D5: Databases and Information Systems
Advanced Topics in Information in IR, WS 2014/15
Dr. Klaus Berberich
Assignment 3, due: **11 Dec 2014**

max planck institut informatik

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

Categorization of Personalization Methods

Problem 1.
In the lecture, we looked at four different methods to personalize web search results. Shen et al. [10] describe two systems to categorize personalization methods, namely according to (a) whether they operate on the client, server, or require a cooperation of the two (**C**, **S**, **CS**) and (b) how much personal information users have to disclose (levels **I**-**IV**). Categorize each of the four methods from the lecture according to the two systems and justify your answer.

Query Difficulty & Locality

Problem 2.
Mei and Church [8] use (conditional) entropy to measure how hard search is and how much knowing the user (via her IP address) can help.

(a) Based on the entropy estimates from the right column in Table 2 from the paper, is it harder to guess a user's identity based on a query she issued or a URL she clicked on?

Assuming that you have their click log data (i.e., plenty of records consisting of a query, a clicked URL, and an IP address) and using the same methodology, how can you accomplish the following tasks?

(b) Identify hard and easy queries. For instance, facebook is considered an easy query, since most users click on `http://www.facebook.com`; causes of economic crisis is considered a hard query, since users click on various of the shown results.

(c) Identify local and global queries. For instance, mensa saarbrücken is a local query, receiving most interest from users located in Saarland; mensa international is a global query, attracting interest from widespread users. Here, you may want to exploit the hierarchical nature of IP addresses and assume that two users are geographically closer if their IP addresses have a longer byte prefix (e.g., `139.15.14.` vs. `139.15.`) in common.

Personalized PageRank (Programming Assignment)

Problem 3.
On the course website you can download an influencer graph among Freebase entities (`freebase-influencers.tsv.gz`) and a description of the dataset (`freebase-influencers-readme.txt`). Vertices in this graph correspond to named entities from Freebase (e.g., Albert Einstein); an edge (u, v) indicates that the named entity u was influenced by the named entity v. As a concrete example, you can see that Albert Einstein was influenced by David Hume. In total there are about 10K named entities in this graph and about 22K influence relationships. Your task is now to implement (Personalized) PageRank on this graph using the power-iteration method described in the lecture. The (Personalized) PageRank score that you compute indicates how influential the corresponding named entity was (on a specified subset of named entities).

(a) When using a uniform random jump vector (i.e., standard PageRank), what are the top-25 most influential named entities in the graph?

(b) What are the top-25 named entities when performing random jumps only to the following ten named entities: John Lennon (3678), Eric Clapton (452), Jimmy Page (3226), Bob Dylan (8479), Jimi Hendrix (1866), Kurt Cobain (1337), NeilYoung (5488), Janis Joplin (182), George Harrison (8855), Miles Davis (6861).

(c) Now pick ten (or more) named entities that influenced you (or that you simply like). What are those named entities; what are the resulting top-25 most influential named entities?

Hints: (i) You can again save a lot of effort by using a library or software that knows how to handle sparse matrices; (ii) Try not to materialize the dense transition probability matrix $\mathbf{P}$, but only keep the sparse matrix $\mathbf{T}$ and the random jump vector $\mathbf{j}$; (iii) There are plenty of dangling vertices in the graph (having no outgoing edges); re-normalize the vector $\boldsymbol{\pi}^{(i)}$ after every iteration, so that $|\boldsymbol{\pi}^{(i)}| = 1$ holds; (iv) Use the parameter choices $\epsilon = 0.2$ and $\delta = 10^{-6}$ for all computations.

Privacy

Problem 4.
Read the paper by Bi et al. [1]. Answer the following questions regarding their approach:

(a) In your own words, how do they map Facebook Likes and search engine queries into a common representation in terms of Open Directory Project (ODP) categories?

(b) What is the AUC score they use? What is plotted in a receiver operating characteristic (ROC) curve? Do the data points correspond to a single system configuration? If not, how are they obtained?

(c) Consider Table 2 from the paper. Why are the AUC scores for religion and political view missing from the table? Are they generally unable to predict those for search engine users?

(d) Assuming that a search engine deployed their approach to infer user traits from queries, but you as a user had a desire to hide your traits. How can you achieve this, or to put it differently, how can you make inference hard for them?