A Topic Extraction Process for Online Forums

Bernardo Pereira Nunes
Department of Informatics - PUC-Rio
Rio de Janeiro, RJ - Brazil
bnunes@inf.puc-rio.br

Alexander Mera
Department of Informatics - PUC-Rio
Rio de Janeiro, RJ - Brazil
acaraballo@inf.puc-rio.br

Ricardo Kawase
L3S Research Center
Appelstrasse 9, 30167 Hannover, Germany
kawase@L3S.de

Besnik Fetahu
L3S Research Center
Appelstrasse 9, 30167 Hannover, Germany
fetahu@L3S.de

Marco A. Casanova
Department of Informatics - PUC-Rio
Rio de Janeiro, RJ - Brazil
casanova@inf.puc-rio.br

Gilda Helena B. de Campos
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.**

***Keywords*-Topic coverage, Topic extraction, Discussion forum, Topic recommendation, Forum assessment**

I. Introduction

With the catch up of the Web 2.0, the World Wide Web became not only a source of information, but also a sharing platform. Any individual with internet access is able to consume and produce content that becomes immediately available for millions. Among the many available communication channels, such as social networks, instant messengers, blogs, etc., forums, in particular, have played a key role in the process of knowledge creation.

However, even though forums clearly leverage the creation of collective intelligence [5], the assessment of users' participation is rather difficult. Depending on the number of students and posts, manual assessment becomes impractical. Previous work addressed the problem of assessing the quality of students' participation [3], [4]. However, they do not take into account whether a particular set of topics were addressed in a thread of a specific discipline.

In this paper, we propose a topic extraction process that combines 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.

II. Motivation

To illustrate the motivation of our research, we describe one scenario where participants of online discussion forums would benefit from our method.

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.

A 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.

III. Topic Extraction and Selection

In this section we present the main steps of a coherent process chain that semantically and statistically selects the most relevant discussed topics in a given online discussion

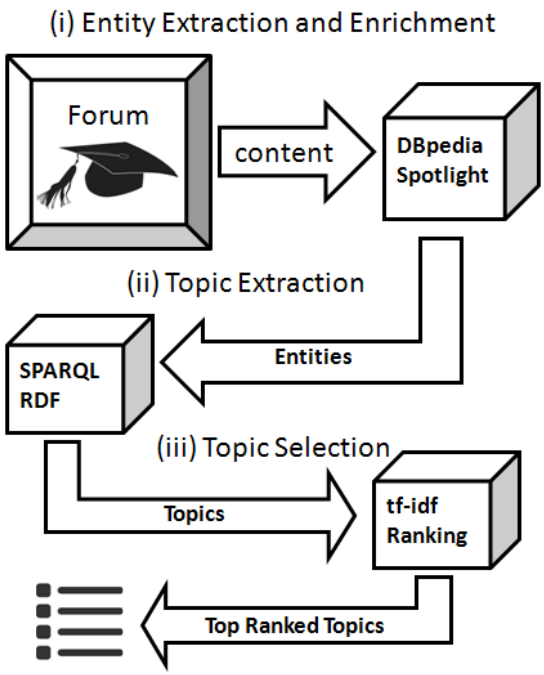

Figure 1. Topic extraction process workflow.

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

We use the DBpedia Spotlight tool[1] 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 (as in [2]). Note that our method is language independent as long as we have a solid repository of entities (such as DBpedia or Freebase[2]) and a proper annotation tool (such as Spotlight).

B. Topic Extraction

For each extracted and enriched entity in the posts, we explore their relationships through the predicate *dcterms:subject*, which by definition[3] 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[4], where we navigate up in the DBpedia hierarchy to retrieve broader semantic relations between the entities and its topics.

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 [1], we take advantage of the co-occurrence of the topics in the different levels to find the most representative ones (see Section III-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.

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 chosen by the user (in our case, the top 10 relevant topics). Finally, the top ranked topics are selected to represent the forum thread topics.

IV. Primarily Results

We applied our proposed method in 97 forum threads containing in total 10,785 anonymized posts provided by the distance education department of a Brazilian university. 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.

V. 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.

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.

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.

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.

[1] http://dbpedia-spotlight.github.io/demo/
[2] http://www.freebase.com
[3] http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=terms#elements-subject
[4] http://pt.dbpedia.org/sparql - DBpedia SPARQL endpoint in portuguese.

References

[1] 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.

[2] B. P. Nunes, A. Mera, M. A. Casanova, and R. Kawase. Boosting retrieval of digital spoken content. In *Knowledge Engineering, Machine Learning and Lattice Computing with Applications*, volume 7828 of *Lecture Notes in Computer Science*, pages 153–162. Springer Berlin Heidelberg, 2013.

[3] 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.

[4] 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.

[5] 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.