Predicting User Visit in Stochastic Web Surfing Checkpoint-2

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

October 2019

Data Collection

We used anonymous web data from www.microsoft.com available in the UCI KDD archive [1]. The data has different training and test set of users-website interaction. Our data consist of following statistical summary:
Training Instances= 32711
Testing Instances =5000
Attributes= 294
Mean vroot visits per case= 3.0

The format of the dataset is ASCII-based sparse data format and each line starts with a letter representing the type of the line. For example: 'A,1277,1,"NetShow for PowerPoint","/stream"' Where: 'A' represents it is an attribute line, '1277' is the attribute ID number of the website (called a Vroot), '1' may be ignored, '"NetShow for PowerPoint"' is the title of the Vroot, '"/stream"' is the URL relative to "http://www.microsoft.com"

Data Preprocessing and Data Visualization

Data cleaning was not necessary for the data as the data had no missing values and no steps had to be taken to clean the data. We processed the data to perform data transformation and created User : Web area pair. We initially performed data preprocessing in which we created a user-website interaction matrix by encoding user visits to a particular website as 1 and else 0. The sample of a tabular representation of our pre-processed data is shown in Table 1 below:

Initial Result of the Core Algorithm

We used our core K-mean algorithm to cluster users and predict website for the users. The following Table 2 shows initial results for some users before

Table 1: Preprocessed User-Website Interaction Data

User	Web 1001	Web 1026..	Web 1034..	Web 1294
10001	0	1	1	0
10002	0	0	0	0
10003	0	0	0	0

fine-tuning of model:

Table 2: K-mean Clustering to predict web visit for first four users

User	Previous Website	Predicted Website
10001	[1038, 1026, 1034]	[1008]
10002	[1008,1056,1032]	[1004]
10003	[1064, 1065, 1020, 1007, 1038, 1026, 1052, 1041, 1028]	[1004]
10004	[1004]	[1004]

In this K-mean clustering, we select K=5 and used 30 % of training data to generate validation set to predict web visit.

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.