Dataset Specifications

The data set that we are using is the YOW User Study Data which was originally collected by the PI at Carnegie Mellon University. The goal of this user study was to collect implicit and explicit data from the users participating in the study on a heterogenous set of documents and a wide range of information about documents and topics. Twenty one subjects participated in this study for four weeks, the implicit and explicit data was collected in the following format:

1 Explicit Data Collection

After the user finished reading an article the browser would redirect user to a form where the user would suggest the class for the article, rate the article basically by providing a scale of the user’s liking for the article. If the article was assigned classes already, the user was asked to determine the relevance of the article with the assigned classes. The user also had the opporunity to assign new classes to the article, the articles were also rated based on measures such as readable, authorative, novel and relevant.

2 Implicit Data Collection

The browser used for this user study recorded the user’s mouse and keyboard activites, the number of seconds the user spent on the article as well as the number of times the user scrolled while reading an article. These methods of implicit data collection were only at work when the mouse pointer was focued within the article, whenever it went out of focus the recording stopped.

From the original data set we considered only the attributes User id, Doc id, relevant and classes for the purpose of our recommendation. Furthermore we divided this data set into two models, the first part contained the User id, doc id and the rating given by the user and the second partition of the data contained the doc id and classes.

The data contained over 7000 article ratings from the given 21 users. We have used all such ratings for the purpose of our recommendation system.

DataSet URL: https://users.soe.ucsc.edu/~yiz/papers/data/YOWStudy/