

ONTOLOGY BASED USER MODELING FOR PERSONALIZED INFORMATION ACCESS

PLABAN KUMAR BHOWMICK*

*Computer Science & Engineering, Indian Institute of Technology, Kharagpur
West Bengal, India-721302
plaban@gmail.com*

SUDESHNA SARKAR

*Computer Science & Engineering, Indian Institute of Technology, Kharagpur
West Bengal, India-721302
sudeshna@cse.iitkgp.ernet.in*

ANUPAM BASU

*Computer Science & Engineering, Indian Institute of Technology, Kharagpur
West Bengal, India-721302
anupam@iitkgp.ac.in*

User modeling is an integral part of any personalized information retrieval system. The user model should be adaptable in order to capture the change in information needs of the users. In this paper, we present an ontology based user modeling strategy in the context of personalized information access. We have adopted a hybrid approach by capitalizing on the features of static and dynamic user profiling strategies. Static user profile specifies the user's interest in a very focused manner and dynamic user profiling adds the feature of adaptability into it. The dynamic user profiling strategy make use of the data sources like usage log and mouse operations that are performed by the users during the browsing sessions. Experiments have performed to evaluate the proposed method for user profiling.

Keywords: Ontology; user modeling; personalization; user profile; usage log.

1. Introduction

Retrieving personalized information from the ever increasing information space of Internet is a broad area of research. Due to the sheer volume of the web, searching for specific information from it has become tedious and time consuming. To alleviate this problem, efforts have been started to customize the view of the web in a user specific way. The field of *Personalized Information Retrieval* deals with the idea of retrieving specific information customized to the need of the user. The model of user is an integral part of any system that provides personalized information. User

*corresponding author

modeling can be described as the process of building the personal preferences of the users in terms of user's knowledge about the world, her behavioral aspects, goals, likes and dislikes. Current research work on personalized information delivery is directed towards representing and constructing the model of the users. The model of the user is generally represented in the form of user profile which captures the personal preferences in a machine processable format. So, the user model can be seen as an abstract entity and the user profile represents an instantiation of the user model for a particular user.

In this context, we refer to a system, called *Samvidha* [Bhowmick *et al.* (2007)], that can be used by school students to query the Internet to retrieve personalized information. The domain of interest is school level topics. In this domain, different users belonging to different grades and abilities will have different requirements. The requirement of a class seven student is somewhat different from that of a tenth standard student. So, the same set of documents may not be understandable to students of every level. Traditional search engines do not address this issue. There is need to develop a system which will personalize the view of the global database depending on the personal preferences of the users. To deliver the appropriate set of documents to the user, the system requires the knowledge of the underlying domain and also the concepts of interest to the user and her level of knowledge. To achieve this, there is need for proper representation of the domain knowledge about different subjects and also knowledge about the user's requirement. In this paper, we will be describing the representation of the domain knowledge in ontological structure and the use of the domain knowledge in user modeling. The knowledge requirement of a user shifts over time. A successful personalized system should be adaptive in order to keep up with the changing information needs of the user. In this paper, we have presented a method for acquiring and updating user models by analyzing the users' information seeking behavior.

In section 2, we provide a brief overview of the works on ontology based user modeling. The ontological representation of the school level topics have been discussed in section 3. In section 4, we describe the user model. Section 5 deals with the method for user profile acquisition. In section 6, the experiments to evaluate the proposed user profile acquisition method. In section 7, we provide some concluding remarks.

2. Related Works

The research works differ in the way they represent the user profile and the adaptivity of the system. A personalized system is said to be adaptive if the system is able to tune itself according to requirement of the user. But even among the adaptive systems, the algorithm for learning users interest varies. Some personalized systems use domain knowledge to model user interest while others do not.

Ontology has been a basis for the construction of a user model [Middleton *et al.* (2002)] in several personalized systems ranging from information delivery systems

to Intelligent Tutoring Systems [Dicheva and Aroyo (2000)]. In this section, we provide a brief discussion of a number of such systems.

In [Pretschner and Gauch (1999)], the user profile is represented as hierarchy of concepts. The concepts are adopted from a reference ontology of 4,400 concepts taking the top level categories from Magellan web site.

myPlanet [Kalfoglou *et al.* (2001)] is an ontology based personalized news delivery system. Simple relationships among the concepts inside the domain have been used to filter out information relevant to the user.

[Razmerita *et al.* (2003)] propose an ontology-based user modeling (OntonUM) architecture. This ontology-based user modeling system integrates three ontologies:

- User ontology: It includes different characteristics of users and their relationships.
- Domain ontology: It captures the domain or application specific concepts and their relationships.
- Log ontology: It represents the semantics of the user interaction with the system.

The overall user model consists of two parts. The explicit part of the user model is provided by the user through a user profile editor and the implicit part is maintained by intelligent services. The user ontology captures metadata about the user's profile including different characteristics: identity, email, address, preferences etc.

STyLE-OLM [Dimitrova (2003)] is an open learner modeling system where the learner participates in the process of constructing her model. The learner is allowed to inspect and discuss the content of the model. It provides an integrated computational framework which includes domain ontology, discourse model and belief mechanism for the construction and maintenance of user model. The model adopted here is an extended overlay learner model which consists of several types of attributes: correct belief, erroneous belief and incomplete belief. The user model is constructed in three basic steps: initialization, interaction and learner model update. The system has been demonstrated in two different domains. The user's interaction is carried out in graphical manner having two modes.

- DISCUSS: The learner can influence their model by discussing their idea about the domain. This process is achieved through a dialog system.
- BROWSE: The learner can inspect the current state of their model.

[Aroyo *et al.* (2006)] provides method for automatic acquisition of user knowledge through an ontology based dialog system. An ontology based dialog agent, called OWL-OLM, interacts with the user to analyze the current state of the users knowledge according to the needs for a particular course task.

[Zhang *et al.* (2007)] proposed a system for constructing user models automatically by monitoring the users browsing behaviors in each session. The system keeps track of the usage logs by means of Semantic Web Usage Log Preparation Model

(SWULPM). The user model consists of personal ontology which is represented through concept graph.

[Zeng *et al.* (2009)] proposed two different approaches for acquisition of user's knowledge requirements about course content in an e-learning system. The ontology of course is represented as concept hierarchy. The first approach relies on interactive question-answer session and the historical session logs are analyzed to determine user's requirements. The second approach is based on users reading behavior logs while reading e-documents. The reader actions considered in this work are underline, highlight, circle, annotation and bookmark.

3. Domain Model

The user interest model in this work is constructed based on the knowledge of the domain. Knowledge structure of education domain has been considered here. The knowledge representation database, ontology, is organized into a three level hierarchical structure as shown in Figure 1.

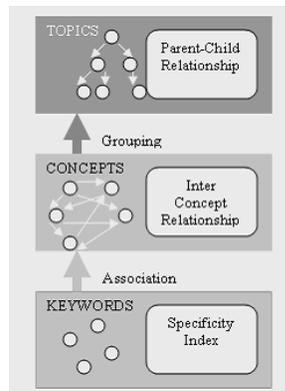


Fig. 1. Three tier representation of ontology.

3.1. Topic-Subtopic Level:

On the top level, the topics share a parent child relationship. This provides a way of generalization from a specific to a more general topic. The hierarchy of the topics is stored as an n-ary tree with the exception that a node may have multiple parents. This is because a subtopic may be placed under two or more topics. For example, in the domain of biology, animal nutrition and plant nutrition are two subtopics of the topic nutrition.

Table 1. Forward and reverse relationships

Forward Relation	Reverse Relation
Has Part	Part Of
Inherited From	Parent Of
Has Prerequisite	Prerequisite For
Part of Procedure	Procedure Contains
Is Caused By	Causes
Functionally Related	Functionally Related

3.2. Concept Level:

A topic consists of several concepts, which form the next level of the ontology and a concept may belong to one or more topics. A set of empirical relations can be defined among the concepts. The relations between the concepts considered in this work are given in Table 1. The concepts in the domain are organized into a di-graph. The existence of an edge between two concepts in the di-graph indicates that the concepts are related.

3.3. Keyword Level:

A concept is associated with several keywords or terms with *specificity index*. The specificity index stores the likelihood of the keyword representing a particular concept. These keywords are used to extract concepts from documents and queries. The association of the keywords to the concepts has several advantages. Firstly, the different keywords having the same meaning are mapped to a common concept removing the synonymous ambiguity of keywords.

Currently, the ontology contains 249 topics, 3369 concepts and 4201 keywords from Physics, Biology and Geography.

4. User Requirement Model

There is a need to model the requirement of the user in order to filter the web documents with respect to the need of the user. As stated earlier, our system looks at the requirement of the user in the school level topics. The features related to this domain are as follows.

- In every class there are some predefined syllabi that can be treated as the learning objective for the class. Generally, a student starts with a small subset of the concept space specified in the syllabi. Gradually, the requirement of the student tends to be saturated to the whole concept space enclosed by the syllabi. This type of modeling is very much similar to the *Overlay Model*.
- The syllabus for a class represents the learning objective of the student in that class. The student is expected to learn each concept whether it

is of interest to her or not. A portion of the concept space may be of higher interest value compared to others. As we focus on the students' requirements, unlike some other systems[Asnicar and Tasso (1997)], we have not included the dislikes of the user in the interest.

The user requirement is represented in the form of user profile. The same ontological structure as the domain knowledge is adopted in the representation of the domain knowledge specific attributes of the user profile. Each concept in the user profile is further annotated with a score revealing how interesting the concepts are. These attributes are used by the filtering module to sort out documents that may be of the user's interest.

5. User Profile Acquisition

In this section, we shall provide details of our user profile acquisition strategies. The identification of data sources is very important.

5.1. Data Sources

Apart from the learning algorithm, it is very important to identify the relevant data sources that will be used to learn the user profile. We have identified some data sources that continuously feed relevant data to the profile learning module. The data sources are described below.

- User provided profile: The users can provide their initial profile during registration procedure and they can further modify the profile later. These profiles are one of the primary sources for user profile acquisition system. Systems like SIFT[Yan and Garcia-Molina (1995)], SmartPush[Kurki *et al.* (1999)], WebWatcher[Joachims *et al.* (1997)] depends solely on the user provided profiles.
- Query history analysis: The queries placed by a user leave some important clues about the user's interest. This source provides a direct evidence of the user interest as it is placed by the user herself. [Yan and Meng (2004)] uses the query history to build the user profiles.
- Links present in a page: Links present in a page sometimes carry some hints about user's interest. In general, every textual link is associated with an anchor text which contains important terms related to the page associated with the link. The links can be categorized into different types according to their access by the users. Letizia[Lieberman (1995)] depends links accessed by the users for recommendation.
- Content scanned by the user: By analyzing the documents accessed by a user, the concepts of her requirement can be discovered. We have considered the features of the domain in analyzing the document content to achieve greater precision in discovering the interest.

Table 2. Relative ordering of the data sources for acquisition

Data Source	Rank	Weight(ϖ)
User provided profile	1	1.0
Query History	2	0.8
Anchor text of links	3	0.6
Content concepts	4	0.4

- Current user profile: Current user profile signifies the data that is present in the current configuration of the user profile before computing the most updated user profile. This source is important because the interest score of a concept largely depends upon the concepts that are present in the previous state of the user profile.
- Usage log: The browsing pattern of a user is stored in the form of usage log. Usage log contains the browsing structure of the user, activities of the user over the browser (mouse click, mouse scroll) and time spent on a specific document. The time spent on a page is measured by activating a timer when the user requests for the page. Reasoning can be done over the collected usage log to achieve effective acquisition of user profile. Several systems[Lieberman (1995); Dai and Mobasher (2003)] have adopted an access pattern based user profiling. In Letizia, some heuristics have been chosen to build the access patterns. The heuristics are saving references to a document, following a link etc. It does not consider the time of access for a page. Time access is important because the interest value of a page largely depends upon it. In the work[Dai and Mobasher (2003)], the access pattern is represented as a set of transaction where each transaction is an ordered pageview-weight pair. The information about the links between the documents is ignored.

As the data sources discussed above are of varying importance, we impose relative weights against them. The user provided profile should get the highest importance because it has been directly provided to the system by the user. When the user places a query, it is evident that she is interested in the concepts present in the concept but the evidence is not as direct compared to the user provided profile. So, query history has been ranked next to the user provided profile. The links in a page possibly contain terms that provoke the users to follow the links. User follows only those links for which the anchor words are of user's interest. This source is more indirect than query history. The terms in the documents scanned by the user are the last ranked data source because this source is the most indirect one as compared to the other sources. The relative ordering and weights for the data sources is shown in Table 2. In the light of the above discussion, the user profile can be captured in three different ways

- The static user profile provided by a user as an initial goal.
- Manual updating of the user profile.
- Monitoring the information access pattern of the user to update the profile automatically.

In Figure 2, we present the architecture for user profile acquisition. When a student

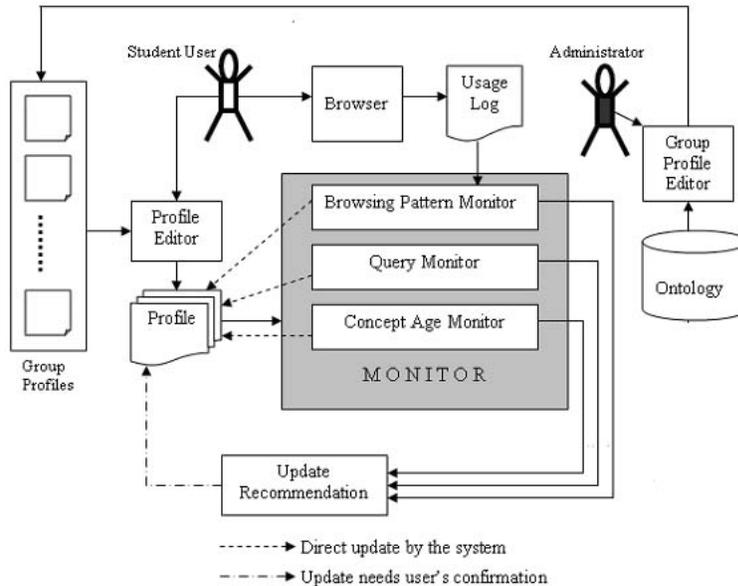
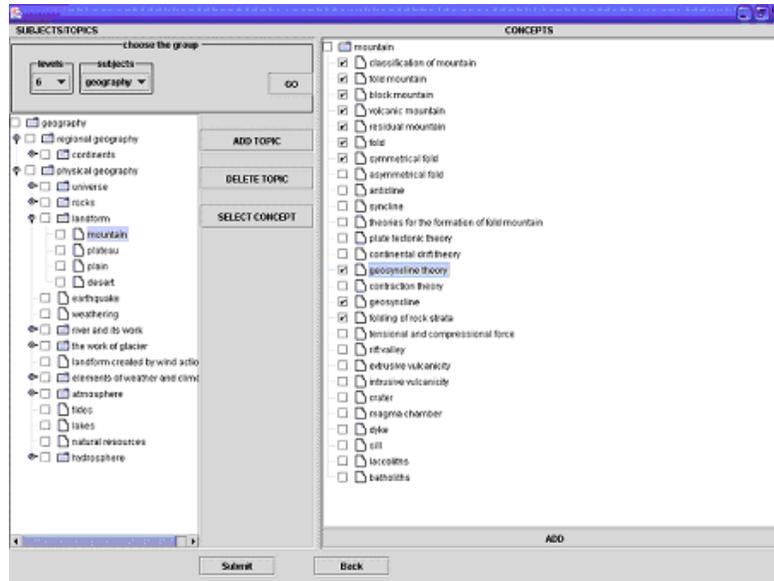


Fig. 2. Profile editing and monitoring architecture

registers into the system, she can provide her initial user profile through the *Profile Editor* (see Figure 3) and can update it later manually. During the creation of initial profile the student can consult some *Group Profiles*. A group profile represents a common requirement of a group of students. Examples of group profiles are *Biology-Class-9*, *Physics-Class-11*. Only the Administrators can create or update the group profiles.

All the user interactions during a session are tracked by three types *Monitors*. Each monitor tries to learn, assigns score to the new concepts or updates score of an existing concept in the user profile. The user interactions are stored in the form of usage log. The *Browsing Pattern Monitor* analyzes usage log of every session to dynamically update the user profile. From the usage log the system builds the access tree of a user in a single session. Each node in this tree is a page that the user has browsed in that session. The concepts present in the links and the actual contents of a page are treated separately here. The *Query Monitor* looks at the queries placed by the user and assigns scores to the concepts present in the query.

Fig. 3. User profile creation and updating interface



The *Concept Age Monitor* looks at the usage of the concepts. Concepts that have not been accessed for a long time may be assumed to lose its importance to the user. The concept age monitor penalizes those concepts by decreasing its score. The learned concepts that do not exist in the current user profile are presented to the user so that the user can confirm the inclusion of the newly learned concepts into her profile. If the learned concepts belong to the user's current profile, only the scores of the concepts are updated without prompting the user.

5.2. Concept Scoring Function

In the user model, each concept present in the user profile is assigned a score which reflects the importance of the concept to a user. The score of a concept in user profile ranges from 0 to 1. The scores are calculated keeping in view the relative importance of the data sources. The concepts that have been provided by the users manually through profile editor are assigned the highest score (1.0). For other sources, we present a general scoring function by which the scores of the concepts learned from different data sources can be calculated.

The scores of the concepts from different data sources depend on different parameters. The concepts directly provided by the user are assigned with the highest score of 1.0. Some scoring formula is applied to the other three data sources namely query history, anchor concepts and concepts present in the document. These sources always increase the score of a concept. All the data sources have some specific parameters that characterize the score increment rate for respective data sources. In

the process of calculating the score of a concept c , all the data sources use three common parameters as discussed below.

- Number of concepts in the domain ontology that are related to c .
- Number of concepts in the current user profile that are related to c .
- Relative weight assigned to different data sources to take into account the relative importance among them

The data source specific parameters are

- Query history: To change the score of a concept that occurs in the query, the only parameter considered is the concept occurrence frequency. The score of the concept increases with the occurrence frequency.
- Anchor text associated with links: We consider the following parameters to calculate the score for an anchor text concept.
 - Type of the link: The links present in a page have been categorized depending on the usage of the links. For each type of link, different weights have been assigned.
 - Access time of the linked page: The access time for the linked page plays an important role in assigning score to the anchor text concepts. The score increases with the increase in the time of access of the linked page.
 - Concept occurrence frequency: The score of a concept increases with the increase in its occurrence frequency.
- Concept in a page: The score of a concept learned from the contents depends on the following parameters.
 - How interesting the page is: If a page is important to the user, the concepts can be considered to be included in the user profile. Whether a page is interesting to the user or not depends on the time of access of that page and the presence of any mouse activity over the browser during the access of the page.
 - Access time of the page: The score of the concept increases with the access time of the page.
 - Concept occurrence frequency: The score of a concept is high if it occurs frequently in the page.

We will like to use a generic scoring function for the data sources and which should have the following properties.

- The score of a concept should increase asymptotically and attain a maximum value of 1.0.

$$\lim_{u \rightarrow \infty} \text{Score}(c, s) = 1 \quad (1)$$

where c is the concept for which the score is being calculated and s is the usage of the concept.

- The score should have the desired behavior with respect to the parameters that we have discussed.
- The rate of change of slopes of different functions for different data sources will depend upon the relative importance of the data sources and should be reflected in the scoring function.

The scoring function has been developed keeping in mind the above mentioned properties and it is given below.

$$S_c = (S_p + (1 - S_p)\xi)\varpi \quad (2)$$

The current score S_c of a concept is an increment over the previous score S_p of the concept. ϖ is the relative weight for the concerned data source (see Table 2). The increment should be so adjusted that it never exceeds the limit of maximum score for a concept. The factor $(1 - S_p)\xi$ is the increment where $0 < \xi < 1$ and the term $(1 - S_p)\xi$ ensures that S_c never exceeds 1. The factor ξ determines the rate of increment and adjusts the increment factor. The increment adjustment factor ξ is defined as

$$\xi = 1 - e^{-\gamma} \quad (3)$$

The value of γ depends on the values of the parameters of different data sources. As discussed earlier, some of the parameters are common to all the data sources and some are different. The term γ is given by

$$\gamma = \beta + \mu \quad (4)$$

where $0 \leq \gamma \leq 2$. The term β is for common parameters. It accounts for the effect of the presence of the related concepts. Let R be the number of related concept of the concept C in the current user profile and R' is the number of related concepts of C in domain ontology. As β is the measure of the background knowledge of the user for the concept C , we call it as the *past knowledge factor*. The β is defined as

$$\beta = \frac{1 + R}{1 + R'} \quad (5)$$

The other parameters are incorporated into the scoring function through the term μ . Frequency of a concept provides the major contribution in the value of μ . This factor has been named as frequency factor of a concept. The value of μ is always limited to 1.0. The formulation of μ for different sources has been discussed later. Next, we validate the intuitively derived formula against the mentioned properties.

- **Decreasing slope:** The slope of the curve for concept scoring function (Equation 2) decreases over time. Let S_t be the score of a concept at time t and S_{t+1} is the score of the concept at time $t + 1$.

$$\begin{aligned} S_{t+1} &= (S_t + (1 - S_t)\xi)\varpi \\ \frac{dS_{t+1}}{dt} &= \varpi(1 - \xi)\frac{dS_t}{dt} \\ \frac{dS_{t+1}}{dt} &< \frac{dS_t}{dt} \quad \text{as} \quad 0 < \varpi(1 - \xi) < 1 \end{aligned} \quad (6)$$

$\frac{dS_{t+1}}{dt}$ is the slope of the curve at time $t + 1$ and $\frac{dS_t}{dt}$ is the slope of the curve at time t . So, we can conclude that the slope of the curve decreases with time.

- **Effect of relative weight of data sources:** The relative importance of the data sources should affect the scoring function. For the data source with higher importance the curve should grow and converge faster than that with lower importance as shown in Figure 4.

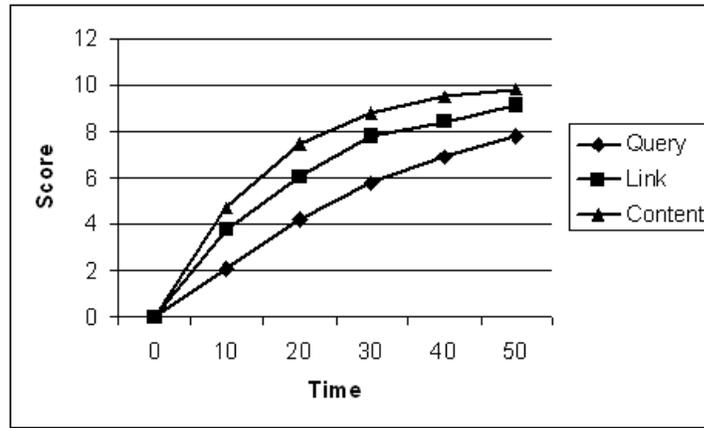


Fig. 4. Effect of relative importance of the data sources on score. Browsing session vs. Score curve for $R' = 2$, $R = 6$; $\mu = 1.0$ for query source, $\mu = 0.7$ for link source and $\mu = 0.5$ for content source

5.3. Query Pattern Monitoring

The query pattern monitor monitors the query placed by the user and assign score to the query concepts depending upon their usage. We assume that a concept occurs once in a query. When a user submits a query, the concepts present in the query are extracted using domain ontology. The scoring function is then applied to the concept vector with the frequency factor μ . The frequency factor μ depends only on the frequency of the concept in the query which is assumed to be 1. After applying the scoring function, we get a list of concepts ordered by their scores.

5.4. Browsing Model and Usage Log

The browsing model of a user is the view of the user's access of the hypertext documents. The usage log is the physical representation of browsing model. The access pattern of the users is an important data source because in most of the cases navigation patterns of the hypertext documents depend on the users intentions. Figure 5 depicts an example browsing model. We have few observations on the

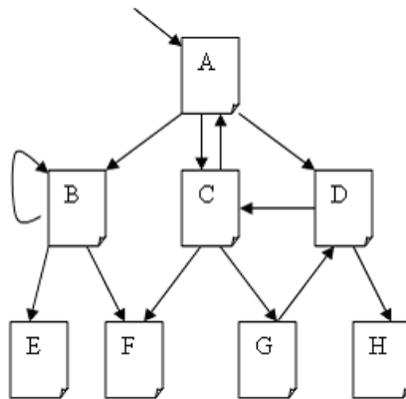


Fig. 5. Graphical representation of an example browsing session.

structure of the hypertext documents.

- One page may refer to the same page. These types of links are called the self referential links. In Figure 5, the page B is being accessed from B itself.
- Hypertext browsers provide the way to go back to the previously viewed page. So, user’s access pattern may include the backward links. In Figure 5, the user goes back to the page A from page C.
- One page may be linked to another page. The users can move from one page to the other linked pages by clicking on the respective links. These types of links are called the forward links. In Figure 5, the page A contains links to page B, C and D.
- One page can be reached through alternate paths. The page H can be reached through following two different paths from the page A.

$A \rightarrow D \rightarrow H$

$A \rightarrow C \rightarrow G \rightarrow D \rightarrow H$

These observations indicate that the access patterns of the users in most of the cases are different form of graphs. We have also made some observations regarding the access behavior of the users.

- When a user finds a topic to be interesting and the current page contains links to that topics, user follows the links.
- The user may access the pages in several depths. In most of the cases, the user goes deeper when the topic discussed in the path is interesting to the user.
- Users spend higher time to the document of her interest as compared to the document she does not like.
- Users perform some mouse operations while they are reading a page.

These observations show how important different features like depth, time of access, mouse activities are in determining user's interest. We have assumed a very simplistic browsing behavior in the part of the user. In this browsing model, all the edges in this graph can be categorized into three types.

- **Self referential links:** Self referential links only provide a way to access a part of the same document. As the same context of the self referenced document has been considered before, there is no need to consider the document again when it is accessed through the self referential links.
- **Backward Links:** This type of links help in accessing previously viewed pages. Back edges have been removed by taking into the account the resumed context. When the user returns back to the parent page, the previous context of the parent page is resumed.
- **Forward Links:** The hyperlinks present in a page can be accessed through this type of links.

Usage logs are important in any system involved in web usage mining activity. We have represented the usage log in the XML format. There are some important clues from usage log that we can exploit:

- The file accessed by the user in her session provides an important clue as the content of the file may contain the concepts of the user's interest.
- The time that a user spends on a particular document is also very useful. Usually, users spend a very little time on a page that does not contain enough relevant or useful information for them and spend a sufficient amount of time for page that contains information of their needs. The idiosyncrasies of the user access behavior should be kept in mind in measuring the time. The user may keep a page opened for a long time while she is busy with some other work. So, two threshold values (upper threshold and lower threshold) have to be set so that effect of these idiosyncrasies can be tackled. Here we set two threshold values, which limit the acceptance of a reasonable access time.
- Mouse activities over the browser provide important information about the users' behavior across the browser [Kamba *et al.* (1995)]. Among them mouse scrolling is very relevant in this context. Time of access for a page may be very large for different reasons
 - User is busy with other work after opening the page.
 - The page is very lengthy.
 - The page is hard enough to take long time for understanding it.

The first case can be handled by upper threshold of time of access. In second and third case, the pages are of importance to the user even if the access time is large enough. These cases can be handled by probing whether there is any mouse scrolling event on the browser.

5.5. Link Analysis

Every link is associated with an anchor text which attracts user to follow that link. Depending on the access of the link by the user, we can identify which of the concepts present in the anchor text can be included in the user profile. Every hyperlink is associated with some anchor text which may provide an idea of what the linked page is all about. A link may contain some concepts which provoke the user to follow the link. After following the link the user may find herself satisfied with the following page. It also may be the case, that the user returns to the parent page without being satisfied. This may occur due to several reasons. The anchor text may contain some ambiguous concepts. In this case, the referred page may talk about some other ambiguous concept that the user has not intended. It may also be possible that the referred page does not belong to the level of the user. The links inside a page may contain some concepts that are not able to grab the attention of the user. In this case, the user may simply ignore the link. These observations motivated us to categorize the hyperlinks in a web page into three different types on the ground of the user's response in accessing web documents.

- ***explored_fruitful***: This category contains the links which have been accessed and are possibly useful to the user. The concepts present in the anchor texts associated with these hyperlinks are of greater importance to the user.
- ***explored_not_fruitful***: This category contains the links that have been accessed but are possibly not interesting to the user. Here, the concepts in the anchor text are interesting enough to provoke the user to explore the link but at the end it does not prove interesting. These concepts must get some lower scores.
- ***unexplored***: These are the links that have not been accessed by the user. The concepts are not attractive enough to prompt user for exploring the link.

To decide upon the type of a link, we consider different parameters like the time of access of the linked page. Time of access has been normalized with the number of words present in the document.

If the page following the link is not present in the browsing tree, it is assigned with the status *unexplored*. If the page following the link is present in the browsing tree, we further categorize it to one of the two categories: *explored_fruitful* and *explored_not_fruitful*. If the normalized time of access of the linked page is within the range of upper threshold and lower threshold, the link is categorized as *explored_fruitful*. One user may spend a large amount of time for several reasons while she is actually reading. In this case the access time may exceed the upper threshold. To decide upon the type of link in this scenario, we rely on the mouse activity over the browser. If the user performs mouse scrolling operation on the browser for this page, then we can conclude that the link is interesting to the user. When the time of

Table 3. Emphasizing factor for different link categories

Category	Emphasizing factor(emph)
explored_fruitful	$\frac{2}{3}$
explored_not_fruitful	$\frac{1}{3}$
unexplored	0

access of the linked page is well below the lower threshold value or the time access exceeds the upper threshold and there are no mouse operations over the browser, the link is categorized into *explored_not_fruitful*. Depending on the importance of the link the concepts present in the anchor text have to be scored. The concepts present in the anchor text of the *explored_fruitful* type of links should get greater weight with respect to the other two categories.

The frequency factor μ for the function of several parameters and is given by

$$\mu = \psi(fr, d, t_a, k) \quad (7)$$

where fr is the frequency of the concept, d is the depth at which the link is accessed, t_a is the normalized time of access of the linked page and k is the category of the link. The depth parameter d is normalized with respect to the maximum depth in the browsing tree. For different types of the links, we assign different emphasizing factor to reflect the relative importance of the link categories. Table 3 gives different emphasizing factors assigned to different categories of links. For each page in the browsing tree, we calculate the frequency factor of each anchor text concepts with the following formula.

$$\mu = fr * (d + t_a + c.emph) \quad (8)$$

5.6. Content Analysis

Here we derive the scores of the concepts that are present in the actual content of a document. The documents consist of several concepts. Among these concepts the system has to learn the concepts that may be of user's interest. Each page in the browsing tree is categorized into interesting or not-interesting category by looking at the access time of the page and mouse activity over the browser. We score the concepts that belong to the interesting pages only.

The frequency factor for the concepts present in the content depends on the frequency of the concept, the depth at which the page is accessed and the time of access of that page. The formula is similar one with the formula for link analysis and is given below

$$\mu = fr * (d + t_a) \quad (9)$$

where fr is the frequency of the concept in a page, d is the normalized depth of the page and t_a is the normalized time of access of the page. fr is computed with the same line of argument in case of anchor text concept. For each page, we obtain

a vector of concept where each concept is associated with its frequency factor. The relative importance for this data source will be incorporated after the accumulation of the concept vectors of all the pages in the browsing tree.

5.7. Score Accumulation

Finally, the concept vectors at individual nodes in the browsing tree are accumulated to get the final concept list learned after a browsing session. The concept vectors are accumulated in a bottom-up manner. The concept vector accumulation process is applied to both types of concept vectors: concept vector learned from anchor texts and concept vector learned from content. The normalized concept vector is given by

$$\bar{C} = \{\mu_1 c_1, \mu_2 c_2, \dots, \mu_n c_n\} \quad (10)$$

$$\mu_i = \frac{\mu'_i}{\max_i \mu'_i} \quad (11)$$

Where μ'_i is the frequency factor of c_i before normalization.

5.8. Concept Age Monitoring

A concept that has not been accessed by the user for a long time is supposed to lose its level of interest to the user and the decrease in the interest level highly depends on the time span during which the concept has not been accessed by the user. This fact has been exploited in personalized the system like ifWeb[Asnicar and Tasso (1997)]. This process is much similar to the process of radioactive decay where the number of atoms present in a radioactive materials decrease over time by means of radiation. So, we can adopt the radioactive decay formula in our system with relevant modifications. The radioactive decay formula is given by

$$N = N_0 2^{-\frac{t}{\lambda}} \quad (12)$$

where N_0 is the number of atoms present initially, N is the number of atoms present at time t and λ is the half-life of the radioactive material.

We use the same formula to calculate the decay in concept score. We define *age* of a concept to be the duration for which it has not been accessed. The age should be normalized with respect to the number of concepts present in the user profile. The more is the number of concepts present in the user profile the less will be the probability of recalling the concept. So, for a concept with given age in a smaller sized user profile will decay faster than that of a concept that is present in a larger sized user profile. We define the ratio of the age and the size of the user profile to be the *persistence factor* (ρ). The persistence factor defines the ability to retain the original score of a concept.

$$\rho = \frac{1}{\left[\frac{\text{age}}{\text{size}(\text{user profile})} \right]^3} \quad (13)$$

The ratio has been raised to the power 3 to adjust the decay rate of a concept. The formula for concept score decay is given below

$$S_c = S_p 2^{-\frac{b}{p}} \quad (14)$$

Where S_c is the current score of the concept, S_p is the previous score of the concept and b is the present browsing session. The score of a concept is considered to decay if the age of the concept exceeds some *threshold age*. In Figure 6, decay curve in a

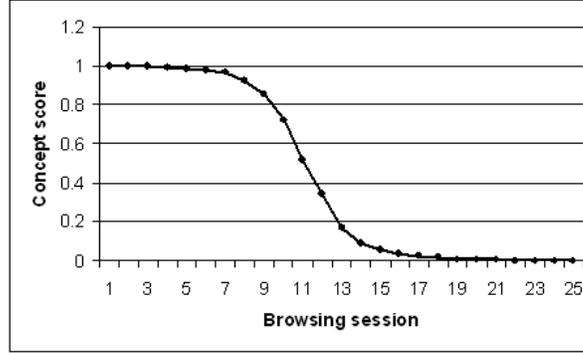


Fig. 6. Concept score decay

simulation run is presented. Size of the user profile is set to 20 and threshold age is set to 5. The initial score of the concept is set to 1.

6. Experiments

The system discovers new concepts that may be of interest to the user. The discovered concepts are ranked according to the level of its interest with respect to the user. The ranked concepts are presented to the user. Our aim is to evaluate the performance of the concept discovery system. Evaluating such a system is a very hard task. We want to evaluate our system on the following points:

- Performance of the user profile acquisition system for users studying different subjects.
- Performance of the system for users with initial user profiles of different size. Two different categories were defined in this regard: one set of users having less than 10 concepts in their initial profile and other category of users with user profiles of size greater than 10 in the initial profile.

6.1. Evaluation Measures

No standard metrics are available for evaluating a concept discovery system. From basic functionalities of the concept discovery system, it is evident that it resembles

a typical information retrieval system where users are presented with documents in response to a query. This motivates us to adopt some of the well established metrics from the set used for evaluating a typical information retrieval system. The evaluation metrics used here are (adopted from [Baeza-Yates and Ribeiro-Neto (2004)])

- **Average Precision (P_a):** Precision (P_{ub}) for a user u and browsing session b is defined as

$$P_{ub} = \frac{\text{No. of relevant concepts}}{\text{No. of recommended concepts}} \times 100\% \quad (15)$$

The average precision (P_a) is the average over all the browsing sessions and users and is given by

$$P_a = \sum_{u=1}^U \sum_{b=1}^B P_{ub} \quad (16)$$

where U is the number of users and B is the number of browsing sessions.

- **Novelty Ratio(η):** Novelty ratio is the ability of the system to provide new items to the user. It is defined as the fraction of the relevant concepts retrieved that was unknown to the user. The average novelty ratio is given by

$$\eta = \sum_{u=1}^U \sum_{b=1}^B \frac{|C_{ub}^n|}{|C_{ub}^o| + |C_{ub}^p|} \quad (17)$$

where C_{ub}^o is the set of recommended concept that was present in the user profile and C_{ub}^p is those which were not present.

6.2. Experimental Setup

Eight students from the 9th standard were selected for our experiment. Biology and Physics were taken as the test subjects. The students were categorized into two subject groups (Physics and Biology) each consisting of 4 students. Each subject group was divided into two user groups: users with small initial profile and users with large initial profile. The students have been monitored for approximately 20 browsing sessions over a period of two weeks. After each browsing session, the updates in the user profiles were recommended to the users. The users' feedbacks were collected.

6.3. Results

The threshold for recommendation is set to 0.2, i.e., concepts with scores greater than 0.2 are presented to the users. With this threshold, we have measured precision of user profile acquisition system averaged over several browsing sessions. Table 4 provides the average precision for all the groups. The average precision over all the

Table 4. Average precision for each group

Group Name	Average Precision (%)
9_physics	83.7
9_biology	81.1

groups is 82.41%.

The novelty ratio of each group is plotted against the browsing sessions. Each figure contains the novelty ratio curves for two candidate users taking one from users with large initial profile and other from user with small initial user profile. The results are shown in the following figures (Figure 7 and Figure 8).

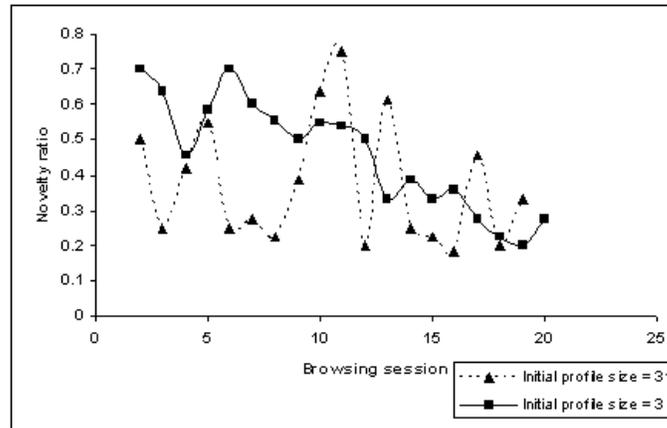


Fig. 7. Novelty ratio of the users in 9_Biology

We make following observations over the novelty curves.

- The novelty ratio depends on the access pattern of the users. The novelty ratio for a user may vary abruptly for the following reasons:
 - In one session, user may search for a topic for which the user profile is well populated. In this case, number of new concepts learned by the system may be less as the documents discuss about those concepts that have been included previously in the user profile.
 - User may search for a topic for which the user profile is not well populated. The number of new concepts learned by the system will be large as most of the concepts encountered in the documents related to the topic may be new to the user.
- With time the user profiles become well populated and the novelty ratio tends to decrease.

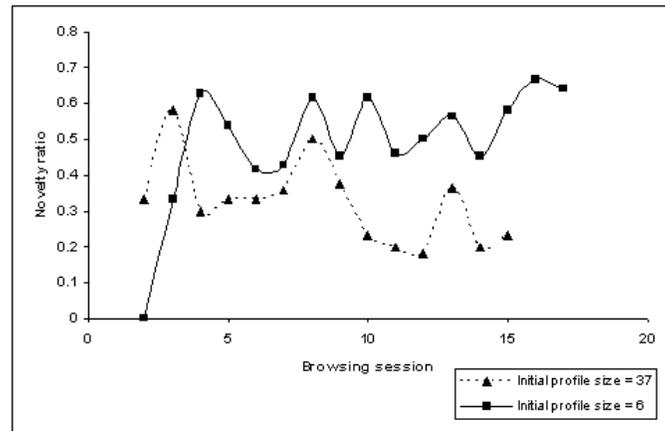


Fig. 8. Novelty ratio of the users in 9_Physics

- In most of the cases, the novelty ratio for a user with large initial profile in a browsing session is less than that of a user with small initial profile.

7. Conclusions

In this work, we have presented an ontology based user profiling strategy to capture the shift in information needs of the users for school curriculum related topics. The ontology for this domain has been represented in a layered structure. The user profile is acquired in two phases. The static phase collects the concepts of user's interest through a profile editor directly. Other implicit sources are utilized by monitoring the activities of the users over the browser. The usage logs collected in the browsing sessions store the link access patterns and the user activities. These usage logs are analyzed to score the concepts. The concept age monitor keeps eye on the usage of the concepts present in the user profile. The scores of the concepts that have not been referred by the user much are decreased.

References

- Aroyo, L., Denaux, R., Dimitrova, V., Pye, M. (2006). Interactive Ontology-Based User Knowledge Acquisition: A Case Study. *European Semantic Web Conference*, 560–574.
- Baeza-Yates, R., Ribeiro-Neto, B.(2004). *Modern Information retrieval*. Pearson Education Pte. Ltd., 482 F.I.E. Patparganj.
- Bhowmick, P.K., Sarkar, S., Chakraborty, S., Sarkar, S., Basu, A. (2007). Samvidha: A ICT system for personalized offline Internet access for rural schools. *Proceedings of the 2nd IEEE/ACM International Conference on Information and Communication Technologies and Development*, Bangalore, December.
- Dai, H., Mobasher, B.(2003). A road map to more effective web personalization: Integrating domain knowledge with web usage mining. *International Conference on Internet Computing*, 58–64.

- Dicheva D., Aroyo, L. (2000). An approach to intelligent information handling in web-based learning environments. *Proceedings of International Conference on Artificial Intelligence*, CSREA Press, 1327–1333.
- Dimitrova, V. (2003). Style-olm: Interactive open learning modeling, *International Journal of Artificial Intelligence in Education* **13**, 35–78.
- Asnicar, F., Tasso, C. (1997). ifweb: a prototype of user model-based intelligent agent for document filtering and navigation in the world wide web. *Proceedings of the workshop Adaptive Systems and User Modeling on the World Wide Web, Sixth International Conference on User Modeling*, 3–12.
- Joachims, T., Freitag, D., and Mitchell, T. (1997). Webwatcher: A tour guide for the world wide web. *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, 770–777.
- Kalfoglou, Y., Domingue, J., Motta, E., Vargas-Vera, M., Shum, S. B. (1999). myPlanet: An ontology-driven web-based personalized news service. *Proceedings of the IJCAI01 workshop on Ontologies and Information Sharing*, 44–52.
- Kamba, T., Bharat, K., Albers, M. C. (1995). The krakatoa chronicle - an interactive, personalized, newspaper on the web. *Proceedings of 4th International WWW Conference*, 159–170.
- Kurki, T., Jokela, S., Sulonen, R. (1999). Agents in delivering personalized content based on semantic metadata, *Proceedings 1999 AAAI Spring Symposium Workshop on Intelligent Agents in Cyberspace*, 84–93.
- Lieberman, H., Letizia: An agent that assists web browsing. *Proceedings of International Conference on Artificial Intelligence*, Morgan Kaufmann, 924–929.
- Middleton, S. E., Alani, H., Shadbolt, N. R., and Roure, D. C. D. (2002). Exploiting synergy between ontologies and recommender systems. *Proceedings of Semantic Web Workshop 2002 At the Eleventh International World Wide Web Conference*, 41–50.
- Pretschner, A., Gauch, S. (1999). Ontology based personalized search. *Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence*, 391–398.
- Razmerita, L., Angehrn, A., Maedche, A. (2003). Ontology based user modeling for knowledge management systems. *Proceedings of the 9th International Conference on User Modeling*, Springer-Verlag, 213–217.
- Yan, T., Garcia-Molina, H. (1995). Sift: A tool for wide-area information dissemination. *Proceedings of the 1995 USENIX Technical Conference*, USENIX Association, 177–186.
- Yu, F. L., Meng, C. W. (2004). Personalized web search for improving retrieval effectiveness. *IEEE Transactions on Knowledge and Data Engineering* **16**, 28–40.
- Zhang, H., Song, Y., Song, H. (2007). Construction of Ontology-Based User Model for Web Personalization. *Proceedings of the 11th international conference on User Modeling*. 67–76.
- Zeng, Q., Zhao, Z., Liang, Y. (2009). Course ontology-based user’s knowledge requirement acquisition from behaviors within e-learning systems. *Computers & Education*, Elsevier Science Ltd. **53**. 809–818