WEB MINING - THE ONTOLOGY APPROACH

Ee-Peng Lim[1)] and Aixin Sun[2)]

[1)] School of Computer Engineering
Nanyang Technological University, Singapore
Email: aseplim@ntu.edu.sg

[2)] School of Computer Science and Engineering
University of New South Wales, Sydney, Australia
Email: aixinsun@cse.unsw.edu.au

ABSTRACT

The World Wide Web today provides users access to extremely large number of Web sites many of which contain information of education and commercial values. Due to the unstructured and semi-structured nature of Web pages and the design idiosyncrasy of Web sites, it is a challenging task to develop digital libraries for organizing and managing digital content from the Web. Web mining research, in its last 10 years, has on the other hand made significant progress in categorizing and extracting content from the Web. In this paper, we represent ontology as a set of concepts and their inter-relationships relevant to some knowledge domain. The knowledge provided by ontology is extremely useful in defining the structure and scope for mining Web content. We will therefore review Web mining and describe the ontology approach to Web mining. The application of these Web mining techniques to digital library systems will also be discussed.

INTRODUCTION

Motivation

The ubiquity of Web can be characterized by the enormous volume and coverage of Web content, the phenomenal number of Web users and businesses, the vast number of computers and devices accessing Web, and the large number of Web-based applications. A survey conducted by OCLC in 2002 revealed that there were 3 million public Websites and 1.4 billion Web pages at that point in time[1]. A ubiquitous Web has certainly led to some fundamental changes to the design of digital libraries. Among them is the search behavior of digital library users. Users today perform more searches using Web search engines than OPAC systems. Increasingly, Web is the preferred or de facto source of information. A May 2004 survey by Nielsen reported that an average surfer went online 30 times for more than 24 hours in total during a month[1]. The ubiquity of Web offers some obvious explanations, namely:

- The coverage of Web content is so large that it is difficult for any traditional digital libraries to match;
- The ability of to browse Web content directly on the users' computers and the ease of downloading them is clearly a big draw; and
- The availability of Web search engine (e.g., Google[2]) and Web directories (e.g., Yahoo![3], DMOZ[4]) has helped tremendously simplified the process of searching Web content.

Nevertheless, Web content is not always easy to use. Due to the unstructured and semi-structured nature of Web pages and the design idiosyncrasy of Web sites, it is a challenging task to develop digital libraries for organizing and managing digital content from the Web. Berners-Lee et al. therefore introduced the idea of

[1] http://www.caslon.com.au/index.htm
[2] www.google.com
[3] www.yahoo.com
[4] www.dmoz.com

Semantic Web which refers to the construction of a machine-understandable semantic layer over the existing Web content so as to support better information processing and Web services [2]. While Semantic Web may take several years to realise, digital library researchers are turning to Web mining techniques to improve the accessibility of Web content[3]. The well established Web mining techniques include Web classification [4] and Web extraction [5,6].

Objectives

Web mining techniques have shown promising performance in research experiments. Their actual deployment in live Web data, in contrast, has been fairly limited due to a lack of background semantics required for processing the text data, links, and other elements in Web pages. In this respect, an *ontology* which gives a conceptual description of the background semantics can serve as a very useful input to the Web mining problems [7]. An ontology refers to a set of concepts and the relationships, together known as ontology entities, describing the information within an application domain. When an ontology is used in solving a Web classification or extraction problem, the results obtained can be associated with the ontology entities making them easier to understand. This is a big advantage because each ontology often represents knowledge agreed upon by users and applications of a domain. For example, within the University domain, {Professor, Student, Course} and {Teach, Register, Supervise} are the common concepts and relationships respectively. University Web pages are likely to centered around these concepts and related concept instances are likely to be linked in one way or another.

As the languages for defining ontologies and using the latter in marking up Web content become well accepted [8], we see an increasing use of ontology in Web mining. In this paper, we will give an overview of *ontology-based Web mining*. In ontology-based Web mining, we are often interested in discovering the instances of concepts and relationships in a given ontology, or using them to discover other useful knowledge. These Web mining techniques can potentially be deployed in a digital library system to enhance the access to Web content. This paper will later present our research on *homepage mining* and *homepage relationship mining* where homepages representing instances of concepts and pairs of homepages representing instances of relationships are to be mined respectively [9,10].

Paper Outline

The rest of the paper is organised as follows. The definition of ontology will first be given. We will then elaborate on ontology-based Web mining. Following that, our research in homepage mining and homepage relationship mining will be described. Finally, we give a conclusion of the paper.

WHAT IS ONTOLOGY?

The term ontology can be defined in many different ways. Genersereth and Nilsson defined an ontology as an explicit specification of a set of of objects, concepts, and other entities that are presumed to exist in some area of interest and the relationships that hold them [11]. As implied by the above general definition, an ontology is domain dependent and it is designed to be shared and reusable. Usually, ontologies are defined to consist of abstract concepts and relationships (or properties) only. In some rare cases, ontologies are defined to also include instances of concepts and relationships[12].

To solve a problem using ontology, a formal definition is required. For the purpose of this paper, we define an ontology to be a set of concepts *C* and relationships *R*. The relationships in *R* can be either *taxonomic* or *non-taxonomic*. For example, Figure 1 depicts a simple University ontology consisting of a set of concepts C_{univ} = *{Person, Faculty, Staff, Student, Department, Project, Course}*, and a set of relationships R_{univ} = *{Department_Of(Person, Department), Member_Of(Person, Project), Instructor_Of(Course, Person), Superclass_Of(Faculty, Person), Superclass_Of(Staff, Person), Superclass_Of(Student, Person)}*. *Superclass_Of* represents the taxonomic relationship while the rest are not. With this definition, the instances of an ontology refer to the instances of its concepts and relationships. If each concept instance exists in the form of a Web page, a relationship instance will then exist in the form of a Web page pair. This view has been adopted in most the Web classification research. On the other hand, if each concept instance exists in

the form an HTML element, a relationship instance will then exist in the form of an HTML element pair. This alternative view is usually adopted in Web extraction research. It is noted that other forms or hybrid forms of concept instances may also exist for some Websites.

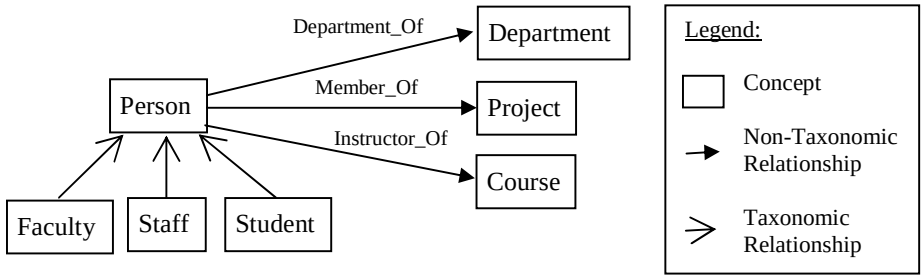

Fig. 1 An University Ontology Example

The construction of an ontology itself is an ongoing research topic. The construction process can be manual with the help of some ontology editing tools, e.g. OntoEdit [13], or automatic using a collection of training documents, e.g.OntoLearn[14]. There are also different languages for representing an ontology so as to support exchanges of ontology knowledge. Su and Ilebrekke gave a good description and comparison of these languages. In this paper, we assume that an ontology is given to conduct Web mining. Since knowledge exchange is outside the scope of our discussion, we do not intend to use any ontology representation language in this paper.

OVERVIEW OF ONTOLOGY-BASED WEB MINING

Overview of Web Mining

Web mining refers to the discovery of knowledge from Web data that include Web pages, media objects on the Web, Web links, Web log data, and other data generated by the usage of Web data. Kosala and Blockeel classified Web mining into: (a) *Web content mining*, (b) *Web structure mining* and (c) *Web usage mining* [3]. Web content mining refers to mining knowledge from Web pages and other Web objects. Web structure mining refers to mining knowledge about link structure connecting Web pages and other Web objects. Web usage mining refers to the mining of usage patterns of Web pages found among users accessing a Website. Among the three, Web content mining is perhaps studied most extensively due to the prior work in text mining. The traditional topics covered by Web content mining include:

- *Web page classification*: This involves the classification of Web pages under some pre-defined categories that may be organised in a tree or other structures.
- *Web clustering*: This involves the grouping of Web pages based on the similarities among them. Each resultant group should have similar Web pages while Web pages from different resultant groups should be dissimilar.
- *Web extraction*: This involves extracting HTML elements, term phrases, or tuples from Web pages that represent some required concept instances, e.g., person names, location names, book records, etc..

Web Mining and Ontologies

In all the above types of Web mining, ontologies can be applied in the following two general approaches:

- *When both ontology and the instances of ontology entities are known*: This usually applies to cases where instances of ontology have been identified among the input Web data. With this additional data semantics, Web mining techniques can discover knowledge that is more meaningful to the users. For example, *ontology-based Web clustering* can use HTML elements corresponding to concept instances as features to derive more accurate clusters. If Web pages are concept instances, *ontology-based Web site structure mining* can derive linkage pattern among concepts from Web pages for Website design

improvements.

- *When only ontology are available as input semantic structures*: Ontologies can also be used as background semantic structures for Web mining. For example, instead of categorizing Web pages into categories, *ontology-based Web page classification* may classify Web pages as concept instances and Web page pairs as relationship instances. This allows Web pages to be searched using more expressive search queries involving search conditions on concepts and/or relationships. In *ontology-based Web extraction*, one may address the problem of extracting both HTML elements as concept instances and finding related pairs of HTML elements.

Digital Library Applications of Ontology-based Web Mining

Ontology-based Web mining, like traditional Web mining, is useful to many different digital library applications. We can group these applications under the following classes:

- *Improved search to Web data*: With additional ontological semantics, Web data can be indexed by their concepts and relationships to support expressive search queries. For example, using the University ontology in Figure 1, we can query faculty member information working on digital library projects by assigning query term "digital library" to the project concept and specifying that faculty related to the qualified projects to be returned in the query results. Such queries may resemble structured database queries except that the data to be dealt with are Web pages. A more expressive query model can support very precise information search and reduce the amount of irrelevant Web information in the results [15].

- *Better browsing capabilities*: Similar to searching, Web pages can be browsed based on their ontology concepts and relationships instead of following Web links only. If Web pages are the concept instances, relationship instances can be created as some virtual links between Web pages. Other than selecting Web pages belonging to concepts of interest, one can thus navigate the virtual links between Web pages enriching the browsing experience in digital library applications[15]. On the other hand, if some text elements are identified as concept instances and their relationships are extracted, they can also be marked up in Web pages to direct user attentions to the more important text passages.

- *Personalization of Web data access*: Personalization aims to find a subset of Web data that matches the interest profile of a user or a group of users. This can be achieved by recommending Web pages or Websites to the user(s), or by filtering Web pages that are of interest to the user(s). For example, this can done by analysing the historical data recording user accesses to Web data, and mining the topics relevant to a user by clustering previously accessed Web pages based on content similarities. When a new Web page is found to be similar to one of the clusters, it can be routed to the user. As Web pages are annotated with ontology entity labels, the grouping of Web pages accessed by a user can be more effectively done leading to more effective content recommendation.

HOMEPAGE MINING AND HOMEPAGE RELATIONSHIP MINING

Homepage mining and homepage relationship mining are two related kinds of ontology-based Web content mining where an input ontology is provided as the domain specific knowledge structure [9,10]. In homepage mining, one aims to find homepages representing concept instances. As the name suggests, homepage relationship mining refers the discovery of homepage pairs as related concept instances.

Web Unit-Based Homepage Mining

In homepage mining, it is assumed that some Web pages are designated as homepages of concept instances and they provide the links to other Web pages that supplement the content in homepages. The latter is known as the *support pages*. Consider the Website of some university, there are usually homepages created for university departments, faculty members, students, courses, and other concepts in the University ontology example. These homepages are also often the targets of Web queries and therefore important to mine. An obvious approach to solve homepage mining problem is to model it as a Web page classification problem.

Nevertheless, Web page classification mainly assigns one or more concept labels to every Web page based on its own content. It does not really consider the context of the Web page consisting of other neigboring Web pages so as to determine if it is a homepage.

In homepage mining, due to the role of homepages and their support pages, we can construct for each homepage a *Web unit* to represent the complete information of the corresponding concept instance. *A Web unit consists of exactly one homepage (also called a key page in [9]) and zero or more support pages.* With this definition, homepage mining can be formulated as a problem of finding Web units representing concept instances. Once Web units are found, so are the corresponding homepages. We therefore call this *Web unit-based homepage mining*.

Figure 2 depicts the Web unit of the G-Portal research project. It consists of a homepage linking to several support pages.

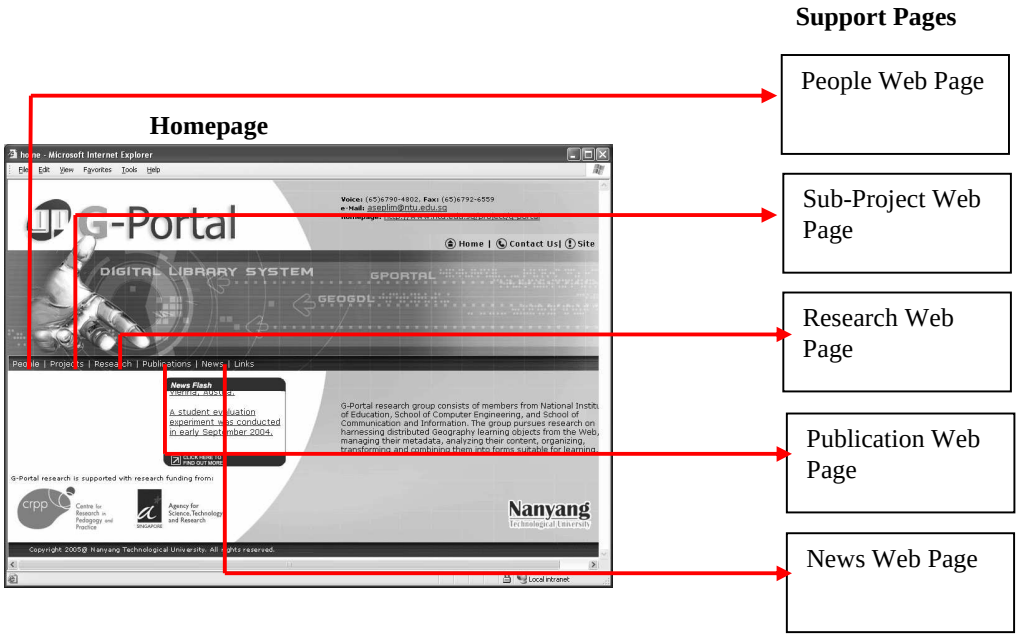

Fig. 2 A Web Unit Example

Web unit-based homepage mining has two main tasks: finding the set of Web pages forming each Web unit, and assigning the Web unit a concept label. The two tasks are complicated by the following two factors:

- The notion of Web unit is subjective as the determination of homepage and support pages of a Web unit may be different for different users.
- Support pages may not always be linked (directly or indirectly) from homepages. For example, some Web pages containing homework assignments may not be linked from a course's homepage and this has some implications in the design of Web unit-based homepage mining methods.

In [9], we proposed an iterative Web unit mining algorithm (iWUM) to constuct and classify Web units. Once the Web units are mined, the homepages will also be discovered. The iWUM algorithm is designed based on the following observations:

- Web pages from the same file directories also known as Web folders are more semantically related than Web pages from different Web folders.
- Support pages of a Web unit are usually reachable from the homepage through some intra-unit links.
- The homepage of a Web unit is usually found at the highest level Web folder containing the pages of the Web unit.
- Two Web units corresponding to instances of the same concept with no recursive relationship seldom have direct links between them.

- Multi-page Web units of the same concept often reside in a set of folders, one for each Web unit and the folders are directly under a common parent folder.
- The homepages of Web units of the same concept are often the link targets of a hub page which may be found at: (a) the folder where the Web units are located if the latter are one-page Web units; (b) the parent (or ancestor) folder of the folder(s) where the Web units are located if the latter are multi-paged.

As depicted in Figure 3, the steps of iWUM can be divided into two main phases, namely the *Web fragment generation phase* and the *Web unit merging phase*. *Web fragments* are Web units or parts of Web units. Like Web unit, a Web fragment consists of a key page and zero or more support pages. However, the key page is not necessary a homepage. iWUM essentially constructs Web fragments and assigns them concept labels, and iteratively combines labelled Web fragments into larger Web units until all the Web units are formed.

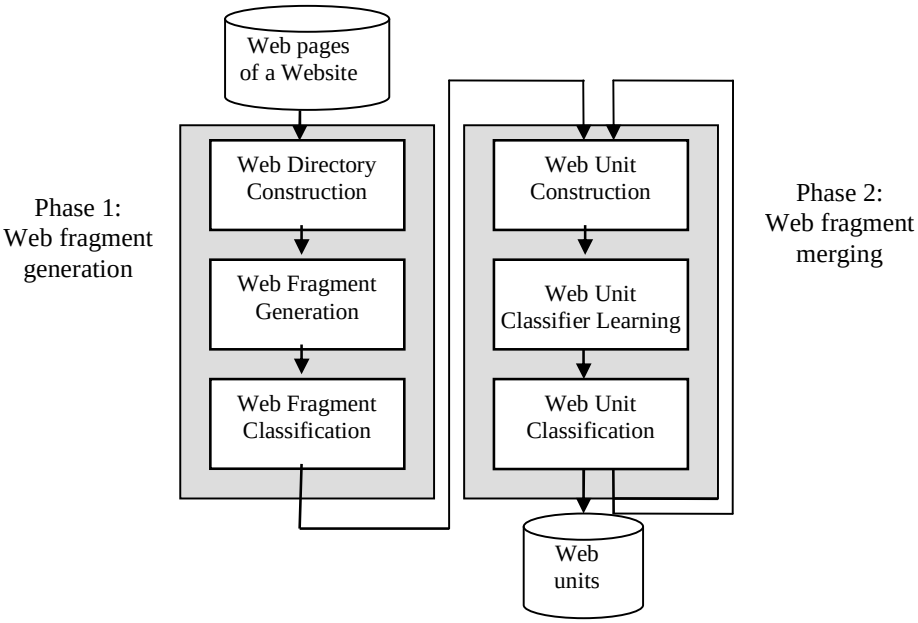

Fig. 3 Iterative Web Unit Mining (iWUM) Method

In Web directory construction, iWUM builds a tree of folders and Web pages from the URLs of Web pages. The generation of Web fragments is done by computing the *connectivity index* of each Web folder so as to indicate the extent to which the Web pages and sub-folders directly under the folder are connected. The more connected they are, the higher is the connectivity index. Starting from the Web folder with the least connectivity index, iWUM constructs Web fragments with some heuristics to determining key and support pages. These Web fragments are then classified with concept labels. Here, the classifier of each concept can be trained using Web units from a Website designated for training. The Web fragments with and without concept label assignments are then used to construct Web units in the Web unit construction step. This step involves merging an unlabelled Web fragment with a labelled Web fragment to form a larger Web unit. Web unit classification further uses the content features and Website structure features to reassign concept labels to all Web units. With new concept label assignments, larger Web units are constructed again and the process repeats until the changes to Web unit labels are minimal.

iWUM has been applied to the well known WebKB dataset in a series of experiments. WebKB consists of more than 4000 Web pages from 4 university Websites and it adopts a University ontology similar to that in Figure 1. It was shown that Web units can be effectively determined using the algorithms and be assigned the correct concept labels. In other words, the homepage mining objective is achieved. The experimental results have also shown that iWUM worked extremely well for fairly well structured Websites.

Web Unit-Based Homepage Relationship Mining

Homepage relationship mining assumes that homepages of concept instances are known, possibly using Web-unit based homepage mining method. It can be formulated as a binary classification. That is, given a relationship $r_k(c_s, c_t)$ where c_s and c_t are the source and target concepts of the relationship r_k respectively, one has to determine whether a pair of homepages $\langle h_s, h_t \rangle$ is an instance of the relationship $r_k(c_s, c_t)$ where $h_s \in c_s$ and $h_t \in c_t$. The general solution to homepage relationship mining consists of three steps, namely *candidate homepage pair generation*, *feature extraction*, *and classifier learning*.

Homepage relationship mining is special in two aspects. For a given relationship $r_k(c_s, c_t)$, one has to generate candidates of hompage pairs by considering pairs of homepages belonging to c_s and c_t respectively. A naïve approach to get all possible pairs however can lead to a very large number. For example, for *m* and *n* homepages of c_s and c_t respectively, there can be *mn* candidate pairs causing much classification overheads. Moreover, the features for representing homepage pairs have to be carefully determined since a poor choice of feature set can lead to poor classification accuracies. In our earlier research, we addressed the above issues by introducing *inter-homepage features* and *zero-filter* [10].

Although a Web page can be represented by a set of features derived from its content, anchor text of incoming links, and HTML tags, a straightforward composition of these features for a pair of homepages does not necessarily work well in homepage relationship mining. Our experiments have shown that simple feature composition yield very poor mining accuracies. We therefore adopt some specially defined inter-homepage features to characterize the background relations between a pair of homepages $\langle h_s, h_t \rangle$. These features include:

- *Navigation features:* The series of links between homepages may suggest the relationship between them. Since we have the notion of Web units, a link can be either an *intra-unit* or *inter-unit* link. Inter-unit links usually capture more information about the relationships and we therefore derive the navigation features from these inter-unit links only. There are altogether 24 different navigation features that can be derived from direct links and indirect links with one intermediate Web page. For example, $h_s \xrightarrow{h-h} h_t$, $h_s \xrightarrow{h-s} h_t$, and $h_s \xrightarrow{h-p-h} h_t$ represent a direct link from h_s to h_t, a direct link from h_s to the support page of Web unit with homepage h_t, and a indirect link from h_s to h_t. The full set of navigation features are therefore:
  $h_s \xrightarrow{h-h} h_t, h_s \xrightarrow{s-h} h_t, h_s \xrightarrow{s-s} h_t, h_s \xrightarrow{h-p-h} h_t, h_s \xrightarrow{h-p-s} h_t, h_s \xrightarrow{s-p-h} h_t, h_s \xrightarrow{s-p-s} h_t,$
  $h_t \xrightarrow{h-h} h_s, h_t \xrightarrow{s-h} h_s, h_t \xrightarrow{s-s} h_s, h_t \xrightarrow{h-p-h} h_s, h_t \xrightarrow{h-p-s} h_s, h_t \xrightarrow{s-p-h} h_s, h_t \xrightarrow{s-p-s} h_s,$
  $h_s \xleftarrow{h-p-h} h_t, h_s \xleftarrow{h-p-s} h_t, h_s \xleftarrow{s-p-h} h_t, h_s \xleftarrow{s-p-s} h_t,$
  $h_s \overline{\xleftarrow{h-p-h}} h_t, h_s \overline{\xleftarrow{h-p-s}} h_t, h_s \overline{\xleftarrow{s-p-h}} h_t, h_s \overline{\xleftarrow{s-p-s}} h_t$
  where $\xleftarrow{h-p-h}$ represents two outgoing links from an intermediate Web page to the two homepages, and $\overline{\xleftarrow{h-p-h}}$ represents two outgoing links from the two homepages to an intermediate Web page.

- *Relative-location features:* Relative-location features are association between two homepages' locations in the Web directory. They include *parent-child, sibling* and *ancestor-descendent.*
- *Common-item features:* Common-item features refer to comon items shared by the pair of homepages. Examples are email addresses, people names, and phone numbers. In our experiments, we have used email addresses as common items.
- *Supplementary features:* These are additional features derived for some inter-homepage features and must be used together with their associated inter-homepage features. In our experiments, we have used anchor terms associated with direct links between homepages.

Having defined the above inter-homepage features, we use them to represent each homepage pair for relationship mining. A zero-filter is adopted to remove those homepage pairs that have homepages are not related to each other by any of the inter-homepage features.

We conducted experiments on the classification method for homepage relationship mining on the WebKB dataset. There were three relationships involved in the experiments, namely *Department_Of(Person, Department)*, *Member_Of(Person, Project)*, and *Instructor_Of(Course, Person)*. The experiments used three universities' Web pages for training and the remaining ones for testing. SVM classifiers were learnt using the training set. It was found that good classification accuracy was achieved using navigation features alone, while other inter-homepage features helped to improve accuracies in lesser extent. It was also found that the results of homepage relationship mining were very dependent on the quality of homepage mining. If homepages could not be determined accurately, the errors would propagate further to homepage relationship mining.

SUMMARY AND CONCLUSIONS

Ontology is a some domain knowledge that could be used to describe information on the Web. This paper summarizes the use of ontology in Web mining. In particular, we focus on how ontology has been incorporated in Web mining. We also use two special Web content mining problems known as homepage mining and homepage relationship mining as examples, and present our solutions. As more ontologies get developed covering a wide of domains and the need for advanced Web search increases, more ontology-based Web mining research will be required. At the same time, we also envisage many of these solution techniques will be implemented in the emerging digital library applications.

REFERENCES

[1] O'Neil, E., Lavoie, B. F., and Bennett, R., "Trends in the Evolution of the Public Web, 1998-2002,", *D-Lib Magazine* 9(4), 2003.

[2] Berners-Lee, T., Hendler, J. and Lassila, O., "The Semantic Web," *Scientific American*. 284(5):35-43, 2001.

[3] Kosala, R., and Blockeel, H., "Web Mining Research: A Survey," *SIGKDD Explorations*, 2(1), June 2000.

[4] Sun, A., Lim, E.-P., and Ng, W. K., "Web Classification using Support Vector Machines," *ACM Workshop on Web Information and Data Management (WIDM'02)*, 2002.

[5] Laender, A., Ribeiro-Neto, B., da Silva, A. and Teixeira, J., "A Brief Survey of Web Data Extraction Tools," *SIGMOD Record* 31(2):84-93, 2002.

[6] Muslea, I., "Extraction Patterns for Information Extraction Tasks: A Survey", *AAAI Workshop on Machine Learning for Information Extraction*, 1999.

[7] Davies, J., Fensel, D., and van Harmelen, F., *Towards the Semantic Web: Ontology-Driven Knowledge Management*, John Wiley and Sons, Ltd., 2002.

[8] Su, X. and Ilebrekke, L., "A Comparative Study of Ontology Languages and Tools," *Conference on Advanced Information System Engineering (CAiSE'02)*, 2002.

[9] Sun A., and Lim, E.-P., "Web unit mining: finding and classifying subgraphs of Web pages," *ACM Conference on Information and Knowledge Management (CIKM'03)*, 2003.

[10] Sun A., and Lim, E.-P., "Web Unit Based Mining of Homepage Relationships," *Journal of American Society for Information Science and Technology (JASIST)*, accepted, 2005.

[11] Genesereth, M. R., and Nilsson, N. J., *Logical Foundations of Artificial Intelligence*, San Mateo, CA: Morgan Kaufmann Publishers, 1987.

[12] Maedche, A., Motik, B., and Stojanovic, L., "Managing Multiple and Distributed Ontologies on the Semantic Web," *The VLDB Journal* 12:286-302, 2003.

[13] Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R. and Wenke, D. "OntoEdit: Collaborative Ontology Engineering for the Semantic Web," *First International Semantic Web Conference 2002 (ISWC 2002)*, 2002.

[14] Navigli, R., Velardi, P., and Gangemi, A., "Ontology Learning and Its Application to Automated Terminology Translation," *IEEE Intelligent Systems*, January/February 2003.

[15] Naing, M.-M, Lim, E.-P., and Chiang, R. H.-L., "Core: A Search and Browsing Tool for Semantic Instances of Web Sites," *Asia Pacific Web Conference (APWeb'05)*, 2005.