An Information Retrieving Service for Distance Learning

Lauro Nakayama, M. Rosa Vicari, and Helder Coelho

Abstract*—The Lifelong Learning environment has specific features and it is strongly supported by Information and Communication Technologies. Our idea consists of a web service to aid the student along the information retrieval and mining process. The service is supported by three different agents. The user profile, the student model and the intelligent information mining process. This information permits us to generate a refined search term according to the student's needs, which occurs during the solution of a problem. The main advantage of this service is a refined search result that can aid the student in his educational activities. This service composes the kernel of the Web educational portal – PortEdu.

Key words - ***Learning Environment, Web Semantic, Virtual Community***

1. INTRODUCTION

LIFELONG learning is crucial in preparing workers to compete in the global economy. Nevertheless, it is important for other reasons as well. By improving people's ability to function as members of their own communities, education and training increase social cohesion, reduce crime, and improve income distribution [20]. We assume that learning is the use and the creation of new operational knowledge [5] that steers our actions. Learning is a social activity in which interactions with the environment (human and artificial agents) play an important role.

Our proposal is a web service to aid the student along the information retrieval and mining process. The service is supported by three different agents. The first one is connected with the user profile, the second is the Student Model, and the third is the intelligent information mining process. This process is based on the student's cognitive information on the course subject and the general information about users preferences. This information permits us to generate a refined search term according to the student's needs, which occurs during the solution of a problem. The student can receive assistence in his educational activities with the refined search result, which is the main advantage of this service. The agents have the facility to design new valued-added tasks. This service composes the kernel of the Web ducational portal – PortEdu – which covers several courses, chat, forum, registration and statistics facilities, just as several well-known distance learning environments have. At the moment PortEdu is being used by 60 regular medicine students in the cardiology area at UFRGS (the Federal University of Rio Grande do Sul). This experiment is the first test to be carried out. We believe serendipity with a new generation of inference engines will impel students to be more involved and active in education.

2. LEARNING ENVIRONMENT

Learning environments can be most any environment in which a person can learn (a traditional classroom, a distance course with occasional face to face meeting, a course on the Web, a learning community, learning at the job).

It should be possible to adapt the learning

Manuscript received September 10, .2004

Rosa Maria Vicari is with Universidade Federal do Rio Grande do Sul, Brasil. (e-mail: rosa@inf.ufrgs.br)

environment to certain characteristics on software interface, content, context and learning. For instance, an inexperienced learner will find difficulty compared to an experienced learner. If the student is an inexperienced learner, then the search result can be a more generic text. If the learner is well experienced, the search result can be a specific text or a particular image.

The content of the course will have specific characteristics. It might be useful to adapt the environment to the context in which it is used. It is possible to take into account such differences as computer literacy, background and cultural distinction. The learning environment should be available for all kinds of users. Therefore, when designing an environment, the context in which environment will be used (country, culture, availability of computers, subject area, type of learner, and experience in learning) should be considered.

E. THE PROBLEM

The exponential growth of the Web and online resources in general has brought forth a real problem: an overload of information on the web. This rapid growth makes it difficult to navigate on the Internet. The capacity to efficiently access what really is relevant for the user becomes crucial for the effective use of the Web, that is, the refinement of the search. In our system, the refinement is done automatically, based on available information in the users and students models (proactive personalization). The student does a high-level information request and receives a distilled reply.

Internet research requires a special ability due to the speed in which the page information is modified along with the diversity of involved people and observations [10]. Navigation needs good sense and intuition. Good sense in order to not be stymied before so many possibilities, knowing how to select what is most important in quick comparisons. Intuition is a radar that we are developing to click the mouse on links that will take us closer to what we are looking for. Intuition enables us to learn by repeated attempts; encountering rights and wrongs.

The focus in this work is to present an agent enabled to recover web information in an intelligent way [8]. The semantic web is intended to complement humans in areas in which they do not perform well, such as rapidly processing large volumes of information or analyzing large text. The proposed work, even with an intense motivation in the technological area, has the objective to verify if intelligent search mechanisms can efficiently collaborate with the students in their learning activities on the Web. The adapted pedagogical model in PortEdu is that of constructivism.

The agent model using the concepts of Piaget's Constructivist Theory, finds inspiration in the Genetic Algorithms model, in the Neural Network [7] and in the principle of The Society of Mind [11].The previous models had already tested this combination, published in [3].

Knowledge assimilation and accommodation tend to become better and more integrated with the cognitive development. The agent has an interface with the environment (sensorial inputs and motor outputs) and some schemas (its cognitive constructions). Thus, the mechanism proposed is able to build its knowledge by interacting with the environment while it carries out its activity. A schema is composed of {*Context*, *Action*, *and Expectation*}. The *Context* is the representation of situations that the schema is able to assimilate. *Action* represents the action that the agent will carry out in the environment if the schema is activated. *Expectation* represents the expected result after the action application.

F. SOFTWARE AGENT SOLUTION

This work is based on the definition that an agent perceives its surroundings by the use of sensors and acts directly in the environment [15]. In order to bestow intelligence for consultation, two PortEdu agents will provide information to the Information Retrieving Agent: the agent that obtains the user profile making available search terms starting with information on students behavior when he interacts with his classmates and uses the Web;

and the Student Model Agent (educational application agent), which has information on the knowledge of each student concerning the pedagogical content at issue.

The User Profile Agent has two characteristics: reactivity and continuity. It is reactive because it perceives all the changes in the student's behavior as in his deportment once away from the foreseen activities in the learning application. That is, it perceives the actions done by the student in PortEdu. It is continuous due to its constant execution in the portal.

The Information Retrieving Agent is cognitive and proactive as it elaborates search plans starting with received information by the User Profile Agent and the students model. It acts when requested by the student or offers help to the student (a search result, for example) when activated by the students model. Figure 1 presents also the relationship between PortEdu and the web learning applications.

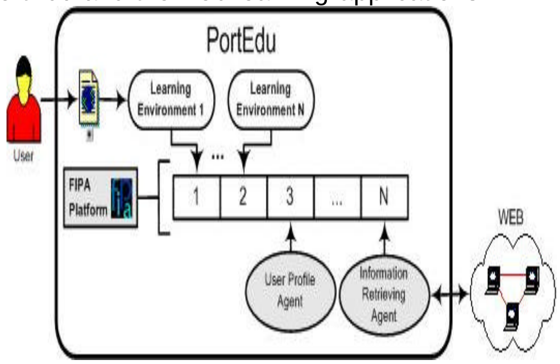

Figure 1 - PortEdu using the FIPA Platform

We propose an architecture using the FIPA (Foundation for Intelligent Physical Agents) platform [4] as a protocol of communication between agents for our environment. The communication among the agents and the learning environments will be based on the FIPA-OS (in the communication structure we use ACL - Agent Communication Language).

3. *USER PROFILE AGENT*

The creation of terms for intelligent search must consider the result to be obtained. In the case of this work, the intention is to aid the student during the use of the learning environment anchored to PortEdu. The aid to the student will be carried out by the obtained contents through the intelligent search mechanism or the indication of a participant in the group that has the knowledge to help him out in the learning of a specific subject.

The information captured by the user profile is:

- Subject: when proposing a search, the information on the context of the student's work is fundamental so that the retrieved document may be useful in the solution context for the student's problem.
- Cognitive context: the Information Retrieving Agent needs to know at what point of the content the student is specifically working on so that it is possible to specify and refine the search. We use a Probabilistic Multi-Agent Environment [19] to test our Information Retrieving Agent. The IT (Intelligent Tutor) is an educational application where the content and the student model are implemented using Bayesian nets. The user profile needs information about the Bayesian net variable where the student is working (solving a particular problem) and about the variable probability information, in order to do a particular search (see Figure 2). These two terms that were obtained from the students model make the difference between the retrieving information process described here and the mentioned techniques in this paper;
- Ontology: The ontology will define which is the best type of content that should be retrieved in each variable. Each educational application could have their proper ontology. An ontology formally defines relations among terms, that in this work has a set of inference rules;
- URL: the User Profile Agent makes available the URL's for the information retrieving agent that were used by other users;
- User history: the User Profile Agent must supply the Information Retrieving Agent the user history related with the chosen web subject. This history is carried out based on the already solved problems by the user in his previous knowledge on the subject, etc.;

- Preferences: the User Profile Agent informs which are the user preferences in relation with the media, the file format, and the text language;
- Navigation sensors: to accompany the user during his navigation and attempt to update the database concerning user interest profile, automatic or interactive;
- General information about the student: name, age, knowledge level (inexperienced to experienced), email, etc.
- Satisfaction: checks along with the user which is his rate of satisfaction concerning the retrieved and available content.

This collection of information is used in the creation of the information search terms in the Web. They represent a differential in the way to search relevant information for the student. With the association of the information of well-known search techniques and the way they are conducted, the retrieved documents constitute the differential of the offered services by our portal. The User Profile Agent will make available the information in a continuous manner to the other agents in the platform and will be receiving information from the learner agent and interface.

The Student Model Agent, as in Figure 2, is who will supply the User Profile Agent, pertinent information on specific knowledge from the educational system in use.

The process of capturing information occurs as follows: the Student Model net is compared with the expert net. The differences constitute the information that directs the search. For example:

- In the absence of a node in the students net, the excluded variable is informed to the User Profile Agent that includes it in the search term;

- When a node that is not part of the solution is included by the student, we find two situations, the first, the search term is constructed with the correct node information present in the expert net (parent nodes up to two levels). In the second, the information of the excluded node is obtained and the information of the correct previous node (parent) is added in the search term.

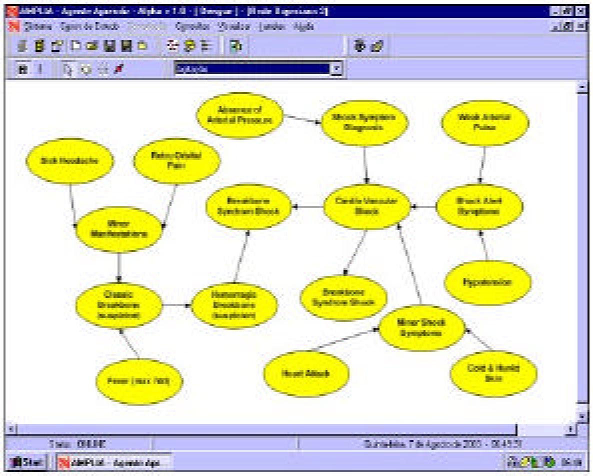

Figure 2 - A Bayesian net that represents the Student Model

Information that we are using in the representation of the Student Model is the level of significance of the node, where each node has a perceptual on how much it represents in the disease diagnosis. This information is used in the final definition of the search term to improve the filtering of retrieved documents (adequate to the context of the error).

Figure 3 presents the relationship between the User Profile Agent, the educational application, and the information retrieving agent.

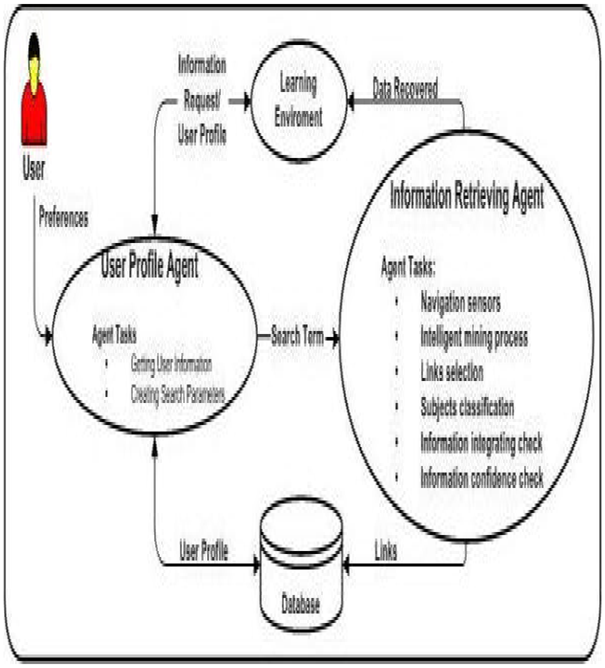

Figure 3 - User Profile and Information Retrieving agents

We may observe that the user profile will be updated at all times (profile will be dynamic). Thus, there is the intention to obtain a closer modeling to that which represents the user at his last instant in the environment and not only a historical profile (some users along the way may change characteristics in their profile).

As a related work, we can mention in [14], which is an adaptive front-end to Google. This work limits itself to model the user's preference during use in the Google portal, not worrying with other aspects of the user. Behavior aspects are understood as the user's developed activities during the navigation in the environment. Another work in this context is *Generic Architecture for User Modeling* [18], who defines the user behavior model in the Internet by making use of an infra-structure (backbone), built on three heuristic levels, user interest, type of documents, and user behavior.

4. INFORMATION RETRIEVING AGENT

Nowadays, there are many applications and prototypes of models based on intelligent agents, such as Search Advisor, Letizia, and InfoFinder. These systems have as an objective to assist in the consumption and organization of the available information on the Web [12]. These applications have the most varied purposes, making more searches from the informed terms by the student, up to accomplishing the personal preferences in learning of each user and, based on this, to bring about information searchers that attend to user .

In our solution the agent differs from the others due to the refined document selection. We compare the user interest profile with the retrieved document, convert the database that contains the examples of positive interest and the retrieved document in vectors where each element represents the weight of the terms in the document and the Student Model information. Calculated by the following method: Term Frequency X Inverse Document Frequency [16,17], we find the angle between one vector and another (see Figure 4). The smallest angle found is our quality judgment. The smaller the angle, the bigger the proximity of the document with the profile and expected subject by the user [2]. Thus, we are using Salton & Chen's models enlarged with the attained information through the user profile and, the Bayesian network variable and probability.

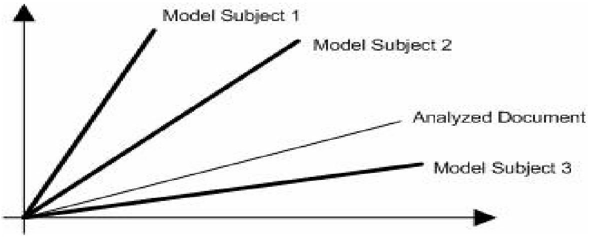

Figure 4 - Comparison between vectors

A vector is created for each retrieved document (model subject N). The analyzed document withholds the obtained information by the User Profile Agent and the information from the Bayesian net concerning the context of the problem that is being solved by the student.

Salton's method is used to calculate the frequency measurement of a word in the document. The weight of W_i of a d_i word [16], on a certain document is calculated by:

$$W i = \left(0,5 + 0,5 \frac{tf\ (i)}{tfmax} \right)\left(\log \frac{n}{df\ (i)} \right)$$

Where:

tf(i) → is the frequency of the term *i* on the document *t*, that is, the number of times that the word d_i appears in the document;

df(i) → is the frequency of the document, that is, the number of documents of the collection which contains the word d_i ;

n → is the total number of documents in the collection;

tf_{max} → is the maximum frequency of a word among all the words in the document.

Once the weight of each analyzed word of the document is obtained, each one is placed in the vector in the corresponding position of the word. The same happens to each set in positive interest models, linked to the subject, as we have both the weight vector for the analyzed document and a vector for each subject on the users interest [2]. The angle between the vectors is by the following formula [1]:

$$Angle = a * \cos \left(\frac{\sum (ai * bi)}{\sqrt{\sum_{vector} (ai)^2} * \sqrt{\sum_{vector} (bi)^2}} \right)$$

Where *a* represents the analyzed vector and *b* represents the vector of a subject of interest. The formula above is applied for each one of the vectors at matter, being that the lowest angle is put into use as it represents a height proximity between the subject and the analyzed document. Each process above will be repeated in the second stage, which is the final classification with information on the Bayesian net with significance level, weight and the probability of hits that will be informed by the User Profile Agent.

The use of attained information by the User Profile Agent, with the educational application in particular, permits to add links in a link repository, with integrity classification, and the judgment for the inclusion of reference in a repository or not. The classification of the document takes in accounts the students and expert (professors) comments. As the Information Retrieving Agent will be offering services in an educational environment, it can automatically retrieve information and offer the student the text content, image, sound and knowledge.

In our application, the navigational sensors will try to obtain the user interest profile and update the database concerning the subjects on user educational interest (User Profile Agent task). The InfoFinder also makes use of stored information in the database to generate consultation, which will be submitted to search (NetClue – Browser/Google), but the information stored by InfoFinder is generic. The attained results on previous searches are stored in a database of candidate addresses and will be selected by the link selector and submitted to the subject classifier to be stored in the link repository (organized by subject). The integrity and the confidence verifier has as its goal to access each one of the stored addresses in the repository to check if they are active and consistent, and in–what rate of reliance they meet at the moment. This operation is necessary in the consulting return in order to better refine the search result and check the confidence of the documents.

The performers represent the encountered replies from the search agent and are offered to the environment with the objective to alter the state of interaction between the student and the educational system that is being used. The Information Retrieving Agent is divided into four great functions:

Navigation: responsible for searching web information that satisfies the user interest profile and adds it to the list with the intention of being selected or not in the future in the repository;

Reach information from the student model: gets information on the cognitive state of the student. The information must be available in the application. This task is offered by the Information Retrieving Agent for the educational application;

Links selection: that chooses, by different discernment, which existing links in the list created by the InfoFinder will be effectively added to the definitive repository;

Subject classification: responsible for the classification by subject, of the selected information and insert them in an organized way in the repository.

Information confidence and integrity verifier: which should, from time to time (or when the user activates it) perform verification on all links stored in the repository with the intention to withdraw inconsistent and eventual out-dated links, or even what does not correspond to user profile. To aid in the links consistency, we will provide a scale with metrics to assist, as for instance, the concepts of academic links. The possibilities for the user, expert or student, to configure this scale as to his single needs (confidence on the extracted information in the repository) will be available.

As a related work we can mention the approach to automatically optimize the quality of retrieving the information in search mechanisms using navigational data [6]. Intuitively, a good information retrieving system should present considerable documents on top of ranking, maintaining the non-relevant documents in the sequence. The Search Adviser, Letizia, and InfoFinder, have in common the apprenticeship of the general profile of the documents. This apprenticeship is done by the use of heuristics and extracts sentences that are representatives of the main topics on each document [9].

Next is an example of message content from the User Profile Agent to the Information Retrieving Agent within PortEdu that will be used for the development of the search term.

```
<search term>
 <user>miletto</user>
 <Subject> Cardiology </Subject>
 <Category> congenital heart disease
</Category>
 <Excluder Node>
 <Description> Rheumatic Fever
</Description>

<Issue>Autoimmune,Systemic</Excluder Node>
 <User Preferences>
 <Language> Portuguese, English
</Language>
 <File Type>
              PDF,DOC,HTM,TXT,JPG,BN
 </File Type>
 <Bandwidth>Broadband </Bandwidth>
 </User Preferences>
 <User Knowledge> ? </User
Knowledge>
<URL>http://educacao.cardiol.br/accsap/
answers%20comentadas%2003.pdf
</URL>
</search term>
```

The example above was made using cardiology terms as a pedagogical context. The message content from the User Profile Agent to the Information Retrieving Agent is based on XML architecture. The message above will be dealt with by the Information Retrieving Agent that will bring up a specific search term for the chosen search tool. In recovering the infomation, the Information Retrieving Agent filters and classifies the results based on the information contained in the message search term and the ontology at the learning environment at issue. Making use of the message search term example shown above and the standard Google search tool, the Information Retrieving Agent develops the following specific search term: Cardiology +“congenital heart disease” +"rheumatic fever". This term is submitted to Google retrieving 10 result pages. The next step is to make the classification and the percolation result based on users preference and the ontology, reducing radically the quantity of resulting links, increasing the level of relevance. The user preference is: Portuguese language and TXT type document. This refined classification is a cognitive context, in which the student is working or needing assistance. This would result in a well-reduced research with two URL’s at the most. Before launching a search, the User Profile Agent certifies if another participant in the group had already requested it. If this occurred, the User Profile Agent notifies the application about the fact that a student has already come up with the same demand and that he could aid the present student in his learning process.

On the other hand, a test was performed on Google with the same subject above, considering the statistics in which a user makes use of two terms at the most in a search mechanism. This way the user would use the term cardiology; cardiopathy, retrieving 210 pages in different contexts from those of interest.

Once the filter and link classification is done, the Information Retrieving Agent communicates to the learning environment that there is available content to complement the information on the topic in which the student is working on and makes available URLs with content, used by other students with similar problems. The precise moment to present this content to the student is determined by the set of learning environment agents, as it depends on the pedagogical model.

Once it is decided by the learning environment agents to deliver the retrieved content to the student, it is necessary to find out if this content was effective in the learning process. In this phase we are working on a good quantity of sections, time-interval, and the importance of the retrieving information (utility weight of the document obtained from the expert and student) for knowledge development. To increase the level of efficiency in future searches, links and contents are stored in a link repository, considering the weight and the comments made by the expert and students.

5. *FINAL CONSIDERATIONS*

The Information Retrieving Agent is now in a test stage using Google integrated on PortEdu. The automatic content retrieving information process, based on user profile information and student model knowledge, is the differential of the traditional search mechanism. The initial tests show that the contextualized search process can really aid the student. For instance, the bigger the significance of the node at matter, in the building of the net, the smaller the number of retrieved documents and these tend to attain significant information for the solution of the problem. The bigger the number of variables (nodes) is obtained in the Bayesian net, the better and more significant is the retrieved documents quality. In our example we are using over two levels (parent nodes) of the conflicting node. Thus, the availability of the intelligent retrieving information process, the main goal of this work, brings a qualified contribution. That is, different from most available search agents, as mentioned before, its functions seek to attend specific Lifelong Learning environment users.

The main difficulties in this project are to anchor technologies and put them together with an operator. The project integrated a variety of previous work developed by the group. This integration was facilitated by the adoption of the FIPA communication platform. The use of software agents engineering techniques permitted the integration of several different software components. The use of the same communication language must be considered.

The initial results that we have obtained with the retrieving information, contextualized and personalized, are encouraged and permit us to believe that this research will be important for learning environments on the Web.

REFERENCES

[1] Bauer, T.;Leake, D. B. **WordSieve:** a method for real-time context extraction. Bloomington: Computer Science Department, Indiana University, 2002.

[2] Chen, L. and Sycara, K. (1998) 'WebMate: A Personal Agent for Browsing and Searching', *Proceedings of the 2nd International Conference on Autonomous Agents (Agents'98)*, New York, pp. 132-139.

[3] Drescher, Gary. Mide-Up Minds: A Construtivist Approach to Artificial Intelligence. MIT Press, 1991.

[4] FIPA – Foundation for Intelligent Physical Agents (2000) 'Agent Management Specification', 05 January, http://www.fipa.org/specs/fipa00023/

[5] Go, F. and van Weert, T. J. (2003) *Regional knowledge networks for Lifelong Learning',* Proceedings of the van Weert, T. J. & Mike Kendall, Amsterdam Ne, pp. Nu.

[6] Joachims, T, (2002), 'Optimizing Search Engine using Clickthrough *Data', Proceedings of the ACM* Conference on Knowledge Discovery and Data Mining (KDD), ACM2002. http://www.cs.cornell.edu/ People/tj/publications/joachims_02 c.pdf.

[7] Kohonen, Teuvo. Self-Organization and Associative Memory. Berlin: Springer-Verlag, 1989.

[8] Krulwich, B. and Burkley, C., (1997) 'The InfoFinder Agent: Learning User Interests through Heuristic Phrase Extraction', *IEEE Intelligent Systems*, Vol. 12, No. 5, pp. 22-27.

[9] MacMillan, I. C. (2003) 'In Search of Serendipity: Bridging the Gap That Separates Technologies and New Markets', 2 July, http://knowledge.wharton.upenn.edu/index.cfm?fa = viewarticle&id=812

[10] Moran, J. M.,(1997) 'How to use the Internet in Education', Information Science Newspaper, Vol 26, No. 2, pp. 146-153 http://www.eca.usp.br/prof/moran/ internet.htm

[11] MINSKY, Marvin. The Society of Mind. New York: Simon & Schuster, 1985.

[12] Pazzani, M., Muramatsu J.,Billsus,D. (1996).'Syskill & Webert:Identifying interesting web sites',*Proceedings of the Thirteenth National Conference on Artificial Intelligence (NCAI96)*, Portland, pp.54-61.

[13] Rosemberg, M. J. (2001) ELearning: strategies for delivering knowledge in the digital age, McGrawHill, New York.

[14] Ruvini, J. D. (2003) 'Adapting to the User's Internet Search Strategy on Small Devices', *Proceedings of the 8th international conference on Intelligent user interfaces (IUI2003)*. Miami, Florida, USA, pp.284-286

[15] Russell, S.,Norvig, P., (1995) *Artificial Intelligence A Modern Approach*, Prentice Hall, Upper Saddle River, NJ, USA.

[16] Salton, G. and Allan J(1995) 'Selective utilization of text traversal', *International Journal Human-Computer Studies*, Vol. 43, pp. 483-497.

[17] Salton G. and Buckley, C.(1988) 'Term weighting approaches in automatic text retrieval', *Information Processing & Management,* Vol. 24, No. 5, pp. 513 – 523.

[18] Sharma, A. (2001) 'A Generic Architecture for User Modelling Systems and Adaptive web services', *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI 2001)*, Seattle, USA, pp. S1.

Nakayama, Lauro. Is a PhD student at Federal University of Rio Grande do Sul – Brazil.

Vicari, M. Rosa. Is a researcher at Instituto de Informática at
Federal University of Rio Grande do Sul – Brazil.

Coelho, Helder. Is a researcher at Computer Science Department at University of Lisbon - Portugal