Research Proposal – MSc in Econometrics Thesis

Recommender system for accomodation-booking website

by

Radim Kašpárek 11669799

Date**: *7.4.2018***

Introduction

Recommender systems exploit ratings provided by users to rank the relevance of items for specific users. However, in many cases it is not feasible to collect explicit feedback from users as this presents a cognitive and time load on users. In such cases, the only information available to recommender systems are the actions of the user, e.g. number of views, time spent on the detail of an item, and save or buy actions of the users. Identifying relevant items to recommend is usually more complicated than in systems which have explicit feedback available, however, the implicit feedback-based recommender systems have significantly more use cases.

Such systems have been shown to be vulnerable to so-called *shilling attacks* in which malicious users with carefully chosen profiles specifically query the system in order to push the predictions of some targeted items. Moreover, these systems show decrease in performance when non-relevant users are included in the algorithm.

In this thesis I would like to construct an algorithm which ranks the item relevance for specific users and test whether pruning out non-standard/malicious sessions improves the accuracy of the algorithm.

Research questions

Is it possible to enhance the performance of implicit feedback-based recommender systems by identifying and pruning out the non-standard/malicious sessions?

Methodology and Techniques

For the baseline recommendation system, collaborative filtering and content-based recommendations will be used. I will test matrix factorization techniques as well as Gradient Boosted Trees.

For the identification of non-standard/malicious searches I will use simple descriptive analysis and decide upon more sophisticated measures later on.

Finally, it will be tested whether the quality of the recommendations has been improved by the *shilling attacks* analysis.

The algorithm performance will be assessed by the Normalized Discounted Cumulative Gain. DCG measures the usefulness of a document based on its position in the result list. The gain of each result is discounted at lower ranks.
The DCG accumulated at a particular rank position *p* is defined as:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

Normalized version of DCG will be used as not all the search queries return the same number of matches.

Data and/or expected results

I will work with a dataset downloaded from Kaggle.com. This dataset was used 5 years ago in a competition called *Personalize Expedia Hotel Searches - ICDM 2013*, it contains 54 columns describing approximately 665 000 searches, the hotels which were shown for each query, the purchase history of the user, the offers of the competition and lastly, the click/book actions in the specific instance. The data are not complete, e.g. purchase history of the user or the competition offers are often unavailable.

I expect the performance of the implemented recommender system to improve upon after identifying the malicious/non-standard searches in the dataset.

Literature

Oard, Douglas W., and Jinmook Kim. "Implicit feedback for recommender systems." *Proceedings of the AAAI workshop on recommender systems*. Vol. 83. WoUongong, 1998.

Lee, Tong Queue, Young Park, and Yong-Tae Park. "A time-based approach to effective recommender systems using implicit feedback." *Expert systems with applications* 34.4 (2008): 3055-3062.

Koren, Yehuda, Robert Bell, and Chris Volinsky. "Matrix factorization techniques for recommender systems." *Computer*42.8 (2009).

Hu, Yifan, Yehuda Koren, and Chris Volinsky. "Collaborative filtering for implicit feedback datasets." *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. Ieee, 2008.

Timetable

1st Month

Activity	Date
Literature review, assembling the code	1. 22.4. – literature review 2. 7.5. – code backbone ready

2nd Month

Activity	Date
Tuning the parameters, writing down the methods section	3. 21.5. – parameters tuned 4. 28.5. – methods section written

3rd Month

Activity	Date
Writing results, conclusion and introduction, polishing the language and form	5. 12.6. – text written 6. 19.6. – polished, ready to hand in

Make sure you are enrolled in Blackboard > MSc in Econometrics: Info, Thesis and Presentations. If this is not the case contact the program director.

I am enrolled in the course:

Master's Thesis Econometrics	☐
Master's Thesis Financial Econometrics	☐
Master's Thesis Mathematical Economics	☐
Master's Thesis Big Data Analytics	X
Master's Thesis Free Track	☐

Supervisor **Signature**

Second marker (suggestion by supervisor)

Thesis coordinator **Signature**

J.C.M. van Ophem