Lecture Notes in Electrical Engineering 215

Kuinam J. Kim · Kyung-Yong Chung *Editors*

IT Convergence and Security 2012

Volume 2

The proceedings approaches the subject matter with problems in technical convergence and convergences of security technology. This approach is new because we look at new issues that arise from techniques converging. The general scope of the proceedings content is convergence security and the latest information technology. The intended readership are societies, enterprises, and research institutes, and intended content level is mid- to highly educated personals.

The most important features and benefits of the proceedings are the introduction of the most recent information technology and its related ideas, applications and problems related to technology convergence, and its case studies and finally an introduction of converging existing security techniques through convergence security. Overall, through the proceedings, authors will be able to understand the most state of the art information strategies and technologies of convergence security.

Electrical Engineering

ISBN 978-94-007-5859-9

▶ springer.com

Kim · Chung *Eds.*

IT Convergence and Security 2012

2

Lecture Notes in Electrical Engineering 215

Kuinam J. Kim
Kyung-Yong Chung
Editors

IT Convergence and Security 2012

Volume 2

Springer

Generation of User Interest Ontology Using ID3 Algorithm in the Social Web

Jong-Soo Sohn, Qing Wang and In-Jeong Chung

Abstract It is feasible to collect individual user interests from social networking services. However, there have been few studies of the interests of domain users. In this paper, we propose an approach for ontology generating the interests of SNS domain users by employing semantic web technology and ID3 algorithm.In our approach, domain ontology is generated by a decision tree, which classifies the domain web pages and the domain users. Experimental test shows ontology of the interests of domains users regarding USA presidential candidates. We expect that our results will be beneficial in the field of computer science, such as recommendations, as well as other fields including education, politics, and commerce. Proposed approach overcomes the problem of domain user classification and lack of semantics by composing decision tree and semantic web technology.

Keywords Semantic web · Ontology · ID3 algorithm · SNS · FOAF · Interest extraction · Social web · Classification · Election · OWL

J.-S. Sohn · Q. Wang · I.-J. Chung (✉)
Department of Computer and Information Science, Korea University,
208 Seochangri, Sejong City, South Korea
e-mail: chung@korea.ac.kr

J.-S. Sohn
e-mail: mis026@korea.ac.kr

Q. Wang
e-mail: wangqing@korea.ac.kr

K. J. Kim and K.-Y. Chung (eds.), *IT Convergence and Security 2012*,
Lecture Notes in Electrical Engineering 215, DOI: 10.1007/978-94-007-5860-5_128,

1 Introduction

The interests of user have an important role in various areas such as recommendations, commerce, and political activities, although less information is provided by users than is available on networks. Thus, the working efficiency is very low if we have to identify the interests of users one by one. The interests of individual users are valuable for personalized services, but they cannot satisfy the requirements of public services because public users lack the universality. By contrast, the interests of public users lack personality, although they possess universality. The interests of domain users are different from individual interests and public interests. They are the common interests of users in the same domain and they have different personalities because users come from a different domain, although they have the same universality because users are in the same domain. Therefore, the value of the interests of domain users is greater than the interests of individual user' and public users in other fields in addition to computer science. However, only a few studies have addressed the interests of domain users. The process of extracting the interests of domain users has the following two problems [1].

- Definingthe user domain. It is simple for a human to define a user domain, but this is a very difficult task for a computer, because machines do not understand semantics.
- Classifying different domain users. Defining a user domain is very difficult for computers. Computers are unable to classify different domain users based on a traditional classification if this requires an understanding of semantics.

Therefore, we propose an approach for extracting the interests of domain users, which exploits semantic web technology for collecting them from SNS. We create a domain ontology that helps computers understand semantics. Semantic web technology is a maturing technological field, which continues to be the focus of much research

Our methodology addressed the difficulties of defining user domains and the classification of different domains to identify the interests of domain users. With our approach, domain ontology is generated by a decision tree that classifies the domain webpages and domain users. Experimental test shows ontology of the interests of domains users regarding USA presidential candidates. The result of our research has great value for computer science and other fields including education, commerce, and political activities.

2 Related Work

Folksonomy is a very important part of web 2.0, which is used to classify SNS such as the Delicious website. This is a new method that allows users to classify information [2],which is also is associated with a semantic problem [3].

Research into this area has included recommendations, user interests, and classification. Illig et al. proposed a tag recommendation method for folksonomy [4]. In [5], Shan et al. described item recommendation based on folksonomy. Szomszor et al. proposed the modeling of user interest based on folksonomy while Kawase et al. proposed a classification for user interest patterns based on folksonomy [6, 7]. Neither classification nor the extraction of user interests of based on folksonomy can achieve classification on their own. They still require other tools or cooperation with traditional classifiers, because folksonomy is associated with the semantic problem.

Few studies have addressed classification methods based on semantics using ontology. Some researchers have focused on the interests of domain users, while others have addressed the interests of individual users. Lipczak proposed a tag recommendation method based on the interests of individual users [8]. Yin et al. proposed the semantic modeling of individual user interests and collaborative tagging systems for individual users [9]. Sasaki et al. proposed a method for extracting preference terms, which excluded unrelated pages from user interests where the user also referred to individual users [10]. White et al. proposed a method for predicting the interests of individual users based on contextual information [11]. However, the interests of domain users have greater value than individual interests and public interests for computer science and other fields.

3 Generation of User Interest Ontology Using ID3 Algorithm

3.1 Generating Decision Trees

Decision tree is a popular classifier, which is characterized by the property that samples are subjected to a sequence of decision rules before they are assigned to a unique class [12]. We acquired training data from the Delicious website to generate a domain decision tree. The domain decision tree is used to create domain ontology. And every pathway in a decision tree can be written using a tag rule such as the Web Ontology Language OWL-DL, so the sub-tree can be translated into domain ontology without loss during classification based on semantics. Decision trees are part of the classification when generating ontology. We use the ID3 algorithm to generate a decision tree for a domain [13]. OWL was developed as a more formal and more powerful ontology language than RDFS. OWL-DL supports users who require maximum expressiveness while retaining computational completeness and decidability [14].

In generate a decision tree for domain within specific limits, we collected a random sample set from the Delicious website as training data, which consisted of the top 5 tags and webpage links, which we stored as a table. The webpage “The Choice: Comment: The New Yorker” is marked by some users on the Delicious website. We extracted the webpage link “The Choice: Comment: The New Yorker”

Fig. 1 An item of training data in the ID3 table

obama	romney	bachmann	politics	election	2012	calss
YES	NO	NO	YES	YES	NO	Obama

and the top 5 tags for Obama, politics, election, New Yorker and an endorsement of this paper, and we recorded them in a table as training data for the decision tree.

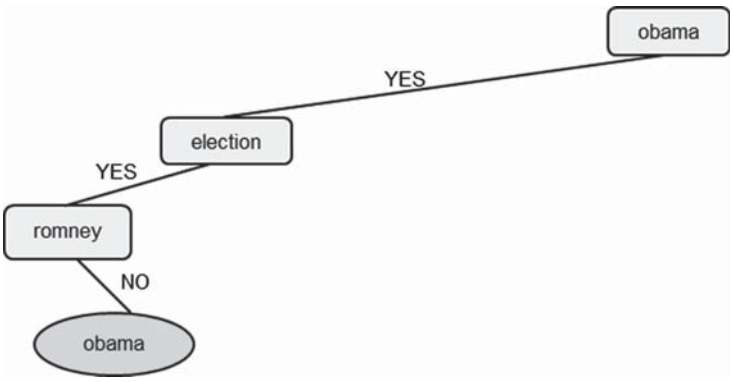
We prepared a table to describe a situation for ID3, which consisted of all the keywords and classes. Next, we determined the relationships between each piece of training data and keywords. In each piece of training data, if any tag matched any keyword, we recorded the cell of the keyword as "YES" in the ID3 table. Cells of keywords were recorded as "NO" when there no tag matched with it. We recorded the cells for Obama, politics, and election as "YES" whereas the others were "NO" in the ID3 table, as shown in Fig. 1. Next, we check the content of the webpage and record its class. Finally, all of the training data is recorded in the table.

Information gain is a measure of the information theory, which characterizes the impurity of an arbitrary collection of examples. The method used for information gain is covered in [15]. After the information gains are calculated for each attribute, we can draw the domain decision tree. Every attribute will have a position based on the results by information gain.

Generated decision tree contains several pathways in different sub-trees that lead to different classes, so the same class may have several pathways. Each pathway consists of several keywords that can be expressed as a tag rule using OWL-DL.

For example, Fig. 2 shows a pathway in the domain decision tree. Based on the calculation of the information gain, the keywords 'obama', election, and 'romney' are attributes in a pathway in the domain decision tree. We can use the OWL-DL to write a tag rule as follows: obama $\sqcap$ election $\sqcap \neg$ romney.Because of the top 5 tags, every webpage on the Delicious website also has a tag rule, which is stored in our webpage database. Based on all of the pathways in the decision tree, we can write tag rules using OWL-DL for every class in the domain ontology.

Fig. 2 Apathway in the domain tree



3.2 Ontology Learning

To infer and classify webpages into different domain classes, the domain ontology has to be trained using the webpage data. The webpage data is used to train the ontology, which is also classified. This is an important part of classification, which we discuss further in this chapter, and it consists of data storage, learning, and inference.

We store webpage data in the database, which consists of the URL of a webpage and tags. We write a tag rule for each webpage to infer the class of the webpage based on the top 5 web page tags using OWL-DL. The learning rules are as followings.

- If the tag rule of the webpage exists in the class "*DeliciousWebpage*", we create a new instance to store the URL and the instance belongs to the subclass "*Condition_n*", which contains the same tag rule. If the tag rule of the webpage does not exist in the class "*DeliciousWebpage* ", we create a new subclass "*Condition_n + 1*" and store the tag rule as a property in "*Condition_n + 1*". We then create a new instance to store the URL and this instance belongs to the subclass "*Condition_n + 1*".
- If we let the tag moves to the class "*Webpagetag*" in the ontology. As with the class "*DeliciousWebpage*", if the tag exists in the class, we create a new instance to store the tag and the instance belongs to the subclass "Tag_n", which contains the same classes such thatUS, USA, and America belong to the same subclass "USA".
- If the tag does not exist in the class "*Webpagetag*", we create a new subclass "*Tag_n* + 1" and create a new instance to store the tag and the instance belongs to the subclass "*Tag_n* + 1".

4 Experiments and Evaluations

We conducted an experiment on a domain containing three US presidential candidates. In this experiment, we selected three USA presidential candidates in 2012 as the test domain. They were the Democrat Party's Barack Obama, and Mitt Romney and Michele Bachmann from the Republican Party. There were 497,849 webpages on Obama, 12,271 webpages on Romney, and 4,570 web pages on Bachmann on theDelicious website. We selected 100 random webpages for each candidate for classification.

We selected sixty web pages of Obama, Romney, and Bachmann from the Delicious website as a training set to make the table for the decision tree.

And we calculated the information gain using the ID3 algorithm and determined all the attributes of the decision tree. Obama, Romney, and Bachmann were the candidates and these three domains were contained in the same tree, as shown in Fig. 3.

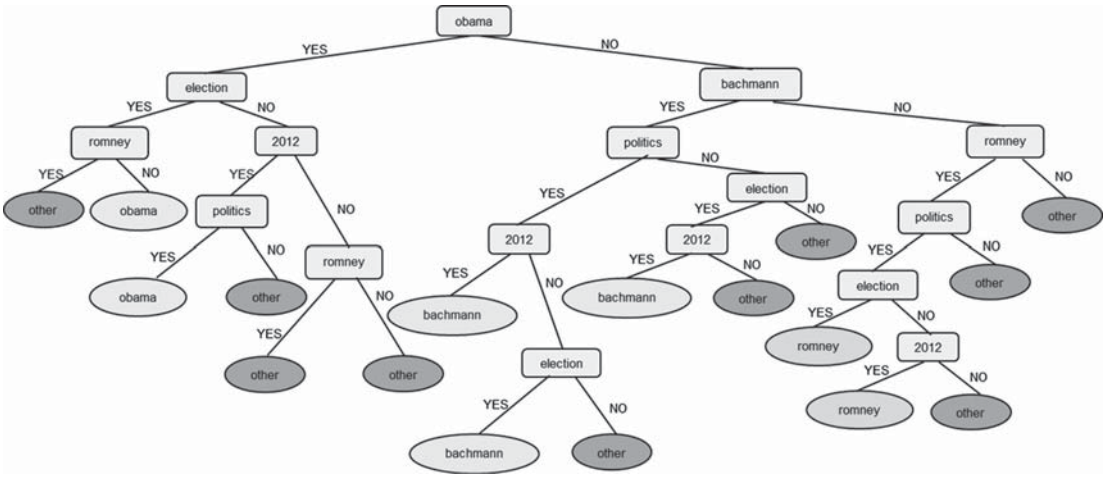


Fig. 3 The decision tree for three domains

In the decision tree, every pathway described by OWL-DL corresponded to a rule. Based on all of the rules in the tree, we produced an ontology for classifying the experimental data and separated the data into three classes, as shown in Fig. 4.

For the justification, we had calculated precision, recall and accuracy [16]. We compared the three types of domain user interests in the experiment with several other types of user interests in our previous studies. We checked 300 web pages and recorded their classes. Table 1 shows the calculation results for the precision, recall, and accuracy of the three domains for their classes. The first column lists the number of quality inbound links for 100 domain web pages. The other three

	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

Fig. 4 Ontology of the domains

Table 1 Results for precision, recall, and accuracy for the candidate Obama class

No.	Precision	Recall	Accuracy
60	0.875	0.824	0.917
65	0.882	0.833	0.923
70	0.895	0.811	0.914
75	0.895	0.811	0.920
80	0.895	0.811	0.925
85	0.900	0.818	0.929
91	0.905	0.826	0.923

columns show the results for precision, recall, and accuracy. The evaluation results showed our approach was reliable for determining the interests of domain users, because it had 91.5 % precision and 93.1 % accuracy for classification.

5 Conclusions

SNS is stimulating the development of the Internet. It accelerates the speed of information sharing between the people, and a vast amount of useful information is available on the interests of users. Users' interests have an important role for computer science and the social sciences. The interests of domain users have greater value, but domain interests have rarely been studied. The process to identify the interests of domain users presents two difficulties: defining user domains and classifying them.

In this study, we addressed these problems of defining user domains and their classification. We proposed an approach based on semantic web technology for identifying the interests of domain users from SNS. Domain ontology was generated by a domain decision tree that classified domain webpages and domain users using ID3 algorithm. Our experiment to determine the interests of domain users on USA presidential candidates showed that our approach had high precision and accuracy. We hope that our research will have value for computer science, such as recommendations, and for other fields such as education, political activities and commerce.

References

1. Zhuge H (2010) Socio-natural thought semantic link network: a method of semantic networking in the cyber physical society perth. In: 24th IEEE international conference on advanced information networking and applications, pp 19–26
2. Zhang T, Lee B, Kang S, Kim H, Kim J (2009) Collective intelligence-based web page search: combining folksonomy and link-based ranking strategy. Computer and Information Technology, 2009, pp 116–171

3. Pi S, Liao H, Liu S, Lin C (2011) Framework for classifying website content based on folksonomy in social bookmarking. In: Intelligent computing and information science, communications in computer and information science, vol. 135. pp 250–255
4. Illig J, Hotho A, Jäschke R, Stumme G (2011) A comparison of content-based tag recommendations in folksonomy systems. In: Knowledge processing and data analysis, Lecture Notes in Computer Science, vol. 6581/2011, pp 136–149
5. ShanS, Zhang F, Wu X, Liu B, He Y (2011) Ranking tags and users for content-based item recommendation using folksonomy. Computing and Intelligent Systems, Communications in Computer and Information Science, pp 32–41
6. Szomszor M, Alani H, Cantador I, O'Hara K, Shadbolt N (2008) Semantic modelling of user interests based on cross-folksonomy analysis. In: The semantic web—ISWC, Lecture Notes in Computer Science, 2008, vol. 5318/2008. pp 632–648
7. Kawase R, Herder E (2011) Classification of user interest patterns using a virtual folksonomy JCDL'11, Ottawa, Canada, ACM 978-1-4503-0744-4/11/06, 13–17 June 2011
8. Lipczak M (2008) Tag recommendation for folksonomies oriented towards individual users. In: ECML PKDD Discovery Challenge, pp 84–95
9. Yin D, Hong L, Xue Z, Davison, BD (2011) Temporal dynamics of user interests in tagging systems. In: Twenty-Fifth AAAI conference on artificial intelligence
10. Sasaki K, Okamoto M, Watanabe N, Kikuchi M, Iida T, Hattori M (2011) Extracting preference terms from web browsing histories excluding pages unrelated to users' interests. In: SAC'11, TaiChung, Taiwan, pp 21–25 March 2011
11. White RW, Bailey P, Chen L (2009) Predicting user interests from contextual information. In: 32nd international ACM SIGIR conference on research and development in information retrieval, ACM New York, USA, pp 19–23
12. Argentiero P (1982) An automated approach to the design of decision tree classifiers. In: IEEE transactions on pattern analysis and machine intelligence, vol. Pami-4, no. 1
13. LópezMántaras R (1991) A distance-based attribute selection measure for decision tree induction. Mach Learn 6(1):81–92
14. Panigrahi S, Biswas S (2011) Next generation semantic web and its application. IJCSI Int J Comput Sci Issues 8(2):385–392
15. Gruber T (2008) What is an ontology. Encyclopedia of database systems, vol. 1. Springer-Verlag
16. vanRijsbergen CJ (1979) Information retrieval, Butterworth-Heinemann Newton, MA