Wanted: A Unified Model for Search in Social Media

Liangjie Hong and Brian D. Davison
Dept. of Computer Science and Engineering, Lehigh University
Bethlehem, PA 18015 USA
{lih307,davison}@cse.lehigh.edu

1. INTRODUCTION

Social Media, including wikis, micro-blogging, forums and social network services, plays an indispensable role in providing a platform to support user-generated and user-disseminated information through social interactions. Social Media provides participants faster and wider information spread and burst compared to traditional media channels, which has attracted significant attention from the research community. In addition, due to its diversity of representation and operation mechanisms, different kinds of tasks are usually supported by different Social Media services or sometimes combinations of them. For example, blogs are usually maintained by individuals while forums are maintained by "moderators" and all users may participate in discussions by writing posts and replies. The differences in design and structure permit them to serve different purposes. Therefore, people use blogs to express personal feelings, opinions and comments but use forums to discuss problems and receive suggestions, though there is no strict line between these two types of Social Media. Due to this diversity, research in Social Media in recent years mainly focuses in two directions: how to effectively support specific tasks that one or several particular Social Media services provide as the example mentioned above, and how to improve the effectiveness of search engines that can take advantage of rich information that Social Media generates. In both directions, research work is usually conducted on a small number of types of Social Media services. In fact, most of them focus on only one type of Social Media. For instance, work has been done to help Community Question Answering portals (CQA) like Yahoo! Answers to retrieve similar question-answer pairs more effectively (e.g., [4]). Researchers also pay attention to social bookmarking systems and try to use them to help users organize information. Blogs are used to track opinions and sentiments while micro-blogging tools like Twitter are used to identify emerging trends and hot topics. Although certain success has been shown in these distinct types of Social Media, little research work has been done to discuss whether or not we need a *unified* way to deal with all types of Social Media, which could then support not-yet-imagined tasks rather than developing separate models for each Social Media service.

Here, a unified model should provide a "baseline" model which is able to capture the basic characteristics among different types of Social Media but not necessarily solve all possible problems. In other words, such a model does not need to be a solution to all tasks but rather a foundation for further development. The benefit of pursuing a unified model for Social Media is twofold. First, since new types of Social Media are emerging every year (e.g., blogs a decade ago, Facebook from 2004, Flickr from 2004, Yahoo! Answers from 2005, Twitter from 2006, etc.) and as their purposes as discussed above may or may not be the same, current research results and conclusions might be not applicable to new services. In fact, little research has been done to discuss the effectiveness of whether current models can or cannot be employed in similar Social Media services. Taking Yahoo! Answers as an example, a number of researchers have focused on how to retrieve similar questions and their potential corresponding answers from the large repository of data. Since forums and discussion boards are also popular platforms for problem-solving, a natural question is whether the models and methods developed for Yahoo! Answers can be applied to discussion boards. As far as we know, no direct answer is available from the research literature. Indeed, as mentioned above, research work focusing on Yahoo! Answers usually took advantage of special features or characteristics that CQAs provide (e.g., [2, 5]), which make them difficult to apply to forums. For example, CQAs usually allow users to pick one answer from a set of posted replies as a best answer to that question, which are usually used as training labels for models and thus many of those models are based on supervised or semi-supervised learning. In addition, users may use "thumbs up" or "thumbs down" to indicate their preferences to the answers. Both of these features are usually not available for forums and discussion boards. Therefore, compared to Question Answering Portals, there are much less successful attempts to retrieve high quality question answer pairs on forums and indeed many approaches developed for CQAs do not work at all for discussion boards. In addition to CQAs, we believe that similar arguments are applicable to other Social Media services.

The second benefit for a potential unified model for Social Media is from the perspective of search engines. Although major commercial search engines are merging a variety of "verticals" all the time, the pace and the quality of introducing new "verticals" are still far from satisfaction. For instance, search engines recently announced their support for micro-blogging services like Twitter (e.g., [1]), which happened three years after the introduction of Twitter and its most recent two years of extreme growth, implying that search engines are suffering difficulty in incorporating the latest information from Social Media. Indeed, search results from Social Media services like forums, social bookmarking, CQAs are still far from satisfying. A partial reason for this phenomenon is the current services-based research orientation and the differences of purposes

and structures of those Social Media services as mentioned before. If we have a unified way to treat all types of Social Media regardless of their original representation and structure, search engines can better utilize the rich content from Social Media and do a much better job than at present.

In this position paper, we would like to initiate discussion about a unified model for search in Social Media and talk about some possibilities that might be worth considering. Instead of proposing a specific unified model, here, we discuss some properties that a unified model might have for Social Media.

2. ROAD TO A UNIFIED MODEL

In this section, we would like to discuss the properties and characteristics a unified model for Social Media might include, and how existing research efforts can be employed.

Before heading to the discussion on individual properties, let us clarify what we mean by a "unified model". Note, here, we are considering the problem of "search in Social Media" and therefore unified model is to retrieve useful information (including entities and their relationships) from Social Media. In particular, we are imagining a *baseline* model that can be extended for a variety of tasks in Social Media, which is independent of specific datasets. The model is essentially a logical layer between actual representations and structures of Social Media (e.g., web pages) and how they can be manipulated by algorithms. In other words, the model is an abstraction of Social Media. The model can act in a similar role as the Relational Model in databases and Protocol Stacks in networking. In traditional Information Retrieval (IR), Vector Space Models and Language Models are arguably such kinds of unified model to deal with all types of retrieval entities, which are usually documents. Two problems exist for current IR models that make them difficult to be applied to Social Media. First, both of them heavily depend on the notion of "documents", which is relatively vague on Social Media. A unified model needs to provide a guidance for the choice of "document" for different types of Social Media. Traditional models provide no guidance for this question. Additionally, traditional models do not explicitly consider the "social" aspect of information. Indeed, neither of these two popular models take the owners of the "documents" and their relationships and interactions into account. Therefore, a unified model for search in Social Media should go beyond traditional information retrieval and explicitly model all important aspects of Social Media.

It seems that a unified model for Social Media could be a simple modification of traditional models. However, we believe that it should be much more than just extending the notion of "document" and incorporating social relationships. In the following discussion, we list several properties that we think important for a unified model.

First, the representation of a unified model would naturally support different paradigms of search. Two basic problems should be considered: the choice of "entities", and their relationships. Traditional IR models treat "documents" as the (only) first class entity in the model. However, in Social Media, content, although still important, is no longer the only entity we care about. How about finding most relevant people? How about finding most frequent relationships (e.g., friends, colleagues, partners) in a group of users? A unified model should consider multiple information retrieval paradigms in the first place. This would also imply that the model possibly needs to treat all entities (e.g., users) equally other than only considering documents as first class entities. Additionally, multiple levels of "documents" are needed to be supported. Taking a web profile page on Facebook as an example—the whole web page can be regarded as a document while a "wall conversation" on that page and a live feed stream can be considered as lower level documents as well. Therefore, a hierarchy of "documents" is needed to be considered in the model. Thus, the relationships between entities are the second basic problem for representation. A unified model would naturally support multiple relationships and the trust of these relationships as well. The links and how they interact with other entities should not be a separate model and at least can be easily embedded into retrieval process.

Second, a unified model would naturally consider "uncertainty" in properties of entities and entity relationships, which are inherent from real-world applications. The model needs to have the ability to "infer" these missing values. For example, many applications are related to the "uncertainty" of certain properties. Whether two accounts on Twitter are operated by the same user? Whether the owner of a Youtube account usually posts spam content? These kinds of questions should be answered more easily through a unified model.

Third, the effectiveness of a unified model can be verified on different types of Social Media. The model can incorporate additional features from particular services but not depend on them (as in the example of CQAs). Some existing research directions are already taking some of them into account (e.g., topic model for document networks [3, 6], which explicitly model content and links between content). We are calling the attention for these properties so that Social Media research may pay more attention on broad models rather than dataset-specific methods.

3. CONCLUSION

While it is certainly debatable whether we need a unified model or not, a unified model would benefit both ordinary users and search engines in that information needs can be better satisfied within the context of Social Media. In this position paper, we argued for such kind of models and discussed possible properties or problems the model should take into account. Though we do not know whether a unified model would necessarily improve the effectiveness of retrieval in terms of performance (ideally, it should!), the pursuit of such a model will open more opportunities to better understand the dynamics of Social Media.

4. REFERENCES

[1] Relevance meets the real-time web, Dec. 2009. http://googleblog.blogspot.com/2009/12/relevance-meets-real-time-web.html.

[2] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the 1st International ACM Conference on Web Search and Data Mining (WSDM)*, pages 183–194, 2008.

[3] J. Chang and D. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 2009.

[4] J. Jeon, W. B. Croft, and J. H. Lee. Finding similar questions in large question and answer archives. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 84–90, 2005.

[5] Y. Liu, J. Bian, and E. Agichtein. Predicting information seeker satisfaction in community question answering. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 483–490, 2008.

[6] R. Nallapati and W. Cohen. Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs. In *Proceedings of the International Conference on Weblogs and Social Media*, 2008.