Non-transparent recommender system evaluation leads to misleading results

Alan Said[1], Alejandro Bellogín[2]
[1]Recorded Future, Gothenburg, Sweden, alansaid@acm.org
[2]Universidad Autónoma de Madrid, Madrid, Spain, alejandro.bellogin@uam.es

ABSTRACT

Recommender systems have become a ubiquitous feature on the World Wide Web. Today, most websites use some form of recommendation to heighten their users' experience. They do so by analyzing what items users have interacted with previously and analyze the interaction patterns generated.

In order to identify whether one recommendation approach performs better than another, various evaluation techniques are employed [1, 3]. These techniques usually focus on how accurate, or close to the user's preferences, the recommendations are; in simple terms: the better the evaluation scores, the better the recommender. However, when comparing the results of one recommender system to another, it is difficult to compare results verbatim due to the many options in design and implementation of an evaluation strategy. Additionally, implementational differences in the underlying recommendation framework can increase the comparison difficulty.

In our work [2], we investigate the discrepancies between common open source recommender system frameworks and highlight the difference in evaluation protocols – even when the same evaluation metrics are employed, evidencing differences in their implementation. In order to create a fair comparison, we additionally generate the same metrics using a transparent evaluation protocol and propose a controlled evaluation protocol where these aspects are clearly specified and tested. Our work shows large differences in recommendation accuracy across frameworks, and even larger differences when comparing the results from the controlled evaluation. The implication is that comparing the results from several recommender systems is often infeasible, unless a strict and transparent protocol is followed.

BODY

Evaluation of recommender systems is often misleading, transparent protocols should be used and reported by industry and academia.

REFERENCES

[1] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, 2004.

[2] A. Said and A. Bellogín. Comparative recommender system evaluation: benchmarking recommendation frameworks. In A. Kobsa, M. X. Zhou, M. Ester, and Y. Koren, editors, *Eighth ACM Conference on Recommender Systems, RecSys '14, Foster City, Silicon Valley, CA, USA - October 06 - 10, 2014*, pages 129–136. ACM, 2014.

[3] G. Shani and A. Gunawardana. Evaluating recommendation systems. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 257–297. Springer, 2011.

Volume 3 of Tiny Transactions on Computer Science