Information Retrieval and User-Centric Recommender System Evaluation

CWI Centrum Wiskunde & Informatica

Alan Said, Alejandro Bellogín, Arjen De Vries, Benjamin Kille
{alan.said, alejandro.bellogin, arjen.de.vries}@cwi.nl, kille@dai-lab.de

UMAP 2013 – Rome, Italy

ERCIM research project at CWI

Focused on novel aspects of recommender systems evaluation from an information retrieval and a user-centric perspective.

Duration and personnel
- January 2013 – March 2014
- 24 person months
- 2 postdocs

Related Events

Workshop on *Reproducibility and Replication in Recommender Systems Evaluation*. In conjunction with RecSys 2013

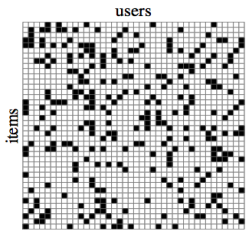
Workshop on *Benchmarking Adaptive Retrieval and Recommender Systems*. In conjunction with SIGIR 2013

ACM TIST Special Issue on *Recommender System Benchmarking*

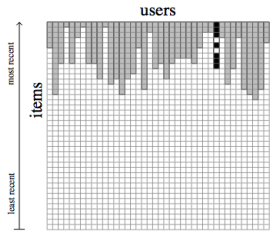
User-centric Evaluation

A **personalized offline evaluation protocol** mimicking real-life deployed recommender systems.
- Each algorithm is trained **once per user** based on a personalized training and test split.
- Candidate test items are selected based on a **personal rating value threshold** (e.g. mean).

Diversity-oriented evaluation methods focusing on:
- non-accuracy-based evaluation metrics
- item feature-centric evaluation
- novelty

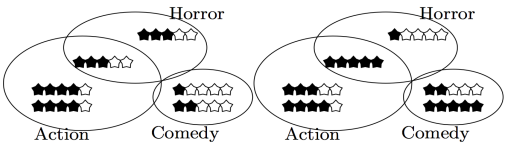


A random 80% – 20% training and test data split for all users. Items do not have to be sorted. One split is performed for the complete set of users.



A proposed split for one user when considering time. Gray items correspond to interactions by other users in the same timespan the candidate user has interacted with the black items.

IR-based Evaluation



(a) Coherent user
$c(u) = -1.28$
$c_w(u) = -0.52$

(b) Not coherent user
$c(u) = -5.95$
$c_w(u) = -2.15$

Example of a coherent vs. not coherent user. Our definition of coherence takes into account the rating's deviation within each item feature, which in this example consists of three genres: action, comedy, and horror.

A **coherence-based function** that is able to predict the user performance
- Correlated with the "magic barrier"
- Improves the total performance of the system when used to group users into difficult and easy ones

Explore **correlations between metrics** typically used in offline experiments (precision) and in real applications (e.g., CTR).

Perform an analysis of how evaluation in recommendation should be performed to obtain **interpretable and comparable results**.

Further Reading

Artist popularity: do web and social music services agree?
A. Bellogín et al. In ICWSM 2013

User-Centric Evaluation of a K-Furthest Neighbor Collaborative Filtering Recommender Algorithm
A. Said et al. in CSCW 2013

Poster Abstract
Information Retrieval and User-Centric Recommender System Evaluation
A. Said, A. Bellogín et al. in UMAP 2013

Acknowledgement

This work is carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no.246016.

ERCIM