Inf Retrieval (2012) 15:179–182
DOI 10.1007/s10791-012-9199-7

Guest editorial: Special issue on information retrieval for social media

Fei Wang · Peng Cui · Gordon Sun · Tat-Seng Chua · Shiqiang Yang

Received: 24 March 2012 / Accepted: 24 March 2012 / Published online: 8 April 2012

1 Information retrieval for social media

As is known to us all, *media* is an instrument on communication, such as newspaper, radio or video. Thus *social media* would be a social instrument on communication. Kaplan and Haenlein (2010) define social media as “a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content.” The key difference between social and conventional media is the user relationships and activities. This clearly poses new challenges to traditional information retrieval technologies including:

- *Evaluation of item similarities*. This is the key for information retrieval. In a social environment, the similarities among items not only involve their contents, but also the users who participate in the life cycle of those items (e.g., post, share, follow, etc.). Moreover, because of the user activities, usually more than one type of items (e.g., image, music, text and video) are involved in a social network, and the information contained in each single item domain is not sufficient (e.g., text in microblogging

F. Wang (✉)
Healthcare Analytics Research Group, IBM T. J. Watson Research Center, Hawthorne,
NY 10532, USA
e-mail: feiwang03@gmail.com

P. Cui · S. Yang
Department of Computer Science, Tsinghua University, Beijing, China
e-mail: cuip@cse.ust.hk

S. Yang
e-mail: yangshq@cse.ust.hk

G. Sun
Tencent Technology, Shenzhen, China
e-mail: gordonsun@tencent.com

T.-S. Chua
Department of Computer Science, National University of Singapore, Singapore, Singapore
e-mail: chuats@comp.nus.edu.sg

system such as Twitter). In this case, how to effectively evaluate item similarities by leveraging both information from user and other related item domains is critical for social media retrieval and recommendation.

- *Data scalability*. In social environment, we not only care about the scale of the items, but also the scale of the users. For example, till the year 2010, Twitter has more than 200 million users accounts, and more than 600 million search queries a day. Facebook has more than 400 million active users, and more than 25 billion pieces of content (web links, news stories, blog posts, notes, photo albums, etc.) shared each month. For the retrieval of social media, we need to consider both the scalability of users and items (and usually there are multiple types of item involved). Thus the scalability challenge is more severe in social media retrieval.
- *Time sensitivity*. Because of the participation of the users, information (e.g., breaking news) will spread much more rapidly in the social environment. This makes the *timing* factor more critical when retrieving social media compared to conventional information retrieval. Therefore analyzing the user network structure and making effective timely recommendation is also a challenge for social media retrieval.

We need to design new information retrieval technologies for social media that can handling the above challenges, and this is also the reason why information retrieval for social media emerges as a new hot research direction.

2 The special issue

Recently, we started some activities to establish the platform for exchanging on data analysis for social media. We have successfully organized the workshop on social and behavioral networked media access (SBNMA) (Ramzan et al. 2011) held in conjunction with ACM Multimedia Conference 2011. This special issue provides a leading focused forum for timely, in-depth presentation of recent advances in algorithms, theory and applications on information retrieval for social media. The selected papers underwent a rigorous extra refereeing and revision process.

As we stated in the introduction, similarity evaluation is still the key to information retrieval for social media. The paper by Markus Schedl presents a systematic and comprehensive evaluation of different term weighting measures, normalization techniques, query schemes, index term sets, and similarity functions for the task of inferring similarities between named entities, based on data extracted from microblog posts. The author analyzes different combinations of choices for those dimensions influencing the similarity calculation process, and investigates in which way they impact the quality of the similarity estimates.

Another important factor for information retrieval is *query*. Effective query manipulation will also affect the performance of social media retrieval. The paper by Dong Zhou, Séamus Lawless and Vincent Wade presents a novel personalized query expansion method for retrieving social media. The query expansion is based on individual user profiles mined from the annotations and resources the user has marked. The underlying theory is to regularize the smoothness of word associations over a connected graph using a regularizer function on terms extracted from top-ranked documents. The experimental evaluations suggest that the proposed approach significantly benefits personalized web search by leveraging users social media data.

The paper by Wouter Weerkamp and Maarten de Rijke proposes to incorporate the *credibility* of information into the blog post retrieval process. Specifically, the credibility of information refers to its believability or the believability of its sources. Following the intuition that more credible blog posts are preferred by searchers, the authors explore the impact of credibility-inspired indicators on the task of blog post retrieval, and utilize the ideas from the credibility framework in a reranking approach. Experimental evaluation results show that credibility-inspired reranking leads to larger improvements over the baseline than combined reranking, but both approaches are capable of improving over an already strong baseline.

Topic model is an important tool for information retrieval. In social environment, it would be interesting to adapt conventional data-driven topic modeling to incorporate user information. The paper by Jingdong Wang, Zhe Zhao, Jiazhen Zhou, Hao Wang, Bin Cui and Guojun Qi proposes a social topic model to automatically recommend Flickr groups by simultaneously exploit media contents and link structures between users and groups. Their model can jointly discover the latent interests for users and Flickr groups and simultaneously learn the recommendation function.

Automatic question answering is a hot research topic in information retrieval which has aroused considerable interests in recent years. Under the *social* umbrella, there emerges some specific interesting problems. The paper by Ian Wakeman, Simon Fleming and Dan Chalmers considers how to ask questions and retrieve answers using the wisdom of the crowd when people are connected together over ad hoc social networks. The authors focus on fully decentralised protocols using ant inspired tactics to route questions towards members of the network who may be able to answer them well. At the same time, both question asking and answering are plausibly deniable with privacy concerns. They show that it is possible to improve answer quality and also reduce the total amount of user attention required to generate those answers.

The paper by Tianyong Hao and Eugene Agichtein considers the problem of finding similar questions in collaborative question answering archives. They propose a precise approach of automatically finding an answer to the questions by automatically identifying “equivalent” questions submitted and answered in the past. Their method generates equivalent question patterns by grouping together questions that have previously obtained the same answers, and these patterns are used as seed patterns to match more questions to extract large number of equivalent patterns by a bootstrapping-based learning method. The learned patterns can be applied to match a new question to an equivalent one that has already been answered, and thus suggest potential answers.

The paper by Matteo Magnani, Danilo Montesi and Luca Rossi introduces a novel search paradigm for microblogging sites resulting from the intersection of Information Retrieval and Social Network Analysis (SNA). This approach is based on a formal model of on-line conversations and a set of ranking measures including SNA centrality metrics, time-related conversational metrics and other specific features of current microblogging sites. The authors compare their proposed approach with other methods on two well known social network sites (Twitter and Friendfeed) showing that the inclusion of SNA metrics in the ranking function and the usage of a model of conversation can improve the results of search tasks.

Sentiment detection is another research topic that is closely related to computational linguistics and information retrieval. It aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. In the Web 2.0 era, due to the expanding volume of available online information such as microblogging messages and review comments, automatic sentiment detection becomes more important to the success of modern e-commerce. In the paper by Dan Zhang, Luo Si and Vernon J Rego, the authors propose a novel method for joint modeling the sentiment space of different domains. Their proposed method is validated on the product review and `twitter` datasets.

The last paper of this special issue, which is by Hanghang Tong, Spiros Papadimitriou, Christos Faloutsos, Philip Yu and Tina Eliassi-Rad, considers the problem of finding *gateways* in a large graph, i.e., find a small subset of nodes that are crucial in connecting the source to the target. For example, given a social network, who is the best person to introduce you to, say, Chris Ferguson, the poker champion? The authors formulate the problem in both Pair-Gateway and Group-Gateway scenarios, and show the problem is sub-modular and can be solved near-optimally. Moreover, the authors also make those algorithms fast and scalable, and validate them on real world datasets.

Acknowledgments We are very indebted to the reviewers who reviewed the papers very carefully. We would also like to thank all the authors who submitted their papers to the special issue. Special thanks to Prof. Charles Clarke for his great help and support in organizing the issue.

References

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, *53*(1), 59–68.

Ramzan, N., Wang, F., Patrikakis, C. Z., Cui, P., Doulamis, N. D., & Yang, S., et al. (2011). Acm international workshop on social and behavioral networked media access (sbnma'11). In *Proceedings of ACM multimedia*, p. 611–612.