Data Analysis for Collaborative Filtering Systems

Toon De Pessemier

Supervisor(s): Luc Martens

I. Introduction

The overabundance of information and the related difficulty to discover interesting content has complicated the selection process for the end-user who gets overloaded with data and risks to get lost. Recommender systems try to assist users in this content-selection process by using intelligent personalisation techniques which filter the information. Most commonly-used recommendation algorithms are based on Collaborative Filtering (CF) techniques which generally provide better results than Content-Based (CB) techniques and require no meta-data of the content [1]. This research investigates the amount of data required for these CF algorithms.

II. Computational Load

The efficiency of personal suggestions generated by CF techniques is highly dependent on the quality and quantity of the available consumption data. Extending data sets with additional consumption data might enrich the user profiles and generally leads to a higher accuracy. Although, if a considerable amount of profile information is already available and detailed personal preferences can be derived, supplementary consumption data may not have any (or a very limited) added value for the accuracy of the recommendations. Therefore, we investigated the minimum amount of data that is required to generate recommendations with a sufficient accuracy.

Toon De Pessemier is a PhD student funded by the PhD fellowship of the Research Foundation Flanders at the WiCa research group. toon.depessemier@intec.ugent.be

III. Method

In this research, we calculated recommendations based on different training sets containing various amounts of consumption history. This way, the minimal amount of consumption data required to acquire recommendations with an acceptable accuracy loss is investigated. Moreover, the optimal size of the neighbourhood containing similar users used for calculating the recommendations is examined. Besides, the thresholds of the neighbourhood selection criteria (e.g. user similarity or profile overlap) are studied. The accuracy of the generated recommendations is compared for the various configurations based on evaluation metrics which are generated by an offline analysis on a test set of consumptions [2].

IV. Conclusions

The results of the data analysis showed that additional consumption data can provide a positive contribution to the accuracy of the recommendations but requires a considerable computation cost. Moreover, reducing the neighbourhood size can reduce the calculation load considerable without a significant accuracy loss. Online content recommenders should weight this calculation costs of additional data / neighbours against the slightly accuracy improvement.

References

[1] Linden, G., Smith, B., York, J. *Amazon.com recommendations: item-to-item collaborative filtering.* Internet Computing, IEEE , vol.7, no.1, pp. 76- 80, Jan/Feb 2003

[2] Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. *Evaluating collaborative filtering recommender systems.* ACM Trans. Inf. Syst., vol.22, no.1, pp. 5-53, Jan 2004