Journal of Southeast University (English Edition) Vol. 22, No. 3, pp. 426 – 429 Sept. 2006 ISSN 1003—7985

Web integration based on classification ontology

Gao Kening Ma Anxiang Zhang Bin

(School of Information and Engineering, Northeastern University, Shenyang 110004, China)

Abstract: In order to eliminate semantic heterogeneity and implement semantic combination in web information integration, the classification ontology is introduced into web information integration. It constructs a standard classification ontology based on web-glossary by extracting classified structures of websites and building mappings between them in order to get unified views. Mapping is defined by calculating concept subordinate matching degrees, concept associate matching degrees and concept dominate matching degrees. A web information integration system is realized, which can effectively solve the problem of classification semantic heterogeneity and implement the integration of web information source and the personal configuration of users.

Key words: information integration; classification ontology; ontology integration; personalization

With the rapid development of Internet/Intranet applications, helping users to obtain important information quickly has become an important service on the web. The web information integration system retrieves information from certain assigned web resources, and offers a unified access pattern to all sources. It also filtrates information according to user features. Challenges in building web integration systems mainly lie in information retrieval and multi-sources integration.

The data on the web is always for users to browse and it is organized in semi-structured html. How to obtain information in html has been widely studied in traditional web information retrieval. There are spider, crawler and robot for general information retrieval and wrapper for each special web source[1–2]. Multi-sources data integration is the core in the web integration system. It processes semantic heterogeneity and conflicts in different sources, and it generates global unified views of different sources in the end[3–4]. Ontology integration has become a focus in web integration because ontology can express semantics in an explicit and formal way. Ontology is used for auxiliary information retrieval[5–6] and global view building[7–8]. Suggested upper merged ontology (SUMO) is created by IEEE Standard Upper Ontology Working Group. The target of SUMO is to develop upper-knowledge ontology, facilitate data communion, information searching, automatic deduction and natural language process, and to provide a basis for building a special domain ontology[9].

In this paper, the web integration system based on classification ontology integration is studied and implemented. When retrieving information, revised-spider obtains site structure at the same time, and then information is classified according to the site structure. Multi-sources integration is processed as follows: a standard classification ontology based on SUMO is built at first, then a local classification ontology is obtained by a site classification system through standardization, and finally, integration of multi-sources is implemented by finding mapping between them.

1 Architecture of Web Integration System

The architecture of an ontology-integration-based web integration system is shown in Fig. 1. The main architecture consists of six modules: information retrieving, pre-processing, integration-processing, storage, presentation and user registration-modules.

The retrieving module consists of two parts: revised spider downloads web pages and records links between pages to obtain site structures. A page block-divided algorithm based navigation system retrieves classification systems used by the site.

The pre-processing module consists of a text extractor and a site structure based classifier. The text extractor extracts a document form html and the classifier classifies document according to its pages' position on the site.

The integration-processing module consists of a standard classification ontology generator, a local classification ontology generator, an ontology mapping and a personalized classification generator.

Received 2006-04-12.

Foundation item: The National Key Technologies R&D Program of China during the 10th Five-Year Plan Period (No. 2004BA721A05).

Biographies: Gao Kening(1963—), female, graduate; Zhang Bin(corresponding author), male, doctor, professor, zhangbin@mail.neu.edu.cn.

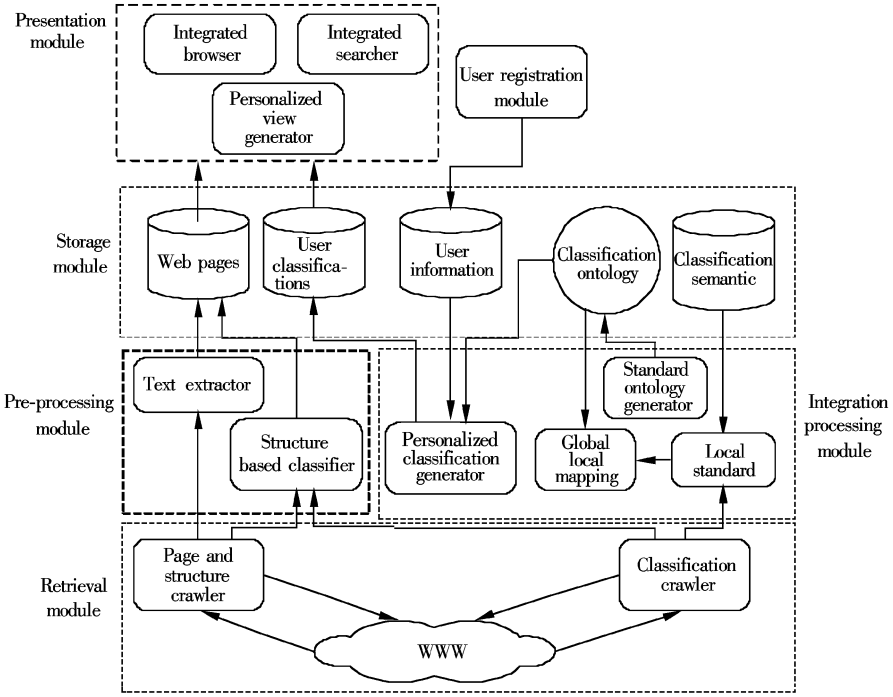

Fig. 1 Architecture of web integration system

The storage module includes five databases. A web page base stores documents extracted from pages and their classes; a classification semantic base stores the class semantic dictionary and thesaurus, which support local classification ontology building from site classification; a classification ontology base stores standard classification ontology; a user information base stores user registration information; and a user classification base stores personalized classification systems.

The presentation module generates a user view based upon a global unified view, using personalized classification filtrating and organizing information.

The registration module accepts user registration, and stores user information into the database.

2 Web Integration Based on Classification Ontology

After information retrieving and pre-processing, we get information which is stored in each site's classification system. We propose to build a SUMO based standard the classification ontology for web information at first; then to standardize the classification system of each source site, so that we can obtain a local classification ontology; finally, by mapping the local ontology to a standard classification ontology, a unified view organized by standard classification ontology is built.

2.1 Standard classification ontology

Information is organized in a certain classification system such as information quantity, objective and owner changing. Different classification systems are adopted. We build a standard web information classification ontology, which is referred as WCO later, based on SUMO. When building WCO, classification systems adopted by portal sites are given most consideration, among huge subjects in portal sites, news is our focus.

WCO consists of a core relationship and a core concept set. The core concept set contains: news, the root concept of WCO, is subordinate to the concept of content bearing objects in SUMO; subclasses of news are international, national, entertainment, and finance etc. The former two are scope-classifying classes, and the others are subject-classifying classes. Subject classes are not enumerated completely. They can be extended during system running.

For later processing convenience, the subject class concept is expressed by words in WordNet, which has the primary dictionary order; the scope class words use standard full names. Fig. 2 is the sketch map of WCO.

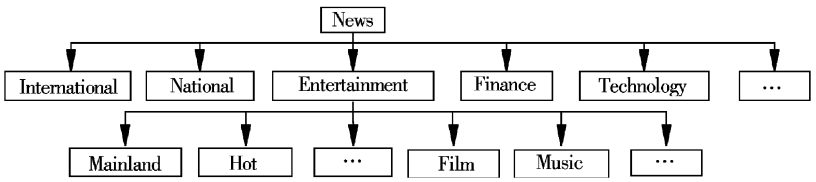

Fig. 2 Sketch map of WCO

2.2 Local classification ontology generation

Based on a web source site classification system, with the support of a classification semantic library, a local classification ontology can be built from site classification system standardization. So different expressions of the same concept in different site classification systems can be diminished. Different concepts in the same expression can also be diminished. We build a classification semantic library : add subject words and their synonym in WordNet into the library, which will be used in subject classes standardization; add time and space scope standard words and abbreviations into the library, which will be used for scope classes standardization.

With the semantic library built in the above way, site classification can be transformed into a local classification ontology in the following ways: for each class in the site classification system, if it is a subject class, we can find the synonym corresponding to it, then let the word with primary dictionary order in the synonym replace the class; if the class is a scope one, let its standard full name replace it. When the scope word cannot be found in library, we will extend the library.

2.3 Mapping local ontology to standard ontology

By defining the ontology concept formally, building the formula to compute a concept matching ratio, and setting the threshold of matching a ratio for the concept match, the mapping can be completed.

Both standard and local classification ontologisms are concept hierarchical, so the concept semantic consists of the concept itself and its position in the hierarchy. The concept semantic in the classification ontology is defined as quadruple CS = (name, F, S, C), where name is the expression word of the concept, F directs to father concept, S directs to set of sibling concepts, and C to set of child concepts.

Definition 1 Mapping rule: For concept C_1 in the local classification ontology, it is mapping with concept C_{gi}, if and only if Match (C_1, C_{gi}) = max(Match (C_1, C_g)), C_g refers to any concept in standard ontology.

Definition 2 String match function: A and B are the two string two test matching,

$$\text{StringMatch}(A, B) = \begin{cases} 1 & A, B \text{ are the same in length and have the same characteristics in the same positions} \\ 0 & \text{Otherwise} \end{cases} \quad (1)$$

Definition 3 Concept sets matching function: S_1 and S_2 are both concept sets. Suppose that radix of S_1 is n, radix of S_2 is m, and $m \geqslant n > 0$,

$$\text{CSetMatch}(S_1, S_2) = \frac{1}{n}\sum_{C_i \in S_1}\left(\sum_{C_j \in S_2}\text{StringMatch}(C_i.\text{name}, C_j.\text{name})\right) \quad (2)$$

where C_i. name refers to the expression of concept C_i. According to definition 2, the members in the set are different from each other, $\sum_{C_j \in S_2}\text{StringMatch}(C_i.\text{name}, C_j.\text{name}) \leqslant 1$, so $\text{CSetMatch}(S_1, S_2) \leqslant 1$.

For two concepts C_1, C_2, whose semantic quadruple CS_1 = (name1, F_1, S_{s1}, C_{s1}), CS_2 = (name2, F_2, S_{s2}, C_{s2}), in classification ontology, we give four concept match confidences and their formulae.

Definition 4 Concept name matching degree,

$$\text{MatchName}(C_1, C_2) = \text{StringMatch}(\text{name1}, \text{name2}) \quad (3)$$

Definition 5 Concept subordinate matching degree,

$$\text{MatchFather}(C_1, C_2) = \text{StringMatch}(F_1.\text{name}, F_2.\text{name}) \quad (4)$$

Definition 6 Concept associate matching degree,

$$\text{MatchSibling}(C_1, C_2) = \text{CSetMatch}(S_{s1}, S_{s2}) \quad (5)$$

Definition 7 Concept dominate matching degree,

$$\text{MatchChildren}(C_1, C_2) = \text{CSetMatch}(C_{s1}, C_{s2}) \quad (6)$$

Based on the above four formulae, match confidence of concepts C_1 and C_2 can be calculated as

$$\text{Match}(C_1, C_2) = \frac{1}{2}\text{MatchName}(C_1, C_2) + \frac{1}{5}\text{MatchFather}(C_1, C_2) + \frac{1}{10}\text{MatchSibling}(C_1, C_2) + \frac{1}{5}\text{MatchChildren}(C_1, C_2) \quad (7)$$

Following the above, a concept in the local classification ontology is mapped to concept in the standard classification ontology. Information in each source, which was once organized in the site classification, is organized in WCO, as WCO is an integrated and unified system, multi-source integration is complete.

3 Conclusion

We have designed and implemented the process of web integration based on ontology integration. The process is capable of multi-site integration. As ontology integration is still in a development and enhancement stage, much work needs to be done to make our

integration a fully integrated ontology, especially concept mapping ontology. We want to embody the mapping rule referred to in this paper into axioms of WCO, so mapping can be established by ontology deduction. There are many aspects which are needed to accumulate rules in the current system. How to accelerate the accumulating process also needs to be addressed in the future.

References

[1] Liu Fang, Yu Clement, Meng Weiyi. Personalized web search by mapping user queries to categories[A]. In: *Proceedings of the Eleventh International Conference on Information and Knowledge Management*[C]. New York: ACM Press, 2002. 558 - 565.

[2] Youns Hafri, Chabane Djeraba. High performance crawling system[A]. In: *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*[C]. New York: ACM Press, 2004. 299 - 306.

[3] Christophides Vassilis, Cluet Sophie, Simeon Jerome. On wrapping query languages and efficient XML integration [A]. In: *Proceedings of ACM SIGMOD Conference on Management of Data* [C]. New York: ACM Press, 2000. 141 - 152.

[4] Knoblock Craig A, Minton Steven, Ambite Jose Luis. The ARIADNE approach to web-based information integration[J]. *International Journal of Cooperative Information Systems*, 2001, **10**(1, 2): 145 - 169.

[5] Decker S, Erdmann M, Fensel D. Ontology based access to distributed and semi-structured information [A]. In: *Semantic Issues in Multimedia Systems*[C]. Boston: Kluwer Academic Publisher, 1999: 351 - 369.

[6] Xu Youzhi, Shen Jie, Chen Zhimin. Ontology-based information retrieval of web services in virtual enterprise [A]. In: *Proceedings of the IEEE International Conference on Services Computing* [C]. Shanghai, China, 2004. 441 - 444.

[7] Corradini F, Mariani L, Merelli E. An agent-based approach to tool integration [J]. *Journal of Software Tools Technology Transfer*, 2004, **6**(3): 231 - 244.

[8] Li Shanping, Yin Qiwei, Hu Yujie, et al. Overview of researches on ontology [J]. *Journal of Computer Research and Development*, 2004, **41**(7): 1041 - 1052. (in Chinese)

[9] Niles Ian, Pease Adam. Towards a standard upper ontology[A]. In: *Proceedings of the Second Conference on Formal Ontology in Information Systems*[C]. New York: ACM Press, 2001. 2 - 9.

基于分类本体的 web 集成

高克宁　马安香　张　斌

(东北大学信息科学与工程学院,沈阳 110004)

摘要: 在 web 信息集成领域,为消除语义异构、实现语义融合,将分类本体引入 WWW 信息集成,设计了一种基于本体集成的 web 信息集成系统. 通过构建标准分类本体以获取局部分类本体,并建立二者间的映射,以获得多源统一视图. 通过计算概念间的统领匹配度、关联匹配度、从属匹配度来完成概念的映射. 实现了基于分类本体的 web 信息集成系统,该系统能很好地解决 web 信息分类语义异构问题,并能实现多 web 信息源的集成以及用户个性化定制.

关键词: 信息集成; 分类本体; 本体集成; 个性化

中图分类号: TP393