Student Search Patterns in a Statistical Web-DB

Joan C. Nordbotten and Svein Nordbotten
University of Bergen

1. Introduction

National Statistical Agencies are increasingly giving free access to Web databases containing national economic and social statistics. The data content, search & retrieval systems, statistical processing programs, and DB usage logs are maintained by the agencies. With the multitude of Web databases available today, it is important to analyze usage patterns so that improved services can be made.

The purpose of this paper is to explore log data as an information source for improving user interfaces. The particular case considered is the use by students in educational organizations of a statistical database to acquire social science data. This study is related to other studies we have made aiming at information for improving the interface to websites [Nordbotten & Nordbotten 1999, 2001a and 2000b].

The current paper presents an analysis of *student usage patterns* in a national statistical Web-DB developed and ran by Statistics Sweden. This database was first made available on the Internet in 1997 and, in September 2000, contained some 800 statistical tables within 19 topic areas, such as health, housing, labor, population, etc. There is also a hierarchical set of *meta-data* describing the tables within each topic area. A special set of usage logs have been created for this database, which include user registration data and entries relating users to the metadata and statistical data they retrieve. In addition, the type of data retrieval used is kept: 1.view to the user display, 2.downloading as a text file, and 3.downloading to a statistical analysis package freely available from the statistical office. The study is based on 2 samples from the DB usage log taken in September 1999 and September 2000, allowing both an identification of usage patterns as well as a study of their development between the 2 time periods.

2. User organization growth

Users of the study Web-DB must initially register to receive a unique user identification. The registration process also gives such data as user category (government, business, education, personal, etc) and geographic location. In Sept. 2000, there were 8316 registered users, 89% of which were organizations from which multiple individuals, or secondary users, could be active. Of these, 497 were educational organizations of which 349 were university/colleges, and 148 high schools (other schools). In addition, there were 1064 private and library users, some of whom have given their user category as education. Since organizational users have multiple secondary users, the number of users is indicating organizations, and not the number of students.

Figure 1 shows the development of number of user organizations registered from September 1999 to September 2000, and it is interesting to see that the organizational interest in taking advantage of this source of information is strong in both educational organization groups which both increased by an factor of 3. Compared with the total growth of registered users, the educational organizations show a modest growth rate, mainly because of the exploding interest among private citizens and economic enterprises came later than in education.

3. Use of the Web-DB

To obtain access to the DB, a student must use the unique identification of his registered organization to authenticate himself. When a student is authenticated and when he is clicking on any any link within the database, a *request* is recorded in the log. The DB log used for this study records the point of time for the request, the organization identification (not the IP address) with each request, and the page or service requested. The identification of the organization made it difficult to separate individual *student-sessions* since frequently there were concurrent student-sessions from the same organization. This is particularly a problem for the educational organizations. In an initial study [Nordbotten, 2000], a *day-session*, defined as all activity from an organization within 1 day, was used. Users from educational organizations had 487 day-sessions in September 2000, which means that each organization was in average active one day that month. A typical student-session is assumed to include an initial request to the DB, followed by several requests for meta-data within the topic of interest and terminated by a request for statistical information in one of the 3 forms described above.

In order to get a clearer view of individual student access patterns to the statistical data, the current study uses a *topic-session*, defined as the sequence of requests from initiation of a topic search to data retrieval in that topic. As earlier, a topic-session is not equivalent to a student-session, in as much as a single student may well retrieve data from multiple topic areas in one session. In the paper, student usage characteristics and trends are discussed and compared to those of the general user community. As in the earlier study, suggestions are given to the design of the log data to support future studies of this kind.

In September 2000, there were 2.916 requests logged for statistical outputs. With the assumptions made above, this indicates 1139 topic-sessions from university and 1777 from high school students. Only 42 university users and 20 high school users of the registered users were active in September 2000. The active university users of the Web-DB was 12% and the active high schools users 14% of the registered users. An explanation of the relative low rates can be that the September is the first month of the fall term. There were in average about 27 sessions per active university and about 88 per active high school. This indicates that use in the high schools were more class organized than in the universities were the students pursued personal needs.

Another interesting characteristic is the *output volume* of requested. In the log, the volume of statistical output is recorded in rows provided. The records show that the university/college students requests smaller outputs, less than 3 rows in average per topic-session compared with the high school students who requested in total 6084 row of output in 1777 topic-sessions. This may be explained with a more goal oriented search by the university students, a hypothesis supported by the fact that the university students have a higher request rate for file transfer than the high school students who had a high rate of request for information as screen displays.

The number of requests between a student's entry to the DB until his request for statistical output reflects how easily the student located the statistical information wanted. These requests are referred to as meta-requests. To measure this number was difficult to for 2 interrelated reasons: the problem of separating concurrent user students from the same registered organization, and the uncertainty associated with determining the start and the end of a session. A useful indicator of the students' activities to locate wanted information may be the average number of meta-requests per output request. For university students the average number of meta-requests per output was 2.6 while the average for high school students was 2.7 meta-requests per output.

Figure 2 presents a summary comparison of the log findings for university and high school user organizations.

4. Preliminary conclusions

During the analysis of the database and the log data we encountered a number of questions which require consideration in connection with interface data capture and design of logging systems:

1. Separation of data for concurrent individual students in the log was difficult because the registered users were organizations (universities/colleges/high schools) with many secondary users (students). One possible solution is to assign a temporary identification to each entrance request, and maintain this for the IP platform until new entrance request is received by the DB. The risk is that the student leaves his computer and a new takes over without making an entrance request.
2. Linking of request data was uncertain partly because of reason mentioned in 1. and partly because the log data used unfortunately did not include referrer page, i.e. the previous page in a chain of requests. In addition to the temporary identification mentioned above, introduction of referring page in the log record will be a great improvement. The referring page is a standard item in the hidden information accompanying a request and can easily be saved in the log.
3. A drawback of all log data is that the times only refer to request time, not to the time work on the request is finished. For a sequence of requests, we can assume that the processing of one request is finished by the time of the next request. The

end of the processing of the last request will, however, be undetermined. An obvious solution to this problem is to introduce a dummy request as a log off procedure, but it cannot be enforced since nothing prevents the user from leaving or turning off his computer without logging off.

4. *User* and user *session* are important concepts the definitions of which depending on the purpose of the studies wanted. Still, these concepts should be discussed with designers of log systems to make the log data as useful as possible for later evaluations.

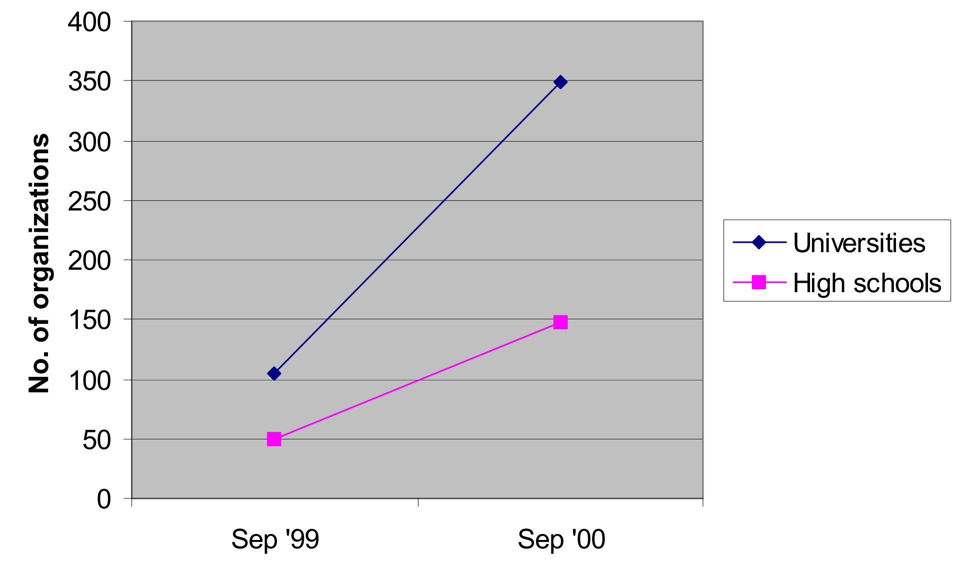
5. References

Nordbotten, J. and S. (1999): SEARCH PATTERNS IN HYPERTEXT EXHIBITS. HICSS-32. Proceedings of The Thirty-Second Annual Hawaii International Conference on System Sciences. IEEE. 1999 ISBN 0-7695-0001-3.

Nordbotten, S. and J. (2001a): PERCEPTION OF STATISTICAL PRESENTATIONS INVESTIGATED BY MEANS OF INTERNET EXPERIMENTS. Hicss-34. Proceedings of the 34th Hawaii International Conference on System Sciences. January 3-6 2001. Institute of Electrical and Electronic Engineers. ……….

Nordbotten, S. and J. (2001b): A STUDY OF THE SSD WEB LOGS FOR SEPTEMBER 1999 AND SEPTEMBER 2000. A Report to Statistics Sweden, Stockholm.

Figure 1: Growth of registered organizations



Registered in DB:	***8316***	**users**
Universities:	***349***	**organizations**
High schools:	***148***	**organizations**
University and highschool use	***487***	**day-sessions**
Topic-sessions (student-sessions	***2916***	**sessions**
Universities:	***1139***	**sessions**
High schools	***1777***	**sessions**
Output volume:	***9217***	**rows**
Universities:	***3133***	**rows**
High schools	***6084***	**rows**
Meta-requests per output request	***2,7***	**requests**
Universities:	***2,6***	**requests**
High schools	***2,7***	**requests**

***Figure 2*: Main characteristics of DB use by students**