

Web Usage Mining for Semantic Web Personalization

Baoyao Zhou¹, Siu Cheung Hui, and Alvis C. M. Fong

School of Computer Engineering, Nanyang Technological University, Singapore
zhouby@pmail.ntu.edu.sg, {asschui, ascmfong}@ntu.edu.sg

Abstract. With the explosive growth of information on the Web, it has become more difficult to access relevant information from the Web. One possible approach to solve this problem is web personalization. In Semantic Web, user access behavior models can be shared as ontology. Agent software can then utilize it to provide personalized services such as recommendation and search. To achieve this, we need to tackle the technical issues on transforming web access activities into ontology, and deducing personalized usage knowledge from the ontology. In this paper, we propose a web usage mining approach for semantic web personalization. The proposed approach first incorporates fuzzy logic into Formal Concept Analysis to mine user access data for automatic ontology generation, and then applies approximate reasoning to generate personalized usage knowledge from the ontology for providing personalized services.

1 Introduction

With the explosive growth of information available on the World Wide Web, it has become more difficult to access relevant information from the Web. One possible approach to solve this problem is web personalization [1]. Web usage mining [2], which aims to discover interesting and frequent user access patterns from web usage data, can be used to model past web access behavior of users. The acquired model can then be used for analyzing and predicting the future user access behavior. Semantic Web [3] provides a common framework that allows data to be shared and reused across application, enterprise and community boundaries. In Semantic Web environment, user access behavior models can be shared as ontology. Agent software can then utilize the ontology to provide personalized user services such as recommendation and search.

Ontology has become an important component for Semantic Web, as it allows the description of the semantics of web content. Various techniques such as Natural Language Processing (NLP) [4], association rules [5], hierarchical clustering [6] and Formal Concept Analysis (FCA) [7] have been investigated for ontology generation. However, majority of these works have focused on generating concept hierarchy for building ontology from text documents. Recently, semantic web personalization [8-10] has become an active research area. However, the current research works create ontology manually and are unable to deal with temporal access behavior. Further, most of them investigated the problem mainly for a specific domain such as e-learning [8,9].

¹ This work was supported by Singapore Millennium Foundation Ltd (www.smf-scholar.org).

To provide semantic web personalization, we need to tackle the technical issues on how to define web access activities, discover hierarchical relationships from web access activities, transform them into ontology automatically, and deduce personalized usage knowledge from the ontology. This paper proposes a web usage mining approach for semantic web personalization. The proposed approach first incorporates fuzzy logic into Formal Concept Analysis [11] to mine client-side web usage data for automatic ontology generation, and then applies fuzzy approximate reasoning [12] to generate personalized usage knowledge from the ontology.

2 Proposed Architecture

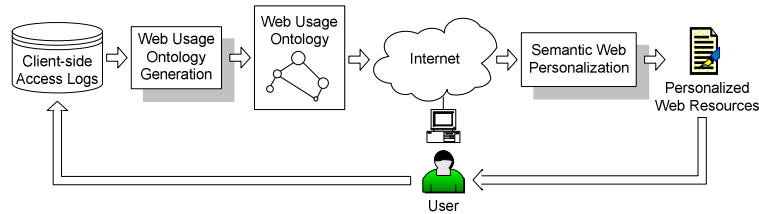


Fig. 1. Proposed architecture for semantic web personalization.

Web logs record user access events from websites as a sequence of requested web pages. We focus on mining client-side access logs, which record access events (involving multiple websites) of a single user or client. Figure 1 shows the proposed web usage mining approach for semantic web personalization which consists of two main components: *Web Usage Ontology Generation* and *Semantic Web Personalization*.

3 Web Usage Ontology Generation

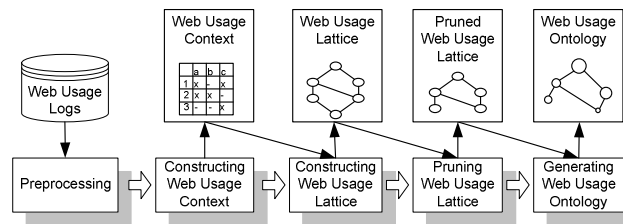


Fig. 2. Web usage ontology generation.

Figure 2 shows the proposed approach for Web Usage Ontology Generation, which consists of the following components: (1) *Preprocessing*; (2) *Constructing Web Usage Context*; (3) *Constructing Web Usage Lattice*; (4) *Pruning Web Usage Lattice*; and (5) *Generating Web Usage Ontology*.

3.1 Preprocessing

Preprocessing is responsible for processing the original web usage logs in order to identify all user access sessions for each individual user. Similar preprocessing tasks for web server logs discussed in [13] can be used in this step.

Let E be a set of unique access events, which represents web resources accessed by users, i.e. URLs of web pages. A user access session $S = e_1e_2\dots e_n$ ($e_i \in E$ for $1 \leq i \leq n$) is a sequence of access events. Each $e_i = (ts_i, te_i, URL_i)$, where ts_i is the start time of event e_i , te_i is the end time of event e_i , and URL_i is the URL accessed by the user in event e_i . Note that it is not necessary that $URL_i \neq URL_j$ for $i \neq j$ in S .

The URLs in web usage logs contain little semantic information about the web contents accessed by users. To overcome this problem, we map each URL into a predefined category or topic such as News, Sports and Entertainment. The category information can be obtained using a web page category classification technique [14]. As such, each user access session can be classified into a sequence of categories. Suppose that the set of event attributes M_C consists of all valid predefined categories, we define the function $Category(e_i) = m_c$ ($m_c \in M_C$) such that it converts each e_i into an event attribute m_c in M_C . In general, the duration ($te_i - ts_i$) of an access event e_i can be used to indicate the level of interest the user has in that web content. Therefore, the total duration for each category accessed can be used for estimating the level of user interest in that category for each user access session. After classifying the categories and computing the duration, each user access session S is transformed into $S^* = (Ts, Te, D)$, where Ts is the start time of the session (i.e. ts_1 in S), Te is the end time of the session (i.e. te_n in S) and the total duration $D = \{d(S, m_i) \mid d(S, m_i) = \sum_{e_i \in S, Category(e_i)=m_i} (te_i - ts_i), m_i \in M_C, 1 \leq i \leq |M_C|\}$.

3.2 Constructing Web Usage Context

We have defined seven real-life time concepts, namely *Early Morning*, *Morning*, *Noon*, *Early Afternoon*, *Late Afternoon*, *Evening* and *Night* to represent temporal attributes for web activities. We have also defined 26 web categories such as *Games*, *Adults*, *Sports* and *Entertainment* as event attributes to describe web access activities. As such, user access behavior can be represented by a set of temporal and event attributes. In addition, we also use fuzzy logic [12] to represent both temporal and event attributes, and incorporate them into Formal Concept Analysis [11] for constructing Web Usage Context.

Definition 1. A fuzzy temporal based *Web Usage Context* is $K = (G, M_T, M_C, I)$, where G is a set of *user access sessions*, M_T is a set of *temporal attributes*, M_C is a set of *event attributes*, $I = R(G, M_T \cup M_C)$ is a fuzzy set on domain $G \times (M_T \cup M_C)$ to represent fuzzy relation between G and $M_T \cup M_C$. Each fuzzy relation $R(g, m) \in I$, where $g \in G$, $m \in M_T \cup M_C$, is represented by a membership value $\mu_R(g, m)$ in $[0, 1]$.

The member function of a user access session $g_i = (Ts_i, Te_i, D_i)$, on a temporal attribute $m_t \in M_T$ is computed as

$$\mu_R(g_i, m_t) = \max(\mu(t, m_t), t \in [Ts_i, Te_i])$$

where $\mu(t, m_t)$ is defined in Figure 3, which is modified from [15].

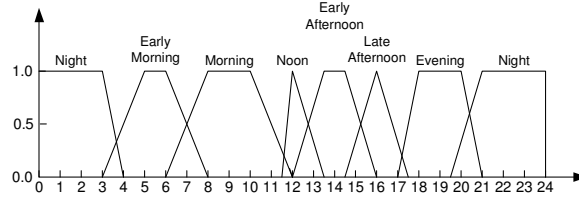


Fig. 3. Member function $\mu(t, m_c)$.

The member function of a user access session $g_i = (Ts_i, Te_i, D_i)$ on an event attribute $m_c \in M_C$ is computed as

$$\mu_R(g_i, m_c) = \begin{cases} 0 & \text{if } z(g_i, m_c) < \frac{1}{2}Z(m_c) \\ \frac{2z(g_i, m_c)}{Z(m_c)} - 1 & \text{if } \frac{1}{2}Z(m_c) \leq z(g_i, m_c) \leq Z(m_c) \\ 1 & \text{if } z(g_i, m_c) > Z(m_c) \end{cases}$$

where

$$z(g_i, m_c) = \frac{d(g_i, m_c)}{Te_i - Ts_i}, \quad Z(m_c) = \frac{\sum_{g_k \in G} d(g_k, m_c)}{\sum_{g_k \in G} (Te_k - Ts_k)}$$

$Z(m_c)$ is defined as the proportion of the duration of accessing a web category m_c in all user access sessions, which indicates the user's global interest of the web category m_c . $z(g_i, m_c)$ is defined as the proportion of the duration of accessing a web category m_c within a user access session g_i , which indicates the user's local interest of the web category m_c .

A Web Usage Context can be represented by a cross table with rows indicating user session IDs and columns indicating the temporal and event attributes. A membership value $\mu_R(g, m) \in [0, 1]$ in row g and column m indicates a fuzzy relation between the session g and attribute m . Table 1 shows an example Web Usage Context, which consists of five user access sessions, three temporal attributes "T1 (A - Late Afternoon)", "T2 (E - Evening)" and "T3 (N - Night)", and three event attributes "C1 (S - Sports)", "C2 (G - Games)" and "C3 (C - Chat)". The relation between a user access session, and a temporal attribute or an event attribute is represented by a membership value in $[0, 1]$, which can be computed automatically as discussed earlier.

Table 1. A cross table of an example Web Usage Context.

SID	T1 (A)	T2 (E)	T3 (N)	C1 (S)	C2 (G)	C3 (C)
S1	0.6	0.5	-	0.8	-	0.6
S2	-	0.4	0.7	-	-	0.9
S3	-	-	0.9	-	0.7	0.5
S4	-	0.8	0.3	0.8	0.6	0.5
S5	-	-	1.0	-	0.9	-

3.3 Constructing Web Usage Lattice

In this step, we construct a Web Usage Lattice from a Web Usage Context.

Definition 2. Given a Web Usage Context $K = (G, M_T, M_C, I)$, we define the set of attributes common to the access sessions in $A \subseteq G$ as $A' = \{m \in M_T \cup M_C \mid \forall g \in A: \mu_R(g, m) > 0\}$, and the set of access sessions which have all the same attributes in $B \subseteq M_T \cup M_C$ as $B' = \{g \in G \mid \forall m \in B: \mu_R(g, m) > 0\}$.

Definition 3. Given a Web Usage Context $K = (G, M_T, M_C, I)$, if there exists a pair (A, B) with $A \subseteq G, B \subseteq M_T \cup M_C$ ($B \cap M_T \neq \emptyset$ and $B \cap M_C \neq \emptyset$), $A' = B$ and $B' = A$, then the fuzzy set based on B , denoted as $\varphi(B)$, is called a *web access activity*, and each $m_i \in \varphi(B)$ has a membership value $\mu_m(m_i) = \min_{g \in A} \mu_R(g_j, m_i)$.

A web access activity represents a temporal access behavior of a user, i.e. it is an implication from temporal attributes to event attributes. The *fuzzy support* of $\varphi(B)$ is defined as

$$Sup(B) = \frac{\sum_{g \in B'} \prod_{m_i \in B} \mu_R(g_i, m_i)}{|G|}$$

and the *fuzzy confidence* of $\varphi(B)$ is defined as

$$Conf(B) = prob(B \cap M_T \mid B \cap M_C) = \frac{Sup(B)}{Sup(B \cap M_T)}$$

where $prob(\cdot)$ is a conditional probability.

Definition 4. Let $\varphi(B_1)$ and $\varphi(B_2)$ be two web access activities, $\varphi(B_1)$ is called *sub-activity* of $\varphi(B_2)$, if and only if $\varphi(B_2) \subseteq \varphi(B_1)$. Equivalently, $\varphi(B_2)$ is called *super-activity* of $\varphi(B_1)$. Such relation is called *hierarchical order* of the web access activities.

Definition 5. A fuzzy temporal based *Web Usage Lattice* of a Web Usage Context $K = (G, M_T, M_C, I)$ is the set of all web access activities with hierarchical order.

Figure 4 shows the Web Usage Lattice of the Web Usage Context given in Table 1. Each node in the figure represents a web access activity with the corresponding fuzzy set of its attributes on the left. Each edge in the figure represents a hierarchical relationship. Note that we have added a virtual node as the root of the lattice.

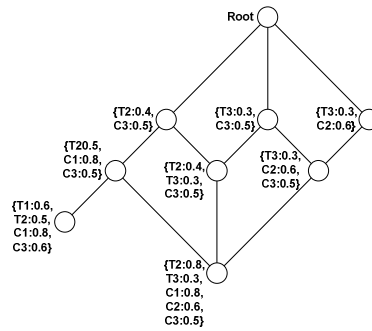


Fig. 4. Web Usage Lattice.

3.4 Pruning Web Usage Lattice

The Web Usage Lattice may be quite complicated and huge due to the large number of web access activities generated. Therefore, it will be very costly to convert the

whole lattice into ontology. To overcome this problem, the lattice should be pruned to retain only those *interesting* web access activities, which are more important for describing the access behavior of a user.

Definition 6. Given a minimum support $MinSup \in [0, 1]$ and a minimum confidence $MinConf \in [0, 1]$, we call a web access activity $\phi(B)$ *interesting*, if $Sup(B) \geq MinSup$ and $Conf(B) \geq MinConf$.

Given a minimum support $MinSup = 0.1$ and $MinConf = 0.15$, the pruned Web Usage Lattice is shown in Figure 5. The fuzzy support and confidence values of each web access activity are shown on the right of the corresponding node. In addition, an activity ID, which is labeled in each node, is assigned to each web access activity.

Based on the pruned Web Usage Lattice, we can derive two kinds of knowledge on user access behavior. The *explicit* knowledge can be extracted from each activity node to represent the user's temporal access patterns. The *implicit* knowledge can be inferred from the hierarchical relations among activity nodes to represent the user's association access patterns. These two kinds of knowledge can be used for semantic web personalization which will be discussed in Section 4.

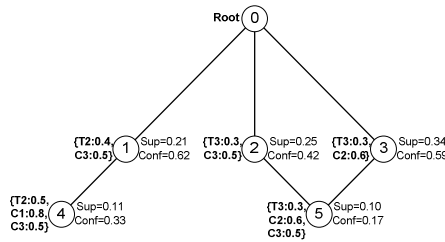


Fig. 5. Pruned Web Usage Lattice.

3.5 Generating Web Usage Ontology

We use OWL (Web Ontology Language) [16] to represent the generated ontology. The body of ontology consists of classes, properties. One of the main components of ontology is taxonomy, i.e. class hierarchy. To generate the Web Usage Ontology from the pruned Web Usage Lattice, we define the following transformation rules:

1. *Classes.* Each web access activity is mapped into an activity class. Note that the root (labeled as 0) in the pruned Web Usage Lattice is a virtual node, thus there is no need for generating the corresponding activity class.
2. *Properties.* Each temporal and event attribute of a web access activity is transformed into a property of the corresponding class. The membership value of each attribute is stored in the corresponding property. Further, the fuzzy support and confidence of each web access activity are also represented as properties named "Support" and "Confidence" respectively.
3. *Class Hierarchy Relations.* Each hierarchical relation between web access activities forms a taxonomy relation between activity classes. The sub-activity relationship in the Web Usage Lattice is transformed into the subclass relationship in the Web Usage Ontology.

Figure 6 shows an example on transforming the activity node 4 given in Figure 5 into the corresponding class definition of “Activity_4” of the Web Usage Ontology.

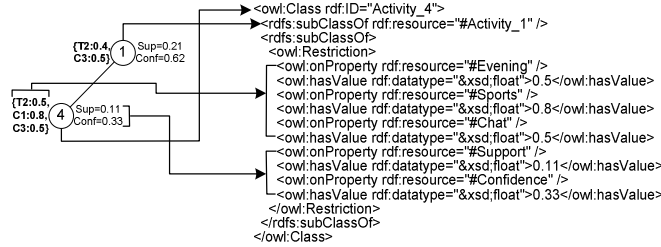


Fig. 6. Transforming an activity node into a class definition of the Web Usage Ontology.

4 Semantic Web Personalization

Web Usage Ontology stores both *explicit* and *implicit* knowledge on user web access behavior. Agent software can utilize the ontology to derive such knowledge. As fuzzy knowledge is stored in Web Usage Ontology, we have applied fuzzy logic techniques into Web Usage Ontology to deduce personalized usage knowledge for semantic web personalization services. This involves extracting activity rules, approximate reasoning from activity rules and providing personalized services.

4.1 Extracting Activity Rules

Knowledge on user access behavior from Web Usage Ontology can be extracted as activity rules. Each activity rule is represented in the form of *conditional and qualified propositions* [12], which are a specific kind of fuzzy propositions. The conditional and qualified propositions are characterized by the canonical form “If x is A , then y is B is S ”, where x and y are variables whose values are in sets X and Y respectively, A and B are fuzzy sets on X and Y respectively, and S is a fuzzy truth qualifier. Such propositions are also known as qualified “IF-THEN” rules. Baldwin [17] defined fuzzy truth qualifier in the universal set $V = \{v \mid v \in [0, 1]\}$ as $T = \{\text{true, very true, fairly true, absolutely true, undecided, absolutely false, fairly false, very false, false}\}$.

Web Usage Ontology gives two kinds of activity rules: *simple activity rules* and *association activity rules*. Simple activity rules can be extracted from the properties of each activity class directly, whereas association activity rules can be inferred from activity classes and the class hierarchy.

Given a Web Usage Ontology, *simple activity rules* of each activity class are in the form of “If x is A then y is B is S ”, where A and B are fuzzy sets of the corresponding temporal properties and event properties of the activity class respectively. We can calculate the fuzzy truth qualifier S using the confidence property (*Conf*) of the activity class and the minimum confidence (*MinConf*) that is used for pruning the Web Usage Lattice as Figure 7. For example, from the activity class “Activity_4” given in Figure

6, a simple activity rule “If 0.5/T2 then 0.8/C1+0.5/C3 is fairly true” can be extracted. As a result, a total of five rules can be extracted from the example Web Usage Ontology.

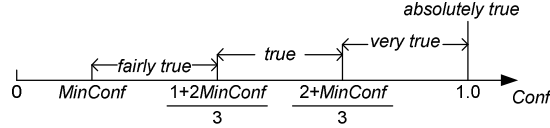


Fig. 7. Calculating the fuzzy truth qualifier using the fuzzy confidence (*Conf*).

Given a Web Usage Ontology, *association activity rules* are in the form of “If x is A then y is B is S ”, where A and B are fuzzy sets of the temporal and event properties of activity classes i and j respectively. Such rules require the activity class j to be the immediate subclass of the activity class i , and the fuzzy confidence $Conf \geq MinConf$. The fuzzy confidence (*Conf*) of association activity rules is equal to the support property of the activity class j divided by that of activity class i . The fuzzy truth qualifier S can also be calculated using the fuzzy confidence of the association activity rule. Therefore, each association activity rule corresponds exactly to one hierarchical relation from the super-activity class to one of its sub-activity classes in Web Usage Ontology. For example, in the Web Usage Ontology given in Figure 6, the relation from the activity class “Activity_1” to the activity class “Activity_4” represents an association activity rule “If 0.4/T2+0.5/C3 then 0.5/T2 +0.8/C1 +0.5/C3 is true”. A total of three association activity rules can be extracted from this Web Usage Ontology.

4.2 Approximate Reasoning from Activity Rules

Based on the two kinds of activity rules, simple and association activity rules, agent software can derive two kinds of personalized usage knowledge using fuzzy approximate reasoning according to:

- A specific time interval such as [19:00:00, 20:00:00] – then we can obtain a ranked list of web content categories that are most relevant to the user’s interests for this time interval; and
- An existing user access session – then we can obtain a web access activity that is most likely to occur. The web access activity can be transformed further into a specific time interval and a ranked list of web content categories.

The agent software can then customize or reorganize web resources for the user according to the ranked list of web content categories for the specific time interval.

Fuzzy approximate reasoning is an inference procedure that deduces imprecise conclusions from fuzzy rules and known facts. The general schema of the qualified “IF-THEN” rule based approximate reasoning has the following form:

$$\begin{array}{ll}
 \text{Rule:} & \text{If } x \text{ is } A, \text{ then } y \text{ is } B \text{ is } S \\
 \text{Fact:} & x \text{ is } A' \\
 \hline
 \text{Conclusion:} & y \text{ is } B'
 \end{array} \tag{1}$$

The qualified “IF-THEN” rule can be either a simple activity rule or an association activity rule.

One approach for approximate reasoning is called *truth-value restrictions method* [12] which is based on a manipulation of linguistic truth values. In this research, we treat all the rules as disjunctive. This means that we obtain a conclusion for all fuzzy rules by calculating the union of the conclusion of each single qualified “IF-THEN” rule. Suppose B'_i is the conclusion from the i^{th} fuzzy rule ($i=1, \dots, n$), then the conclusion for all fuzzy rules is $B' = \bigcup_{i=1}^n B'_i$.

A specific time interval T_p can be transformed into a fuzzy set of time concepts using the member function given in Definition 1. Given the fuzzy set as the fact x is A' , and using simple activity rules as rules in schema (1), the conclusion B' is a fuzzy set of web content categories F_C , in which the fuzzy membership value of each category indicates its importance to the user. Generally, larger membership values indicate higher priorities. Therefore, membership values can be used for ranking web content categories. F_C can be regarded as a ranked list of web content categories L_C .

A specific user access session can also be transformed into a fuzzy set of both time concepts and web content categories using the member function given in Definition 1. Given the fuzzy set as the fact x is A' , and using association activity rules as rules in schema (1), the conclusion B' is another fuzzy set of time concepts and web content categories, which can be divided into two parts, i.e. a fuzzy set of time concepts F_T and a fuzzy set of web content categories F_C . A specific time interval T_p can then be calculated from F_T . And F_C can be regarded as a ranked list of web content categories L_C .

From the above discussion, the personalized usage knowledge for both simple and association activity rules can be represented as a specific time interval T_p and a ranked list of web content categories L_C .

4.3 Providing Personalized Services

After deriving the personalized usage knowledge from approximate reasoning of activity rules, agent software can then customize and reorganize web resources for the users for the specific time interval T_p based on the ranked list of web content categories L_C . Assume that we have obtained [19:00:00, 20:00:00] and {C1:1.0, C2:0.0, C3:0.5} as personalized usage knowledge after approximate reasoning. If the agent software needs to provide personalized search service, then the URL links to web contents related to C1 (Sports) will be highlighted to the user with higher priority in the search result list during the time period [19:00:00, 20:00:00]. If the agent software intends to perform personalized web recommendation, then web resources involving C1 (Sports) and C3 (Chat) will be recommended as the content that are more likely to be accessed by the user during the time period [19:00:00, 20:00:00].

5 Conclusion

In this paper, we have proposed a web usage mining approach for semantic web personalization. In the proposed approach, we incorporate fuzzy logic into Formal Con-

cept Analysis to mine the client-side web usage logs for automatic generation of Web Usage Ontology. Then we extract fuzzy activity rules from Web Usage Ontology and deduce personalized usage knowledge from the activity rules using approximate reasoning. The derived personalized usage knowledge can be used for supporting semantic web personalization services. The performance of the proposed approach is currently under evaluation using web usage data from a group of research students in the Database Technology Lab, Nanyang Technological University, Singapore.

References

1. Eirinaki M., and Vazirgiannis M., Web Mining for Web Personalization, *ACM Transactions on Internet Technology*, 3(1) (2003) 1-27.
2. Kosala R., and Blockeel H., Web Mining Research: A Survey. In: *ACM SIGKDD Explorations*, 2 (2000) 1-15.
3. Berners-Lee T., Hendler J., and Lassila O., The Semantic Web. *Scientific American*, (2001).
4. Todirascu A., Beuvron F., Galea D. and Rousset F., Using Description Logics for Ontology Extraction. In: *Workshop on Ontology Learning*, at the 14th European Conference on Artificial Intelligence, Berlin, (2000).
5. Maedche A., and Staab S., Ontology Learning for the Semantic Web. *IEEE Intelligent Systems*, 16(2) (2001) 72-79.
6. Clerkin P., Cunningham P., and Hayes C., Ontology Discovery for the Semantic Web using Hierarchical Clustering. In: *Workshop on Semantic Web Mining*, at ECML/PKDD, (2001) 27-38.
7. Stumme G., and Maedche A., Ontology Merging for Federated Ontologies on the Semantic Web. In: *Workshop on Ontologies and Information Sharing*, at IJCAI, Seattle, USA , (2001).
8. Peter D., Nicola H., Wolfgang N., and Michael S., The Personal Reader: Personalizing and Enriching Learning Resources Using Semantic Web Technologies. In *Proc. of AH 2004*, LNCS 3137, pp. 85-94, (2004).
9. Peter D., Nicola H., Wolfgang N., and Michael S., Personalization in Distributed eLearning Environments, In *Proc. of WWW2004*, New York, USA, pp. 170-179, (2004).
10. Kumar S., Kunjithapatham A., Sheshagiri M., Finin T., Joshi A., Peng Y., and Cost R.S., A Personal Agent Application for the Semantic Web, *AAAI Fall Symposium on Personalized Agents*, North Falmouth, (2002).
11. Ganter B., and Wille R., *Formal Concept Analysis: Mathematical Foundations*. Springer, Heidelberg, (1999).
12. Klir G.J., and Yuan B., *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Published by Prentice Hall, (1995).
13. Cooley R., Mobasher B., and Srivastava J., Data Preparation for Mining World Wide Web Browsing Patterns. *Journal of Knowledge and Information Systems*, 1(1) (1999) 5-32.
14. Lee P.Y., Hui S.C., and Fong A.C.M., Neural Networks for Web Content Filtering. *IEEE Intelligent Systems*, 17(5) (2002) 48-57.
15. Olaru D., and Smith B. Modelling Daily Activity Schedules with Fuzzy Logic. In: *Proc. of the 10th Intl. Conf. in Travel Behaviour Research*, Lucerne, Switzerland, (2003).
16. Web Ontology Language (OWL), Available at <http://www.w3.org/2004/OWL/>.
17. Baldwin J.F., Fuzzy Logic and Fuzzy Reasoning, *Intl. Journal of Man-Machine Studies*, (1978).