Project 4 Requirements

Charley Ferrari

July 4, 2016

Project 4 Requirements

I will be using Yahoo's Delicious popular URLs and tags dataset. This dataset represents 100,000 URLs that were bookmarked on Delicious, each of which were saved at least 100 times.

This is a highly contextual database. Time data can be inferred by the date that it was first bookmarked by a delicious user is indicated, and Delicious allows users to tag URLs with topics, which gives categories to the URLs.

Using this data, I hope to be able to recommend URLs within the dataset not viewed by users, and allow a framework to incorporate new URLs. Because URLs aren't actually rated, this will be done in a binary way (1 for a URL that a user has tagged, and 0 for a URL not tagged.)

My technique will have to be slightly different than what I have used for numerical ratings. In previous cases, numerical ratings give a clearer picture of what has been rated, with 0's referring to items that have not been rated. The input data is a sparse matrix, which is filled in.

Because we simply have URLs that are tagged and not tagged, we won't technically have a sparse matrix. Nonetheless, using contextual information of the URLs, I believe I will be able to calculate a similarity matrix between different URLs, and rank the remaining URLs as they would be viewed by different users.

This sort of information can be used by a company like Yahoo to keep users within the delicious ecosystem. As they see relevant recommendations, users will be incentivized to use the system more, and provide more recommendations to make the system stronger.