Enhancing Online Discussion Forums with Topic-Driven Content Search and Assisted Posting

Damiano Distante[1], Alejandro Fernandez[2], Luigi Cerulo[3] and Aaron Visaggio[3]

[1] Unitelma Sapienza University, Rome - Italy
damiano.distante@unitelma.it
[2] LIFIA, CIC/F.I., National University of La Plata, La Plata - Argentina
alejandro.fernandez@lifia.info.unlp.edu.ar
[3] University of Sannio, Benevento - Italy
lcerulo@unisannio.it,visaggio@unisannio.it

Abstract. Online forums represent nowadays one of the most popular and rich repository of user generated information over the Internet. Searching information of interest in an online forum may be substantially improved by a proper organization of the forum content. With this aim, in this paper we propose an approach that enhances an existing forum by introducing a navigation structure that enables searching and navigating the forum content by topics of discussion. Topics and hierarchical relations between them are semi-automatically extracted from the forum content by applying Information Retrieval techniques, specifically Topic Models and Formal Concept Analysis. Then, forum posts and discussion threads are associated to discussion topics on a similarity score basis. Moreover, to support automatic moderation in websites that host several forums, we propose a strategy to assist a user writing a new post in choosing the most appropriate forum into which it should be added. An implementation of the topic-driven content search and navigation and assisted posting forum enanchment approaches for the Moodle learning management system is also presented in the paper, opening to the application of these approaches to several realdistance learning contexts. Finally, we also report on two case studies that we have conducted to validate the two approaches and evaluate their benefits.

Keywords: Online discussion forums, information search, information extraction, text mining, topic modeling, navigability, searchability, assisted posting, e-learning, learning management systems, Moodle.

1 Introduction

Online discussion forums represent one of the main sources of user generated content (i.e., social media) and asynchronous communication means in the form of message posts over the Internet. Most visited websites, including blogs and social networks, use forums to support user interaction and knowledge sharing.

In several domains ranging from e-commerce [22][12], to news [18], and healthcare [27], discussion forums constitute rich and widely accessed repositories of information for Internet users.

As an example, software developers forums are an effective source of information where programmers search for and describe solutions to specific problems[4]. In e-learning contexts, discussion forums enable asynchronous communication student-to-student, and teacher-to-student, e.g., to support collaborative learning and group work [26][15]. Whatever the forum domain, discussions held in a certain period of time become a source of information for any user accessing the forum afterwards.

In general online forums organize messages into a chronological order. A user starts a new discussion by posting an initial message, other users post their replies or comments to it, and the list of messages forms a *discussion thread*. If users are allowed to reply to other users' replies in additional to the original message, discussions take the form of trees, with discussion branches.

The effectiveness of a discussion forum as information source mainly depends on the forum richness in information, but also on the forum organization and on the searching paradigm users can adopt to find contents of their interest.

Search features usually provided with online discussion forums are limited to full-text search which returns a list of forum messages that include (and/or do not include) one or more of the query keywords in their body and/or their title. Such a search feature may return too many or too few results (depending on the forum size and the query keywords) and may miss messages which are semantically related to the query keywords but do not actually include them [1].

Hierarchical graphs constitute an effective paradigm to represent users' knowledge [34]. In a previous work [7] we have introduced an approach to improve information retrieval and content navigation in online discussion forums by introducing in them a complementary hierarchical topic-driven navigation structure. Information Retrieval (IR) techniques, specifically Topic Models [4] and formal concept analysis (FCA) [10], are used to discover discussion topics and hierarchical relations between them in the forum content. Then, forum messages and discussion threads are associated to discussion topics based on a similarity score, thus to enable searching and navigating them on a topic-driven basis, additional to conventional chronological order and full-text search approaches.

In this paper we present an implementation of this approach as a plugin for the Moodle learning management system which makes the topic-driven navigation approach accessible and evaluable in several e-learning contexts. We also present a case study that provides a first qualitative assessment of the benefits of topic-driven navigation and search of forum content, with respect to traditional full-text search.

There are scenarios, such as large communities of interest or learning spaces, that call for the organization of interactions into multiple forums. Creating multiple forums aims at making each of them more focused, and manageable in

[4] An example of such forum is the Microsoft MSDN Developer Network forum. http://social.msdn.microsoft.com/Forums/en/categories/

terms of frequency of updates. The Moodle English Community[5], for example, organises discussions in 57 forums. Even if the conversation space is split in multiple forums, some topics can be cross-cutting. For example, "Scorm", the Sharable Content Object Reference Model, is the main focus of the "Scorm" forum[6] of the Moodle English Community. However, the topic is also discussed in the "Comparisons and advocacy", "General Help", and "General Developers" forums, among others. Topic-driven navigation helps users discover and navigate such connections.

In these scenarios, the lack or absence of moderation normally results in a poor organization of the forums content, with duplication of content and difficulties in finding relevant information among them. The organization of content in forums is the outcome of the choices users make when posting new messages. They choose the forum to post to, and decide whether to start a new discussion or contribute to an existing one. As the result of these choices, forums turn more or less cohesive, and topics become more or less scattered. To mitigate this problem, in this paper we also propose a strategy to assist a user while writing a new message (i.e., creating a new discussion) in deciding in which forum to post it. We evaluate our strategy against forums of the Moodle public community[7] and against on-line discussions in Stack Exchange[8], a network of question and answer websites on diverse topics that counts with a community focused curation process.

This paper is a revised and extended version of our earlier work presented in [9]. Particularly, new contributions of this paper are: *i.* an extended version of the topic-driven forum enhancement approach that now allows analyzing several forums at once to build one single topic-driven navigation structure that spans all their content; *ii.* an approach to assist users in choosing the most appropriate forum/thread to which to post a new message, thus to support the automatic moderation of a forum; *iii.* the implementation of both approaches via our TDForum plugin for the Moodle learning management system.

The rest of the paper is organized as follows. Section 2 describes our approach, earlier introduced in [7], to enhance an existing forum with topic-driven content search and navigation capabilities. Section 3 discusses the approach to assist users in selecting the most appropriate forum in which to add a new post. Section 4 presents the implementation of both these approaches for the Moodle[9] learning management system. Section 5 reports on a case study conducted to qualitatively assess the benefits of the topic-driven forum enhancement approach in searching forums for information of interest for the user. It also reports on a experiment that assesses the performance of our assisted posting approach. Section 6 overviews related work, while Section 7 draws conclusions and introduces future works.

[5] `https://moodle.org/course/view.php?id=5`
[6] `https://moodle.org/mod/forum/view.php?f=365`
[7] `https://moodle.org/course/`
[8] `http://stackexchange.com`
[9] `www.moodle.org`

2 The Topic-Driven Forum Navigation Enhancement Process

The topic-driven forum navigation enhancement process, introduced by Cerulo and Distante in [7], is shown in Figure 1. It consists of four main steps represented in the figure as rectangles and described briefly in the following subsections.

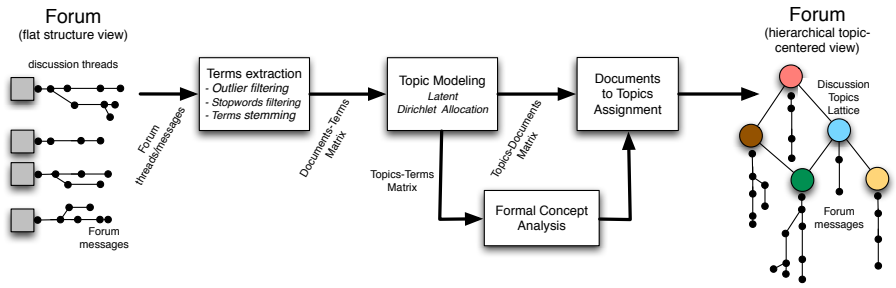

Fig. 1: The topic-driven forum navigation enhancement process [7].

2.1 Terms extraction

We represent a forum message as a vector of indexing terms, $\{t_1, \ldots, t_m\}$, extracted, from the corpus of n messages, $\{d_1, \ldots, d_n\}$, through a standard text analysis pipeline usually adopted in Information Retrieval that comprises: outlier filtering, stopwords filtering, and stemming [1].

The outcome of this step is a document-term matrix $\mathbb{DT}$, where each element $\{\mathbb{DT}\}_{jp}$ is the *tf-idf* of the term t_p in the forum message d_j [1].

2.2 Topic modeling

Topic modeling, in particular Latent Dirichlet Allocation (LDA), is a statistical technique that is able to extract frequently co-occurring terms, known as *topics*, from a corpus of documents [4]. The input is the document-term matrix, $\mathbb{DT}$, obtained from the previous task, while the output is a topic-document matrix, $\mathbb{TD}$, and a topic-term matrix, $\mathbb{TT}$. The number of topic k is a parameter that controls the granularity of the topics and must be fixed a priori.

Intuitively, the top terms of a topic are semantically related and represent some real-world concepts. For example the concept related to problems e-mails setup is represented by the terms “mail”, “problem”, “setup”. The topic membership of a document describes which concepts are present in that document. Table 1 shows and example of topic-document and topic-term matrices.

topic	topic-term (top terms)	topic-document d_1	d_2	d_3	d_4
z_1	*problem, email, setup*	0.6	0.7	0.1	
z_2	*problem, email, connection, setup*	0.3	0.1	0.1	0.5
z_3	*problem*	0.1			
z_4	*problem, video, decoder, setup*	0.1	0.2	0.8	0.5
z_5	*problem, video*	0.1	0.2		

Table 1: Examples of topic-term and topic-document matrices

2.3 Formal concept analysis

Using the topic membership of a term, we prune a topic lattice by means of Formal Concept Analysis (FCA). FCA is a computational way to derive a concept hierarchy or formal ontology from a collection of objects and their properties [10][3].

We model the topics as the objects of a formal context and the terms as their attributes. The relation R of the formal context is computed from the topic-term matrix $\mathbb{TT}$ by means of a decision threshold h_T, *i.e.*, a term (attribute) t_p belongs to a topic (object) z_i, $(z_i, t_p) \in R$ iif $\{\mathbb{TT}\}_{ip} \geq h_T$.

As a clarification example consider the formal context shown in Table 3 and the topic lattice obtained from such a formal context shown in Figure 2. Topics are mapped on circles and hierarchical relationships are represented by arcs. Large circles are mapped on topics extracted with the topic modeling approach, while small circles are intermediate topics extracted with the formal concept analysis. The lattice shows the hierarchical relationships between topics. In the lattice the top most topic (z_3) is the most general topic. A path starting from the top most topic is a more specific topic. For example z_2 is reachable by the path from z_3 (problem), setup, z_1 (email), and z_2 (connection), and represent the more specific topic of problems related to the email connection setup.

	problem	*email*	*connection*	*video*	*decoder*	*setup*
z_1	×	×				×
z_2	×	×	×			×
z_3	×					
z_4	×			×	×	×
z_5	×		×	×		

Table 2: The formal context obtained from the topic-term matrix shown in Table 1

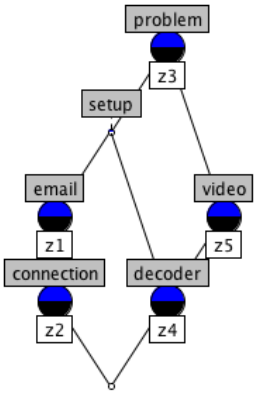

Fig. 2: The topic-lattice pruned from the formal context shown in Table 2.

2.4 Documents to topics assignment

During this step each document (*i.e.* forum message/thread) is mapped onto the topics to which it is more likely to belong by estimating the probabilities of each topic for that message (topic-document matrix). For this purpose we adopted the topic-document matrix $\mathbb{TD}$ and a decision threshold h_D, *i.e.*, a document d_j belongs to a topic z_i iif $\{\mathbb{TD}\}_{ij} \geq h_D$.

2.5 Parameter setting and accuracy evaluation

Selecting the number of topic k is one of the most problematic modeling choice in topic models [31]. We adopt a metric, introduced by Meilă [20] for clustering comparison, that measures the *Variation of Information* as the entropies the mutual information associated with cluster assignments. Intuitively, the entropy measures the uncertainty of allotting an item to a cluster, while the mutual information measures the reduction of such uncertainty when the allocation in the other cluster is known. Following the approach adopted by Wallach *et al.* [31] the assignment of documents (forum messages or threads in our context) to topics can be assimilated to a sort of cluster assignment. In our previous work [7] we showed that above a certain value of k no significant increment of Variation of Information can be observed in a specific context. We consider such a value the optimal number of topic in that context.

We evaluated the document assignment task to check whether the documents assigned to topics by the Latent Dirichlet Allocation were congruent with the semantics of their content [2]. In our previous work [7] we addressed this question with a controlled experiment obtaining in average a precision ranging between 52% and 74%.

3 Assisted Posting

To assist users in deciding where to submit their posts, we change the accustomed posting workflow. Instead of first deciding where to post and then writing the message, users first write and then let an algorithm suggest the most adequate forum. The algorithm ranks available forums according to some measure of relevance. The user can decide to submit the post to the highest ranked forum, or to a different one. To calculate the relevance of a post to a given forum we have explored two alternatives. The "One-like-this" approach assumes that the post is most relevant to the forum that contains the most similar post. The "Centroids" approach assumes that the post is most relevant to the forum that, as a whole, is most similar to the content of the post. To operationalise these alternatives, we looked into the theory of Vector Space Models [24].

3.1 Vector Space Models: One-like-this and Centroids

As explained in Section 2.1, the extraction of terms from posts results in a matrix $\mathbb{DT}$, where each element $\{\mathbb{DT}\}_{jp}$ is the *tf-idf* of the term t_p in the forum message d_j. Thus, row $\{\mathbb{DT}\}_j$ in the matrix, represents the *tf-idf* vector for forum messaje d_j. The messages in a forum constitute as a space of vectors where similar posts have similar vectors (according to some similarity function such as Cosine Similarity). Both strategies we propose to assess the relevance of a post to a forum build upon this model. They therefore require, as the first step, the construction of the *tf-idf* vector $\{\mathbb{DT}\}_n$ for the new post d_n.

To find the forum that contains the post most similar to the one the user attempts to submit (One-like-this), we compute the Cosine Similarity between vector $\{\mathbb{DT}\}_n$ and the *tf-idf* vectors of all messages in available forums. For each forum we take the similarity coefficient of the most similar post. Forums are ranked according to these coefficients, suggesting the user to post to the one with the highest value. Although this strategy is intuitive and straightforward to implement, it incurs in high computation cost for large forums.

We define similarity between a post and a forum in terms of the Cosine Similarity between the post's *tf-idf* vector and the centroid of the latter. During the indexing process, we compute the centroid of each forum by taking that average between the *tf-idf* vector of all its messages. To find the most similar forum to the new post d_n we compute the similarity coefficients between the $\{\mathbb{DT}\}_n$ vector and the centroids of all available forums. Forums are ranked according to these coefficients, suggesting the user to post to the one with the highest value. The number of comparisons required at the time of posting linearly depends on the number of forums available and not on the number of posts. This approach incurs in high computation cost, for large forums, to calculate centroids. However, the larger the number of messages in a forum, the less impact a few new messages have in its centroid's position. Therefore, centroid calculation can be performed periodically as an off-line batch process.

3.2 Assisted Posting Workflow

Figure 3 summarizes the modified posting workflow and the key elements of the assistance algorithm. First, the user writes the post in a form similar to the one used by regular Moodle Forums. When the user submits the post, its content is analysed and it's *tf-idf* vector created. Cosine Similarity is used to compare the vector to those of the centroids of each forum in the TDForum activity. Centroids are updated periodically, independently of the posting workflow. The similarity coefficients for each forum are used to rank the suggestions of the most adequate forum according to the new post's content. Although the system's recommendation is to post to the highest ranked forum, the user makes the final decision.

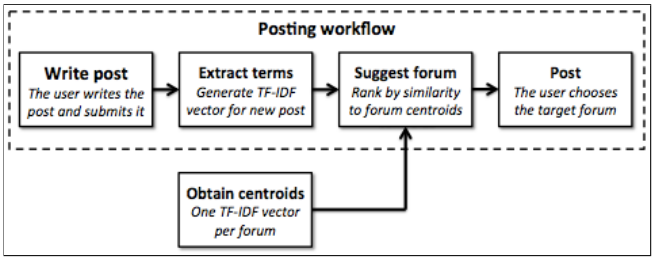

Fig. 3: Posting workflow modified to include assisted selection of the most appropriate forum.

4 TDForum: A Plugin for the Moodle Learning Management System

Topic-Driven Forum (TDForum) is a Moodle plugin (particularly, an *activity module*) that implements the topic-driven forum navigation enhancement approach described in Section 2 for the Moodle open-source learning management system.

In Moodle, *activity* is a general name for a group of features in a course. Usually an activity is something that a student will do that interacts with other students and/or the teacher. Assignments, quizzes, surveys, workshops, chats, and forums are examples of activities that can be created in a course and that are provided in Moodle by default. Each activity is implemented by a software module (plugin) located in the *mod* sub-folder of the Moodle instance. Additional activities can be included by installing the corresponding Moodle plugin[10].

From a source code point of view, each Moodle activity module consists of a series of mandatory files (e.g., install.xml, lib.php, and view.php) used to install

[10] A rich and up-to-date list of Moodle plugins can be found in the Moodle Plugins Directory at http://www.moodle.org/plugins

the module and integrate it within the Moodle system, and other files specific to the plugin.

Figure 4 shows the architecture of the TDForum Moodle plugin that we developed. In the figure, we can distinguish the components representing the plugin *front-end* (the graphical interfaces that Moodle users interact with), and those that are part of the plugin *back-end*.

The plugin front-end comprises the components *Main View* and *Discussion Topics View* corresponding to the two possible views on the forum content: (*i*) standard chronological list of discussions augmented with discussion topics and scores, and (*ii*) navigable hierarchical discussion topics graph. The last view is built using the *JavaScript InfoVis Toolkit*[11]. It also includes the *Admin User Interface* component which lets administrators manage the forum data processing and customize the visualization plugin parameters.

The plugin back-end contains the components implementing the forum analysis and indexing process described in Section 2 to build the additional topic-driven navigation structure. In particular, the *Process Controller* controls the process by executing the commands provided through the plugin admin user interface. It also exports forum content from the Moodle database into a local temporary csv text file and imports the data on the new navigation structure from the local filesystem into the Moodle database.

The *Data Processing* component includes the following sub-components:

- *Data Preprocessing*: a Perl script which extracts threads and messages from the csv file into separated text files and performs terms extraction and text filtering such as stopwords and stemming (cf. Section 2.1).
- *Topics Identification and Documents to Topics Assignment*: a R[12] script which uses the Topic Model library[13] to perform discussion topics identification and documents to topics assignment. The matrices Topics-Terms and Topics-Documents of the detected forum discussion topics and scores associated to them are generated in this step (cf. Sections 2.2 and 2.4).
- *Formal Concept Analysis and Topics Graph Export*: this component uses the FcaStone[14] Formal Concept Analysis command-line utility to generate the lattice representative of the hierarchy of topics and to export the topics graph used in the graph view of the plugin (cf. Section 2.3).

The TDForum activity implemented by our plugin offers the same features provided by a standard Moodle forum (particularly, a *main view* which lists forum discussions and messages organized in a chronological order, the functionality of posting new messages or replying to existing ones, full-text search of messages, etc.) and adds to them a *discussion topics view* which acts as a topic-driven navigation index to the forum content.

The *main view* (Fig. 5) presents the list of discussion threads of the forum in a chronological order and adds to each of them the list of discussion topics in it

[11] http://philogb.github.io/jit/
[12] http://cran.r-project.org/
[13] http://cran.r-project.org/web/packages/topicmodels/
[14] http://fcastone.sourceforge.net/

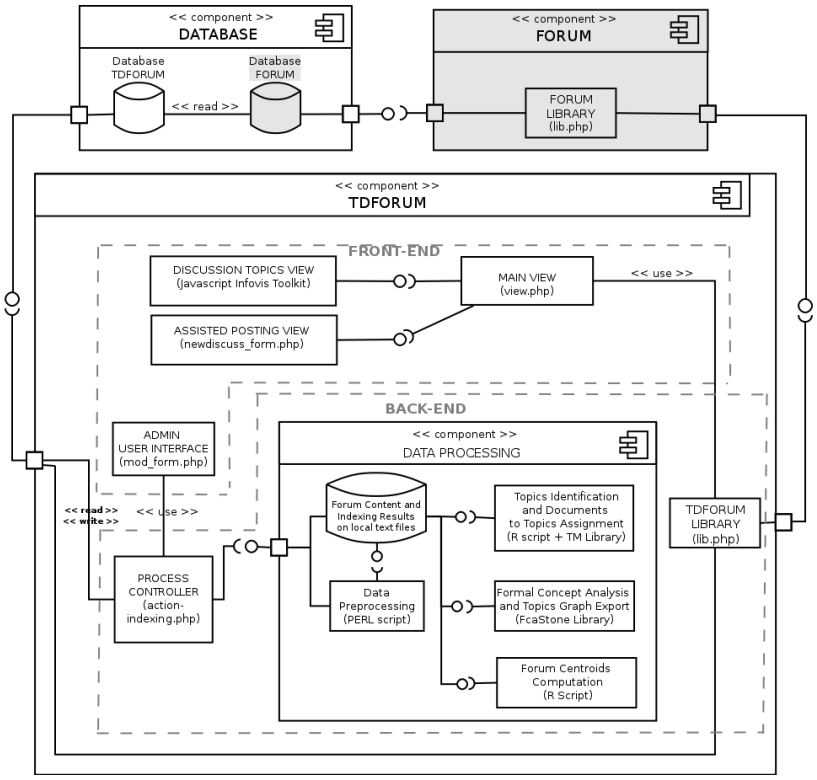

Fig. 4: Architecture of the TDForum Moodle plugin (with a gray background color, standard Moodle components).

identified, and the calculated similarity score (column 'Discussion topics' in the figure). Score values range between 0 and 1 (with 1 representing the maximum similarity value) and the list of topics associated to a discussion is ordered by score. By right-clicking one of the topics of the list, the user can search for discussions or messages which are related to the selected topic. The results of this search is presented sorted by decreasing values of score.

The *discussion topics view* (Fig.6) shows the list of discussion topics found by the analysis process for the considered forum (scrollable list on the left side of the figure) and a graph that the user can pan and zoom which highlights the hierarchical relations between the identified topics. The user can navigate the discussion topics graph or the topics list and once she finds a topic of her interest she can retrieve the list of discussions/messages associated to it with a click.

The plugin has been designed to extend a standard Moodle forum and, at the same time, to be independent from it. As such, if it is installed, applied on a forum, and then deactivated, none of the content of the original forum are lost, nor the additional messages/discussions that will have been added in it after the plugin instantiation. Moreover, the current version of the plugin now allows to

simultaneously index posts from various forums, thus supporting topic-driven navigation across them. Figure 5 displays the main view for a TDForum labeled "Holistic view forum", that indexes messages from two forums of the Moodle.org online english community, "Teaching with Moodle" and "Forum (Activity)". Messages are sorted by date and time thus, messages from both forums appear alternated. The second column (Forum) indicates to which forum each post belongs to. To post a new message (as there are two forums involved) the user is required to first specify the target forum. The common approach is to choose a forum from the drop down list, and then press the button labeled "Add a new discussion topic". This will start the standard Moodle posting workflow on the selected forum.

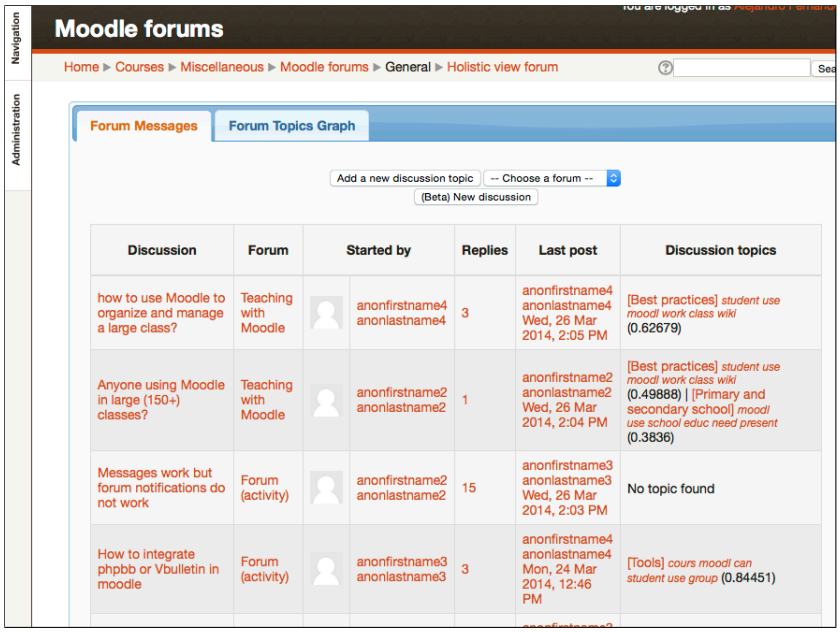

Fig. 5: The main view of TDForum showing the list of forum discussion threads enhanced with discussion topics and scores associated to them.

4.1 Assisted Posting

The "Assisted posting view" component in the architecture diagram shown in Fig. 4 implements the user interface functionality to support assisted posting. When the "(Beta) New discussion" button is clicked in the *main view* (Fig. 5), a new posting workflow starts. The user writes the new post in a form identical to that commonly use by Moodle forums. However, the "Post to forum" button invoques the term extraction functions provided by the "Data Preprocessing"

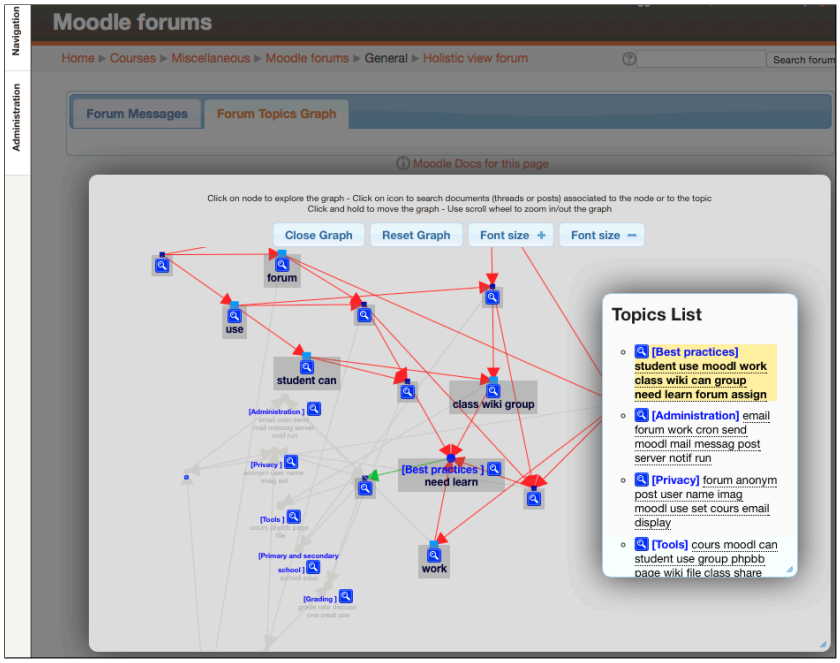

Fig. 6: The Discussion Topics view of TDForum showing the topics list and the hierarchical topics graph.

component. The resulting *tf-idf* vector is fed to the similarity function, in the "Forum Centroids Computation" component, that compares it to the centroids of all available forums and produces the list of forums ranked by relevance. The function additionally filters out all those forums the user does not have permission to post to. The user then selects the destination forum and proceeds to the next step in the posting workflow where forum specific checks, such as permission to include attachments, are made. Then, a final step confirms the user's submission. The "Forum Centroids Computation" component in the architecture diagram provides a centroids computation function that is periodically called by a Moodle sync task. The "Admin User Interface" component of the Plugin provides configuration forms that allow administrators to set the time between centroid updates.

5 Case Study

We evaluated qualitatively that, with the topic-driven approach, searching and browsing tasks of forum contents can be significantly improved with respect to traditional full-text search. The case study has been conducted on forums inside an instance of the Moodle learning management system. The context is composed by a reduced version of 2 forums extracted from the Moodle user

and development communities (Table 3). The *Installation Help Forum* includes all discussions about user difficulties with first Moodle installations or errors happening during the installation process, or with migration to different OSs, or to newer Moodle versions. The *General Help Forum* includes discussions about problems not included into other Moodle community forums, such as problems with database access, file upload, block modules and student enrollment.

Forum	# threads	# posts	# users	time period
Installation Help	253	777	78	May 1, 2013 – May 24, 2013
General Help	115	714	107	Jul 15, 2013 – Aug 8, 2013

Table 3: Case study context

We evaluated effectiveness in 11 searching tasks in terms of (*i*) the number of items (forum posts) the user had to inspect in order to satisfy the information need, and (*ii*) the time spent to accomplish the task (Table 4). The nature of the 11 searching tasks has been defined by the first two authors of this paper. For each task the search goal, *i.e.,* the expected posts that should be retrieved, is known beforehand. The other two authors performed the searching tasks with two complementary approaches: (*i*) full-text search, and (*ii*) topic driven navigation. The first approach is accomplished with the default full-text search engine implemented in the Moodle platform which performs a full-text search from a set of user defined keywords. By default, the keywords are linked by an AND operator and the system retrieves the list of all posts containing all searched terms. The second approach is executed with the TDForum Moodle plugin introduced in Section 4. While performing the tasks, we collected the number of items inspected and the time needed to achieve the search goal. With the Moodle full-text search the number of inspected items is computed by counting the number of posts examined before finding the correct expected posts. With the topic-driven navigation approach the number of inspected items is the sum of two quantities: the number of links followed to reach the closest topic in the Discussion Topics View of the TDForum plugin and the number of posts examined before finding the correct expected posts.

Table 5 reports the results obtained by executing the evaluation protocol on the 11 tasks. The table reports for each search goal the number of inspected items and the time spent to find the correct posts. For Moodle full-text search we also reported the number of search attempts (queries) performed with different search keywords necessary to reach the goal. In general the number of items inspected with full-text search is in average higher than the number of items inspected with TDForum (14 vs 9). The time necessary to obtain the correct answer is in average less in TDForum (137 sec. vs 170 sec.) because with full-text search more time is spent to choose the correct search keywords. The difference is not statistically significant due to the limited number of samples, thus further experiment are necessary to draw more general conclusions.

ID	Search goal	Adopted Keywords
1	Retrieve the 10 posts related to css problems in the General help forum	*css*, *problem*
2	Retrieve the 5 posts related to login issues in the General help forum	*login*, *problem*
3	Retrieve the 3 posts related to uploading files problems in the General help forum	*file*, *not*, *upload*
4	Retrieve the 4 posts related to changing Moodle fonts in the General help forum	*change*, *font*
5	Retrieve the 5 posts related to not sent enrollment email in the General help forum	*enrollment*
6	Retrieve the 3 posts related to editing Moodle theme in the General help forum	*change*, *theme*
7	Retrieve the 5 posts related to web hosting in the General help forum	*moodle*, *web*, *hosting*
8	Retrieve the 10 posts related to Moodle upgrading problems in Installation help forum	*problem*, *moodle*, *upgrade*
9	Retrieve the 5 posts related to editing admin password in the Installation help forum	*admin*, *password*
10	Retrieve the unique post related to missing files after Moodle migration in the Installation help forum	*missing*, *files*, *after*, *migration*
11	Retrieve the 2 post related to slower system after upgrade in the Installation help forum	*css*, *problem*

Table 4: Search tasks definition

5.1 Assisted Posting Performance

In order to validate the approach and assess how well each of the alternatives performed, we evaluated and compared the "Centroids" and the "One-like-this" algorithms on existing forum data. To conduct the evaluation we obtained posts from two forums in the Moodle Community, namely "Teaching with Moodle" [15], and "Forum"[16], (a forum about the *Forum activity*). In total they had 422 posts. In addition, we conducted an evaluation on forums obtained from Stack Exchange. We built two datasets from it. One of them consists of two forums with closely related topics, namely "Webapps" [17] and "Webmasters" [18]. It contains 34,248 posts. The other dataset combines four forums with different topics, namely "Beer" [19], "Aviation" [20], "Politics" [21], and "Space Exploration" [22]. It contains 4,642 posts.

[15] `https://moodle.org/mod/forum/view.php?id=41`
[16] `https://moodle.org/mod/forum/view.php?id=732`
[17] `http://webapps.stackexchange.com`
[18] `http://webmasters.stackexchange.com`
[19] `http://beer.stackexchange.com`
[20] `http://aviation.stackexchange.com`
[21] `http://politics.stackexchange.com`
[22] `http://space.stackexchange.com`

Task ID	Moodle full-text search # queries	# items	time	TDForum search # items	time
1	2	15	201	20	254
2	1	5	109	5	92
3	1	17	131	8	168
4	3	12	230	7	135
5	3	13	225	11	187
6	5	40	275	9	113
7	1	5	134	4	62
8	2	42	287	10	131
9	1	4	122	6	90
10	1	1	86	16	192
11	1	0	70	6	85
average	2	14	170	9	137

Table 5: Case study results (time in seconds)

We evaluated both algorithms using cross validation. A percentage of the messages in each forum, selected randomly, was set apart as the test set, the rest used for training. We used both algorithms to obtain the most appropriate target for each of the posts in the test set and we recorded the number of times the algorithm suggested a forum different from the one the post was in. Table 6 depicts the results we observed. To cope with variability in the results, we performed 6 runs for each percentage split and reported the average results. The "Centroids" approach performed consistently better, by a large margin, than the simpler "One-like-this". Moreover, the "Centroids" approach performed substantially better when the forums had different topics.

Data-set	Split (train - test)	Centroids	One-Like-This
Moodle community	30% - 70%	9.03%	54.49%
Moodle community	50% - 50%	7.02%	47.66%
Moodle community	70% - 30%	6.68%	48.62%
StackExchange similar	30% - 70%	16.03%	23.19%
StackExchange similar	50% - 50%	15.35%	22.65%
StackExchange similar	70% - 30%	15.49%	22.44%
StackExchange different	30% - 70%	4.79%	43.14%
StackExchange different	50% - 50%	3.89%	37.12%
StackExchange different	70% - 30%	4.07%	33.80%

Table 6: Evaluation results for assisted posting - Error percentages

6 Related Work

Recently, on-line education systems are becoming widespread tools adopted by both historical and newly founded educational institutions. E-learning and e-teaching are new contexts for education where large amounts of information are generated and ubiquitously available. Most of generated information has the form of free text without a structure crucial for automating knowledge retrieval.

Data Mining has been historically used to extract knowledge from free text [1]. Knowledge extraction from e-learning systems, in particular from user generate data, has been introduced in [6, 13]. Patterns of system usage by teachers and learning behavior by students has been investigated in [29]. Data clustering was suggested to promote group-based collaborative learning and to diagnose students incrementally [5].

Web Mining techniques to meet some of the current challenges in distance education was presented in [28] where a clustering of forum messages are in fact grouped into similar discussion topic classes. Association Rules mining has been widely adopted in e-learning, in particular recommendation systems [33, 32], learning material organization [30], student learning assessments [23], course adaptation to the students behavior [14], and evaluation of educational websites [25]. In educational research the development of cooperative learning and knowledge sharing inside student groups constitute recent research trends[16]. To this aim, Web technologies should grasp the opportunities raised by mixing the Social and the Semantic Web [11] and on adopting Semantic and Artificial Intelligence techniques for discovering information objects and restructure large digital collections [19]. Concept maps and their use for navigation in educational contexts has been investigated in the recent past by different authors. As a representative of this research effort we cite the work of Dicheva and Aroyo [8]. In this work the authors propose a framework and a set of tools for the development of ontology-aware repositories of learning materials. While the idea and use of *concept maps* is similar to our topic-driven navigation structure, in our approach topics are extracted from free text in a semi-automatic way, by leveraging information retrieval techniques and then validated by the user, while concepts have to be manually defined by the authors of the learning materials in the work of Dicheva and Aroyo.

Information retrieval and topic modeling have been used in the context of on-line conversations, for example, to help users in on-line communities obtain faster and better answers. This has been achieved by routing new messages to those users that are potential experts in the topic of the request and more likely to answer [35]. In this work, the authors compute expertise by using both the content of messages and the structure of the network. Moreover, posting assistance has been a research topic in the context of Issue Tracking Systems. Jalbert *et al.* [17] propose a mechanism to automatically classify duplicate bug reports upon creation, thus saving developer time. They use surface features, textual semantics, and graph clustering to predict duplicate status. Similarly, Nguyen *et al.* [21] use a combination of traditional information retrieval strategies and topic modeling to identify duplicate bug reports even when different terms are

used. Although these works share owr goal of keeping conversation focused and cohesive, our approach is not concerned with duplicates neither with obtaining prompt replies. Still, future work could incorporate such concerns.

7 Conclusions and Future Work

Online discussion forums are one of the main asynchronous communication means and repositories of user generated content over the Internet. Learning management systems (LMSs), such as Moodle, use forums to support interaction and collaboration between students and students-to-teachers. Discussions taken place in a forum at some time represent a source of information for users accessing the forum afterwards. However, the effectiveness of a forum as a source of information for its users, additionally to be closely related to its richness in content, is also influenced by the way its contents are organized made searchable.

In this paper we presented an approach and a plugin for the Moodle LMS that enhances content navigation and information search in online discussion forums with a topic-driven navigational paradigm. The approach enables the automatic recovery of a lattice of discussion topics from the forum content, and the introduction of an additional navigation structure and graphical user interface which enable navigating and searching forum contents by topics of discussion. The plugin also supports the indexing of multiple forums into a single topic-driven enhanced forum, which is useful when conversation is split in various independent forums, with crosscutting topics.

While the approach has proven correctness for both the identified topics and the document-to-topics assignment [7], in this paper we have also shown with a case study that the additional navigation structure significantly improves the search of information stored in forum discussions.

In order to keep forums focused and cohesive, the plugin integrates a functionality that assists users in choosing the most appropriate forum to post to. Evaluation of the underlying assisted posting algorithms showed high accuracy, specially when dealing with forums discussing different topics.

In the future we aim to apply our approach in the context of social networks, in order to explore how it could improve social organization and user interaction. As a matter of fact, social networks are increasingly used in e-learning as side means for connecting students and teachers. Additionally, we plan to explore the applicability of "assisted posting" at a thread granularity level, to identify specific threads to post to within a given forum or a set of forums. This would increase cohesion of discussions and, potentially, reduce content duplication.

References

1. Baeza-Yates, R.A., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1999)
2. Bakalov, A., McCallum, A., Wallach, H.M., Mimno, D.M.: Topic models for taxonomies. In: Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '12, Washington, DC, USA, June 10-14, 2012. pp. 237–240 (2012)

3. Birkhoff, G.: Lattice theory. In: Colloquium Publications, vol. 25. Amer. Math. Soc., 3. edn. (1967)
4. Blei, D.M.: Introduction to probabilistic topic models. Communications of the ACM (2011), http://www.cs.princeton.edu/~blei/papers/Blei2011.pdf
5. Castro, F., Nebot, A., Mugica, F.: Extraction of logical rules to describe students' learning behavior. In: Proceedings of the sixth conference on IASTED International Conference Web-Based Education - Volume 2. pp. 164–169. WBED'07, ACTA Press, Anaheim, CA, USA (2007), http://dl.acm.org/citation.cfm?id=1323159.1323189
6. Castro, F., Vellido, A., Nebot, A., Mugica, F.: Applying data mining techniques to e-learning problems. In: Jain, L., Tedman, R., Tedman, D. (eds.) Evolution of Teaching and Learning Paradigms in Intelligent Environment, Studies in Computational Intelligence, vol. 62, pp. 183–221. Springer Berlin Heidelberg (2007)
7. Cerulo, L., Distante, D.: Topic-driven semi-automatic reorganization of online discussion forums: A case study in an e-learning context. In: Global Engineering Education Conference (EDUCON), 2013 IEEE. pp. 303–310 (March 2013)
8. Dicheva, D., Dichev, C.: Tm4l: Creating and browsing educational topic maps. British Journal of Educational Technology 37(3), 391–404 (2006), http://dx.doi.org/10.1111/j.1467-8535.2006.00612.x
9. Distante, D., Cerulo, L., Visaggio, C.A., Leone, M.: Enhancing online discussion forums with a topic-driven navigational paradigm: A plugin for the moodle learning management system. In: Proceedings of the 6th International Conference on Knowledge Discovery and Information Retrieval. pp. 97–106. KDIR 2014, Scitepress (2014)
10. Ganter, B., Wille, R.: Formal concept analysis: mathematical foundations. Springer (1999)
11. Ghenname, M., Ajhoun, R., Gravier, C., Subercaze, J.: Combining the semantic and the social web for intelligent learning systems. In: Global Engineering Education Conference (EDUCON), 2012 IEEE. pp. 1 –6 (april 2012)
12. Gruen, T.W., Osmonbekov, T., Czaplewski, A.J.: eWOM: The impact of customer-to-customer online know-how exchange on customer value and loyalty. Journal of Business Research 59, 449456 (2006)
13. Hanna, M.: Data Mining in the e-Learning Domain. Campus-Wide Information Systems 21(1), 29–34 (2004)
14. Hogo, M.A.: Evaluation of e-learning systems based on fuzzy clustering models and statistical tools. Expert Syst. Appl. 37(10), 6891–6903 (Oct 2010), http://dx.doi.org/10.1016/j.eswa.2010.03.032
15. Hrastinski, S.: What is online learner participation? a literature review. Computers & Education 51(4), 1755 – 1765 (2008)
16. Jakobsone, A., Kulmane, V., Cakula, S.: Structurization of information for group work in an online environment. In: Global Engineering Education Conference (EDUCON), 2012 IEEE. pp. 1 –7 (april 2012)
17. Jalbert, N., Weimer, W.: Automated duplicate detection for bug tracking systems. In: Dependable Systems and Networks With FTCS and DCC, 2008. DSN 2008. IEEE International Conference on. pp. 52–61 (June 2008)
18. Li, Q., Wang, J., Chen, Y.P., Lin, Z.: User comments for news recommendation in forum-based social media. Information Sciences 180, 49294939 (2010)
19. Martin, A., Leon, C.: An intelligent e-learning scenario for knowledge retrieval. In: Global Engineering Education Conference (EDUCON), 2012 IEEE. pp. 1 –6 (april 2012)
20. Meila, M.: Comparing clusterings by the variation of information. In: Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. pp. 173–187 (2003)

21. Nguyen, A.T., Nguyen, T.T., Nguyen, T.N., Lo, D., Sun, C.: Duplicate bug report detection with a combination of information retrieval and topic modeling. In: Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering. pp. 70–79. ASE 2012, ACM, New York, NY, USA (2012), http://doi.acm.org/10.1145/2351676.2351687
22. Otterbacher, J.: Searching for product experience attributes in online information sources. In: Proceedings of the International Conference on Information Systems (ICIS 2008). Association for Information Systems (December 2008)
23. Romero, C., Ventura, S., Bra, P.D.: Knowledge discovery with genetic programming for providing feedback to courseware authors. User Modeling and User-Adapted Interaction 14(5), 425–464 (Jan 2005), http://dx.doi.org/10.1007/s11257-004-7961-2
24. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM 18(11), 613–620 (Nov 1975), http://doi.acm.org/10.1145/361219.361220
25. dos Santos Machado, L., Becker, K.: Distance education: A web usage mining case study for the evaluation of learning sites. In: 2003 IEEE International Conference on Advanced Learning Technologies (ICALT 2003), 9-11 July 2003, Athens, Greece. pp. 360–361. IEEE Computer Society (2003)
26. Stefan, H.: A theory of online learning as online participation. Computers & Education 52(1), 78–82 (2009)
27. Sudau, F., Friede, T., Grabowski, J., Koschack, J., Makedonski, P., Himmel, W.: Sources of information and behavioral patterns in online health forums: qualitative study. Journal of medical Internet research 16, e10 (2014)
28. Sung Ho Ha, Sung Min Bae, S.C.P.: Web mining for distance education (2000)
29. Tang, T., McCalla, G.: Smart Recommendation for an Evolving e-Learning System: Architecture and Experiment. International Journal on e-Learning 4(1), 105–129 (2005)
30. Tsai, C.J., Tseng, S.S., Lin, C.Y.: A two-phase fuzzy mining and learning algorithm for adaptive learning environment. In: Proceedings of the International Conference on Computational Science-Part II. pp. 429–438. ICCS '01, Springer-Verlag, London, UK, UK (2001), http://dl.acm.org/citation.cfm?id=645456.654828
31. Wallach, H.M., Mimno, D.M., McCallum, A.: Rethinking lda: Why priors matter. In: Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. pp. 1973–1981 (2009)
32. Yang, Q., Sun, J., Wang, J., Jin, Z.: Semantic web-based personalized recommendation system of courses knowledge research. In: Proceedings of the 2010 International Conference on Intelligent Computing and Cognitive Informatics. pp. 214–217. ICICCI '10, IEEE Computer Society, Washington, DC, USA (2010), http://dx.doi.org/10.1109/ICICCI.2010.54
33. Zaíane, O.R.: Building a recommender agent for e-learning systems. In: Proceedings of the International Conference on Computers in Education. pp. 55–. ICCE '02, IEEE Computer Society, Washington, DC, USA (2002), http://dl.acm.org/citation.cfm?id=838238.839230
34. Zhang, K., Peck, K.: The effects of peer-controlled or moderated online collaboration on group problem solving and related attitudes. Canadian Journal of Learning and Technology / La revue canadienne de lapprentissage et de la technologie 29(3) (2003)
35. Zhou, Y., Cong, G., Cui, B., Jensen, C.S., Yao, J.: Routing questions to the right users in online communities. In: Proceedings of the 2009 IEEE International Conference on Data Engineering. pp. 700–711. ICDE '09, IEEE Computer Society, Washington, DC, USA (2009), http://dx.doi.org/10.1109/ICDE.2009.44