**Experimental Datasets**

The datasets for performing experiments and evaluating the recommendation systems have been obtained from <http://www.reddit.com>. Reddit is a popular social news website, wherein, users frequently post links to news articles, images, internet memes and other popular and interesting articles on the web. The site consists of various sub-reddits, where each sub-reddit defines a category. For example, programming, java, politics, technology, music are some of the popular sub-reddits of Reddit. Users are allowed to up-vote or down-vote the links based on whether the links are of interest to them or not. A user can cast a maximum of one up-vote or one down-vote on each link.

For our experiments, we consider each sub-reddit as a category, and we have obtained datasets from Reddit, where each line is in the format:

*account-id link-id sub-reddit-id +1/-1*

*account-id*, *link-id* and *sub-reddit-id* are hashes of the username, link and the sub-reddit respectively. +1 indicates an up-vote on the link by the user, whereas -1 indicates a down-vote on the link by the user.

Reddit would not disclose the actual usernames, links up-voted and down-voted and sub-reddits due to privacy reasons. Since we are recommending categories to users, the *link-id* is irrelevant. Moreover, for each unique pair of *account-id* and *sub-reddit-id*, we need a representation that indicates whether the affinity of the user towards links belonging to the category. This is obtained by calculating the total number of up-votes and down-votes for every (*account-id*, *sub-reddit-id*) pair.

To prevent those users that are more active or sub-reddits that have received more number of votes from dominating the predictions, the number of up-votes and down-votes for every (*account-id*, *sub-reddit-id*) pair is normalized to fit in the range [-1.0, 1.0]. Min-Max normalization is performed using the following formula:

Applying Min-Max normalization for the up-votes and down-votes to the range [-1.0, 1.0], we obtain the following formula:

The raw dataset is now converted to lines of the form:

*account-id sub-reddit-id affinity*

Here, an affinity of 1.0 denotes that the user has up-voted all the links he has seen in that category (sub-reddit). An affinity of -1.0 denotes that the user has down-voted all the links belonging to that category. An affinity of 0.0 denotes the user’s indifference to links of that category, or equal number of up-votes and down-votes to links belonging to that category. Heuristically, an affinity value that is greater than 0.5 indicates a positive recommendation, while affinities less than 0.5 indicate the category is not of much interest to the user.

From the above normalized dataset, we have chosen two datasets for our experiments - one with 100 attributes (categories) and a larger dataset with 200 attributes. The two datasets, *attr100* and *attr200* respectively, are represented in a sparse manner in the files as pairs of (*account-id*, *sub-reddit-id*), along with the affinities. They are converted into a matrix, where each row represents a user and each column represents a category. The entries in the matrix for which the affinities are not present in the dataset, are filled with 0.0, as this indicates that the user has not voted links belonging to the corresponding category.

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Number of rows (users)** | **Number of columns (attributes or categories)** |
| attr100 | 9160 | 100 |
| attr200 | 21679 | 200 |

Dimensions of the experimental datasets