Topic Extraction and Selection: Online Discussion Forums at a Glance

Bernardo Pereira Nunes[*], Alexander Mera[*], Ricardo Kawase[†], Marco A. Casanova[*], Gilda Helena B. de Campos[‡]
[*]*Department of Informatics - PUC-Rio - Rio de Janeiro, RJ - Brazil*
{bnunes, acaraballo, casanova}@inf.puc-rio.br
[†]*L3S Research Center, Leibniz University Hannover, Germany*
kawase@l3s.de
[‡] *Department of Education - PUC-Rio - Rio de Janeiro, RJ - Brazil*
gilda@ccead.puc-rio.br

***Abstract*—Forums play a key role in the process of knowledge creation, providing means for users to exchange ideas and to collaborate. However, educational forums, along several others online educational environments, often suffer from topic disruption. Since the contents are mainly produced by participants (in our case learners), one or a few individuals might change the course of the discussions. Thus, realigning the discussed topics of a forum thread is a task often conducted by a tutor or moderator. In order to support learners and tutors to harmonically align forum discussions that are pertinent to a given lecture or course, in this paper, we present a method that combines semantic technologies and a statistical method to find and expose relevant topics to be discussed in online discussion forums. We surveyed the outcomes of our topic extraction and selection method with students, professors and university staff members. Results suggest the potential usability of the method and the potential applicability in real learning scenarios.**

I. Introduction

Over the past decade, the World Wide Web became an important source of information and knowledge. The diversity and engagement of independent users and communities contributed to the creation and proliferation of a rich set of content available in different communication channels (such as social media, real-time channels, blogs, forums, etc) as well as in formats (such as text, audio and video).

In particular, online discussion forums have played a key role in the process of knowledge creation, providing means for its users to exchange ideas, form opinions, position themselves and collaborate. As an outcome of the importance of online discussion forums is Wikipedia[1], where for each Wikipedia article there is a forum-based page[2] that relies on the collaboration, discussion, consensus and collective effort of its users to keep Wikipedia constantly updated and curated.

Due to the benefits generated by users' participation in forums, most online courses combine educational materials and online discussion forums. However, even though forums clearly leverage the creation of collective intelligence [10], the assessment of users' participation is rather difficult. Depending on the number of students and posts, manual assessment becomes impractible. Previous work addressed the problem of assessing the quality of students' participation [8], [7]. However, they do not take into account whether a particular set of topics were addressed in a thread of a specific discipline.

Furthermore, different backgrounds in online discussion forums may lead a discussion to unforeseen directions, needing external support to realign the discussed topics of a forum thread. This task is often conducted by a tutor or moderator. But, as we show in this paper, on the average, 50% of forums discussing a specific subject with different audience or tutor/moderator cover distinct topics. Even though forum discussions are often different, a set of specific topics must be addressed to achieve the course goals. Therefore, if a given forum does not cover a set of expected topics, the assessment of the students who participated in distinct forums (with the same subject) might be hampered, since the acquired knowledge depends on the topics discussed in the forum.

In this paper, we combine semantic technologies and a statistical method to find, expose and recommend relevant topics to be discussed in online discussion forums. Briefly, with the help of semantic tools, the proposed method first performs named entity recognition (NER) and topic extraction, followed by a statistical approach that selects and ranks the most representative topics. The method outputs the topmost representative topics discussed in a specific forum as well as a set of suggested topics to be discussed. We used 97 online forums from a Brazilian university to validate and assess our method.

The main contributions of this paper are: (i) high-level assessment of tutor/moderator progress; (ii) topic recommendation; (iii) forum coverage; (iv) equalise knowledge acquisition by students after forum participation in specific topics; and (v) provide a better overview of the discussed topics.

The rest of this paper is organised as follows. Section II discusses and compares relevant research. Section III describes the use of forums in our context and motivation. Section IV introduces the topic extraction and selection

[1] http://www.wikipedia.org

[2] http://en.wikipedia.org/wiki/Help:Using_talk_pages

method. Sections V and VI present the evaluation setup and results of our method, respectively. Finally, Section VII discusses our outcomes and outlines future works.

II. Related Work

Li and Wu [6] combine approaches involving sentiment analysis and text minining to detect hotspot forums within a certain time span. Their method assists users to make decisions and predictions over polarised groups of messages in online forums. Despite not performing topic extraction in the hotspot forums, the emotional polarity information for each topic extracted would help users on understanding how a given topic is addressed in a discussion.

Cong et al. [1] present an approach for finding question-answer pairs in online forums based on Labeled Sequential Patterns (LSPs) and graph-based propagation model. While the creation of patterns for interrogative senteces is made using part-of-speech tags, the answers are detected and ranked using KL-divergence language model. Again, our approach is complementary to their approach, since our approah would serve as a filter for finding question and answers based on topics. Conversely, our method would benefit of this approach by identifying key posts in a online discussion.

Online forums play a key role in the student skills development as shown by Scaffidi et al. [9]. Their study focuses on the types of posts that facilitate discussions and collaboration amongst novice developers. The study of user behavior in online forums help to promote active interaction amongst users and therefore the construction of collective knowledge. We believe that the introduction of new topics to be discussed by such community of users could trigger new discussions and hence new knowledge.

Desanctis et al. [4] provide an interesting discussion about e-venues for learning such as video-conferenced classrooms, online communities and group discussion spaces. Although each venue influences the learning process of a particular group, they all have in common the need to bring new discussions that promote the development of knowledge of the participants. For instance, online communities usually last more than private group discussion spaces, since new participants with fresh questions can drop in at anytime. Thus, in order to maintain the group discussion, the recommendation of new topics for discussion would foster longer interactions amongst participants and knowledge refreshment.

Evidently, online discussions can also be fueled by tutors responsible for bringing new topics and questionings for the discussion. Previous studies [2] have shown that tutored venues can improve both retention and performance of the participants. In this paper, we use the tool for assisting tutors on addressing new topics relevant to the discussion.

III. Motivation

To illustrate the motivation of our research, we describe two scenarios where participants of online discussion forums would benefit from our method. Both scenarios result from the need of the staff from a Brazilian university to assess online discussion forums.

As online discussion forums are fundamental in the learning process, most of the online courses take advantage of their use to meet specific goals. However, assessing student participation in forums is not a simple task, and due to the high number of posts, it can become impracticable. Hence, in order to maintain the quality of teaching and student experience, the university staff members required a tool to track the discussion progress.

The first scenario described by the university staff members is that tutors constantly overlook the discussion of relevant topics in favor of a better flow in the forum. However, although the discussion flow is of utmost importance, tutors must conduct the forum in such a way that specific topics must be addressed and, at the same time, preserve the discussion flow. Hence, the university staff members are interested in the analysis of forums to check if particular topics were covered in the discussion. In this way, they can ensure that all participants had similar experience and learning situations that can contribute to the next activities. In the case that a set of topics are not covered, they would like to intervene and extend the forum closure or create a new forum thread to discuss the missing topics.

As a consequence of the first scenario, the second scenario aims at fostering the discussion with suggestions that may assist students in the discussion. For many reasons, some forums lack of interaction and students must be encouraged to participate. In this manner, university staff members believe that a recommendation tool would promote the discussion and help to reach the forum's goal.

Therefore, the current work assists the university staff members and students to have a better overview of what is happening in the forum to take the right action and create learning situations that can improve the learning experience of the students.

IV. Topic Extracion and Selection

In this section we present the main steps of a coherent process chain that semantically and statiscally selects the most relevant discussed topics in a given online discussion forum. The process chain is composed of three steps described as follows: (i) Entity Extraction and Enrichment; (ii) Topic Extraction; and (iii) Topic Selection.

A. Entity Extraction and Enrichment

When dealing with online discussion forums, we are essentially working with unstructured data, which in turn hinders data manipulation and the identification of atomic elements in texts. To alleviate this problem, information extraction (IE) methods, such as Named-Entity Recognition (NER) and name resolution, are employed. These tools automatically extract structured information from unstructured data and link to external knowledge bases in the Linked Open Data cloud (LOD), such as DBpedia[3].

For instance, after processing the following sentence using an IE tool: “I agree with Barack Obama that the whole episode should be investigated.”, the entity “Barack Obama” is annotated and classified as *person* and linked to the DBpedia resource ⟨http://dbpedia.org/resource/Barack_Obama⟩, where structured information about him is available.

We use the DBpedia Spotlight tool[4] to extract and enrich entities found in the posts within a forum thread. DBpedia Spotlight adds markups with semantic information surrounding atomic elements (entities) in the forum posts. Note that our method is language independent as long as we have a solid repository of entities (such as DBpedia or Freebase[5]) and a proper annotation tool (such as Spotlight).

B. Topic Extraction

Given as starting point the entities that were found in the previous step, the topic extraction step begins by traversing the entity relationships to find a more general representation of the entity, i.e., the topics.

An entity is conventionally represented as a RDF (Resource Description Framework) triple in the form of (Subject, Predicate, Object), where each triple represents a fact, and the predicate names the relationships between the subject and the object. For example, a triple is (“Barack Obama”, “isPresidentOf”, “United States of America”). Furthermore, a set of RDF triples form a directed and labeled graph, where the nodes are a set of subjects and objects and the edges are represented by the predicate.

Thus, for each extracted and enriched entity in the posts, we explore their relationships through the predicate *dcterms:subject*, which by definition[6] represents the topic of the entity. In that sense, to retrieve the topics, we use SPARQL query language for RDF over the DBpedia SPARQL endpoint[7], where we navigate up in the DBpedia hierarchy to retrieve broader semantic relations between the entities and its topics. As it is shown in the following SPARQL query, we use the predicate *skos:broader*.

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

 SELECT DISTINCT ?l1 ?l2 ?l3 ?l4
 WHERE {
        <entity_uri> dcterms:subject ?l1 .
        ?l1 skos:broader ?l2 .
        ?l2 skos:broader ?l3 .
        ?l3 skos:broader ?l4
      } LIMIT 1000;
```

The variable *entity_uri* represents the entity in which we are interested in retrieving the topics extracted from the posts in a forum thread, while the variables *l1* to *l4* represent the topics that will be retrieved from the entity. Thus, given an entity, the topics of an entity are retrieved through the predicate *dcterms:subject* and *skos:broader*. The latter predicate is used to get a more general representation of the topic. This strategy will help us on finding the topics that best cover a forum thread.

Note that an entity/concept can be found in different levels of the hierarchical categories of DBpedia, and hence this approach would lead us to retrieve topics in different category levels. However, as in [5], we take advantage of the co-occurrence of the topics in the different levels to find the most representative ones (see Section IV-C).

C. Topic Selection

Finally, in this last step, we select the most representative topics extracted from the posts that belong to a forum thread. For this, we rely on *tf-idf* (term frequency - inverse document frequency) score to statistically measure the importance of a topic in a forum thread.

Typically, *tf-idf* is used on information retrieval and text mining to measure the importance of a word to a document in a collection. However, in this paper, we adapted this metric to take into account entities and topics extracted from the posts instead of words.

Thus, to select the most representative topics, we compute *tf-idf* score twice, one for the entities extracted from the forum thread (i.e. the most representative entities in the collection) and another for the topics extracted from the entities (see Section IV-B).

Basically, to compute the term frequency (*tf*), we count the number of occurrences of an entity e in a post $p \in P$. As for the inverse document frequency (*idf*), we compute the (*idf*) score by dividing the total number of posts $|P|$ by the number of posts containing the entity $|P_e|$, see Eq. 1.

$$tfidf(e,p,P) = tf(e,p) \times idf(e,P) \tag{1}$$

where *tf* is the raw frequency of a term in a post $tf(e,p) = frequency(e,p)$, and *idf* is the measure of commoness/rareness of an entity in a collection P given by the following equation: $idf(e,P) = log(\frac{|P|}{|P_e|})$.

After computing the *tf-idf* score for each entity, the topmost representative entities are selected. From the selected entities, the topics are extracted according to the process described in Section IV-B.

With the topics in hands, we then compute the *tf-idf* score over the topics extracted from the entities and decreasingly rank them. Again, the topmost representative topics for a given forum thread are selected. Note that the number of topics that represent a forum is choosen by the user (in our case, the top 10 relevant topics). Finally, the top ranked topics are selected to represent the forum thread topics.

[3] http://www.dbpedia.org

[4] http://dbpedia-spotlight.github.io/demo/

[5] http://www.freebase.com

[6] http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=terms#elements-subject

[7] http://pt.dbpedia.org/sparql - DBpedia SPARQL endpoint in portuguese.

V. Evaluation Setup

Over the course of our study, real data from online discussion forums were used to perform a comprehensive evaluation of our method. Our method was evaluated using 97 online discussion forums containing in total 10,785 anonymised posts provided by the distance education department of a Brazilian university. All selected forum threads to the evaluation occurred at least twice. Furthermore, each professor assessed the suggested topics from forums conducted by themselves.

Our main objectives included a thorough assessment of the recommendation of topics based on previous online discussion forums as well as the assessment of the selected topics that cover a forum discussion. For this, we submitted 3 questionnaires to 11 students, 4 professors and 3 coordinators of the distance education department to gather different perspectives and views of the proposed method.

The questionnaires were divided into three different categories of questions, namely *perceived usefulness*, *perceived ease-of-use* and additional suggestions. Basically, the questions followed the technology acceptance model (TAM) proposed by Davis [3], arguably the most influential "Technology Acceptance Theory". A Likert scale of 5-point of agreement and frequency were applied on the questionnaire.

Briefly, this theory states that there are two key aspects to measure users' intention to adopt a new technology, the *perceived usefulness* and *perceived ease-of-use*. Perceived usefulness (PU) refers to "the degree to which a person believes that using a particular system would enhance his or her job performance", while perceived ease of use (PEOU) refers to "the degree to which a person believes that using a particular system would be free of effort" [3].

Each questionnaire was divided in 6 PU questions, 6 PEOU questions and additional 3 opinion mining questions where we asked participants for further suggestions. Note that, in the case of university staff members, the the topics were assessed over two randomly choosen forum threads, since they do not participate of the forum. Thus, a list of topics discussed in the forum and a list of suggested topics for each forum thread was available for their evaluation. As they are staff members of the university, they also have access to the forum discussions in case the would need additional information.

VI. Results

The results of the questionnaires are summarised in Figure 1. The error bar charts show that all participants reported a high positive perception for the proposed topics, the implications and the applicability of the results. In particular, professors had a slightly better acceptance, when compared to the other tiers of participants. The coefficient of internal consistency Cronbach's α of 0.65 for PU and 0.72 for PEOU indicated a good reliability of the results. These questionnaire results suggest the potential usability of our proposed topic extraction and selection method.

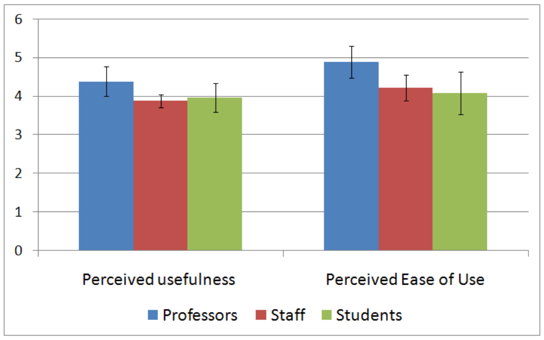

Figure 1. Error Bars for survey questions regarding perceived usefulness and perceived ease-of-use.

Regarding the suggestions included in the questionnaires, we observe that the most controversial question referred to whether or not the recommended topics should be available for professors, students or both. All professors suggested that the topics should be available only to them. All staff members suggested that topics should be available to both. Interestingly, students did not come to a common agreement. While the majority (64%) agreed that suggested topics should be available to professors and students, 36% opined that topic suggestions should be available only to professors.

We believe that the controversy is raised by the different backgrounds each group of participants had and the understanding they had of the topics. Staff members, who are not effectively involved in the online forums, assumed that the discussed topics should come out from an agreement between professors and students. On the other hand, the opinion of professors that a tool should present topics recommendation directly to them in fact reflects their need to control those around them. Finally, the split students' opinions lie in the fact that some students are still skeptical that online educational forums can smoothly evolve without proper moderation.

Unlike the questionnaires given to students and university staff members, professors' questionnaire had an additional question regarding whether other professors can benefit from the suggested topics. The results reported that 75% of the professors strongly agree that *other* professors would take advantage of the suggested topics.

Finally, all university staff members and professors who participated in the survey (strongly) agree that the assessment of students would be facilitated if disparate forums addressed the same topics. Likewise, all university staff members (strongly) agree that the proposed method would help in the assessment of the professor regarding the coverage of topics addressed in the forums. Nevertheless, 88% of all participants agree that the use of such method should be optional.

VII. Discussion and Outlook

We presented a method for automatically generating topics that represent a forum thread in distance learning environments. Basically, we combined semantic and statistical techniques in a coherent process chain to extract, select and rank the most relevant topics of a forum.

Our experiments showed that most professors, university staff members and students are willing to use our proposed approach in future forums. Moreover, 75% of the professors reported that other professors would benefit from the suggested topics.

Reviewing a sample of 97 forum threads, we verified that, on the average, 50% of the topics discussed in disparate forums addressing the same subject are different. This situation resulted in a concern with regard to the topics addressed in the forums and the post assessment of the students. A priori, students in disparate forums covering the same subject should have a similar experience and learn the same topics.

Thus, providing a method to overview the topics discussed in different forums will help university staff members, such as course coordinators, to rapidly intervene in forums that topics are being overlooked.

In theory, the use of the proposed method would bring more control of what is being taught in a forum and, therefore, ensure quality. In practice, this can be different and some considerations arise out of the purpose of the use of the proposed method by a few interviewed respondent.

A first consideration lies in the freedom of professors in guiding forums. As every professor has its own teaching style and may also have a different point-of-view when they approach a subject, the concern of having to address specific topics in a forum might decrease the creativity and engagement of some professors. On the other hand, assistant professors may also take advantage of the suggested topics to guide the forum.

Another consideration regards the use of the proposed method is in the use of the topics by students, in case topic suggestions are also available for them. In the same time a topic suggestion may trigger an insight or make some students more confident, other students may stick only to the suggested topics and inhibit discussions of various other relevant topics.

A last consideration raised due to space restrictions is dependent on the type of the forum and course being taught. In many courses, the subjects change over time and using automatically suggested topics from previous forums threads might hinder the discussion flow. Although the main topics would hold out in future discussions, the list of suggested topics must be from time to time manually updated.

In general, the proposed method aims at assisting university staff members, professors and students to have a better overview of what is being discussed in the forum and, therefore, enable professors to take more informed actions to preserve discussion flow, improve students' experience and ensure topics coverage.

Our method also provides to the university staff members the possibility of assessing forum coverage, tracking what students are learning in different forums and, in some cases, detecting deviations in the topics addressed in the forums. To adopt or not the proposed method in an online course, we believe that it depends on the instructional design of the course. The set-up of the course is crucial to determine which methods or tools must be used.

As for future work, we plan to expand the method to accept external topic suggestions. For instance, professors involved in the course can also add topics to the discussion. Furthermore, we also plan to create a Moodle plugin.

References

[1] G. Cong, L. Wang, C.-Y. Lin, Y.-I. Song, and Y. Sun. Finding question-answer pairs from online forums. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 467–474, New York, NY, USA, 2008. ACM.

[2] J. A. Cottam, S. Menzel, and J. Greenblatt. Tutoring for retention. In *Proceedings of the 42nd ACM technical symposium on Computer science education*, SIGCSE '11, pages 213–218, New York, NY, USA, 2011. ACM.

[3] F. D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, pages 319–340, 1989.

[4] G. DeSanctis, A.-L. Fayard, M. Roach, and L. Jiang. Learning in online forums. *European Management Journal*, 21(5):565 – 577, 2003.

[5] B. Fetahu, S. Dietze, B. P. Nunes, D. Taibi, and M. A. Casanova. Generating structured profiles of linked data graphs. In E. Blomqvist and T. Groza, editors, *International Semantic Web Conference*, volume 1035 of *CEUR Workshop Proceedings*, pages 113–116. CEUR-WS.org, 2013.

[6] N. Li and D. D. Wu. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, 48(2):354 – 368, 2010.

[7] M. Pendergast. An analysis tool for the assessment of student participation and implementation dynamics in online discussion forums. *SIGITE Newsl.*, 3(2):10–17, June 2006.

[8] C. Romero, M.-I. López, J.-M. Luna, and S. Ventura. Predicting students' final performance from participation in on-line discussion forums. *Comput. Educ.*, 68:458–472, Oct. 2013.

[9] C. Scaffidi, A. Dahotre, and Y. Zhang. How well do online forums facilitate discussion and collaboration among novice animation programmers? In L. A. S. King, D. R. Musicant, T. Camp, and P. T. Tymann, editors, *SIGCSE*, pages 191–196. ACM, 2012.

[10] A. L. Veerman, J. E. B. Andriessen, and G. Kanselaar. Collaborative learning through computer-mediated argumentation. In *Proceedings of the 1999 Conference on Computer Support for Collaborative Learning*, CSCL '99. International Society of the Learning Sciences, 1999.