Second Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2011)

Iván Cantador
Escuela Politécnica Superior
Universidad Autónoma de Madrid, Spain
ivan.cantador@uam.es

Peter Brusilovsky
School of Information Sciences
University of Pittsburgh, USA
peterb@pitt.edu

Tsvi Kuflik
Information Systems Department
University of Haifa, Israel
tsvikak@is.haifa.ac.il

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering, retrieval models*

General Terms
Algorithms, Human Factors, Experimentation, Performance.

Keywords
Recommender systems, information heterogeneity, information integration.

1. MOTIVATION AND GOALS
In recent years, increasing attention has been given to finding ways for combining, integrating and mediating heterogeneous sources of information for the purpose of providing better personalized services in many information seeking and e-commerce applications. Information heterogeneity can indeed be identified in any of the pillars of a recommender system: the modeling of user preferences, the description of resource contents, the modeling and exploitation of the context in which recommendations are made, and the characteristics of the suggested resource lists.

Almost all current recommender systems are designed for specific domains and applications, and thus usually attempt to make best use of a local user model, using a single kind of personal data, and without explicitly addressing the **heterogeneity** of the existing personal information that may be freely available (on social networks, homepages, etc.). Recognizing this limitation, among other issues: a) **user models** could be based on different types of explicit and implicit personal preferences, such as ratings, tags, textual reviews, records of views, queries and purchases; b) recommended **resources** may belong to several domains and media, and may be described with multilingual metadata; c) **context** could be modeled and exploited in multi-dimensional feature spaces; d) and ranked **recommendation lists** could be diverse according to particular user preferences and resource attributes, oriented to groups of users, and driven by multiple user evaluation criteria.

The aim of the International Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec) is to bring together students, faculty, researchers and professionals from both academia and industry who are interested in addressing any of the above forms of information heterogeneity and fusion in recommender systems. We are interested in raise awareness of the potential of using multiple sources of information, and look for sharing expertise and suitable models and techniques.

RecSys'11, October 23–27, 2011, Chicago, Illinois, USA.
ACM 978-1-4503-0683-6/11/10.

Another dire need is for strong datasets, and one of our aims is to establish benchmarks and standard datasets on which the problems could be investigated. In the second edition of the workshop, we make available on-line **datasets** with heterogeneous information from several social systems. These datasets could be used by participants to experiment and evaluate their recommendation approaches, and con be enriched with additional data, which may be published at the workshop website[1] for future use.

2. TOPICS OF INTEREST
Topics of interest at HetRec included, but were not limited to:

- Fusion of **user profiles from different representations**, e.g. ratings, text reviews, tags, and bookmarks
- Combination of **short- and long-term user preferences**
- Combination of **different types of user preferences**: tastes, interests, needs, goals, mood
- **Cross-domain recommendation**, based on user preferences about different interest aspects, e.g. by merging movie and music tastes
- **Cross-representation recommendation**, considering diverse sources of user preferences: explicit and implicit feedback
- Recommendation of **resources of different nature**: news, reviews, scientific papers, etc.
- Recommendation of **resources belonging to different multimedia**: text, image, audio, video
- Recommendation of **diverse resources**, e.g. according to content attributes, and user consuming behaviors
- Recommendation of resources annotated in **different languages**
- **Contextualization of multiple user preferences**, e.g. by distinguishing user preferences at work and on holidays
- **Cross-context recommendation**, e.g. by merging information about location, time and social aspects
- **Multi-dimensional recommendation** based on several contextual features, e.g. physical and social environment, device and network settings, and external events
- **Multi-criteria recommendation**, exploiting ratings and evaluations about various user/item characteristics
- **Group recommendation**, oriented to several users, e.g. suggesting tourist attractions to a group of friends, and suggesting a TV show to a family

[1] HetRec 2011, http://ir.ii.uam.es/hetrec2011

3. KEYNOTE

Yehuda Koren, from Yahoo! Labs, gave a presentation entitled "I Want to Answer, Who Has a Question? Yahoo! Answers Recommender System".

Yahoo! Answers is currently one of the most popular question answering systems. The author claims, however, that its user experience could be significantly improved if it could route the "right question" to the "right user." While some users would rush answering a question such as "what should I wear at the prom?", others would be upset simply being exposed to it. Community Question Answering systems in general and Yahoo! Answers in particular, all need a mechanism that would expose users to questions they can relate to and possibly answer.

In the presentation, the author proposed to address this need via a multi-channel recommender system technology for associating questions with potential answerers on Yahoo! Answers. One novel aspect of the presented approach is exploiting a wide variety of content and social signals users regularly provide to the system and organizing them into channels. Content signals relate mostly to the text and categories of questions and associated answers, while social signals capture the various user interactions with questions, such as asking, answering, voting, etc. The approach fuses and generalizes known recommendation approaches within a single symmetric framework, which incorporates and properly balances multiple types of signals according to channels. Tested on a large scale dataset, the model exhibits good performance, clearly outperforming standard baselines.

4. DATASETS

In HetRec 2011, we made available three datasets with various types of user preferences about resources belonging to different domains –movies, Web pages, and music tracks– and having diverse meta-information. The datasets are publicly accessible at the workshop website, and are hosted by GroupLens Research group[2], from University of Minesota, USA. In the subsequent subsections, we briefly describe the datasets.

4.1 hetrec2011-movielens-2k

This is an extension of MovieLens10M dataset, which contains personal ratings and tags about movies. From the original dataset, only those users with both ratings and tags have been maintained.

In the dataset, MovieLens[3] movies are linked to Internet Movie Database[4] (IMDb) and RottenTomatoes[5] (RT) movie review systems. Each movie does have its IMDb and RT identifiers, English and Spanish titles, picture URLs, genres, directors, actors (ordered by "popularity"), countries, filming locations, and RT audience' and experts' ratings and scores. Table 1 shows some statistics about the dataset.

Table 1. Data statistics of hetrec-movielens-2k dataset

2113 users
10197 movies
20 movie genres (avg. 2.0 genres/movie)
4060 directors (1 director/movie)
95321 actors (avg. 22.8 actors/movie)
72 countries (1 country/movie) Locations (states, regions, cities, etc.): avg. 5.3 locations/movie
13222 tags 47899 tag assignments (avg. 22.7 tas/user; avg. 8.1 tas/movie)
855598 ratings (avg. 404.9 ratings/user; avg. 84.6 ratings/movie)

4.2 hetrec2011-delicious-2k

This dataset was obtained from Delicious[6] social bookmarking system. Its users are interconnected in a social network generated from Delicious "mutual fan" relations.

Each user has bookmarks, tag assignments, i.e. tuples [user, tag, bookmark], and contact relations within the dataset social network. Each bookmark has a title and URL. Table 2 shows some statistics about the dataset.

Table 2. Data statistics of hetrec-delicious-2k dataset

1867 users
69226 bookmarked URLs 38581 bookmarked principal URLs (e.g. www.delicious.com for http://www.delicious.com, http://www.delicious.com/
104799 bookmarks (avg. 56.1 bookmarked URLs/user; avg. 1.5 users/bookmark)
53388 tags 437593 tag assignments (avg. 234.4 tas/user; avg. 6.3 tas/URL)
7668 bi-directional user relations (avg. 8.2 relations/user)

4.3 hetrec2011-lastfm-2k

This dataset was obtained from Last.fm[7] online music system. Its users are interconnected in a social network generated from Last.fm "friend" relations.

Each user has a list of most listened music artists, tag assignments, i.e. tuples [user, tag, artist], and friend relations within the dataset social network. Each artist has a Last.fm URL and a picture URL. Table 3 shows some statistics about the dataset.

Table 3. Data statistics of hetrec-lastfm-2k dataset

1892 users
17632 music artists (singers, composers, bands)
11496 tags (avg. 18.9 tags/user; 8.8 tags/artist) 186479 tag assignments (avg. 98.5 tas/user; avg. 15.9 tas/artist)
92834 user-listened artist relations (avg. 49.1 artists/user; 5.3 users/artist)
12717 bi-directional user relations (avg. 13.4 relations/user)

[2] GroupLens Research group, http://www.grouplens.org

[3] MovieLens – movie recommendations, http://movielens.umn.edu

[4] Internet Movie Database, http://www.imdb.com

[5] Rotten Tomatoes – movie critic reviews, http://www.rottentomatoes.com

[6] Delicious social bookmarking, http://www.delicious.com

[7] Last.fm Internet radio, http://www.lastfm.com