Towards a process to support solving the content selection problem from online community forums

Dárlinton B. F. Carvalho[1], **Ricardo M. Marcacini**[2],
Carlos J. P. de Lucena[1], **Solange O. Rezende**[2]

[1]Pontifical Catholic University of Rio de Janeiro (PUC-Rio)
Department of Computer Science – Rio de Janeiro, Brazil

[2]Mathematical and Computer Sciences Institute - ICMC
University of Sao Paulo (USP) – São Carlos, SP, Brazil

{dcarvalho,lucena}@inf.puc-rio.br,{rmm,solange}@icmc.usp.br

Abstract. *There are plenty of public content available on the Internet, especially in online communities, enabling researchers to study society in new ways. Since the qualitative content analysis of online forums is very time consuming, the following problem arises: how to select the content to be analyzed? This paper introduces a new process to support solving this problem. This process is based on unsupervised machine learning techniques and provides consolidated and structured results. This includes measurements and a content exploration method. A tool that helps to apply the proposed process was created and is presented as well.*

1. Introduction

The use of the Internet as a communication medium brings a new paradigm in the information sharing in our society. Online communities [Preece and Maloney-Krichmar 2005] are important meeting points for people in this virtual world, leaving records of their discussions in the online fora. Newcomers can use the body of knowledge present in these discussions to learn new things, and so researchers can study about these communities, analyzing the same content.

Influenced by ethnographic principles, Kozinets [Kozinets 2009] says that the Netnography observation happens in users' natural habitat, being relevant for its users, detailed and contextualized in community, and retrieved in non-intrusive ways, enabling an opportune, effective and efficient processing. In this way, users are not summoned to participate in a reactive fashion (e.g. online surveys), which enables the analysis of freely constructed manifestations. An opportunistic approach that explores social media in order to conduct such studies is presented in [Carvalho et al. 2012]. A technique that helps to perform this analysis is the "Discourse of the Collective Subject", a qualitative technique with roots in the Theory of Social Representations [Lefevre and Lefevre 2006]. The aim of this technique is to identify collectives, aggregated by central ideas, and describe them through a created discourse based on a patchwork from the collective members' speeches, synthesizing the discourse as one collective subject.

Since the amount of data in online forums is huge and most of the qualitative research techniques are very time consuming, the following problem arises: how to select the content to be analyzed? This paper introduces a new process to support solving this

problem. The problem of content selection can be summarized as the task to find discussion in online community forums in which content looks promising to answer study research questions. This problem has two distinct objectives, the first being to maximize the number of discussion participants (i.e. users) and the second being to minimize the number of topics to be analyzed (i.e. content volume). It is important to say that the discussion analysis must be performed considering the context of the topic, because the analysis of one interesting post requires the comprehension of the whole discussion as presented in other posts of the topic. Therefore, the solution of the problem of content selection for analysis is a set of topics selected from the forum of an online community.

In order to solve this problem, we propose a process, based on unsupervised machine learning techniques, that presents processed and structured results, including measurements and a content exploration method, to support final users (e.g. social researchers) in the task of selecting content for analysis. This process is based on hierarchical text clustering.

The hierarchical clustering strategy can be classified as agglomerative or divisive. In the agglomerative hierarchical clustering, initially each document is a singleton cluster and, in each iteration, the closest pair of clusters is unified, until they form only one cluster. In the other strategy, the divisive hierarchical clustering starts with a cluster containing all documents, which are iteratively divided into smaller clusters until there remains only singleton clusters. Experimental evaluations show that the algorithm UPGMA (agglomerative) and the Bisecting-kmeans (divisive) achieve the best results in textual data [Zhao et al. 2005]. However, it is noteworthy that these clustering algorithms were proposed to solve general-purpose tasks. Consequently, several studies have investigated text clustering approaches for specific applications. For instance, text clustering for online community analysis has recently gained the attention of the research community, because of the need to organize automatically the huge volume of texts published daily. The content selection problem from online forums is an aspect that, from the best of our knowledge, has not been explored so far. The proposed process adds to this body of work. The details of how a tool based in this process attempts to solve this task are presented in the next section.

2. A tool to support solving the content selection problem

The software tool developed to support the content selection from online communities is an extension of the Torch – Topic Hierarchies [Marcacini and Rezende 2010]. Our tool provides techniques for text preprocessing and hierarchical clustering algorithms. In addition, we developed a module for posts and topics recommendation and a visual interface to explore the clustering results from topics and posts.

The posts and topics collected from the online forums communities have several attributes. The attributes considered for each post were the text message of the post, the publication date of the message and the post author. These attributes are usual in online forums and were chosen to allow a wider application of the tool. Each post belongs to a particular topic of the online forum and the forum has many topics. Thus, the attributes considered for representation of the topics were the topic title, the period of existence (publication date of the first and last post) and the number of participants in the topic.

After collecting a set of posts and topics, the tool performs the textual data pre-

processing. The first step is the stopwords removal where pronouns, articles and prepositions are discarded. Then, the terms are simplified by using the Porter Stemming algorithm [Nogueira et al. 2008]. Thus, morphological variations of a term are reduced to its radical. A feature selection technique based on document frequency obtains a reduced and representative subset of terms. The term co-occurrence network obtained from preprocessed texts is the first structure available to the users for exploring the textual content of the online communities. Our tool allows the users to analyze significant relationships among terms through an interactive interface.

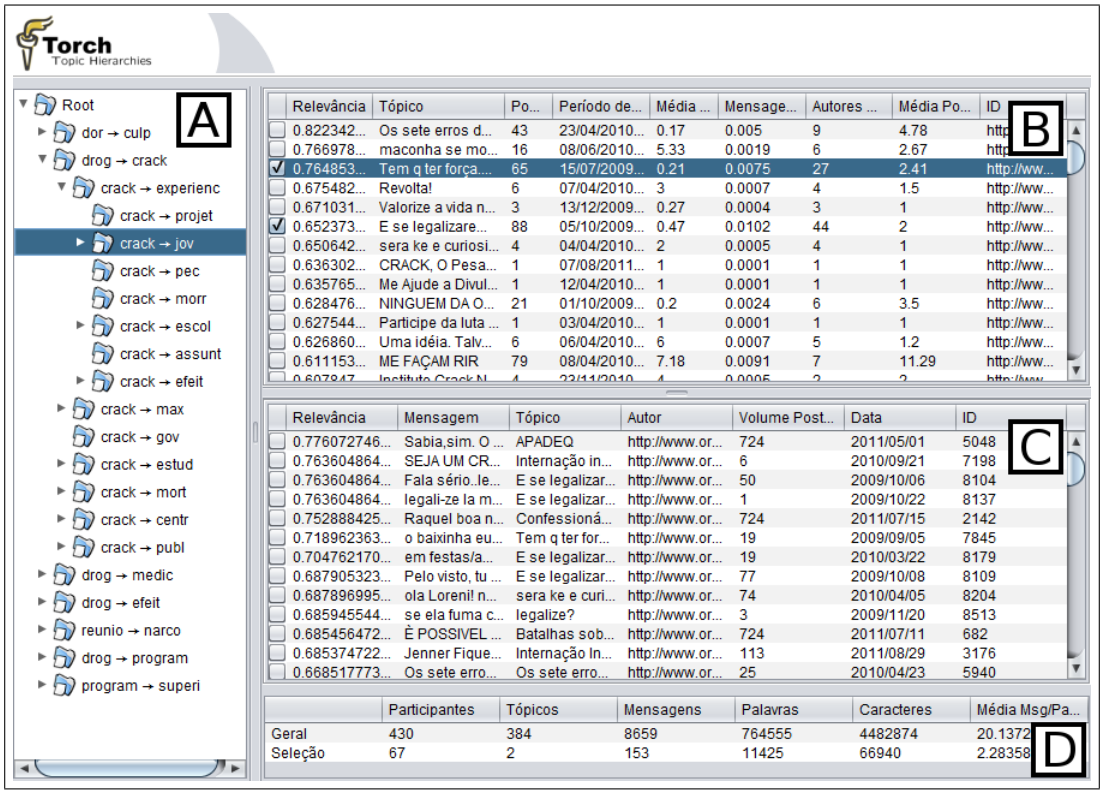

Figure 1. Screenshot of the tool built from the proposed process

Figure 1 shows the main interface of the tool created to support solving the content selection problem. The term co-occurrence network was summarized with hierarchical clustering. Thus, the various topics discussed in the forums are presented to the user in succinct themes (Figure 1A). When the users select a theme, the most relevant topics (Figure 1B) and posts (Figure 1C) are presented according to the ranking strategy. The user can select content through check boxes at the right side of the topic or post names. The selection of a post in the lower part (Figure 1C) automatically includes all posts of its topic, because the comprehension of the discussion requires all posts of the topic. The problem solving progression is described by a set of measures (Figure 1D) that describe the number of participants, that should be maximized, the content volume (topics, posts, words, characters), that should be minimized, and a relationship of both measures (median of posts per participant). The first line shows the measurement considering the whole forum content and the second has the selection measurement. It is up to the users to decide when the current solution (i.e. selection) is satisfactory, and these measures help them to make this decision.

3. Conclusion

The content selection problem is a pivotal factor for social scientists that embrace new ventures in conducting research based on the vast content available on the Internet, especially regarding online forum content. It has two objective goals, maximize the number of selected participants and minimize the content volume to be analyzed. These are conflicting goals. The problem solution is also driven by the research interests, that are not measurable (so far). Without the tool, the researchers rely only on the general metrics about the content of the topics, or they must look at the whole forum content to perform the content reducing task.

This paper introduces a process proposal to support solving the problem of content selection from online forums. The proposed process aims to support the researchers to tackle this problem in a smarter way, leveraging them with the best machine learning techniques available so far. These results include measurements and a content exploration method. As an application of the proposed process, a tool based on that process was created to aid researchers to apply it. The created tool is already useful, but it is still in an early stage of development. Further research in the measures used to calculate the posts similarities can also provide better results to the user, and improve the process.

ACKNOWLEDGEMENTS

This work has been sponsored by CNPq (Brazilian Council for Research and Development), process 142620/2009-2, and FAPESP (State of Sao Paulo Research Foundation) – process 2010/20564-8 and 2011/19850-9.

References

Carvalho, D., Madeira, W., Okamura, M., Lucena, C., and Zanetta, S. (2012). A practical approach to exploit public data available on the internet to study healthcare issues. In *Proceeding of XXXII Congresso da Sociedade Brasileira de Computação (CSBC) – XII Workshop de Informática Médica*, page to appear.

Kozinets, R. (2009). *Netnography: Doing Ethnographic Research Online*. Sage Publications Ltd, London.

Lefevre, F. and Lefevre, A. M. C. (2006). The collective subject that speaks. *Interface - Comunicação, Saúde, Educação*, 10(20):517–524.

Marcacini, R. M. and Rezende, S. O. (2010). Torch: a tool for building topic hierarchies from growing text collection. In *WTA'2010: IX Workshop on Tools and Applications. In 8th Brazilian Symposium on Multimedia and the Web (Webmedia)*, pages 133–135.

Nogueira, B. M., Moura, M. F., Conrado, M. S., Rossi, R. G., Marcacini, R. M., and Rezende, S. O. (2008). Winning some of the document preprocessing challenges in a text mining process. In *IV Workshop on Algorithms and Data Mining Applications, XXIV Brazilian Symposium on Database*, pages 10–18.

Preece, J. and Maloney-Krichmar, D. (2005). Online communities: Design, theory, and practice. *Journal of Computer-Mediated Communication*, 10(4):article 1.

Zhao, Y., Karypis, G., and Fayyad, U. (2005). Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10(2):141–168.