Google news personalization: scalable online collaborative filtering

Description of collaborative filtering for generating personalized recommendations for users of Google News.

3 approaches:

1. Collaborative filtering using MinHash clustering
2. Probabilistic Latent Semantic Indexing (PLSI)
3. Convisitation Counts

The project combines recommendations from different algorithms using a linear model. This approach is content-agnostic and consequently domain independent.

Project requirements:

* Scalability: several million unique visitors as well as several millions news stories.
* Item churn: In Google News the underlying item set undergoes churn every few minutes and at any given time the stories of interest are the ones that appear in last couple hours. Therefore, any model older than few hours may no longer be of interest and partial updates will not work.

Problem: Google News aggregates news articles from more than 4.500 news sources worldwide.

\*\*A user click on an article is treated as a positive vote for the article.

Algorithm: Mix of memory based (item convisitation) and model based (2 clustering techniques a.PLSI and b.MinHash) algorithm to generate recommendations.

1. Each of these algorithms assigns a numeric score to a story.
2. Scores given by each of these algorithms are combined to obtain a ranked list of stories.
3. Top *K* stories are chosen from this list as recommendations for the user.

Putting the algorithms into a real-time recommendation system requires the following three main components:

1. An offline component responsible for periodically clustering users based on their click history.
2. A set of online servers responsible for performing two main types of tasks:
   1. Updating user and story statistics
   2. Generating news story recommendations.
3. And 2 types of data tables:
   1. User table UT
   2. Story table ST

Evaluation

We use 3 tests datasets for the study:

1. Movies lens dataset: 943 users, 670 movies and 54.000 ratings.
2. NewsSmall dataset: Subset of click received in Google News 5000 users, 40.000 unique items and 370.000 clicks
3. NewsBig dataset: Subset of click received in Google News with more records 500.000 users, 190.000 unique items and 10’000.000 clicks.

\*\*We randomly divide the datasets into a training set and test set. 80% train and 20% test.

Results: The algorithms, although more scalable, do not incur a loss in quality and on the contrary do better in terms of quality.