Topic-Driven Semi-Automatic Reorganization of Online Discussion Forums: A Case Study in an E-Learning Context

Luigi Cerulo[1] and Damiano Distante[2]
[1]Faculty of Science, University of Sannio, Benevento - Italy
[2]Faculty of Economics, Unitelma Sapienza University, Rome - Italy
[1]lcerulo@unisannio.it, [2]damiano.distante@unitelma.it

***Abstract*—Online discussion forums represent, nowadays, one of the main asynchronous communication means and information sources over the Internet. The forum paradigm is adopted by the most followed websites, such as social networks and blogs. The effectiveness of discussion forums as information source, *i.e.*, the capability to satisfy their users' information needs, depends on their information richness first, but also on how their are organized and effectively moderated. Forums organized and moderated by topics of discussion tend to host messages on related subjects and, overall, provide a classification of message threads which eases information search.**

In this paper we propose a semi-automatic approach to detect topics of discussion in a forum and to enhance its organization by providing a hierarchical topic-driven navigation view on its messages. We adopt Information Retrieval (IR) techniques, such as topic modeling, and formal concept analysis (FCA) to identify discussion topics and to provide a hierarchical topic-centered view on messages.

We tested the validity of our approach on four forums of the e-learning platform of an Italian distance-learning university which provides around 20 moderated and unmoderated main forums followed actively by almost 5000 users, including students and teachers, each year. We validated the topics identification and messages to topics allocation process with a specific empirical experiment obtaining promising results.

Keywords-Online discussion forums; Topic-driven reorganization; Information Retrieval; Clustering; e-Learning, Learning management systems.

I. Introduction

An Internet forum is an online discussion place where people can hold conversations in the form of posted messages. It represents, nowadays, one of the main asynchronous communication means and information sources over the Internet in several domains ranging from e-commerce [1][2][3] to news [4] and healthcare [5]. The most followed websites, such as blogs and social networks, adopt the forum paradigm to enable interaction between their users and to support their communities in sharing information and creating knowledge.

In general, a discussion forum has a flat structure, but could be explicitly organized also in a hierarchical or tree-like fashion with a number of sub-forums, each of which may include several discussion topics. A posted message can open a new discussion or could be a reply to a previous posted message. Each initial post with the series of messages it receives as replies act as a conversation usually named *discussion thread*.

In specific contexts, discussion forums may constitute a wide and rich repository of information. For example, developer forums may be an effective repository of good programming practices, where, usually non expert, programmers search for typical coding solutions [6]. In e-learning contexts, discussion forums are used to enable asynchronous communication between students, and between teachers and students, for a number of purposes [7][8], including: (*i*) virtual classrooms creation and interaction support; (*ii*) collaborative learning and group work; and (*iii*) knowledge and information sharing. Discussions taken place during a certain period of time become a source of information for any user accessing the forum afterwards and can be made available for a long time. As an example, answers provided by teachers and tutors to learners' questions may be useful to other students facing the same problems or wondering the same questions. Discussion forums are provided as a communication means by basically any Learning Management System (LMS), such as Moodle[1], Blackbard[2], and Docebo[3], and are usually related to classes, or to wider educational environments, such as Departments and/or Faculties.

The effectiveness of discussion forums as information source, *i.e.*, their capability to satisfy their users' information needs, in addition to depend on their information richness, depends primarily on how they are organized and effectively moderated [9]. In this regard, the Internet netiquette guidelines [10] suggest preserving the original discussion topic of each discussion thread in a forum and starting new discussion threads as new discussion topics arise. Such policies are usually governed and implemented in a forum by a set of individuals including forum administrators and moderators which may recall unobservant users posting *off-topic* messages. However, this activity is time and effort consuming, as it requires domain experts reading and judging the content

[1]www.moodle.org
[2]www.blackboard.com
[3]www.docebo.com

of each new posted message. As a consequence, many forums on the Internet are moderated only occasionally and not in depth, or are completely unmoderated.

Unmoderated and general discussion forums tend to host a potentially high number of messages about very different topics; messages talking of a given topic may be spread in different discussion threads, and discussion threads may contain messages discussing more than one topic. In these forums, information search might be difficult, information noise can be high (*e.g.*, the same discussions may appear several times or be split in different threads in the forum), and, overall, they can causes wasting of time for both consumers and providers of the information. Conversely, forums well organized and moderated tend to host messages on related subjects and, overall, provide a classification of messages into topics of discussion corresponding to discussion threads. This drastically improves the retrieval of useful information and reduces the chances for similar discussions taking place several times in the forum, resulting in higher user satisfaction. As mentioned above, however, moderation is an expensive and time consuming task, as it requires continuous human interventions to move messages to the appropriate forum or to create new threads/forums for new topics when needed.

In this paper we propose a semi-automatic approach to detect discussion topics in a forum enhancing its organization with a complementary hierarchical topic-centered navigation view on its discussion threads and messages. The method works both on a completely unstructured (*i.e.*, flat) forum or on a set of partially structured forums, to possibly offer a topic oriented view of their discussion threads. We adopt information retrieval (IR) and topic modeling techniques to extract significant topics from discussion messages. Then, we organize such topics into an optimal topic-lattice structure by means of formal concept analysis (FCA).

We tested the validity of our approach on the forums of the e-learning platform of an Italian distance-learning University (Unitelma Sapienza[4]) which provides almost 20 main forums (including moderated and unmoderated forums) followed actively by almost 5000 users, including students and teachers. We validated both the topics identification and the messages-to-topics allocation methods of our forum reorganization approach with a specific empirical study obtaining promising results. In particular, automatically suggested topics and their hierarchical organization in sub-topics has been considered meaningful and appropriate for each of the analyzed forums by experts of the considered educational organization. Precision and recall for the message-to-topics allocation method have also shown good averages.

The rest of the paper is organized as follows. The next Section describes the proposed topic-driven forum reorganization process. Section III discusses the validity of our approach within the Unitelma e-learning forums case study. Section IV overviews related work and Section V draws conclusions and outlines directions for future research we aim to conduct.

[4]www.unitelma.it

II. METHODS

The topic-driven forum reorganization process consists of four main steps depicted as rectangle boxes in Figure 1. In the first step standard information retrieval preprocessing, such as stopping, filtering, and stemming, are applied on original forum messages [11]. Then, topic models are adopted to identify relevant discussion topics. An optimal topic lattice structure is pruned from the topic-terms matrix by means of formal concept analysis. Each forum message is mapped onto the topics to which it is more likely to belong by estimating the probabilities of each topic for that message. In the following, each of the process main steps is described more in detail.

A. Data preprocessing steps

In Information Retrieval (IR) a *document* is the unit of information containing free text that is retrieved during an information retrieval task and satisfies a user information need. In our context, depending on the chosen granularity level, a document may be a single forum message or an entire thread of forum messages. We represent each document as a vector of indexing terms, $\{t_1, \ldots, t_m\}$, extracted, from the corpus of n documents, $\{d_1, \ldots, d_n\}$, through a standard text analysis pipeline:

1) *outlier filtering*, the process for filtering out very frequent (more than 90%) or very rare words (less than 1%);
2) *stopwords filtering*, the process for filtering out words belonging to a predefined vocabulary, e.g. Italian first names.
3) *stemming*, the process for reducing inflected words to their stem.

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 document d_j. The *term frequency-inverse document frequency* (*tf-idf*) is a common accepted statistics of how important a term is in a corpus [11].

B. Discussion topics extraction

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. The input of the topic modeling task is the document-term matrix, $\mathbb{DT}$, obtained from the previous task. Topic modeling approaches discover automatically a set of k topics $\{z_1, \ldots, z_k\}$ and the mapping between terms, topics and documents [12][13]. The number of topic k is a parameter that controls the granularity of the topics and must be fixed a priori. The outcome of a topic modeling task is a topic-document matrix, $\mathbb{TD}$, and a topic-term matrix, $\mathbb{TT}$. The elements $\{\mathbb{TD}\}_{ij}$ describe the topic membership values of topic z_i in document d_j, while the elements $\{\mathbb{TT}\}_{ip}$ describe the term membership values of term t_p in topic z_i.

More formally, each topic is defined by a probability distribution over all of the unique words in the corpus. Given two Dirichlet priors, α and β, a topic model fits a generative probabilistic model from the term occurrences in the corpus. The fitted model is able to capture an additional

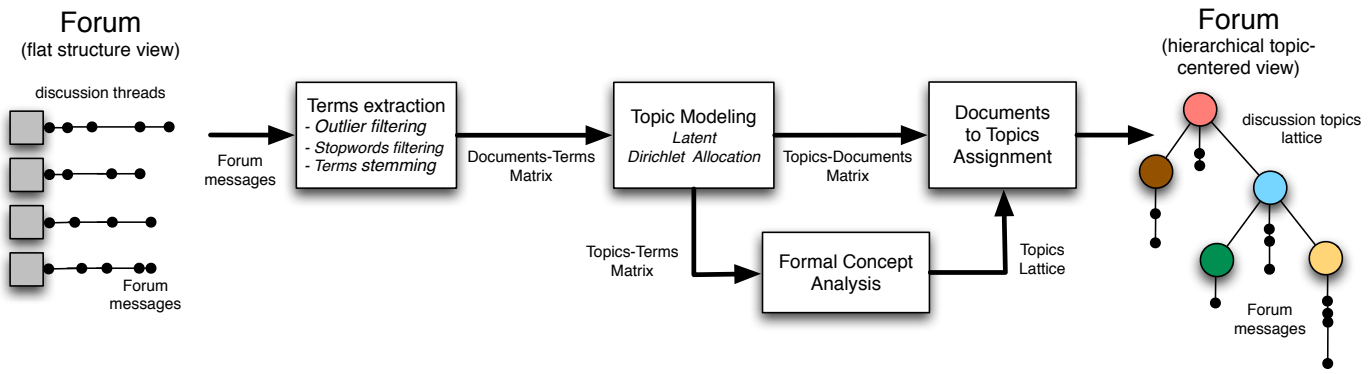


Fig. 1. The topic-driven forums reorganization process.

	Top terms	d_1	d_2	d_3	d_4
z_1	*computer, network, problem*	0.6	0.7	0.1	0.0
z_2	*mail, send, problem, computer, network*	0.3	0.1	0.1	0.5
z_3	*computer, network, course*	0.1	0.2	0.8	0.5

TABLE I
EXAMPLES OF TOPIC-TERM AND TOPIC-DOCUMENT MATRICES

	network	*problem*	*mail*	*send*	*computer*	*course*
z_1	×	×			×	
z_2	×	×	×	×	×	
z_3	×				×	×

TABLE II
THE FORMAL CONTEXT OBTAINED FROM THE TOPIC-TERM MATRIX SHOWN IN TABLE I

layer of latent variables which are referred as topics. Based on priors a topic model generates a topic probability distribution, $p(z = z_i|d_j) = \{\mathbb{TD}\}_{ij}$ for $i = 1, \ldots, k$, to each document d_j, and a term probability distribution, $p(t = t_p|z_i) = \{\mathbb{TT}\}_{ip}$ for $p = 1, \ldots, m$, to each topic z_i. Choosing the right parameter values for k, α, and β depends on the size of the corpus and the desired granularity of the topics [14].

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

Using the topic membership of a term, we prune a topic lattice by means of formal-concept analysis, while using the topic membership of a document we assign each forum message to the correspondent relevant topics as described by the following steps.

C. Topics lattice pruning

An optimal topic lattice structure is pruned from the topic-terms matrix, $\mathbb{TT}$, 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 [15]. The idea behind FCA is based on the notions of *formal context* and *formal concept*, defined as follows:

Definition 1. *A* **formal context** *is defined as a triple* (Ω, Δ, R)*, where* $R \subseteq \Omega \times \Delta$ *is a binary relation between a set of formal objects* Ω *and a set of formal attributes* Δ*.*

Definition 2. *A* **formal concept** *is a maximal collection of formal objects sharing common formal attributes. It is defined as a pair* (O, A) *with* $O \subseteq \Omega$ *and* $A \subseteq \Delta$ *such that:*
i) $\forall o \in O, \forall a \in A \implies (o, a) \in R$*;*
ii) $\forall o \notin O \implies \exists a \in A, (o, a) \notin R$*;*
iii) $\forall a \notin A \implies \exists o \in O, (o, a) \notin R$*.*

The set O of a formal concept is named the *extend* of the concept, while the set A is named its *intent*. The following notion of *subconcept* is adopted to construct a concept lattice:

Definition 3. *A formal concept* (O_1, A_1) *is defined as the* **subconcept** *of the formal concept* (O_2, A_2) *iif* $O_1 \subseteq O_2$ *or* $A_2 \subseteq A_1$*.*

The set of all formal concepts of a formal context form a partial order and define a complete lattice [16].

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 II; Figure 4 shows the topic lattice obtained from such a formal context. In it, each topic is mapped on a circle.

D. Documents assignment

During this step each document 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

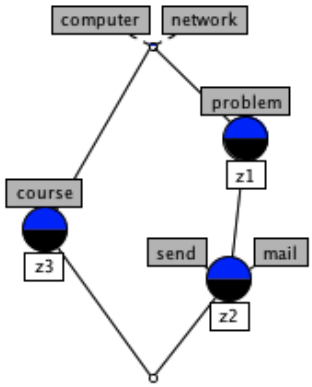


Fig. 2. The topic-lattice pruned from the formal context shown in Table II.

decision threshold h_D, *i.e.*, a document d_j belongs to a topic z_i iif $\{\mathbb{TD}\}_{ij} \geq h_D$. The pruned topic lattice structure is adopted to build the topic-driven navigation view.

III. Case study

The purpose of the conducted case study was to test the validity of our approach. Intuitively, an approach is valid if with it we can be observed a significant improvement of one of more characteristics of interest. In this case we are interested in evaluating whether the reorganized forum structure constitutes a valuable alternative for typical forum interaction tasks. A user searching a forum for a solution to a given problem or for a discussion taken place on a topic of her interest will take advantage from a topic-driven structured forum if the following two conditions are satisfied:

1) the topic lattice provided as an access structure to forum messages is able to guide the user through the topics actually discussed in the forum;
2) forum messages are associated to the topics they actually talk about.

It follows that the aspects we aim to evaluate in our case study is to what extend the automatically extracted topics are able to reveal:

1) the correct number of topics actually discussed in the forum; and
2) the topics actually discussed in a given document (*i.e.* forum message or thread).

We evaluated the first aspect quantitatively and the second quantitatively. To estimate the optimal number of topics k we adopt a metric that measures the Variation of Information between two topic solutions. To qualitatively assess the document-to-topics association process, we conducted an experiment in which we compare those associations with those assigned manually by a number of forum users obtaining an estimation of the accuracy of the association.

We implemented each step of our topic-driven forum reorganization process with specific software tools. In particular, for the information retrieval steps we adopted the *tm* R package which provides standard natural language processing functions. For words stemming we adopted the *Snowball* R package which provides the support for a number or languages, including Italian. Topic analysis was performed with *topicmodels* R package. More information about such packages and tools can be found at the R-project website http://www.r-project.org. The topic lattices were pruned with the java Concept Explorer *conexp* (http://conexp.sourceforge.net/.

In this section we introduce the subject that we used for our case study and report the results of the evaluation procedure.

A. The subject: UniTelma Sapienza e-learning forums

To validate our approach, we applied it to reorganize in a topic-centered view four of the e-learning forums of an Italian distance-learning University (Unitelma Sapienza). As in many other distance-learning environments, online forums represent in Unitelma Sapienza one of the asynchronous communication means between students, and students to teachers. Online discussion forums are provided by the adopted learning management systems, and in Unitelma it is possible to count about 20 main moderated and unmoderated forums (we do not count in this number smaller forums, such as those associated to each class), which are followed actively by almost 5000 users including students, teachers, learning tutors, and administrative and technical support staff. Table III reports the main discussion topics and some statistics on the four analyzed forums, including the number of discussion threads, the number of total messages, the number of terms extracted by our analysis process, and the average number of terms per message.

The "General discussions" forum is the largest forum including 10609 messages organized into 2031 discussion threads with an average of 46.67 terms per message. This forum hosts discussion on generic topics (so it is open to any discussion topic), and for this reason it is the largest forum in Unitelma. The "Technical issues" forum is conceived to host requests for technical support and reports of technical issues in online services offered by the organization, such as email and online video lessons. Finally, the remaining two analyzed forums are also general purpose forums, but associate to smaller learning contexts, such as specific degree and post-graduate curricula.

All the forums where observed in a period spanning from January 2011 to July 2012, *i.e.*, their content is represented by the messages posted by users in the above period of time.

B. Optimal number of topics

Selecting the number of topic k is one of the most problematic modeling choice in topic models [14]. Ideally if the Latent Dirichlet Allocation has sufficient topics to model the corpus of documents, the assignment of of terms to topics should be relatively invariant to an increase of k. We adopt a metric, introduced by Meilă [17] for clustering comparison, that measures the *Variation of Information*, $VI(C_1, C_2)$, between two clusters assignments. VI has several interesting properties and is able to measure the amount of information loss by C_1

Forum	Main topics	# of users	# of threads	# of messages	# of extr. terms	avg terms/msg
1	General discussions	≈5K	2031	10609	495202	46.67
2	Technical issues	≈5K	1033	2087	76483	36.64
3	Post-graduate course 1	≈2K	100	1089	58647	53.85
4	Post-graduate course 2	≈1.5K	101	1091	58860	53.95

TABLE III
UNITELMA FORUMS STATISTICS (CONTENT RESULTING FROM ACTIVITY IN THE PERIOD FROM JANUARY 2011 TO JULY 2012).

and gained by C_2 when going from assignment C_1 to C_2. Formally, $VI(C_1, C_2)$ is defined as:

$$VI(C_1, C_2) = H(C_1) + H(C_2) - 2I(C_1, C_2)$$

where $H(C_1)$ and $H(C_2)$ are the entropies associated respectively with cluster assignments C_1 and C_2, and $I(C_1, C_2)$ is the mutual information between cluster assignments C_1 and C_2 [17]. 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.* [14] the assignment of documents (forum messages in our context) to topics can be assimilated to a sort of cluster assignment where the probability that a document d_j is assigned to a topic z_i is proportional to:

$$p(z_i) \propto \sum_{j=1}^{n} p(z = z_i | d_j)$$

and the probability that a document d_j is assigned to topic z_i in the first topic assignment and to $z'_{i'}$ in the second topic assignment is proportional to:

$$p(z_i, z'_{i'}) \propto \sum_{j=1}^{n} p(z = z_i | d_j) p(z = z'_{i'} | d_j)$$

With those probabilities we compute the Variation of Information of topics assignment by increasing the number of topics from k to $k + 1$. We show 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. To compute such a value we searched the optimal regression lines that best fit the curve of Variation of Information. The condition of best fitness is found when the sum of squared residuals is minimum.

Figure 3 shows the Variation of Information obtained for each forum for different number of topics. The optimal number of topics can be observed at the intersection of the optimal regression curves. The optimal number of topics observed in each forum are respectively, 20, 8, 22, and 20. This is almost expectable as forum 2 is related to technical issues, thus it has more restrictive and specific discussion topics, while forum 1, 3 and 4 are more general, thus allowing for more heterogeneous discussions.

Figure 4 shows the topics lattice obtained for 'Technical issues' forum. The lattice represents main terms included and characterizing each topic, but also hierarchical relationships between groups of terms, which actually represent sub-topics and macro-topics. Some of the most relevant topics that can be easily recognized are those characterized by the following sets of terms:

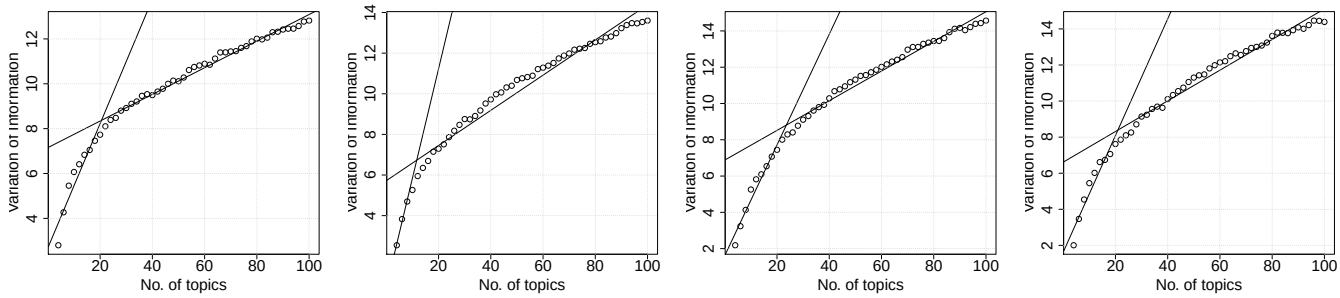
- *not*, *able*, *access*, *webmail* (issues in accessing the webmail service);
- *not*, *problem*, *lesson*, *audio*, *video*, *listen*, *view*, *watch*, *connection* (issues in viewing online video lessons);
- *not*, *mail*, *send*, *test*, *exam*, *intermediate*, *examination* (difficulties in sending via e-mail the test of an intermediate examination).

C. Documents to topics assignment evaluation

We wanted to check whether the documents assigned to topics by the Latent Dirichlet Allocation were congruent with the semantics of their content [18]. To address this question we designed and issued a survey to 10 undergraduate students volunteers of the Unitelma computer science class. Each participant was presented with a random list of 10 documents and the optimal set of topics discovered by topic models for the forum from which the documents were extracted. Each topic was represented with its most relevant terms and all users were asked to assign the documents to the most relevant topics. We repeated such tasks 4 times (thus to manually assign to topics around 100 documents for each forum) and computed the average fraction of documents that were correctly assigned by topic models. The automatic document to topics assignment process considers a message to be associate to a topic if its posterior probability is above a threshold of 0.6 (calibrated empirically).

Figure 5 show a screenshot of the tool developed to support the documents-to-topics assignment evaluation process. By using a browser, the tool allows to view for each selected topic (top-right frame), the list of the most relevant terms (frame below the list of topics), the list of 10 documents randomly selected among those assigned to the topic (bottom-right frame), and to view the content of each document (central frame). The tool also shows the topic-term lattice obtained from the analysis phase for the considered forum.

With the above described experiment, we could only assess the precision (*i.e.*, the fraction of corrected assigned documents) of the automatic assignment process, while evaluating the recall would have required to manually check the assign-



(a) Forum 1 (20 topics) (b) Forum 2 (8 topics) (c) Forum 3 (22 topics) (d) Forum 4 (20 topics)

Fig. 3. Optimal number of topics identified in each forum.

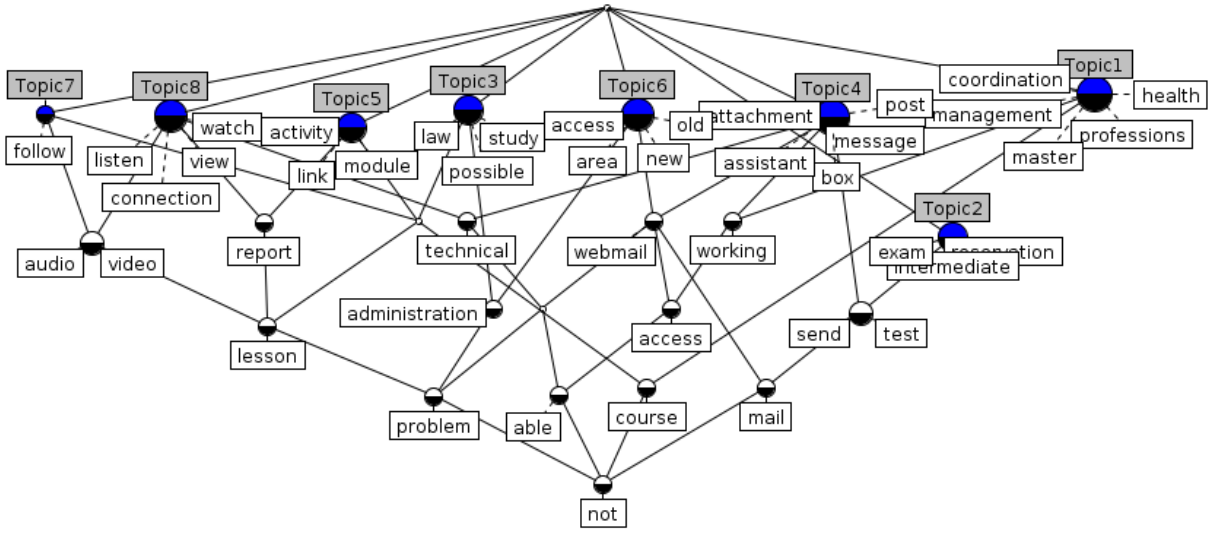


Fig. 4. The topics lattice obtained with Forum 2 ("Technical issues").

Forum	Precision
1	0.62
2	0.74
3	0.54
4	0.52

TABLE IV
AVERAGE PRECISIONS OF THE DOCUMENTS-TO-TOPICS ASSIGNMENT PROCESS ON UNITELMA FORUMS.

ment of all documents of a forum. Results reported in Table IV shows that the method performs better when the number of identified topics is smaller, *i.e.*, on forums which are not general discussion forums (this may also be due to the fact that it was easier for the subjects of the experiment to check and associate messages to topics). Overall, the resulted precision was never below 0.5 and as average it was around 0.60, which means that around 2 out of 3 messages are correctly associated to identified forum discussion topics.

IV. Related work

The explosion of online education, e-learning, and e-teaching, as a new context for education where large amounts of information describing the continuum of the teaching-learning interaction are generated and ubiquitously available. This makes available a plenty of information but at the same time promotes unstructured information and chokes the educational system without providing any articulate knowledge to its stakeholders.

Data Mining has been recently used to extract knowledge from e-learning systems by analyzing the information available and generated by their users [19], [20]. Patterns of system usage by teachers and learning behavior by students has been investigated in [21]. Data clustering was suggested to promote group-based collaborative learning and to diagnose students incrementally [22].

A review of the possibilities of the application of Web Mining (Web usage mining and clustering) techniques to meet some of the current challenges in distance education was presented in [23]. The approach presented in this paper performs a sort of clustering as forum messages are in fact

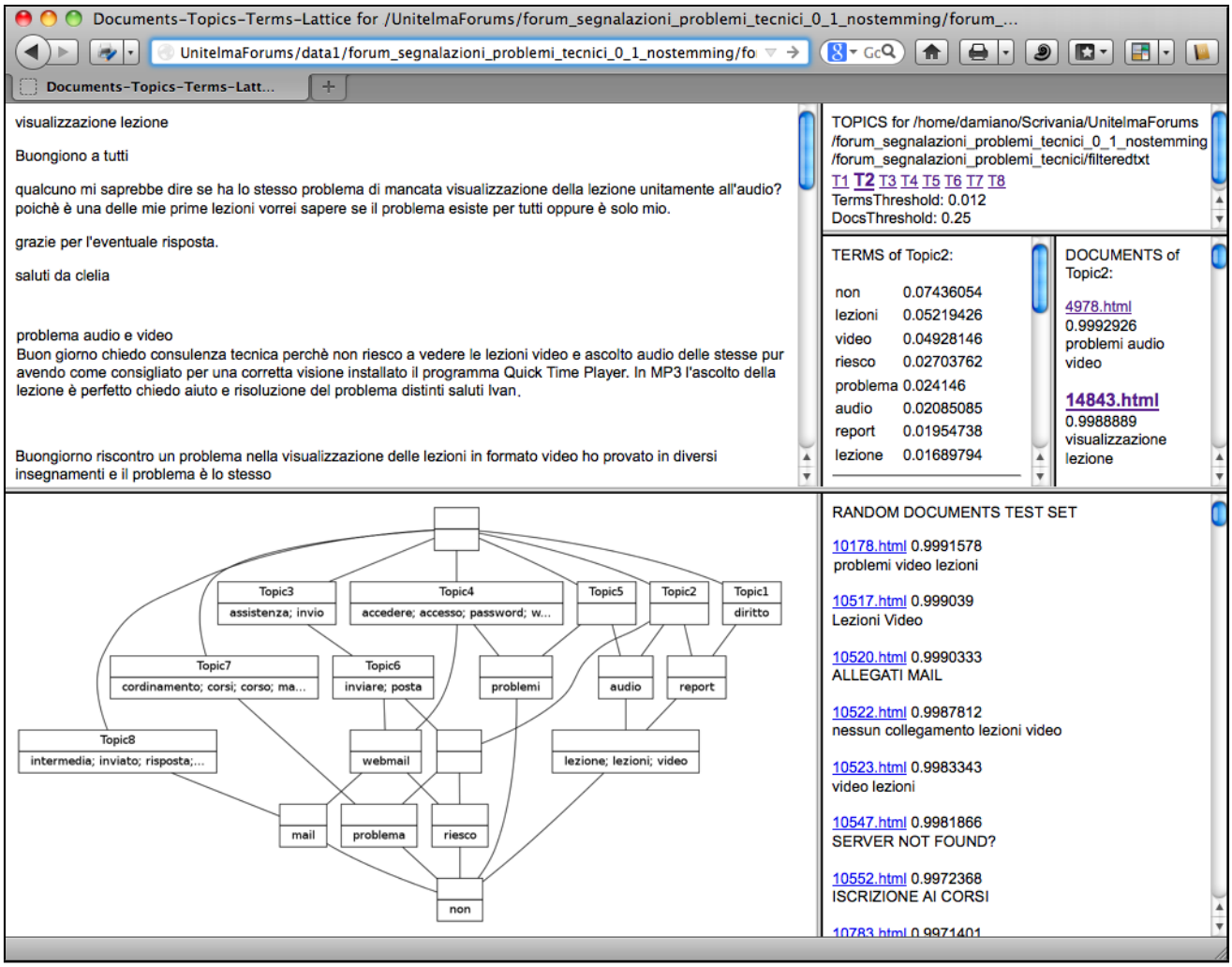


Fig. 5. A screenshot of the tool supporting the evaluation of documents-to-topics assignment process.

grouped into similar discussion topic classes. Association Rules for classification has been widely adopted in e-learning, in particular recommendation systems [24], [25], learning material organization [26], student learning assessments [27], course adaptation to the students behavior [28], and evaluation of educational websites [29]. Recent research trends in educational research recommend the development of cooperative learning and knowledge sharing inside student groups [30]. To this aim, the next stage of Web developments for education works on the opportunities raised by mixing the Social and the Semantic Web [31] and on adopting Semantic and Artificial Intelligence techniques for discovering information objects and restructure large digital collections [32].

V. Conclusions and Future Work

An important shift in educational focus is that for remembering large amounts of information it is necessary to develop the ability to quickly find the relevant information. Discussion forums represent one of the main asynchronous communication means offered by any learning management system and discussions taken place and stored in them represent a source of information for learners accessing the forum afterwards.

Their effectiveness as information sources, *i.e.*, the capability to satisfy users information needs, depends on their information richness first, but also on how discussion are organized and effectively moderated. In this paper we proposed an approach to restructure the native flat structure of a forum. It is able to detect discussion topics in a forum and to enhance its organization by providing a hierarchical topic-driven access structure to its messages. The method adopts information retrieval techniques, including topic models and formal concept analysis, to automatically discover discussion topics, organize them in a hierarchical way, and associate forum messages and discussion threads to the most appropriated identified topics. The hierarchy of topics is intended to offer an alternative access structure enabling surfing the forum content by topics, instead of just by time of conversations, thus to ease information search and reducing the chances for redundant discussions to appear. We have presented the results from a case study that we have conducted to validate both the

appropriateness of the identified topics and the correctness of the assignment of forum messages to them. The study has shown that the approach is valuable and promising.

Threats to validity of the obtained results lie mainly in the size of the conducted case study (which involved four forums of a single e-learning environment) and in the need for assessing the benefits of the additional forum navigation structure to forum users. To address such threats we aim to perform an experiment in which we will apply the approach to a larger number of forums from different e-learning contexts and to survey representative sample of users of each forum on the perceived usefulness and usability of the new complementary forum access structure.

Additional research directions in which we aim to extend our work are: i) improve the precision of the approach by better tuning the different available input parameters; ii) adopt fold-in techniques to apply the approach in an incremental way as new messages are added to a forum; iii) extend the approach in order to classify a new message over current identified topics as it is added to the forum; iv) implement the approach as a Moodle plug-in in order to be smoothly integrated in one of the most adopted open-source learning management system on the market.

REFERENCES

[1] B. Bickart and R. M. Schindler, "Internet forums as influential sources of consumer information," *Journal of Interactive Marketing*, vol. 15, pp. 31–40, 2001.

[2] T. W. Gruen, T. Osmonbekov, and A. J. Czaplewski, "eWOM: The impact of customer-to-customer online know-how exchange on customer value and loyalty," *Journal of Business Research*, vol. 59, p. 449456, 2006.

[3] J. Otterbacher, "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, paper 207.

[4] Q. Li, J. Wang, Y. P. Chen, and Z. Lin, "User comments for news recommendation in forum-based social media," *Information Sciences*, vol. 180, p. 49294939, 2010.

[5] W. Macias, "Health-related message boards/chat rooms on the web: Discussion content and implications for pharmaceutical sponsorships," *Journal of Health Communication*, vol. 10, p. 209223, 2005.

[6] (2012) Microsoft msdn developer network forums. [Online]. Available: http://social.msdn.microsoft.com/Forums/en/categories/

[7] S. Hrastinski, "What is online learner participation? a literature review," *Computers & Education*, vol. 51, no. 4, pp. 1755 – 1765, 2008.

[8] H. Stefan, "A theory of online learning as online participation," *Computers & Education*, vol. 52, no. 1, pp. 78–82, 2009.

[9] K. Zhang and K. Peck, "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*, vol. 29, no. 3, 2003.

[10] (1995) Network working group rfc 1855 - netiquette guidelines. [Online]. Available: http://www.ietf.org/rfc/rfc1855.txt

[11] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, March 2003.

[13] D. M. Blei, "Introduction to probabilistic topic models," *Communications of the ACM*, 2011. [Online]. Available: http://www.cs.princeton.edu/~blei/papers/Blei2011.pdf

[14] H. M. Wallach, D. M. Mimno, and A. McCallum, "Rethinking lda: Why priors matter," in *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009*, 2009, pp. 1973–1981.

[15] B. Ganter and R. Wille, *Formal concept analysis: mathematical foundations*. Springer, 1999.

[16] G. Birkhoff, "Lattice theory," in *Colloquium Publications*, 3rd ed. Amer. Math. Soc., 1967, vol. 25.

[17] M. Meila, "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*, 2003, pp. 173–187.

[18] A. Bakalov, A. McCallum, H. M. Wallach, and D. M. Mimno, "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*, 2012, pp. 237–240.

[19] F. Castro, A. Vellido, A. Nebot, and F. Mugica, "Applying data mining techniques to e-learning problems," in *Evolution of Teaching and Learning Paradigms in Intelligent Environment*, ser. Studies in Computational Intelligence, L. Jain, R. Tedman, and D. Tedman, Eds. Springer Berlin Heidelberg, 2007, vol. 62, pp. 183–221.

[20] M. Hanna, "Data Mining in the e-Learning Domain," *Campus-Wide Information Systems*, vol. 21, no. 1, pp. 29–34, 2004.

[21] T. Tang and G. McCalla, "Smart Recommendation for an Evolving e-Learning System: Architecture and Experiment," *International Journal on e-Learning*, vol. 4, no. 1, pp. 105–129, 2005.

[22] F. Castro, A. Nebot, and F. Mugica, "Extraction of logical rules to describe students' learning behavior," in *Proceedings of the sixth conference on IASTED International Conference Web-Based Education - Volume 2*, ser. WBED'07. Anaheim, CA, USA: ACTA Press, 2007, pp. 164–169. [Online]. Available: http://dl.acm.org/citation.cfm?id=1323159.1323189

[23] S. C. P. Sung Ho Ha, Sung Min Bae, "Web mining for distance education," pp. 715–719 vol.2, 2000.

[24] O. R. Zaíane, "Building a recommender agent for e-learning systems," in *Proceedings of the International Conference on Computers in Education*, ser. ICCE '02. Washington, DC, USA: IEEE Computer Society, 2002, pp. 55–. [Online]. Available: http://dl.acm.org/citation.cfm?id=838238.839230

[25] Q. Yang, J. Sun, J. Wang, and Z. Jin, "Semantic web-based personalized recommendation system of courses knowledge research," in *Proceedings of the 2010 International Conference on Intelligent Computing and Cognitive Informatics*, ser. ICICCI '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 214–217. [Online]. Available: http://dx.doi.org/10.1109/ICICCI.2010.54

[26] C.-J. Tsai, S.-S. Tseng, and C.-Y. Lin, "A two-phase fuzzy mining and learning algorithm for adaptive learning environment," in *Proceedings of the International Conference on Computational Science-Part II*, ser. ICCS '01. London, UK, UK: Springer-Verlag, 2001, pp. 429–438. [Online]. Available: http://dl.acm.org/citation.cfm?id=645456.654828

[27] C. Romero, S. Ventura, and P. D. Bra, "Knowledge discovery with genetic programming for providing feedback to courseware authors," *User Modeling and User-Adapted Interaction*, vol. 14, no. 5, pp. 425–464, Jan. 2005. [Online]. Available: http://dx.doi.org/10.1007/s11257-004-7961-2

[28] M. A. Hogo, "Evaluation of e-learning systems based on fuzzy clustering models and statistical tools," *Expert Syst. Appl.*, vol. 37, no. 10, pp. 6891–6903, Oct. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2010.03.032

[29] L. dos Santos Machado and K. Becker, "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*. IEEE Computer Society, 2003, pp. 360–361.

[30] A. Jakobsone, V. Kulmane, and S. Cakula, "Structurization of information for group work in an online environment," in *Global Engineering Education Conference (EDUCON), 2012 IEEE*, april 2012, pp. 1 –7.

[31] M. Ghenname, R. Ajhoun, C. Gravier, and J. Subercaze, "Combining the semantic and the social web for intelligent learning systems," in *Global Engineering Education Conference (EDUCON), 2012 IEEE*, april 2012, pp. 1 –6.

[32] A. Martin and C. Leon, "An intelligent e-learning scenario for knowledge retrieval," in *Global Engineering Education Conference (EDUCON), 2012 IEEE*, april 2012, pp. 1 –6.