Ontology-Centered Personalized Presentation of Knowledge Extracted From the Web[1]

Stefan Trausan-Matu[1], Daniele Maraschi[2], and Stefano Cerri[2]

[1]Computer Science Department, "Politehnica" University of Bucharest
Romanian Academy Research Centre for Machine Learning, Computational Linguistics, and Conceptual Modeling, Calea 13 Septembrie nr. 13, Bucharest
ROMANIA
trausan@cs.pub.ro, http://www.racai.ro/~trausan
[2]LIRMM, rue Ada, Montpellier, FRANCE
cerri@lirmm.fr, maraschi@lirmm.fr

Abstract. The paper presents an approach for the dynamic generation of a complex structure of personalized web pages for learning purposes, reflecting the ontology of the considered domain. The need of assuring a holistic character for the body of knowledge induced in the learner's mind is emphasized. This is very important in the learning processes, especially nowadays, in the context of the huge amount of information available on the web and of its permanent evolution. The approach permits the adaptation of the content of the generated web pages to the incoming information from the web. New information is extracted, annotated and coherently integrated in the body of knowledge in order to keep the holistic character of the body of knowledge. Personalization is achieved by filtering the semantic network according to the learner model, which keeps the list of concepts known or unknown by the learner. The approach was used in an EU INCO Copernicus project for computer aided language learning

1. Introduction

One of the main goals of any human learning process consists in the progressively construction of a body of declarative knowledge for the considered domain (a "model" of the domain) in the mind of the learner. This must also be, of course, one of the main goals of any Intelligent Tutoring System ("ITS") or learning environment.

What we consider very important, especially nowadays, in the context of the huge amount of information available on the web and of its permanent evolution, is the need of

[1] in S.Cerri, G.Gouarderes (eds.), Intelligent Tutoring Systems 2002, Springer, Lecture Notes in Computer Science number 2363, pp 259-269.

assuring a holistic character of the constructed body of knowledge. The learning process must induce the sense of the whole in the learner's mind. This simplifies the understanding of complex sets of concepts, being in consonance with the cognitive ergonomic rule of reducing the cognitive load [11].

A second way to facilitate learning is the usage of metaphors. Every professor, even in the most abstract domains, uses metaphors (consider, for example, the "trees" in mathematics and programming). Metaphors may appear every day and they are needed for a correct understanding of a concept. An important problem is that metaphors are not easy to tackle by a foreign speaker of a language [16] (e.g. "to sustain a loss" [16]). This problem of metaphors becomes more important in the context of the explosion of documents (and potential new metaphors).

Another important idea of the paper, related with the above ones, rose especially after the success of the web. It is a reality that everyday some new relevant information might appear on the web and must be included in the learning process. This new information must be coherently integrated in the learned body of knowledge in order to keep the its holistic character.

The skeleton of the above-mentioned knowledge body may be considered as a semantic network of the main concepts involved in that domain. These concepts are usually taxonomically organised, have several attributes and relations connecting them with other concepts. From a knowledge-based perspective, we might say that the learner must articulate in his mind the so-called ontology of the domain.

The word "ontology" is used in philosophy to denote the theory about what is considered to exist. Any system in philosophy starts from an ontology, that means from the identification of the concepts and relations considered as fundamental. In artificial intelligence, the same word is now often used as "a specification of a conceptualization.... an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents" [7].

The approach presented in this paper follows the previous ideas, combining agents that search for information, text mining (for metaphors) techniques, learner modeling, and personalized web page generation. The visible result is the fact that the semantic network of the concepts from the domain ontology is mapped on a network of personalised web pages automatically generated. Personalization is achieved by filtering the semantic network according to the learner model, which keeps the list of concepts known or unknown by the learner. The generated web pages contain relevant excerpts from documents continuously retrieved from the web.

The idea of assuring the inclusion in the learning process of the latest information available on the web is provided by searching relevant documents, extracting necessary knowledge and including it in the generated web pages. The domain ontology has a determinant role in all these three activities this central role providing also the needed holistic character.

The novelty of the approach presented in the paper resides on the dynamic generation of a complex structure of personalized web pages that reflects the domain ontology (having a holistic character). Another novel feature is the possibility of continuously

updating the content of the generated web pages, considering the incoming information from the web. Other approaches [2], [10] provide adaptability starting from local policies ("adaptive navigation support" – [2]), driven by learner's model and goals. They do not follow any holistic principles. They also do not process (by selection and annotation, as in our approach) new information and do not include it in a complex structure. They rather provide support for relevant web pages in a given context [3], [8]. Eventually, the usage of metaphors for enhancing understandability is also new.

Our approach was used for the development of an ITS distributed on the web (with three servers, in France, Romania, and Bulgaria) under the INCO Copernicus project LarFLaST ("Learning Foreign Language Scientific Terminology" – see http://www-it.fmi.uni-sofia.bg/larflast/). This project had as main objectives to provide a set of tools available on the World Wide Web for supporting Romanian, Bulgarian and Ukrainian people to learn foreign (English) terminology in finance.

The next section presents our approach for extracting new, relevant information from the web. Gathered information is further edited for semantic (knowledge) annotation. Eventually, annotated documents are used to extract excerpts to be included in the generated web pages. The third section exemplifies the usage of this framework in the LarFLaST project. The fourth section presents the dynamic generation of ontology-centred, personalized web pages that include information obtained with the framework discussed in the previous two sections.

2. Intelligent search, annotation, and usage of information from the web

The World Wide Web (WWW or the "web") is a huge hypermedia on Internet, browsable with very simple, direct manipulation interfaces. Its explosive growth in only several years is the best prove of its usefulness. Two of the causes of this phenomenon are, probably, the ease of "publishing", of communicating something through text and/or images on the web. From the other direction of the communication process, it is very easy for everybody to explore the network of web pages. As a consequence, we notice the extremely dynamic character of information nowadays, the availability of the Web today having definitely changed the information scenario. What happens is that the time between the appearance of new information in some domain and the use of this information by people has extremely shortened comparatively with some years ago. Therefore, information may become obsolete very quickly or be replaced by some other Information. A good tutoring system should consider this scenario, and consequently be able to update its Information continuously.

The process of extracting and using the most relevant information from the web involves three phases: information acquisition performed by searching the web, knowledge identification by semantic editing and the usage of this knowledge, as in the below figure.

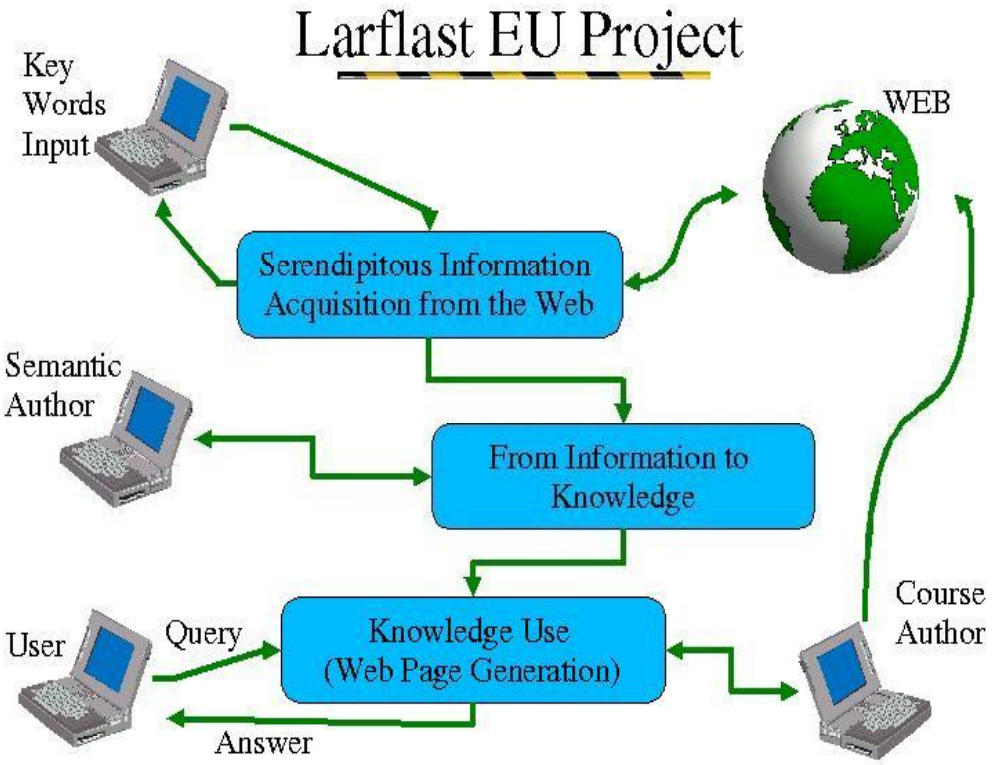

Fig. 1. Information acquisition, annotation and usage in Larflast

The domain ontology plays important roles in each of the three mentioned phases. The keywords used by a web spidering agent [4] for the search of relevant documents are obtained from the domain ontology. The process controlled by this agent searches on the web by means of activating a number of search engines. During this phase, data mining techniques may be applied in order to better select automatically the fit between the requested information and the retrieved one.

The set of raw Web pages is stored in a DBMS of XML (XHTML) pages (generated by transforming HTML pages with JTIDY) in order to facilitate the crucial activity of knowledge construction by the "semantic author" [18]. This semantic author is either a human agent (e.g. the professor) that edits the documents by means of a special, "semantic editor" or a natural language processing program that transforms the XML documents.

We have built two "semantic editors". The first one may be used remotely (accessed through a Web browser). The second semantic editor (which will be presented in the next section) was built for a specific purpose (metaphor identification and annotation) and is used only offline.

The authored XML documents are stored in a XML database, to be subsequently used by the web page generator or by the "course author" for composing the Course. By applying XSL files to XML authored files, the "semantic author" may visualise the appearance of each knowledge unit in order to check the interface with the learner.

The XML web-based editor is generic with respect to any chosen XML mark-up choice. For these reasons, the same editor may be used both by the "semantic author" and the "course author" (described hereafter).

Once the knowledge units have been made available on the XML DBMS, the course author (by using the same XML Web editor) may compose the course at his-her choice. At the end, the course will be published and delivered on the Web. Another possibility, described in the next section, is the dynamic generation of web pages.

3. Metaphor identification, annotation, and usage as aid for learning financial terminology

An instance of the process discussed in the previous section is used for the identification, annotation, and usage of metaphors for aiding learning foreign finance terminology [14]. This approach was implemented in the LarFLaST project.

Metaphors are often used to give insight in what a concept means, like in the following example: "Stocks are very sensitive creatures" (NYSE, New York Stock Exchange web page, http://www.nyse.com/). Such insight cannot be obtained in knowledge-based approaches centred on taxonomic ontologies. For example, these systems will explain the concept "stock" in terms of its super-concepts like "securities", "capital", "asset" or "possession". Its attributes and relations with other concepts may provide more details.

Metaphor processing was used for aiding language learning in the LarFLaST project and it involved the three phases discussed in the previous section:

- gathering relevant texts from the web,
- identification (acquisition) of metaphors in the selected texts and their XML mark-up of the identified metaphors,
- personalised usage of the metaphors.

The following picture illustrates the whole process:

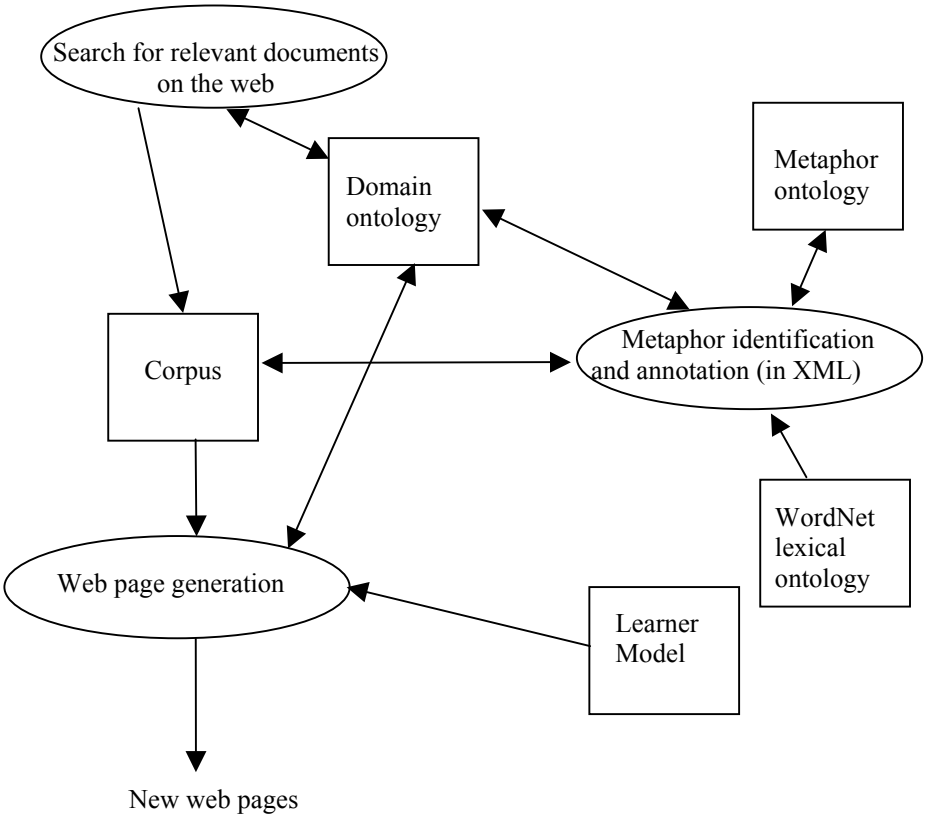

Fig. 2. Metaphor identification and web page generation

The DBMS (corpus) of relevant texts is the raw material for the second phase, metaphor identification and annotation. This phase may be considered similar to a knowledge acquisition process, in which XML semantic annotations are added in the texts. A specialized acquisition tool, written in Java, supports this semantic editing (the second type of semantic editor mentioned in the previous section). The interface for this tool is presented in the next figure.

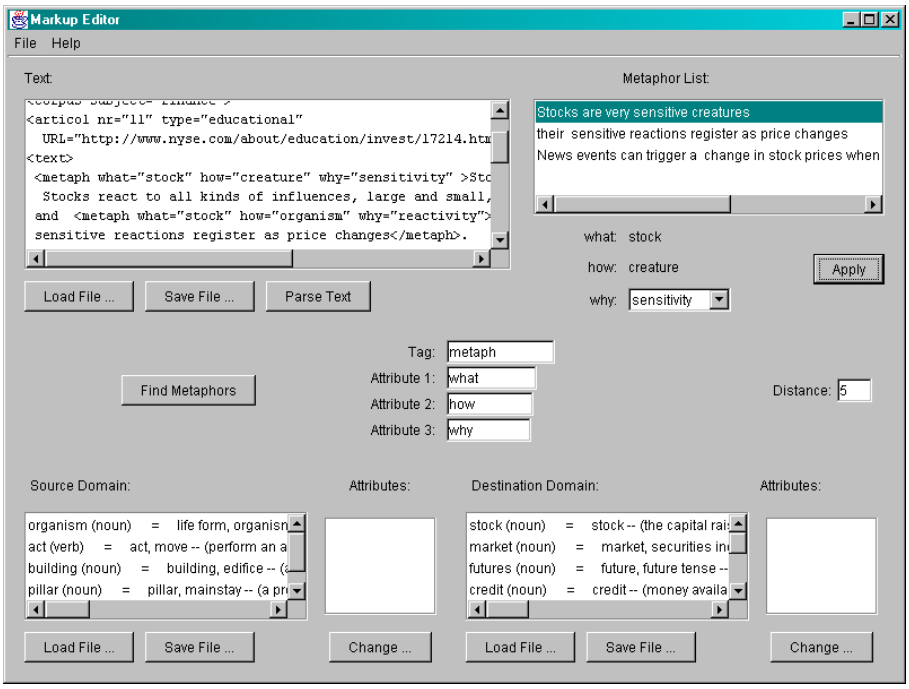

Fig. 3. Semantic metaphor editing

Metaphor identification in the corpus is performed by searching for occurrences of pairs of concepts from source (metaphor) and destination (domain) ontologies. These concepts are loaded from XML files and are displayed in the two bottom text areas in the figure. The lexical ontology WordNet (http://www.cogsci.princeton.edu/~wn) is used for extending the search to related (synonyms, hyperonyms, hyponyms, meronyms) concepts. Identified metaphors are listed in the top right text area. If the user approves, the selected proposed metaphor is annotated in the corpus.

The metaphor-annotated corpus may be further used for several purposes. In LarFLaST, the corpus is used in the dynamically generated Web pages for metaphor examples and explanations tailored to a learner model, in a given context (see next section). For this purpose, XSL descriptions are used for the personalized web pages.

A concept ontology editor is used for editing concept ontologies, and for controlling the process of metaphor search by stating which relations from WordNet are considered. This editor allows the user to add, remove or modify a concept. For a specific concept you can set its part of speech, its WordNet sense number, its attributes and the related concepts, which will be considered for metaphor identification. The ontologies may be saved in XML files, to be used as source and domain ontologies for finding new metaphors (see the above figure).

The next figure illustrates the concept ontology editor:

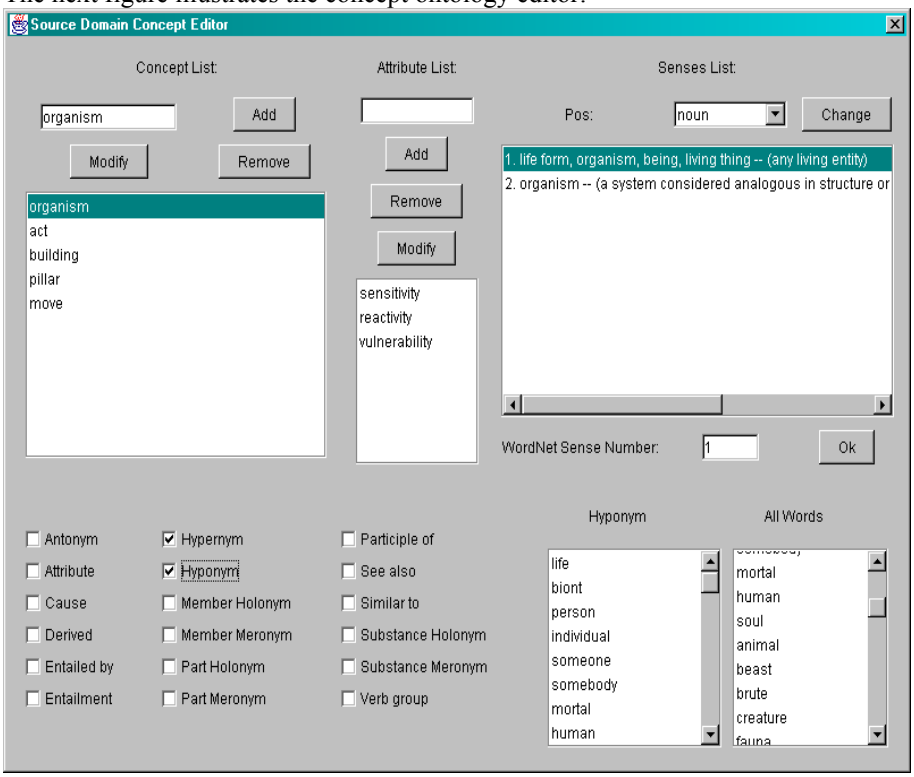

Fig. 4. Concept editor

4. Ontology-centered presentation on the web

An intelligent tutoring system is a knowledge-based, interactive computer program that can be used by a learner as a personal, tireless teacher, which adapts to the learner's cognitive particularities and his/her individual progress. A major emphasis is paid in ITS to the choice of domain and learner models.

The domain model (the knowledge base or ontology) is one of the most important parts in such applications. In most ITS, the domain model is build once for all, at define time, before the conversations with the learners will occur. Usually, the learner model is developed in correspondence with the domain model. By consequence, the structure of the learner model is as well decided once for all, at define time.

In our approach we challenge this view, and attempt to transform the construction and the delivery of knowledge into dynamic processes, continuously updated by incoming Information, on the domain (from the Web) and on the learner. The central issue remains to be able to tune the generation of domain knowledge according to what can be derived from the learner's model as well as to update permanently this model as a result of learner testing.

In LarFLaST, the learner model ("LM") includes correct, erroneous and incomplete learner's beliefs (about which he is or not aware), misunderstandings and misconceptions about concepts [5]. From a knowledge-based perspective, LM includes what knowledge has the learner, what knowledge misses or has been wrongly acquired: LM may be inferred starting from the analysis of the results at tests or from other data, as the path followed by the learner during browsing [5].

WWW is a perfect place for learning. The synergetic integration of ITS with the Web might provide a learning environment able to change totally the way we learn. This idea is supported by the fact that hypertext was introduced by Douglas Engelbart, in the early sixties, as a "Conceptual Framework for Augmenting Human Intellect" [6]. Moreover, Theodor Nelson, who coined the term "hypertext", defined it as the hyperspace of concepts from a given text (Nelson 1995) or "a system for massively parallel creative work and study … to the betterment of human understanding" [9].

A natural consequence of the above idea is the usage of the web "hyperspace of concepts" for facilitating the conceptualisation and understanding in learning processes. For assuring a best conceptualisation, the conceptual map of the considered domain (the ontology) should be filtered according to the learner model. After the filtering, the concepts considered as relevant and the relations among them are mapped in a network of web pages. Each concept is mapped to a Web page and each relation to a link, according to explicit (e.g. "is-a", "part-of", "agent", "instrument" etc.) or implicit (e.g. "similar") relations [12]. The result is that the ontology, the network of generated web pages, and the conceptual map to be induced in the mind of the learner have the same (semantic network like) structure. As mentioned in the introduction, this mapping assures the holistic character of the knowledge body in the mind of the learner.

This idea was used in the GenWeb system, for the dynamic generation of highly structured web pages for learning functional programming [12] and in the INCO Copernicus project LarFLaST. The web page generator is written in the Lisp-based knowledge-based framework XRL [1]. Personalization is obtained by dynamically tailoring the content of the Web pages according to each learner's model [12], [13]. That means, for example, that explanations refer only to known concepts, while the structure of the generated collection of Web pages is centred on unknown concepts.

Information about metaphors (the third phase, the "usage" of metaphors from the above section) is added in the generated pages. This information is personalized through the selection only of the relevant metaphors accordingly to the learner model. The attributes of the XML metaphor annotation are used for this purpose.

Three examples of dynamically generated Web pages are given in the fig.5. They are personalized accordingly to the learner model [5], from which a fragment is shown below:

```
know(john,money_market,[[b_def,2025,80],[b_def,20,80]],u_1_d_1,1,none,4).

not_know(john,investment,[[b_def,2006,80]],u_1_d_1,1,none,5).

not_know(john,investment,[[b_def,2006,80]],u_1_d_1,2,none,6).

know_wrongly(john,secondary_market,[[b_def,2004,80]],u_1_d_1,3,none,7).

know_wrongly(john,financial_market,[[b_def,2001,80]],u_1_d_1,2,none,14).

not_know(john,open_market,[[b_def,2016,80]],u_1_d_1,1,none,15).

not_know(john,open_market,[[b_def,2016,80]],u_1_d_1,2,none,16).
```

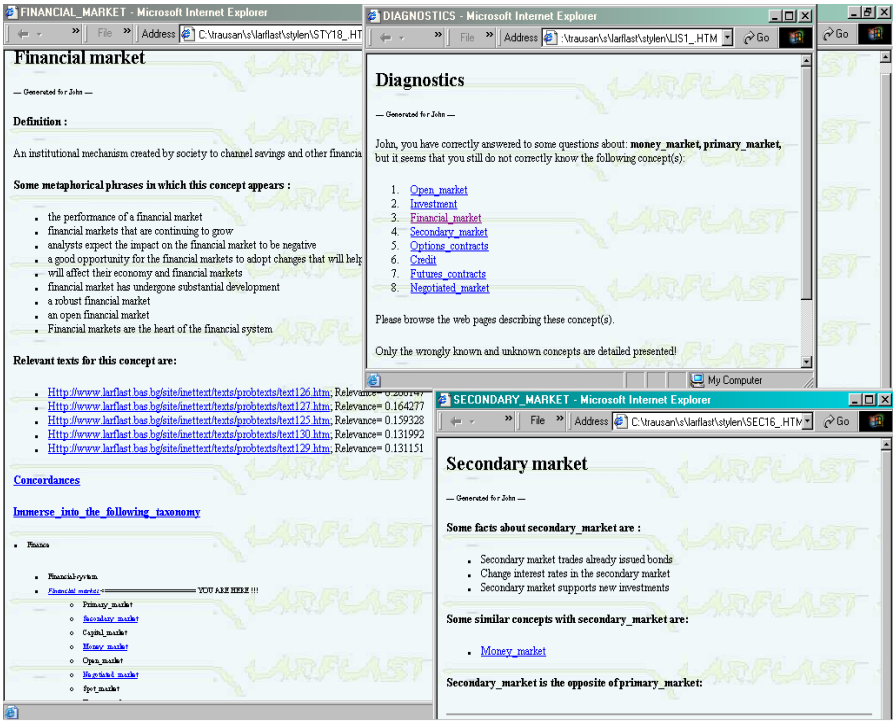

Fig. 5. A screenshot with dynamically generated, personalized web pages

The left (biggest) window in the figure contains some metaphorical phrases extracted with the framework presented in the second section. This window contains, in the bottom, the structure (the taxonomy) of domain concepts. Links are generated only to the concepts unknown or wrongly known by the learner (listed in the upper right window). A web page from the taxonomy is illustrated in the third, bottom-right window.

5. Conclusions

The approach presented in this paper is centred on the domain ontology. This ontology is used as a start point in the serendipitous search. The same ontology is used for the XML semantic annotation of the retrieved documents. The ontology is used for the retrieval of relevant metaphors from the XML annotated documents. The structure of the dynamically generated web pages reflects the domain ontology. In addition, the ontology is driving the construction of the learner's model and the filtering of the amount of concepts and facts presented. This omnipresence of the ontology induces the holistic character discussed in the introduction, assuring the coherence of the presentation, which has direct effects on the learning process.

The good effects of the structuring of the generated web pages according to the ontology were emphasized in a study performed with students in the end of the LarFLaST project. They remarked (not being asked explicitly about it) that the taxonomic organisation is very helpful.

Other systems for learning on the web usually have a static domain model from which they construct (not dynamically) web pages. Even if they have some dynamic characteristics, like adaptive hypermedia [3], [17] or planning the content of the presented material [10], [15], they miss a holistic character. Adaptive hypermedia is obtained, for example, by local policies ("adaptive navigation support" – [3]) like flexible link sorting, hiding or disabling or by conditionally showing text fragments etc. [3]. Planning the content of the generated web pages is also not concerned with a global, holistic approach, but more with local decisions based on the learner model.

The permanent inclusion of new information gathered and annotated from the web is another novel feature, not included in other systems. Existing approaches only provide intelligent recommend interesting web pages, according to the user profile [3], [8]. They do not permit the inclusion of relevant facts in the structure of ontology-centred structure.

Eventually, the usage of metaphors is a novel approach. It has a lot of implications discussed in detail in [14]

References

1. Barbuceanu, M. and Trausan-Matu, S. (1987) Integrating Declarative Knowledge Programming Styles and Tools in a Structured Object Environment, in J. Mc.Dermott (ed.) Proceedings of 10-th International Joint Conference on Artificial Intelligence IJCAI'87, Italia, Morgan Kaufmann Publishers, Inc.
2. deBra, P., Brusilovsky, Houser, G.J., Adaptive Hypermedia: From Systems to Framework, ACM Computing Surveys, 31(4), 1999.
3. Breese, J.,Heckerman, D., Kadie, C., Empirical Analysis of Predictive Algorithms for Collaborative Filtering, Procs. Of 14th Conf. On Uncertainity in AI, Morgan Kaufmann, 1998
4. Cerri, S.A., Loia, V., Maffioletti, S., Fontanesi, P., and Bettinelli, A. (1999) Serendipitous acquisition of Web knowledge by Agents in the context of Human Learning. In: Proceedings of

THAI-ETIS : European Symposium on Telematics, Hypermedia and Artificial Intelligence, Varese, Italy.
5. Dimitrova, V., Self, J., Brna, P., 'Maintaining a Joinly Constrcted Student Model', in S.A.Cerri (ed.), Artificial Intelligence, Methodology, Systems, Applications 2000, Springer-Verlag, ISBN 3-540-41044-9, pp.221-231.
6. Engelbart, D.C. (1995) Toward Augmenting the Human Intellect and Boosting our Collective IQ, Communications of the ACM, vol.38, no. 8, pp. 30-33.
7. Gruber, T., What is an Ontology, http://www-ksl.stanford.edu/kst/what-is-an-ontology.html
8. Lieberman, H., Letizia: An Agent That Assists Web Browsing, International Joint Conference on Artificial Intelligence, Montreal, August 1995.
9. Nelson, T.H. (1995), 'The Heart of Connection: Hypermedia Unified by Transclusion', Communications of the ACM, vol.38, no. 8, pp. 31-33.
10. Siekmann, J., Benzmuller, C, and all, Adaptive Course Generation and Presentation, Proceedings of the International Workshop on Adaptive and Intelligent Web-based Educational Systems, Montreal, 2000
11. Thuering, M., Hannemann, J., Haake, J.M., Hypermedia and Cognition: Designing for Comprehension, Communications of the ACM, vol.38, no. 8, pp. 57-66, aug. 1995.
12. Trausan-Matu, St. (1997) 'Knowledge-Based, Automatic Generation of Educational Web Pages', in Proceedings of Internet as a Vehicle for Teaching Workshop, Ilieni, June 1997, pp.141-148, See also http://rilw.emp.paed.uni-muenchen.de/99/papers/Trausan.html
13. Trausan-Matu, St. (1999) 'Web Page Generation Facilitating Conceptualization and Immersion for Learning Finance Terminology', in Proceedings of RILW99. See also http://rilw.emp.paed.uni-muenchen.de/99/papers/Trausan.html
14. Trausan-Matu, St. (2000) 'Metaphor Processing for Learning Terminology on the Web', in S.A.Cerri (ed.), Artificial Intelligence, Methodology, Systems, Applications 2000, Springer-Verlag, ISBN 3-540-41044-9, pp.232-241.
15. Vassilieva, J., http://julita.usask.ca/homepage/AIED'97.ps
16. Vitanova, I. (1999) English for Finance. Understanding Money and Markets, Research report, Larflast project, http://www-it.fmi.uni-sofia.bg/larflast/
17. Weber, G., Specht, M., User Modeling and Adaptive Navigation Support, in WWW-based Tutoring Systems, http://www.psychologie.uni-trier.de:8000/projects/ELM/Papers/UM97-WEBER.html
18. Wittman, P., Evolution de l'activite editoriale face a un nouvel ordinateur: le Web, Eng. These, Univ. Montpellier, France, 2002