The Use of Ontology in Semantic Analysis of the Learner's Queries on the Web through Search Engines

Samia Ait Adda[1], Amar Balla[2]
National High School of Computer
Ouad Smar, Algiers, Algeria
s_ait_adda, a_balla {@esi.dz}

ABSTRACT: *Search engines have become an integral part of our information environment. Gradually more they are replacing the role of libraries in facilitating information discovery and access. The learner is one of the most remarkable users of the search engines on the Web. The purpose of this analysis is to identify the domain concepts that are most asked on the Web and to store them into the leaner's model Knowledge as concepts not well mastered. During this article, we want to underline the inherent reasons that push learner to search for domain concepts in the outer of the learning platform by the use of the search engine; so we introduce an hypothesis which says that all concepts sought on the web by use of the search engine can be considered as knowledge poorly or badly acquired by learners and requires support both by the course author to restructure and adapt the course content and from tutor to monitor the learner on these identified concept.*

Keywords: E - Learning, Search Engine,Track, Semantic Indexing, Learner Model

Received: 11 June 2014, Revised 28 July 2014, Accepted 31 July 2014

1. Introduction

The World Wide Web (WWW) has immense resources for all kind of people for their specific needs. Using search engines (e.g. Google, Bing, Yahoo!) to locate Web information is probably the most common application we use every day [1]. The use of the search engines is the one of the most practical ways to access these documents on the web. This action is also commonly made by learners, user of the internet through the online learning systems, during their learning process in order to supplement or enrich their knowledge on some concept of the thought course.

Thus, in this article we interest on semantic analysis of learners query on the Web. The objective of this analysis is to identify the domain concept sought by the learner in the Web and to store it in the learner model as concept and knowledge not well understood.

In addition, we make the assumption which says that all concepts researched and asked on the Web can be considered as wrong or bad acquired concepts by learners, and requires thus more attention and consideration, both by the tutor to assist and help learners on these concepts, and by the designer of the course, to restructure and to further enrich the educational content articulating these concepts previously identified by the analysis. This analysis allows us to simply recognize the learners with difficulties and concepts which pose a problem for them.

The paper will be structured as follows, in the first section we present some works that are carried on the analysis of the user requests on the web, then we are interested in some semantic web tools, such as ontology and the possibility, thanks to these tools to exploit the contents of learner query to detect concepts badly acquired by the learner, further we detail and explain our approach. Finally we present some results which we have achieved through an implementation of our approach on a considered domain.

2. Related Work

A diverse range of papers and work report the results of studies of the information–seeking, retrieval behavior observed in search engine environments, classification and ranking of result [2]. However, in this paper we will refer only to some work which deals the modeling of the user who seeking information.

In the paper [3], the authors have proposed an approach to personalized query expansion based on a semantic user model. They discussed the representation and construction of the user model which represents individual user's interests by semantic mining from user's resource searching process. They exploited the user model to provide semantic query expansion service in e-Learning system.

Authors in [4] have shown an approach for personalized retrieval in an e-learning platform, that takes advantage of semantic Web standards to represent the learning content and the user/learner profiles as ontologies, and that re-ranks search results/ lectures based on how the contained terms map to these ontologies. The important aspect of their approach is the combination of an authoritatively supplied taxonomy by the colleges, with the data driven extraction (via clustering) of a taxonomy from the documents themselves.

In [5], the authors perform the personalized semantic search and recommendation of learning contents on the learning Web-based environments to enhance the learning environment. Semantic and personalized search of learning content is based on a comparison of the learner profile that is based on learning style, and the learning objects metadata. This approach presents both the learner profile and the learning object description as certain data structures.

This paper [6] proposed a new method for the personalized search, using click-through data as the personal data. Firstly, uses the semantic statistical of word frequency method to extract the query expansion terms and recommended to the user. Secondly, improves the Naive Bayesian classifier and combines SVM to make users' personalized learning models, then provides personalized re-sort results by user models.

In paper [7] authors investigated the search personalization problem and presented an ontology-based framework which automatically learns the user's search interests based on the combined analysis of the user's past click through data and current queries. In first they proposed the use of a topical ontology for identifying the topic importance of Web pages and associate them with user clicks on search results. Then, based on this association they presented a method for actually learning the user interests based on both the user-issued queries and their relationship to the users past topic preferences. Finally, they proposed a method to rank search results based on the learned user interests.

The study in [8] authors proposed an ontological approach for semantic-aware learning object retrieval. The proposed ontological approach has two significant novelties: a fully automatic ontology query expansion algorithm for inferring and aggregating user intentions based on their short queries.

In [9], authors introduced a method for learning and updating a user profile automatically. The proposed method belongs to implicit techniques. It processes and analyzes behavioral patterns of user activities on the web, and modifies a user profile based on extracted information from user's web-logs. The method relies on analysis of web-logs for discovering concepts and items representing user's current and new interests.

We may distinguish that in all these studies, the authors are based on Web log files (tracks) to recover the data that represent the activity of a user on the web (implicit modeling). These data from the web are then filtered, analyzed and mapped on ontology. These works are intended to update the user profile, and more exactly, its interests, preferences and intentions, always with the goal of ensuring personalization of information retrieval and the recommendation of the learning object.

In our case we focus to model the knowledge of the learner and not its interests or preferences, so we consider that all domain concepts searched from the internet using the search engine may be considered as knowledge bad or no mastered by the learner and stored it in leaner profile as lacks or deficient. To do this, the concepts of the studied domain are modeled through ontology of the taught course, as we describe it below.

3. Domain Ontology

The Semantic Web [10] is an understandable and navigable space by both human and software agents. It introduces an additional meaning to the navigational data of the classical web, based on a formal ontology and controlled vocabularies through semantic links. In standpoint of e-learning, it can help learners to locate, access, querying, processing and evaluating learning resources across distributed heterogeneous network, or assist teachers in creating, using, locating, or the sharing and exchanging learning objects. Ontology [11] includes a set of terms, knowledge, including vocabulary, semantic relations, and a number of logic-inference rules for some particular domain. The ontology applied to Web creates thus the Semantic Web [12].

Ontologies [13] facilitate the sharing and reuse of knowledge, i.e. a common understanding of diverse content by persons and machines.

The use of ontology in our case consists in the conceptual indexing of the researched query and on top of that indexing, the most searched domain concepts on Web by learners will thus detected. This ontology also represents the structure of the learner's model, since it is part of the domain model, i.e. the domain ontology in our case.

In our case of study, we consider that ontology is composed of a set of concepts and relations between these concepts. A unique identifier is assigned for each concept; these concepts are labeled with one or several terms. The domain ontology model that we propose is shown in Figure 1. The considered educational resources are described by a set of metadata (LOM) [14].

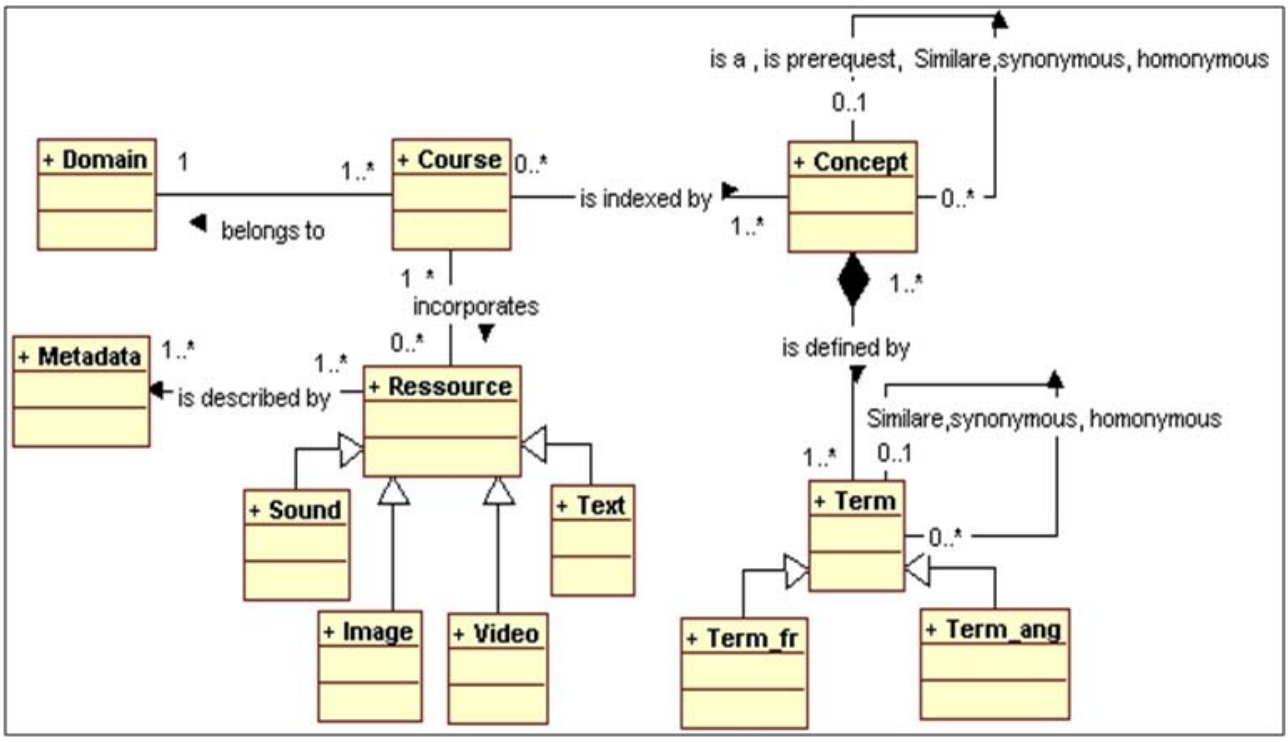

Figure 1. The model of domain ontology

4. The Proposed Approach

To detect knowledge of domain supposed poorly assimilated by learners and learners with difficulties, we propose an approach that consists in constructing vector relative to the concepts look up on the Web by learners using the research engine.

These concepts are taken from the query of learner made on the Web and then indexed according to the concepts of the taught course through the domain ontology (semantic indexing). The most researched concepts by a learner or group of learners are then highlighted. Our goal is twofold. On the one hand we try to detect the most researched concepts of domain by learners, in the other hand we want to identify learners who have used these concepts via the engine research to make inquiries about some domain concepts.

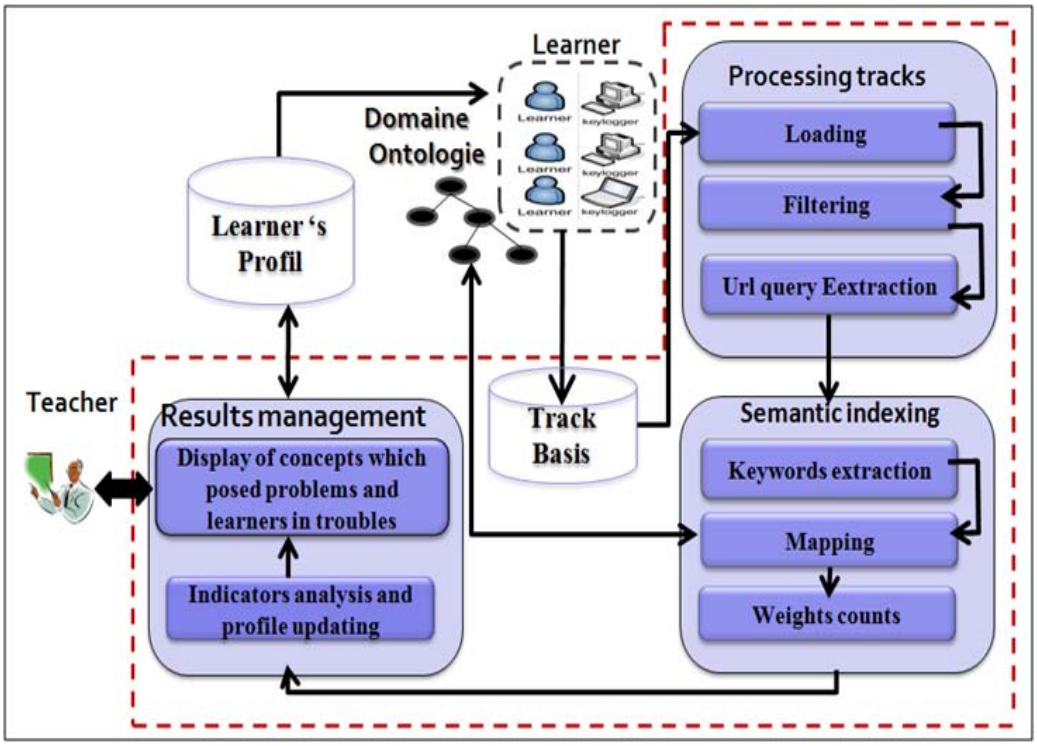

Figure 2. Architecture of the proposed solution

The approach that we propose (cf. Figure 2) is divided into three basic processes: (1) processing tracks, (2) Semantic Indexing of query and (3) the Management and processing of results, this is what will be detailed in the following of this paper.

4.1 Precessing Traks

In traditional learning, interaction between the learner and the teacher are manifold through educational materials, gestures and words. The teacher can, according to his observations, change the course of his career to suit the different profiles of its learners [15].

In distance learning, these observations are derived from traces collected (log file). Jermann [16] defines a digital trail as a set of observations on the interaction of the learner with a system. It is defined as an observable time sequence of the browsed pages modeled by the URL and a set of actions on these pages.

The collect on the client side provides data relating to the path of the learner, whether inside the learning platform, or in outside of it, i.e., browsing the course or the Web, communicates via forum or chat, running various applications on the local machine. We are interested in our case particularly on the navigation data Web.

The data collected on the client side is called log files or the track files. Voluminous and very meticulous that are, it is difficult or impossible to interpret them as they are by humans. It is therefore indispensable to perform some processing on it to the make it interpretable by the cleaning process.

The cleaning consists in filtering the insignificant and superfluous data from the track file. In our context we keep only the URL captured by the browser with the domain name corresponding to that of the search engine (e.g., Google, Lycos, and Yahoo), we

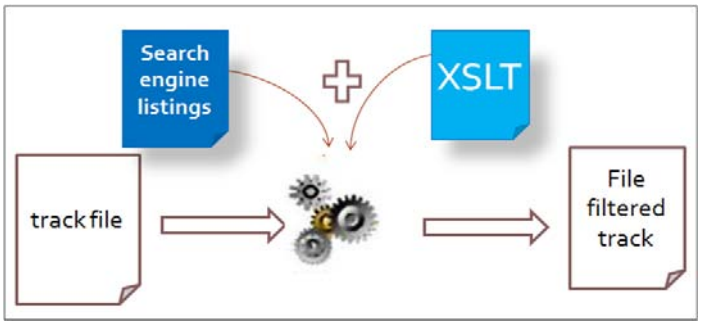

Figure 3. The filtering track's file corresponding to the Search Engine URL

use do this the XSLT file.

For example a researched on Google engine with the request "*session and cookies in php*" will look in the address bar of the browser as:

"https://www.google.com/search?q=session+and+cookies+in+php&ie=utf-8&oe=utf&aq =t rls = org.mozilla:en-US:official&client=firefox-a" So, this URL will be captured in the File filtered track, with the date and the IP address of the learner machine.

In this kind of URL, keywords of the query which were written on a search engine are easily located after some parameter corresponding for each search engine; Table 1 summarizes some search engine and the corresponding parameters.

Search Engine	Google	Yahoo	Lycos	Netscape.com	Ask.com	Aol.com	Webcrawler.com	Altavista.com
parameter	q	p	query	query	q	query	search/web/	q

Table 1. Corresponding Parameters of Some Search Engine

Also in the previous example the keywords of the query are located immediately after the parameter value q which is followed by the various keywords separated with a plus sign

4.2 The Semantic indexing

We propose to use domain ontology to build the semantic index of keywords of the learner request. The process of query's indexing is handled through three main steps: (1) Identifying ontology concepts, (2) Assigning concepts to terms (key word) and (3) Weighting concepts.

In the following, we present these steps.

4.2.1 Concept Identification: The purpose of this step is to identify ontology concepts that correspond to the query words. Concept identification [5] is based on the overlap of the local context of the analyzed word with every corresponding domain ontology entry. Concepts are referred in the query with simple or compound words (term). The concept identification algorithm is given in Figure 4.

In the ontology, a set of terms is used for labeling concepts and relationships between concepts. That set forms the vocabulary of the ontology. To respond nevertheless in case if the processed term is ambiguous, a disambiguation step is so necessary.

4.2.2 Term Disambiguation: The operation of terms conceptualization consists on assigning to each term *t* a concept *c* in ontology. Among the cases of ambiguity that can be introduced, the polysemic terms and homographs terms. So we distinguish these two particular situations, semantic and linguistic ambiguities.

i) Semantic or polysemy Ambiguity: This is where a number of concepts are defined by the same term, i.e., the same term may be the label of several concepts in the ontology. For example, the term "*circuit*" has seven senses in WordNet [17] as name and one sense as a verb. It can thus refer to eight different concepts. In this case, we proceed as follows. For an ambiguous term t_i,

we seek in document a label of a concept C_k in relation, in the ontology, with a concept C_i indicated by the ambiguous term t_i. If C_k exists we take C_i as the concept designated by the term t_i.

ii) Language or homography ambiguity: two terms belong to different languages may have the same form in a text; this relationship can be seen as a relation of multilingual homonymy. For example the word "*variable*" exists either in French and English language. In case if t_i is homograph, we seek in document an unambiguous term t_k in standpoint of language at the proximity of the term t_i. Therefore, the language of t_k defines the language of the term t_i.

Input: query Q.

Output: Vector of all ontology concepts belonging to terms of query Q.

Procedure

Let w_i be the next word to analyze in the query q. We define the context $sent_i$ which is the set of terms of query that contains the word occurrence w_i being analyzed.

Compute $V_i = \{C_1, C_2, ..., C_n\}$ the of ontology entries containing wi.

Each $C_j \in V_i$ is represented by a multiword or mono-word term.

Rank concepts C_j **in** set V_i in **where**: $|C(1)| > |C(2)| > || \ldots > |C(n)|$ // || denotes the concept length, in terms of the number of words in the corresponding terms.

For each element C_j **in** V_i **do**

Get common words between $sent_i$ and representative term of C_j, which is the intersection

$N = \cap (sent_i, C_j)$

If $|N| < |C_j|$ **then** the concept-sense is not within the context.

End If

If $|N| = |C_j|$ **then** the concept-sense C_j is within the context $sent_i$.

Add C_j to the set of vectors' element (index) associated to query Q.

End If

End For

Figure 4. The algorithm of Words Mapping into Concepts

4.2.3 Concept weighting: The extracted concepts are weighted according to a method more general than *tf*idf* named $Cf_c * idf$ (concept-frequency-inversed query frequency). In this method each extracted term represents necessarily a concept of the ontology since we used ontology to identify them. For a concept C its frequency in a query depends on the frequency of the word itself [18]. It is calculated as follows:

$$idf_c = log \frac{N}{f_c} + 1 \quad (1)$$

$$Cf_c = \sum_{tm \in t(c)} tf_{tm} \quad (2)$$

Where: $t(c)$ is the set of terms corresponding to different concept C. The weight of each concept in a query q is so calculated as follows:

$$Cfidf = Cf_c \times idf_c \quad (3)$$

4.3 Representation of the semantic Learner Knowledge

The learner's models are cognitive models which allow to provide relevant information for a learning system in order to adapt learning to the knowledge, competences, features, preferences and objectives of apprenticeship to learner in particular domain (these which is taken in the learning environment) [19].

There are five popular and useful features for an individual learner's representation, these are: the learner's knowledge, interests, goals, background, and individual traits [20]. The user's knowledge of the subject being taught or the domain represented in hyperspace appears to be the most important user feature. For existing learning adaptative system, the knowledge is frequently the only user feature being well-modeled [21], [22].

So, corresponding to each learner, we obtain set of sessions (i.e. Consist on the connection time of a leaner on his personal space in the e-leaning platform). Let C be a set of n Concept of the Domain ontology: $C = \{ C_1, C_2, \ldots, C_n\}$, and let L be a set of m learners registered in a specific course within the e-learning environment, $L= \{L_1, L_2, \ldots, L_m\}$, the learner knowledge model LK_i corresponding to the learner $L_i \in L$ is represented by a set of p sessions S^i_j extracted from log file : $LK_i = \{ S^i_1, S^i_2, \ldots\ldots, S^i_p\}$ where each S^i_j is a subset of k weighted requested concept C_i, $S^i_j : <w(C^{S^i_j}_1), w(C^{S^i_j}_2), w(C^{S^i_j}_3), \ldots\ldots, w(C^{S^i_j}_k)>$ where each $(C^{S^i_j}_k) = Cl$ for some $L \in \{1, 2, \ldots, n\}$, and $w(C^{S^i_j}_k)$ is the weight associated with the requested concept $(C^{S^i_j}_k)$ in the session S^i_j corresponding to the i^{th} learner compute with the Equation (3).

The learner's knowledge component LK_i can be represented as a matrix Ml (p, n), where p is the total number of learner's sessions and n the cardinality of concept of the domain ontology. Therefore, concepts with a high weight by estimating a threshold α, that we will fixed by experimentation, will be reviewed as problematic domain concepts for a learner. Therefore, the tutor may intervene to help the learner on these concepts.

Once learners' models are delimited properly, we apply a second treatment based on a collaborative analysis in order to detect the most concepts which is asked on the web by a group of learners based on their queries.

So the learner's knowledge of groups will be presented as matrix $Mg(m, n)$ where m is the total number of learners who participate in the learning process. Thus, the weight Wk of each concept Ck in session sessions S^i_j will be computed by the sum of the weights of all learners.

$$Wk = \sum_{i=1}^{m} w(C^{S^i_j}_k) \quad (4)$$

Similarly to the first treatment, the concepts greater than a threshold γ, which we also determine the value by evaluation, will be considered as wrong developed concepts in the course. To this end, the designer of course can review the content of resources that explain these identified concepts, and further enrich its course on these concepts.

Accordingly, the learner model in the proposed system is built in ontological representation since we use domain ontology. Learner model is built by mapping of learner's knowledge information and the concept in domain ontology; converting the browsing contents of the learner's knowledge into the form of ontology concept, and using these ontology concepts to construct learner's knowledge ontology. The figure 5 shows an algorithm of building and acquiring of the learner's knowledge model relative to the browsing concepts on the Web:

Input:

δ: the set of concepts of domain ontology that was sought by learner during the learning session $Id_{session}$

O: Concept in domain ontology

Output: The learner's knowledge model (UKM)

For each C_i in δ

Get concept identifier Id_{Ci} from O and its weight by using equation 3

Save concept identifier, $Id_{session}$, Web as aspect and its weight as user knowledge value in UKM

End For

Figure 5. The algorithm of learner's knowledge Model acquiring

5. Experimentation

5.1 The Test Collection

For our experiments, we have proceeded to test on group of computer science students in the second years, with the number of 27. We have proposed to them *PHP* course, shown in *eFAD* (www.ufc.efad.dz) platform and modeled with the ontology of *SKOS* format [23] (Simple Knowledge Organization System). The experiments were established in three sessions of one hour. The PHP course is mainly composed of 8 top concepts and 49 sub-concepts. To consolidate our experiment, we have conceived a questionary paper which we have distributed to students, asking them to place concepts which pose problem to them. At the end of the test, a written assessment was performed for all learners.

5.2 Evaluation of results

The result of the experimentations consists of 27 corpus of each learner plus the general corpus. Therefore, we constituted a number of 48 requests for all of learners participated in the Test. The following diagram shows the score characterizing the main domain concepts for each learner seeking information by use of the search engine:

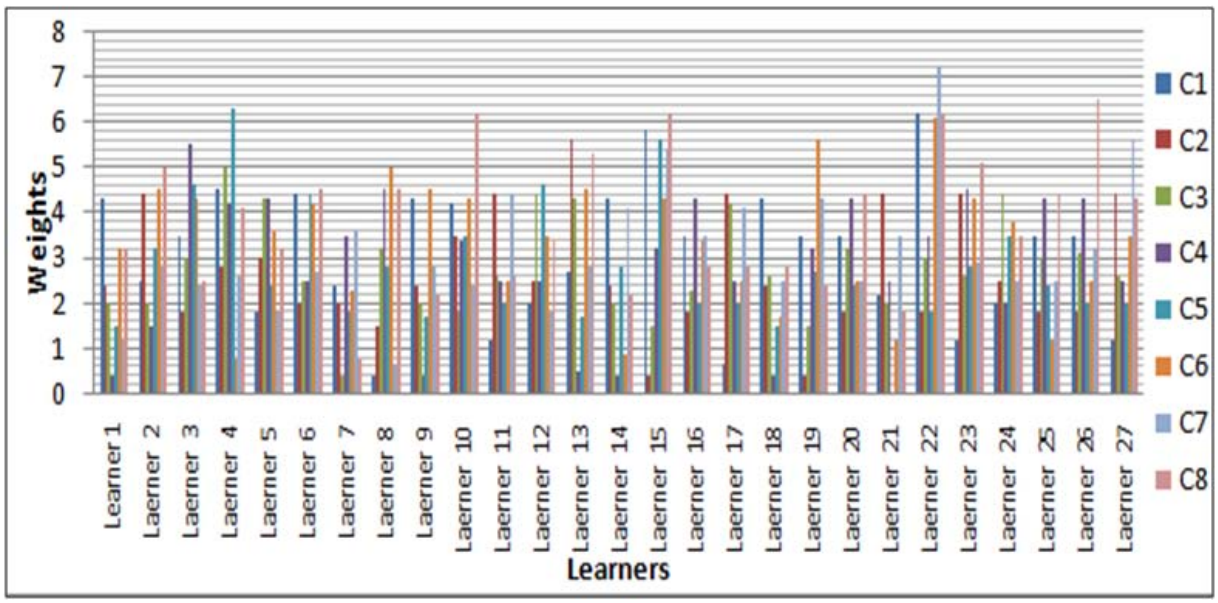

Figure 6. The Researched Weights of the Main Domain Concepts

The challenge of this test was to find the concepts insufficiently mastered by each learner. That is to say; the threshold á, that we set to evaluate the most concepts which posed a problem for student *j*, is estimated by the median of the weights of concepts researched by this learner. This threshold is different from learner to learner; accordingly we have counted 27 values of this threshold (Figure 6). Indeed, we found that the concepts *C*4 and *C*6 have posed a problem for some students who are recognized by the following process.

As that is signaled, a written assessment was performed for each learner on each concept of the domain as well as a questionnaire which we have asked them to indicate the concepts not mastered. Therefore, the diagram in the Figure 7 shows a comparison of different results obtained from the assessment, the weight of researched concepts and questionnaire responses.

As for the threshold g, which we considered to estimate the concepts which are badly defined in course, it is determined by comparing the result with that obtained through the written assessment and questionary responses, the value is fixed at 0.32. As a result, we detect that 5 learners have problems with some concepts of the domain (learners 2, 10, 14, 18, 22).

Finally, we have made the correlation ratio between mark and the requested weight of concepts domain (Figure 8) which is measured at -0.51. This value on concepts interprets a causal relationship between the two variables, concept visit weight and assessment mark of concept which are due to the learner's knowledge state on these concepts. This confirms as it should be, the hypothesis raised in the beginning of this paper.

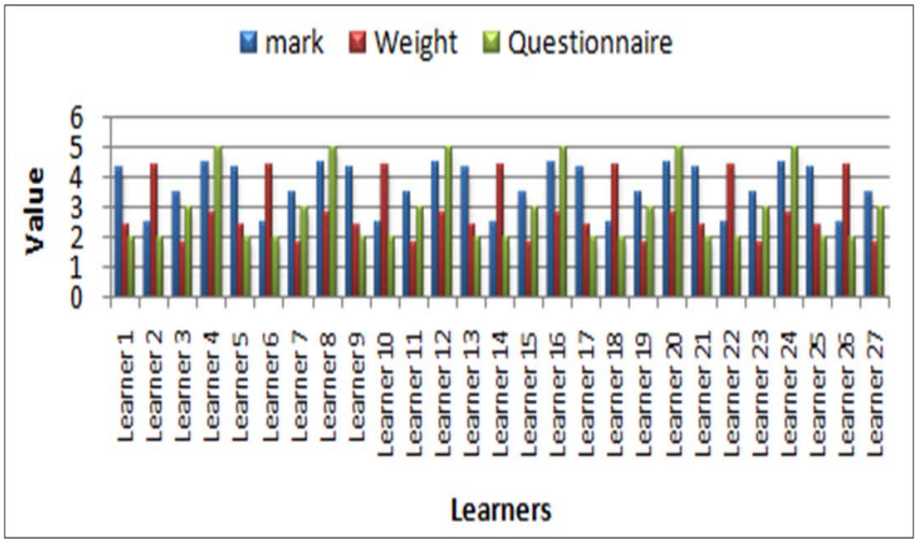

Figure 7. Synthesis of the Mark-Weight-Questionnaire results for each learner

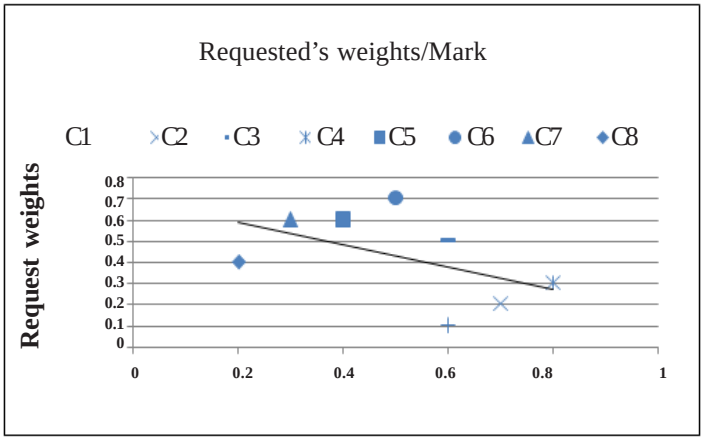

Figure 8 : Cloud digramme of Mark-Weight

6. Conclusion

The educational adaptive systems provide good support for learners on their individual characteristics. It can also provide information on the needs and deficiencies of these learners, either for the tutor or designer of the course, even for evaluation, monitoring and customizing the process and strategy of learning. In fact, the learner model must be developed for each student, containing information about the history of his interactions, objectives and knowledge badly acquired.

In this article, we have highlighted the need to analyze the request edited by learners in the web by use of the search engine during the learning sessions and we have proposed an approach for semantic analysis that we have presented and explained which permits to detect domain concepts that were difficult for learners, by comparing the content of their requests with a domain ontology of the studied course. An experiment was carried out on a group of students taking a PHP course, and has enabled us to validate the proposed approach and to set some parameters. This result needs to be further refined by additional tests, which we are presently conducting.

References

[1] Raval, V., Kumar, P., Kosta, Y. (2012). SEReleC# - C# Implementation of SEReleC: A Meta Search Engine based on Combinatorial Search and Search Keyword based Link Classification, CUBE, September 3–5, Pune, Maharashtra, India.

[3] Xiaojian, L., Shihong, C. (2009). Personalized Query Expansion Based on Semantic User Model in e-Learning System. Sixth International Conference on Fuzzy Systems and Knowledge Discovery. IEEE.

[4] Leyla, Z., Olfa, N. (2008). Semantic Information Retrieval for Personalized E-learning.20th IEEE International Conference on *Tools with Artificial Intelligence*. 1082-3409/08, IEEE.

[5] Salah, T., Khaled, B., Naveed, M. F. A (2013). Personalized Semantic Retrieval and Summarization of Web Based Documents. *In*: (IJACSA) *International Journal of Advanced Computer Science and Applications*, 4 (1).

[6] Cheqian, C., Kequan, L., Heshan, L., Shoubin, D. (2010). PERSONALIZED SEARCH BASED ON LEARNING USER CLICK HISTORY. Proc. 9th IEEE Int. Conf. on Cognitive Informatics (ICCI'10). IEEE.

[7] Stamou, S., Ntoulas, A. (2008). Search Personalization through Query and Page Topical Analysis, Kluwer Academic Publishers. Printed in the Netherlands.

[8] Ming-Che, L., Kun, T., Tzone, W. (2008). A practical ontology query expansion algorithm for semantic-aware learning objects retrieval. *Computers & Education* 50, 1240–1257. Elsevier Ltd.

[9] Reformat, M., Koosha, S. (2009). Updating User Profile using Ontology-based Semantic Similarity, FUZZ_IEEE, Korea, August 20-24, IEEE.

[10] Dicheva, D. (2008). Ontologies and Semantic Web for E-Learning, *In*: Handbook on Information Technologies for Education and Training, Springer Berlin Heidelberg.

[11] Siti, U., Rohiza, A., Shakirah, M. (2010). Ontology of Programming Resources for Semantic Searching of Programming Related Materials on the Web. IEEE.

[12] Berners-Lee, T., Hendler, J., Lassila, O. (2001). The semantic web, A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientic American*, p. 3543.

[13] Zschocke, T., DeLeon, J. (2010). Towards an Ontology for the Description of Learning Resources on Disaster Risk Reduction. *In*: CCIS 111, p. 6074, Springer-Verlag Berlin Heidelberg , Article in a conference proceedings. p. 7-12,

[14] IMS Global Learning Consortium. IMS Meta-data Best Practice Guide for IEEE 1484.12.1- 2002 Standard for Learning Object Metadata (2006).

[15] Ben, M., Sassi., Laroussi, M. (2012). Towards learners' tracks standardisation, frantice.net, numéro 5, september, www.frantice.net

[16] Jermann, P., Soller, A., Muehlenbrock, M. (2001). From mirroring to guiding: A review of state of the art technology for supporting collaborative learning. *In*: Proceedings of the First European Conference on Computer-Supported Collaborative Learning (p. 324-331).

[17] Fellbaum, C. (2010). WordNet. In R. Poli et al. Theory and Applications of Ontology: Computer Applications, (p. 231-243). 231-243, Springer Science+Business Media B.V.

[18] Dragoni, M., Pereira, C., Tettamanzi, A. (2010). An Ontological Representation of Documents and Queries for Information Retrieval Systems, IEA/AIE, Part II, LNAI 6097, p. 555–564, Springer- Verlag Berlin Heidelberg.

[19] VanLehn, K. (1988). Student models. *In*: M. C. Polson and J.J. Richardson (eds.): Foundations of intelligent tutoring systems. Lawrence Erlbaum Associates, Hillsdale. p. 55-78.

[20] Brusilovsky, P., Henze, N. (2007). Open corpus adaptive educational hypermedia. *In*: Brusilovsky, P., Kobsa, A., Neidl, W. (eds.): The Adaptive Web: Methods and Strategies of Web Personalization. Lecture Notes in Computer Science, 4321. Springer-Verlag, Berlin Heidelberg New York, 4 194.

[21] Rueda, U., Larrañaga, M., Arruarte, A., Elorriaga, J. A. (2006). DynMap+: A Concept Mapping Approach to Visualize Group Student Models, First European Conference on Technology Enhanced Learning, ECTEL'06, W. Nejdl & K. Tochtermann (Eds), p. 383397, Cret, Greece.

[22] Bull, S., Gardner, P., Ahmad, N., Ting, J., Clarke, B. (2009). Use and Trust of Simple Independent Open Learner Models to Support Learning within and across Courses, *User Modeling, UMAP*, Houben, G. -J., et al. (Eds.), LNCS 5535, p. 42–53, Springer-Verlag Berlin Heidelberg.

[23] Tuominen, J., Frosterus, M., Viljanen, K., Hyvnen, E. (2009). ONKI SKOS Server forPublishing and Utilizing SKOS Vocabularies and Ontologies as Services. *In*: European Semantic Web Conference (ESWC 2009).