New Approaches to Log Mining and Applications to Collaborative Filtering

Ján SUCHAL*

Slovak University of Technology
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
suchal@fiit.stuba.sk

In era of information overload we do seek help in recommendation systems aiding us to focus our attention to items like products, articles [1] or websites, that might be relevant to our needs based on prior knowledge. Typically this is done by mining knowledge about our interests from logs of our previous activity.

Our work focuses on two main goals strongly related to recommendation engines. Novel approaches to log mining and the potential usage of these implicit data in recommendation systems, especially collaborative filtering exploiting implicit negative feedback data. In general, while interpreting implicit feedback from logs can be a challenging task [2], mining and interpreting negative implicit feedback from positive visit logs is much more challenging.

First, we present a method for mining sources and cascading graphs of viral visits from raw logs. Such information can be useful for evaluation of marketing targeting to detect influencers and potential sources of viral traffic. We detect users that visit pages via viral recommendations (instant messaging, email...) and look for users that visited such pages before and estimate probability of viral recommendation source. A probabilistic viral cascade graph can be reconstructed from such data. We also categorize types of sites for which our method can be used and experiment on real world dataset containing the massively viral start of foaf.sk service.

Second approach focuses on mining negative interests of users from basic server logs in the domain of news articles. We propose two different methods for mining negative feedback, the first is based on time-based identification of articles which users do not read, second is based on detecting such articles, whose title (or even an abstract) was probably seen by the user, but did not raise enough of interest to actually visit and read it. Such data can be used in addition to positive interest that are normally used for generating recommendations. Experiments on live traffic on largest Slovak news portal www.sme.sk show that incorporating such feedback into collaborative filtering

* Supervisor: Pavol Návrat, Institute of Informatics and Software Engineering

Spring 2011 PeWe Workshop, April 1, 2011, pp. pp. 33–34.

recommender gains 8.5% higher click-through rates and lowers recommendation rejection rate by 5% when compared to baseline collaborative filtering algorithm [3].

Finally we present a novel method for linearly scalable nearest-neighborhood based collaborative recommender system using specially prepared fulltext indices [3]. Evaluation is done on datasets from largest Slovak news portal sme.sk and github.com recommendation contest. Comparison with graph-based spreading activation recommendation method shows comparable results in means of relevance (20% precision on top 10 list) and with superior scalability characteristics.

Future work focuses on combining content-based and collaborative recommendation approaches into a superior hybrid approach exploiting positive and negative implicit feedback.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

[1] Das, A.S., Datar, M., Garg, A., Rajaram, S.: Google news personalization: scalable online collaborative filtering. In *Proc. of the 16th Int. Conf. on World Wide Web*. New York, NY, USA, ACM, pp. 271–280, 2007.

[2] Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In *Proc. of the 28th Int. ACM SIGIR Conf. on research and development in information retrieval*. New York, NY, USA, ACM, pp. 154–161, 2005.

[3] Suchal, J., Navrat, P.: Full Text Search Engine as Scalable k-Nearest Neighbor Recommendation System. In Bramer, M., ed.: Artificial Intelligence in Theory and Practice III. Volume 331 of IFIP Advances in Information and Communication Technology. Springer Boston, pp. 165-173, 2010.