Porto Alegre, v.14, n.2, jul./dez. 2011.
ISSN impresso 1516-084X ISSN digital 1982-1654

FindYourHelp: an expert finder module on Virtual Learning Environments

FindYourHelp: um módulo localizador de especialista em Sistemas de Gestão de Aprendizagem

Abstract
This paper discusses the creation of an additional module for the Moodle environment. FindYourHelp enables automatic identification of experts who make their contribution to discussion forums. Such an approach differs from previous solutions in that it is based on applying text mining techniques as a supplementary analysis of students' participation in the existing environment. A feasibility study of such a solution is provided and has demonstrated the tool's applicability.
Keywords: Virtual Learning Environment. Text Mining. Expert Finder Tool.

Resumo
Este artigo discute a criação de um módulo adicional para o ambiente Moodle, que permita a identificação automática de especialistas que contribuam em seus fóruns de discussão. O diferencial da solução desenvolvida consiste em aplicar técnicas de mineração de textos como forma complementar à análise de participação dos estudantes já existente no ambiente. Um estudo de viabilidade da solução foi desenvolvido e evidenciou a aplicabilidade da ferramenta em relação a seus objetivos iniciais.
Palavras-chave: Ambiente Virtual de Aprendizagem. Mineração de Textos. Busca por Especialistas.

SANTOS, Marcos Lapa; SALVADOR, Laís Nascimento; CRUZES, Daniela Soares. FindYourHelp: an expert finder module on Virtual Learning Environments. *Informática na Educação*: teoria & prática, Porto Alegre, v. 14, n. 2, p. 95-112, jul./dez. 2011.

Marcos Lapa Santos
UNIFACS

Laís Nascimento Salvador
Universidade Federal da Bahia

Daniela Soares Cruzes
NTNU

1 Introduction

The advent of virtual learning environments (VLE) as a support to collaborative discussion groups in several undergraduate institutions has boosted the creation of a large amount of information circulating and stored in major academic databases. Such environments "[...] connect people and link knowledge through discussion topics creation, messages posting in forums, chats, online content management tools (WIKI), among other features [...]" (SANTOS; SALVADOR, 2009, p. 2).

The great advantage of learning in a VLE is the assumption that learning takes on a different meaning when it relies on collaborative action and approximates people that help each other aiming at knowledge construction and reconstruction (MOORE; KEARSLEY, 2004). At the same time, the contemporary world is going through a phenomenon called *hyper-specialization* characterized by people

Porto Alegre, v.14, n.2, jul./dez. 2011.
ISSN impresso 1516-084X ISSN digital 1982-1654

searching for continued and specialized education as a means of being inserted in the labor market, which is increasingly competitive.

Studies on Knowledge Management show that in order to maximize the contribution to the environment in which they operate, these people should try to share their knowledge with others, avoiding knowledge to become stagnant with the one who holds it (NONAKA; TAKEUCHI, 1995). However, an antagonistic problem that appears is related to the difficulty in finding experts to find solutions in their domain areas (SANTOS; SALVADOR, 2009).

According to Mattox, Maybury and Morey (2010), many social groups have encountered significant questions that could be answered if the right person to answer such questions were found in time. With respect to virtual learning groups, there are some situations in which it is also important to know who is, in fact, positively contributing to the course. This information is useful, for example, to help teachers and coordination teams to make decisions about leadership and evaluate the need for the creation of subgroups.

Searching for experts with the appropriate skills and knowledge in a specific research field is an important task when it comes to academic activities. For teachers it is important to:

1. Identify which student has greater affinity with certain subjects, or those who contribute most to the construction of collective knowledge within the group,
2. Motivate their participation in the group, which is usually translated into gains for the students and the teacher.

Expert Finding is the area of research that addresses the task of finding the right person with the appropriate skills and knowledge (BALOG; RIJKE, 2010). The search for experts can be purely manual, such as the analysis of document authoring, email, or the participation of people in social networks. Nevertheless, it is interesting to have a way to automate such a process. An Expert Finder tool is basically a machine that finds experts in certain subjects, usually taking into account aspects such as the number of publications on specific issues, number of citation by other authors, number of joint communities and others (BECERRA-FERNANDEZ, 2006).

Seo and Croft emphasize that:

> Identification in online communities is of importance for the following two reasons: online communities can be viewed as knowledge databases where knowledge is accumulated by interactions between the members […] On the other hand, in terms of communication dynamics, online communities are spaces where non-experts can communicate with experts. (SEO; CROFT, 2009, p. 1)

Our goal is to analyze the adoption of algorithms and techniques of text mining as a way of supporting the search for experts in research/academic discussions groups. A related work can be observed in Maybury, D□Amore and House (2001), which focus on collaborative communities instead of VLE. This paper discusses, therefore, the creation of an open source solution that aims to identify students who may be considered experts within discussion forums in the Moodle environment. The results of a case study, in which such a module was evaluated for three different subjects, will also be presented. This paper is an extended version of (SANTOS; SALVADOR; CRUZES, 2010) presented at XXI *Brazilian Symposium* on Computer in *Education* (SBIE).

This work is organized as follows. Section 2 presents the area of search engines by experts and some related work. Section 3 describes our proposal, the module FindYourHelp for the Moodle environment. Section 4 discusses some results of the experiment applied. Finally, conclusions and future work are presented in Section 5.

2 Search Engine for experts

An Expert Finder tool is basically a machine that finds experts in certain subjects, usually taking into account aspects such as the number of publications on specific issues, number of citation by other authors, number of joint communities and others (BECERRA-FERNANDEZ, 2006).

Many search engines have been developed nowadays. According to Jung *et al.*, (2007, p. 56) "[...] sources to find experts are various documents, programs, emails, databases, quotes, communities, among others [...]". Maybury (2006) in turn complements the idea of these authors noting that the above sources can also be composed of self-statements, summaries and web pages.

Some examples of search engine for experts are presented in Maybury (2006), see Fig. 1. It is worth noticing that these examples are all commercially available solutions, and none of them constitute free software or open source.

Such tools can be analyzed taking into consideration not only their sources, but also 1) type of processing that the software applies, such as automatic ranking of experts, extraction of entities from text (people, places, organizations), social network analysis (determining relationships between experts), support for foreign languages other than English and, finally, authorship identification (documents publications); 2) type of query supported by the tool: including keywords, boolean query, natural language, or taxonomic content view; 3) how results are displayed: ordered experts list, lists of documents or artifacts produced or used by experts, and concepts related to the expertise that the user is searching for, 4) system properties: interoperability support provided by the system or kinds of privacy support provided by the tool.

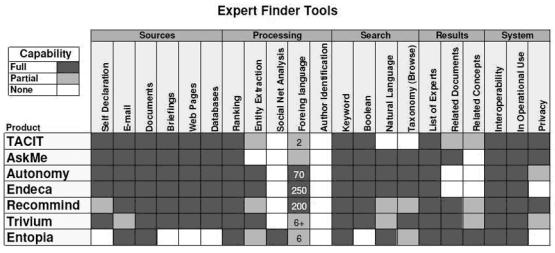

FIGURE 1 – Search Engine for experts
SOURCE: Adapted Maybury, 2006, p. 18.

Most listed tools offer considerable support to the diverse sources of expert search (Fig. 1), however, they differ widely when it comes to the support level provided to processing. In this context, it is important mention that they all have expert ranking and none of them has support for authorship identification. It is also possible to observe that, at some level, all of them have support for foreign languages other than English, Endeca tool, for example, supports 250 different languages. From Fig. 1 it is also possible to note that queries based on *keywords* and *Boolean queries* are predominant among tools, as for example AskMe, Autonomy and Endeca tools, which support all kinds of queries.

With regard to displaying results, all tools focus on listing experts. In this aspect, Entopia tool deserves special attention because it provides comprehensive support for all display types. As for properties, all the tools analyzed have broad support for interoperability and are in operational use.

Finally, a point that should be highlighted is that none of these tools uses the idea of analyzing messages posted by people in forums of discussion to infer what they know and whether they interact more, thus identifying their specialties. This is the contribution of the solution proposed in this paper.

In the next subsection we extend this analysis by addressing three important issues:

Porto Alegre, v.14, n.2, jul./dez. 2011.
ISSN impresso 1516-084X ISSN digital 1982-1654

VLE (Virtual Learning Environments), Expert Finding systems, and the use of text mining techniques in order to find experts.

2.1 Virtual Learning Environments and Expert Finding

The focus of this research is the search for experts in virtual learning environments (VLE). Some of such environments have been analyzed: AulaNET (GEROSA *et al.,* 2004), BlackBoard (BLACKBOARD, 2008), Moodle (MOODLE, 2008), Dokeos (DOKEOS, 2008) and Sakai (SAKAI, 2010). Table 1 compares the main strategies used by the VLE in order to verify user participation, specifically, the feature Message Content Analysis, that address the idea presented before of analyzing the content of messages posted by people in forums of discussion.

The obtained information about these environments reveal that most have ways to check the participation level of its users, such as: quantitative analysis about messages sent, estimative of user time dedicated to the course, written report of participation in discussion, contacts with academics, among others. However, none of the environments have demonstrated a way to analyze messages content conveyed in their information bases.

We have chosen the Moodle VLE in order to implement our proposal. The word Moodle was originally an acronym for Modular Object-Oriented Dynamic Learning Environment, which is mostly useful to programmers and education theorists. Moodle is an Open Source Course Management System (CMS), also known as a Learning Management System (LMS) or a Virtual Learning Environment (VLE). It has become very popular among educators around the world as a tool for creating online dynamic web sites for their students.

Moodle can be installed on any computer that can run PHP[1], and can support an SQL type database (for example MySQL). It can be run on Windows and Mac operating systems and many flavors of linux (for example Red Hat or Debian GNU). There are many institutions that already use this tool to VLE, in Brazil, more than 2900 sites are registered in the Moodle.org site. Worldwide there are 45916 sites from 199 countries registered (MOODLE, 2008b).

2.2 Text Mining for Expert Finding

Text Mining (TM), also called Textual Data Mining, Documental Information Mining, or

1 For further information about PHP, see: <www.php.net>

TABLE 1 – VLE comparisons

Environment	Technology	License type	Strategy to verify the participation level	Message content analysis?
AulaNet	Java/J2EE	GPL (Free)	Quantitative Analysis	No
Moodle	PHP	GPL (Free)	Quantitative Analysis	No
Blackboard	Not analyzed	Proprietary	Not Analyzed	Not analyzed
SAKAI	Java	Educational Community License	Quantitative Analysis	No
Dokeos	PHP	GPL (Free)	Quantitative Analysis	No

SOURCE: Prepared by authors.

Discovery of Textual Database is a technology for analysis of large collections of unstructured documents. It consists of the extraction of rules, patterns or trends from large volumes of texts written in natural language, usually for specific purposes. Many authors have applied text mining techniques for expert finding. Although Text Mining is a wide area, this paper will discuss only some techniques for text categorization that can be applied for expert finding, in other sources: Sim, Crowders and Wills (2006), Yang, *et al*. (2008) discusses an approach to locating an expert through the application of information retrieval and analysis processes to an organization's existing information resources; Dermatini (2007), proposes a finding experts searching in the content of Wikipedia articles; and Jung *et al*. (2007), propose an experts-finding method based on identity resolution and full text analysis, and further extract topic-centric information.

Campbell *et al*. (2003) use text mining algorithms in order to understand what the main subject of each e-mail is and to find out the contact persons of the email author, with this information they perform an expert ranking based on the person's contribution to that specific subject, Campbell *et al*. (2003) use an email analysis approach, where they analyze the email content in a corporation to find the experts who most contribute on specific subjects. They justify the importance of e-mail analysis:

> Email is a valuable source of expertise. It provides an easy-to-mine repository of communication between people in the social network, and it contains actual demonstrations of expertise (e.g., answering a question on some topic) as well as knowledge of expertise (e.g., decision of who should be asked the question). Both the content of email and the pattern of communication contain information about who knows what in an organization. (CAMPBELL *et al*., 2003, p. 1)

Our proposal also focuses on a similar analysis as Campbell *et al*. (2003), but it involves the mining of forum message content instead of email content. For this task, there are some approaches and techniques that can be applied on the message texts to categorize them into predefined subjects.

2.2.1 Some text categorization techniques

Wang and Taylor (2007, p. 395) highlight two keywords-based methods "[...] commonly used in various information retrieval and text mining applications [...]", the Latent Semantic Indexing – LSI (DEERWESTER *et al*., 1990) and the Vector Space Model – VSM (SALTON, 1975). An important aspect observed in these two methods is the fact that they perform a keyword based document search which is what we are interested in this research.

The Latent Semantic Indexing (LSI) is a specific indexing method based on the Latent Semantic Analysis (LSA) which is used in expert finding tasks. Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text (LANDAUER; DUMAIS, 1997 *apud* LANDAUER; FOLTZ; LAHAM, 1998). LSI uses a description of terms and documents based on the latent semantic structure for indexing and retrieval. Heerem and Sihn describe how LSA can be used in an expert finder tool:

> XPERTFINDER uses the Latent Semantic Analysis (LSA). This method is based on a vectorisation of the compared documents. Each vector can be used to represent one term in the document ... All vectorised documents of the topic-related reference texts form a vector space as columns of the matrix A, the so-called semantic space. (HEEREM; SIHN, 2002, p. 43)

Porto Alegre, v.14, n.2, jul./dez. 2011.
ISSN impresso 1516-084X ISSN digital 1982-1654

The VSM (SALTON, 1975) is another method commonly used in text categorization, and it was the method selected for this initial research because of its simplified implementation and adherence to fast analysis of short messages, the focus of our proposed tool. The VSM has two main steps: (i) transforms the document in a terms vector where each term has a value/weight associated; (ii) compares the terms to a category based on angle distance in Euclidian Space. In our proposal we use VSM in the following way:

- The message terms are the keywords from the posted message, extracted by the tokenizer;
- The categories are composed by terms that describe them, the category terms. So the message terms are compared to the category terms in order to carry out the expert identification.

Although VSM uses the same principle as LSA (transforms each text in a document vector and carries out the processing), it does not perform the semantic analysis step. This is why VSM shows better results than LSI regarding performance (MANNING; PRABHAKAR; SCHÜTZE, 2004). However, it has a disadvantage regarding to polysemy and synonym, which must be treated with auxiliary structures, *e.g. thesauri* and dictionaries.

3 FindYourHelp

In this work, we propose the creation of an expert search system to operate in discussion groups within virtual learning environments (VLE), entitled FindYourHelp. Our purpose is to enhance interaction and collaboration among discussion forums participants by pointing out possible experts in some subject (or matter) of group interest. The approach is based on the analysis of forum messages content in order to identify the participants that most have contributed in some subject (or matter) inside the group.

The original idea was to use FindYourHelp only in academic scenarios such as research groups or e-learning courses, but this solution can be easily adapted to a business context, for example collaborative workgroups supported by a groupware system.

The FindYourHelp module comprises three main operation stages in the use of VLE: the first is the definition of a hierarchy of categories; the second is the messages categorization at the time it is posted on a forum; the third is the expert ranking validation by the teachers and the visualization of the ranked Expert List.

A general view of the system operation is shown in Figure 2:

1. In order to evaluate the posts we use a tree of categories created by an expert, usually a teacher or the course coordinator. These categories are related to the subject predefined by the teacher – the forum discussion subject. FindYourHelp uses this subject hierarchy as a source of authorized information for a

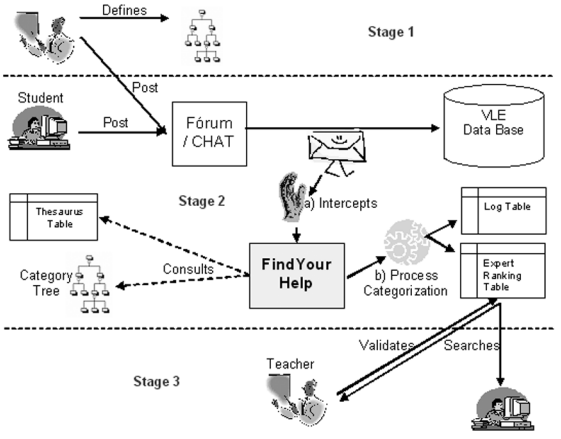

FIGURE 2 – FindYourHelp Operation
SOURCE: Prepared by authors.

future comparison with the messages that will be posted on the VLE.

2. After the message is posted in a forum, the categorization task is carried out by a text mining algorithm that: a) intercepts the messages for automatic pos categorization; b) reviews its contents in order to discover the message subject based on the tree of categories and saves the information about expert ranking and user logging.
3. With the expert ranking completed, FindYourHelp activates a human assisted validation task, where the teacher validates the inferred information about experts and enables everyone to find them later.

3.1 Architecture

We have chosen the Moodle VLE in order to implement our proposal. There are some advantages in using Moodle as our first experimental platform: (i) it is a broadly used environment in education institutions (ii) it is an open source solution, therefore new plug-ins can be created to extend its functionalities.

In our research, we found a feature in Moodle that also aims at analyzing the participation of discussion group members, but that does not analyze the content of the posts, focusing only on collecting the statistics of number of posts made by users. In order to complement this analysis, we used an automatic categorization technique in FindYourHelp module. This categorization is applied on topics previously defined by teachers of each subject. In this way, FindYourHelp is a plug-in added to Moodle environment that enhances the existing data structure to promote the search for experts in discussion forums messages. Figure 3 shows the FindYourHelp component architecture.

The remainder of this section describes the technical details of the FindYourHelp plug-in. This presentation is based on the implementation of its internal functionalities.

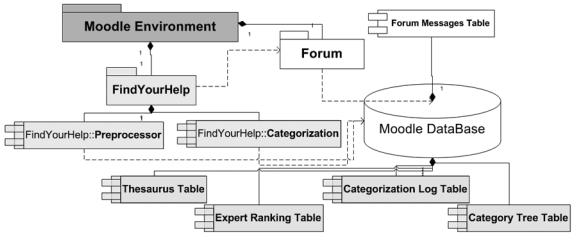

FIGURE 3 – FindYourHelp Architecture
SOURCE: Prepared by authors.

3.1.1 Environment setup

The first step in the FindYourHelp module is the set initialization of two tables: Thesaurus and Category Tree. These tables are added to Moodle data base as shown in Figure 3.

The Thesaurus is an auxiliary table that optimizes the processing of words with similar meanings in the text. In our case, this table was populated with imported data from OpenThesaurusPT project (OPENTHESAURUS, 2009). This project maintains a dictionary that lists words with a similar or related meaning in the Portuguese language.

Besides the Thesaurus setup, a forum participant – usually teacher or coordinator with good knowledge about the subjects of e-learning course – has to construct a category/terms hierarchy *i.e*. taxonomy of the main topics covered in the course. This hierarchy is loaded into Category Tree table (see Figure 3). Figure 4 shows the module screen that displays, for example, the categories related to game development:

Porto Alegre, v.14, n.2, jul./dez. 2011.
ISSN impresso 1516-084X ISSN digital 1982-1654

FIGURE 4 – Hierarchy of Terms in the game development example
SOURCE: Prepared by authors.

3.1.2 Preprocessor component

The Preprocessor component prepares the captured text from the message forum for the Categorization task. At the moment the message is posted, FindYourHelp intercepts its content and forwards this information to the Preprocessor that:

1. Removes numbers, symbols and punctuation,
2. Generates message "tokens" (divides the message into an array of terms),
3. Removes stopwords (words that do not interfere in text analysis, such as articles and prepositions),
4. Applies the process Stemming (reducing terms to its common radical), in this case the algorithm proposed by (COELHO, 2009), and
5. Applies synonym reduction by using a thesaurus as a support source for the algorithm.

After executing these steps, the text is prepared for analysis by the categorization algorithm.

3.1.3 Message categorization – Text mining

Once the preprocessing task has been carried out, the text mining algorithm generates a bag of words – a terms vector -- associated to the analyzed message. The message categorization algorithm is applied in order to confront this terms vector with the terms vector (or the category tree) related to the subject registered in the environment.

During this stage we use the weighting method TDIDF (Term Frequency / Invert Document Frequency) (ROBERTSON, 2004) to assign weighted values to each term vector related to the message content. TDIDF is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. Soucy and Mineau (2005, p. 2) identify the TFIDF as "[…] the most common weighting method used to describe documents in the Vector Space Model, particularly in IR problems […]". After that, the algorithm produces what we call *weighted terms vector*. Based on this structure, the algorithm decides whether the message is closer to the concept/category A or B present in the subject taxonomy. This decision is made based on the relationship between the cosine similarity technique applied to the *weighted terms vector* of each category with the same method applied to *weighted terms vector* of the processed message.

The text message categorization algorithm has the following steps and uses the classical approach of calculating the cosine of vectors on the Vector Space Model (SALTON, 1975) (Categorization component in Figure 3).

1. Generates a vector of TFIDF message values, as follows:

a) For each term:

i) Computes the term frequency (TF) within the message,

ii) Computes the inverse term fre quency (IDF) considering the number of existing categories,

iii) Stores the product of these two va lues in another vector (TF * IDF);

2. For each category:

a) Generates TFIDF category vector (similar to the item 1.a) and then com putes the cosine similarity between TFIDF category vector and TFIDF mes sage vector,
b) Stores the value calculated in a vector of similarity measures.

After executing these steps, the Categorization component calculates the similarity degree between the posted message and the categories defined in the taxonomy. The last and most important step is then to decide which categories are associated with the message. This decision is made by comparing the obtained similarity degree with a cutoff point, identified during the implementation of the case study.

When the algorithm categorizes a message it creates an entry to the Expert Ranking adding up to the score of its author (students or any other environment user) (see Figure 2). This score will be useful for further queries, thus forming the rankings of experts. During this process, the Categorization Log table is also updated with data related to the posted message, such as the number of typed words, and its category. The algorithm analyses the number of author typed words as a tiebreaker indicator in the expert definition process. Once the ranking process has finished, the Expert Ranking table contains the information about the most active participants and the experts are displayed for the FindYourHelp user.

The algorithm used by the Categorization component takes into account the main terms that represent a category, and their existence/ frequency within the text in the post. However, it is possible that a text describes terms that are related to more than one category directly, in this case, the score for the post author will consider all categories i.e. a single message can score in more than one category.

In our case study, we observed that correctly categorized messages always reach a similarity score greater than reported value and the mistakenly categorized messages always scored lower. So, it is important to mention that, to categorize a message, the similarity value between a message vector of terms and its category vector should be greater than or equal to a cutoff value identified during the case study.

3.1.4 FindYourHelp user interaction

As can be seen in Figure 2, FindYourHelp user interaction includes: a) defining categories hierarchy (only users with teacher or administrator profile can perform this task) b) validating information, concerning automatically identified experts, extracted by the tool (done by teachers who are responsible for each subject), and c) assembling a list of visualization experts, grouped by subject and based on a score for each posted message.

FindYourHelp User interaction starts with the definition of the categories hierarchy where only users with teacher or administrator profile can perform this task. We can see an example of this hierarchy in Figure 4. Also in this figure we can see that the user has two options to construct the discipline taxonomy: (i) *Add New Root Category* when he/she starts the creation of taxonomy, adding the root category (ii) *Add Child Category or Topic* when he/she wants do add a child category to a preselected parent category. Each category has a name and an optional description. Figure 5 shows the form for filling information on the categories.

In order to provide credibility to the results presented by the tool, a human assisted validation functionality was developed (Figure 6). Only users with teacher or administrator profile can access this functionality, which serves as a complement to the information extracted automatically by FindYourHelp

Porto Alegre, v.14, n.2, jul./dez. 2011.
ISSN impresso 1516-084X ISSN digital 1982-1654

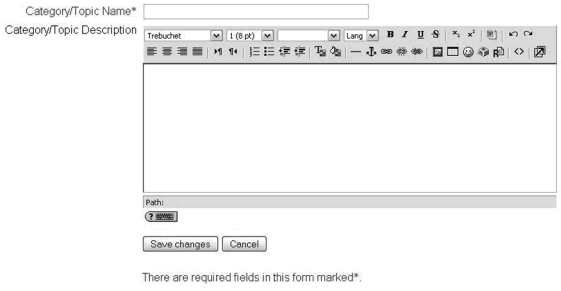

FIGURE 5 – Category Form
SOURCE: Prepared by authors.

module. We recommend that this validation should be performed after many interactions in forum as the teacher thinks it is sufficient to identify the experts. Figure 6 shows how the user can accept or discard a participant automatically identified as an expert by the tool. When the data on this screen is confirmed, accepted students receive greater weighting on the list of experts.

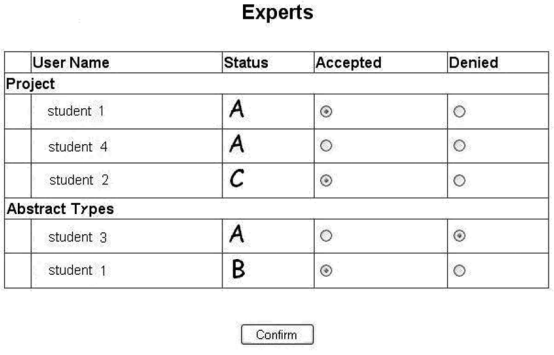

FIGURE 6 – Expert validation screen
SOURCE: Prepared by authors.

The experts' visualization feature adopts an idea based on the knowledge tree proposed by Lèvy (LÈVY, 2001). In this approach, a tree containing the categories is generated in the initial screen and if the user selects one category, among the existing ones, a list containing the experts on a specific issue is presented. In such a list, the experts are separated into three groups A, B and C (see Figure 7). These groups divide the participants as follows:

Group A: Participants who most contributed on the topic selected by the user to date. The criterion adopted for this grouping: participants with scores higher than 90% from the highest computed score for a specific issue.

Group B: Participants who contributed moderately, so far, compared to Group A, on the selected topic. The criterion for this grouping is: participants with scores higher than 70% and less than or equal to 90% of the highest computed score for a specific issue.

Group C: Participants who contributed less significantly when compared to groups A and B on the selected topic to date. In this grouping participants with scores higher than 50% and less than or equal to 70% from the highest computed score for a specific issue are listed.

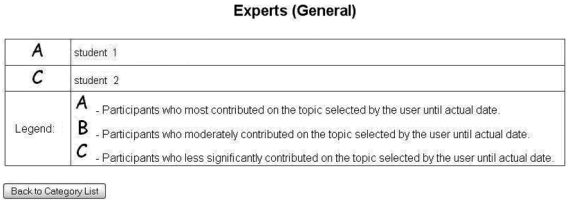

FIGURE 7 – Expert's List grouped by category
SOURCE: Prepared by authors.

4 Case Study

During the development of this work, we ran a case study of the FindYourHelp module in to verify its functionalities and measure the following aspects of the plug-in:

1. Number of messages correctly categorized automatically (analysis if its content really matches the category identified);

2. Number of correctly discarded messages by the algorithm (analysis if its content does not match any predefined category);
3. Evaluate if the teachers agree with the indication produced by FindYourHelp e.g. if the identified experts are in fact experts according to teacher perception.

4.1 Data Collection

The objects of study in our case study are three undergraduate courses: Object Oriented Programming Language I, Advanced System Design II and Interactive Technologies Applied to Education. All of them are traditional offline courses that use discussion forums and online resources to support the interaction among students. Table 1 shows the period and the number of messages posted in each course. We analyzed 325 messages in total and to get better reliability of the analysis, these messages were added to the database environment in the same order they actually happened.

TABLE 1 – Courses analyzed by the feasibility tool

Undergraduate Course	Period	Number of Posted Messages
1 –Object Oriented Programming Language I	18/08/2009 to 10/12/2009	32
2 – Advanced Systems Design II	07/02/2009 to 18/06/2009	76
3 – Interactive Technologies Applied to Education	22/01/2004 to 31/01/2004	217

SOURCE: Prepared by authors.

During the study, we noted that in the two first courses, there were no many messages posted in the discussion groups about the content of the course it self, most of the messages argued unimportant aspects of the course, such as: new task dates, news about absences or evaluations, and so on. We then decided to analyze a third course to perceive the interest of the group in discussing aspects of the course content given by the teacher.

Each course's teacher collaborated to build up the categories' hierarchy related to their course in the FindYourHelp and this also helped to get their feedback about the tool usage afterwards. Figure 8 shows an example of a hierarchy of terms that comprises the main subjects of the course Object Oriented Programming Language I, defined together with the teacher responsible.

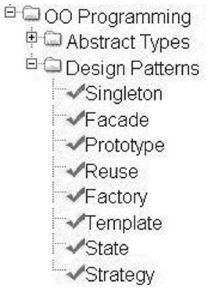

FIGURE 8 – Hierarchy of terms in one of the analyzed courses
SOURCE: Prepared by authors.

It is important to note that this process is very important for the algorithm behavior, specially when it comes to categorizing or dismissing a message, because this hierarchy generates a vector of terms for each category, which will be compared with the vectors of every posted message by applying the cosine similarity technique in Vector Space Model proposed by Salton (1975).

4.2 Analysis of the Forum Messages

The categorization algorithm was run twice for each subject during the study, because we

Porto Alegre, v.14, n.2, jul./dez. 2011.
ISSN impresso 1516-084X ISSN digital 1982-1654

noticed that after the first run the algorithm related messages with similarity values too small for some categories caused a large percentage of category error messages (around 10% in one subject, 20% in subject 2 and 15% in the third).

We then noted, through a random sampling of messages, that the similarity value of correctly categorized messages was always greater than 0.2, while those for messages categorized with error were generally below this.

The algorithm was then modified to consider a cutoff of similarity greater than or equal to 0.2 and after the second run on every message posted in the forums, we noticed the following results summarized in the table 2.

The messages were categorized correctly in more than 87% of the messages in all three courses. During the study, however, we realize that most errors found in the course with less accuracy (course 2) were related to messages with source code in its content. This type of message was not expected as input for the algorithm, and so words were concatenated and pits were removed improperly.

A strong point to be emphasized in the proposed solution algorithm is the fact that it managed to discard irrelevant messages with a degree of accuracy consistently above 94%. During the interviews with the teachers for each subject, we came up with a possible solution to this problem, to use of terms consisting of more than one word to characterize a category. Some terms such as: *Discussion List* or *Abstract Classes*, for example, have a greater meaning to the categories to which they were associated when analyzed because it was only one term instead of 2 separate. Statistically, these words may appear alone in messages with other semantic connotations, and in this case, they would be erroneously contributing to relate a person to a category.

4.3 Interviews with Teachers

This research had the collaboration of three teachers; they created the hierarchy of category in their disciplines and analyzed each message that was categorized or rejected by the tool during the message post. Two teachers are computer scientists and the third teacher is a Bachelor in Education, however, the course is about technology for interactivity. We interviewed these teachers and asked

TABLE 2 – Comparative results for subjects

	Subject 1	**Subject 2**	**Subject 3**
Central Theme	Programming Language	Systems Design	Education and Technology
Total of Participants	12	33	31
Total of Messages	32	76	217
Year	2009	2009	2004
Duration	Around four months	Around four months	Nine days
Categorized Messages	OK = 12 (92%) ERROR = 1 (8%)	OK = 35 (87%) ERROR = 5 (13%)	OK = 118 (90%) ERROR = 13 (10%)
Discarded Messages	OK = 19 (100%)	OK = 34 (94%) ERROR = 2 (6%)	OK = 80 (96%) ERROR = 3 (4%)
Does teacher agree with the algorithm?	Yes, Completely	Yes, Completely	Yes, Completely

SOURCE: Prepared by authors.

them about their impressions of the tools. Theses interviews were supported by a questionnaire (See Appendix A and B).

In this interview, everyone thought interaction with the tool to be very simple. One of them recommended an improvement so as to copy and move terms from one category to another and the others recommended adding terms of more than one word in the hierarchy.

Two teachers considered the FindYourHelp as a very reliable tool given the results and one of them classified it as reliable. Consequently, it is important to emphasize the unanimity among the teachers that the tool correctly identified which students were specialists in their groups.

5 Discussion

One of the most visible contributions of this work is to provide to the academic community an alternative to the automatic search of participating experts in specific issues within a VLE. The solution described in this paper is open source and provides an analysis of postings made in VLE discussion forums. This work also advances the state of the art in VLE by developing methods for automatically categorizing messages according to the degree of similarity between these and the categories defined by the teachers of each course, with an emphasis on the technique of cosine distances.

For the Brazilian community this work adds an analysis of the applicability of a thesaurus available for Portuguese (OPENTHESAURUS, 2009). This thesaurus is used to reduce the size of the set of the most significant terms of messages from the forums.

There are performance limitations in FindYourHelp that impacts the performance of the algorithm. The tool was developed in PHP and as PHP has no native support for parallel processing threads of execution (multithreading) and so we could not improve the execution time of the tool by using parallel processing.

Another limiting factor is the inability to analyze a greater diversity of courses in Moodle, as not all institutions allow access to this information. Our next goal is to use the plug-in in some courses in progress. We intend to run a more controlled experiment to analyze the impact of the use of the plug-in in the motivation of the students to participate in the course and help others.

Finally, the solution proposed in this paper was satisfactorily accepted and evaluated by the teachers involved with the courses in the case study, which indicates that the solution may be found useful in other situations and other institutions.

6 Conclusion

This article demonstrated the application of mining techniques in discussion forums as a way of supporting the search for expert participants by examining the content of their messages. The use of discussion forums allows participants to reflect better on what is posted in the group, which creates a tendency to improve the quality of messages conveyed in this kind of tool compared to messages written in a chat, for example (MOORE; KEARSLEY, 2004). Indeed, this leads us to review the content of this type of message.

The proposed solution aims to give visibility to the most active participants of the forum, allowing other participants to have faster access to suggested contacts to solve their problems. Teachers can also use such information as a support to their teaching in each classroom.

The design of FindYourHelp was explained and a preliminary analysis of its feasibility was

Porto Alegre, v.14, n.2, jul./dez. 2011.
ISSN impresso 1516-084X ISSN digital 1982-1654

provided by the application of a case study involving three different courses and their teachers. The tool meets its initial goals and was positively evaluated by the teachers participating in the study.

Some possible improvements identified in this study will be fixed in a future version, such as allowing the use of terms consisting of more than one word in the hierarchy of categories and increasing the thesaurus related terms written in Portuguese with English synonyms. Some related work shall also be investigated to improve the algorithm for categorization of messages, eg. test the remover of suffixes (PORTER, 1980) compared with the proposal of (COELHO; RENGO; BURIOL, 2009) applying fuzzy logic in the algorithm decision about which categories are closer to the posted message.

References

BALOG, K.; RIJKE, M. Determining Expert Profiles (With an Application to Expert Finding). In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE – IJCAI'07, 20., 2007, Hyderabad, India. *Proceedings*. [S.l.: s.n.], 2007. P. 2657-2662. Disponível em: <http://www.ijcai.org/papers07/Papers/IJCAI07-427.pdf> Acesso em: 07 jan. 2010.

BECERRA-FERNANDEZ, I. **Searching for experts on the Web: A review of contemporary expertise locator systems.** *ACM Transactions on Internet Technology* (TOIT), New York, v. 6, n. 4, p.333-335, nov. 2006.

COELHO, A.R.; ORENGO, V.M.; BURIOL, L.S. *Removedor de Sufixos da Língua Portuguesa*: RSLP. Porto Alegre: UFRGS/IF, 2007. Disponível em: <http://www.inf.ufrgs.br/~arcoelho/rslp/html/index.html> Acesso em: 20 nov. 2009.

DEERWESTER, S.; DUMAIS, S.T.; FURNAS, G.W.; LANDAUER, T.K.; HARSHMAN, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, Washington, v. 41, n. 6, p. 391-407, 1990.

DEMARTINI, G. Finding Experts Using Wikipédia. In: INTERNATIONAL EXPERTFINDER WORKSHOP, 2., 2007, Busan, Korea. *Proceedings*. [S.l.: s.n.], 2007, p. 33-41.

FARIA, E.T. *et al*. (Org.). *Educação Presencial e Virtual*: espaços complementares essenciais na escola e na empresa. Porto Alegre: EDIPUCRS, 2006.

HEEREN, F.; SIHN, W. XPERTFINDER: Message analysis for the recommendation of contact persons within defined topics. In: AFRICON CONFERENCE IN AFRICA, IEEE AFRICON, 6., 2002, George, South Africa. Proceedings. [S.l.]: IEEE, 2002. V. 1, p. 41-46.

JUNG, H.; LEE, M.; KANG, I.-S.; LEE, S.-W.; SUNG, W.-K. Finding Topic-centric Identified Expertsbased on Full Text Analysis. In: INTERNATIONAL EXPERTFINDER WORKSHOP, 2., 2007, Busan, Korea. *Proceedings*. [S.l.: s.n.], 2007, p. 56-63.

LANDAUER, T.K.; FOLTZ, P.W.; LAHAM, D. Introduction to Latent Semantic Analysis. *Discourse Processes*: a multidisciplinary journal, Norwood, n. 25, p. 259-284, 1998.

LÉVY, P. *Ciberculture*. 1. ed. Minnesota: University of Minnesota Press, 2001.

MANNING, C.D.; PRABHAKAR, R.; SCHÜTZE, H. *An introduction to information retrieval*. Cambridge: Cambridge University Press, 2009.

MATTOX, D.; MAYBURY, M.; MOREY, D. *Enterprise Expert and Knowledge Discovery*. [S.l.: s.n.], 2010. Disponível em: <http://www.mitre-corporation.org/work/tech_papers/tech_papers_00/maybury_enterprise/maybury_enterprise.pdf > Acesso em: 22 jan. 2010.

MAYBURY, M.T. *Expert Finding Systems*. Bedford: MITRE, 2006. (Mitre Technical Repport, MTR 06B000040)

MAYBURY, M.T.; D'AMORE, R.; HOUSE, D. Expert Finding for Collaborative Virtual Environments. *Communication of the ACM*, New York, v. 44, n. 12, p. 55-56, dec. 2001. Disponível em: <http://doi.acm.org/10.1145/501338.501343> Acesso em: 22 jan. 2010.

MOODLE. *Modules and Plugins*. S.l.: MOODLE, 2008. Disponível em: <http://moodle.org/mod/data/view.php?id=6009> Acesso em: 21 jan. 2010.

MOORE, M.G.; KEARSLEY, G. *Distance Education*: A Systems View. 2. ed. Belmont, California: Wadsworth Publishing, 2004.

NONAKA, I.; TAKEUCHI, H. *The Knowledge-Creating Company*: How Japanese Companies Create the Dynamics of Innovation. Oxford: Oxford University Press, 1995.

PORTER, M.F. An algorithm for sufix stripping. *Program*, v. 14, no. 3, p. 130-137, july 1980. Disponível em: <http://telemat.die.unifi.it/book/2001/wchange/download/stem_porter.html> Acesso em: 16 jan. 2010.

SALTON, G.; WONG, A.; YANG, C.S. A Vector Space Model for Automatic Indexing. *Communications ot the ACM*, New York, v. 18, n. 11, p. 613-620, nov. 1975.

SANTOS, M.L.; SALVADOR, L. do N. FindYourHelp: um módulo de busca por especialistas no ambiente Moodle. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 20., 2009, Florianópolis. *Anais*. Florianópolis: Biblioteca Universitária da UFSC, 2009. Disponível em: http://wwwexe.inf.ufsc.br/~sbie2009/anais/artresumidos.html Acesso em 05 out. 2010.

SANTOS, M.L.; SALVADOR, L. do N.; CRUZES, D. FindYourHelp*:* an expert search module on Moodle. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 21.; 2010, João Pessoa. *Anais*. João Pessoa: Sociedade Brasileira de Computação, 2010. V. 1, Disponível em: http://www.ccae.ufpb.br/sbie2010/anais/Artigos_Completos_files/76138_1.pdf Acesso em 05 out. 2010.

Porto Alegre, v.14, n.2, jul./dez. 2011.
ISSN impresso 1516-084X ISSN digital 1982-1654

SEO, J.; CROFT, W.B. Thread-based Expert Finding. 2009. Paper presented in SIGIR 2009 Workshop on Search in Social Media – SSM, 2009, Boston, Massachusetts. Disponível em: <http://maroo.cs.umass.edu/pub/web/getpdf.php?id=893> Acesso em: 16 jan. 2010.

SIM, Y.W.; CROWDER, R.M.; WILLS, G.B. *Expert Finding by Capturing Organisational Knowledge from Legacy Documents*. 2006. Paper presented in IEEE International Conference on Computer & Communication Engineering (ICCCE '06), 2006, Kuala Lumpur, Malaysia. Disponível em: <http://eprints.ecs.soton.ac.uk/12509/> Acesso em: 20 jan. 2010.

WANG, J.Z.; TAYLOR, W. *Concept Forest*: A New Ontology-assisted Text Document Similarity Measurement Method. In: IEEE/WIC/ACM INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE, 2007. *Proceedings*. [S.l.]: IEEE, 2007. P. 395-401.

YANG, K.-H.; CHEN, C.-Y.; LEE, H.-M.; HO, J.-M. EFS: Expert Finding System based on Wikipedia Link Pattern Analysis. In: IEEE INTERNATIONAL CONFERENCE ON SYSTEMS, MAN AND CYBERNETICS, 2008, Singapure. *Proceedings*. [S.l.]: IEEE, 2008. P. 631-635.

This work was supported by Instituto Nacional de Ciência e Tecnologia para Engenharia de Software (INES) with financial support of CNPq and FACEPE under the projects 5773964/2008-4 and APQ-1037-1.03/08.

Recebido em 22 de junho de 2010.
Aprovado para publicação em 05 de outubro de 2010.

Marcos Lapa Santos
Universidade Salvador – UNIFACS, Salvador/Ba – Brasil. E-mail: marcoslapa@gmail.com

Laís Nascimento Salvador
Universidade Federal da Bahia – UFBA, Salvador/BA – Brasil. E-mail: laisns@dcc.ufba.br

Daniela Soares Cruzes
Norwegian University of Science and Technology – NTNU, Trondheim/Noruega. E-mail: danielascruzes@gmail.com

APPENDIX A – Questionnaire 01 – Experiences with VLE

1. Please report your experience with teaching, considering the following questions: how much experience do you have in teaching, when did you graduate, which course?

2. In some point in your education (undergraduate, graduate, master's, doctoral or postdoctoral), did you have contact with a technology which facilitated the knowledge building in virtual spaces, for example have you participated in collective knowledge building in discussion forums?

3. Do you have experience with VLE? How long do you work with VLE?

4. In your teaching experience, do you use educational resources such as forums, chats, blogs, among other virtual tools? Why?

5. In case of affirmative on question 4, please answer:

a. What are the benefits and difficulties did you perceive during such interaction in the context of the teaching-learning process?

b. Regarding the mailing lists, discuss mediation strategies used to incentive the participation of students.

Porto Alegre, v.14, n.2, jul./dez. 2011.
ISSN impresso 1516-084X ISSN digital 1982-1654

APPENDIX B – Questionnaire 02 – Perceptions on The FindYourHelp

Name:

What is your background?

What is the name (s) Institution (s) of Higher Education (s) which now you have some teaching duties?

Which courses do you teach?

Please report your experience in the use FindYourHelp. Please focus on the assembling of the hierarchies of categories/terms in your course. Was the interaction with the tool satisfactory? Do you have any suggestions of improvement to this process?

On the automatic identification of experts. What do you think about the degree of accuracy of the tool?

Excellent (High) = 5, Very Good = 4, Satisfactory = 3, Fair = 2, Poor (Low) = 1

On the analysis done on each individual message posted (categorization and disposal), how much did the tool categorize right?

() 1 - No time () 2 - Few times () 3 - Moderately () 4 - Often () 5 - Always

In general, how reliable was the result of the process imposed by FindYourHelp?

() 1 - Unreliable () 2 - Low reliability () 3 - Dependable () 4 - Very Reliable () 5 - Completely reliable

In general, what are your recommendations for improvements on the usage of the FindYourHelp?