Scientific Report for ELIAS Short Visit Grant

Anne Schuth, anne.schuth@uva.nl
ISLA, University of Amsterdam, The Netherlands

May 14, 2012

I hereby report on my recent visit to the Yahoo! Research Labs in Barcelona that was funded by the ELIAS Network Programme.

Purpose of visit Increasingly, people convey their interests through their online presence. Users leave digital traces in the form of updates, comments, images, geo-locations, likes, +1's and in a multitude of other ways. Interests are also communicated more implicitly by subscribing to updates—tweets, posts, statuses—of *other* users on various platforms. By subscribing to other users, a user reveals interest in what that other user is saying or doing. I am in the early stages of introducing a framework for combining all these expressions of interest.

Specifically, I intend to build on recent work of De Francisci Morales et al. [2012], who brought two sources of information together to recommend news articles: news streams and micro blogs.

The online aspect—the living lab setting—is crucial since things are never static on the web: new friendships are formed while others die out, the opinion of certain friends on a topic might become more important than it used to be. Also, completely new sources of evidence might appear and then should be incorporated as soon as possible without the need of re-learning the whole model. We mimic the online nature by replaying the user history—obtained from a browser toolbar—step by step. Research in this direction—involving a living lab—can only be performed there where access to this commercially valuable and sensitive user data is available.

The Yahoo! Barcelona Research Lab was—and is—the ideal location to visit for my evaluation plans since this is where the data required for the experiments resides. Moreover, this is where De Francisci Morales et al. [2012], the authors who wrote the work on which I build, perform their research. The purpose of the visit was sharing the evaluation framework, knowledge and exploiting the toolbar data present.

Description of work The work I carried out during the two weeks of visit consisted of extending the data sets of tweets, news and toolbar data mentioned in De Francisci Morales et al. [2012] from a single month to a full year. Furthermore, the work of Meij et al. [2012] has been consolidated in the Hadoop framework used at Yahoo!.

On the theoretical side, numerous fruitful discussions took place on how to view news recommendation in the light of Reinforcement Learning, following for example Hofmann et al. [2013].

Description of results Concrete results of the visit are:

- the extended and processed datasets consisting of tweets and news with semantic annotations and clicks extracted from the toolbar data;
- a tool that can provide these semantic annotations on a large scale;
- knowledge of, and hands-on experience, with the Hadoop framework on a large cluster;
- insights in how a news recommendation system can be fit into a reinforcement learning setting.

Future collaboration and projected publications We plan on consolidating the work carried out during my visit. We are targeting the Sixth ACM International Conference on Web Search and Data Mining—WSDM '13—with a full paper for which the applicant and the host institution intend to collaborate.

References

G. De Francisci Morales, A. Gionis, and C. Lucchese. From chatter to headlines: harnessing the real-time web for personalized news recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 153–162. ACM, 2012.

K. Hofmann, S. Whiteson, and M. de Rijke. Balancing exploration and exploitation in listwise and pairwise online learning to rank for information retrieval. *Information Retrieval Journal*, 2013.

E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, New York, NY, USA, 2012. ACM.