Activity 1. Web browsing Probability question

Computer Scientists are being called more and more frequently to provide computer log data that can be used to find out how users interact with the Internet (see, for example, Sen and Hansen (2003), Cadez et al. (forthcoming), Huberman et al (1998)). In addition to that, the number of Internet user surveys is growing at an amazing speed (see, for example, Wellman and Haythornthwaite (2002)). The following problem uses a data set published in the UCI KDD Archive (Kederman, 2003) to obtain data on the different pages visited by users who entered the msnbc.com page on September 28, 1999.

Data Characteristics

The data on which this activity is based comes from Internet Information Server (IIS) logs for msnbc.com and news-related portions of msn.com for the entire day of September 26, 1999 (Pacific Standard Time). Each sequence in the dataset corresponds to page views of a user during that twenty-four hours 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 levels 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".

To acquaint yourself with the raw data, go to **Data URL**
http://kdd.ics.uci.edu/databases/msnbc/msnbc.data.html

The total number of users is $n = 989818$. All together, they did a total of 5,688,612 page hits.

Activity

The table below describes how many of the users visited the corresponding page category.

Summary Table

Pages	number of users that visited it	proportion of users
Frontpage	338056 users	
News	264016	
Tech	196461	
Local	228143	
Opinion	50326	
On-air	218560	
Misc	89708	
Weather	95615	
Health	90192	
Living	50606	
Business	57597	
Sports	112183	
Summary	76948	
bbs	119138	
Travel	29200	
MSN-news	2082	
MSN-sports	11006	

Annswer the following questions:

1. Compute the proportion of users that visited each page (column 3).
2. Which page was the most popular? Which was the least popular?
3. Is this table a relative frequency table?
4. Suppose we draw a user at random. What would be the probability that this user hit the Travel page? What would be the probability that she or he hit the Sports page?
5. Would it be possible to compute the probability that a user hit at least one page based on the summary table? YES or NO? Why or Why not?

References

(1) Cadez, I., Heckerman, D., Meek, C., Smyth, P., White, S. (2003). Model-based clustering and visualization of navigation patterns on a Web site. Accepted for publication on Journal of Data Mining and Knowledge Discovery, 7(4).
(2) Hansen, M.H. and Sen, R.(2003). Predicting Web User's next access based on log data. Journal of Computational and Graphical Statistics, Vol 12, No. 1, March, p. 143-155.
(3) Heckerman, D. The UCI KDD Archive

http://kdd.ics.uci.edu

Irvine, CA: University of California, Department of Information and Computer Science.
(4) Huberman, B.A., Pirolli, P.L.T., Pitkow, J.E., Lukose, R.M. (1998). Strong Regularities in World Wide Web Surfing. Science, Vol. 280, 3 April.
(5) Wellman, B. and Haythornthwaite edts.(2002). The Internet in Everyday Life. Blackwell Publishing.