Online Knowledge Triage: Searching, Detecting, Labelling and Orienting User Generated Content

Project OCKTOPUS

Jean-Michel Dalle, UPMC / CRG, jean-michel.dalle@upmc.fr
Catherine Faron-Zucker, Univ. Nice Sophia Antipolis, faron@polytech.unice.fr
Fabien Gandon, Inria, fabien.gandon@inria.fr
Mathieu Lacage, Alcmeon, ml@alcmeon.com
Zide Meng, Inria, zide.meng@inria.fr

ABSTRACT

This position paper provides an overview of the OCKTOPUS project whose goal is to increase the social and economic benefit of user-generated content, by transforming it into knowledge which can be shared and reused broadly.

Categories and Subject Descriptors

[**Information Systems**]: Social networking sites

1. INTRODUCTION

Since the end of the 90's and along with the success of the "social" web, online communities have progressively and collaboratively produced massive amounts of user-generated content. While some of these communities are highly structured and produce high-quality content (e.g., Wikipedia), the quality of discussions found within less structured forums remains highly variable. Coupled with their explosive growth, the loosely structured nature of online forums makes it hard to retrieve relevant and valuable answers to user search queries, and subsequently diminishes the social and economic value of this content. In this context, we summarize results of the OCKTOPUS project (Online Content and Knowledge Triage: Optimizing the Productivity of User Search) that takes advantage of newer data mining techniques while properly assessing 1. the organizational traits of online communities, 2. the available structure of online discussions, and 3. the temporal dynamics of the content and structure, to improve the automatic classification and triage of unstructured online content. In section 2, we give an overview of how a demonstration platform was built, thus providing a visible and easily understandable output of the project that can be used to input a newly-formulated question, search online forums for a similar already-answered question, and display a unique user-generated answer associated with these similar questions. Then, sections 3 and 4 highlight selected results with respect to using automatic classification and to analyzing the activity and structure of Q&A communities, most notably in relation to topics and to the emergent organization of online communities.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1109/ASONAM.2014.6921608

2. DEMONSTRATION PLATFORM

The OCKTOPUS demonstration platform [1] is built around a custom crawler designed to perform deep focused incremental crawls of numerous Q&A and forum sites. The data is extracted from HTML pages, and then, it is normalized to large per-site in-memory databases where triage to mine for gold happens. Finally, gold is exported to a set of per-language SOLR indexes against which queries are executed to retrieve candidate answers for new incoming questions. The triage algorithm was initially built around an extensive dataset of human-annotated Q&A pairs fed to an SVM together with learning features based on intrinsic metadata such as Q&A votes, views, whether answers were accepted or not, user scores, etc. This first-generation algorithm was then extended to learn from user interactions with the system itself and include newer post and user quality metrics as described in the next sections.

3. STEERING PEER PRODUCTION

In the past few years, numerous researchers in both management and computer science have been specially interested in understanding how incentive mechanisms could be used in peer production communities in order to enhance the quantity and the quality of knowledge produced. Indeed, in most recent attempts to harness the so-called "wisdom of crowds", traditional online community approaches based mostly on online discussion networks and on a mix of heterogeneous motivations, the outcome of which results in more less balanced knowledge structures [2]) have been blended with other mechanisms that more intentionally and more strongly rely on intrinsic motivations, if not on financial incentives as in microwork labor markets [4]. The well-known Q&A web site StackOverflow (SO) is a particularly prominent example in this new direction as it seems to have implemented a particularly successful blend, even if dedicated mostly to developers. Moreover, it has been easily shown that incentive mechanisms such as "reputation scores" and "badges" actually could influence and possibly steer the behavior of users in SO. However, how these newer mechanisms could affect peer production per so, i.e. the fact that users

[1] https://alcmeon.com/ocktopus/demo.html

tend to dynamically edit content previously contributed by others, a phenomenon known to be key for many online knowledge producing communities such as open-source software and Wikipedia and others, has not been studied up to now, even in SO, whereas editing answers previously contributed by others is definitely an option offered to contributors, if not at all strongly incentivized compared to other tasks. Following [3] we study this editorial process oriented towards maintenance with a dataset of the 40K most voted questions and their 2 best answers (based on up-votes) in SO, and with a sub-dataset of 10K questions where best answer pairs were almost-synchronous. We match these questions and their associated answers with topics automatically inferred from question tags using results from section 4. We first characterize edits and observe that answer editing by other authors than the initial contributor mostly deals with *late* editorial maintenance, the frequency of which tends to have decreased over time in the global SO community. We further observe that this maintenance by others, within a given topic, actually decreases with the average reputation of the initial authors of answers within the topic, whereas this average reputation does not depend on the age of a topic as measured by the average date when answers associated with it were initially contributed. In addition, the later the maintenance with respect to the initial answer, the lower the reputation of the editor. Quite surprisingly, we also observe that maintenance by others actually increases for more recent topics, contrary to the general trend observed. Furthermore, authors of answers of more recent topics are also more recent members of the SO community. Therefore, rather than simple peripheral participation, these observations altogether suggest that newer topics could serve as attractors for the contributions of newcomers, both in terms of answers and of editorial maintenance. If confirmed, this phenomenon would certainly point towards some limitations with regard to the relevance of SO's (at least current) incentive policies and maybe also towards more general issues with respect to online peer production activities.

4. COMMUNITIES AND INTERESTS

Certain kinds of online communities such as Q&A sites, have no explicit social network structure. Therefore, many traditional community detection techniques do not apply directly. For instance in SO, a user submits a question, assigns 1∼5 tags to indicate the key points of this question. Other users who are interested in the question may provide answers to the question or comments to other answers. The only social graph in SO is this question-answer graph. Similarly to [5], we applied the Latent Dirichlet allocation(LDA) [1] to construct a users-topic-tags model to detect latent topics of interest from user acquired tags and then cluster users into different topics. The results were encouraging, however, we identified three problems: the complexity of the probabilistic model was prohibitive; the original LDA model is not enough to extract temporal and expertise information; the detected probabilities distributions cannot be compared with each other. To solve these problems, we first proposed TTD (Topic Trees Distribution) a simple method to detect topics and overlapping communities. From the observation of StackOverflow dataset, we confirmed the natural intuition that high frequency tags are more generic and low frequency tags are more specific, and most of the low frequency tags are related to a more generic tag. This inspired an algorithm to build a tag tree and perform a spectral clustering on the affinity matrix to group root nodes in topics and compute the probability for a tag to be related to a topic. This in turn provides an efficient approach for extracting topic from Q&A to detect communities of interest. We compared three detection methods we applied on a dataset extracted from SO. Our method based on topic modeling and user membership assignment was shown to be much simpler and faster while preserving the quality of the detection [6][7][8]. In current work and in order to extract even more information from user-generated content, we extend the previous model with Temporal Topic Expertise Activity (TTEA) to jointly model topics, trends, user expertise, and user activities.

5. CONCLUSION

OCKTOPUS explored several new directions to analyze and reuse content from Q&A web sites. One final objective of the project is to study if and how the different approaches proposed can be integrated together in social Web applications to improve the management of the communities.

Acknowledgment to ANR-12-CORD-0026 Ocktopus grant

6. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[2] J.-M. Dalle and P. A. David. The allocation of software development resources in open source. In J. Feller, B. Fitzgerald, S. A. Hissam, and K. R. Lakhani, editors, *Perspectives on Free and Open Source Software*, pages 297–326. The MIT Press, 2005.

[3] J.-M. Dalle, M. Devillers, and M. den Besten. Answer editing in stackoverflow. In *Proceedings of the 2nd International Conference on Knowledge Commons*. NYU Engelberg, 2014.

[4] J.-M. Dalle, T. Lacroix, M. Lacage, and M. den Besten. A direct empirical study of the determinants of online work supply in amazon mechanical turk. In *Proceedings of the 3rd Internet, Policy and Politics Conference*, IPP 2014. University of Oxford, 2014.

[5] D. Li, B. He, Y. Ding, J. Tang, C. Sugimoto, Z. Qin, E. Yan, J. Li, and T. Dong. Community-based topic modeling for social tagging. In *Proc. ACM CIKM*, CIKM '10, pages 1565–1568, New York, 2010. ACM.

[6] Z. Meng, F. Gandon, and C. Faron-Zucker. Simplified labeling of overlapping communities of interest in question-and-answer sites. In *The 2015 IEEE/WIC/ACM International Conference on Web Intelligence*, Singapore, Singapore, Dec. 2015.

[7] Z. Meng, F. L. Gandon, C. Faron-Zucker, and G. Song. Empirical study on overlapping community detection in question and answer sites. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM, Beijing, China, August 17-20, 2014*, pages 344–348, 2014.

[8] Z. Meng, F. L. Gandon, C. Faron-Zucker, and G. Song. Detecting topics and overlapping communities in question and answer sites. *Social Netw. Analys. Mining*, 5(1):27:1–27:17, 2015.