A Taxonomic relationship Learning Approach for Log Ontology Content Event

Sun Ming, Shang Qinghong, Chen Bo
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, China
sunm@uestc.edu.cn

Journal of Digital Information Management

ABSTRACT: *To construct the log ontology is one of the main tasks of semantic Web usage mining. In order to discover the hierarchy of users' visit interesting for web sites, we propose a taxonomic relationship learning approach for content events on log ontology. In this method, event is used to express the users' visiting action, and content event is the semantic behavior of users' visiting to the page content of sites. This method extracts the taxonomic relationship of content event by web document cluster based on swarm intelligence, which combines web content mining and web usage mining. This method improves the results of semantic Web usage mining and provides more decision-making for optimizing the structure of Web sites. The simulation experimental results show that this method is effective and quite feasible to solve practical problems.*

Categories and Subject Descriptors:
H.3.5 [Online Information Services]; Web-based services; [**I.2.7 Natural Language Processing];** Text analysis

General Terms: Web Mining, Ontology, Taxonomy

Keywords: Log Ontology, Semantic web Mining, Content event Taxonomic Relationship, Swarm Intelligence

Received: 8 November 2011, Revised 3 January 2012, Accepted 10 January 2012

1. Introduction

Although web data mining based on ontology can improve the efficiency and quality of mining, the ontology described by knowledge on Internet is very deficient. How to construct web usage ontology on semantic web becomes a mandatory requirement of web usage mining [1].

Ontology learning is the main approach for helping domain specialists and knowledge engineers to get log ontology [2, 3]. Most of researches focus on how to construct the web content ontology [4, 5], which neglects the semantic expression of user usage information. Traditional methods of web mining and ontology learning are discussed in this paper. Based on web document, user request and user access information, a new taxonomic relationship learning method for log ontology content event is proposed in this paper. Compared with the dominating ontology learning tools, this method increases the accuracy and recall rate of learning results. These results provide not only a data set which can be used directly on semantic web, but also a foundation for the mapping and combination of log ontology between different sites. It is also used for the following web mining such as frequent pattern discovery and modes analysis.

2. Event and Log Ontology

User behavior of web site visiting is diverse because web site is related to background knowledge in different area. Currently, log ontology definition advances the user behavior to the semantic level, which makes the analyzer to use abundant ontology language for the research of web using knowledge. This method is inflexible and limitary, because it considers the commonness and relationship of user behavior by the system abstraction. It can't satisfy the requirement of data mining for complex semantic web. This paper analyzes the user behavior and their commonness used for area application. Log ontology is defined based on the core concept.

Analysis focused on user behavior and strategy is abstracted by semantic information. The concept event is used in this paper to describe the semantic information of user access.

Definition 1 Event is the formal description of all the visiting tasks to the business model of a web site.

Event includes the commonness and characteristic of web using. Event is described by event properties, which are collected from visiting information of page content and log files. Event properties describe the characteristic of user visiting from different levels. On the common level Event can be divided into two types: Atom Event and Complex Event according to the aim of site visiting, interest and different semantic strategy. Atom Event describes the user behavior of one-time request to the site. Complex Event is based on the users' visiting mode of holistic information, which is propitious to know users' semantic behavior for analyzer.

1. Atom Event (EC): describes the user behavior of one-time request to the site. The request includes content event and service event. Atom Event can be divided into two types:

Content Event (EC): $EC \subset EA$, EC is the semantic behavior of users' visiting to the page content of sites. It can express the users' interest to the content of a particular web page.

Service Event (ES): $ES \subset EA$, ES is the semantic behavior of users' access of web service. It means users utilize the web service provided by sites to satisfy the specified request.

2. Complex Event (EX): EX is an ordered sequence of EC and ES. $EX = < e_1, e_2, ..., e_n > ei EC \cup ES$, $i = 1, 2, ..., n$

The above classification makes the results of web data mining more reasonable and efficient. It is similar to the definition of application event by Stumme [6], which classify events by the type of user requests. The difference is our methodology analyzes events from both the commonness and field, which can describe users' semantic behavior more accurately.

Events define users' semantic behavior. Mining and research on this behavior is more efficient and thorough with the inference rules. Log ontology belongs to the web visiting area ontology. It describes the semantic web, which based on the concept of event uses formal constructor of knowledge [7].

Definition 2 Log Ontology (LO) is a tetrad, which defines the common behaviors of users' access to the web sites and their relationship. $CLO:=\{\varepsilon \leq \varepsilon, R, F\}$o $\varepsilon \subset E$ is the common events set, $\varepsilon = \{EX, EA\}$, $EA = \{EC, ES\}$; $\leq \varepsilon$ is the Taxonomic relationship of ε, R is non-Taxonomic relationship of ε, $R = \{hasPart, previous\ Of\}$, F is common relationship function, $F = \{EX \times EA \rightarrow hasPart,\ EA \times EA \rightarrow previous\ Of\}$.

3. The Taxonomic relationship learning of Content Event

Definition 3 Event taxonomic relationship, is - a event E_1, $E_2 \in E$, if the connotation of Event E1 contains the connotation of Event E2, and the extension of Event E1 is contained by the extension of Event E2, the relationship between E1 and E2 is call taxonomic relationship, which can be marked by is - $a(E_1, E_2)$.

$$is - a(E_1, E_2) = def\ \forall x(inst(x, E_1) \rightarrow inst(x, E_2))$$

The Taxonomic relationship learning of content events, which is from the perspective of application, focuses on how to find the cluster connection of users' access on semantic level. This paper uses the web document cluster method based on swarm intelligence [8, 9]. It clusters and groups page eigenvector on bottom-up view by the similarity calculation, then revises the taxonomic relationship of content event by cluster results. The similarity of page eigenvector is defined by vector cosine as Equation 1 (m is the dimension of page eigenvector):

$$sim(TVpi, TVpj) = \frac{TVpi \,.\, TVpj}{|TVpi\ |.|TVpj|} = \frac{\sum_{k=1}^{m} fw_k(pi) \,.\, fw_k(pj)}{\sqrt{\sum_{k=1}^{m} fw_k(pi)^2} \cdot \sqrt{\sum_{k=1}^{m} fw_k(pj)^2}} \quad (1)$$

In order to increasing the efficiency of vector space model and decreasing the vector dimension, we can choose some best weights from the eigenvector.

The main idea of swarm intelligence is putting the objects to be clustered on a two-dimension grid randomly [10]. Each object has a random starting point. Ants move on the grid and measure the swarm similarity of current object on the local environment. It transforms the swarm similarity to the probability of moving object by transition functions. Whether to choose an object is decided by this probability. Ant collaborations makes all objects belong to the same class be clustered in the same space. Swarm similarity is an integration of objects to be clustered and all of other models on the local environment. The basic formula of swarm similarity measure for web page cluster is as Equation 2.

$$f(o_i) = \sum_{o_i \in Neigh(r)} \left[1 - \frac{d(o_i, o_j)}{\alpha}\right] \quad (2)$$

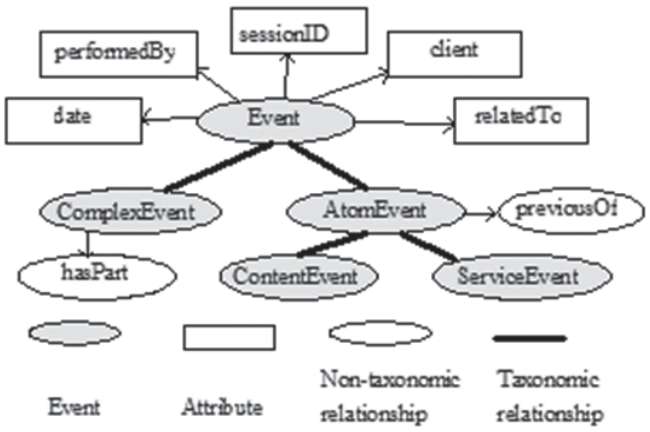

Figure 1. The basic structure of LO

$Neigh(r)$ is the local environment, which represents a circle

area (radius r). $d(o_i, o_j)$ is the distance of two objects o_i and o_j, which is described by the similarity of page eigenvector $sim(TVpi, TVpj)$. α is swarm similarity modulus. It is the critical parameter to measure the swarm similarity. It affects not only the number of clusterbut also the convergence rates of cluster algorithm. Convenient for choosing the parameter we revise the formula as Equation 3.

$$f(TVp_i) = \sum_{TVp_j \in Neigh(r)} \left[1 - \frac{10 \times sim(TVp_i, TVp_j)}{\alpha}\right] \quad (3)$$

This improve can ensure that α is an integer from 1 through 10. Probability transition functions transform swarm similarity to simple object cluster mode function. Its range is from A to B. Probability transition functions are two opposite curves, P_p and P_d.According to the principle of probability curve, we transform the quadratic curve to a line with slope k. The definition is as Equation 4 and Equation 5.

$$P_p = \begin{cases} 1-\varepsilon, & f(TVp_i) \le 0, \\ 1 - k \times f(TVp_i), & 0 < f(TVp_i) \le 1/k, \\ 0, & f(TVp_i) > 1/k \end{cases} \quad (4)$$

$$P_d = \begin{cases} 1-\varepsilon, & f(TVp_i) \ge 0, \\ k \times f(TVp_i), & 0 < f(TVp_i) \le 1/k, \\ 0, & f(TVp_i) \le 1/k \end{cases} \quad (5)$$

4. Simulation Experiments

In order to examine the taxonomic relationship learning method of content event, the prototype system of this method id developed with Java language and Text2Onto, which is a open source software for ontology learning. Text2Onto is the most important tools for ontology learning, which can discover the concepts, the taxonomic and non-taxonomic relationship between concepts. According to the different corpuses, the accuracy of Text2Onto on concept learning is from 70% to 90% in general, and the accuracy of Text2Onto on the relation of classification is from 60% to 85% in general. Unfortunately, Text2Onto cannot extract content events and the taxonomic relationship between content events from the web site usage data set.

Our experimental environment includes Intel Core2 2.4G, 4G RAM, Windows 2008 Server, J2SDK1.6. Testing data is from the following websites: http://www.aifb.uni-karlsruhe.de/Projekte and http://www.3ffire.com. CO1 and CO2 are the two corpuses after pre-processing [11], as shown in Table1.

The simulation experiment is to evaluate the extraction of the content events and their taxonomic relationship on CO1 and CO2. Precision, Recall and F. measure, which are widely adopted in the Information Extraction field, are used to evaluate the results of extraction. Precision is the percentage of the correctly extracted concepts in the all extracted concepts. Recall is the percentage of the extracted concepts in the all concepts of the corpus. F measure is the weighted geometric mean value of Precision and Recall. The calculation formulas are as Equation 6 to Equation 8.

$$Precision = \frac{Corrected_{extracted}}{all_{extracted}} \quad (6)$$

$$Recall = \frac{Corrected_{extracted}}{all_{extracted}} \quad (7)$$

$$F\text{-}measeure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

When extracting content events by swarm-intelligence-based page eigenvector cluster, the parameter a is stetted to 8. The experimental results are shown from Table 2 to Table 5.

Experimental results show that the accuracy of content

Corpus	Data set	Data Structures	The number of Component	Component dimensions
CO1	Page theme set	set	527	1
	log file	sequence	4653	1/31/6.571 (min/max/average)
	Request Set	tipple	55	2
	Page eigenvector space	VSM	32	527
CO2	Page theme set	set	1013	1
	transaction file	sequence	24533	1/27/8.879 (min/max/average)
	Request Set	tipple	1143	2
	Page eigenvector space	VSM	1116	1013

Table 1. The corpus for content event taxonomic relationship learning

	Correct extracted events	All extracted events	Reference	Precision log events	Recall	F-measure
EC	36	40	37	90%	97.3%	93.5%

Table 2. The evaluation of content events extraction from CO1

	Correct relationship	Total relationship	Reference relationship	Precision	Recall	F-measure
Taxonomic relationship of EC	21	27	24	77.8%	87.5%	82.4%

Table 3. The evaluation of content events taxonomic relationship extraction from CO1

	Correct events	Total events	Precision
Content events	1098	1116	98.4%

Table 4. The evaluation of content events extraction from CO2

	Correct relationship	Total relationship	Precision
Taxonomic relationship of EC	2147	2823	76.1%

Table 5. The evaluation of content events taxonomic relationship extraction from CO2

event extraction is more than 90% in different types of web sites. The reason is that the method of content event learning uses the static page URL mapping. The accuracy doesn't reach 100%, which due to the extraction of some nonsense pages from the site, such as "error.html".

The accuracy of content event taxonomic relationship extraction is between 75% and 80%, which is more than the results of other ontology learning method because of the swarm-intelligence-based page eigenvector cluster method. Due to the incompletion of page taxonomic relationship, the accuracy of content event taxonomic relationship extraction is also under 80%. The evaluation of recall and F-measure is satisfactory; because we use swarm similarity to the probability page subjects instead of web page cluster vector page eigenvector is limited to 8. In general, the experimental results show that this method can improve the results of ontology learning.

5. Conclusion

A taxonomic relationship learning approach for log ontology content event is proposed in this paper, which integrates web content mining and web usage mining. Based on site documentation, dynamic pages, user request set, Web log file and their own structure characteristics, this method makes use of swarm intelligence Web document cluster to extract the taxonomic relationship of content event. The results of this method can express the semantic features of the site usage data, which can meet users and domain experts' needs to mine the Semantic Web and can provide guidance for learning and generating Web content ontology. The simulation experiment results show that the method is both feasible and effective.

6. Acknowledgment

This work is supported by Key Projects in the National Science & Technology Pillar Program during the Eleventh Five-year Plan Period (NO.2009BAH46B0302).

References

[1] Thakur, M., Kumar, Y., Silakari, G. (2011). Query based Personalization in Semantic Web Mining. I*nternational Journal of Advanced Computer Science and Application*s, (2), 177-123.

[2] Gorodetsky, V., Samoylov, V., Serebryakov, S. (2010). Ontology based context dependent personalization technology, International Conference on Web Intelligence and Intelligent Agent Technology, 278-283.

[3] Vuljani, D., Rovan, L., Mirta Baranovi, (2010). Semantically enhanced web personalization approaches and techniques, 32nd IEEE International Conference on Information Technology Interfaces (ITA'10), 217-222

[4] Stojanovic, L., Stojanovic, N., Gonzalez, J. et al. (2003). OntoManager-a system for the usage-based ontology management. In: On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. Heidelberg: Springer Berlin, 858-875.

[5] Stojanovic, N., Gonzalez, J., Stojanovic, L. (2003). Ontologer-a system for usage-driven management of ontology-based information portals. In: International Conference On Knowledge Capture. New York: ACM, 172-179.

[6] Berendt, B., Hotho, A., Stumme, G. (2002). Towards Semantic Web mining. *In:* Proc. of the 1st International Semantic Web Conference on The Semantic Web (ISWC). London, UK: Springer, 264-278.

[7] Gacitua, R., Sawyer, P. (2008). Ensemble methods for ontology learning-an empirical experiment to evaluate combinations of concept acquisition techniques. In: Proc. of the 7th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2008). Los Alamitos: IEEE Computer Society Press, 328-333.

[8] Gulla, J. A., Brasethvik, T. (2008). A hybrid approach to ontology relationships learning. Accepted for 13th International Conference on Applications of Natural Language to Information Systems. London, 115-120.

[9] Annappa, B., Chandrasekaran, K., Shet, K. C. (2010). Meta-Level constructs incontent personalization of a web application , IEEE International conference on Computer & Communication Technology(ICCCT'10), 569-574.

[10] Soucy, P., Mineau, G. W. (2005). Beyond TFIDF weighting for text categorization in the vector space model. *In:* Proc. of IJCAI, 1130-1132.

[11] Wenshan, W., Haihua, L. (2010). Base on rough set of clustering algorithmin network education application , IEEE International Conference onComputer Application and System Modeling (ICCASM 2010), 481-483.