Predicting User Visit in Stochastic Web Surfing

Sachin Mohan Sujir, Himanshu Vinod Nirmal, Krishna Prasad Neupane
Rochester Institute of Technology

October 2019

Problem Definition

Web surfing is now common and used in every sector to access the information. Its extensive use is due to its easiness, flexibility, and ability to provide immediate information to the user. The problem of our project is to recommend the websites that a user may visit based on what other websites he or she visited. Recommending the websites that a target user may visit based on the website visits made by the other users of the cluster to which our target user belongs to. This will help the user to search/choose faster with the recommendations provided by the model we create. So, finding the possible new visit areas of the web for each is the area to focus on.

Solution Approach

To predict user visits of the websites we will be using unsupervised learning algorithms. The unsupervised algorithm clusters the users who visited most of the similar websites and then we predict the websites for the users based on their clusters. We will use K-means clustering as a core algorithm. We will also apply the matrix factorization technique of a collaborative filtering approach as an extra work to recommend the website to the users. The core algorithm that we will be using is the K-means Clustering algorithm to cluster the web areas. We are also planning to use Hierarchical Clustering, Density-based Spatial Clustering of Applications with Noise to compare performance between different algorithms. We use the Unsupervised learning model as we do not have class labels and users will be clustered together based on their similarity of website visits of other users.

The target users are general users who use the website. Recommendation has become a trend today and users expect recommendations for faster access and search. An active user will get suggestions to visit other websites of similar content thus giving them more options to explore and gather more information. This recommendation will be based on the active user's previous search and the similarity with users who have similar searches. The active user will be suggested with a website based on other users(voting- based on other users who have visited that area) who belongs to the cluster or user nearest to the mean of the cluster

Data Collection

We will be using anonymous web data from www.microsoft.com available in the UCI KDD archive. The data has different training and test set of users-website interaction. The dataset contains 32711 training data and 5000 data reserved for testing. The dataset has 294 attributes(Web Areas), which is called vroots in the dataset. Every user is identified by a unique number and each attribute is an area of www.microsoft.com. The mean of website visit per case is 3.0(approx). https://kdd.ics.uci.edu/databases/msweb/msweb.html [1] is the link to the data set we will be using.

References

[1] Seth Hettich and SD Bay. The uci kdd archive [http://kdd.ics.uci.edu/databases/msweb/msweb.task.html]. irvine, ca: University of california. *Department of Information and Computer Science*, 152, 1999.