Recommendations for Items Using Web Crawl Techniques, Tag Cluster Data, Cosine Similarity Calculation, and Wikipedia Title Matching

James Xue, SEAS ’17, Computer Science, Columbia University, james.xue@columbia.edu

Supervising Faculty, Sponsor, and Location of Research
Professor Kotaro Nakayama, Matsuo Web Laboratory, The University of Tokyo

Abstract
Asia Trend Map, a unique website displaying thousands of Japanese and international cartoons, comics, books, games, and more. Sponsored by the Ministry of Economy, Trade, and Industry in Japan, its aim is to visualize the current and 6-months-ahead popularity of these titles. The goal of this project is to create more engaging content for users on the Asia Trend Map website. In particular, recommendations are a good way to link to different content in our site, and for users to explore new titles. To do this, I investigated collaborative filtering, a technique used to make automatic predictions about the interests of a user by collecting preferences. However, I eventually used web crawl techniques to obtain tag data for several thousand titles from Anikore, a popular anime website in Japan. Then, I performed a cosine similarity calculation to identify the four most similar titles corresponding to each title. Then, I used a Google Search Query to match Wikipedia titles to the titles stored on the back end of Asia Trend Map. Finally, I presented the recommended titles on the website.

After comparing the accuracy and speed of this approach on a dataset of over 8000 tags, I concluded that tag by item clustering proved to be more accurate than other alternative approaches, in about the same amount of time. The accuracy was also confirmed through several human tests as well.

Keywords
Data mining, collaborative filtering, item clustering, recommendations