Modeling of User Interest Based on Its Interaction with a Collaborative Knowledge Management System

Jaime Moreno-Llorena, Xavier Alamán Roldán, and Ruth Cobos Perez

Dpto. de Ingeniería Informática, EPS
Universidad Autónoma de Madrid
28049 Madrid, Spain
{Jaime.Moreno,Xavier.Alaman,Ruth.Cobos}@uam.es

Abstract. SKC is a prototype system for knowledge management in the Web by means of semantic information without supervision and tries to select the knowledge contained in the system by paying attention to its use. This paper explains user activity analysis in order to find out their interest for knowledge elements in the system, and the application of this interest for users classification and knowledge identification for their interest, inside and outside SKC. As a result a model for user interest based on interaction is obtained.

Keywords: user interest model, user interaction, user profiling, data mining, knowledge management, CSCW.

1 Introduction

Information overload is one of the problems of the ICT use extension. The Web is the most general and significant example of this phenomenon. We think network knowledge management systems have the most important characteristics of the systems with this problem, but these systems are more scalable and have more controllable parameters, so they may be used as an experimental model for research. For example, the Web could be seen as a global knowledge management system.

Our hypothesis is that there are several hidden aspects in the systems affected by the information overload which can contribute positively to the solution of this problem. On one hand, taking advantage of the excess energy of the active elements that are involved in the systems, such as users, services, applications and other entities related to them. On the other hand, using the properties of both the elements and the activities related to the systems affected by the problem, eg. the network, the active entities mentioned above, both the information and the knowledge involved, or the processes and the interactions of that elements and activities.

To investigate this hypothesis we have used, as an experimental platform, a knowledge management system called KnowCat [1][4]. This is a groupware system that facilitates the management of a knowledge repository by means of user community interaction through the Web. This can be done without supervision by using information about user activities and their opinion about the documents that are part of the knowledge base -eg. through votes or notes-. The knowledge repository is constituted

J.A. Jacko (Ed.): Human-Computer Interaction, Part III, HCII 2009, LNCS 5612, pp. 330–339, 2009.

fundamentally by two components: the Documents, that are the basic knowledge units; and the Topics, that are structured hierarchically as a Knowledge Tree. Each document describes the topic where the document is located in the tree. Each topic appears once in the tree, and it could include several documents that describe it and it could include some subtopics too. Each system instance is a KnowCat node that deals with a specific subject and has a user community and its own knowledge repository organized in a knowledge tree.

The task carried out by KnowCat could be improved by reducing the necessity of explicit displays of users' opinion about the knowledge, as well as exploiting the implicit displays of opinion demonstrated by users' activity and the features of all the elements involved in the process -eg. user's interaction with the system-. Such improvements could be generalized to other knowledge management systems, as the Web. This paper deals with how to represent the implicit interest shown by the users' activity towards the knowledge elements of the KnowCat's repository so that it can be used in the system. In order to corroborate the design hypothesis of the proposed approach, a prototype has been developed on KnowCat, which has been called Semantic KnowCat (SKC) [10]. SKC includes, among other things, Client Monitor [11] (CM) that is in charge of obtaining information about the user's activity on the system's client side, information to be used for analyzing the user's interest to the knowledge, and SKC incorporates a Analysis Module (AM) [12] that is in charge of analyzing the system knowledge repository with data mining techniques to describe its elements by means of Words Weight Vectors (WWV) [2].

This paper starts from a mechanism for monitoring users' activity [11] built into the CM of SKC prototype mentioned above. That mechanism makes possible to establish each user's interest per the elements of the knowledge repository. The monitoring mechanism registers the users' interaction intensity towards the knowledge elements by using the user interface events of the system. This paper shows how the data obtained with that mechanism allows representing each user's interest inside the knowledge management system, using a User's Interest Vector (UIV), which represents the distribution of user's interest among the knowledge elements of the system. This paper also shows how the monitored data are good to describe each user's interest inside and outside of the system by means of a User's Interest Words Weight Vector (UIWWV). This new vector is obtained starting from the corresponding user's UIV and the Words Weight Vectors (WWV) that describe the knowledge elements of the system. These UIWWVs allow comparing the users' interests with any knowledge element represented by a generic WWV [2]. The proposed approach has been validated with an experiment series carried out with the system KnowCat in educational activities in the Universidad Autónoma de Madrid (Spain).

Several approaches have been proposed for users' interest modeling for personalization in data recovery field. These go from manual personalization [9] [6], by means of direct indication of preference, to complete automatic modelling [7][5], through monitorization and analysis of user behaviour; going through a demonstration modeling [8], by means of a presentation in the resource system which users consider interesting. Many of the above-mentioned approaches use vector models for representing and comparing text documents, some of which also use WWVs to represent users' interest [13]. Others use conceptual clustering algorithms and use user activity with the processed documents to establish user interest for these documents [13] [5]. Others are in charge of modeling the progress of user interest throughout the time [8].

The proposed approach combines some of the strategies and techniques mentioned above to achieve an interest model based on user interaction with the system on client's side but which can nevertheless be used on the server side. The intention of the model presented is to facilitate comparison of user interest and selection of knowledge elements which are interesting for those inside and outside the system. The final objective of this proposal is to improve knowledge collaborative management through the system, taking advantage of user's normal activity but without disrupting it.

2 User Interest Vector

In a previous paper [11] we have shown how an impression on user's interest towards items that constitute the knowledge handled in one instance -node- of SKC prototype can be obtained by measuring the intensity of users' interaction with the latter.

The corresponding process is based on the activity register analysis (LOG) of the Web Server that supports the system. In this file a line is annotated per each resource requested to the server -HTML pages, images, etc.- in chronological order. The registered data may be configured in the Server, but it usually includes data such as URL resource, the moment of the request or the requester IP address, among others.

Normally, Web servers –such as the one that supports the KnowCat system- only write in the LOG file when they deal with resource request, for this reason they have no information about what happens to them after they have been served. The Client Module (CM) of the SKC prototype corrects this inconvenience by collecting data on user's activity on the Client and sending it to the Server at a frequency appropriate to the activity received. Since data is sent invoking a program from the Web Server itself, the corresponding calls are registered in the LOG with this data. A CM activity data register LOG line is a register line like the others, but it refers to a characteristic resource *"infoSituation.pl"* and includes encoded data, among them, for instance, the user identifier that generated it or values observed for the activity indicators taken into account in each period, such as mouse movement, scrolling or computer keystroke.

By analysing LOG file lines and taking into account the CM lines included in this file and the design of the system user interface -by using both techniques [3]: activity register mining Web and Web site structure mining -, we can possibly get an idea of what happened with the resources served throughout the time and of user interest for knowledge base items to which the resources refer to. The basic process analysis is described in detail in the above mentioned paper [11]. In general terms, it consists of establishing regular monitorization cycles by days and sessions of the user, and to appreciate observed interaction for the knowledge items involved in each period.

Taking as a starting point this basic analysis, we may establish how intensive user community interaction is, as a whole, on each knowledge item, what may give us an idea of the group interest for items and, somehow, of the value that the group gives them [11]. In addition, when users are requested to identify themselves to use the system, the CM register lines include the references of those that have caused them, as we have seen in the previous example. With this information it is possible to follow-up each user activity throughout the time, even through several sessions. This has been done to establish user interest indication (UII) and user interest vectors (UIV).

In particular, taking the results obtained from the basic analysis outline as our starting-point, the values assigned to each of the registered knowledge items in monitorization cycles are accumulated by users. As a result, a list of all knowledge items accessed by the user labeled with an interaction intensity indicative value (III) maintained with them is obtained for each user. A couple of interesting elements may be withdrawn from this list. On the one hand, by adding the IIIs per user, the latter may be assigned the representative interaction intensity indicative value maintained with the system, which may be used as indication of the interest shown by the individual in his/her activity with the node, IIU, and used to compare it with the level of interest shown by other members of the community. On the other hand, by dividing every III of each user into the sum of these, a vector representing contribution of each item relative to the interest shown by the user may be obtained. These are the user interest vectors, UIVs, with which it is possible to establish comparisons among interests shown by members of the community who use the system.

3 User Interest Words Weight Vector

In a previous paper [12] the process used in SKC prototype to create knowledge item descriptors handled by the system was explained, through words weight vectors (WWV) [2] based on the texts related to these items. The way to use these descriptors for establishing a relationship among the corresponding knowledge items was also shown. The following summarizes the above in general terms. WWVs are term lists which have been given a weight. Terms are obtained from significant words from the starting text, putting into groups the different ways in which they may appear -number of nouns, verb tenses, etc.- The weight is established taking into account the frequency of the terms in the original text and the frequency of these in the general use of the language, so that the most frequent words in the text and less common in the language have greater weights than the most common ones in the usual use of the language, especially if they are not very numerous in the text. In order to compare the items with each other, a similarity among WWVs is established, in this specific case the cosine of the angle that forms these vectors. The level of similarity among WWVs is between zero and one, closer to the unit, the greater similarity is among the vectors and closer to zero, the less similar they are. These are typical data recovery techniques [2] and are very popular currently for its use in well-known search engines as Google.

Until now, in SKC prototype, the WWVs related to the knowledge items were based on texts that were considered to be totally linked to the respective items, eg. text documents were their own descriptions, topics were assigned a descriptive text composed by a combination of descriptions of all the documents and subtopics that were included in these topics, users' descriptions as authors were obtained from document linking created by them, etc. This strategy is not likely to be appropriate for user interest treatment, since each individual's interest is distributed unequally among several knowledge items. The user interest vector, UIV, of each individual shows how its interest is distributed among various knowledge items, indicating for each of them a level of interest, which is given as a fraction of unity of its participation of the total. Taking as a starting point each UIV and the knowledge items WWVs to which the first one refers to, it is possible to create one descriptor per user representative of the interest content shown by the corresponding user in the system.

The new descriptor is the user's interest words weight vector, UIWWV. It is actually a WWV in which words weight is established based on the weight of these in the knowledge item words weight descriptors and on the proportion of these items in users' interest. UIWWVs, as it happens for UIVs and WWVs, depend on the period taken into account for their calculation and they progress with time depending on how users' interest is shown by the system knowledge items or the description of the items involved vary -eg. when documents are updated or new documents or subtopics are added to the topics-. UIWWVs are the same as other WWVs and may be compared with each other, as well as being useful to establish a relationship among users and any element of this type that a descriptor may have associated. Specifically, within the SKC prototype field, it makes sense to determine interest similarities among users on one or more instances -nodes- in the system, and also to establish possible users' interest for node themes or any knowledge tree topic theme -list of topics-. In addition, as it is easy to establish WWVs for elements external to the system and accessible through the Internet, as long as they are textual or have some description of this type associated, UIWWVs may be used to establish links with them. In this sense, it could be interesting to discover available resources in other knowledge repositories – eg. different SKC nodes or Wikis - within users' interest area inside the knowledge node field to which these belong to, or to identify those which are outside this area.

4 Experiments Carried Out

In order to prove the viability of the proposals mentioned above, support has been incorporated into the Analysis Module (AM) of the SKC prototype for users' interests. Experiments with KnowCat nodes aimed to test the new prototype SKC functionalities in teaching activities in the Universidad Autónoma of Madrid (Spain) have been carried out. To that end, two KnowCat nodes have been used, one on Computer Systems (CS) and another on Automata Theory and Formal Languages (ATFL). In both cases, activities with specific aspects for the AM tests have been designed with support for users' interest, which were developed throughout 2006-2007. The CS node continues with the development of the subject matter initiated one year before. During the new academic year around 90 students have contributed about 160 documents distributed among 40 new topics, which are added to about that many existing ones and over 180 documents previously developed. The ATFL node gathers the documents made by the students based on others contributed by a teacher throughout an academic year in a knowledge tree on topics related to the corresponding subject. Around 90 students have worked on 6 topics and almost 450 documents in total.

4.1 Identification Experiments of User Interest

The first experiment uses the CS node so that students can develop the subject outline based on the notes and references provided during the year. The idea is to develop papers that are useful for preparing the development of one of the topics during the subject test. The experiment has two stages: In the first stage each student is assigned a couple of topics so that they can prepare the corresponding papers and publish them in the system; in the second stage, each student is assigned three more topics so that they can check, make comments and mark the work carried out by their classmates in

the first stage. After this process, the system arranges the papers according to quality in the corresponding knowledge tree. As a result, students have at their disposal 40 new topics developed for the exam preparation and can get up to one more point in the subject's final mark, depending on their participation in the activity. In order to make follow-up possible, students must identify themselves at the beginning of each session on the system.

At the end of the experiment the LOG file registered during the development has been processed. Based on it, user interest vectors (UIVs) have been established for two activity intervals: the first, from the beginning of a time interval until the end of the first stage, paper preparation stage on the topics assigned; and the second, from the beginning of the second stage until the end of this stage, revision interval of the papers included in the new topics assigned. To illustrate this experiment, the second interval UIVs are shown in Figure 1, in which the UIV of each user is represented by a column of colour blocks. Each block shows the users' relative level of interest for each topic, so that the lighter colours indicate a higher degree of interest than the darker colours. In the diagram, each column corresponds to one user and each row to a topic and the topics assigned for checking to students are marked with a cross. Some users have no topics assigned at all, because they are not students. In some cases there are more topic assignments than expected, these are errors in the assignment process that were corrected during the development of the experiment, but remained registered in the LOG.

As we can see in Figure 1, the topics assigned to users in the second interval are found to be the most interesting to the vast majority of them, since the assignment marks are usually placed on some lighter colour block in each column. This phenomenon is also obvious in the first interval. Another phenomenon that turns the attention in the diagram is students' little apparent interest for the topics they have not been assigned. This could be due to experiment conditions, activity periods observed and characteristics of the group involved. The experiment has been designed to cause an artificial interest in users for specific topics at certain moments and make the necessary activity easy to monitorize. Students' motivation in the periods observed is to get an extra mark for participating properly in the activity, but not for using the material prepared for studying -students usually prepare their exams in the last minute-.

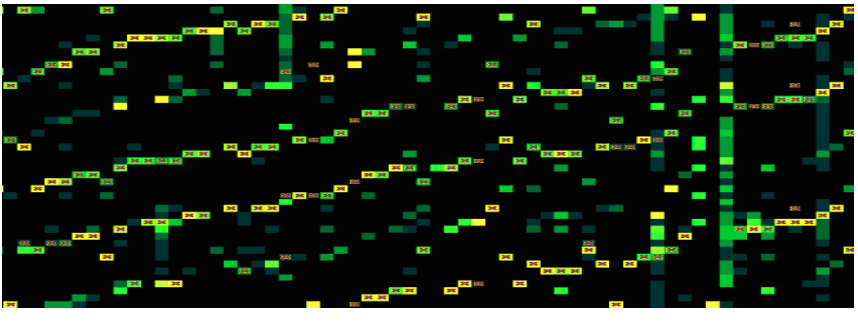

Fig. 1. UIVs of the second activity stage in the experiment with CS node

Fig. 2. UIVs of the third activity stage in the experiment with ATFL node

The second experiment carried out uses the ATFL node mentioned above. The activity consists of making five short summaries on some other topics of the subject by using the reference papers provided by the teacher through the system. There are two weeks for the preparation of each paper. At the beginning of each period one working topic per student must be chosen, so that the students do not repeat the same topic throughout the activity and there are no more than a maximum number of students per topic and interval. The summaries delivered in each period remain placed in the system with restricted access, only authorized to the teacher in charge of their assessment. According to the work carried out, each student can get up to one point more in their subject mark. As in the previous case, participants in the experiment must identify themselves at the beginning of each session to allow their activity follow-up.

After the experiment the system LOG was processed to establish User's Interest Vector (UIV) for the five document delivery periods. The UIVs corresponding to the third stage are shown in Figure 2. Each column corresponds to one user and each row to one topic. As in the previous case, the topics assigned to the students in the illustrated interval are marked with a cross and some users have no topics marked because they are not students. The UIVs representation is the same as in the previous experiment and the same chromatic code is used, in which the lighter shades of colour the higher degree of interest. In analogous way to what happened in the previous experiment, students appear to be more interested in topics which are clearly among the ones assigned to users in each period. On this occasion, the phenomenon is more obvious than in the previous case, given the characteristics of the experiment. Firstly, the knowledge tree has been prepared for the activity, presenting a number of topics adjusted to the latter. Secondly, for summary preparation the documents published in the corresponding topics of the knowledge tree itself must be revised. Thirdly, students are not allowed to access the documents delivered. In addition, the only incentive for participating is the extra mark for the fulfilment of the assigned tasks. Lastly, owing to all this, students are hardly interested in topics that have not been assigned to them in each interval.

4.2 Content Identification Experiments of Users' Interest Inside and Outside SKC

In the above experiments we have shown the use of User's Interest Vector (UIV) to represent how members' attention of a community is distributed among the knowledge node elements they belong to. UIVs have no sense outside the context where they have been defined, because they refer in an explicit way to specific elements of the latter and do not provide directly any information that can be used outside this context. Now we are going to test the benefits of User's Interest Words Weight Vector (UIWWV) to make use of the information previously collected in the UIVs in a more general manner and with application inside and outside the node where they were produced.

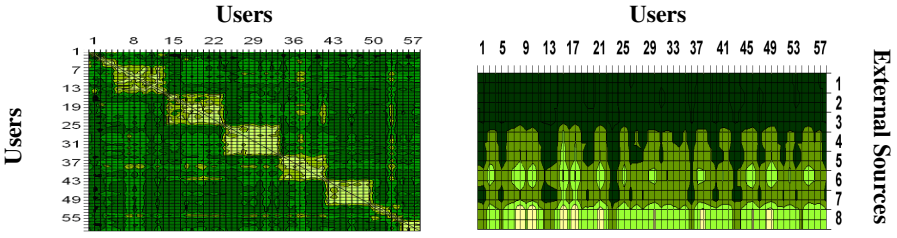

Fig. 3. Grouping of users by interest through their VPPIUs (left) and classification of external knowledge sources by user interest through VPPIUs (right)

The first experiment carried out on this occasion consists of comparing among themselves the UIWWVs generated from the UIVs obtained in the first of the previous experimental, which was carried out with a CS node and was introduced in section 4.1. This activity consisted of proposing students two topics first, to carry them out in each paper, and after another three topics, so that they would evaluate the work provided by different users. As a result, during the activity, students have appeared to be more interested in the assigned topics than in the rest of the contents in the knowledge tree node used, which was really the intention. Topics were divided into eight packages and were assigned to students in a systematic manner; therefore the topic group which has been established for each of them can be known. The packages included a varied selection of CS forty topics taken into account for the activity. The corresponding UIWWVs have been calculated according to interest distribution - represented by the UIV of each user- and to the WWVs of the papers involved, as explained in section 3. Users' UIWWVs have been compared among themselves, to check if they made clear the common interests users, using typical information retrieval techniques [2] mentioned above. The result has been shown in a graph (see Figure 3 left) where all the participant users are represented in both x and y axes, arranged by the assigned topic groups. Similarity value is shown with small colour blocks, being a lighter colour the greater the represented coefficients are. As we can see in the graph, the higher values appear grouped around the diagonal forming blocks. In proposed conditions, this indicates that the UIWWVs allow to identify users who are interested in the same topic groups, using generic descriptors that may be used to compare them with any object that has a WWV associated. However, we can only clearly see five blocks in the diagonal instead of eight, as would have been expected from the number of topic groups established. This may be due to the number of users assigned to each package that have taken part effectively in the activity, and to the loss of significant characteristics in some of the UIWWVs of those packages, as a consequence of specific combinations in diverse topics.

The second experiment carried out consists of applying the UIWWVs generated in the previous experiment to identify knowledge sources outside the Computer System node used there. For this reason, several knowledge virtual sources have been established on different topics based on documents accessible through the Web -such as papers on other KnowCat nodes or on Wikipedia and teachers' notes-. Specifically, eight virtual repositories have been prepared on the following topics: (1) Philosophy, (2) Current History, (3) Biology, (4) Data and Information Structure, (5) Automata

and Formal Languages, (6) Software Engineering (7) Operating Systems and (8) Computer Systems. As we can see, the last five repositories deal with subjects related to the field of Computer Science, whereas the three first ones have nothing to do with this area. The last repository specifically refers to Computer Systems, which is the field of node Computer Science, which students have had to work on and for which they were compelled to show interest in an artificial manner. At no given time documents included in the reference node -where UIWWVs come from- have been used.

Based on the documents in each repository a WWV representative of the latter has been obtained. These WWVs have been compared with users' UIWWVs generating a similarity graph (see Figure 3 right). In the graph each row corresponds to one user and each column to a repository -identified by the number assigned to its topic in the list of topics mentioned in the above paragraph-. The same as in graphs of other previous experiments, the greater similarities have been represented by lighter shades and the minor ones by darker shades. As we can see, the WWVs of the topics that are less related to Computer Science have much minor similarities with UIWWVs than the ones of the so mentioned knowledge area, something which seems to be quite reasonable. Within the WWV's of topics related to the reference node, the one in Operating Systems has greater similarity with some users' interest than other topics, what makes sense if we look at the reference node contents. Lastly, it is obvious that the greatest similarity is concentrated in the virtual repository on Computer Systems, what was also to be expected.

Under the conditions of the experiment, we may conclude that representation of user interest through UIWWVs allows to identify knowledge repositories that seem to be objectively linked to the interest shown by the users. As a matter of fact, in the conditions of these experiments the UIWWVs are WWVs representative to the themes of each user's particular interest and are similar to WWVs of topics that form Knowcat node knowledge trees. Therefore, these vectors may be useful to compare themes that they represent to WWVs of diverse elements: documents, topics or users of the node itself as well as other KnowCat nodes, or even any external source represented by WWVs. However, UIWWVs seem to be very sensitive to user interest dispersal as it happens with UIVs and its use may decline if the context of its definition is not handled in some way, for example maintaining different UIWWVs for each user in different contexts.

5 Conclusion

In this paper, a model to represent user interest in a Web system for collaborative knowledge management without supervision has been presented. This model uses some User Interest Vectors (UIVs) to represent user interest for the elements that constitute the system knowledge repository. With these vectors it is possible to compare users' interests inside the system. In addition, this model uses some User Interest Words Weight Vectors (UIWWV) to represent user interest in such a way that it may be compared with Words Weight Vectors that represent any data element inside and outside the system. In addition, the results of a series of experiments aimed to test the practical application of the proposed approach have been shown. As a result of these experiments, we may confirm that: (1) UIVs allow to adequately identify the focal points of user interest in the system knowledge repository, (2) UIWWVs allow to

identify users with similar interests and (3) data elements related to users' interests inside and outside the system. In conclusion, the model proposed contributes to the work area that concerns us, with a new process to determine the key words that may represent users' interest and its importance; thereby starting from the intensity of user interaction with data elements through the system.

Acknowledgements. This research has been partially financed by the Spanish Ministry of Science and Technology, through TIN2007-64718 and TIN2008-02081/TIN projects, and by the Spanish Agency for the International Cooperation (AECI) through A/7954/07 project.

References

1. Alamán, X., Cobos, R.: KnowCat, AWeb Application for Knowledge Organization. In: Chen, P.P., et al. (eds.) ER Workshops 1999. LNCS, vol. 1727, pp. 348–359. Springer, Heidelberg (1999)
2. Baeza, R., Ribeiro, B.: Modern Information Retrieval. Addison-Wesley, Reading (1999)
3. Chang, G., Healey, M., McHugh, J., Wang, J.: Mining the World Wide Web: An introduction search approach. Kluwer, Dordrecht (2001)
4. Cobos, R.: Mechanisms for the Crystallisation of Knowledge, a proposal using a collaborative system. Doctoral dissertation. Universidad Autónoma de Madrid (2003)
5. Godoy, D., Amandi, A.: Modeling User Interests by Conceptual Clustering. Information Systems 31(4-5), 247–265 (2006)
6. Gudivada, V.N., Raghavan, V.V., Grosky, W.I., Kasanagottu, R.: Information Retrieval on the World Wide Web. IEEE Internet Computing 1(5), 58–68 (1997)
7. Kim, S., Fox, E.A.: Interest-Based User Grouping Model for Collaborative Filtering in Digital. In: Chen, Z., Chen, H., Miao, Q., Fu, Y., Fox, E., Lim, E.-p. (eds.) ICADL 2004. LNCS, vol. 3334, pp. 533–542. Springer, Heidelberg (2004)
8. Lam, W., Mostafa, J.: Modeling user interest shift using a Bayesian approach. Journal of the American Society for Information Science and Technology archive 52(5), 416–429 (2001)
9. Lieberman, H.: Letizia: an agent that assist web browsing. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, pp. 924–929 (1995)
10. Moreno, J., Alamán, X.: A Proposal of Design for a Collaborative Knowledge Management System by means of Semantic Information. In: Navarro-Prieto, R., et al. (eds.) HCI related papers of Interacción 2004, pp. 307–319. Springer, Dordrecht (2005)
11. Moreno, J., Alamán, X.: SKC: Measuring the user's interaction intensity. In: Fernández-Manjón, B., et al. (eds.) Computers and Education: E-learning, From Theory to Practice, pp. 123–132. Springer, Heidelberg (2007)
12. Moreno, J., Alamán, X.: SKC: Digestión de Conocimiento. In: Proceedings of the VII Congreso Internacional INTERACCION 2007, Zaragoza, pp. 281–290 (2007)
13. Zhengwei, L., Shixiong, X., Qiang, N., Zhanguo, X.: Research on the User Interest Modeling of personalized Search Engine. Wuhan University Journal of Natural Sciences, 893–896 (2007)