msnbc.com anonymous web data

Data Type

Discrete sequence

Abstract

This data describes the page visits of users who visited msnbc.com on September 28, 1999. Visits are recorded at the level of URL category (see below) and are recorded in time order.

Sources

David Heckerman (heckerma@microsoft.com)

Data Characteristics

The data comes from Internet Information Server (IIS) logs for msnbc.com and news-related portions of msn.com for the entire day of September, 28, 1999 (Pacific Standard Time). Each sequence in the dataset corresponds to page views of a user during that twenty-four hour period. Each event in the sequence corresponds to a user's request for a page. Requests are not recorded at the finest level of detail---that is, at the level of URL, but rather, they are recorded at the level of page category (as determined by a site administrator). The categories are "frontpage", "news", "tech", "local", "opinion", "on-air", "misc", "weather", "health", "living", "business", "sports", "summary", "bbs" (bulletin board service), "travel", "msn-news", and "msn-sports". Any page requests served via a caching mechanism were not recorded in the server logs and, hence, not present in the data.

Other Relevant Information

- Number of users: 989818
- Average number of vitis per user: 5.7
- Number of URLs per category: 10 to 5000

Data Format

Here are the first 20 lines of the ascii data file:

```
% Different categories found in input file:

frontpage news tech local opinion on-air misc weather msn-news health living business msn-sports sports summary bbs travel

% Sequences:

1 1
2
3 2 2 4 2 2 2 3 3
5
1
6
1 1
6
6 7 7 7 6 6 8 8 8 8
6 9 4 4 4 10 3 10 5 10 4 4 4
1 1 1 11 1 1 1
12 12
1 1
```

Each category is associated--in order--with an integer starting with "1". For example, "frontpage" is associated with 1, "news" with 2, and "tech" with 3. Each row below "% Sequences:" describes the hits--in order--of a single user. For example, the first user hits "frontpage" twice, and the second user hits "news" once.

Past Usage

I. Cadez, D. Heckerman, C. Meek, P. Smyth, S. White, "Visualization of navigation patterns on a Web site using model-based clustering," Journal of Data Mining and Knowledge Discovery, accepted for publication.

Acknowledgements, Copyright Information, and Availability

This data is avaliable thanks to msnbc.com

The UCI KDD Archive
Information and Computer Science
University of California, Irvine
Irvine, CA 92697-3425
Last modified: 20 Feb 2003