Anonymous Ratings from the Jester Online Joke Recommender System

Dataset 1: Over 4.1 million continuous ratings (-10.00 to +10.00) of 100 jokes from 73,421 users: collected between April 1999 - May 2003.
Dataset 2: Over 1.7 million continuous ratings (-10.00 to +10.00) of 150 jokes from 59,132 users: collected between November 2006 - May 2009.
Dataset 2+: An updated version of Dataset 2 with over 500,000 new ratings from 79,681 total users: data collected from November 2006 - Nov 2012

Freely available for research use when acknowledged with the following reference:

Eigentaste: A Constant Time Collaborative Filtering Algorithm. Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Information Retrieval, 4(2), 133-151. July 2001.

As a courtesy, if you use the data, I would appreciate knowing your name, what research group you are in, and the publications that may result.

Dataset 1

Over 4.1 million continuous ratings (-10.00 to +10.00) of 100 jokes from 73,421 users: collected between April 1999 - May 2003

Save to disk, then unzip to obtain Excel files:

- jester_dataset_1_1.zip: (3.9MB) Data from 24,983 users who have rated 36 or more jokes, a matrix with dimensions 24983 X 101.
- jester_dataset_1_2.zip: (3.6MB) Data from 23,500 users who have rated 36 or more jokes, a matrix with dimensions 23500 X 101.
- jester_dataset_1_3.zip: (2.1MB) Data from 24,938 users who have rated between 15 and 35 jokes, a matrix with dimensions 24,938 X 101.

Format:

1. 3 Data files contain anonymous ratings data from 73,421 users.
2. Data files are in .zip format, when unzipped, they are in Excel (.xls) format
3. Ratings are real values ranging from -10.00 to +10.00 (the value "99" corresponds to "null" = "not rated").
4. One row per user
5. The first column gives the number of jokes rated by that user. The next 100 columns give the ratings for jokes 01 - 100.
6. The sub-matrix including only columns {5, 7, 8, 13, 15, 16, 17, 18, 19, 20} is dense. Almost all users have rated those jokes (see discussion of "universal queries" in the above paper).

The text of the jokes can be downloaded here: jester_dataset_1_joke_texts.zip (92KB)

Format:

1. 100 files
2. Each file has title init_.html, where _ is 1 to 100
3. The titles correspond to the ID's of the jokes in the Excel files above

Dataset 2

Over 1.7 million continuous ratings (-10.00 to +10.00) of 150 jokes from 59,132 users: collected between November 2006 - May 2009

Save to disk, then unzip: jester_dataset_2.zip (7.7MB)

Format:

- jester_ratings.dat: Each row is formatted as [User ID] [Item ID] [Rating]
- jester_items.dat: Maps item ID's to jokes

Note that the ratings are real values ranging from -10.00 to +10.00. As of May 2009, the jokes {7, 8, 13, 15, 16, 17, 18, 19} are the "gauge set" (as discussed in the Eigentaste paper) and the jokes {1, 2, 3, 4, 5, 6, 9, 10, 11, 12, 14, 20, 27, 31, 43, 51, 52, 61, 73, 80, 100, 116} have been removed (i.e. they are never displayed or rated).

Dataset 2+

An updated version of Dataset 2 with over 500,000 new ratings from 79,681 total users: data collected from November 2006 - Nov 2012

Save to disk, then unzip: jester_dataset_2+.zip (5.1MB)

Format:

- In this dataset we stripped out users that did not respond to the gauge set of question. The data is formated as an excel file representing a 66336 x 151 matrix with rows as users and columns as jokes.
- 10 of the jokes don't have ratings, their ids are: { 1, 2, 3, 4, 6, 9, 10, 11, 12, 14 }.
- Each rating is from (-10.00 to +10.00) and 99 corresponds to a null rating (user did not rate that joke).

Note that the ratings are real values ranging from -10.00 to +10.00. As of May 2009, the jokes {7, 8, 13, 15, 16, 17, 18, 19} are the "gauge set" (as discussed in the Eigentaste paper) and the jokes {1, 2, 3, 4, 5, 6, 9, 10, 11, 12, 14, 20, 27, 31, 43, 51, 52, 61, 73, 80, 100, 116} have been removed (i.e. they are never displayed or rated).

Other Collaborative Filtering Datasets:

- The MovieLens Dataset: 1,000,000 integer ratings (from 1-5) of 3500 films from 6,040 users.
- The EachMovie Dataset: 2,811,983 integer ratings (from 1-5) of 1628 films from 72,916 users.
- The BookCrossing Dataset: 1,149,780 integer ratings (from 0-10) of 271,379 books from 278,858 users.

Papers using the Jester Dataset (or a subset from 18,000 users)

For further information please contact:

Ken Goldberg
goldberg at berkeley dot edu
Prof of IEOR and EECS
4135 Etcheverry Hall
University of California
Berkeley, CA 94720-1777
(510) 643-9565 (phone)
(510) 642-1403 (fax)