An Approach for Identification of User's Intentions During the Navigation in Semantic Websites

Rafael Liberato Roberto and Sérgio Roberto P. da Silva

Universidade Estadual de Maringá, Av. Colombo 5790, zona 07,
Maringá – PR – Brasil
{liberato,srsilva}@din.uem.br

Abstract. The growing need for content customization in websites has fostered the development of systems which try to identify the user's navigation patterns. These may be, normally, identified by means of log file analysis. However, this solution does not identify the semantic intention behind user's navigation. This paper provides an approach to incorporating semantic knowledge to the process of identifying the user's intentions in the navigation of a website with semantic support. The capture of the user's intentions is achieved by the semantic enrichment of the log files and the use of and approach that takes into account the linguistic and cognitive aspects in the development of the user model.

Keywords: User Model, Semantic Web, Web Personalization.

1 Introduction

A user's website navigation is strongly related to his interests and necessities. However, most of today's websites do not take this into account. A solution to this problem would use of personalization mechanisms. A personalized website may include new index pages, provide personalized search results, dynamically create recommendations (such as new links), or even define new layouts for a webpage. Several projects have come out aiming to improve the user's interaction with the website such as, for instance, Letizia [14], WebMate [5], PersonalWebWatcher [17] and OBIWAM [11], which construct a user model analyzing the webpages visited by the user and making adaptations on the pages he has visited.

Today's personalization mechanisms, in general, utilize a navigational behavior analyses to create a user model [4] [10] [13], extracting behavioral patterns by means of machine learning techniques. These models can be represented in several different ways. Letizia [14] produce a list of webpages, while QuickStep [16] creates a list of concepts of interests, and Persona [21] creates ontologies. It is interesting to note that these models are constructed by using a great variety of learning techniques such as, vectorial space models [5], probabilistic models [17] or clustering [18]. However, as seen in [17], the exclusive use of the navigation data can be problematic, because there may be difficulties when there is not enough data to extract patterns related to certain categories of the domain, or when new webpages, which have not been visited already by the users yet, are added to the website.

E. Franconi, M. Kifer, and W. May (Eds.): ESWC 2007, LNCS 4519, pp. 371–383, 2007.

The incorporation of information related to the content of the webpages or the website's structure, i.e., data concerning its semantic relations, provides a way of overcoming the problems mentioned above, improving the personalization process [8] [10]. A common approach in this context is the integration of the characteristics of the webpages contents with a classification defined by the users [6] [19]. Generally, in this approach, keywords are extracted from the website's content and are used to index or classify its webpages in several categories of content. Thus, these approaches would permit the recommendation engines to indicate pages to a user based not only in the similarity among the users' navigation patterns, but (or alternatively) on the similarity of the content of the webpages as well. Some projects have adopted the integration of the similarity of the webpages' content to enrich the process of user model construction [6] [19] [9]. SEWeP [9], for example, has as main characteristic the creation of C-logs (concept-logs), a semantically enriched log file based on the extraction of keywords of the webpages. After they have been found, the keywords are mapped to the concept of a taxonomy created to model the domain concepts. Each register of log file is improved with the concepts representing the semantic of the respective URI. However, even these systems may not be able to capture more complex relations among the information as, for example, relations originating from a deeper semantic level, which are based on attributes and properties of the concepts involved. To do so, a richer model is necessary, one which is able to represent the semantically richer relations as, for example, a domain ontology.

In the context of the semantic web [3], webpages must be understood not only by people, but also by the machines, in the form of computational agents. One way of making this idea viable is through the usage of ontologies associated to the websites [12]. These ontologies would permit the software agents to reason about the relations of the websites' content and, by doing so, to make it possible to improve the attention given to the necessities of the user.

This article suggests an approach to associate the benefits provided by the semantic structuring offered by the semantic web proposal, by means of ontologies associated to the websites, along with the analyses of the user's navigational behaviors, creating what is called a semantic log. In this way, it proposes a process of user model creation based on the identification of possible interests of a user when s/he is navigating a website with semantic support. This process is based on the analyses of a semantic log, using the domain ontology available for the website. To do so, an algorithm was developed in order to classify the user's intentions. This algorithm is based on the idea that all the concepts involved on the website are candidates to be the user's intentions. To determine the real relevancy of each concept, a set of parameters was defined to ponders the influence of the linguistic control over the user's power of expression in the form of a segmentation of the domain ontology, and the influence of a concept in the user's interaction applying the idea of cognitive force, derived from the theory of spreading activation [1] [2], and developed under the cognitive psychology to explain how human memory works.

This article is organized as follows. In section 2 we describe the algorithm responsible for identifying the possible user's interests. In section 3 we discusses a comparative analyses done by means of a simulation of the proposed algorithm with algorithms based on the frequency and the classic algorithm of the Naïve Bayes classifier. Finally, section 4 describes the results and our conclusions.

2 The User Modeling

The creation a user model is always a very complex task, for the set of parameters to be considered is always large. The choice of these parameters along with the machine learning techniques which are applied makes the difference in the quality of the model.

To define which parameters should be used in our project we take into account the linguistic and cognitive aspects which affect the users in the expression of their interests. In this way, we are initially going to discuss the effect that the vocabulary available to the user has on his form of expression. Then, we are going to discuss the effect that the semantic relationship among the concepts has on the human memory and how that can affect the user's expression and the identification of his intention when navigating a website.

2.1 The Linguistic Aspects That Affect the User Model

A domain ontology consists of a set of concepts and relationships that describe the knowledge about a domain of interest. It can also be seen as a vocabulary definer for a controlled language that permits the users to express themselves about the domain.

However, when developing a software application, be it for the desktop or for the web, not all the concepts present in the ontology domain needs to be involved. This partition of domain ontology, which we can call the **application model,** has indirect influence on the expression of the user's interests about the domain.

In the case of a website, the problem goes further, for each webpage uses only a part of the knowledge contained in the application ontology, limiting even more the user's power of expression. To reflect this new reduction in the user's vocabulary, we use a **presentation model**, which represents the segmentation of the application model, which is presented in each webpage. Thus, the partition of the knowledge for a website is structured as illustrated on Fig. 1.

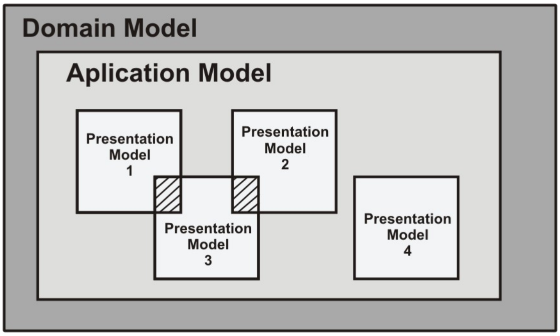


Fig. 1. Segmentation of knowledge in a web application

It is important to emphasize that, in this project, the presentation model is related, mainly, to the information content present in the webpages. It does not take into account the effect that the esthetic presentation has on the user's attention and that can also direct his intentions. The introduction of the presentation model tries to contemplate the fact that a person can only express something for which he has a vocabulary. This fact has an indirect influence on the creation of a model of the user's interests, but that should be considered. Thus, the presentation model is necessary to model the limitations imposed by the language to the user's power of expression.

Trying to consider, partially, the effect that the layout has on the presentation model, we have defined a **status (S)** parameter, in which we isolate three main components in a webpage, based on its level of prominence, which are: the "Main Menu", the "Secondary Menu" and the "Body" of the webpage, as illustrated in Fig. 2. As each component possesses a level of importance in the webpage, drawing the user's attention in different forms, we attribute differentiated weights for the concepts involved to each component.

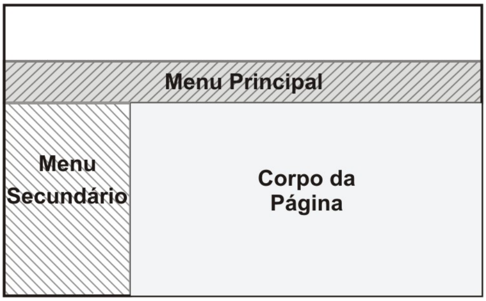


Fig. 2. Parts of the page with different status

2.2 The Cognitive Aspects That Affect the User Model

One of the dominant theories to explain the semantic processing in the cognitive psychology is known as spreading activation [1] [2] [20]. This theory tries to explain how the information recovery of the human brain works [1] [20]. It considers that the human memory is organized in a semantic network form, proposing that when a concept becomes the **focus** of our attention all the concepts associated to it are also activated. That is to say, the activation of a concept spreads itself to all the concepts associated to it. This activation spreading helps to explain how the remembrance of a topic can bring related topic to the mind.

According to Collins and Loftus [7], to better explain the cognitive process of spreading activation it is also necessary to consider **activation strength** for each existent association with the focus concept. In this way, it is possible to amplify or reduce the cognitive force in the spreading process.

In this project, the domain ontology, used for the semantic log construction as well as the models of application and presentation, is represented by a semantic network expressed in the OWL language [15]. If we considerer that by choosing a link in a semantic website page the user will be activating a concept (the **focus concept**) in the semantic network which composes it, it is very reasonable to apply the concept of cognitive strength and spreading activation to evaluate the real interest of the user.

2.3 Linking Navigational Patterns and Semantic Content

Aiming at aiding the user's intentions modeling, in the development of semantic website pages we have adopted a strategy to monitor in a transparent way, the user's interactions with the website, creating a proper log file. Thus, each website page visited inserts a register in the log file storing the date and time of access, and its address. However, in the context of websites personalization, as seen in [17], using data originating from the user's navigation can bring difficulties, especially when there is not enough data to extract patterns related to certain domain categories, or when new website pages are added which have not been visited by the user yet. In this way, due to the fact that all information on the website pages we are interested are based on the domain ontology, the concepts involved in the information of the pages visited, also are inserted in the log file, creating what we call a **semantic log**. Thus, the construction of the semantic log associates the semantic information of website's content with the user's navigation behavior.

It is important to emphasize that, in the scope of this project; we will be limiting our discussion to websites constituted of intranets portals which have been constructed with technology for semantic support since the beginning. We know that the treatment of the websites constructed with modern technology is very relevant, but it will be the target of this project at the moment.

2.4 An Identification Algorithm of a User's Interests

The algorithm proposed here takes into consideration the linguistics and cognitive aspects that influence the process user model creation. Thus, it uses the presentation model and the idea of spreading activation, considering the cognitive strength of each concept, in order to select the concept that expresses the present user's interest.

The algorithm has as entrance the *semantic log* and the presentation model of the page that the user is in. The presentation model has two functions. The first is taking into account the linguistic limitation imposed by the content presented on the webpage. As the concepts present in it are more strongly activated in the user's memory, focusing its scope of expression, they are pondered more strongly than the other concepts of application model, when determining the probability of a concept expresses the user's intention. The second is to take into account the cognitive aspect of the spreading activation in the human memory. The cognitive strength increases or decreases the strength of the activation of the concept in the spreading process. To determine this cognitive strength for each concept is necessary to determine the following parameters: $\boldsymbol{R}$, which define the number of relations that the concept in question possesses, and $\boldsymbol{S}$, that defines the status of the concept in the presentation model, as mentioned before in section 2.1. The relations that a concept possesses were classified in two groups: **#**$\boldsymbol{R_{aplic}}$ provides the number of relations with other concepts in the application model (concepts that do not appear in the presentation model), **#**$\boldsymbol{R_{apress}}$ provides the number of relations with other concepts of the presentation model. Thus, the parameter $\boldsymbol{R}$ is defined by the formula $\boldsymbol{R}=\mathbf{P_{aplic}}$ **#(**$\mathbf{R_{aplic}}$**).+** $\mathbf{P_{apres}}$ **#(**$\mathbf{R_{apres}}$), were $\mathbf{p_x}$ is a specific weight for each type of relation, respecting the first function of the presentation model. The parameter $\boldsymbol{S}$ is defined by the position in which the concept occupies in the webpage layout, as seen in Fig. 2. The cognitive strength of a concept is defined by the formula ***F* = *S* * *R***.

Fig. 3 illustrates the parameters that must be determined for any concept of presentation model during the process of spreading activation. Besides the cognitive strength, differentiated weights were defined for three different types of concepts: the **focus concept** – associated to the *link* chosen by the user and that, therefore, has a high possibility of being his real interest, has a higher value; the **directly connected concepts** to the focus concept – which have a good probability of being the user's real interest, have a value dependent on its cognitive strength, as previously defined; and the **disconnected concepts** of the focus concept, which are not involved in the spreading process and are considered **noise** (once they do not share the user's attention) have a residual value, since they must not be ignored, as it will be shown in the Section 3.

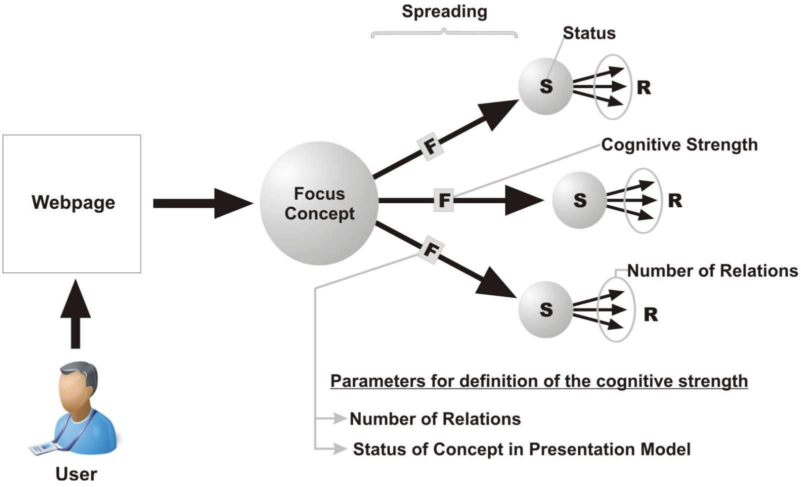


Fig. 3. Parameters in the spreading process

Thus, at each user's interaction, represented by a register of the semantic log, the algorithm identifies the focus concept, the connected concept and the noises in the presentation and application mode and attributes the weights correspondent to each of them and defines the probabilities of interest of each concept. This probability is defined by the formula:

$$P(C_i \mid I_1 \ldots I_j) = \sum_{k=1}^{j} (Tc_i)_k + (F_i)_k \tag{1}$$

where:

- P = probability of user's interest to a determined concept;
- C = the concept in question among all the domain ontology concepts;
- I = the present interaction in the semantic log;

- *Tc* = the value correspondent to the type of concept (Focus Concept, Directly Connected Focus and Disconnected Concept);
- *F* = the cognitive strength of concept (*F* = 0, if the type of concept is a Disconnected Concept).

To obtain the concept that represents the greater interest the following formula is used:

$$\arg\max_{i}(P(C_i \mid I_1 ... I_j)) \tag{2}$$

where ***I*** varies for all the concepts and ***j*** is the total of interactions. In this way, the algorithm keeps an updated list of the concepts with their respective probabilities of interest.

3 An Evaluation of the Proposed Algorithm

In this section we present an experimental study based on simulated empiric tests, in which navigations are defined with pre-defined interests. The objective of these tests is to verify the validity of the approach of semantic knowledge integration to navigation data in the identification process of the user's interests, considering the linguistics and cognitive aspects of the process of user model creation.

To compare our results, tests were also performed using classification algorithms based on frequency and on a Bayesian approach. It is important to emphasize that the parameters used on the proposed approach were defined in an empiric form. However, as future work, we intended to develop a learning algorithm to identify their ideal values.

Thus, in the Bayesian approach, we have used the Naïve Bayes Classifier algorithm, using just positive learning examples, since it is not possible to obtain negative examples in this case without an explicit intervention of the user. The positive examples were extracted by means of the simulated semantic log, in which each entrance was considered as a positive example. In the classification approach based on frequency, just the frequency of visits to the webpage is considered. However, as the frequency is obtained by means of semantic log, it will be considered in fact the frequency of concepts involved on the webpage.

For the execution of the tests a semantic website was developed for the Computer Department of the State University of Maringá, in which the navigations were applied. For each navigation, one hundred interactions with their respective interest concepts were defined. As the result of navigating in the semantic website, a semantic log was generated and departing from it, and from the presentation model, the three approach previously mentioned were applied.

In the tests, each navigation was defined to characterize a determined interest. As an example, the objective of one of the pre-defined navigation was to search for "publications" and "events" in which some "professors" are involved. To find the desired professor it is necessary to access the research projects in which s/he takes part. When the professor is found, her/his publications as well the events in which s/he took part can be visualized. In this example, the greatest concentration of accesses will be found in the process of searching the professor, that is to say, most of

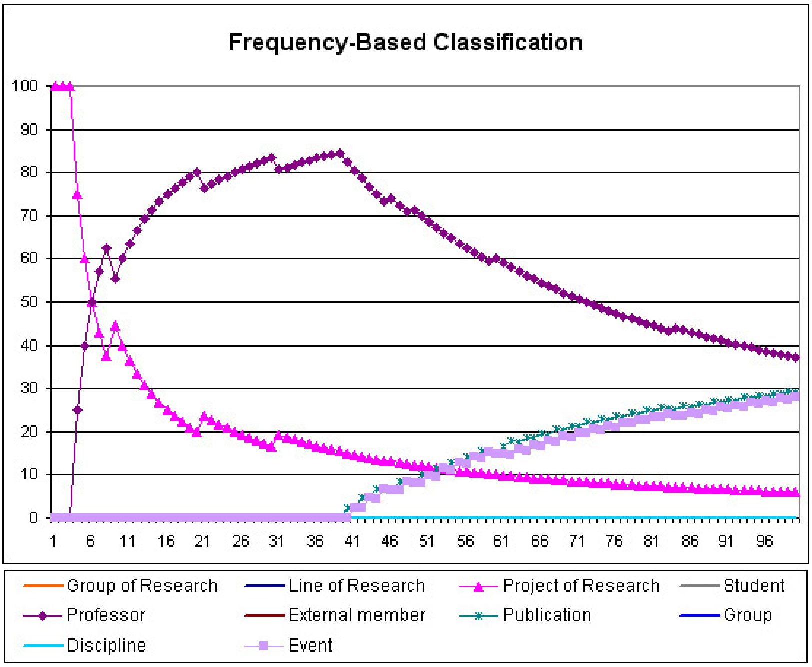


Fig. 4. Interests Percentages progress in the navigation for the algorithm based on frequency

the accessed pages belong to the professors. In this way, the concepts pre-defined as of greater navigation interest were “Professor”, “Publication” and “Event”.

The Bayesian classification approach and the frequency-based approach, presented similar results. In the frequency-based classification, the concepts identified as of greater interest are the concepts involved on the most accessed webpages. Thus, initially we had a very high value (100%) for the concept “Research Project” and a null value for the others, with a better convergence for the real value with the increase of interactions number, as illustrated in **Fig. 4**. In the Bayesian classification, this total preference for just one concept is corrected, as illustrated in **Fig. 5**, for the Naïve Bayes Classifier considers a similar probability for all concepts in the beginning, but with the increase in the interactions number its classification converge to the same result.

Thus, both approach identified the same pre-defined concept for the navigation, that is: “Professor”, “Publication” and “Event”, as being the ones of the user’s greatest interests. The “Professor” concept becomes the user’s greatest interest. This happens due to the high access to the professors’ pages. On the other hand, the frequency-based classification do not take into account all the concepts involved in the webpage, which were not accessed, attributing a null value to the percentage of interest. The same happens in the Bayesian classification, however it attributes a very small value to the percentage of interest, as illustrated in Fig. 6.

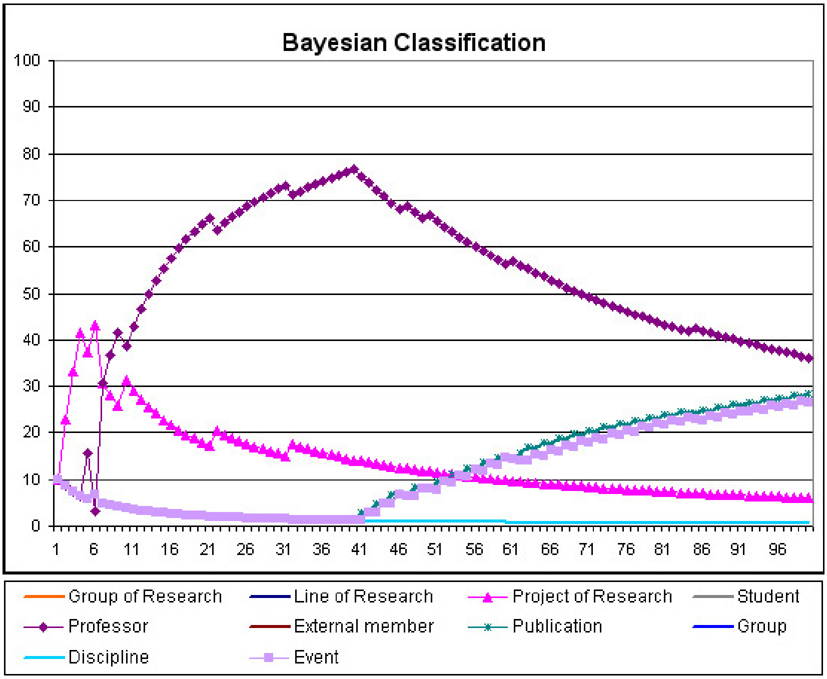
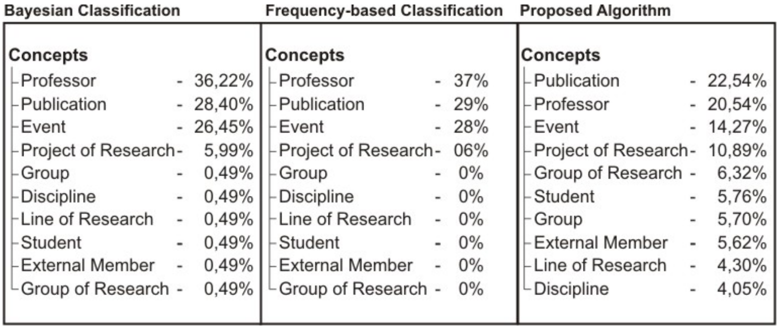


Fig. 5. Interest Percentage Progress in the navigation for the Bayesian algorithm



Bayesian Classification		Frequency-based Classification		Proposed Algorithm	
Concepts		Concepts		Concepts	
Professor	- 36,22%	Professor	- 37%	Publication	- 22,54%
Publication	- 28,40%	Publication	- 29%	Professor	- 20,54%
Event	- 26,45%	Event	- 28%	Event	- 14,27%
Project of Research	- 5,99%	Project of Research	- 06%	Project of Research	- 10,89%
Group	- 0,49%	Group	- 0%	Group of Research	- 6,32%
Discipline	- 0,49%	Discipline	- 0%	Student	- 5,76%
Line of Research	- 0,49%	Line of Research	- 0%	Group	- 5,70%
Student	- 0,49%	Student	- 0%	External Member	- 5,62%
External Member	- 0,49%	External Member	- 0%	Line of Research	- 4,30%
Group of Research	- 0,49%	Group of Research	- 0%	Discipline	- 4,05%

Fig. 6. Preferences resulted from the Bayesian, the Frequency-based and the Proposed algorithm

One characteristic, which we consider negative in these two approaches, is that they do not take into account the concepts involved on the webpages that were not accessed yet. These concepts have a great semantic relation to the present user's interest concepts and, due to that semantic proximity, can be good candidates to become the user's interest focus in a posterior navigation.

Our algorithm also identified as main user's interest the concepts that were predefined for the navigation, as illustrated in **Fig. 7**, however we have found some interesting differences.

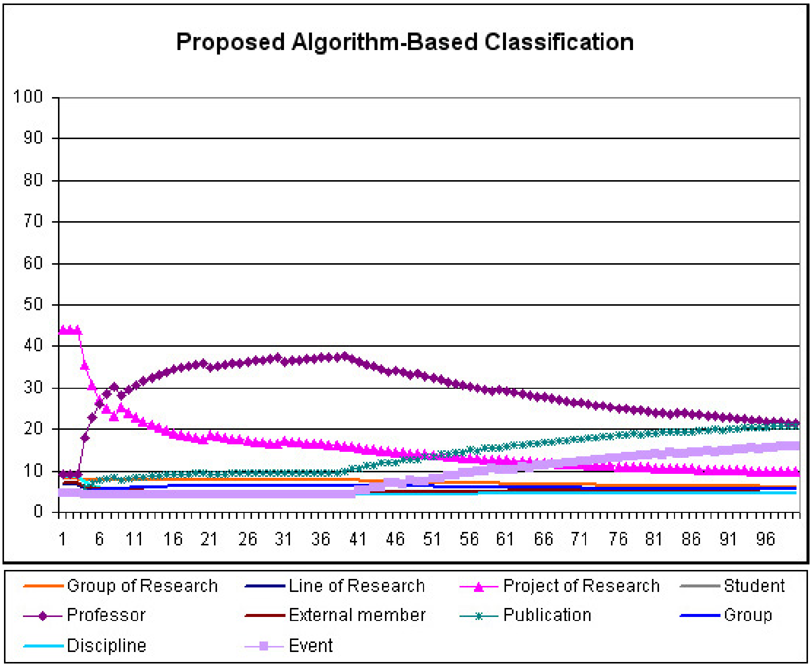


Fig. 7. Progress in the percentage of interest in the navigation when applied to the proposed algorithm

First, the "Publication" concept was identified as being of the user's greater interest, even having the less access number in the navigation. This fact is very relevant, considering that during the navigation, the process of searching the professor is secondary, for s/he is only the way to find the "Publications" and "Events" of the desired professor. In this way, taking in to account the linguistics and cognitive aspects of the process of user model creation, the proposed algorithm was able to identify this semantic relation and ponder on the "Publication" concept (according to its cognitive strength) each time that the "Professor" concept was accessed. As can be seen in **Fig. 6**, the difference between the percentage of interest of the "Publication" and "Professor" concepts became small which confirms an equivalent interest between the concepts. The same do not happen in the other approaches, in which there is a considerable difference between the "Publication" and "Professor" concepts.

Secondly, we can observe that the percentages graphic suffers a flattening. This phenomenon can be explained by the fact that in our algorithm, the concepts involved on the non-accessed pages were also considered, this occurs due to their semantic relations with the concepts of greater interest and their cognitive strength. This consideration of the concepts which do not appear directly on the pages raises the percentage of interest for these concepts and permits a faster response to the changes of the user's future interests.

Finally, we can consider that the results of this experiment were satisfactory, conforming our expectations that the use of the cognitive and linguistics aspects make a great difference in determining the user’s intention

4 Conclusion and Future Projects

In this article, we present an approach to integrate semantic knowledge and navigation data in the process of identification of possible user’s interests. For this, we have created a semantic log associating navigational patterns to concepts defined in a domain ontology. We have developed an algorithm using the linguistics and cognitive aspects that affect the process of user models creation. We identified that the form with which the concepts are disposed in the webpages, limit the user’s capacity of choice, influencing indirectly the expression of her/his interests. In this way we consider the effect that the segmentation of the application model, i.e., the presentation model, has over the concepts presented in the webpages. The presentation models along with the semantic log act as the basis for the proposed algorithm. In order to take into account the way the concepts are organized in the human memory, and how the remembrance of a certain concept can bring to the memory the related concepts, we apply the idea of spreading activation and cognitive strength, to consider the degree of relevance of the concepts in the knowledge model.

In our analysis of the proposed approach, our algorithm presented very positive aspects in relation to other well known approaches (the Frequency-based classification and the Bayesian classification and the). One of the relevant points was to make it possible to identify a concept that has not had the largest number of access as being it the concept of greatest interests by the user. This behavior shows that the spreading activation considered with the different status of the accessed concepts, defined by the presentation model, permits the compensating the deviation caused by greater number of access resulted from a secondary path used to achieve the main objective. This does not happen in the other approaches, in which these concepts have a considerable difference of percentage of interest, in which the concept with greater number of access prevails. Another important fact is the determination of greater values for the probability of the non-visible concepts, which makes possible to obtain a better answer in the change of the user’s interests possible.

This article also shows interesting research problems to be discussed and future projects. Most urgent among them is the development of learning techniques to identify ideal values of the parameters for the cognitive strength consideration and of the individual values of the different types of concepts. We also envisage a great deal of work to be done to consider the influence of the aesthetic aspects of the webpages.

Acknowledgements

Special thanks to CNPq – (Scientific and Technological Development National Counsel for financial support to the Post Graduation in Computer Science Program of the State University of Maringá - UEM).

References

1. Anderson, J. R. (1983a). A spreading activation theory of memory, *Journal of Verbal Learning and Verbal Behavior*, 22.
2. Anderson, J. R. (1983b). *The architecture of cognition. Cambridge*, MA: Harvard University Press.
3. Berners – Lee, t., Hendler, J., Lassila O.(2001). The Semantic Web. *Scientifc American*, May 2001.
4. Brusilovsky, P., (2001). *Adaptive Hypermedia, User Modeling and User-Adapted Interaction*, Kluwer Academic Publishers, Netherlands, pp. 87-110.
5. Chen L., Sycara K. (1998). Web Mate: A Personal Agent for Browsing and Searching. In *Proceedings of the 2nd International Conference on Autonomous Agents and Multi Agent Systems, AGENTS '98*, ACM, pp. 132 – 139.
6. Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D. and Sartin, M.(1999). Combining Content-based and Collaborative Filters in an Online Newspaper. In *Proceedings of the ACM SIGIR '99 Workshop on Recommender Systems: Algorithms and Evaluation*. University of California, Berkeley, Aug.
7. Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic priming. *Psychological Review*, 82, 407-428.
8. Dai, H. and Mobasher, B. (2003). A Road map to More Effective Web Personalization: Integrating Domain Knowledge with Web Usage Mining. *Proc.of the International Conference on Internet Computing 2003 (IC'03)*, Las Vegas, Nevada, June 2003.
9. Eirinaki, M.; Vazirgiannis, M., Varlamis, I. (2003) SEWeP: using site semantics and a taxonomy to enhance the Web personalization process. *KDD 2003*: 99-108.
10. Eirinaki, M. and Vazirgiannis, M. (2003). Web Mining for Web Personalization, *ACM Transactions on Internet Technology (TOIT)*, February 2003/ Vol.3, No.1, 1-27.
11. Gauch S., Chaffee J., Pretschner A. Ontology Based Personalized Search. *Web Intelligence and Agent Systems* (in press).
12. Hendler, J., Berners-Lee, T. and Miller, E. (2002). Integrating Applications on the Semantic Web, *Journal of the Institute of Electrical Engineers of Japan*, Vol 122(10), October, 2002, p. 676-680. http://www.w3.org/2002/07/swint.
13. Lei, Y., Motta, E. and Domingue, J. (2004). Modelling Data- Intensive Web Sites with OntoWeaver, in *International Workshop on Web Information Systems Modelling (WISM 2004)*, Riga, Latvia.
14. Lieberman, H. (1995). Letizia: An Agent That Assists Web Browsing. *In Proceedings of the International Joint Conference on Artificial Intelligence*, Montreal, CA.
15. MCGUINNESS, D. L.; VAN HARMELEN, F.(2006). OWL Web Ontology Language Overview. W3C Recommendation 10 February 2004. Available in: http://www.w3.org/TR/2004/REC-owl-features-20040210/, Acess in: jan. 2006.
16. Middleton, S., De Roure, D., Shadbolt, N. (2001). Capturing knowledge of user preferences: ontologies in recommender systems. *In Proceedings of the 1st International Conference on Knowledge Capture (K-Cap2001)*, Victoria, BC Canada.
17. Mladenic, D. (1999). Text-learning and related intelligent agents. Revised version In *IEEE Expert*, special issue on Applications of Intelligent Information Retrieval.
18. Mobasher, B., Daí, H., Luo, T., Sung, Y. and Zhu, J. (2000). Integrating Web Usage and Content Mining for More Effective Personalization, in *Proc. of the International Conference on E-Commerce and Web Technologies (ECWeb2000)*, Greenwich, UK, September 2000.

19. Pazzani, M. A (1999). Framework for Collaborative, Content-Based and Demographic Filtering. *Artificial Intelligence Review*, Dec. 1999, pp. 393-408.
20. Quillian, M. R. (1968). Semantic memory. In M. L. Minsky (Ed.), *Semantic Information Processing*. Cambridge, MA: MIT Press.
21. Tanasa, D. and Trousse, B. (2004). Advanced data preprocessing for intersites web usage mining. *IEEE Intelligent Systems*, 19(2):59-65, March-April 2004.