mining system. Another reason is that they ignore the important role of domain ontology mapping.

Ontology mapping can well solve heterogeneous problem of different web resources. In this paper we present two models of semantic web mining based on domain ontology mapping. In the first model we firstly use web technology to semantic annotate web resources based on domain ontology mapping rules, and then we again use web technology to get relevant semantic knowledge rules, it is a cyclic process in order to update semantic knowledge rules continuously; In the second model we firstly use web technology to get initial knowledge rules,

and then we get semantic knowledge rules by fuzzy mapping between the initial knowledge rules and domain ontology mapping rules, this is also a cyclic process which aims to get semantic knowledge rules continuously. And the experiment proves that the effects of two models in improving the degree of semantic results of the web mining.

The rest paper is organized as follows: Section 2 and Section 3 introduce two models of semantic web mining based on domain ontology mapping. Section 4 gives the experiment results. The final section concludes and highlights future works.

2. Introduction of the First Model

The first model is on the premise of semantic web resources which uses web mining technology to semantic annotate web resources, and then based on this generates semantic knowledge rules. The first model shows in Fig.1.

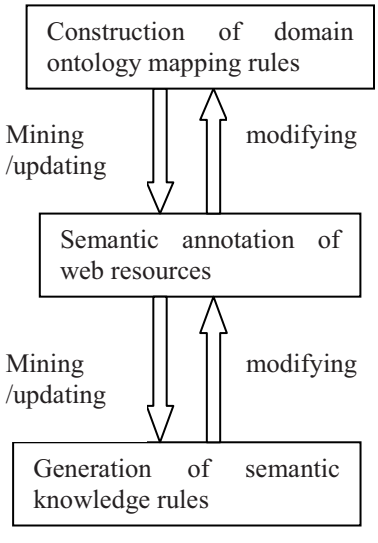

Fig. 1: Semantic web mining based on the first model.

2.1. Construction of Domain Ontology Mapping Rules

Mapping rules means to construct equivalent, homonymy, overlapping, hyponymy, whole-part and opposite semantic relations between heterogeneous domain ontology concepts[5]. In this paper we construct mapping rules by referring to the semantic similarity algorithm of HowNet[6]. Mapping rules construction process shows as follows:

- According by HowNet algorithm rules that overall semantic similarity of concepts should establish on parts of semantic similarity of concepts, and which divides every concept into several original concept atoms. And according to the degree contribution to the overall semantic similarity, every original concept atom can be divided into primary, other, relational and symbol part in descending order. Then the overall semantic similarity of concepts equal to weighted average of four parts of original concept atoms. Where s_1 and s_2 are concepts which need semantic mapping; $sim(s_1, s_2)$ is the semantic similarity of concepts s_1 and s_2 ; β_i （ 1≤i≤4 ） is an adjustable parameter; moreover, $\beta_1+\beta_2+\beta_3+\beta_4=1$, $\beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$, which reflects the degree contributions to the overall semantic similarity in descending order form Sim_1 to Sim_4. And $sim_j(s_1, s_2)$ is respectively semantic similarity of four parts of original concept atoms. the formula is as follows:

$$sim(s_1, s_2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} sim_j(s_1, s_2) \quad (1)$$

- Many homonymy concepts can

be merged according by synset in WordNet. The semantic mapping of other concepts using the algorithm rules of HowNet. If $sim(s_1, s_2)$ =1, and the semantic similarity between corresponding concept nodes linking s_1 and s_2 also equals to 1, then the relation of these concepts are equivalent; if $sim(s_1, s_2) \in (0,1)$, and the semantic similarity between corresponding concept nodes linking s_1 and s_2 also belongs to this interval, then the relation of these concepts are hyponymy or Whole-Part ; if $sim(s_1, s_2)$ =0 , and the semantic similarity between corresponding concept nodes linking s_1 and s_2 also equals to 0, then the relation between concepts are opposite.

2.2. Semantic Annotation of Web Resources

Concepts and their relations can be semantic annotated by web mining technology referring to the domain ontology mapping rules, and then web resources can be understood under unified semantic environment. The illustration of flowchart gives as Fig.2.

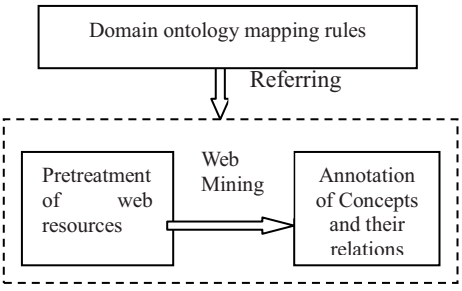

Fig. 2: Flowchart of web resources semantic annotation.

Firstly, we process web resources. Noisy information such as advertisements, broken links should be removed from web texts; Body texts and meta tags should be extracted form web texts, and then we organize meta tags into text sets; Terms can be extracted form body texts using Chinese lexical analysis system (ICTCLAS), so body texts can be expressed by vector-space model, then high-dimensional vector-space model should be reduced in order to enhance semantic relationship between concepts using Latent Semantic Indexing(LSI) method of singular value decomposition(SVD) technique[7].

Secondly, we semantic annotate concepts. Obtaining concepts is the process of clustering terms. We transpose pretreated text vector-space model into "term-text" matrix N, and then use plane partition method[8] to divide term sets into several clusters, and the center of clusters are concepts which will be annotated. Specific algorithm is as follows:

- Determine the generation number of terms clusters K (K<n).
- Get K number upper concepts to be seeds for clustering from matrix N, $S = \{s_1,\ldots,s_j,\ldots,s_k\}$, and $V(s_j)=(W s_j(d_1),\ldots, W s_j(d_i),\ldots, W s_j(d_n))$.
- Calculate the similarity between every term t_i and every concept seed s_j, the value of $sim(t_i,s_j)$ equal to the cosine of vector V（t_i）and V（s_j）,formula is as follows:

$$sim(t_i, s_j) = \frac{\sum_{i}^{n} W_{t_i}(d_i) * W_{s_j}(di)}{\sqrt{\sum_{i=1}^{n} W_{t_i}(d_i)^2} * \sqrt{\sum_{i=1}^{n} W_{s_j}(d_i)^2}} \quad (2)$$

- Select concept seed with maximum similarity of arg max $sim(t_i,s_j)$,the term t_i can be classified to the cluster C_j with the center of concept seed s_j, then get one cluster of term

sets C＝{c_1,…,c_j,…c_k}.

- Redefine the center of each cluster.
- Repeat steps (2), (3), (4), (5), until the center of each cluster will not be replaced, the instance t_i is no longer reallocated.
- Supplement, modify and update the annotation results with assistance of experts referring to the heterogeneous domain ontology mapping rules.

Finally, we semantic annotate concepts relations. In this paper we find hierarchical relation, equivalent relation and opposite relation form web resources by using association mining and statistical methods. Specific algorithm is as follows:

- Set the minimum support threshold Smin and the minimum confidence threshold Cmin, then use Apriori algorithm to find all of frequent term sets W＝{t1,…,ti,…tn}, and directly generate strong associate rule sets R ＝ {r1,r2,…,ri,…},ri={ti $\Rightarrow$ tj},where ti,tj∈W, and P（ti$\cup$tj）>Smin,P(tj|ti)>Cmin.
- For ti,tj∈W, if they meet conditions of ti$\Rightarrow$tj 和 tj$\Rightarrow$ti, then the relation of them is equivalent or opposite[9].
- For ti,tj∈W, and {ti $\Rightarrow$ tj}∈R,P(tj|ti)>P(ti|tj),if the probability of which text sets contained tj is a subset of text sets contained ti is greater than the probability of which text sets contained ti is a subset of text sets contained tj , then the relaton of ti and tj is hierarchy, ti is the upper concept of tj .
- Select strong association rules set R1 from R with hierarchical relations, R1 ＝ {r1,…,ri,…rn}, where ri={ti $\Rightarrow$ tj}, and P(tj|ti)>P(ti|tj). Then form R1 select strong association rules training set Ris-a and Rwhole-part. Calculated the value of maximum, minimum range, and average respectively of training sets Ris-a and Rwhole-part based on P（tj|ti）as a standard of learning hierarchical relations.
- For any ri={ti$\Rightarrow$tj}∈R1, if P（tj|ti）∈Ris-a[Pmin,Pmax],and P（tj|ti）$\sim$ Ris-a ($\overline{p}$), then ri should be allocated to the relation of hyponymy, and the whole-part relation can be inferred using the same algorithm.
- Supplement, modify and update the annotation results with assistance of experts referring to the heterogeneous domain ontology mapping rules.

2.3. Generation of Semantic Knowledge Rules

Finally, we can get semantic knowledge rules. The state of web resources can be improved from machine-readable to machine-understandable by semantic annotating. Based on this level we implement web mining technology to get semantic mining rules naturally.

3. Introduction of the Second Model

The second model aims to continuously implement fuzzy semantic mapping between initial knowledge rules and domain ontology mapping rules in order to change the initial knowledge rules into semantic knowledge rules. The second model shows in Fig. 3.

3.1. Fuzzy Semantic Mapping

The basis of fuzzy semantic mapping is fuzzy logic theory[10], which is a rule that simply make input of one space mapping to output of another space. For the membership calculation of fuzzy mapping, we still refer to semantic similarity

algorithm of HowNet. And in this paper we make the value of membership approximately equal to semantic similarity value in order to reduce calculation complexity. The specific algorithm is as follows:

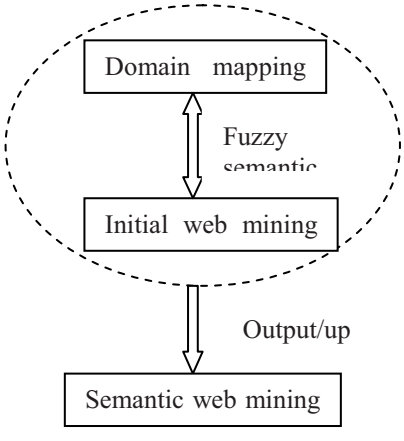

Fig. 3: Semantic web mining based on the second model.

- Initial knowledge rules should be pretreated firstly based on synset in WordNet in order to correct spelling errors, and merge homonymy terms.
- For initial knowledge rules getting from web cluster mining, firstly calculate semantic similarity between the center point of every cluster and every concept in domain ontology mapping rules, and set semantic similarity threshold $\alpha \in [0,1]$, the greater of value α , the higher of semantic similarity, then the center point of every cluster can be attributed to the corresponding conceptual level; secondly calculate semantic similarity between other points and center point of this cluster, and set semantic similarity threshold $\beta < \alpha \in [0,1]$, then other points can be attributed to sub-conceptual level or instance sets according by descending order of value β ; finally the relations of concepts between different initial knowledge rules can be organized into equivalent, hyponymy, whole-part styles referring to concept relations of domain ontology mapping rules. And initial knowledge rules of web classification mining can be improved to semantic level in the same way.
- For initial knowledge rules getting from web association mining, firstly respectively calculate semantic similarity between terms on both sides of association rules and every concept in domain ontology mapping rules, and set semantic similarity threshold α' , where $\beta < \alpha' < \alpha \in [0,1]$, then terms can be attributed to the conceptual level with the highest semantic similarity; Secondly calculate semantic similarity of terms on both sides of association rules, and set semantic similarity threshold $\gamma \in [0,1]$, if $\gamma = 1$, then the relation of terms is equivalent; if $\gamma = 0$, the relation is opposite; if $\gamma \in (0,1)$, the relation is hierarchical; finally further organize the relation of concepts referring to domain ontology mapping rules.

3.2. Semantic Knowledge Rules Updating

For the new initial knowledge rules can be improved to semantic level in accordance with algorithm introduced by section 3.1; For the generated semantic knowledge rules can be updated regularly under the guidance of experts referring to the domain ontology mapping rules.

4. Experiments

4.1. Experiment of the First Model

We select eight heterogeneous domain ontology form agricultural field, and construct mapping rules of them according by the algorithm of section 2.1. Part of mapping rules gives in Table 1.

Several Key Mapping	The Concept Relations Mapping Rules
Equivalent mapping	Solanum lycopersicum=tomato；cauliflower=cabbage；sweet potato=Chinese potato；potato=Solanum tuberosum；garden=park；fisheries=aquatic products；greenhouse=hothouse .....
Hyponymy mapping	vegetables ⟶ tomato；flowers ⟶ cactus；garden ⟶ weeping willow；fisheries ⟶ shrimp；animal husbandry ⟶ cattle；fertilizer ⟶ organic fertilizer；seeds ⟶ cotton seed…..
whole-part mapping	vegetables ⟶ vegetable seeds；fruits ⟶ pulp；flower ⟶ petal；weeping willow ⟶ willow leaves；cattle ⟶ milk；greenhouse ⟶ heating equipment；fish ⟶ fish scales…..
Opposite mapping	vegetables ⟷ fruits；flowers ⟷ forest；animal husbandry ⟷ fishery；fresh fruit ⟷ dried fruit….

Table 1: Part of mapping rules.

Then we select twenty texts form agriculture-related web sites randomly. Firstly we use web mining technology to get initial knowledge rules form texts, and improve them to semantic level with assistance of experts in the filed based on heterogeneous domain ontology mapping rules; Secondly we get texts pretreated to clear noise, segment words, reduce dimensions, then use plane partition algorithm to annotate concepts for every term cluster, and use association mining and statistical algorithm to annotate concept relations; Last we implement web mining based on web resources with semantic annotation to get semantic knowledge rules. Then we can evaluate the results of semantic web mining using precision and recall rate[11] of information retrieval method, and take semantic knowledge rules mutually annotated by experts as an example, where the precision and recall rate can be defined as follows:

$$precision = \frac{\text{mining } the\ correct\ number\ of\ semantic\ rules}{\text{mining } the\ all\ number\ of\ semantic\ rules}$$

$$racall = \frac{\text{min}ing\ the\ correct\ number\ of\ semantic\ rules}{the\ all\ correct\ number\ of\ semantic\ rules}$$

In this experiment, we get 23 numbers of initial knowledge rules using general web mining methods, where the number of clustering and classification rules is 14, the number of association rules is 9; Experts will modify these initial rules and assign one value to them according by their experiment and semantic similarity in descending order. For example, if the term of initial knowledge rules is expressed semantic contains fully which will assign value 1, and we can assign other rules with values 0.95, 0.83,…..., these values are irregular containing the semantic similarity opinions of experts; Finally we get the precision and recall rates of general web mining separately equals to 19.3% and 8.7%. In order to reduce the complexity of the contrast, we also extract 23 numbers of semantic knowledge rules based on the first model methods, and the same numbers of clustering, classification and association rules. Experts modify them and assign them values in the same way as modifying general web mining rules. Finally we get the precision and recall rates separately equals to 83.2% and 78.9%. Form these data, we can clearly see that web mining based on the first model can greatly improve the semantic level of mining

knowledge rules. The compared results show in Table 2.

Mining Ways	Comparing index	
	Precision	Recall
General web mining	19.3%	8.7%
Web mining based model 1	83.2%	78.9%

Table 2: Compared results based on the first model.

4.2. Experiment of the Second Model

We still take the twenty texts in agriculture field as experiment subjects used in the first model. On one hand we firstly use web mining technology to get initial knowledge rules, and then improve them to semantic level with assistance of experts based on domain ontology mapping rules; on the other hand we implement fuzzy semantic mapping between initial web mining knowledge rules and domain ontology mapping rules according to algorithm of section 3.1 in order to get semantic knowledge rules, and respectively set threshold $\alpha = 0.53$, $\beta = 0.31$, α' =0.42. So in order to ensure the consistency of the experiment, in the experiment of the second model we also select the same numbers of knowledge rules, and experts also revise them in the same way which put in the first model. he compared results show in Table 3.

Mining Ways	Comparing index	
	Precision	*Recall*
General web mining	19.3%	8.7%
Web mining based model 2	75.6%	68.9%

Table 3: Compared results based on the second model.

From table 2 and table 3, we learn that the semantic web mining based model 1 and model 2 can also greatly improve the semantic degree of general web mining knowledge rules. But the semantic web mining based on model 1 can get better degree of semantic knowledge rules than the semantic web mining based on model 2, which reason is that the theoretical basis of model 1 is semantic web, and web mining based on web resources with semantic annotation can naturally get semantic knowledge rules; but the theoretical basis of model 2 is fuzzy logic theory, and there have errors in making fuzzy semantic mapping between initial knowledge rules and ontology mapping rules, so whatever in precision and recall rate the degree semantic results based model 2 will lower than the model 1. However in terms of time complexity of algorithm put in model 1 and model 2, the semantic web mining based on model 2 can spend less time than model 1, which the reason is that semantic annotate web resources is a huge engineering and so far there are no good methods of semantic annotation; but the fuzzy semantic mapping can get semantic results using semantic similarity method, and this method is already mature. So the time complexity of algorithm based on model 2 is O（N+M）, the time complexity of algorithm based on model 1 is O（N^2+M^2）, that N is the number of concepts getting form twenty texts, M is the number of concept relations getting form twenty texts.

5. Conclusions

Due to the web resources is massive, heterogeneous, and difficult to understand, the general web mining cannot meet the needs of human and computer, the barrier of communication freely between heterogeneous computer systems has been the main problem of development of semantic web. This paper proposes two models about semantic web mining based on heterogeneous domain ontology mapping in order to improve the degree of

semantic results of general web mining, solve the barrier of communication freely between heterogeneous computer system and contribute to the development theory of semantic web. Experiments show that the semantic web mining based on two models can greatly improve the degree of semantic results. But our research still has shortages, on one hand the algorithm of every model has high time complexity, with more manual intervention; on the other hand the experimental texts in experiment take low coverage. In the future, we will continuously optimization algorithm and expand our experimental coverage.

Acknowledgment

This paper is supported by the National Natural Science Foundation of China (NO. 71073121) , the MOE (Ministry of Education) Project of Key Research Institute of Humanities and Social Sciences at Universities (NO. 08JJD870225).

References

[1] Till Plumbaum, Tino Stelter, and Alexander Korth， “Semantic Web Usage Mining: Using Semantics to Understand User Intentions,” in G.-J,LNCS 5535, Houben,Eds.Heidelberg: German,2009,pp.391-396.

[2] Yuefeng Li and Ning Zhong, “Mining Rough Association from Text Documents for Web Information Gathering,” in Transactions on Rough Sets, vol. VII, LNCS 4400, J.F.Peters,Eds.Heidelberg: German,2007, pp.103-119.

[3] Ming Yi, Jinlong Zhang, Weihua Zhang, “An Approach of Personalization for Electronic Commerce Website,”in Journal of China Society for Scientific and Technical Information, vol Ⅴ,2005,pp.567-572.

[4] Jie Tang, Bangyong Liang, Juanzi Li, Kehong Wang, “Risk Minimization Based Ontology Mapping,” LNCS 3309, Chi and K.-Y. Lam, Eds. Heidelberg: German, 2004,pp.469-480.

[5] Jianjiang Lu, Yafei Zhang, Zhuang Miao, Po Zhou,The Semantic Web Principle and Technology, Science Press:Beijing ,2007,pp.57-81.

[6] Yi Jiang, You Ding, “An Improved Computation Method of Word Semantic Similarity Based on HowNet,”in Journal of Chongqing University of Posts and Telecommunication, vol Ⅲ,2009,pp533-537.

[7] Wendong Zhang, Yihu Yi, “To Construct the Set of Synonyms and Association Words Using Latent Semantic Analysis and the Mining of Association Rules,” Computer Engineering and Science,vol Ⅵ,2007,pp.104-116.

[8] Jiawei Han, Micheline Kamber,Data Mining Concepts and Techniques, China Machine Press:Beijing ,2007,pp.137-148.

[9] Xin Zhao, Zhi Cai, “A Chinese Conception Sets-Generating Algorithm Based on Association Rules,”in Computer Science. vol.31,2004, pp.175-176.

[10] Raymond Y.K.Lau, Chapmann C.L.Lai, Yuefeng Li, “Mining Fuzzy Ontology for a Web-Based Granular Information Retrieval System,” LNCS 5589, P. Wen, Eds.Heidelberg:German,2009, PP.239-246.