SNS에서 Decision Tree를 이용한 온톨로지 생성 방법

Qing Wang, 손종수, Yun-Feng Pei, 정인정
고려대학교 컴퓨터정보학과 지능정보시스템 연구실
{wangqing, mis026, chung}@korea.ac.kr

Ontology Generation Method on SNS Using Decision Tree

Qing Wang, Jong-Soo Sohn, Yun-Feng Pei, In-Jeong Chung
Intelligent Information System Lab. Dept. of Computer and Information Science, Korea University

Abstract

Ontology is used to describe and represent an area of knowledge. To date, the domain experts and ontology engineers have usually taken responsibility of the construction of ontology. However, ontology experts are in short supply who can describe the enormous real world knowledge. The birth of SNS heralded a new era of development of the network. In the mean time, a huge volume of useful information abounds in SNS. In this paper, we propose an approach toward generating a great variety of domain ontology without experts using all-encompassing information on SNS. In our approach, we create a domain decision tree to generate the domain ontology based on the training data extracted from SNS websites. The result demonstrates that our methods can successfully generate domain ontologies. It is our hope that the result of our research can enable the generation of domain ontology easier for diverse fields.

1. Introduction

The birth of SNS (Socical Networking Services) proclaimed that the development of network is entering a new era [1].

Ontology is used to describe and represent an area of knowledge [2]. However, the construction of ontology largely depends on a cooperation between the limited number of domain experts and ontology engineers [3]. For the description of enormous real world knowledge, the experts in generation of ontologies are in short supply. In this paper, we suggest an approach which generates a great variety of domain ontologies using all-encompassing information on SNS. In our approach, we create a domain decision tree to generate the domain ontology based on the training data extracted from SNS websites. The result shows that the use of our approach can enable the successful generation of a domain ontology, making a domain ontology generation easier for a variety of diverse fields.

2. Making decision tree

Decision tree learning is one of the most popular classification methods and many algorithms have existed such as information gain and gain ratio by Quinlan in 1993 [4] and gini index by Breiman et al in 1984 [5].

In order to make the decision tree for domain, within a limit, we collect a sample set randomly from delicious website(http://delicious.com) as training data, which consists of top 5 tags and webpages links and store it as a table.

2.1 ID3 algorithm

We made a table to describe a situation for ID3, which consists of all the keywords and classes. Then we judged the relationship between the training data and keywords. In each piece of training data, if there is any tag matching with any keyword, we recorded the cell of keyword as "YES", otherwise "NO", in the ID3 table. We checked out the classes of webpages and record the class for each piece of training data in ID3 table. We calculated the information gained from (1), (2), (3), (4) to select the position of each attribute in the tree. Let S be a set consisting of data samples. Assume that the attribute has m distinct class values, that is $C_i(\text{for } i = 1,...,m)$. s_i is the ith element of S

of class C_i . p_i is the probability that an arbitrary sample belongs to class C_i. The expected information needed to classify a given sample is provided as follows:

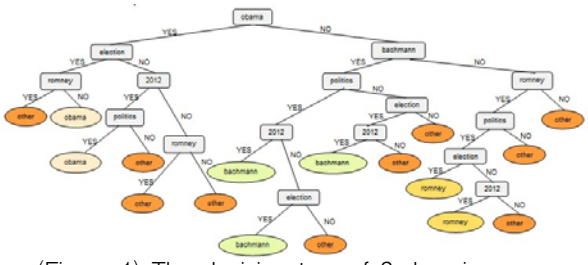
$$I(s_1,s_2,...,s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (1)$$

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j}+...+s_{mj}}{s} I(s_1,s_2,...,s_m) \quad (2)$$

$$I(s_{1j},s_{2j},...,s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2(p_{ij}) \quad (3)$$

$$Gain(A) = I(s_{1j},s_{2j},...s_{mj}) - E(A) \quad (4)$$

We calculated the information gain and determined all attributes of the decision tree: Obama, Romney and Bachmann as candidates of US presidents in 2012, in which 3 domains contained in the same one tree as shown in Figure 1.

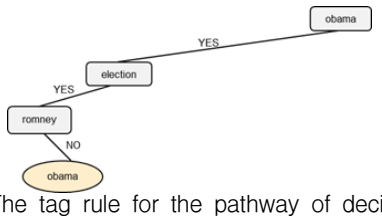


(Figure 1) The decision tree of 3 domains

3. Generating the domain ontology

Ontolgy is a formal knowledge representation and specification as a set of concepts within a domain and the relationships between the concepts. It could be widely used in many fields of information architecture, and thus the generation of domain ontologies is crucial.

In the domain decision tree, there are several path ways in different sub-trees to go to different classes. Each pathway consists of several keywords and can be expressed as a tag rule by OWL-DL.



(figure 2) The tag rule for the pathway of decision tree

For example, Figure 2 shows a pathway in the domain decision tree. According to the calculation result of the information gain, the keyword Obama, election and Romeny play roles as attributes in a pathway of the domain decision tree. We can use the OWL-DL to write a tag rule for it: $obama \sqcap election \sqcap \neg romeny$

In the decision tree, every pathway was described by OWL-DL corresponding to a rule. On the basis of all rules from the tree, we made an ontology and separated into three classes; Obama, Romney Bachmann, which we needed as shown in Figure 3.

TBox/RBox

(1) ObamaWebpage ≡ DeliciousWebpage
(2) ObamaWebpage ≡hastag ∃ ((Obama ⊓ Election) ⊔(Obama ⊓ _2012 ⊓ Politics))
(3) ObamaWebpage ≡ ¬ Romney
(4) BachmannWebpage ≡ DeliciousWebpage
(5) BachmannWebpage ≡ hastag ∃ (Bachmann ⊓ ((Politics ⊓ (_2012 ⊔ Election)) ⊔(_2012 ⊓ Election)))
(6) BachmannWebpage ≡ ¬ Obama
(7) RomneyWebpage ≡ DeliciousWebpage
(8) RomneyWebpage ≡ hastag ∃ (Romney ⊓ Politics ⊓ (Election ⊔ _2012))
(9) RomneyWebpage ≡ ¬ Obama
(10) RomneyWebpage ≡ ¬Bachmann
(11) [Romney, Politics, Election , Obama, Bachmann, _2012...] ⊏ ∃ WebpageTag
TagOf ¯ ≡ hasTag

ABox
<Webpage_1,..., Webpage_n>: hasTag
<romney, politics, election , obama, bachmann, _2012_,...,>:TagOf

(Figure 3) Domain ontology by OWL-DL

4. Conclusion

In this paper, we propose an approach using the domain decision tree to generate the domain ontologies without experts based on SNS. We extracted a mount of data randomly from the Delicious website as training data to make a domain decision tree based on ID3 algorithm. Each pathway of decision tree can be expressed as a tag rule by OWL-DL to generate the domain ontology. The result of our research demonstrates that it can easily and successfully generate a domain ontologies by applying the use of SNS dada.

Reference

[1] Rémy Magnier-W, Michiko Y, Tomoaki W. "Social network productivity in the use of SNS". Journal of Knowledge Management, Vol. 14, 2010, pp.910 - 927.

[2] Liyang Y. "Introduction to Semantic Web and Semantic Web services", 2007, TK5105.88815Y95

[3] Aguado C, Montiel-Ponsoda G, Suárez-Figueroa E, Mari C. "Approaches to ontology development by non ontology experts". ISMTCL, 1-3 2009.

[4] Milija S, Boris D, Milos J, Milan V, Dragana B-Vujaklija, Zoran O. "Reusable components in decision tree induction algorithms". Comput Stat, Springer-Verlag 2011.

[5] J.R. Quinlan."Simplifying decision trees". International Journal of Man-Machine Studies,Volume 27, Issue 3, September 1987, Pages 221-234