Training personalization with Knowledge Technologies and Contextualization

Davor Orlič
Centre for knowledge transfer in information technologies
Jožef Stefan Institute
davor.orlic@ijs.si

Mitja Jermol
Centre for knowledge transfer in information technologies
Jožef Stefan Institute
mitja.jermol@ijs.si

Abstract

The state of the art in educational technology at the moment is limited to few technological platforms and devices which cannot allow learners to gain the most out of education content and context. In this article we will present some technologies for educational contents that will break the barrier of learning applications and will demonstrate some advanced knowledge technologies combined with contextualization tools which can do on the fly personalization for users of specific knowledge domains.

1   Introduction

The current state of the art in educational technologies is mostly based on user environment personalization and not content personalization. Technological platforms and learning devices enable only superficial or design oriented personalization which affects only the learners viewing, interaction and not full educational experience. In order to shift the focus centre from superficial or design oriented to deep knowledge derived experience we developed a new set of applications that will provide the learner with a completely new learning experience namely a personalization of content for his specific needs based on a deeper profiling set of information. In order to do this we developed a suite of tools containing a service oriented text enrichment tool, a user targeting data mining tool and a recommendation content system based on user preferences. We will also present relevant results from a questionnaire on training in academic and business institutions which points out a need for more advanced technology enhanced functionalities and a more holistic user/learner oriented use of educational content within the current state of technology enabled education. Finally we will describe a typical learner using the proposed new system in an admittedly hypothetical mode.

2   Background

The potential of these tools can solely be used if there exists a sufficiently large pool of educational content. As content can only be created in educational institutions we aimed at establishing an umbrella initiative that could create a large enough sustainable pool of knowledge and educational content. For this purpose the Knowledge for all Foundation will be created, which will join together an alliance of established international institutions into a forum for discussion and dissemination of advances in innovations in technology enabled education at the university level in established and emerging nations.

Knowledge for All Foundation for the purpose of this article theoretically could mainly try to bridge and give additional value to two large education oriented consortiums namely the OpenCourseWare Consortium (http://www.ocwconsortium.org/) and Matterhorn Opencast Project (http://www.opencastproject.org/). OpenCourseWare is an initiative that produces free and open digital publications of high quality educational materials, organized as courses and is a collaboration of more than 200 higher education institutions and associated organizations from around the world creating a broad and deep body of open educational content using a shared model. Matterhorn is an open source project working within the Opencast Community to develop an end-to-end, open source platform that supports the scheduling, capture, managing, encoding and delivery of educational audio and video content. In this way OCWC brings a large pool of educational institutions which

produce value content in textual form and Matterhorn can be a possible solution for captioning video content within those institutions. The dissemination of the video output from these two projects could appropriately be channeled and provided by the educational video repository VideoLectures.NET (http://videolectures.net/). The main purpose of this web page is to provide free and open access of a high quality video lectures presented by distinguished scholars and scientists at the most important and prominent events like conferences, summer schools, workshops and science promotional events from many fields of Science. The technologies for training personalization taken into consideration in this article will be implemented in the Matterhorn Opencast project as VideoLectures.NET is also one of the core partners within this project.

Knowledge for All Foundation at the moment has a programme of five pillars of activity namely:

- Infrastructure: ICT Matterhorn - Interoperability, Channels, Semantics
- Science: Journal and conferences
    o Online scientific video journal to global university
- Education: courses and content
    o Quality assurance – peer reviewed content
- Research:
    o facilitating the systems, accessing the content, enabling interaction
    o IPRs, multilingualism, standards
    o Business models (added value models)
- Other continent connections: case study in engagement and interaction

As we try and combine the two initiatives there is no apparent focus on the user's needs or perception of how users will manage to obtain reasonable knowledge from this pool of textual and video content. As text and video represent knowledge or basic fragments of it, there is a need to extract and properly display in a personalized way the information users might be looking for or need. Therefore the personalization of the users experience must be forwarded and provided with new existing technologies not yet used in the educational sphere as such. These are based on the following Knowledge Technology techniques:

- Text-Mining, Data/Web-Mining, Stream Mining
- Network-Analysis
- Statistical machine learning
- Semantic-Web, Context technologies
- Complex data visualization
- Cross-modal analytic
- Semantic Web Services
- Language technologies
- Knowledge formalization, Reasoning

Within the framework of Knowledge for All Foundation the main purpose is to use the before mentioned Knowledge Technology techniques for providing content knowledge services such as:

- On-the-fly personalization, contextualization and recommendation
- Video scene recognition, automatic annotation and categorization
- Semantic and multilingual search
- Accessibility, Internationalization (subtitles, transcripts)
- Advanced presentation services with direct user involvement
- Textual, graphical, video (audio) content integration services and enrichment

In this article we will further present and focus on one section of services namely On-the-fly personalization and contextualization.

3 Educational Training Needs Analysis and Users Study

In order to understand the real situation and basic needs of users and learners we present a selected set of questionnaire results of the users study that was aimed at getting information about current training habits of educational institutions and about their training needs and preferences. For this purpose we used an existing survey that was made by the Centre for knowledge transfer in information technologies, JSI for a European

Research project COIN [1] (Collaboration & Interoperability for Networked Enterprises). The plan in Knowledge for All Foundation is that more detailed assessments should be made in order to refine strategies and approaches. Based on the results from this analysis the technological activities of Knowledge for all Foundation and plan have been developed.

3.1 Questionnaire

To the extent that it is practicable and meaningful, the user needs and preferences for training in knowledge intensive organization have been assessed for the organizations according to several criteria:

- Organization type (enterprises, SMEs, industrial associations, professional communities, research organizations and academic institutions),

- Organization focus: industry, service, mixed

- Geographical location: west Europe, east Europe, rest of the world

Since quite an extensive research has been done in the past by European Research project ECOLEAD (only academic institutions in 2006), ToolEast (only industries in 2007) and currently in ACTIVE (industries and academic institutions in 2008), we have used these results and enrich them with some additional data that was gathered from the questionnaires sent to some of the target group representatives that were not contacted before. Those were in particular national industry associations (Chambers of Commerce and industry clusters, enterprises and academic organizations from the countries that were not taken into consideration before) and organizations inside leading Universities in the world that are responsible for Technology Enhanced Learning from the two communities OpenCourseWare Consortium and Matterhorn Opencast Project.

Because of the two distinct worlds of academia and business we decided to prepare two questionnaires one for the businesses and the other for academia. The aim of these questionnaires was to get the information about:

- Existing educational programs in the academic institutions and methods of learning,

- Existing training behavior but also training needs and preferences of business type of organizations.

We have asked business organizations what type of learning methods they use in their training. Figure 1 show that the most used methods are traditional training seminars, self-learning and consultations and also ICT (Information and communication technologies) supported learning. What is interesting also is the information that the institutional training is not that important. When the organizations answered with "other" to this question some of them proposed additional methods that were not mentioned in the answers. 85% of answers were proposing collaborative type of learning (collaborative learning, collaborative problem solving, group learning, and learning by help of social software).
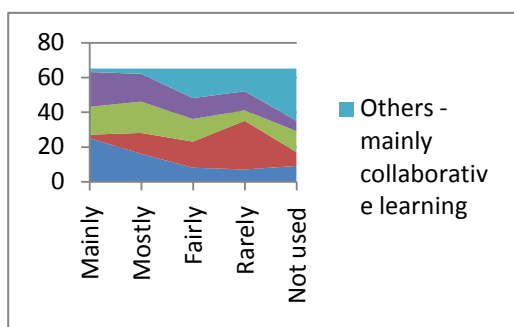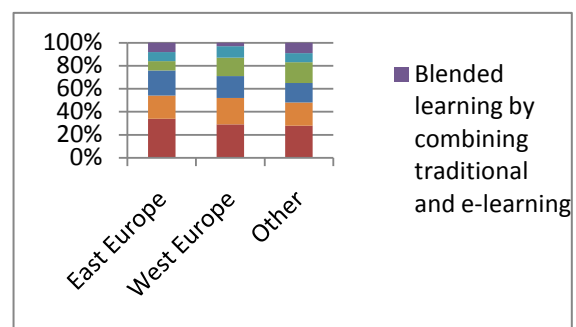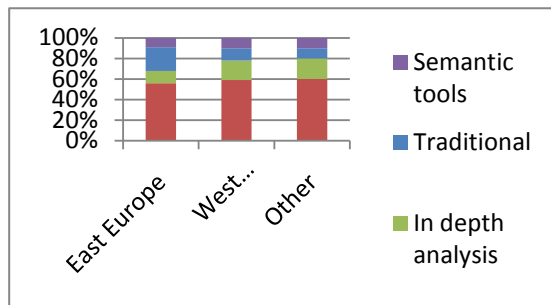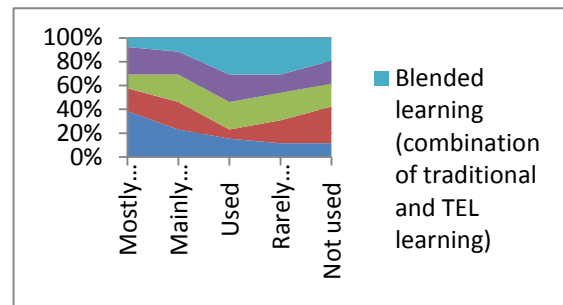

Figure 1: Learning methods used

Figure 2: Preferred methods of learning

When asked about their preferred type of training events the answers as showed in Figure 2 were: self learning, face to face consultations in the company as well as online training (traditional workshops and e-learning) are the most important training activities. The offline learning (self-learning) is also quite important. The final question was asking about the type of methods, tools and techniques that organization use for the company structure, knowledge resources and social structure analysis. The answers show what was expected. Most of the

organizations are using traditional and statistical methods. Very few are using tools for more in-depth analysis or semantically enhanced tools.

Figure 3: Type of tools used for analytics

Figure 4: Used TEL methods in Academia

The first question asked to Academic institutions was about the technology enhanced learning methods that they might use. Answers show that there is not only high level of awareness about TEL (Technology enhanced learning) but also that many institutions already use in their normal operation TEL methods. The most used is ICT-supported self-learning that is mainly demonstrated as Web based learning. Surprisingly, learning through social software is also very popular. There is also some distance learning courses as well as online seminars.

We have also used results from the market studies made in ECOLEAD [2], PROLEARN [3], Tencompetence [4], Prolix [5] and Kaleidoscope [6]. Findings from the questionnaire that are interesting for Knowledge for All Foundation are as follows:

- learning from experience is preferred
- lecture-style delivery is unpopular
- only Web based/multimedia learning is not attractive
- the preference is to learn outside of the workplace
- courses longer than one day are difficult to attend
- training costs are usually too high

4   Knowledge technologies solutions

The answers in the questionnaire point in the direction that standard learning techniques, blended learning approaches, and similar are favorable in institutions but show a lack in penetration and not enough content retrieval to enhance and gain the maximum out of the educational potential of the content at hand. There is a clear need for stronger exploitation of technology enhanced learning (TEL). The ultimate goal of every TEL system is to support personalization of the complete learning process that includes: personalization of content, methods, guiding, motivation and learning goal. Since we are dealing mainly with the established training channel VideoLectures.NET, a predefined set of learning methods and self-learning approach, our focus is to personalize content objects (videos) based on user needs and preferences. In order to understand user needs and preferences on one side and provide learning environment and content adaptation on the other, advanced knowledge and context technologies needs to be applied. This is why we applied a set of tools that enables better representation, organization and exchange of information and knowledge and thus showing a more personalized approach for user solutions. We propose a suite of tools that are able to understand user behavior, the content that they are accessing and construct a user model accessed competences. Here we introduce a personalization process that is based on the three sets of tools that has been proven with their limited purpose in the real case implementations for (1) website visitors modeling and segmentation, (2) content recommender and (3) content contextualiser. Personalization loop as seen through knowledge technologies has a strong focus point on online users' dynamics and can depict a very clear picture of the actual status, needs and desires regarding real time on the fly needs.
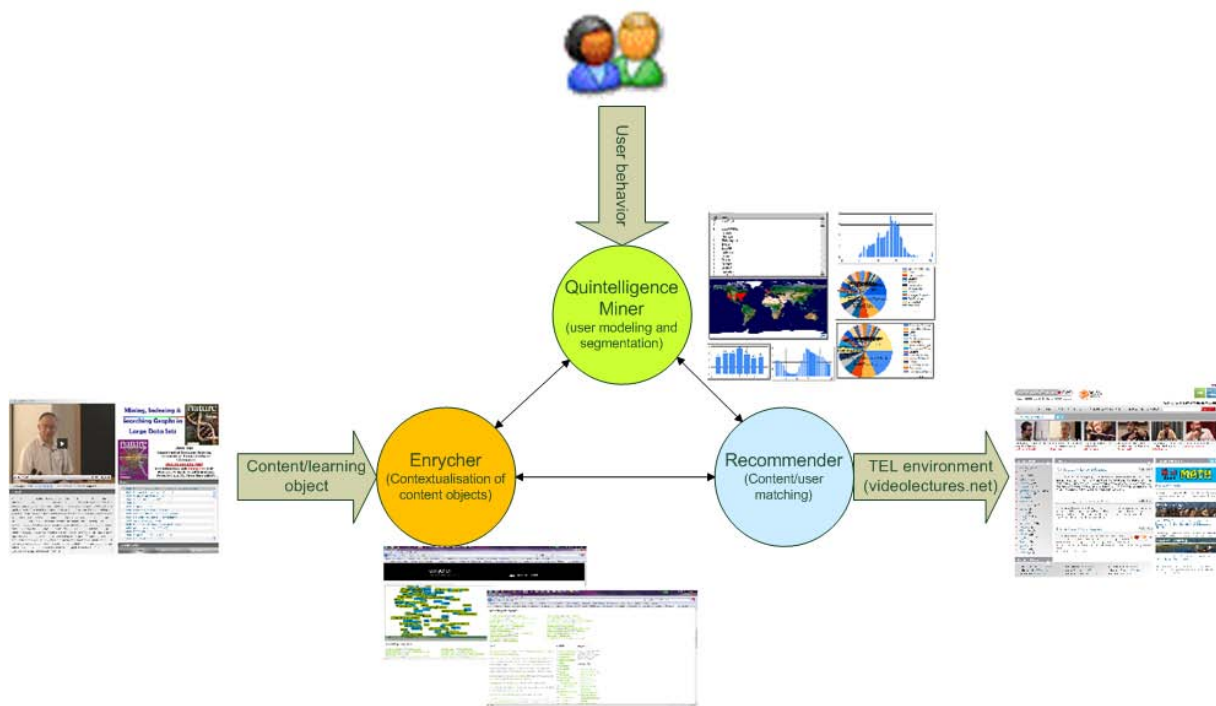
Figure 5: Developed personalization loop

In the chapters below we describe in more detail each of the three personalization toolsets.

4.1 User/Learner Targeting

Quintelligence Miner (QM) is a decision support environment that integrates several data mining techniques and OLAP-style splits of large-scale data stores containing structured information (e.g. customer information). QM is being used by online publishers to model and understand the users, their needs and preferences and behavior characteristics. It modeling is based on the content users are reading in every web service session. It can model a particular user, cluster them in the interest groups and develop the most used user scenarios which is in particular important for the personalization of training.

A unique part of the tool is the ability to handle unstructured data such as text. In default setting, the analytic user can filter and split the data along any dimension in structured (e.g. gender or age) and unstructured data (e.g. topic, keywords, named entities). The system can aggregate all other dimensions for each split and visualize those using standard techniques for structured data (e.g. pie charts, histograms, world map). Aggregation of unstructured data is done using text mining techniques such as clustering, feature extraction and text visualization. An additional result of each split is a machine learning model which can be used for database fields (e.g. gender of a customer based on their shopping or reading behavior).

QM presents a successful approach for user modeling and understanding user behavior which is the entry point and first part of functionalities in the personalized training loop.
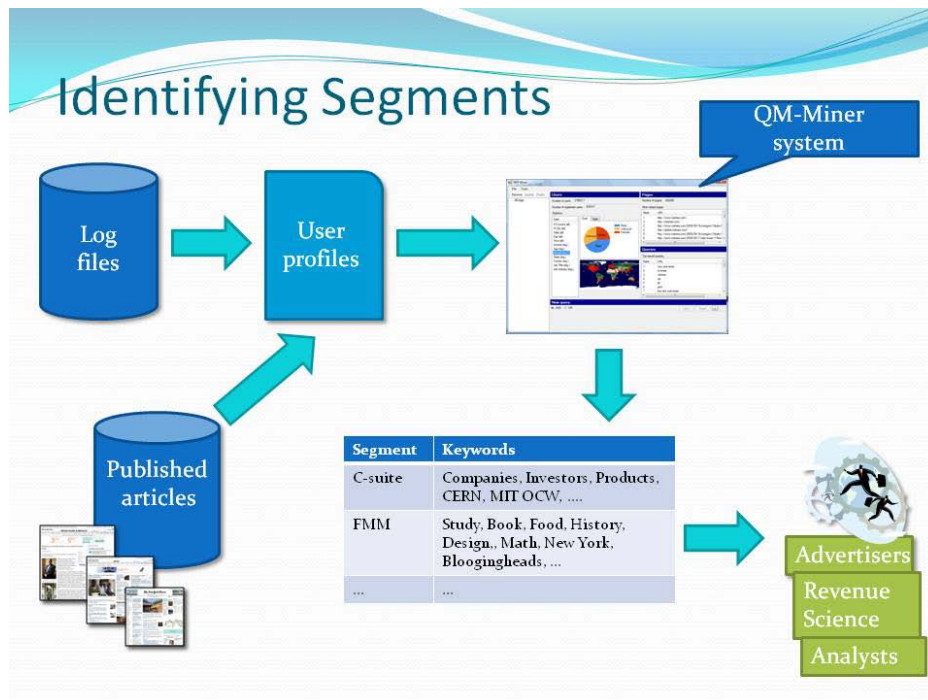
Figure 5: Quintelligence Miner Architecture

4.2   Recommendation Content System Based On User/Learner Preferences

The recommendation system currently used in the online video educational repository VideoLectures.NET (http://videolectures.net/) is another functionality that reflects a personalized service delivery to the user/learner.

The growth of information on World Wide Web makes users/learners more difficult to search for relevant information as the amount of product in e-education increases rapidly. Learners greatly suffer from searching for interested products or content. To avoid this problem the recommendation system helps users in finding the right resources they are looking for in a certain environment. Data collection in the recommendation system is managed with content-based filtering which includes observing the content that the user views online, analyzing the content/user viewing times, storing data about the content that a user views in log files format, analyzing the user's social network and discovering similar likes and dislikes. More explicit collecting of data includes the activities within collaborative filtering monitoring such as users rates on specific content on a sliding scale. The system makes crossover calculations by using a specific set of algorithms which combine users view history and server log files and the textual, temporal, visual metadata related to the specific content, in this case video lectures and which eventually show a logical and semantically based recommendation set for the user. Therefore by using a specific set of knowledge content a suggested and contextual preference set can be made therefore again personalizing the training experience.

This recommendation system presents an approach which comes after the user modeling and behavior understanding as it derives the users' behavior once the user reaches the content. It presents the second part of functionalities in the personalized training loop.
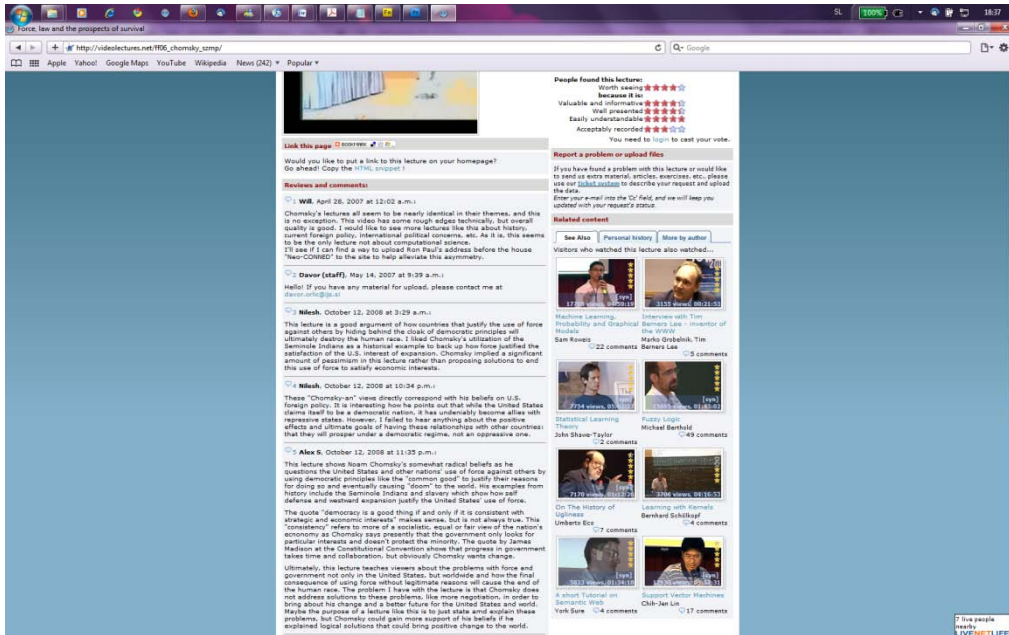
Figure 6: Recommendation system on VideoLectures.NET

4.3 Service oriented text enrichment

Enrycher (http://enrycher.ijs.si) is a set of web services that enable automated content enrichment and knowledge extraction functionalities. Enrycher services form an umbrella part of the basic services of innovative tools that enable building up many complex knowledge extraction scenarios that are needed in the environment of training personalization. Enrycher consist of several levels of basic technologies:

- Language processing services: Sentence splitting, Tokenization, Part of speech tagging, Entity extraction
- Entity level processing services: Named entity extraction, Anaphora resolution, Co-reference resolution, Semantic entity resolution
- Entity graph processing services: Triplet extraction
- Document level processing services: Semantic graph visualization, Taxonomy categorization, Content summarization

The schema that is used in the inter-service communication is abstracted to the point that it is able to represent:

- Document-wide metadata: identifier, document wide semantic attributes (e.g. categories, summary)
- Text: sentences, tokens, part of speech tags
- Annotations: entities and assertion nodes, identified in the article with all identified instances, possibly also with semantic attributes (e.g. named entities, semantic entities)
- Assertions: identified <subject, predicate, object> triplets (where subjects, predicates and object themselves are annotations.)

The datasets that can be used is any type of textual information. Since these are the basic services for dealing with textual information, many new applications can be built on top of it. In the case of training, these services are related to competency extraction and management, contextualized search over distributed information sources, information categorization, knowledge extraction and formalization, document linking and context preservation for the archives, educationally relevant topic detection and many others. We have already implemented some more general services that were also used in European Research Project COIN platform like (1) visual analytics services, (2) semantic integration of texts and ontologies, (3) question answering service and (4) story link detection service. Another potential use scenario is from a related domain of computational linguistics as evaluating local discourse coherence of text or extracting knowledge from large-scale document collections, such as news corpora, training course documents, etc.
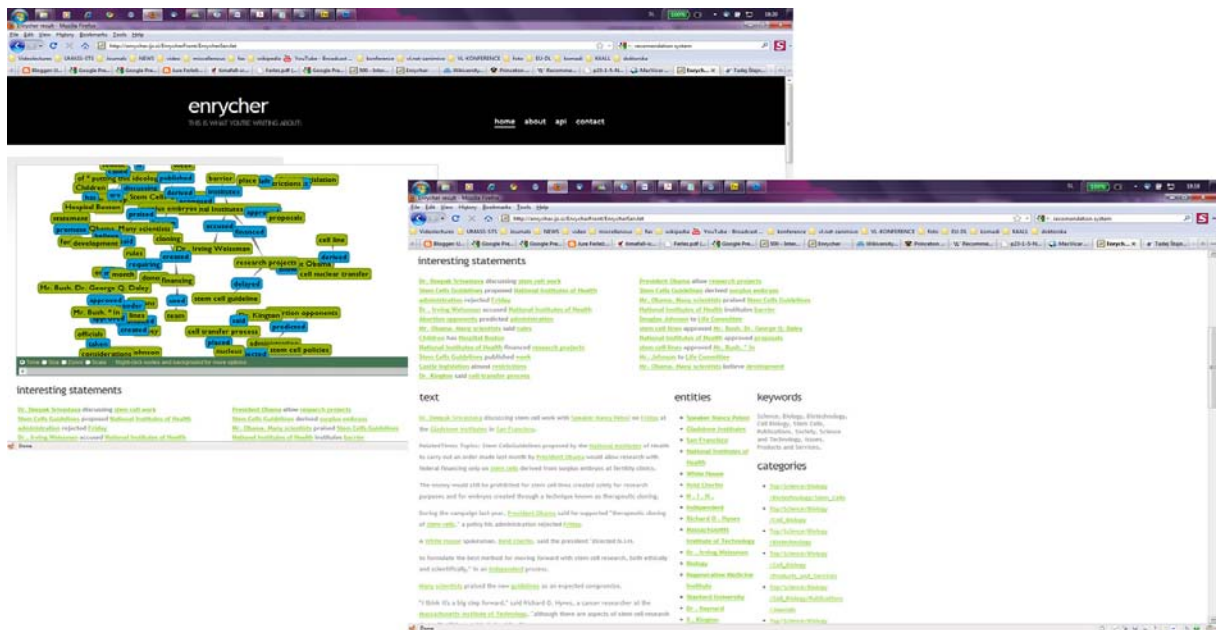
Figure 7: Enrycher

5 Personalization learning environment test case

As the main purpose of this paper is helping the learner in new and unique ways we would like to propose a hypothetical scenario of a learner trying to access a course on an educational web page that consists of video, audio and textual content. For the purpose of this paper we decided to hypothetically use VideoLectures.NET (http://videolectures.net/) as the testing page although we feel that these tools should be implemented and used on a more broad and higher level repository or unique set of pages with educational content. The test page has been pre-monitored for a certain period in order to accumulate a sufficient minimum amount of user data in order to understand the user culture and trends that inhabit the page.

Suppose we know that a user comes to this particular web page every day and browses its content. The content the user is reading and watching in every web service session is monitored and stored; we know on the fly the users geographical location, country location, type of personal computer, browser, internet connection, gender, social status (student, worker), financial income, native language, etc. By this user/learner targeting we later model this particular user, cluster him in an interest group of similar users and get information about his browsing experience through the page via hints (not direct questions), that way we develop a set of most used content scenarios that he will decide to take. This is the user modeling and user segmentation part.

A this point we know his details and content viewing habits so we can use an automatic recommending system in order to show him what his interest target of users group saw based on the same viewing habits. If his viewing habits show that he is from a technical university, is an English speaker, has a good internet connection, he is able to stream video, and is ultimately interested in Physics and additionally the system knows that he already watched the introductory courses of MIT professor Walter H. Lewin, namely "MIT 8.01 Physics I: Classical Mechanics - Fall 1999" thus recommends him the "CERN Colloquium on Fundamental Constants in Physics and their Time Dependence" as the system supposes that based on the content he has seen so far, his knowledge on the subject should be adequate enough to see the CERN Colloquium. Here the content and user matching takes place.

Once the user decides that his video session is over, he tries to find out more on the subject of Physics, so he decides to browse and use a text enriching function over the term "Physics"; the result he gets is a visualized, contextualized search over distributed information sources (lecture description, physics course text and descriptions, news articles on the same web page or on related pages), and get a delivered information categorization of the text over this specific chosen topic. The text before him is enriched with semantic attributes, sentences, parts of speech, that can follow up into different search and topic contexts. Here the contextualization of learning objects takes place.

This scenario can continue invariably as it represents the personalized ongoing learning loop where the learner gets driven into the content with the help of contextualized personalization tools. As this all happens on the fly, this means that for each decision the learner makes, the tools adopt a suitable solution that keeps on driving the user toward his needs.

6 Conclusion

In this paper we have presented a specific framework for training personalization based on advanced knowledge technologies combined with contextualization tools which can do on the fly personalization for users of specific knowledge domains. All three technologies take into consideration that there is an existing pool and deep body of structured open educational content which has an immense knowledge potential and is not fully exploited towards learners needs. The Centre for knowledge transfer in information technologies at the Jožef Stefan Institute is familiar with these technologies, has implemented them on several test cases and proposes for the first time such an analytic and innovative approach towards training. As these technologies have in practice been used separately for specific purposes when merged on one location and in one tool suite, they can become a powerful technological resource for the broader learning community.

7 References

[1] COIN 2008 D2.3.1. Training set-up and assessment deliverable, edited Marko Conte, Uninova, 2008 COIN IP (FP6 IP 216256)

[2] ECOLEAD 2005   D72.5 Training set-up and assessment, edited Jermol M., JSI, 2005, ECOLEAD IP (FP6 IP 506958)

[3] PROLEARN 2004   D1.03 Learner Models for Web-based Personalised Adaptive Learning: Current Solutions and Open Issues, Aroyo, L., 2004, Prolearn NoE (IST 507310)

[4] TENCOMPETENCE 2007 M2.1 Initial Requirements Report, Miguel, A, 2007, TENCOMPETENCE IP (IST-2005-027087)

[5] PROLIX 2006 D1.1. PROLIX Requirements Analysis Report. Herrmann, K., 2006, PROLIX IP (IST-FP6-027905)

[6] KALEIDOSCOPE 2007   D15.2 Case Study on social software in distributed working environments, Kiesslinger, B., 2007, KALEIDOSCOPE NoE (IST 507310)