D5: Databases and Information Systems
Advanced Topics in Information in IR, WS 2014/15
Dr. Klaus Berberich
Assignment 2, due: **27 Nov 2014**

Please submit your solution as a PDF to `atir2014@mpi-inf.mpg.de` *by the indicated due date!*

User-User Collaborative Filtering vs. Item-Item Collaborative Filtering

Problem 1.
In user-user collaborative filtering, we can only determine the similarity between two users, if they have rated at least one common item. Likewise, in item-item collaborative filtering, we can only determine the similarity between two items, if there is at least one user who has rated both. It is typically assumed that $|U| >> |I|$, i.e., there are by far more users than items.

(a) Determine the probability that two users have rated at least one common item. Assume $|U| = m = 10^7$, $|I| = n = 10^5$, that every user has rated exactly $k = 10$ items, and that users determine their ratings independently from each other.

(b) Under the same assumptions, what is the probability that two items have been rated by at least one common user?

Association Rules

Problem 2.
Consider the following (binary) utility matrix of 10 users stuck in the 1970s. We want to generate recommendations for a new user $\mathbf{u}_{11}$ who likes `Queen`, `Cream`, and `Zepp` using the approach based on association rule mining outlined in the lecture.

	Beatles	Yes	Queen	ABBA	Hendrix	Dylan	Cream	Stones	Wham!	Zepp
$\mathbf{u}_1$	1		1		1				1	
$\mathbf{u}_2$					1		1	1		1
$\mathbf{u}_3$	1				1					1
$\mathbf{u}_4$		1	1	1		1			1	
$\mathbf{u}_5$	1						1	1		1
$\mathbf{u}_6$			1	1		1			1	
$\mathbf{u}_7$	1		1	1					1	
$\mathbf{u}_8$		1	1							1
$\mathbf{u}_9$				1	1		1			1
$\mathbf{u}_{10}$	1		1	1		1		1		

(a) Identify all frequent itemsets for a minimum support threshold of 20%. From these, determine all association rules having a confidence of at least 50%.

(b) Using the association rules identified in (a) generate the top-3 recommendations for $\mathbf{u}_{11}$.

Collaborative Filtering and Evaluation (Programming Assignment)

Problem 3.
We provide a dataset from the MovieLens project (`http://www.movielens.org`) on the course website. It consists of 100,000 ratings for 1,682 movies provided by 943 users. Familiarize yourself with the format of the data by reading the `ml-100k-README.txt` file.

(a) Determine the 10 most similar movies to `#64 The Shawshank Redemption` and `#56 Pulp Fiction` using Pearson correlation and cosine similarity.

(b) Determine the top-10 recommendations for user `#101` assuming $k = 5$ using item-item collaborative filtering with Pearson correlation and cosine similarity.

(c) Rate 20 movies from those contained in the dataset and perform the following steps: (i) Generate your top-10 recommendations assuming $k = 5$ and judge them (have you seen any of the recommended movies? did you like it?) (ii) Randomly split your ratings (50/50 training/test) and report the mean absolute error on the test data.

Hint: You can save a lot of effort for this assignment by using a suitable library (e.g., `numpy` for Python or `mtj` for Java) or software (e.g., `Octave/Matlab` or `R`) which provide you with data structures for sparse matrices and operations on them.

Recommending Named Entities

Problem 4.
How would you build a recommender system for general named entities including movies like `Pulp_Fiction`, albums like `Revolver`, but also cities `Akureyri` and brands like `Ben_&_Jerry's`? Assume that you have Wikipedia (including its revision history) and a knowledge base derived from it (e.g., DBpedia) available. What kind of recommendation approach would you use? How would measure similarity between different named entities (possibly in more than one way)? How would you find out about users' utility values? Please describe your idea in detail.

Explicit Semantic Analysis for Cross-Lingual Information Retrieval

Problem 5.
Cross-lingual Information Retrieval (CLIR) targets the scenario where users can understand documents written in one language (e.g., English) but are more comfortable issuing queries in another language (e.g., their native language). Typical approaches to CLIR rely on translating queries and/or documents based on dictionaries. Can you come up with an approach to CLIR that does not require dictionaries but makes use of explicit semantic analysis? Please describe your idea in detail.

Entity Search with Type Cues

Problem 6.
Read the paper by Neumayer et al. [6], which we did not discuss in the lecture. Don't panic it's short and an easy read!. Answer the following questions:

(a) How is their approach different from the papers by Balog et al. [2] and Elbassuoni et al. [5], which we discussed in the lecture, in terms of how queries, data, and results look?

(b) Type cues (e.g., albums, players, movies) are quite common in queries and relatively easy to spot. Can the approach by Neumayer et al. [6] take these into account? If not, describe how you would adapt their structured entity model using predicate-folding (Section 2.3) and their hierarchical entity model (Section 4) to take these into account. To this end, assume that the query q has been split into type cues q_T and descriptors q_D. The query $q = \langle$ bayern munich players jail$\rangle$, as a concrete example, would be split into $q_T = \langle$ players $\rangle$ and $q_D = \langle$ bayern munich jail $\rangle$. Reuse notation from their paper.