International Journal of Computer Science and Applications
© Technomathematics Research Foundation
Vol. 7 No. 2, pp. 30 - 51, 2010

A FRAMEWORK FOR MANUAL ONTOLOGY ENGINEERING FOR MANAGEMENT OF LEARNING MATERIAL REPOSITORY

PLABAN KUMAR BHOWMICK*

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

DEVSHRI ROY

Computer & Informatics Center, Indian Institute of Technology, Kharagpur
West Bengal, India-721302
droy.iit@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

India is a country of diversity in language, culture and socio-economic conditions. In order to make use the cutting edge computer aided education technology in this setting, efforts have to be made for proper management of the learning materials by addressing this diversity. In this paper, we present a framework for manual ontology engineering in education domain for managing learning materials of the curriculum related requirements of school students. We identify an effective way of structuring the knowledge about education domain in a layered architecture. The layered architecture allows us to clearly demarcate the roles of topics, concepts and actual words in a multilingual setting. With the help of the constructed ontology, we provide scheme for indexing educational materials into different layers of ontology for the management of learning materials. Domain specific and personalized search interfaces based on the created index structure have also been provided.

Keywords: knowledge representation; ontology management; personalized retrieval; repository management; computer aided learning; information filtering; ontology building tool.

*corresponding author

1. Introduction

Among many roles of computers towards social causes, computer aided education is gaining importance in recent era. It has an important role to play in developing countries like India for the following reasons.

- Shortage of teachers specifically in rural areas.
- Higher dropout rate due to traditional and boring text book teaching.
- Harvesting the benefits provided by multimedia based learning materials.

Diversity in languages and culture in India is one of the main hindrances in percolating computer aided education in this country. Developing language specific solution is not effective to address this problem. There is a need to develop a common framework that will unify this diversity. A framework that make use of the participation of the local experts may help this cause. This motivated us to develop a participatory authoring environment that will help the knowledge experts (e.g., teachers) to design vernacular based course structure and attach localized learning materials into the course structure.

Internet is a diverse repository of learning materials ranging from text to multimedia contents. Making use of these contents may help in supplementing the shortage of electronic contents. But Internet is an ever increasing database of structured and unstructured data ranging from audio, video and text. In response to a query, a huge number of results are returned by a general purpose search engine. In most of the cases, the results are ranked independent of the user and the domain. For example, in response to the query *reflection*, a standard search engine returns the documents from varying domains like physics, Java, marketing etc. This is because the keyword *reflection* has different meanings in different context. If a system has to provide only relevant documents, it needs to disambiguate the sense of the query words and identify the exact concepts the query words refer to. This requires the knowledge about the domain in concern to be stored in a machine processable way.

Ontology refers to the shared understanding of a domain of interest and is represented by a set of domain relevant concepts, the relationships among the concepts, functions and instances. The most commonly used definition of ontology among the knowledge engineering community is of Gruber[Gruber (1993)]: *An ontology is a formal, explicit specification of shared conceptualization.*

Conceptualization is the abstract representation of a real world entity with the help of domain relevant concepts. Ontology should be *formal* so that it becomes machine understandable and it should have to enable *shared* communication across the communities. Ontology can be viewed as a vocabulary containing formal description of terms and a set of relationships among the domain relevant concepts.

Currently, ontology has emerged as a very important discipline as its usefulness has been demonstrated in varying types of applications which include information organization and extraction, personalization, natural language processing, artificial intelligence, knowledge representation and acquisition. Ontology is going to play a

major role in the evolution process of the WWW to the *Semantic Web*, the second generation web.

Ontology management tools provide the facilities and environments to build a new ontology from scratch, modify existing ontologies, reuse other ontologies and also provide a visual interface for viewing ontology. The authoring interface must be simple enough so that a domain expert having very scanty technological expertise can build the ontology of her domain efficiently. Visual interface is very important as through the visual interface other domain experts can cross validate one created ontology easily. Ontology representation language is an important issue in developing ontology building tools as one of the primary objectives of ontology is to enable shared communication and a set of standard ontology representation languages helps in this regard. The knowledge requirement in school education domain may vary with the needs of target student groups. So, knowledge structure developed for one student group may not address the requirement of another student group. Again the learning materials are effective when they are tuned to the cultural orientations of the target groups. These issues are very important in the context of e-learning and intelligent tutoring system. For vernacular based education, rendering of knowledge structure in local language is very important. So, a framework should be provided where the experts can easily maintain knowledge structures in vernaculars and attaching learning materials which are localized culturally. This participatory authoring of knowledge structure may prove to be effective in countries like India having diversity in language and culture.

In the next section, we describe different knowledge modeling paradigms and knowledge structure management tools that did not address the following issues.

- Knowledge structure model tuned to education domain.
- Manual and automatic indexing of learning materials into the knowledge structure.
- Multilingual knowledge structure authoring.
- Easy interface for domain experts having little technological expertise.

Above mentioned shortcomings of the existing knowledge structure management tools motivated us to develop a new one. In this paper, we provide a multilingual framework for management of knowledge structures of such domains in a participatory way.

2. Related Works

In this section, we shall provide a brief introduction to the previous works related to this work.

2.1. *Related Works in Ontology Modeling*

Several ontologies have been developed in various domains and for varying purposes. They differ in the way the ontology is structured, the ontology representation lan-

guage that has been used to represent the ontology and the application domain they are targeted for. These ontologies can be categorized into several groups as follows:

Top level ontologies define the concepts from the universe like entity, event, feature, etc. It is articulated in the sense that distinction is made where it is necessary. In Sowa's Top Level Ontology[Sowa (1995)], the categories and distinctions are derived from the sources like logic, linguistics, artificial intelligence and philosophy. The ontology has a lattice structure where the top level concept is of universal type and the bottom level concepts are absurd type. The primitive concepts are taken from the set: independent, relative, mediating, contonuant, occurant. The purpose of Standard Upper Merged Ontology[Niles *et al.* (2001)] is to make computers utilize the ontology in the applications needing data interoperability, information search and retrieval, natural language processing. It is implemented in DAML+OIL [McGuinness *et al.* (2002)]. Other top level ontologies are Upper Cyc ontology[Lenat (1995)], Wordnet top level ontology[Gangemi *et al.* (2001)].

Domain Ontologies capture the knowledge of the domain. Several domain ontologies for different domains have been developed. CHEMICALS[Fernandez-Lopez *et al.* (1999)] is an ontology representing the domain of chemical elements and crystalline structures. This ontology is represented with Ontolingua. This has been used in the projects like OntoGeneration and ChamicalOntoAgent. The domain of pollutant chemical materials is represented by the ontology Environmental Pollutants. It captures the pollutant chemical materials in different media like water, air soil etc. Ontology representation language here is XML.

Linguistic Ontologies capture the semantics of the grammatical units. Wordnet[Goerge (1995)] is the largest lexical database for English. It is categorized into synsets each representing one lexical concept. The synsets are related to each other by a set of linguistic relationships like hypernymy and hyponymy, meronymy and holonymy, hynonymy and antonymy. Wordnet represents lexical entries into five categories: nouns, verbs, adjectives, adverbs and functional words. It is used in natural language processing based applications. SENSUS[Knight and Luk (1994)] is a natural language based ontology developed for the machine translation project by ISI. It is a hierarchically structured concept base. The Ontology Base at the upper level represents the generalization needed in the process of translation. The middle level represents the model of the world by storing English word senses and the bottom most level have concepts representing anchor points for different applications.

2.2. *Related Works in Knowledge Structure Management Tools*

Ontolgy building tools provide framework for manual or automatic ontology engineering. Some of these ontology building tools have been described below.

Protege[Gennari *et al.* (2002)] helps knowledge engineers and domain experts to perform knowledge management tasks. It includes support for class and class hierarchy with multiple inheritances, slots having cardinality restrictions, default values, inverse slots, metaclass and metaclass hierarchy. It supports easy naviga-

tion through the class hierarchy through tree controls. The knowledge model is OKBC compatible. The distinguishing features in Protege are the scalability and extensibility.

OntoEdit[Sure *et al.* (2002)] provides an environment for the development and modification of ontologies with the help of a graphical user interface. The concept hierarchy can be edited or created in which the concepts may be abstract or concrete. The decision of making direct instances of a concept depends upon the type of the concept. The support for handling synonymous concepts is evident. It has support for several plugins for including domain lexicons, inference engine and import-export facilities.

WebODE[Vega (2000)] is an advanced ontological engineering workbench that provides varied ontology related services, and gives support to most of the activities involved in the ontology development process. It has been developed using a three tier model having Client Tier, Middle Tier and Database Tier. The main components of WebODE ontology are concepts, groups of concepts, taxonomies (single and multiple inheritance) ad-hoc relations, constants, formulae, instances (of concepts and relations) and references.

LinkFactory[Ceusters and Martens (2001)] is an ontology management system that provides a way to create and manage large scale, complex, multilingual and formal ontologies. It has been used to develop the medical linguistic knowledge base LinKBase. The LinKFactory Server, and the LinKFactory Workbench (client-side component) are two major components of the system. The user can customize the tool to view and manage the ontology. Different views of the ontology like Concept tree, Concept criteria and full definitions, Linktype tree, Criteria list, Term list, Search pane, Properties panel, Reverse relations are managed through Java beans. The server is responsible for storing the ontology physically into a relational database implemented in Oracle. The access to the database is abstracted by some intuitive APIs like get-children, find-path, join concepts, get terms for concept X.

Two other tools relevant to this context are: OilEd[Bechhofer *et al.* (2001)], Ontolingua Server[Farquhar *et al.* (1996)].

3. Motivation

The domain of our interest is the school education domain, i.e., school curriculum related topics. This domain is very structured compared to the other domains. The specific features of this domain are

- The school curriculum related topics are organized under several subjects like Biology, Geography, Physics, Chemistry, History etc.
- Each subject consists of several chapters. Each chapter deals with a topic. Each chapter may again be divided into several sub-chapters. For example, the chapter *nutrition* in Biology may consist of two sub-chapters *plant nutrition* and *animal nutrition*.
- Each chapter or sub-chapter contains materials that include discussions on

various concepts. These. For example, the chapter *light* in Physics contains *ray*, *incident ray*, *reflection*, *refraction*, etc.
- It is often the case that the same concept is dealt with in different chapters. For example, the concept *reflection* belongs to both the chapters *light* and *sound*.
- Every concept refers to a semantically distinct entity. The concepts in a domain are related to each other through different relationships. For example, the concept *reflection* is related to the concepts *incident ray*, *plane mirror*, *reflected ray*, *angle of incidence*, etc. We often find that documents containing the concept *reflection* contain some of these other related concepts.
- Different types of relationships may exist. For example, *microscope* has part *objective*, *mechanical reaction* is prerequisite for *Newton's laws of motion*. The association between the related concepts may vary. For example, association between the concept *shunt* and *ammeter* is higher than that between *ammeter* and *parallel circuit*.
- The phenomenon of synonymy is very common. So the same concept may be referred to by several terms. For example, the terms *DC* and *direct current* refers to the same concept that indicates non-alternating current.
- The same concept can be represented by different words in different languages.

Analyzing the specific features of the domain, we have identified the requirements for representing the domain knowledge. The requirements are

- The representation should be non-monolithic consisting of distinct layers for different entities (chapter-subchapter, concepts, and terms) in the domain.
- The representation should provide efficient means to map the entities of one layer to the other layer.
- To identify relevant documents for school related topics, a system not only needs to have the taxonomy of the topics but also some other important features. For example, the system should analyze the documents to judge whether it is understandable by the targeted student or group.
- Finding information at the concept level is very important to reduce the redundancy occurring due to the synonymous ambiguity between the terms.
- Different types of relationships may be used differently in systems that make use of the domain knowledge.
- For applications like multilingual tutorial system, the ontology has to be developed in a multilingual environment.

The ontology management tools discussed in section 2 provide robust and multifaceted framework for the management of ontologies of different domains. These tools are widely accepted in different knowledge management communities. But there are several issues that have not been addressed when we focus on the Indian

perspective. We aim at providing a tool to manage domain for any subject belonging to school curriculum. As stated earlier, these domains are mostly well structured. The knowledge management tool for these domains should meet some basic needs which are listed below.

- The majority of the school teachers in rural India have got a very limited exposure in the usage of the computer systems. But the teachers are the best experts to build the domain knowledge for school curriculum related topics. So, an intuitive and very easy to use interface has to be provided through which the actual domain experts with very limited computer usage skill can build a very robust ontology in different subjects. In this context, an important issue is the tradeoff between the simplicity of the user interface versus its flexibility. Simple interfaces are easy to learn and most suitable for the novice users but it lacks the flexibility. Flexible interfaces on the other hand, are more powerful but it takes more time to learn the interface and needs a great deal of expertise in the part of the user. Here, we have focused on the simplicity of the user interface keeping in view the targeted users.
- The role of visualization of information is well understood in the design of human-computer interface. The visual representation of the objects sometimes helps in rapid communication of the information compared to other representation like textual representation. So, there is a need for intuitive visualization of the ontology to reduce the congnitive load. The visualization can be represented as a graph, or a tree structure.
- Ontology provides a well defined structure of the underlying domain i.e., the structure of the knowledge is reflected through ontology. The actual information or detailed properties of the entities are stored in a very limited way. In the applications like tutoring systems, the detailed information about the entities in the domain may have to be stored with the structure of the knowledge. This requirement leads to the need for indexing of information at every level of the ontology hierarchy.
- In a multilingual country like India, there is a need for an environment through which the domain knowledge can be specified and rendered in vernacular.

Comparing the above mentioned requirements and the features present in the ontology management tools described in section 2, we can enumerate the shortcomings of the described tools.

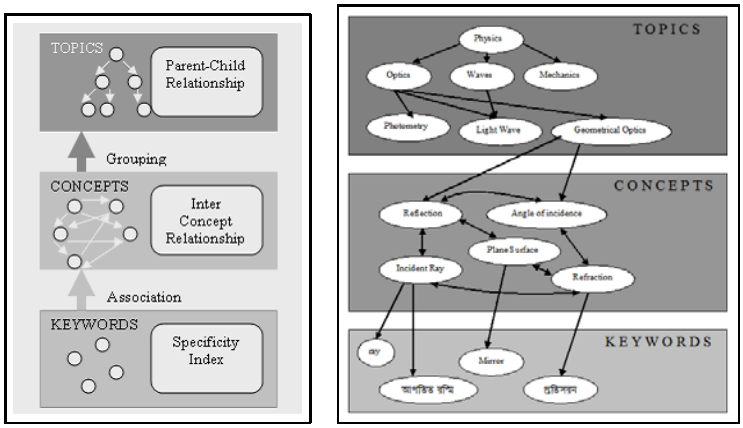
- *Requirements for knowledge structure modeling:* The previously described systems are not tuned to provide the environment where specially structured ontology in education domain can be authored intuitively. Protege supports only the class-subclass relationships. LinkFactory provides some specialized relationships among the concepts. Again these relationships are

not sufficient to define knowledge structure of education domain.

- *Automatic content indexing*: They do not provide the facility of indexing contents with the entities present in ontology. This is important for efficient and automated content organization.
- *Manual content indexing*: As specified earlier, content of the learning materials may be influenced by culture, geographical features and other factors that are specific to a particular locality or community. The learning materials may be best utilized if they are authored in vernaculars. A framework that enables the organization of locally specific learning materials in a participatory manner may be effective in vernacular based education.
- *Multilingual framework*: No system except LinkFactory provides the platform where knowledge can be authored in multilingual environment. LinkFactory is designed to author ontologies in European languages whereas here our focus is on authoring knowledge structure in Indian languages.

4. Knowledge Model

To meet the above mentioned requirements we have proposed an three tier knowledge model to represent the domain knowledge in education domain.



(a) Three tier structure of ontology (b) A small section of physics ontology

Fig. 1. Ontology model and example ontology in education domain.

The properties and characteristics of the entities however do not need to be stored in the system. The structure of knowledge is to be stored, and not the knowledge itself. The knowledge representation database is organized into a three level hierarchical structure as shown in Figure 1(a). The concept level in the middle tier represents the ontology of the domains consisting of domain concepts and domain dependent inter-concept relationships. The topmost layer is a collection of

school curriculum related topics. The topics share a parent child relation. The concepts in the ontological layer are grouped into topics. This reflects that a topic is described by the concepts that are grouped in that particular topic. The third layer consists of raw keywords. The keywords are associated with concepts by defining some association values. A small section of physics ontology is shown in Figure 1(b).

4.1. *Concept Level*

This layer consists of the ontological concepts. A set of empirical relations can be defined among the concepts in a domain. We notice that if a concept is of significance in a document, it is usually the case that the document contains a number of references to related concepts. The breadth and depth of the ontology is used by the ranking algorithm because concepts that are directly and remotely connected to the concepts in the query are used for the calculation of the document scores. In fact the occurrence of related concepts is taken as a very strong indication of the relevance of the document. Pages that do not contain related concepts are suspect and may be spurious. The relations that are stored in the ontology become very important for this reason. In order to keep the system simple the relations must be broad and general. The relation list chosen must also cover most important forms of relations that occur so that the ranking process has a sufficiently good ontological web. For example, if a document contains material relevant to reflection in optics, it will have references to some of the related concepts like light, ray, mirror, lens, angle of incidence, etc.

To capture the strength of a relation, we introduce the notion of distance between two concepts. This distance between two concepts is not symmetric. These distances have been devised and tuned experimentally for each domain. 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. Each edge is assigned a weight depending upon the relation by which two concepts are related by this edge. The weight is an indication of the strength of the relationship.

A set of 11 relations is found to cover most types of relations between two concepts. Thus the lengths of the relations are different in different domains. The relations are explained below:

- *Has Part & Part Of*: These relationships reflect the meronym and holonym relation between two concepts. For example, in Physics domain, the concept *microscope* contains *objective* as its part.
- *Inherited From & Parent Of*: The hypernym and hyponym relations are reflected in these relationships. For example, *electron microscope* inherits some properties from the more generalized concept *microscope*.
- *Has Prerequisite & Prerequisite For*: Sometimes, to grasp some idea about a concept, we need to know some other concepts. These two categories of concepts are connected via Has Prerequisite & Prerequisite for relations. For example, to learn the concept *lens* we should have some idea about the

concept *image*.

- *Functionally Related*: To derive the concept *radius of curvature* we need to derive the concept *focal length* and vice versa. Thus these concepts are related by this relation. In this case, the reverse relation is same as the forward relation.
- *Part Of Procedure & Procedure Contains*: In this relation, a concept is a part of procedure represented by the other concept. For example, in Biology, the concept *anaphase* is a part of the process of *cell division*.
- *Is Caused By & Causes*: In this type of relationship, one concept is the effect of the occurrence of the other connected concept. In the domain of Biology, the disease *ricket* is caused by the insufficiency of *Vitamin D*.

The relations explained above provide a way of storing the structure of a domain without storing any information about a particular concept. These relations make it possible to find the concepts that are close to a particular concept and this information may be used in many ways.

4.2. *Topic-Subtopic Level*

4.2.1. *Level Structure*

On the top level, the topics share a *containment* 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. In this knowledge model, the chapter-subchapter organization is reflected through topic-subtopic relations.

4.2.2. *Topic-Concept Relation*

One topic in the school curriculum is described by a set of concepts. So in our knowledge structure, a subset of ontological concepts are grouped into a particular topic.

4.3. *Keyword Level*

4.3.1. *Level Structure*

This level contains a set of keywords of each domain. These keywords are associated to concepts in the ontology. 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. Secondly, the keywords from several languages with the same meaning can be mapped to a

common concept. So, multilingual search can be provided by the help of a concept based search technique.

4.3.2. *Keyword-Concept Relation*

The keywords are associated to the concepts with a specificity index. This specificity index reflects the likelihood of the keywords representing a particular concept.

5. Ontology Management Tool

In this section, we describe our ontology management tool. This tool provides an easy and intuitive environment for authoring or modifying knowledge structure in the education domain. The tool consists of the following interfaces.

5.1. *Knowledge Structure Builder Interface*

It provides the interface through which knowledge structures for new subjects can be created from scratch and existing knowledge structure of one subject can be modified. Figure 2 provides a screenshot of this interface. For the layered representation of knowledge, this interface has been divided into four different sections. Each section provides the facilities for editing one hierarchy. The *Topic management interface* provides the menus to create, modify topics, adding concepts to a topic and adding documents to a topic. *Concept modification interface* provides the menus to modify concepts, adding keyword to concept, adding document to concept. The Related concept interface provides the view of the related concepts. The *Keyword management interface* provides the menus to modify, delete keywords, adding document to keyword. The Search window provides options search for topics, concepts and keywords.

5.2. *Knowledge Structure Browser Interface*

This interface provides the facility of browsing the knowledge structure in a top-to-bottom fashion. This interface can be rendered through the language selected by the user in the initial setting. The interface is shown in Figure 3. This interface contains six main windows:

- *Topic Browser*: Topics can be browsed through this window by a tree view. All the topics are categorized into a root topic called Subject. Each node in the topic tree can be expanded or collapsed if it has some topics as its children. Each single click on a topic node invokes two events: the Concept Browser gets refreshed, all the concepts under this topic are displayed and the documents indexed to this topic appear in the Document Window.
- *Concept Browser*: All the concepts related to a topic are rendered in a list view through this window. One click on each concept node invokes three

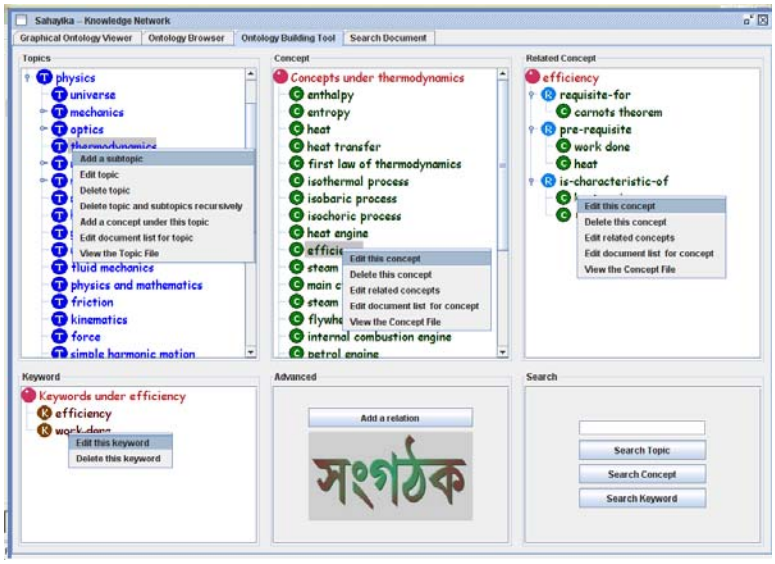


Fig. 2. Knowledge structure builder interface in English.

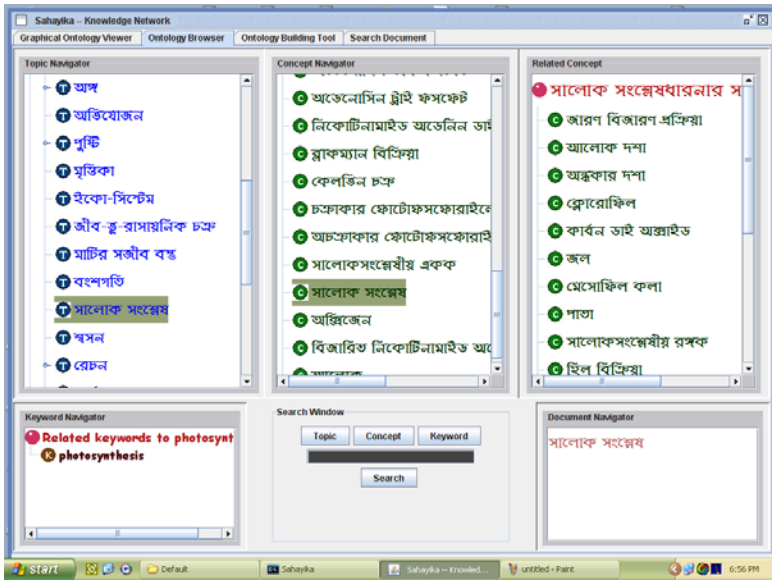


Fig. 3. Knowledge structure browser interface in Bengali.

different events: all the concepts related to the clicked concept appear in the Related Concept Browser, the keywords associated to the concept are displayed through the Keyword Browser and the documents for the concept

are rendered through the Document Window.

- *Related Concept Browser*: Related concepts to a particular concept are displayed in a list view through this window. The events associated with this window are same as that of the concept window.
- *Keyword Browser*: The keywords associated with a concept are displayed through this window in a tree view.
- *Document Window*: The documents related to the entities present in the ontology can be accessed through this window.
- *Search Window*: Through this window, users can specify search for different entities in the knowledge structure.

5.3. *Visualization Interface*

The visual interface gives graphical view through which the knowledge structure can be navigated visually. This is shown in Figure 4. Different components of the visual interface are

- *Selection Panel*: Two types of graphical views are provided in this interface: Topic View and Concept View. Through the Topic View the topic hierarchy is displayed and the concept graph can be displayed through Concept View.
- *Tree View Panel*: Tree view of the ontology is rendered through this panel. The selected topic of the concept in the tree is displayed graphically in the Graph Navigation Panel.
- *Graph Navigation Panel*: This panel produces the visual representation of the concept or topic passed to it either through Selection Panel or Tree View Panel. The ellipses filled with red color represent topics and those filled with blue are concepts. Every relationship has been attached with a color and the direction of the arrow signifies the direction of the relationship.

5.4. *Database and Knowledge Structure Access APIs*

The knowledge structure database is stored in two forms. There is a backend relational database which stores the knowledge structure for local use and it is also stored in XML format so it can be exported to other formats. The database is implemented in MYSQL. We have identified and implemented some basic objects and methods to access the knowledge structure database.

5.5. *Multilingual Interface*

The knowledge structures can be authored in any Indian languages having Unicode support. ITRANS[a] notation is used for entering the language specific representation of an entity of the ontology. To add a new language to the tool, one has to provide the

[a]http://www.aczoom.com/itrans/

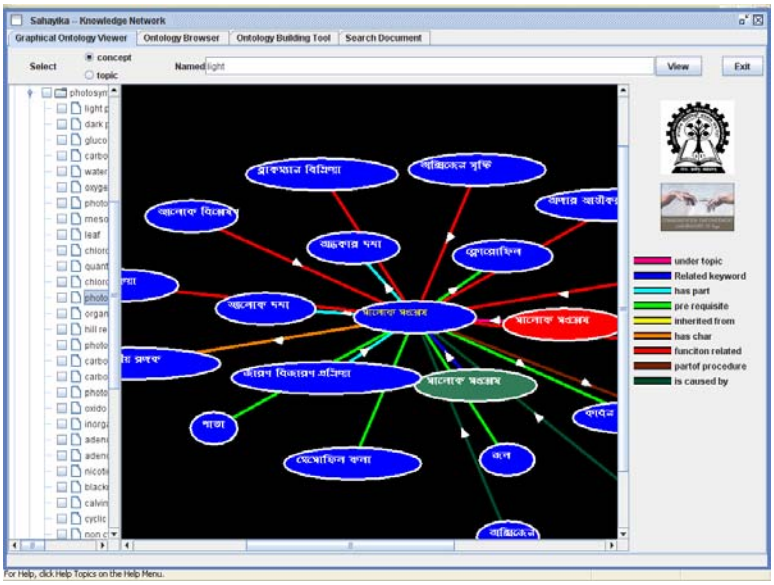


Fig. 4. Knowledge structure visualization interface in Bengali.

ISCII to Unicode mapping file and the font through which an entity will be rendered in the particular language. Currently, we have developed ontologies for Physics, Biology and Geography in two languages English, Bengali. Presently, the knowledge structure database consists of 249 topics, 3369 concepts and 4201 keywords.

6. Management of Learning Material Repository

In this section, we provide means to manage learning material repository with the help of constructed ontology

6.1. *Indexing of Documents in Ontological Levels*

In order to provide search and browsing facility in the learning material repository, the materials need to be indexed. Here, we provide ways to index the learning materials in different levels of ontology.

6.1.1. *Keyword Based Indexing*

Keyword based indexing is performed by computing the term frequencies of all the domain relevant keywords in the document. The normalized term frequency (tf_{ij}) of the i^{th} keyword in the j^{th} document is computed as:

$$tf_{ij} = \frac{f_{ij}}{\sum_i f_{ij}} \tag{1}$$

where f_{ij} is the term frequency of the i^{th} keyword in the j^{th} document. Inverse document frequency of the i^{th} keyword with respect to all the documents present in the repository is given by

$$idf_i = \log \frac{N}{n_i} \tag{2}$$

where N is the total number of documents in the repository and n_i is the number of documents in which the i^{th} keyword occurs. The significance of i^{th} keyword with respect to document j is given by

$$sig_{ij} = tf_{ij} * idf_i \tag{3}$$

6.1.2. *Concept Based Indexing*

According to our domain model, the keywords are associated with the concepts with a specificity index. Once we have calculated the term frequency of each domain relevant keywords for a document, we can calculate the significance of the concepts for the particular document. For each document, we derive a vector

$$K = \{(f_1, k_1), (f_2, k_2), ..., (f_m, k_m)\} \tag{4}$$

where f_m is the term frequency of the keyword k_m. The frequency of concept C_i in document j (fc_{ij}) is the cumulative frequency of keywords (from vector K), that are indicative of C_i.

$$fc_{ij} = \sum_l s * f_l \tag{5}$$

where s is the specificity index of (C_i, k_l) association. The normalized frequency nf_{ij} of C_i is given by

$$nf_{ij} = \frac{fc_{ij}}{\sum_i fc_{ij}} \tag{6}$$

The presence of related concepts of a particular concept in a document is a strong evidence for the concept to be a candidate index concept of the document. So, we consider the contribution of the related concept also by the following formula

$$nf_{ij} = nf_{ij} + \frac{R_{ij}}{R_i} \tag{7}$$

Where R_{ij} is the number of related concepts of C_i present document j and R_i is the number of related concepts of C_i present in the ontology. The significance of concept C_i is given by

$$sigc_{ij} = nf_{ij} * \log \frac{N}{n_i} \tag{8}$$

where N is the total number of documents in the repository and n_i is the number of documents in which C_i occurs.

6.1.3. *Topic Based Indexing*

A topic is said to a candidate index for a document if there is sufficient overlap between the concepts present in the document and the concepts that belongs to the topic in ontology. The significance of a topic for a document j is given by

$$sigt_{ij} = \sum_i sigc_{ij} \qquad \text{for all } C_{ij} \in T_C \cap T_D \tag{9}$$

where T_C is the set of ontology concepts that belongs to topic t_{ij} and T_D is the set of concepts in the present document.

6.2. *Search Interface*

The users can search learning materials on the given input query in the search interface and also can navigate through the topic hierarchy for accessing learning materials on different topics. System accepts keywords as input query from the users for searching learning materials. The search interface make use of the created index structure presented in section 6.1 for retrieving relevant learning materials in response to the user query. The search interface developed by us is depicted in Figure 5.

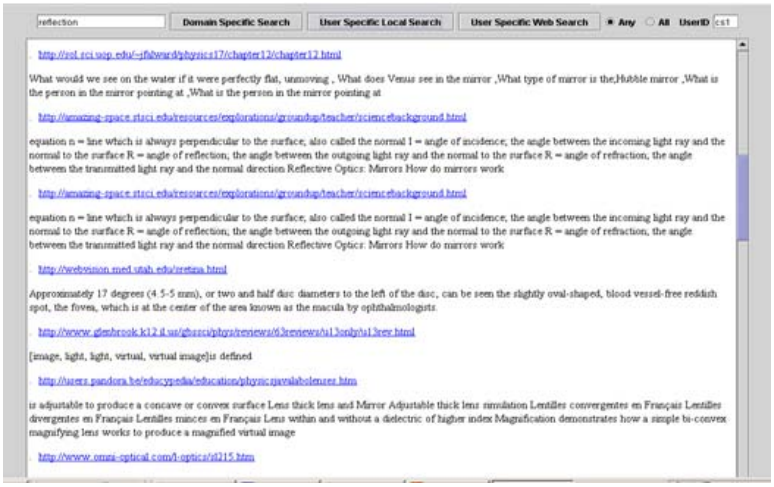


Fig. 5. Interface for searching learning materials

6.2.1. *Domain Specific Search*

Our system uses the ontology for domain specific information filtering [McCallum *et al.* (1999); Cohen (1998); Craven (1998); Tryfonopoulos *et al.* (2009)] to retrieve the documents relevant to the input query. The input query is a term or a list of

terms. To perform the domain specific search, each term in the input query list is mapped to its corresponding concepts. A term can map to a single concept or multiple concepts. When a term has multiple meanings in different domains; it can map to multiple concepts. The ontology maintains a dictionary of terms and the list of associated concepts for each term. If a term in the input query list corresponds to a single concept, that single concept is selected. If a term maps to more than one concept of different domains, then the learner's feedback is taken to select the concept of the particular domain for which he is interested. The concept list is forwarded to the content retrieval module for document retrieval. Documents of learner's interested domain are filtered out from the document set.

For filtering the domain specific documents, the significance of each of the concepts present in the document is computed as discussed in section 6.1.2 concept based indexing. The significance of each concept is computed considering the occurrences of the query concept as well as the occurrences of the related concepts of the query concepts present in the document. The *relevance score* of each of the document with respect to query is computed based on the value of concept significance. The *relevance score* represents the relevance of the document to the given query. Documents are ranked based on the value of *relevance score* and presented to the user.

We have compared the performance of our system with Google. We are presenting here some representative results of our evaluation with queries from Physics domain. The users are asked to rank the first 20 results returned by the Google search engine in response to the users query. For each document the user specify whether the document is relevant or irrelevant to him. The same test is performed over our system with the same set of queries. Figure 6 depicts the comparative precision of the returned results by our system against the results returned by Google with feedbacks from 10 users with 10 queries.

6.2.2. *User Specific Personalized Search*

A document may be relevant to a query but it may not be understandable to the learner. This can happen if for example the document contains many concepts that are unknown to the learner. If a document contains too many concepts that are unknown or outside the scope of the learner's curriculum then the document may not be understandable to the learner. For the same query input the relevance of a document will be different for different learners having different knowledge levels. So we have incorporated the facility of user specific personalized search.

The system keeps a profile of the user's interest to meet the user's need. Firstly, the system keeps track of the user's requirement including the user's interests. This is referred to as the user requirement. The students belonging to the same class have a common set of requirement that is defined by the curriculum and this common set is a part of the total domain knowledge, which reflects the knowledge requirement for a specific user group. We represent this requirement of knowledge for a specific

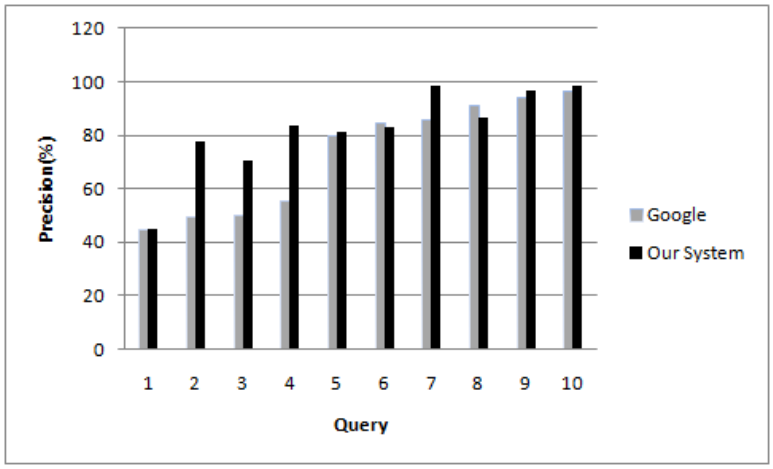


Fig. 6. Comparison of precision (Queries used in this study: 1. regelation, 2. laws of reflection, 3. centigrade scale, 4. inertia of motion, 5. angular acceleration, 6. newton's first law, 7. angle of reflection, 8. projectile motion, 9. electric charge, 10. gravitational unit of force).

group in a Group Profile. This is a representation of the syllabus of a class for a particular subject.

Individual interest of a user can vary from the predefined Group Profiles. Again, different students have different state of knowledge levels. System maintains an individual user profile for an individual student. The student's current state of knowledge is captured in the individual user profile. It keeps track of the concepts already learned by the user or known to the user. This is referred to as the user state. Each user profile is stored in two levels, the topic level and the concept level. A profile includes a set of topics. For each topic, it includes the concepts known to the user. We have provided a user interface (Figure 7) that helps the student to create the profile.

To obtain personalized search we define a score called the *perceptive score* (P) to each the document along with the relevance score. The *perceptive score* reflects the extent of match between the concepts present in the document and the known concepts present in the learner's state. We compute this score only for those documents that are relevant to the given query.

To compute this score, we look at the concepts present in the document and the concepts present in the user state. We find the proportion of the concepts of the document, which are known to the user. We extract the list of unknown concepts. An unknown concept is easier to perceptive, if the learner knows most of the prerequisite concepts of that concept. So we look at the prerequisite concepts of the unknown concepts.

Let $C = \{C_1, C_2, \ldots, C_n\}$ be the concepts present in the document D. The concepts in set C be divided into two categories i. e. known concepts and unknown

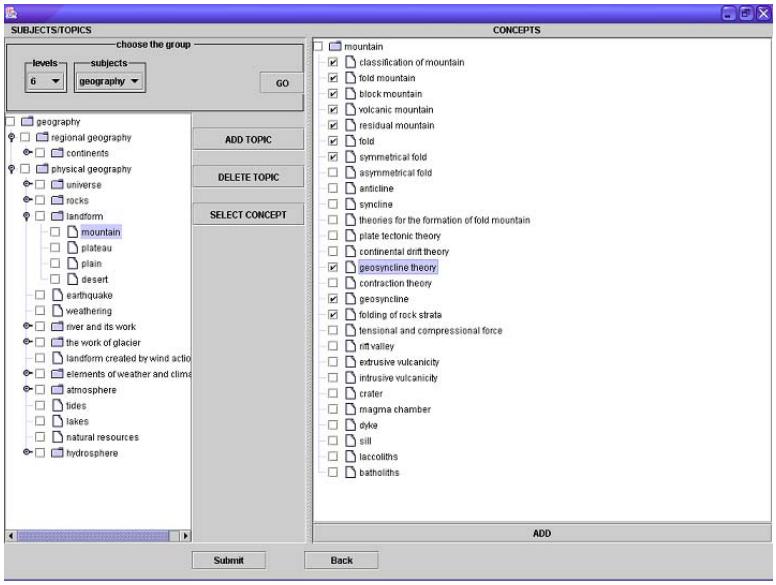


Fig. 7. Interface for creating user profiles

concepts. The known concepts are those that are known to the user. Similarly, the unknown concepts are those that are not known to the user.

Let the set K contains the known concepts and the set U contains the unknown concepts. We obtain all the prerequisite concepts for each of the unknown concepts from the ontology. Next, we check whether the obtained prerequisite concepts are known to the learner consulting the user profile. We partition set U into two subsets U_1 and U_2. The subset U_1 contains the concepts, whose all prerequisites are known to the user. $U_2 = U - U_1$ is the subset of U, which contains the concepts, whose at least one prerequisite is unknown to the user. Let the $size(K)$, $size(U_1)$, $size(U_2)$ give the total number of concepts present in the set K, U_1, U_2 respectively. The *perceptive score* (P) is computed as follows:

$$P = \frac{\alpha * size(K) + \beta * size(U_1) + \gamma * size(U_2)}{size(C)} \tag{10}$$

where $\alpha = 1.0$, $\beta = 0.5$ and $\gamma = 0.0$.

Each document is given a score namely document score. The document score is an estimate of the relevance of the document with respect to a query and the learner state. We compute the document score by combining the *relevance score* (R) (discussed in section 6.2.1) and the *perceptive score* (P). Only those documents are chosen, which have understandability score above threshold value. For the chosen documents, the document score is given by

$$\text{Document score} = \text{Relevance score}(R) + \text{Perceptive score}(P) \tag{11}$$

The documents are ranked using the document score and presented to the user.

To evaluate the performance of the user specific personalized search, many queries from the collected set were processed out by our system. For the same given query, the ranking of the documents varies according to the learner's knowledge level. Here, we show the results obtained by our system for the query Law's of reflection for two students whose knowledge level differs from each other.

The query Law's of reflection was forwarded to the Google search engine and first 100 documents were further processed by our system. The system first filtered out the domain specific documents and the filtered documents were re-ranked using documents score (equation 11). The Table 1 and Table 2 show the first few documents presented by our system to two different students with user ids S1 and S2 for the same query reflection. The known concept space of student S1 and S2 for the topic Optics of subject Physic are given below:

Known concept space of S1 = {*light, plain mirror, concave mirror, image, real image, virtual image, reflection, normal, angle of incident, angle of reflection, total internal reflection, refraction*}

Known Concept Space of S2 = {*light, plain mirror, reflection, normal, angle of incident, angle of reflection*}

Table 1. Top 6 output results shown to a student with user id S1 for query reflection

System Ranking	**Google Ranking**	**Document Score**	**Document URL**
1	19	1.2205	`http://physics.bu.edu/~duffy/PY106/Reflection.html`
2	21	1.1419	`http://www.physicsclassroom.com/Class/refln/reflntoc.html`
3	12	1.1408	`http://id.mind.net/~zona/mstm/physics/light/rayOptics/reflection/reflection1.html`
4	35	1.0524	`http://hyperphysics.phy-astr.gsu.edu/hbase/phyopt/reflectcon.html`
5	98	1.0096	`http://www.geom.uiuc.edu/education/calc-init/rainbow/reflection.html`
6	54	0.9172	`http://www.gcse.com/waves/reflection.htm`

For the same query, the ranking of documents in the output results varies for students with User ID S1 and S2 according to their knowledge levels. The known concept space of student S1 is more as compared to the student S2. If we go through the documents shown to students, we find that the top ranked document shown to student S1 includes discussion of the concept reflection along with other concepts refraction and total internal reflection, whereas the top ranked documents for the student S2 are different and includes discussion mainly about those concepts which

Table 2. Top 6 output results shown to a student with user id S2 for query reflection

System Ranking	Google Ranking	Document Score	Document URL
1	35	1.1008	http://hyperphysics.phy-astr.gsu.edu/hbase/phyopt/reflectcon.html
2	98	1.0908	http://www.geom.uiuc.edu/education/calc-init/rainbow/reflection.html
3	54	1.0096	http://www.gcse.com/waves/reflection.htm
4	20	0.8385	http://science.jrank.org/pages/4871/Optics-Reflection-refraction.html
5	67	0.8340	http://theory.uwinnipeg.ca/physics/light/node4.html
6	21	0.7661	http://www.physicsclassroom.com/Class/refln/reflntoc.html

are there in the known concept space of student S1.

7. Conclusions

In the current education scenario, computer aided learning has received a huge response. To be effective, proper course sequencing and integration of learning materials into the system is required. For this, the knowledge of the concerned domains has to be represented. In countries like India, where vernacular based education has proved to be effective, multilingual access interfaces are of great importance. In this paper, we have provided a framework that enables the knowledge experts to build domain knowledge in their vernaculars. The developed ontology has been utilized to index learning materials into different ontological levels. The created index structure has been used for providing more accurate search in eduction domain. The framework also provides the ways to perform personalized search by consulting the user profiles and the created index structures.

References

Bechhofer, S., Horrocks, I., Goble, C. Stevens, R. (2001). OilEd: a Reasonable Ontology Editor for the Semantic Web. *Working Notes of the 2001 International Description Logics Workshop (DL-2001)*, 1–9.

Ceusters, W., Martens, P. (2001). LinKFactory: an Advanced Formal Ontology Management System. *Proceedings of Interactive Tools for Knowledge Capture Workshop, KCAP 2001*, pp. 21-23, Oct. 2001.

Cohen, W. W. (1998). A Web-Based Information System that Reasons with Structured Collections of Text. Proceedings of Second International Conference on Autonomous Agents (Agents-98), ACM Press, 400–407.

Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., Slattery, S. (1998). Learning to Extract Symbolic Knowledge from the World Wide Web.

AAAI'98/IAAI'98: Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence, 509–516.

Farquhar, A., Fikes, R., Rice, J. (1996). The Ontolingua Server: a Tool for Collaborative Ontology Construction. *International Journal of Human-Computer Studies*, **46**, 707–727.

Fernandez-Lopez, M., Gomez-Perez, A., Pazos-Sierra, A. Pazos-Sierra, J. (1999). Building a Chemical Ontology Using Methonology and the Ontology Design Environment. *IEEE Intelligent Systems and their Applications*, **14**, 37–46.

Gangemi, A., Gunario, N., Oltramari, A. (2001). Conceptual Analysis of Lexical Taxonomies: The Case of Wordnet Top-Level. *Proceedings of the International Conference on Formal Ontology in Information Systems*, 285–296.

Gennari, J. H., Musen, M. A., Fergerson, R. W., Grosso, W. E., Crubzy, M., Eriksson, H., Noy, N. F., Tu, S. W. (2002). The evolution of Protege: an Environment for Knowledge-based Systems Development. *International Journal of Human Computer Studies*, **58**, 89–123.

Goerge, M. A. (1995). Wordnet: A Lexical Database for English, *Communication of ACM*, **38**, 39–41.

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, **5**, Academic Press Ltd. 199–220.

Knight, K., Luk, S. K. (1994). Building a Large Knowledge Base for Machine Translation. *Proceedings of the American Association of Artificial Intelligence Conference*, 773–778.

Lenat, D. B. (1995). CYC: a Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM*, **38**, 33–38.

McCallum, A., Nigam, K., Rennie, J., Seymore, K. (1999). A Machine Learning Approach to Building Domain-Specific Search Engines. *Proceedings 16th International Joint Conference Artificial Intelligence (IJCAI -99)*, 662–667.

McGuinness, D. L., Fikes, R., Hendler J., Stein, L. A. (2002). DAML + OIL: An Ontology Language for the Semantic Web, *IEEE Intelligent Systems*,**17**, 72–80.

Mukherjee, A. (2002). Build Robots Create Science - A Constructivist Education Initiative for Indian Schools, *Proceedings of Development by Design* , Bangalore, India.

Niles, I., Pease, A. (2001). Towards a Standard Upper Ontology. *Proceedings of the International Conference on Formal Ontology in Information Systems*, 2–9.

Sure, Y., Angele, J., Staab, S. (2002). OntoEdit: Guiding Ontology Development by Methodology and Inferencing. *Proceedings of the 1st International Conference on Ontologies*, Databases and Applications of Semantics for Large Scale Information Systems, 1205–1222.

Sowa, J. F. (1995). Top-level Ontological Categories. *International Journal of Human-Computer Studies*,**43**, 669–685.

Tryfonopoulos, C., Koubarakis, M., Drougas, Y. (2009). Information filtering and query indexing for an information retrieval model. *ACM Transactions on Information Systems (TOIS)*,**27**:2, 1–47.

Vega, J. C. A. (2000). WebODE 1.0: User's Manual. Laboratory of Artificial Intelligence, Technical University of Madrid.

Wang, R., Li, K., Martonosi, M., Krishnamurthy A. (2004). Distance Learning Technologies for Basic Education in Disadvantaged Areas, *Proceedings of the 8th Global Chinese Conference on Computers in Education*.