Dynamic Profiling in a Real-Time Collaborative Learning Environment

Jaakko Kurhila[1], Miikka Miettinen[2], Petri Nokelainen[2], and Henry Tirri[2]

[1] Dept. of Computer Science
P.O. Box 26
FIN-00014 University of Helsinki
Finland
tel. +358 9 191 44664
Jaakko.Kurhila@cs.helsinki.fi
[2] Complex Systems Computation Group
Helsinki Institute for Information Technology
Finland
{Miikka.Miettinen, Petri.Nokelainen, Henry.Tirri}@hiit.fi

Abstract. EDUCO is system for collaborative web-based learning. Collaboration is enabled by real-time social navigation and support for social interaction via synchronous and asynchronous discussions. However, the system and the use of the system in collaborative learning could benefit from dynamic profiling as well as active recommendations for the students working in the learning environment. The paper describes the system and discusses the generation of profiles and recommendations.

1 Introduction

EDUCO is a system for on-line collaborative learning. EDUCO employs a form of *social navigation* [10] by visualizing the actions of other participants currently present in the learning environment. Since the actions of the users are visible to the other users in real-time, the social navigation is *direct* [1].

Many of the contemporary systems incorporating social navigation use *collaborative filtering*. It means that these "systems provide the user with recommendations of their likely interest in data items on the basis of 'interest matches' derived from ratings from the set of users" [2].

The approach to social navigation taken in EDUCO brings the *feel* of live companions into web-based learning [8]. The feel can promote pedagogically meaningful communication and collaboration. To amplify the social aspects of learning and to make the system adaptive for different needs, effective profiling of the students and active recommendations are appropriate features.

One approach to profiling of the users is to use data gathered with questionnaires. Kurhila et al. [7] introduced an adaptive questionnaire designed both for the construction of profiles and adaptive questioning. However, the scope of this paper is to examine the possibility of constructing dynamic profiles based on the users' behavioral patterns without any prior data. The profiles can be utilized

in EDUCO to support collaborative learning and to enable adaptive recommendation of potential documents to different users.

2 Description of EDUCO

From the users' point-of-view, EDUCO consists of six different views for various activities, a document area and a discussion area (Fig. 1). The views are map, chat, search, alarm, preferences and help. They are presented in a tool resembling a handheld computer (upper-left corner in Fig. 1, now in "map" view). The largest area on the right-hand side is the document viewing area for documents gathered into an Web-course in EDUCO. The space below the EDUCO (bottom-left corner in Fig. 1) is reserved for the asynchronous discussions.

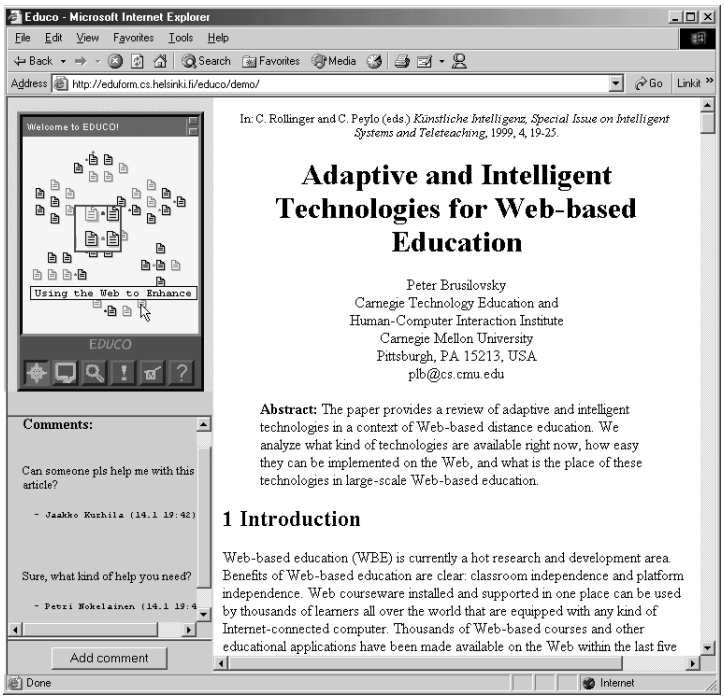

Fig. 1. The user interface of EDUCO. Map view presents documents gathered in EDUCO.

"Map view" presents documents ordered into clusters of in EDUCO. Documents are presented as paper-icons and the users are presented as dots. The colour of a dot indicates a group membership. The dot is located next to the document the user is currently viewing. When a user places the mouse pointer on top of a document or a dot representing a user, a tool tip text appears showing the name of the person or the title of the document. In Fig. 2, the pointer is on

a document called "Where did all the people go?". Double clicking a document opens the document into the right-hand side of the browser window and moves the dot representing the user to a corresponding location on the map view of every user in EDUCO.

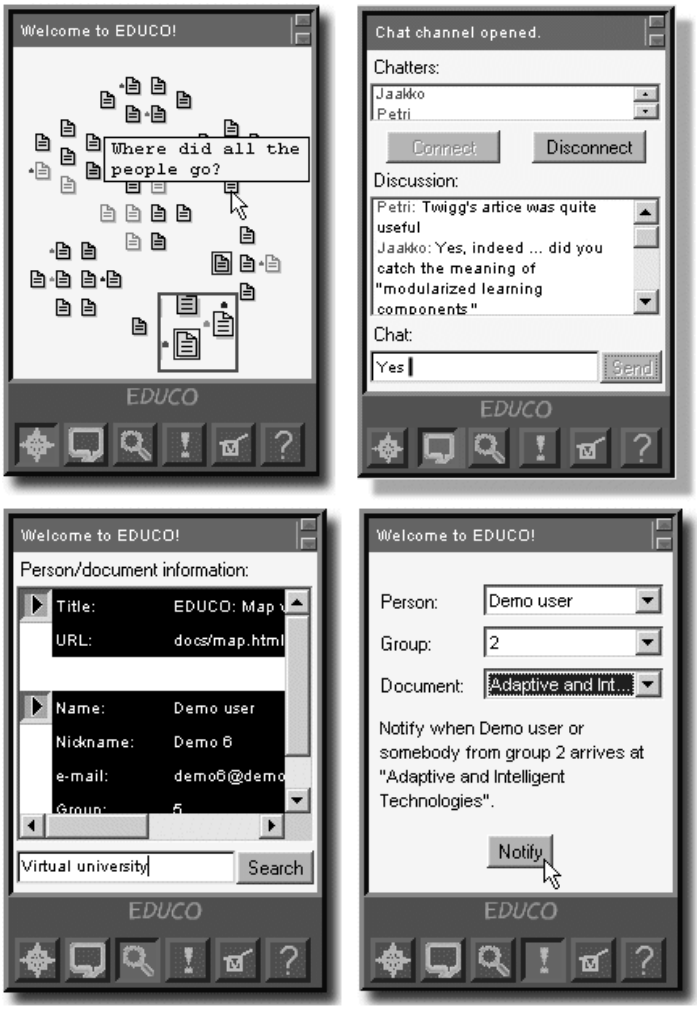

Fig. 2. The "map" (top-left), "chat" (top-right), "search" (bottom-left) and "alarm" (bottom-right) views of EDUCO.

The documents change their colour on the map depending on how much the users have viewed the document relative to the other documents. The colours range from bright to dimmed, indicating heavily viewed and nearly ignored documents, respectively. This way the user can get the historical navigation infor-

mation at a glance and does not have to be on-line all the time to know where the other users have navigated and where they have stayed for long periods of time. The change in the colour of an individual document is determined by the distance of its moving average for the last 24 hours from the same average for all the documents.

Other functions of EDUCO include a chat and a bulletin-board for asynchronous discussions. Any user can easily initiate a chat discussion with other users simply by clicking the corresponding user symbols and then clicking the "Connect" button in the chat view (Fig. 2). Asynchronous discussions are document-specific. Users of EDUCO can write a comment when viewing a document. The comment is visible to users navigating to that document. Other users can comment on the comment, thus continuing the chain of comments as illustarted in Fig. 1.

The third view is "Search". Users can search other users and documents in an instance of EDUCO. When a user searches a user or a document, the results are shown textually in search view (Fig. 2) and graphically in map view by highlighting the corresponding user or document.

"Alarm view" gives users the possibility to set up an alarm that is triggered if the requested condition occurs. For example, if a user seeks another user also interested in a certain document, he or she can tell the system to give a notifying message when someone else arrives to the document (Fig. 2).

The last two views are "Preferences" and "Help". While viewing "Preferences", the user is allowed to change personal settings in the system. Help view provides information about the system usage in general.

3 Educo architecture

From a technological point of view, EDUCO consists of a server, a Java applet for every user and a number of CGI-scripts. The most important task of the server is to keep track of the state of the distributed system and inform the clients as changes occur. An example of typical change in the system is a navigation step: If one of the users moves to another page, the new location has to be sent to everyone currently present in EDUCO. This type of implementation of communication without delays requires that the clients maintain an open socket connection to the server throughout the session.

To avoid copyright issues and to make the use of EDUCO simpler for the EDUCO-administrator (i.e., the course teacher), we have taken the approach that the documents (HTML-files) for a particular instance of EDUCO do not need to be copied to the EDUCO server. Instead, they can be located anywhere on the Web. To operate properly, the server still needs to know which document the user is reading to be able to send that information to all the other users in the environment. This has to work even when the users navigate along the hyperlinks in the documents and are not using the map view by double-clicking the document symbols. We have solved this problem by using the EDUCO server as a proxy. The documents are routed through the server instead of being sent

to the client directly from their actual location. Two additional operations are required: clients are informed about the new location of the user and all of the links in the document are changed so that they point to their destination indirectly through the proxy. If the user then navigates to another document along one of the links, the same procedure is repeated.

4 Dynamic profiling and document recommendations

4.1 Data set

The data set used for evaluating the profiling and recommendation possibilities in EDUCO was collected in the Fall 2001 during a course entitled "Web-based learning" given at the University of Helsinki, Finland. The course was a web-based course, and the use of EDUCO was mandatory. The course was an advanced course in Computer Science. Twenty-four students participated in the course, some of them adult learners with varying backgrounds and degrees but most of them were Computer Science majors.

The type of the course was a "seminar" which means that the students have to pick a topic, prepare a 10-page paper on a topic and present it to the teacher and other students in the course. However, there were also small weekly assignments to complete. These assignments required navigation in EDUCO, and interaction with the other participants in the course. The weekly assignments introduced the students to the concepts and issues essential to the course.

The course material was organized to six different document clusters in EDUCO. The clusters were: 1) Implications of Web-based education on the society, 2) History of Web-based education, 3) Web-based education research, 4) Pedagogical issues, 5) Adaptive educational systems and 6) Learning environments (i.e. course delivery systems). The document cluster sizes varied from six to nine, giving a total of 43 documents.

The course included only two face-to-face meetings. The first was an initial meeting where the structure and requirements for the course were explained and the EDUCO system was introduced. The second face-to-face meeting was the final meeting where the students presented their papers. Everything else between the initial and final meeting was conducted on-line using EDUCO.

Because of the small student population participating the course, it was possible that there will not be enough students at the same time using EDUCO. We wanted to make sure that students will see other live users in the environment, so we fixed a primary time slot for the students to visit EDUCO. However, the time slot was not restrictive in any way. In practice, the majority of the students visited EDUCO right after the weekly assignment was published on Mondays, and that was also the primary time slot.

4.2 Profiling the users

A significant amount of useful information about the students' interactions with the system was accumulated during the course. Exact data about the times

students viewed the documents and word histograms of chat discussions as well as the use of "search" and "alarm" was logged. Besides being of value to researchers studying the social aspects of group formation and collaboration [8], the data can be used for adaptation.

It could be possible to profile the users based on the data gathered by observing the social behaviour (chatting, comments, searching people and setting alarms) of the users. Besides social activity, dynamic profiling of the users can be based on other sources of information. The approach taken in this paper is to study the use of navigation and viewing times of documents, i.e. *navigational patterns*, as a basis for effective profiling.

The idea of dynamically profiling the users based on their navigational patterns defines some requirements for the technical implementation. We would like to form, update and discard clusters dynamically based on the data available at a particular point in time. However, there is no need to reconsider the whole clustering every time one of the students moves to another page. It is sufficient to update the profiles on a daily or weekly basis. The data set used in our preliminary analysis made weekly updating a natural choice. Most of the activity took place on Mondays, when the students were given small assignments due on the same day. For the reason, the experimental clustering was conducted with 8 different data sets, each consisting of the data accumulated by the end of the particular session.

The amount of the consolidated data was small from a statistical point of view, since there were only 24 students attending the course. This will likely to be the case with other courses of the same kind, so an important requirement for the clustering method is its ability to give useful results even with small data sets. In other applications it is often appropriate to balance the complexity of the model with the amount of data, but placing all the students to one cluster would not be helpful for our present purposes. On the other hand, algorithms that determine the number of clusters automatically would in principle be preferable.

As discussed above, the original idea was to profile the students on the basis of the amount of time they had spent viewing particular topics (i.e. groups of documents). However, some of the visits to the documents appearing in the logged data lasted for several hours. It is obvious that the user had left the browser window open rather than studied the document actively for the whole time. To prevent this type of data from distorting the profiles, we ignored additional time after a certain limit. The limit was set to 10 minutes, affecting about 9% of the data. It is certainly possible that more time was spent on active studying, but a relatively low limit is appropriate for evaluating the distribution of the viewing time over topics rather than individual documents.

Ideally, the clustering algorithm should learn incrementally from a small set of training examples, adjusting the number of clusters as needed. A method called COBWEB [3] attempts to address these issues. The clusters are represented as probability distributions calculated from the instances assigned to each cluster. The clustering process is guided by a heuristic evaluation function known as *category utility* [5] that attempts to maximize both the similarities within clusters

and differences between clusters. When encountering a new instance, COBWEB calculates the category utility for various alternative modifications of the previous clustering. New clusters are formed and old ones are splitted and merged as needed. Incremental learning, in addition to automatic determination of the number of clusters and applicability to small data sets, made COBWEB an attractive solution for the problem at hand. However, the method did not perform adequately on our data set. Almost every student was assigned to a different cluster, which made the results useless for the purpose. The problem, known as overfitting, is an issue in almost all forms of machine learning. Fortunately, better performance can be achieved with other methods.

One practical solution is the k-means algorithm [6]. The data vectors are first assigned randomly to k clusters, and the center of each cluster is determined by calculating the mean of the data vectors assigned to it. The centers and the clustering are then redefined repeatedly until a stable configuration is found. More specifically, each data vector is moved to the cluster that maximizes its probability, after which the cluster means are re-calculated. When none of the assignments changes any more, the algorithm has converged to a local optimum. Alternative solutions can be created by restarting from a new randomly generated clustering.

When using the k-means algorithm, the number of clusters has to be fixed beforehand. This is not a major problem, however, since the suitable range is rather narrow. In order to constitute student profiles of practical value, the clusters should not be too big, and our focus on the social aspects of learning implies that we do not prefer clusters of one or two people either. Considering the available data set of 24 users, the optimal cluster size iis in our opinion 5 to 6, suggesting that there should be 4 clusters. Trials with 3 and 5 clusters also produced satisfactory results.

After the clusters have been constructed on the basis of topic level viewing times, it is useful to examine differences within the clusters more carefully. In particular, it seems possible to implement a small scale recommendation system if the viewing statistics and the clustering algorithm are indeed successful in capturing useful information about the interests of the students. Some students in a cluster may have spent a reasonable amount of time viewing a particular document, and assuming that viewing time indicates interest, the document could be recommended to others in the same cluster who have not looked at it yet. The viability of the idea can be tested by looking if such differences exist, and if they have predictive power in the sense that sooner or later everyone in the cluster ends up viewing the interesting document. We currently base the recommendations on the following heuristic rule: if more than one third of the people in the cluster have viewed a certain document 5 minutes or one fourth of the people 10 minutes, the document is recommended to the others.

4.3 Empirical results

The clustering was conducted for 8 different data sets, each successive set including the additional data of one more week. The viewing times of the documents

were transformed into a discrete 3 point scale. Based on some experimentation, the ranges were set to 0–10 minutes, 10–20 minutes and more than 20 minutes. The k-means algorithm was executed with the amount of clusters fixed to 4.

About the clusters. Because of the small data sets (only 24 cases and 6 variables), it was straightforward to see that the results of the clustering appeared to be "good" in the sense that items within clusters were remarkably similar to each other and clusters were reasonably different from each other. From the eight periods, the only confusing exception was period 3. For some reason, all of the items were in one cluster, and the others were empty. In period 4, which contained the same data and one more week, the results looked good again. The third period was left out from the other analyses, since it seemed to be an odd exception among otherwise consistent results.

Figure 3 shows as an example of the clusters identified in period 1. The bars indicate the average amount of time spent viewing each of the topics of the course. As can be seen from the figure, there are clear differences between the clusters regarding both the total amount of time spent in the system and its distribution over the topics.

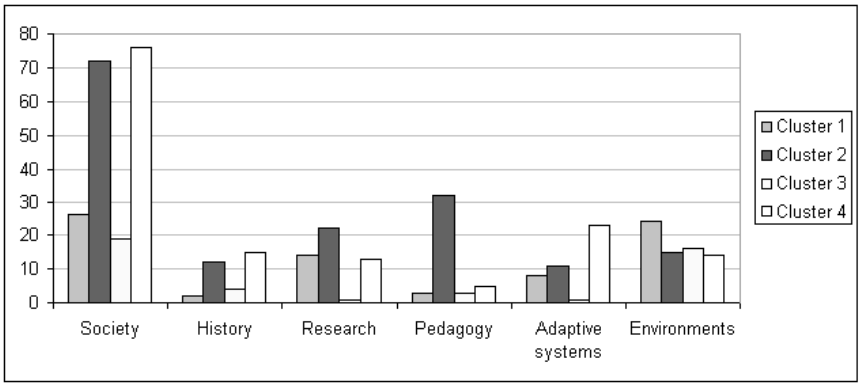

Fig. 3. Average viewing times of the topics in the first period.

The students in clusters 2 and 4 had been online much more than the ones in clusters 1 and 3. In all clusters the most popular topic was the "Implications of Web-based education on the society". The documents discussing those issues accounted for 34 to 52% of the total viewing time. "Learning environments" appeared to be another focus of interest in clusters 1 and 3. Their main difference is that the topic having the third longest viewing time is "Web-based education research" in cluster 1 and "Adaptive educational systems" in cluster 3. The students in cluster 2 seemed to be more interested in education than technology as documents related to society, pedagogy and research comprised 77% of the

total viewing time. In cluster 4 the most popular topics were "Implications of Web-based education on the society" and "Adaptive educational systems".

Stability of the clusters. In order to evaluate the stability of the clusters, we calculated the proportion of the students who stayed in the same cluster for two successive periods. Each cluster of the first period was paired with one of the clusters of the second period in such a way that the overall degree of similarity, defined in terms of the overlap between the two clusterings, was maximized. On average, the similarity was 78%. The similarities of individual pairs of clusterings varied between 46% and 96%, largely reflecting the rate of the accumulation of data. The more additional data was received during the period, the more the clusters tended to change. Only 5 people of the 24 stayed in the same (gradually evolving) cluster for the whole time. It seemed that there were no major variations between individuals regarding the difficulty of profiling them, since nobody was moved to another cluster more than 3 times.

Recommending the documents. The idea of recommending documents within clusters was also tested in a simulation. Since each clustering generates a set of recommendations independently, there is no straightforward way of comparing the results of different periods. For this reason, we did our analysis on one data set chosen from the middle of the course (period 4). Significant differences in the viewing times of individual documents were first identified using the criteria discussed above. The system would have given 42 recommendations in total. Only 3 of these were found in the smallest cluster, which contained only 2 students. It is obvious that the variation among 2 people cannot be sufficient for the purpose. The other clusters received 12-15 recommendations each. For individual students the number ranged from 0 to 6, with an average of 1.8. People who had spent the least time in the system were generally in the upper end and the most active readers in the lower end.

The recommendations generated at the middle of the course were also compared to the final data set, which included the actual viewing times of each document. In particular, it was interesting to see if the students eventually found their way to the potentially meaningful material without guidance. The proportion of recommended documents, in which the particular students spent more than 5 minutes during the latter part of the course, was 18%. Since 31% of all viewing times were above 5 minutes, the students were actually less likely to indicate interest in the recommended documents than all documents in general. However, in 67% of the cases the recommended document was not visited at all. Therefore, it seems possible that the recommendation facility would be helpful in finding relevant material.

5 Conclusions

EDUCO is a learning environment with built-in support for real-time on-line social collaboration. The underlying principles of EDUCO could benefit from

dynamical user profiling and active recommendations of potentially interesting documents to users. The paper presented a method to provide the profiling and recommendations using a data set gathered from an actual web-course where EDUCO was used. Analyzing the data suggests that merely the navigation and viewing of documents are enough to provide sufficient data for meaningful and efficient profiling of the learners, and the adaptive recommendations are possible from the same data.

Examining the navigational patterns of the users of EDUCO elaborates the idea that human thinking can be viewed to occur at least partly outside the mind [4]. In the case of EDUCO, *insight* (new learning method or view into the course material) can come from *outside* (new tools for direct social navigation) the mind. In a way, McCalla et al.[9] address the same issue when they state that intelligent tutoring systems of the future do not have a learner model as a single distinct entity. Instead, they will have a virtual infinity of models, computed as needed during the learning process.

References

1. Dieberger, A.: Social Navigation in Populated Information Spaces. In A. Munro, K. Höök and D. Benyon (Eds.), Social Navigation of Information Space, pages 35–54. London: Springer (1999).
2. Dourish, P.: Where the Footprints Lead: Tracking Down Other Roles for Social Navigation. In A. Munro, K. Höök and D. Benyon (Eds.), Social Navigation of Information Space, pages 15–34. London: Springer (1999).
3. Fisher, D.H.: Knowledge Acquisition via Incremental Conceptual Clustering. Machine Learning 2, 139–172 (1987).
4. Gigerenzer, G.: Adaptive Thinking. New York: Oxford University Press (2000).
5. Gluck, M.A. and Corter, J.E.: Information, uncertainty, and the utility of categories. Proceedings of The 7th Annual Conference of the Cognitive Science Society, pages 283–287 (2001).
6. Jain, A.K., Murty, M.N. and Flynn, P.J.: Data Clustering: A review. ACM Computing Surveys 31, 264–323. (1999).
7. Kurhila, J., Miettinen, M., Niemivirta, M., Nokelainen, P., Silander, T. and Tirri, H.: Bayesian Modeling in an Adaptive On-Line Questionnaire for Education and Educational Research. Proceedings of The 10th International PEG2001 Conference, pages 194–201 (2001).
8. Kurhila, J., Miettinen, M., Nokelainen, P. and Tirri, H.: EDUCO - A Collaborative Learning Environment using Social Navigation. To appear in Proceedings of Adaptive Hypermedia and Adaptive Web-based Systems (AH2002)(2002).
9. McCalla, G., Vassileva, J., Greer, J. and Bull, S.: Active Learner Modelling. Proceedings of the Intelligent Tutoring Systems (ITS2000), pages 53–62. Berlin: Springer (2000).
10. Munro, A., Höök, K. and Benyon, D.: Footprints in the Snow. In A. Munro, K. Höök and D. Benyon (Eds.), Social Navigation of Information Space, pages 1–14. London: Springer (1999).