ANALYSIS ON EINS ACCOUNTING DATA

Introduction

Information providers are facing a highly competitive environment. The advent of the Internet seems to render the whole world’s knowledge freely available to the individual at a click of the mouse. New technologies, new services and new competitors appear on Internet everyday. In this dramatically changing scenario, it is very important to improve the understanding of customer’s behaviour and requirements.

We address the need of an information provider that offers a subscription-based unified access to Scientific and Technological databases: EINS (European Information Network Services). The service collects, for accounting purposes, data on system usage through an application server. These data are much richer than web logs data usually collected by web servers and permit the recognition of the single events that take place inside a user session.

During a session, the user can perform different actions: search databases, print the results in different formats, use special features like zoom (frequency analysis of indexing terms) and expand (examination of sections of the database's alphabetic dictionary or specific indexes). A different fee is requested for each action. The fee for the same action can differ depending on the database in use, too. Logs relative to these actions are recorded in the accounting file. Moreover, the user can set queries (SDI) that run automatically at prefixed times. This activity is recorded in a separate log file. We also used the registry file that contain administrative information about the users such as the national centre they belong to, the paying status, the nationality, the age of the contract.

We show how these data can be successfully mined to extract knowledge on user behaviour. We apply a clustering algorithm to the accounting data, summarized at the user level, to obtain a customer segmentation. The results show the existence of different strategies of usage of the service resources. These segmentation results serve to focus the attention on specific set of customers to understand and address their common needs, opportunities or threats. The obtained user profiles provide the basis for targeted marketing campaigns, customer retention and service improvement.

Objective

This analysis aims at identifying the main typologies among EINS users.

Description of the analysis

The data analysed come from the log file of event data, collected for accounting purposes, during the month of June 2000. The available fields were:

user file event items year yearday month day startime endtime uniprice format

By means of this information, and of the demographic information (user registration), it was possible to describe each user in term of:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>user</td>
<td>User identification code</td>
<td>Id</td>
</tr>
<tr>
<td>sessioni</td>
<td>Number of sessions (days of connection to the service)</td>
<td>Active</td>
</tr>
<tr>
<td>duratamg</td>
<td>Average length of session (minutes)</td>
<td>Active</td>
</tr>
</tbody>
</table>
We considered a sequence of sessions made on the same day as a unique session. As a matter of fact, we observed that sessions are ended and immediately started again, by the same user. Some of these real sessions are probably started only because of system failure and may be considered as a unique working session.

As a consequence of this choice, the variable “number of sessions” counts the days the user has been connecting to the service during the month, while the variable “number of real sessions” counts the “begin” commands the user made during the month.

In many cases the event “logoff” is missing and, as this event is the only one that determines how long the session lasted, the connection time and average length of each session are underestimated.

For the same reason, the percentage of connection time devoted to each data base is not very accurate and we decided to define which data base mostly interests the user in terms of percentage of documents extracted.

To find out the main customer typologies, we need to group the customers by behaviour. To do so, some variables have been assigned an active role in defining the similarities between users, others have been assigned a supplementary role (they will only contribute to the description of the identified groups) and others have not been used (neutral role).

**Results**

Among the 190 users that connected to the EINS service at least once during the month of June 2000, 35 different typologies have been discovered. Some users make a typology on their own (a group consisting of a single user) as they showed a different behaviour from all the other users.

Here’s a description of the first 10 groups (accounting for 81% of the users).

**“A” typology (17% of users): unsuccessful explorers**

They only made one session during the considered period of time, connecting to only one data base, and making a few searches. They didn’t extract any document, the connection time is very low (6.3 minutes avg) and the expense is the lowest (4.5 ECU avg).

**“B” typology (14% of users): CHEMABS users**
They only access the CHEMABS data base, they extract relatively few documents (96 avg) but spend a lot (145,4 ECU avg). Most of them (67%) only make one session.

“C” typology (12% of users): **INSPEC users**
They only access the INSPEC data base. The number of documents retrieved and the level of expense are medium (84 documents and 36,9 ECU avg).

“D” typology (11% of users): **intensive users / exploiters**
These users make very frequent sessions (an average of 7 session-days during this period), of medium length. They perform many searches (206 avg) on different data bases (NASA 83%, INSPEC 83%, PASCAL 58%, METADEX 50%, MEDLINE 33%, …), extracting quite a lot of documents (272 avg), but the level of expense is medium (43 ECU avg). This group of users has the maximum connection time (125 minutes avg). Half of them are non paying users.

“E” typology (8% of users): **thrifty users (?)**
They make more than one session (2 avg) and access to more than one data base, mainly CHEMABS (71%), but the usage is very low (only 10 searches, only 32 documents and 20 ECU avg). Each session has a very short length and the total connection time is very low.

“F” typology (8% of users): **less experienced users (unskilled)**
With a very small number of searches (6), they retrieve a huge amount of documents (382), generating an average expense of 263 ECU (which is the highest among the identified groups). They mainly use one data base and the session length is the highest (55 min. avg).

“G” typology (8% of users): **best users (ideal)**
As the “intensive users” (D typology), they make many sessions and work on many different data bases. Among them, CHEMABS (100%), AGRISUP, NTIS, …They do the highest number of searches (228 avg) and retrieve a very high number of documents (311 avg). With respect to typology “D”, they also have a very high session length and a very high level of expense (147,6 ECU avg).

“H” typology (8% of users): **skilled “quick and dirty” searchers**
Only one session, only one data base, they have the smallest number of searched terms (2,7 avg) and a small number of retrieved documents (22). The connection time and level of expense are very low.

“I” typology (7% of users): **“quick and dirty” searchers (2)**
Very similar to the “H” typology, they are even quicker (3.8 minutes, which is the lowest average connection time of all groups), but access to more data bases, search more select terms (10,4) and retrieve less documents (3).

“L” typology (7% of users): **occasional very good users**
They mainly make one session (for this reason they are named “occasional”) that lasts long (25 min.) and use many data bases (CHEMABS 100%, PASCAL 80%, …). They retrieve many documents (163) and have a very high level of expense (137 ECU).

These identified typologies have been labelled with respect to some distinguishing characteristic, anyway CHEMABS and INSPEC users are as good as those labelled “best
users" or “ideal users”, they are only using a lower number of data bases, and the “quick and dirty searchers” are often very well experienced users.

The image maps the groups on two axes that have been identified by a factorial analysis as independent linear combinations of the available numeric variables. The horizontal axis might be called (system) usage as it is a combination of the number of real sessions (begin), the number of data bases and the number of select terms. The vertical axis is the expense axis, as it is a combination of the number of documents retrieved, the average session length and the average expense.

Conclusion

The accounting data provide a valuable source of information on users behaviour, that can be exploited in various ways.

This analysis has been useful to gain insight into the data. In particular, it pointed out a segment of users that, despite being one of the most profitable, show a critical behaviour: the F typology (the “unskilled” users). Retrieving a very high number of documents by means of query that are not focused enough, might generate unsatisfied users. A training course might be useful for these customers.

Future analysis should take into account the real session (identified by the begin command and the payment of the session fee), in this way we can eliminate those very short sessions that are likely to be done by mistake and the typology “A” (unsuccessful explorers) would probably disappear. Other events should also be taken into account (zoom, expand, offline prints, …).

A more in depth analysis could be done on sessions, rather than users, in order to find out the different usage pattern, or strategies, that the same user may adopt in different sessions.
By taking into account a longer period (one year) we should also be able to investigate the behaviour of all EINS users and obtain larger groups that could be the target of direct marketing campaigns by each National Centre.

Beside segmentation, other data mining techniques can help analysing the accounting data to extract valuable information. Association analysis can identify the rules in database usage (for example that 80% of NTIS users, also use the COMPENDEX database) and suggest cross-selling opportunities. Decision trees or neural networks can be applied to build an attrition model for detecting the customers that are likely to defect (churn analysis). Beside the prediction power, it is very useful to understand the characteristics of defectors and non-defectors: if defectors mostly use a specific database, the information content of this database might be unsatisfactory and should be improved.