ABSTRACT
In this paper we introduce how to improve a Web-based hybrid recommender system developed with a collaborative bookmark management system approach using semantic capabilities. The system combines content analysis and the development of virtual clusters of students and of educational sources. It provides facilitation in the use of huge amount of digital information stored in a distributed learning environment on the basis of the student’s personal requirements and interests. By adopting a hybrid approach, the system is able to effectively filter relevant resources from a wide heterogeneous environment like the Web, taking advantage of the common interests of the students and also maintaining the benefits provided by content analysis. The resources are processed to extract senses (not just words) from the documents. Then, the classification, recommendation and sharing phases take advantage of the word senses to classify, retrieve and suggest documents with high semantic relevance with respect to the student and resource models. Documents are represented using metadata model.

Categories and Subject Descriptors
H.3.3 [Information search and retrieval]: Clustering, Information filtering, Relevance feedback.

Keywords
Recommendation system, distance learning system, document sense, content analysis.

1. INTRODUCTION
Filter and recommend relevant Learning Objects (LOs) can be useful to address issues like trying to determine the type or the quality of the information suggested from a personalized learning environment. In this context, standard keyword search is of very limited effectiveness. For example, it cannot filter for the type of information (tutorial, applet or demo, review questions, etc.), the level of the information (aimed at secondary school students, graduate students, etc.), the prerequisites for understanding the information, or the quality of the information.

The paper describes how to appropriately gather different agent-based modules which would help students classify domain specific information found on the Web and saved as bookmarks, to recommend these documents to other students with similar interests and to notify periodically new potentially interesting documents. The system is developed to provide immediate portability and visibility from different user locations, enabling the access to personal bookmark repository just by using a web browser.

To obtain a personalized view of (relevant) information a number of Recommender Systems (RS) have been implemented. The RS recommendations are generated using two main techniques: content-based information filtering, and collaborative filtering [2].

If, by one hand, a content-based approach allows to define and maintain an accurate user profile, that is particularly valuable when a user encounters new content, on the other hand it has the limitation of dealing only with textual resources. Moreover, content-based techniques do not depend on having other users in the system but they suffer certain drawbacks, including requiring a source of content information, and not providing much in the way of serendipitous discovery.

Differently, in a collaborative approach, resources are recommended based on the rating of other users of the system with similar interests. As there is no analysis of the item content, collaborative filtering systems can deal with any kind of item, not just limited to textual content. This way, users can receive items with content that is different from the one received in the past. But for a collaborative system to work well several users must evaluate each item. So, new items cannot be recommended until some users have taken the time to evaluate them and new users cannot receive recommendation until the system has acquired some information about the new user in order to make personalized predictions. These limitations often referred to as the sparsity and start-up problems [8].

By adopting a hybrid approach, the system is able to effectively filter relevant resources from a wide heterogeneous environment like the Web, taking advantage of the common interests of the users and also maintaining the benefits provided by content analysis.

Usually, this idea has been widely and successfully developed for specific domains, like movie or film recommendations, and rarely used for recommending LOs. Our system uses a hybrid approach and suitable representations of both available information sources and user’s interests in order to match as accurate as possible, user information needs, as expressed in his query, and available information. In the next paragraph, we describe how to consider LO’s structure to properly represent learning material and extract metadata information.
The starting point is the use of statistical information extraction and natural language parsing techniques to automatically derive classificatory and metadata information from primarily textual data (web pages, Word, postscript or similar documents, etc.). While still challenging for large ontologies, text classification methods which semantically categorize an entire document are now relatively well-understood, and provide a good level of performance.

The paper is organized as follows. Firstly, it illustrates the recommendation process that offers interesting opportunities to introduce the need for personalized criteria. In fact, personalized information classification and filtering facilities are essential basics to use the huge amount of digital information according to the student’s personal requirements and interests. In this way, data can be obtained about which material is proving to be most effective in raising student achievement. Following, the paper introduces how to extract and use senses from the documents. Finally, the conclusion of the paper is presented.

2. THE E-LEARNING RECOMMENDATION

In our system, we have adopted a representation based on the Vector Space Model (VSM), the most frequently used in Information Retrieval (IR) and text learning. Since the resources of the system are Web pages, to obtain a vector representation it has been necessary to apply a sequence of contextual processing to the source code of the pages. To filter information resources, according to user interests, we must have a common representation for both the users and the resources. This knowledge representation model must be expressive enough to synthetically and significantly describe the information content. The use of the VSM allows to update the user profile in accordance to consulted information resources [10].

The system includes a process of classification and recommendation feedback, in which the user agent learns from the student and consequently adapts itself according to the changes in his interest; this gives the agent the chance to be more accurate in the following classification and recommendation steps. Thus, a high number of students using the system would make the following agent’s actions more accurate. This allows the system to be capable of reflecting continuous ongoing changes of the practices of its member, as required by a cooperative framework.

To guarantee a personalized framework, the system needs to construct and maintain user profiles. For a particular user, it is reasonable to think that processing a set of correctly classified relevant and inappropriate documents from a certain domain of interest, may lead to identify the set of relevant keywords for that domain at a certain time. Thus, the user domain specific sets of relevant features, called prototypes, may be used to learn how to classify documents. In particular, to consider the peculiarity of positive and negative examples, we define positive prototype for a class $c_j$, a user $u_i$ at time $t$, as a finite set of unique indexing terms, chosen to be relevant for $c_j$, up to time $t$. Then, we define negative prototype as a subset of the corresponding positive prototype, whereas each element can be found at least once in the set of documents classified as negative examples for class $c_j$. Positive examples for a specific user $u_i$ and for a class $c_j$, are represented by the documents explicitly registered or accepted by $u_i$ in $c_j$, while negative example are either deleted bookmarks, misclassified bookmarks or rejected bookmarks that happens to be classified into $c_j$.

According to traditional IR, we must state our classification problem as a combination of Text Categorization and Relevance Feedback. Detailed information about technical characteristics of described system can be found in [3] and [1].

The automatic recommendation of relevant learning objects is obtained considering student and learning material profiles and adopting filtering criteria based on the value of selected metadata fields. Our experiments are based on SCORM compliant LOs. For example, we use the student’s knowledge of domain concept to avoid recommendation of highly technical papers to a beginner student or popular-magazine articles to a senior graduate student. For each student, the system evaluates and updates his skill and technical expertise levels. The pre-processing component developed to analyze the information maintained in LOs is able to produce a vector representation based on term weighting that can be used by the collaborative recommendation system described in previous paragraphs.

In SCORM, the organization and learning resources must be included with the course and placed in an XML file with the name imsmanifest.xml. The structure required for this file is detailed in the SCORM content aggregation specification[1]; it consists of four sections:

- a preamble section containing XML pointers to the schemas required for validating this file,
- a metadata section contain global course information, such as its title or its description,
- an organizations section describing course sequencing, and
- a resources section listing all the files used in the course

To obtain the loading of some didactical source and its classification, we analyze the imsmanifest.xml file to extract .htm and .html files and examine the content. We consider the following metadata to provide the corresponding technical level:

- difficulty: represents the complexity of the learning material, ranging from “very easy” to “very difficult”;
- interactivity level: represents the interactive format, ranging from “very low” (only static content) to “very high”;
- intended end user role: represents the user type (for example student or teacher);
- context: represents the instructional level necessary to take up LO.

The difficulty level is explicitly loaded into our database (in most cases, LMSs use this value). Difficulty and the other values are combined to characterize technical level of learning material ranging from 0 to 1 and representing how demanding is the LO. If some of these fields are not present in the manifest file (they are not required), we consider their average value.

Our system also considers the user’s skills to express cleverness as regards different categories. This value ranges from 0 to 1 and it depends initially on the context chosen from the user during

---

1 [http://www.adlnet.org](http://www.adlnet.org)
his/her registration (primary education, university level, and so on). During the creation of a new category (for example, when a lesson is saved) we consider the user’s skill value equal to the resource technical level, presuming that if a user saves a learning material then he could be able to make use of it. The user’s skill level is updated when a new resource is saved, taking into account its technical level and the user’s skills in that category. Starting value for user’s skills parameter, its update frequency, the increment or decrement value and the difference between technical level and user’s skills necessary to obtain a recommendation outcome from the following experimental tests. They are easily adaptable, though.

Despite SCORM is widely used in learning environments, its presence in the web is very little; furthermore, most of the LOs that are published are not free. So, we have created a SCORM compliant learning material using the abstract of hundred and hundred of papers in .html version from scientific journals published on the web. We have linked an imsmanifest file to each paper. Then, we have simulated ten users with different initial profiles (based on the field of interest and the skill level) and saved, in four turns, ten learning resources for each user, obtaining 400 LOs.

Figure 3 and 4 show the index.html and the imsmanifest.xml file corresponding to recommended learning object highlighted in Figure 2.

The precision value of recommendation phase, calculated considering the resource difficulty and the user’s skills, exceeds 80%. It is important to note that the recommendation is made using skill and technical levels (the resource difficulty is one of the parameters, even if the most meaningful); in fact, users rarely know the technical level of a recommended resource; they rather know about its difficulty. Moreover, the categories of recommended learning material correspond to the user’s interests, in almost all of the tests carried out.

Fig. 1 shows the personal bookmark page, after the student has logged into the system: the bookmarks are organized in the categories automatically proposed or chosen during the registration of an interesting page (a) and the user can check the received recommendation (“you have a new recommended learning object” in (b)). The student can either accept or reject when he is notified of such recommendation (Fig. 2).

3. SEMANTIC CAPABILITIES

The Semantic Web will add meaning, or semantics, to Web content in order to make it easier to find and use information for both humans and machines. Adding formal semantics to the Web will aid in everything from searching the web, to resource discovery, to the automation of all sorts of tasks [7].

The system we have designed consists in a Web application that provides an on-line bookmarking service. The InLinx (Intelligent Links) system is the result of the integration of three filtering components, corresponding to as many functionalities:
Figure 2. Learning object recommendation

Figure 3. index.html page of recommended learning object
support to bookmark classification (content-based filtering): the system suggests to the user the more suitable category in which he can save the bookmark, based on the document’s content; the user can accept the suggestion or change the classification, selecting another category that he considers best for the given item;

• bookmark sharing (collaborative filtering): the system checks for newly classified bookmarks and recommends them to other users with similar interests. Recipient users can either accept or reject the recommendation when they receive the notification;

• paper recommendation (content-based recommendation): periodically, the system checks if a new issue of some online journal has been released; then, it recommends the plausible appealing documents, accordingly to the user profiles.

Our aim is to improve the three components of the recommender system utilizing concepts for classification, recommendation and document sharing. Generally, recommender system uses keywords to represent both the users and the resources. Another way to handle such data is to use hierarchical concept categories. This will enable users and system to search, handle or read only those concept of interest to them in a more general manner, providing semantic possibility. For example, synonymy and hyponymy can reveal hidden similarities, potentially leading to better classifiers and recommends. Also the typical search engine is not even aware of the existence of the conceptual hierarchy.

As the user saves the documents, the system builds the resource representation as a semantic network whose nodes represent senses (not just words) of the documents. Then, the classification, recommendation and sharing phases take advantage of the word senses to classify, retrieve and suggest documents with high semantic relevance with respect to the user and resource models.

Applied to each document that the system manage, this means that we want to find a subset of words, which helps to discriminate between concepts. In this way, two documents or two users characterized using different keywords may result similar considering underlying concept and not the exact terms. We documents are the collection text written in natural language. To extract important information from documents, we use the following feature extraction pre-process. Firstly, we label occurrences of each word in the document as a part of speech (POS) in grammar. This POS tagger discriminates the POS in grammar of each word in a sentence. After labelling all the words, we select those labelled as noun and verbs as our candidates. We then use the stemmer to reduce variants of the same root word to a common concept and filter the stop words.

A vocabulary problem exists when a term is present in several concepts; determining the correct concept for an ambiguous word is difficult, as is deciding the concept of a document containing several ambiguous terms. To handle the word sense disambiguation problem we intend to use distance or similarity measures based on WordNet. Budanitsky and Hirst [4] give an overview of five measures, and evaluate their performance using a word association task [9].

WordNet [6] is an online lexical reference system, in which English nouns, verbs, adjectives, and adverbs are organized into synonym sets. Each synset represents one sense, that is one underlying lexical concept. Different relations link the synonym sets, such as IS-A for verbs and nouns, IS-PART-OF for nouns,
etc. Verbs and noun senses are organized in hierarchies forming a “forest” of trees. For each keyword in WordNet, we can have a set of senses and, in the case of nouns and verbs, a generalization path from each sense to the root sense of the hierarchy. WordNet could be used as a useful resource with respect to the semantic tagging process and has so far been used in various applications including Information Retrieval, Word Sense Disambiguation, Text and Document Classification and many others.

Generally, the proposed distance or similarity measures are applicable to the hyponymy relations (the IS-A or HAS-PART relation in WordNet), that is to the syntactic categories of noun and verb. After the nouns and verbs sense disambiguation based on WordNet, our documents are represented using a matrix storage format containing a row for each WordNet subject code and a column for each WordNet object code. The position i, j of the matrix maintains a list, eventually empty, containing the WN verb code that relate the subject i and the object j and the number of times the triple is present in the document. For example,

\[
\text{SUBj} - \text{LIST}\{\text{VERBi}, j \#\text{occurrences}\} - \text{OBJj}
\]

GATE provides a number of useful and easily customizable components, grouped together to form the ANNIE (A Nearly-New Information Extraction) component. These components eliminate the need for users to keep re-implementing frequently needed algorithms and provide a good starting point for new applications [5]. These components implement various tasks from tokenization to semantic tagging and co-reference, with an emphasis on efficiency, robustness, and low-overhead portability, rather than full parsing and deep semantic analysis.

4. CONSIDERATIONS

The paper shows how the integration of content-based and collaborative approaches in a virtual community setting has a great impact on the quality of the service. Thanks to the bookmark sharing and recommendation facility, the system contributes to human collaborative works, as it supports group collaboration among people involved in a work process, independently of time and space distance, and learns from positive and negative experience in group practice. In fact, recommendation systems can help learners by collaboratively eliminating irrelevant information, operating like mediators between the sources of information, the learning management system, and the learners. We believe that the described system can find application in any context in which the group collaboration is a requisite, and a Web-based learning system is an ideal application domain.

The paper addresses issues like trying to determine the type or the quality of the suggested information. The automatic recommendation of relevant learning objects is obtained considering student and learning material profiles, adopting filtering criteria based on the value of selected metadata fields and capturing not only structural but also semantics information. Our experiments to test the system's functionality are based on SCORM compliant LOs; we use artificial learners to get a flavor of how the system works.

Summarizing, the key elements of the described system could be highlighted as follows. The system provides immediate portability and visibility from different user locations, enabling the access to personal bookmark repository just by using a web browser. The system assists students in finding relevant reading material providing personalized learning object recommendation. The system directly benefits from existing repositories of learning material providing access to open huge amount of digital information. The system reflects continuous ongoing changes of the practices of its member, as required by a cooperative framework.

5. REFERENCES


