

FindYourHelp: an expert search module on Moodle

Marcos L. dos Santos¹, Laís do N. Salvador², Daniela S. Cruzes³

¹Engineering and Technology Area – University Center of Bahia (Estácio-FIB)
Salvador – BA – Brazil

²Computer Science Department – Federal University of Bahia (UFBA)
Salvador – BA – Brazil

³IDI-NTNU, NO-7491, Trondheim – Norway

marcoslapa@gmail.com, laisns@dcc.ufba.br, dcruzes@idi.ntnu.no

***Abstract.** This paper presents FindYourHelp, a solution that enables automatic identification of experts who have most contribute into Virtual Learning Environments discussion forums. The proposal is based on applying text mining techniques as a supplementary analysis of student's participation. Some technical details are discussed, as well as, a feasibility study of such solution.*

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 and SALVADOR, 2009).

Searching for experts who have 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. Expert Finding is the area of research that addresses the task of finding the right person with the appropriate skills and knowledge (BALOG and RIJKE, 2010).

Our goal is to analyze the adoption of algorithms and techniques of text mining as a means of supporting the search for experts in research/academic discussions groups. This work discusses, therefore, the creation of a module which 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 organized as follows. Section 2 presents the area of search engines by experts and text mining. 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. Expert Search Engines

The literature shows that many search engines have been developed. According to Jung, (et al 2007, p. 56) “sources to find experts are various documents, programs, emails, databases, quotes, communities, among others”. Maybury (2006) 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 expert search engines 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.








		Expert Finding Tools																						
		Sources				Processing			Search		Results		System											
		Self declaration	Email	Documents	Briefings	Resumes	Web pages	Databases	Behavior/searches	Ranking	Entity Extraction	Social Net Analysis	Foreign Language (#)	Author Identification	Keyword	Boolean	Natural Language	Taxonomy (Browse)	List of Experts	Related Documents	Related Concepts	Interoperability	In Operational Use	Privacy
PRODUCT																								
	TACIT	Full	Full	Full	Full	Full	Full	Full	Full	Partial	Partial	Partial	2	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full
	AskMe	Full	Full	Full	Full	Full	Full	Full	Full	Partial	Partial	Partial	70	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full
	Autonomy	Full	Full	Full	Full	Full	Full	Full	Full	Partial	Partial	Partial	250	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full
	Endeca	Full	Full	Full	Full	Full	Full	Full	Full	Partial	Partial	Partial	200	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full
	Recommind	Full	Full	Full	Full	Full	Full	Full	Full	Partial	Partial	Partial	6+	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full
	Trivium	Full	Full	Full	Full	Full	Full	Full	Full	Partial	Partial	Partial	6	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full
	Entopia	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full	6	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full

Figure 1. Search Engine for experts (Maybury, 2006, p. 18)

As can be seen in Fig. 1, most listed tools offer considerable support to the diverse sources of expert search, however, they differ widely when it comes to the support level provided to processing. It's also possible to note that queries based on “keywords” and “Boolean queries” are predominant among tools. 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.

Finally, is important to note that none of these tools uses the idea of analyzing messages posted in discussion forums to infer what their authors know and whether they interact more thus identifying their specialties.

2.1 Text Mining for Expert Finding

The text mining task used for expert finding is a usual research field. Many authors have applied text mining techniques for this purpose. Although Text Mining is a wide area, this paper will discuss only some techniques for text categorization that can be applied for expert finding. 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 key-word based document search which is what we are interested in this research.

The Latent Semantic Indexing is a specific method from the Latent Semantic Analysis (LSA) area used in expert find tasks, described by Heerem and Sihm (2002, p.43): “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”.

The VSM is another method commonly used in text categorization, and it was the selected method in this research because its simplified approach and adherence to the analysis of short messages, the focus of our proposed tool. The VSM uses the same principle as LSA, but it does not perform the semantic analysis step. It has a disadvantage related to handle polissemey and synonym which must be treated with auxiliary structures, i.e. thesauri and dictionaries.

3. FindYourHelp

This work proposes the creation of an expert search system to operate in discussion groups within VLE, entitled FindYourHelp. We intend to enhance approximation among discussion forums participants by pointing out possible experts in some subject (or matter) of group interest. FindYourHelp approach is based on the forum messages content analysis in order to identify the participants that have most contributed in some subject inside the group. A general view of the system operation is shown in Figure 2:

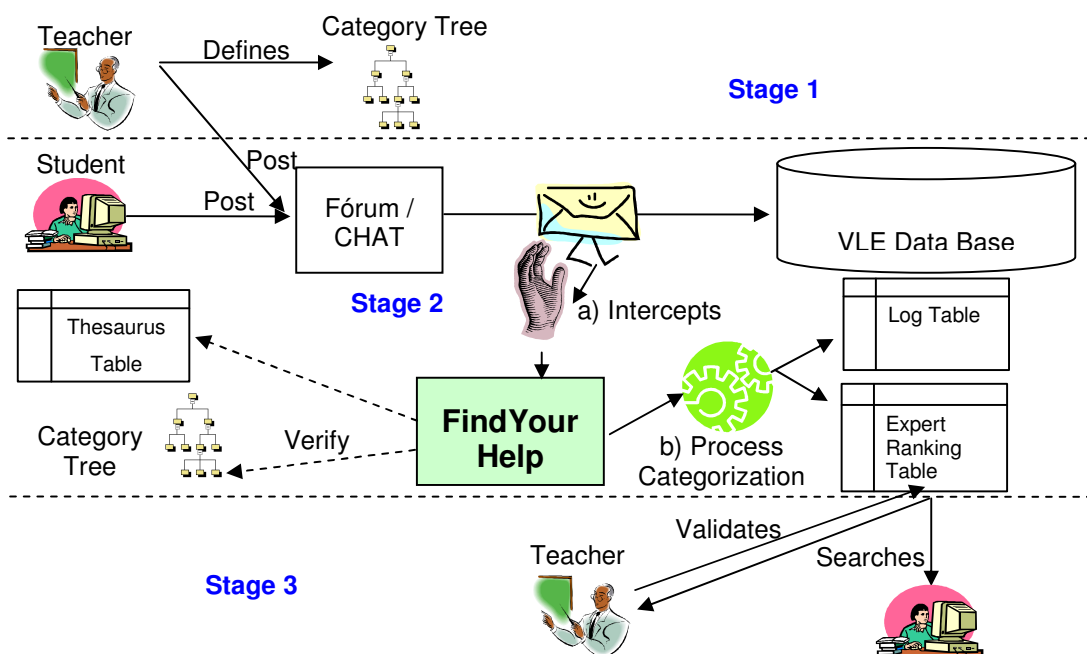


Figure 2. FindYourHelp Operation Overview.

The main idea is to use FindYourHelp 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.

As we can see in Figure 2, The FindYourHelp module comprises three main operation stages in the use of VLE: the first is the category hierarchy definition; the second is the message 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.

- (i) To evaluate the posts better, we use a categories tree 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 future comparison with the messages that will be posted on the VLE.
- (ii) After the forums message posting, 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 categories tree and saves the ranking and logging information.
- (iii) These reviews provide information for the ranked expert list implementation supported by a human assisted validation (that will be discussed 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 post content, focusing only on collecting the statistics of number of posts made by users. So, FindYourHelp is a plug-in added to Moodle environment that support the find expert task inside the discussion forums messages. See its component architecture in Figure 3:

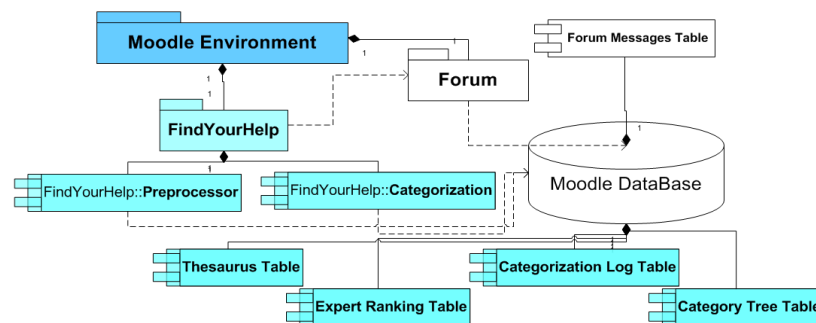


Figure 3. FindYourHelp Architecture.

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.

3.1.1 Environment Setup

For the proper functioning of the categorization, the module needs to be configured. Two tables, added to Moodle data base, are setup initially: Thesaurus and Category Tree.

The Thesaurus is an auxiliary table that optimizes the processing of words with similar meanings in the text. This table was populated with imported data from OpenThesaurusPT project (OPENTHESAURUS, 2009). This project maintains a dictionary that lists words which have 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. a 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 oriented object programming:



Figure 4. Terms hierarchy for a game development example.

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 Preprocessor that: (i) removes numbers, symbols and punctuation; (ii) generates message “tokens”; (iii) removes stopwords; (iv) applies the process Stemming (reducing terms to its common radical), in this case the algorithm proposed by (COELHO, 2009) and then (v) 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 executed, the text mining algorithm generates a bag of words – a terms vector – associated to the analyzed message. Following this, 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) to assign weighted values to each term vector related to the message content. After that, the algorithm produces “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):

- (i) 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 frequency (IDF) considering the number of existing categories; (iii) Stores the product of these two values in another vector (TF * IDF)
- (ii) For each category: (a) Generates TFIDF category vector (similar to the item 1.a) and then computes the cosine similarity between TFIDF category vector and TFIDF message vector and (b) Stores the value calculated in a vector of similarity measures.

After executing these steps, the Categorization component obtains the similarity degree between the posted message and the categories defined in the taxonomy. Now this component can execute its last (or most important) step which is to decide which categories are associated to the message by comparing the obtained similarity degree with a cutoff point, identified during the implementation of the case study.

When a message is categorized by the algorithm, its author builds up a score in the Expert Ranking table (see Figure 3) which will be useful for further queries, thus forming the expert ranking. 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 (see section 3.1.4).

The algorithm used by the Categorization component takes into account the main terms that represent a category, and their existence/frequency within the post text. 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.

3.1.4 FindYourHelp User Interaction

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, which is 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.

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.

In order to provide credibility to the results presented by the tool, a human assisted validation approach for each specialist was created. This functionality can be accessed only by users with teacher or administrator profile and serves as a complement to the information extracted automatically by FindYourHelp module. We recommend that this validation be done after as many interactions in forum as the teacher thinks sufficient to identify the experts, so its moment vary by the discipline rhythm.

Experts

User Name	Status	Accepted	Denied
Project			
student 1	A	<input checked="" type="radio"/>	<input type="radio"/>
student 4	A	<input type="radio"/>	<input type="radio"/>
student 2	C	<input checked="" type="radio"/>	<input type="radio"/>
Abstract Types			
student 3	A	<input type="radio"/>	<input checked="" type="radio"/>
student 1	B	<input checked="" type="radio"/>	<input type="radio"/>

Figure 5. Expert validation screen

As the screen in Figure 5 shows, the user can accept or discard a participant automatically identified as a specialist by the tool. When the data on this screen is confirmed, accepted students receive greater weighting against the denied on the list of experts.

The expert’s visualization feature will adopt an idea which is based on to knowledge tree (LÈVY, 2001). In this approach, a tree containing the categories is generated in the initial screen (see Figure 4) and if the user selects one existing category, 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. 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; participants with scores higher than 70% and less than or equal to 90% from 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.

4. Case Study

During the development of this work, we ran a case study of the FindYourHelp module in order to verify its function and measure the following aspects of the plugin: (i) Number of messages correctly categorized automatically (analysis if its content really matches the category identified); (ii) Number of correctly discarded messages by the algorithm (analysis if its content does not match any predefined category); (iii) Assess if participant scores match the evaluation of expertise from the point of view of the teacher responsible.

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.

Table 1. Subjects 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

We analyzed 325 messages in total (see Table 1) and to get better reliability of analysis these messages were added to the database environment in the order they actually happened. Each teacher collaborated to build up the hierarchy category related to their course inside the environment and this also served as to get their feedback about the tool usage afterwards. Figure 3 shows an example of a terms hierarchy that comprises the main subjects of the course Object Oriented Programming Language I, defined together with the teacher responsible.

It is important to observe that this process is very important for the algorithm behavior, especially 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

After the execution of categorization algorithm we noticed the following results:

Table 2. Comparative results for subjects

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%);	Ok = 35 (87%);	Ok = 118 (90%);

	Error = 1 (8%);	Error = 5 (13%);	Error = 13 (10%);
Discarded Messages	Ok = 19 (100%)	Ok = 34 (94%); Error = 2 (6%);	Ok = 80 (96%); Error = 3 (4%);
Does the teacher agree with identified experts?	Yes, Completely	Yes, Completely	Yes, Completely

As can be seen the categorization of posts hit a percentage greater than or equal to 87% in all subjects. During the study, however, we realize that most errors found in the subject with less accuracy (subject 2) were related to messages with source code in its content. This had not been foreseen by the algorithm, which in these cases came to concatenate words and remove punctuations 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%.

4.3 Interview 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 graduates in computing and the third graduated in Education (Teaching), however, in her subject she deals with areas of technology for interactivity.

Interviews were carried out with these teachers in order to analyze the FindYourHelp performance concerning to the messages relating to their subject and to get their impressions about the tool: 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.

Also during these interviews, we came up with a possible improvement to FindYourHelp performance, this was the 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.

5 Conclusion

The design of FindYourHelp was explained and a preliminary analysis of its feasibility was done by the application of a case study involving three different courses and their teachers. We found that the tool meets its initial goals and was positively evaluated by the teachers participating in the study.

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.

A 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

plugin in some courses in progress. With this we intend to run a more controlled experiment to analyze the impact of the use of the plugin in the motivation of the students to participate in the course and help others. Some problems identified by the solution feasibility 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 can be investigated to improve the algorithm for categorization of messages, eg. to test the remover of suffixes (PORTER, 2010) compared with the proposal of (COELHO, RENO and BURIOL, 2009) applying fuzzy logic in the algorithm decision about which categories are closer to the posted message.

References

- BALOG, Krisztian and RIJKE, Maarten. *Determining Expert Profiles (With an Application to Expert Finding)*. IJCAI'07: Proc. 20th International Joint Conference on Artificial Intelligence, Hyderabad/India, 2007 Disponível em: <http://www.ijcai.org/papers07/Papers/IJCAI07-427.pdf> Acesso em: 07 jan 2010
- COELHO, Alexandre R.; ORENGO, Viviane M.; BURIOL, Luciana S. *Removedor de sufixos da língua portuguesa: RSLP*. 2007 Disponível em: <http://www.inf.ufrgs.br/~arcoelho/rsrp/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, 41(6), 391-407. 1990
- GEROSA, M.A.; RAPOSO, A. B.; FUKS, H.; LUCENA, C. J. P. de. *Uma arquitetura para o desenvolvimento de ferramentas colaborativas para o ambiente de aprendizagem aulanet*. In: Anais do Simpósio Brasileiro de Informática na Educação – SBIE, Manaus – AM, 2004
- HEEREN, Frank and SIHN, Wilfried; *XPERTFINDER: Message analysis for the recommendation of contact persons within defined topics*, IEEE Africon 2002
- LÉVY, Pierre. *Ciberculture*, University of Minnesota Press, 1st edition, 2001
- MAYBURY, Mark T. *Expert Finding Systems*. Mitre Technical Report, 2006
- MOODLE, *Modules and Plugins*. Disponível em: <http://moodle.org/mod/data/view.php?id=6009> Acesso em 21 jan 2010.
- PORTER, M. F. *An algorithm for suffix stripping*. Disponível em: http://telemat.die.unifi.it/book/2001/wchange/download/stem_porter.html Acesso em: 16 jan 2010
- SALTON, G., WONG, A and YANG C. S. *A Vector Space Model for Automatic Indexing*. Communications of the ACM, New York, v.18, 1975
- SANTOS, Marcos L. and SALVADOR, Laís, do N. *FindYourHelp: um módulo de busca por especialistas no ambiente Moodle*, XX Simpósio Brasileiro de Informática na Educação, Florianópolis - SC, 2009
- WANG, James Z.; TAYLOR, William. *Concept Forest: A New Ontology-assisted Text Document Similarity Measurement Method*. in Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp 395-401, 2007