Wiki course builder: A system for retrieving and sequencing didactic materials from Wikipedia

Conference Paper · June 2015

DOI: 10.1109/ITHET.2015.7218041

3 authors:

Carla Limongelli
Università Degli Studi Roma Tre
98 PUBLICATIONS **861** CITATIONS
SEE PROFILE

Fabio Gasparetti
Università Degli Studi Roma Tre
78 PUBLICATIONS **1,146** CITATIONS
SEE PROFILE

Filippo Sciarrone
Università Degli Studi Roma Tre
93 PUBLICATIONS **1,292** CITATIONS
SEE PROFILE

Wiki Course Builder: a System for Retrieving and Sequencing Didactic Materials from *Wikipedia*

Carla Limongelli, Fabio Gasparetti and Filippo Sciarrone
Department of Engineering
Roma Tre University
Via della Vasca Navale 79, Rome, Italy
Email: {limongel, gaspare, sciarro}@dia.uniroma3.it

***Abstract*—The designing and delivering of a new online course is a crucial task for teachers that have to face two main problems: building, or retrieving, and sequencing learning materials. Retrieving learning materials requires a great effort and a waste of time, while sequencing them requires an accurate didactic project. On the other hand, thanks to the Internet, teachers and instructional designers today can search and retrieve learning materials from Learning Objects Repositories freely available on the Web, such as *Mertlot* or *Ariadne*. In this paper we investigate the possibility of using the Wikipedia free encyclopedia, that is the biggest repository of educational material which is visited daily by about sixty million people with its 49 millions of registered people. It is a matter of facts that teachers consult this encyclopedia to arrange, integrate or enrich their courses. So here we propose a system, now at its early stage of development, aiming at supporting teachers to build courses basing on Wikipedia only. The system retrieves learning materials form Wikipedia and sequences them on the basis of the links embedded in the Wikipedia HTML pages, following a course building process based on the Grasha teaching styles and on a social didactic approach. A first questionnaire has been submitted to a sample of teachers with encouraging results.**

I. Introduction

Building a new course is a hard and expensive task for teachers that have to cope with two main issues: firstly retrieving, or creating, learning materials compliant to some given learning goals, and secondly delivering them to learners in a correct learning sequence. The first issue requires a big effort, in terms of both design time and searching activities. In fact this step spans from the designing of the concept map to the building of new learning materials through software didactic tools like *Microsoft Powerpoint*[1] *Articulate*[2] or *Ispring*[3] and many others. On the other hand, Internet with its huge amount of information acts as a big repository where people can find and retrieve documents in several formats such as html, plain text, pdf, flash, etc. etc. . and teachers can search for didactic materials that suit their own didactic needs in several Learning Object Repositories (LORs), such as Merlot[4] or Ariadne[5], freely available on the Web. Intelligent crawlers can also autonomously retrieve additional resources by navigating the web [1]. Users can login into these systems and input their query through a Graphic User Interface (GUI): the system returns a ranked list of links to learning objects allowing for their download. Finally the user can use the learning object as is or after having modified it. Moreover, it is a matter of fact that most of the teachers, about 87%,[6] use the online *Wikipedia*[7] encyclopedia in their didactic activities. The reliability of Wikipedia (primarily of the English-language edition), has been also assessed: an early study in the *Nature* journal said that in 2005, Wikipedia's scientific articles came close to the level of accuracy of the *Encyclopedia Britannica* [2].

The second issue, i.e., the delivery phase, is another complex task of the course building process: the selected materials are to be delivered in a correct learning sequence. This step requires the design of semantic relationships between learning materials, which can be represented by a directed graph where edges represent relevant didactic relationships, i.e., prerequisite relationships, while nodes represent the learning materials themselves.

Here we propose a system at its early stage of development, called *Wiki Course Builder*, with the aim to help teachers to build a new course starting from Wikipedia HTML pages. The system retrieves and sequence HTML pages on the basis of their out-coming embedded links. We address the above-mentioned two critical issues for supporting teachers to retrieve and sequence new learning materials with a little effort. We address the retrieving problem making use of the *Wikipedia* repository as the repository used by our system. Wikipedia is an online encyclopedia, collaborative, multilingual and free, born in January 2001 and supported by the *Wikimedia Foundation Inc.*, a no-profit US foundation. This repository forms a huge semantic graph of relevant content. We chose it as the system repository for its popularity and usefulness among students, instructional designers. A lot of people regularly use it as the starting point for knowledge building and for identifying general didactic goals. In our system, through a suitable GUI, first the user can input one or more keywords concerning the topic he/she is working on. Secondly, the system analyzes the most relevant HTML pages, returned by its embedded search engine, together with a proposal of a first sequencing of them. The concept of relevance of a Wikipedia page is related to a set of teaching styles it is tagged with, following the Grasha model [3]. The sequencing is performed basing on

[1] http://www.microsoft.com
[2] https://www.articulate.com/
[3] https://www.www.ispringsolutions.com
[4] www.merlot.org
[5] www.ariadne.org
[6] www.pewinternet.org/2013/02/28/how-teachers-are-using-technology-at-home-and-in-their-classrooms/
[7] www.wikipedia.org

the links among the retrieved Wikipedia pages, as stored in the Wikipedia repository, and on the Grasha teaching styles. Each time a teacher selects a retrieved page, it is tagged with the her teaching styles, represented by a weighted vector. As time goes by, this form of knowledge is acquired and exploited so that the repositorys content is automatically filtered. In other words, a social filtering on Wikipedia is put in practice by clustering the documents depending on the teachers way of learning. The well-known cold start problem of collaborative approaches is partially overcome by exploiting the first visits of the resources as soon as they become available. This is a straightforward and lightweight approach, well suited for Communities of Practice (CoPs), where multiple users access and filter documents according to their preferences [4], [5], [6], [7], [8].

A further problem that is addressed by our approach is the quality of the content stored in the repository. Our hypothesis is that the selection performed by the community of users can assist future users by ignoring less relevant materials. This proposal is currently in a first development phase that will consider real scenarios where users will explicitly suggest the precision of the filtering process.

The remainder of the article is structured as follows. Section II draws some important related work; Section III shows the system, i.e., the Wiki Course Builder system. In Section IV a first evaluation of the system is presented. Finally in Section V some conclusions are drawn.

II. Related Work

To our knowledge, there is no much literature that considers the possibility to retrieve and sequence Wikipedia pages to build a new course while there is some literature concerning the retrieval of didactic material and the teacher model like the Grasha one. Gasparetti *et al.* [9] propose an early attempt to exploit the Wikipedia content in order to determine prerequisite relationships among learning objects. For instance, intelligent crawlers can autonomously retrieve additional resources by navigating the web [1]. One of the first works that have investigated Wikipedia as a learning support is [10], however from a different prospective. In particular, in this work is discussed the following research question: publishing on Wikipedia encourages students towards a collaborative and involving learning. Then, [11] highlights the didactic potential of wikis that actively involve learners in their own construction of knowledge, on the basis of a collaborative approach. The nearest approaches to our work are presented in [12], [13] and [14]. They extract semantic relationships among Wikipedia pages in order to link them automatically. Our approach is complementary to this one: starting from links among Wikipedia pages we understand the semantic didactic relations among them. The use of a the Grasha model can be found in [15], where the teacher model is represented by a teaching experience and a dynamic semantic network composed by the retrieved and used learning materials by a community of teachers. In [16] a clustering of teachers is proposed for a community of teachers. Other methods to build courses and training by using a modeling technique can be found in [1], [17], [18], [19], [20], [21].

III. The Architecture of the System

In this Section we show the overall architecture of the system. It has been developed as a classic 3-tier web application. As already said in Section I, the system is currently at its very early stage of development and not all its modules are completely running and ready to be used. Here we present the part of the system shown in Fig. 1, focussing on those modules concerning the retrieving and sequencing processes only. In its final form, the system will be composed of the following modules:

- *Home page*: it is the page showing all the functionalities and presenting the web site;
- *Community*: this module manages the CoP behind the use of the system. This module will allow teachers for sharing learning materials, i.e., those Wikipedia pages used by the community: besides, through this module the teacher can insert her teaching styles according to [3]. These teaching styles will be used by the search engine;
- *My Courses*: this module allows for the courses management. The user can revise her courses, changing the sequences, export the courses, etc. etc.
- *Build Course*: it is the module that we show in this work. The teacher can insert her concept-term(s), i.e., the term(s) representing the concepts that she wants to teach and launch the search engine on the Wikipedia database that will return a sequence of Wikipedia pages, compliant with her teaching styles.

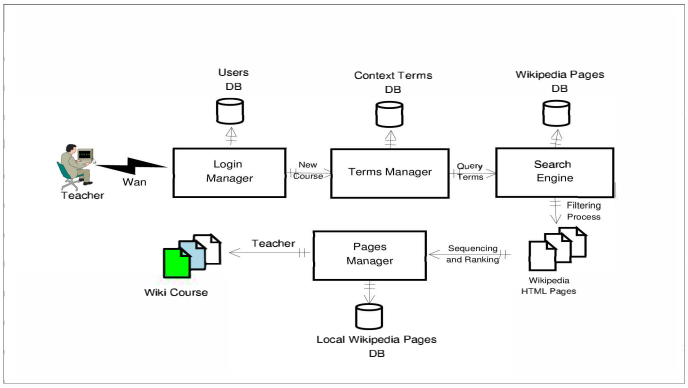

Fig. 1. The functional architecture of the system.

A. The Login Module

This module manages the security of the accesses to the system. In Fig. 2 a screen-shot of this module is shown. The user can register and join the community of teachers. A database of users is built and managed by this module. When the user registers to the system, she is required to fill in a form with some personal data. It is in this step that the user is required to take the Grasha-Riechmann Teaching Style Survey[8]. It consists of 40 5-points Likert-scale questions, such as: *Sharing my knowledge and expertise with students is very important to me* and *I give students negative feedback when*

[8]available at: http://longleaf.net/teachingstyle.html

their performance is unsatisfactory. The questions aim at modeling the teacher by means of the following five dimensions:

- *Expert*: the teacher has the knowledge and the experience that students need;
- *Formal authority*: the teacher maintains her/his institutional role;
- *Personal model*: the teacher bases her/his teaching by personal example and establishes a model for thinking and acting;
- *Facilitator*: the teacher emphasizes personal interactions between students and teacher;
- *Delegator*: the teacher develops student's ability so that they can act autonomously.

where each dimension is measured by an integer number $d \in [1, 7]$. In this way, when the user login into the system, she is modeled by the set of the aforesaid five dimensions. Subsequently, this data will be used by the search engine.

Fig. 2. The Login Module.

B. The Terms Manager

The goal of this module, embedded in the *Build Course* module, is to help teachers to launch the query to submit to Wikipedia. This process is performed by means of the following steps, as shown in Fig. 3:

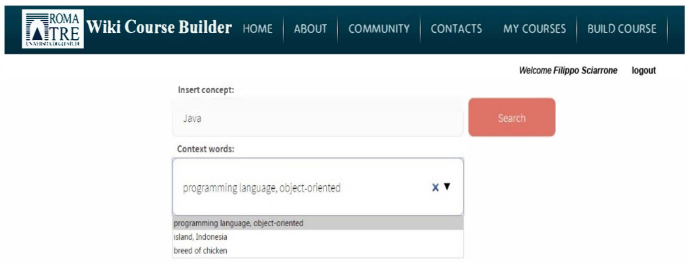

Fig. 3. The terms manager.

- The teacher inserts the concept-terms, i.e., those terms related to the topic to teach in the new course;
- The system proposes other sets of related terms to complete the query, if already exist in the context terms database, otherwise these terms could be manually added by the teacher herself, if the context terms database does not contain at that time related terms as is in the first uses of recommending systems for the *cold start problem* [22]. This mechanism is based on the contribution given by all the queries entered previously by all the teachers of the CoP. In Fig. 3 we show an example of the module at work: the teacher enters the term *java* and the system proposes, in a separate combo-box, the context terms sets: *programming language, object oriented*, *island,indonesia* and *breed of chicken*. The teacher can select the right set of terms that will complete the query. We propose this solution to other classic query expansion mechanisms for its simplicity and to not overload the complexity of the system. On the other hand this task is not a complex and time consuming task for the teacher;
- The module stores in the local database the terms entered by the user together with the added terms;
- The query is passed to the Search Engine module.

A particular aspect of this mechanism is its social aspect: the CoP of teachers contribute to the query expansion process by means of those terms already entered by all members. In this way, a sort of a social semantic network of terms and concepts grows with the use of the system, allowing for a strong characterization of the community of practice, i.e., the CoP.

C. The Search Engine

This module retrieves and ranks the Wikipedia pages from the Wikipedia database. The pages selection process starts with the query formed by the Terms Manager Module which is given in input to the Wikipedia search engine. Each Wikipedia HTML page returned by Wikipedia is filtered both by a content-based and a social-based filtering mechanism, as shown in Fig. 4:

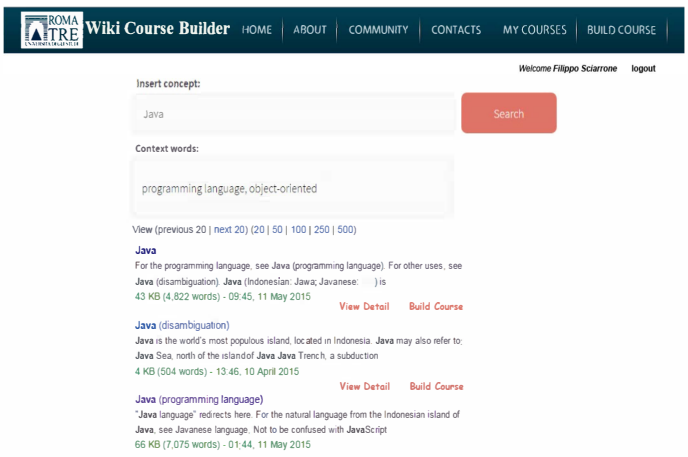

Fig. 4. An example of the Wikipedia pages retrieved and ranked by the search engine.

- *Content-based filtering process*: the retrieved pages are filtered by means of the cosine similarity between the query formed by the Terms Manager module and

the HTML pages, with a TFxIDF terms-weighting technique, together with the use of the vector model, a classic technique used in the information retrieval area to rank documents (see for example [23]). The system will use this technique, which is not based on didactic features, only at the time t_0, that is when the retrieved pages are all *cold items*

- *Social-based filtering process*: the retrieved pages are filtered using the teaching styles of the teacher who launched the query, that is a teacher-style based metric is used to rank the retrieved pages. This mechanism is based on the fact that every time the teacher uses a retrieved page in her course, this page is tagged with the teaching style of the teacher that selected it. A similar approach has been undertaken in the web domain by Biancalana *et al.* [24], [25], [26]. In this way, by the use of the system, each used Wikipedia page (i.e., a link to it) will be stored in a local database together with a set of 5-ples, one for each teacher that used it, with its user-occurrency that is if the teacher used that page n-times, that page will be tagged n-times with the user teaching styles. The metric used to perform the distance D between the user teaching styles TS_k^u and a generic document TS_i^d is based on the euclidian distance:

$$D_{u,d} = \sqrt{\sum_{i=1}^{5}(TS_i^u - TS_i^d)^2} \qquad (1)$$

The user can benefit of the choices already made by other members of the community on the same documents, strengthening the social aspects of the page selection process.

D. The Pages Manager

This module proposes the retrieved and ranked pages to the user in a suitable graphic way, as shown in Fig. 4. The user has at her disposition an interactive *Google-like* environment where, for each retrieved page, she can see the complete page (*View Details* function) or launch the course building process (i.e., the *Build Course* function). An example of the use of the first function is shown in Fig. 5 where a Wikipedia page is analyzed. The second function, i.e., the *Build Course* function, starts the process of course building from the root page that is performed through the following steps:

1) The Wikipedia root page is parsed with the use of suitable Wikipedia API [9]. All the linked pages are evaluated and the most promising linked page is selected on the basis of its reputation in the community (see the previous subsection). This process starts again with a predefined tree depth;
2) All the sequenced retrieved pages are displayed as shown in Fig. 6: the user can delete a page and change the proposed sequencing order;
3) Finally the user can save her course.

[9] en.wikipedia.org/w/api.php

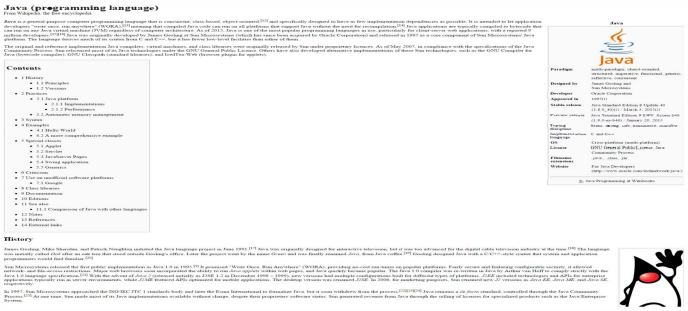

Fig. 5. An example of the *View Details* function.

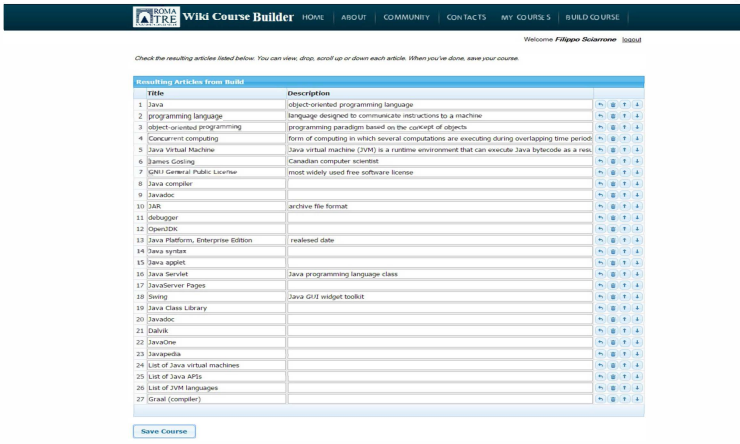

Fig. 6. Simulation Results.

IV. A First Evaluation of the System

In this Section we show a first evaluation of the system. We submitted a happy sheet questionnaire, composed of 5 questions, to a sample composed by 15 teachers of a technical high school with the aim to test a first feeling of teachers with respect to the system. The sample was required to read the goals of the system and to evaluate its interface, i.e., the Graphic User Interface. As we mentioned in Section III, at this moment the system is in its early stage of development and the sample could mainly evaluate the rationale behind the system.

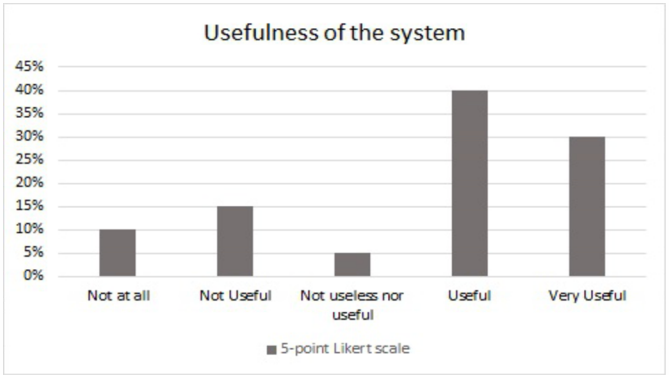

Fig. 7. The answers to the usefulness of the system.

A question was about the usefulness of the system: *Do you feel useful the system?* while another question was: *How*

do you feel the GUI of the system?. The answers to the first question are shown in Fig. 7. 70% of the sample considered the system useful and very useful. In Fig. 8, are summarized the answers concerning the assessment of the system GUI. Finally, the 80% of the sample judged such interface as simple-very simple to use.

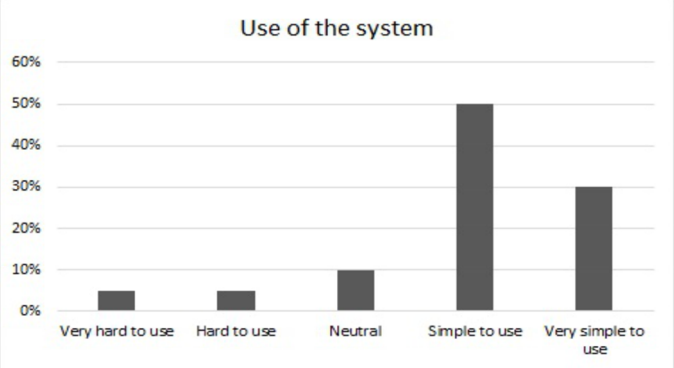

Fig. 8. A first evaluation of the system GUI.

V. Conclusions and Future Work

In this paper we presented *Wiki Course Builder*, a system capable to help teachers build and manage courses composed of *Wikipedia* HTML pages. The teacher is required first to insert the concept - term and second to insert some terms in order to help the search engine disambiguate the Wikipedia pages. Subsequently the system returns some Wikipedia pages sequenced by means of some simple metrics, based on the social reputation of all the pages linked to a starting page. The teacher can accept the system results or change them both in terms of HTML pages and in terms of a different sequencing. All the used pages are stored in a local database where they are tagged with the teaching styles of the teacher that used them. In this way, a Community of Practice can grow and users, i.e., teachers, can benefit of the each other work. Currently the system is at its very early stage of development, as a 3-tier web architecture developed in java language and using the *wikipedia-miner* toolkit to interface itself to the Wikipedia database. As a future work we plan to implement all the parts of the system and in particular all those social-based modules to improve the growth of the CoP. Moreover, teacher interests can be modeled by the analysis of the navigated HTML pages [27], [28]. Profiles of those interests can be combined with the teachers' queries in order to improve the filtering process.

Acknowledgment

The authors would like to thank Alessandra Milita and Andrea Tarantini for their significant contribution in the development of the system.

References

[1] A. Micarelli and F. Gasparetti, "Adaptive focused crawling," in *The Adaptive Web*, ser. Lecture Notes in Computer Science, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Berlin, Heidelberg: Springer-Verlag, 2007, vol. 4321, pp. 231–262.

[2] J. Giles, "Special report internet encyclopaedias go head to head," *Nature*, vol. 438, pp. 900–901, 2005.

[3] A. Grasha, "Teaching with style: The integration of teaching and learning styles in the classroom," *Teaching Excellence*, vol. 7, pp. 31–34, 1996.

[4] E. Wenger, "Communities of practice: Learning as a social system," *Systems thinker*, vol. 9, no. 5, pp. 2–3, 1998.

[5] A. Sterbini and M. Temperini, "Selection and sequencing constraints for personalized courses," in *Proc. IEEE Frontiers in Education, FIE*, 2010, pp. T2C1–T2C6.

[6] C. Limongelli, G. Mosiello, S. Panzieri, and F. Sciarrone, "Virtual industrial training: Joining innovative interfaces with plant modeling." in *ITHET*. IEEE, 2012, pp. 1–6.

[7] C. Limongelli, F. Sciarrone, M. Temperini, and G. Vaste, "Lecomps5: A web-based learning system for course personalization and adaptation," in *Proceedings of IADIS 2008, Amsterdam, The Netherlands, July 22-25, 2008. Proceedings*, 2008, pp. 325–332.

[8] M. De Marsico, A. Sterbini, and M. Temperini, "A strategy to join adaptive and reputation-based social-collaborative e-learning, through the zone of proximal development," *International Journal of Distance Education Technologies (IJDET)*, vol. 11, no. 3, pp. 674–681, 2013.

[9] F. Gasparetti, C. Limongelli, and F. Sciarrone, "Exploiting wikipedia for discovering prerequisite relationships among learning objects," in *Proceedings of the 14th International Conference on Information Technology Based Higher Education and Training, ITHET 2015*, 2015.

[10] A. Forte and A. Bruckman, "From wikipedia to the classroom: Exploring online publication and learning," in *Proceedings of the 7th International Conference on Learning Sciences*, ser. ICLS '06. International Society of the Learning Sciences, 2006, pp. 182–188.

[11] K. R. Parker and J. T. Chao, "Wiki as a teaching tool," *Interdisciplinary Journal of Knowledge and Learning Objects*, vol. 3, pp. 57–72, 2007.

[12] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, ser. IJCAI'07. Morgan Kaufmann Publishers Inc., 2007, pp. 1606–1611.

[13] D. N. Milne and I. H. Witten, "Learning to link with wikipedia," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008*, 2008, pp. 509–518.

[14] F. Gasparetti, C. Limongelli, and F. Sciarrone, "A content-based approach for supporting teachers in discovering dependency relationships between instructional units in distance learning environments," in *Proceedinges of the 17th International Conference on Human-Computer Interaction HCI 2015, Los Angeles, CA, USA 2-7 August 2015*, 2015.

[15] C. Limongelli, F. Sciarrone, and M. Temperini, "A social network-based teacher model to support course construction," *Computers in Human Behavior (in press)*, 2015.

[16] C. Limongelli, M. Lombardi, A. Marani, and F. Sciarrone, "A teaching-style based social network for didactic building and sharing," in *Artificial Intelligence in Education - 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings*, 2013, pp. 774–777.

[17] M. De Marsico, A. Sterbini, and M. Temperini, "A framework to support social-collaborative personalized e-learning," in *M. Kurosu (Ed.) Human-Computer Interaction (Part II), Proc. HCI Int. LNCS 8005*, Springer, Ed., 2013, pp. 351–360.

[18] G. Gentili, A. Micarelli, and F. Sciarrone, "Infoweb: An adaptive information filtering system for the cultural heritage domain." *Applied Artificial Intelligence*, vol. 17, no. 8-9, pp. 715–744, 2003.

[19] G. Gentili, M. Marinilli, A. Micarelli, and F. Sciarrone, "Text categorization in an intelligent agent for filtering information on the web." *IJPRAI*, vol. 15, no. 3, pp. 527–549, 2001.

[20] F. Gasparetti, A. Micarelli, and F. Sciarrone, "A web-based training system for business letter writing," *Knowledge-Based Systems*, vol. 22, no. 4, pp. 287–291, 2009.

[21] C. Limongelli, M. Lombardi, A. Marani, and F. Sciarrone, "A teacher model to speed up the process of building courses," in *Human-Computer Interaction. Applications and Services - 15th International Conference, HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part II*, 2013, pp. 434–443.

[22] G. Shani and A. Gunawardana, "Evaluating recommendation systems,"

in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Springer US, 2011, pp. 257–297.

[23] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Infromation Retrieval.* Addison Wesley, 1999.

[24] C. Biancalana, F. Gasparetti, A. Micarelli, and G. Sansonetti, "Social semantic query expansion," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 4, 2013.

[25] F. Sciarrone, "An extension of the q diversity metric for information processing in multiple classifier systems: a field evaluation." *International Journal of Wavelets, Multiresolution and Information Processing IJWMIP*, vol. 11, no. 6, 2013.

[26] M. De Marsico, A. Sterbini, and M. Temperini, "The definition of a tunneling strategy between adaptive learning and reputation-based group activities," in *Proc. 11th IEEE Int. Conf. on Advanced Learning Technologies, ICALT*, 2011, pp. 498–500.

[27] F. Gasparetti, A. Micarelli, and G. Sansonetti, "Exploiting web browsing activities for user needs identification," in *Computational Science and Computational Intelligence (CSCI), 2014 International Conference on*, vol. 2, March 2014, pp. 86–89.

[28] A. Sterbini and M. Temperini, "Dealing with open-answer questions in a peer-assessment environment," in *Proc. ICWL 2012, LNCS 7558*, 2012, pp. 240–248.