JOURNAL OF KOREA MULTIMEDIA SOCIETY, VOL. 10, NO. 6, JUNE 2007(pp. 754-759)

A Web-Based Domain Ontology Construction Modelling and Application in the Wetland Domain

Jun Xing[†], Min Han[††]

ABSTRACT

Methodology of ontology building based on Web resources will not only reduce significantly the ontology construction period, but also enhance the quality of the ontology. Remarkable progress has been achieved in this regard, but they encounter similar difficulties, such as the Web data extraction and knowledge acquisition. This paper researches on the characteristics of ontology construction data, including dynamics, largeness, variation and openness and other features, and the fundamental issue of ontology construction - formalized representation method. Then, the key technologies used in and the difficulties with ontology construction are summarized. A software Model-OntoMaker (Ontology Maker) is designed. The model is innovative in two regards: (1) the improvement of generality: the meta learning machine will dynamically pick appropriate ontology learning methodologies for data of different domains, thus optimizing the results; (2) the merged processing of (semi-) structural and non-structural data. In addition, as known to all wetland researchers, information sharing is vital to wetland exploitation and protection, while wetland ontology construction is the basic task for information sharing. OntoMaker constructs the wetland ontologies, and the model in this work can also be referred to other environmental domains.

Keywords: Ontology learning, Web mining, Domain ontology, Wetland protection, Meta learning machine

1. INTRODUCTION

Ontology describes knowledge in a formalized language that can be understood by machines, and aims to ultimately remove the communication barrier of information and knowledge between human and machine, between machines themselves. Hence, it is widely applied in every field that involves the communication of information and knowledge. The typical cases range from Medicine, Electronic Commerce, Real Estate, Financial Accounting, Manufacturing, to Environment. In environmental domains, Pruvis [1] applied ontology in the NZDIS project; Ceccaroni and Cortés [2] adopts ontology to build an EDSS system. Within the environmental domains, wetland protection has always been much emphasized; Zacharias [3] attempts to protect wetland by means of GIS, RS, etc. Can ontological technologies be also applied to protect wetlands? The answer is a definite yes.

The goal of this work is to automatically acquire the domain terminology and the mutual relation from Web documents, and then to construct ontology based on the obtained concepts and their relations. To solve the difficulties involved in the automated ontology construction methods, a Web-based domain ontology automated construction model is designed - OntoMaker. This model fully explores the enormous information on WWW network sources, and applies automated (also semi-automated) data mining technologies in the construction of ontology. This methodology

※ Corresponding Author : Jun Xing, Address : School of Electronic and Information Engineering Dalian University of Technology, Dalian 116023, P.R.China, TEL : +82-411-84707847, FAX : +82-411-84707417, E-mail : xingjun_tom@tom.com

Receipt date : May. 28, 2007, Approval date : Jun. 29, 2007.

† School of Electronic and Information Engineering, Dalian University of Technology

†† School of Information Science and Engineering, Dalian Institute of Light Industry(E-mail : minhan@dlut.edu.cn)

※ This research is supported by the project (50139020) of the National Natural Science Foundation of China. The support is appreciated.

has its unique features; therefore, it should not simply reproduce the normal ontology construction methods, or it merely apply the data mining technologies. To solve this issue, the further research on the web data characteristics and the formats presentation methods of ontology is firstly carried out. Concerning the Web-based ontology automated construction, after the exploration of large volume of documents, we systematically sort and conclude the data source features, ontology form presentation, the research status and their research core issues. The specific contributions of this work include:

(1) The improvement of generality. Meta Learning Machine mechanism is adopted to dynamically cite different ontology learning methods in different domain data. The results can reach optimum.

(2) The combined processing of structural, semi-structural and non-structural data. Via the structural (also semi-) data analysis over the web pages, the linkage structure of the web pages is firstly recorded, and then they are treated as normal text. This approach preserves the consistency of data processing and also avoids the loss of useful information in the data.

(3) Wetland, the third natural resources on earth, titled as "the kidney of the earth", has received worldwide attention. As researchers in wetland, we have carried out substantive works on the wetland protection and exploration [4,5]. To use and protect wetland, the information sharing is a vital groundwork in the construction of wetland ontology. We have attempted to manually construct wetland domain ontology and achieved the basic stages of the project. Now we try to explore the automation of ontology building with the development of software, which is named OntoMaker. The model can not only be applied in wetland protection, but also be used in other projects in environmental domains.

The organization of this work follows the logic flow of what to learn, where to learn and how it may be learnt. Section 2 focuses on ontology and data source, mainly concerning the issues of input and output, that is, what to learn and where to learn it. Section 3 depicts the architecture of OntoMaker and its core technologies. Section 4 presents the application of OntoMaker in the domain of wetland. Section 5 gives the conclusions.

2. DATA ONTOLOGY AND DATA SOURCE

To successfully solve the learning task of ontology, it must be determined what to learn, where to learn and how to learn. This section addresses the two former issues firstly.

2.1 Ontology

The ultimate goal of learning is to construct ontology. The major work includes identifying what is ontology, what components ontology cover, and how ontology can be understood by computers. Neches et al. [6] were the first to give the definition of ontology. The later definition by Studer [7] was the most widely accepted in the field of artificial intelligence; Perez [8] on the ground of Studer's definition, concluded the five basic modeling meta language for depicting ontology: classes & concepts, relation, functions, axioms, instances.

2.2 Web Data Sources

Among diverse media, the Internet has the largest stock of information. As indicated by the number searched by google search engine, Web has over 8 billion web pages of information distribution space. It is also the most frequently updated information source, nearly synchronized with real world understanding development. It truly mirrors the reality via text (including multimedia). The essence of ontology is to depict the real world. Its representation gives the reflection of objective reality, independent of the language, thus the best

reflection of the objective real world. This matches the information provided by the Internet perfectly. Hence, it is the best choice as the information source for automated ontology construction.

3. ONTOMAKER MODEL AND THE CORE TECHNOLOGIES

Based on the above analysis, the Structure design of OntoMaker is shown in Fig. 1. It includes:

- preliminary Ontology Building: the preliminary ontology is structured by experts in the domain. Its function includes: (I) providing the searching key words for the search engine; (II) serving as the basis for objective ontology expansion; (III) serving as the reference standards for classification and clustering.
- Data Source Collecting: to download related web pages from Internet, based on the domain expert advised elementary ontology. Two methods can be applied. (I) using Crawler technologies to download web pages; (II) using the API developer package from Google to search for related sites, and downloading user selected pages. OntoMaker adopts the latter approach.
- Data Pre-processing: to transform the formalized presentation of the downloaded web data sets, allowing computers to participate in the processing. This covers two aspects: merger and two times dimension reducing. It is assured then that the processed data would keep the integrity of the information and the minimized information loss.
- Meta Learning Machine: to make the software generally applicable, the classification and clustering algorithm must bear the nature of dynamic flexibility for various data sets. This part is difficult and crucial.
- Web Page Classification: with reference to the classification algorithm selected by the meta learning machine and the conceptual standards in the elementary ontology, the task of web page classification can be implemented.
- Web Page Clustering: with reference to the clustering algorithm selected by the meta learning machine, the classified data sets are re-classified. As the preliminary ontology does not involve relations, this process is a clustering process.
- Concept and Instance Extraction: to calculate the feature value of each clustering as the labeled concepts; thereafter, to determining concepts with mutual information methodology; to recognize the knots of leaves as instances via Wordnet.

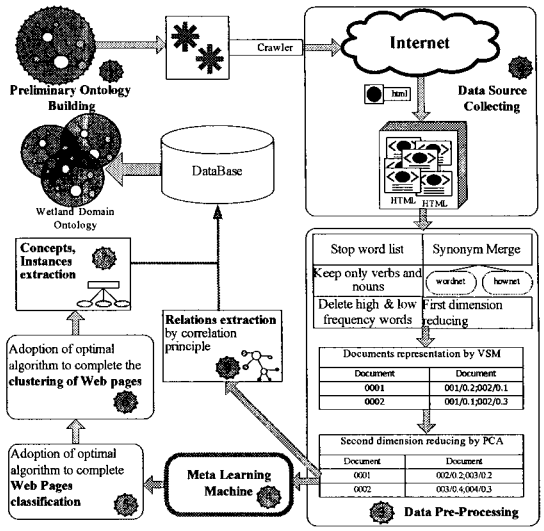

Fig. 1. Architecture of the OntoMaker.

We will illustrate below the key part of the model: meta learning machine.

To guarantee the software generally applicable, the selection of classification and clustering algorithms must be dynamically flexible when processing various data sets. This part is crucial and challenging. OntoMaker offers a learning machine that can find the most suitable classification and clustering algorithms. OntoMaker firstly calculates the data features of the data sets, and then takes into account the data features and quality requirements, finally via a case-based learning method - k-NN algorithm to select the optimal algorithm. The reason for why select k-NN algorithm is that, in practice, the relative performance of the learning

machine when processing different training data cannot be easily captured, especially when the data are gradually enlarged. Other methods fall short in solving this problem. K-NN algorithm solves this problem well. In this work, measurement based on statistics and information theory is applied to model the features of the data sets; the concrete definition can be referred to reference [9]. The data features set in this work includes:

DC1: the total number of data records, denoting the extension potential of the algorithms;

DC2: the ratio of numeric attributes to all attributes, denoting the processing capacity of algorithms for various numeric attributes and text attributer;

DC3: the ratio of missing data, denoting the processing capacity of algorithms for missing data;

DC4: the ratio of partially described data. The partial description of data is a phenomenon that captures recent attentions as one of the new cases of data loss. This ratio denotes the processing capacity of algorithms for partially described data brought by incomplete semantics;

DC5: the ratio of abnormally distributed attributes to all attributes, denoting processing capacity of algorithms for abnormally distributed data;

DC6: the entropy distribution of the values, denoting the complexity level of the problems;

DC7: the average mutual information of values and attributes, measuring the level of meaningful information contained in text attributes;

The flow chart is shown in Fig. 2: The existing meta leaning machine can only cope with single quality factor (usually as the degree of certainty). Pavel [9] discussed the case of two quality factors (degree of certainty and time), but the processing is too complicated. OntoMaker applies k-NN algorithm to perform the learning. The meta learning issue can describe the following: given a stock of algorithms and a set of data F, whose feature vector is $DC^F=(DC_1^F,DC_2^F,...,DC_8^F)$, assuming the quality requirements set by the users are in the vector

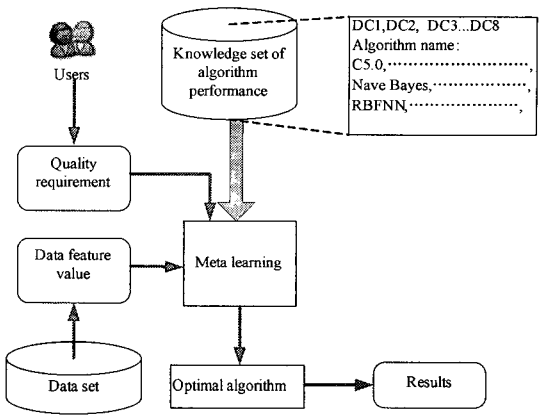

Fig. 2. Meta learning machine structure.

$QF^F=(QF_1^F,QF_2^F,...,QF_{10}^F)$, quality factor weight vector as $W=(W_1^Q,W_2^Q,...,W_{10}^Q)$, $0\le W_i^Q\le 1$, $\sum_1^{10}W_i^Q=1$; If the user has no specific requirements for the quality factors QF_i or they only give some constraints, then $W_i^Q=0$; the distance between data set F and F′ is $Dist(F,F')=\sum_{i=1}^{8}\frac{|DC_i^F-DC_i^{F'}|}{\max(DC_i)-\min(DC_i)}$. The goal of the meta learning is to find k units of record $\vec{R}$ in the algorithm performance knowledge stock, where data set is $F\vec{R}$, and quality vector is $QF^{F\vec{R}}$, meeting users' quality requirements, while minimizing average $Dist(F,F\vec{R})$ and $|W\cdot(QF^F-QF^{F\vec{R}})^T|$, T denoting the transformation of the matrix. Firstly, we merge the quality vector and the data set features vector into a new vector, then we applies k-NN algorithm for overall decision-making. The merged distance function is:

$$fDist(F,R\vec{R})=\bar{w}\cdot Dist(F,F\vec{R})+(1-\bar{w})\cdot|W\cdot(QF^F-QF^{F\vec{R}})^T| \quad (1)$$

Here in, $\bar{w}$ denotes the ratio of data set similarity to quality similarity in the whole decision-making. Two situations might happen: (1) data features are similar, but the quality does not meet the requirements; (2) the data set feature vary greatly but the quality meets the requirements well. The setting of $\bar{w}$ value can decide to give priority to (1) or (2). The above process offers a satisfied solution for selecting algorithms for the meta learning machine.

4. THE CONSTRUCTION OF WETLAND PROTECTION ONTOLOGY BY ONTO-AKER

4.1 Preliminary Wetland Protection Ontology Building

As shown in Fig. 3, the construction of a wetland preliminary ontology should incorporate wetland types, wetland resources, wetland destruction factors, wetland protection measures, and zoology character change, and other concepts.

4.2 Data Source Collection

Referring to the keywords provided by the wetland preliminary ontology, related web pages are downloaded from the Internet. Due to the size of paper page, we experiment with a medium pool of pages. Fig. 4 shows a sample of 100 related websites, and the downloaded 3,000 web pages.

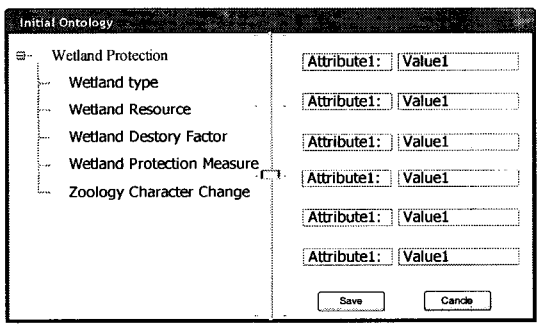

Fig. 3. Preliminary wetland protection ontology.

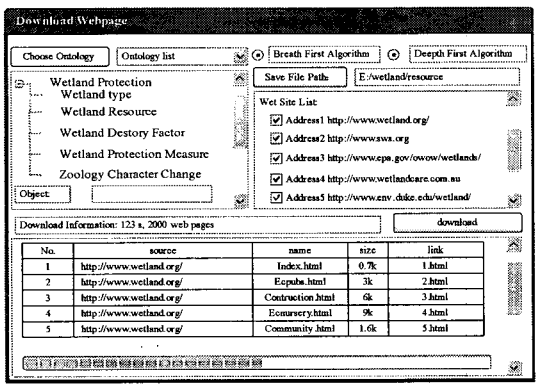

Fig. 4. Data source collection.

4.3 Ontology Construction

First of all, the data of 3, 000 web pages from over 100 sites are pre-processed, and a 300-dimension vector space matrix is obtained. Then, the meta learning machine performs the filtering and selection; classification is performed by the algorithm of Navie-Bayes and clustering is performed by the algorithm of K-means. Finally, the feature values of the clustering results are extracted by the methodology of mutual information, that is, the extraction of concepts. The results are shown in Fig. 5.

5. CONCLUSIONS

This work, based on the researches and reported by current international colleagues, proposes a Web-based ontology construction model - OntoMaker, and demonstrates in detail the related concepts, technologies, core subjects and methodologies. The research on Web-based ontology construction will exert great influence on the practical application of ontologies, especially for the instant realization of next generation WWW web. The Web-based ontology construction has just started its germinal stage, and needs further research. Yet, along with the development of Web data mining technologies and the standardization of ontology formalized description, the progress

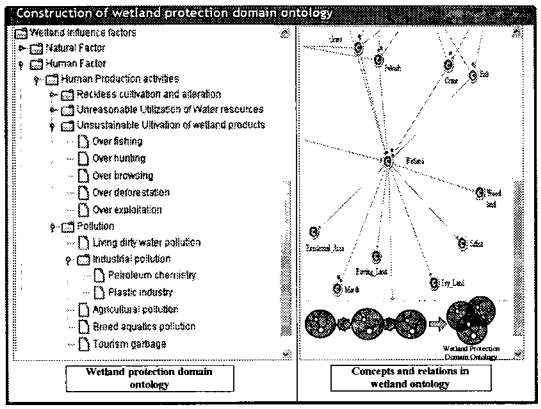

Fig. 5. Wetland concepts by OntoMaker.

can be accelerated. We will continue to perfect OntoMaker system, especially over the problem of relation extraction. In addition, we will continue to study on the wetland ontology construction and its practical use.

REFERENCES

[1] M. Purvis, S. Cranefield, R. Ward, M. Nowostawski, D. Carter, and G. Bush, "A multi-agent system for the integration of distributed environmental information," *Environmental Modelling & Software*, Vol. 18, No. 6. pp. 565-572, 2003.

[2] L. Ceccaroni, U. Cortés, and M. Sànchez-Marrè, "OntoWEDSS: augmenting environmental decision-support systems with ontologies," *Environmental Modelling & Software*, Vol. 19, No. 9. pp. 785-797, 2004.

[3] I. Zacharias, E. Dimitriou, and Th. Koussouris, "Integrated water management scenarios for wetland protection: application in Trichonis Lake," *Environmental Modelling & Software*, Vol. 20, No. 2, pp. 177-185, 2005.

[4] M. Han, L. Cheng, and Q.Liu, "Marsh information extraction based on knowledge discovery," *Remote Sensing for Land & Resources*, Vol. 1. pp. 43-47, 2004.

[5] M. Han, L. Cheng, and H. Meng, "Application of four-layer neural network on information extraction," *Neural Networks*, Vol. 16. pp. 547-553, 2003.

[6] R. Neches, R.E.Fikes, T. R. Gruber, et al., "Enabling Technology for Knowledge Sharing," *AI Magazine*, Vol. 12, No. 3, pp. 36-56, 1991.

[7] R. Studer, VR. Benjamins, and D. Fensel, "Knowledge Engineering, Principles and Methods," *Data and Knowledge Engineering*, Vol. 25, No. 1-2. pp. 161-197, 1998.

[8] A.Gomez-Perez and Benjamins, "Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem-Solving Methods," *Proc. of IJCAI99's Workshop on Ontologies and Problem Solving Methods:Lessons Learned and Future Trends. San Francisco: Morgan Kaufmann*, pp. 65-78, 1999.

[9] B. B.Pavel, S. Carlos, and P.D.C Joaquim, "Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results," *Machine Learning*, Vol. 50, No. 3. pp. 251-277, 2003.

Xing, Jun

9. 2004. now Dalian University of Technology, Dalian, P.R. China Degree: Ph.D. candidate

3. 1999~07. 1996 Northeastern University, Shenyang, P.R. China Degree: Master of Science

9. 1996~07.1992 Northeastern University, Shenyang, P.R.China Degree: Bachelor of Science

Since 3. 2004 Lecturer, School of Information Science & Engineering, Dalian Polytechnic University, Dalian, P.R.China

Main research activity: 3S system, Ontology Building and Web data mining.

Han, Min

9. 1996~7. 1999 Kyushu University, Fukuoka, Japan Degree: Ph.D.

9. 1990~7. 1993 Dalian University of Technology, LiaoNing, P.R.China. Degree: Master of Science

9. 1982~07. 1986 Dalian University of Technology, LiaoNing, P.R.China. Degree: Bachelor of Science

Since 9. 2004 Professor, School of Electronic and Information Engineering, Dalian University of Technology, Dalian, P.R.China

Main research activity: 3S system, expert system, Neural network and chaos. She is a senior member of IEEE.