On the relationships between user profiles and navigation sessions in virtual communities: a data-mining approach

Simone Garatti[1*], Sergio M. Savaresi[1] , Sergio Bittanti[1], Luca La Brocca[2]

[1]*Dipartimento di Elettronica e Informazione, Politecnico di Milano, Piazza L. da Vinci, 32, 20133 Milano, ITALY.*
{sgaratti,savaresi,bittanti}@elet.polimi.it

[2]*European Virtual Community Division, Tiscali S.p.A., Viale Trento, 39, 09123, Cagliari, ITALY*

Abstract: In this paper the analysis and Data-Mining of a large data-set related to a very popular Italian Virtual Community is presented. The Community is constituted by more than half-million registered users, each characterized by a unique nickname and a personal "profile" filled during a registration procedure, on a voluntary basis. Two data-sets have been considered: the Data-Base of the Users (nicknames and profiles), and the log-file of the server hosting the Community web-site. This work is constituted by three main parts: 1) analysis and clustering of the User Data-Base; 2) sessionization of the log-file and clustering of the navigation session database; 3) correlation of User clusters and navigation session clusters. This analysis provides a complete and full-rounded picture of the Virtual Community.

Keywords: Data Mining; Virtual Communities; profiled users; web-log files; sessions; unsupervised clustering; PDDP; heterogeneous data;

* Corresponding Author: Simone Garatti, Dipartimento di Elettronica e Informazione, Politecnico di Milano, Piazza L. da Vinci 32, 20133, Milano, ITALY. Phone: +39.02.2399.3650; Fax: +39.02.2399.3412 ; e-mail: sgaratti@elet.polimi.it.

1. Introduction and problem statement

A Virtual Community is a group of persons interacting and communicating through Internet, that is, all services for data exchange are provided by a web-site, usually called the Community web site. Due to the rapid growth of the number of Internet users, Virtual Communities are now very popular in the world-wide-web (especially for entertainment purposes) and many of them can be joined for free. In these cases, the web site is usually managed by an internet service provider which is rewarded in terms of advertising and by the availability of a huge amount of data from which useful information can be retrieved ([4,8,10]).

This paper deals with the analysis and Data-Mining of the data related to a very popular Italian Virtual Community (the Virtual Community "*People*" - http://people.tiscali.it/ - of *Tiscali S.p.A*, the 2nd biggest European Internet Service Provider). This Virtual Community is very large, as witnessed by the high number of daily accesses to the web site (about 500.000 page views), and the community web site has a complex structure, providing a large variety of web services like forums, chats, message post, programs downloads, on-line games etc.

Even if the access to the web site of the virtual community "*People*" does not require an advance registration, users are strongly encouraged to by simply filling a profile-form (multiple choices among a predefined list of items) and choosing a nickname which has to be unique for each single user. The profile is a sort of virtual ID for the user. After the registration, the users have the opportunity to log-in to the web site, before performing a navigation session, so as to make their personal data available to other users for data-exchanging. The number of registered users was over 550.000 at the time of this work.

The analysis presented herein is based on two (very large) data-sets:

- The Data-Base (DB) of the Users (i.e. the data set collecting all user nicknames and profiles).
- 1-week log-file of the servers hosting the Community web-site (i.e. a text file where all accesses to the web site are registered).

Needless to say, these two data-sets are extremely different: they deliver complementary pieces of information, and they must be processed and analyzed through completely different techniques.

The main goal of this work (which is also its main original contribution, from a methodological point of view) is to establish the relationships between these two heterogeneous data-sets. This is inherently a challenging task, and – to the best of our knowledge – this is one of the first attempts documented in the Data-Mining literature to "merge" and to find relationships between user profiles and web-navigation behaviors of a very large Virtual Community ([14,15]).

The search for relationships between half a million Users and millions of page-views cannot be faced directly, starting from the raw data. The basic idea and methodological approach proposed in this work are as follows:

- The User DB is analyzed and clustered into a small number (12) of clusters; each cluster represents a "prototype" of User (Section 2);
- The log-file of the web server is first sessionized and then analyzed and clustered (using an unsupervised bisecting divisive clustering approach) into 8 clusters; each cluster represents a "navigation behavior"

(Section 3);

- Thanks to the huge dimensional reduction of the two data-sets (the User DB has been reduced to 12 items; the 1-week log-file has been reduced to 8 items), it is possible to find the association map between User clusters and navigation session cluster (Section 4). Note that this can be done since more than 10% of the page-views registered in the log-file contain the nickname of the User, stored in a *cookie*. This allows the linking between the User DB and the log-file.

This not-trivial Data-Mining analysis – although preliminary – provides a very general and full-rounded picture of the Virtual Community Users.

2. Analysis and clustering of the User DB

Fig. 1 displays a filled profile-form, as it is shown in the Community web site.

The profile is made of 14 fields: the first is the nickname, i.e. a personal ID characterizing uniquely each single user, while the other 13 fields specify, respectively, the age, the gender, the spoken language, the job, the country, the zodiac sign, the favorite place to live, the favorite music, the favorite hero, the best loved object, the sexual orientation of the user, the user plans and wishes in the case he was a millionaire, and the choice of the person to stay with on a desert island. As it appears, the User profile is mainly "entertainment oriented" although there are also some socio-demographical fields like Age, Gender, Location and Job.

The User Data-Base is the collection of all the user profiles and it has a simple structure. In fact, the bulk of the User DB is condensed into a single table, where each row represents a user, while each column corresponds to a profile field. In Table 1 a small portion of the User DB is displayed (the field "Zodiac sign" has not been considered since no relevant information was expected within this field).

Note that when completing the profile-form, the user can select among a finite (well-defined) set of items for each field, and only the numeric code of the item selected is stored in the data-base. As an example, see Table 2 where the numeric codes for the items of the field "Age" are displayed.

Note also that, according to the Tiscali policy of minimizing user restrictions, the profile can be left completely blank or can be only partially filled. Therefore, for each field, there is always an item meaning that the considered field has been left undefined. By convention, this item is encoded by "1" for all the fields in the User Data-Base.

Let us observe, finally, that, due to the finiteness of items for each field, data are *categorical* (each field item represents a category, not a quantity; e.g. *job*=2 means that the user is a student) and *non-ordinal* (in general, no ordinal relations can be established among users, and neither among items of a single field; for example saying that a student is greater than an engineer is in fact meaningless).

Before performing the clustering of the User DB so as to obtain a small number of representative user profile prototypes, a preliminary statistical analysis has been performed with the twofold purpose of revealing some per-se interesting features of the user community, and identifying guidelines and key-points for the following clustering procedure. Next subsections 2.1 and 2.2 discuss the main results of this statistical analysis, while

subsection 2.3 is devoted to the clustering of the User DB.

2.1 Preliminary analysis

According to the indications expressed by the Management of the *Tiscali Virtual Community Division*, the preliminary analysis of the User Data-Base has been performed by focusing on:

- The item selected by a User within each field.
- The willingness of a User to fill a specific field during the registration procedure. This is a very interesting piece of information since the profiling is made on a voluntary basis.

As a first step, the entire data-set of the User DB has been transformed into a real-valued matrix M, of size 550.000×12. The element M_{ij} of M represents the item of the *j*-th field selected by the *i*-th User.

We then proceeded as follows:

- The amount of Users having the whole profile undefined (each field equal to "1") has been computed.
- For each field and among Users with at least a field defined, the amount of Users having "1" in the considered field has been computed.

The results are displayed in Fig.2 and in Fig.3, respectively.

It is interesting to observe that:

- A small number of Users (less than 15%) leave the profile completely blank. This reveals that only a small part of subscribers are pure "visitors" without any interest in establishing relationships with other Users.
- Gender is – by far – the most "filled" field. This confirms that this kind of Virtual Community is mainly seen as a means for meeting (dating…) people. Note also that geographical region is quite voted.
- Age, language, sexual orientation, and "alone with on a desert island" are the less voted fields.

2.2 Dependences Analysis

The second step of the analysis of the matrix M has been the search for hidden relationships (also known as *Association Rules* in Data-Mining – [15]) between the 12 fields of the profile. This analysis is very useful to reduce the number of relevant variables for data description, and it provides a new piece of information on user preferences as well.

In the present situation, it has not been possible to resort to the standard linear correlation analysis between pairs of columns, since this analysis works for quantitative data only. In fact, it is not possible to compute the mean value of *categorical* and *non-ordinal* data (which is the mean value between a student and an engineer?) and the data correlation and the data covariance are not defined as well.

In this context, as suggested in [15], the search for the association rules can be done by computing a "dependence" index based on the so called *average mutual information.* For other possible solutions – as e.g. the so-called "Correspondance Analysis" – see e.g. [1].

The *average mutual information index* $I(h,k)$ can be computed for each couple of fields h and k, an it

measures the dependence degree between h and k. $I(h,k)$ is defined as follows (see [15]):

$$I(h,k) = \sum_{ij} \ln\left(\frac{p(h=i,k=j)}{p(h=i)p(k=j)}\right) \cdot p(h=i,k=j),$$

where i and j take all the possible item values for the fields h and k, respectively, and $p(h=i,k=j)$ is the empirical probability of the event $h=i$ and $k=j$ (analogous definitions hold for $p(h=i)$ and $p(k=j)$). Probabilities $p(h=i,k=j)$, $p(h=i)$ and $p(k=j)$ for all possible i, j have been computed by a sample statistics based on an exhaustive search in the *h-th* and *k-th* column of the data set.

In our analysis we used a normalized version of $I(h,k)$, i.e. the dependence index

$$\Gamma(h \mid k) = \frac{I(h,k)}{I(h,h)}.$$

$\Gamma(h \mid k)$ returns values in $[0,1]$ and similarly to $I(h,k)$ it measures the dependence between h and k. More precisely, $\Gamma(h \mid k)$ measures the information level on h which can be obtained from the knowledge of k. For example, if h and k are independent random variables, then $\Gamma(h \mid k) = 0$ since the knowledge of k gives no information on h; on the contrary, if $k = h$, then $\Gamma(h \mid k) = 1$, since k completely describes h. Note that $\Gamma(h \mid k)$ is not symmetric in general ($\Gamma(h \mid k) \neq \Gamma(k \mid h)$) since the information on h given by k may be different from the information on k given by h. As a matter of fact, suppose that h could be equal to 1 or 2 while k could be equal to 1 or 2 if $h = 1$ or to 3 and 4 if $h = 2$. In this case, once k is known, the value of h can be exactly predicted, while the knowledge of h leaves an uncertainty on the possible value of k. It is perhaps worth to note that, assuming a uniform distribution for h and k in this example, we obtain that $\Gamma(h \mid k) = 1/2$ while $\Gamma(k \mid h) = 1$.

The dependence index $\Gamma(h \mid k)$ has been computed for the data set of user profiles. However, we have not considered the entire M but a subset of it, say $\tilde{M}$, which has been obtained by removing the users with all fields left undefined. In fact these users introduce a strong dependence relation between the columns of M and such dependence "hides" the association rules in the relation to with other users.

The results of this field-dependence analysis are condensed in Fig.4, where $\Gamma(h \mid k)$ is plotted as follows:

- Each cell has been coloured proportionally to the value of $\Gamma(h \mid k)$, where h is the field on the row and k is the field on the column; the darker the cell is, the closer $\Gamma(h \mid k)$ is to 1 (hence a dark cell means strong dependence).
- The values on the diagonal (all equal to 1 by definition) have been set to 0, in order to enhance the colour contrast of the plot.
- The field “Language” has not been plotted since no interesting results have been found on this field. This is not surprising: “Language” is usually left blank and, when filled, it is usually set to “Italian” (this can be verified by inspecting the data set), since the Virtual Community is an Italian Community.

The dependences plotted in Fig.4 reveal many interesting things. Among others:

- Age depends on most of the other fields (e.g. the age of the users can be easily predicted from the choice of the "Personal Hero"), but, at the same time, many fields can be predicted from the age of the user. This is somehow expected and suggests that the age is a good field for clustering. Note that the strongest dependence is between age and job.
- Gender is strongly dependent on fields like "Job", "Personal Hero", "What I like" and "Alone With".
- "Region" seems to be independent of other fields, i.e. it represents a piece of information which cannot be predicted from other fields.
- "Sexual Orientation" and "Alone With on a desert island" also show strong correlations with many other fields.

The map in Fig.4 delivers many interesting association rules between the 12 considered fields. It represents, per-se, an interesting result.

2.3 Clustering of the User DB

The third step of the User DB analysis has been the partition of the data set into a limited number of clusters, each one collecting users with similar features. In other words, each cluster represents a sort of user prototype, so that the partition gives a simple but relevant picture of the typology of users registered to the community.

According to the marketing goals of the web site provider and to the results of the previous analysis (see subsections 2.1 and 2.2), the clustering has been done using 3 fields only: gender, age, and geographic region. These three fields get the best trade-off between the following requirements:

- They are relevant from a socio-demographic point of view.
- They are largely filled by the users.
- They are informative enough so as to contain the main features of the user DB.

Also because the data were categorical and non-ordinal data, the clustering has been performed through a hierarchy of univariate decisions on the chosen representative fields (see [15]). In other words, the clustering has been performed by building a *classification tree* ([9,15]) such that each internal node specifies a subset-membership test on a singular field, between gender, age and region. Each row-vector **u** in M (i.e. each user) descends a unique path from the root node to a leaf node depending on how the values of the fields of **u** match the tests at the internal nodes. The set of users reaching the same leaf node is a cluster of the data-set and the characteristics of each cluster are determined by the path connecting the root-node of the classification tree to the corresponding leaf node.

The choice of the nodes and paths in the classification three has been guided by the indications (and requirements) given by Management of the *Tiscali Virtual Community Division.* The objective was to obtain clusters which were suitable for marketing purposes of Tiscali. In Fig.5 the classification tree is shown while Fig.6 displays the size and the characteristics of the 12 clusters obtained by applying the clustering algorithm to M . This partition is a simple socio-demographic partition, actionable for target marketing.

3. Analysis and clustering of a 1-week web-log file

The second data set analyzed in this work has been the log-file of the servers hosting the Virtual Community web-site, collected during a week (from 00:00 of Monday, to 24:00 of Sunday) of January 2002. The log-file analyzed in this work was a standard log-file delivered by an Apache 1.3 web server ([2]) and it was quite huge (about 2.7 GBytes). In Fig.7 a small sample of this file is shown.

Each item (row) of the log-file represents a single "page-view" of a navigation session performed by a single user and it contains, among others, the following data (see Fig.7):

- IP address of the remote host (User) retrieving the web page.
- Complete time-stamp.
- URL of the web page requested by the remote host.
- A *cookie*, containing the nickname of the user. Note that the cookie is stored in the log-file only if a registered user is logged-in (only a small subset of functions of the web-site is restricted to logged-in registered users).

The main objective of this second part of the work was to establish (by means of a clustering procedure) which kind of navigation sessions the users of the community are used to perform while visiting to the community web site. However, retrieving this information from the log-file has required some non-trivial pre-processing due to the structure of the log-file. In fact, a log-file is a raw text-file which stores, along with the IP address, the time stamp, the visited URL and the nickname of the user, many other additional data which were not relevant for our purposes. It may also happen that, due to time-delays and routing problems in the web, the web-server receives multiple requests for the same page by a single user; as a consequence the same row can be replicated many times in the log-file. Moreover, since the log-file just collects each single access to the web site a sessionization (reconstruction of the navigation sessions from the basic data on the page-views) is usually required.

The treatment and the sessionization of the log-file are discussed in the next subsection 3.1 and 3.2, while the clustering is presented in subsection 3.3.

The analysis of navigation patterns is a well known and studied issue in data-mining literature. See e.g. [4,8,10] for some interesting approaches.

3.1 Preliminary analysis

First of all, all unnecessary data have been removed from the log-file through a direct inspection of it. Then the log-file has been stored into a single table of a Data-Base. Each record of the table is a "page-view" registered by the web-server (see Table 3). The table has 10 fields: 4 fields are used for the IP address; 4 fields are used for the complete time-stamp; 1 field for the URL of the requested web-page; 1 field for the nickname (if present). Once the data have been stored in a data-base structure, the page-view redundancy has been easily removed through a proper SQL query on the data-table.

As already stated, the structure of the community web-site is very complicated (there are a lot of nested sub-

pages and links to other domains). As a result the total amount of different URLs registered by the web-server is over 60.000. Thus, with the purpose to ease the management and the interpretation of the data set, all the different visited URLs have been manually grouped into 30 sets, each one corresponding to a thematic area or a service provided by the web site (this URLs classification partially corresponds to the directories tree of the web site), and the "URL" column in the data-base has been replaced by a column of numeric codes representing the set to which each URL belong. The base URLs and the coding of the 30 sets of URLs are displayed in Table 3.

Using the obtained single-table Data-Base, a preliminary statistical analysis has also been done:

- The sample distribution of page-views distinguished in pages with known nickname (logged-in registered Users) and pages with unknown nickname (non-logged Users) has been computed.
- The sample distribution of the page views on the 30 sets of URLs has been computed.

These results are displayed in Fig. 8. Some basic but fundamental observations can be drawn:

- Only the 12% of the page views is made by logged-in users. This is somehow expected since most of the web-site can be seen by not-logged users (only file uploads, profile changes, and message posting on the forum strictly require to log-in). This is in accordance to the *Tiscali* policy of minimizing restrictions on the user navigations.
- The distribution of the page-views on the 30 sets of URLs is much skewed: most of the page views are condensed into 8 sets of URLs. It is interesting to note that the pages related to *messenger* and *chat* are very popular. Another peculiar thing which is worth to be noted is that a large number of hits are on "not-Tiscali" pages. This can be explained by the fact that each User can put in his/her profile the link to his/her personal home page (in a fashion similar to the famous www.geocities.com Virtual Community), which often is hosted on not-Tiscali domains. This confirms that personal web-pages are intensely visited during web navigation.

3.2 Sessionization

Starting from the raw data-base extracted from the 1-week log-file, the "sessionization" of the page views has been the following step. Sessionizing a log-file is known to be a tricky and subtle task, which requires some heuristic and a-priori assumptions (see e.g. [3,11,22]).

The rules used in this work to reconstruct sessions from the log-file DB are the following:

- A session is constituted by a set of time-contiguous URLs requested by the same host (namely by the same IP address).
- A timeout of 15 minutes has been used to recognize two different sessions (two URLs requested by the same IP address, separated by more than 15 minutes are assumed to belong to different sessions).

All sessions have been extracted from the log-file by an exhaustive search in the data base (this search has been made possible by sorting the data so as to consecutively place in the DB all the time-contiguous page-views performed by a single user).

From the obtained navigation sessions, a real-valued matrix S of session visiting times has been built. It is a

31×460.000 matrix where each element $S_{i,j}$ represents:

- The number of seconds spent on the the *i*-th URL in the *j*-th session of the week, for *i*=1..30 (corresponding to all the URL sets determined in Section 3.2).
- A numeric code for the nickname of the user performing the *j*-th session (in case of not logged-in user (no nickname available) the code 0 has been used), when *i*=31.

In Table 4 a sample of the matrix *S* is displayed.

Note that the visiting time-length for each page-view can be computed considering the difference between the complete time-stamps of two subsequent page-views in the same session. The only exception is the last visited page to which a nominal visiting time of 30 seconds has been assigned. This is the main "a priori" assumption we used for sessions reconstruction, and, as introduced at the beginning of this subsection, the reconstruction of the matrix *S* partially depends on it.

The main problem with this assumption is related to the fact that the arbitrarily chosen visiting time for the last page-view could be too small in some cases, leading to an evaluation error of the importance of the corresponding URL.

However, it should be noted that this error is avoided every time a row *j* has an element $S_{i,j}$, for some *i,* equal to 30. This clearly means that, with high probability, the corresponding URL was the last one in the corresponding session, and is even more true for sessions with visiting time 0 for all columns except one with visiting time equal to 30. This situation occurs when a user accesses to the web site and he remains for the whole session on a single page, as, for example, when visiting the pages devoted to chatting or messaging.

It is perhaps worth mentioning that (see the following subsection 3.3) the one-single-page sessions described above are frequent in the session matrix *S*. This confirms that assuming a visiting time of 30 seconds for the last page-view has not been misleading.

3.3 Clustering

The third step of the log-file analysis has been the clustering of the session matrix *S* obtained in Subsection 2.2. Since no indications on which kind of sessions the users were used to performed, we resorted to an unsupervised clustering i.e. the data matrix is partitioned into a number of sub-matrices, without a-priori or external information, but according to the similarity (distance) between data only. The basic rule is that the partition should maximize the similarity among the elements of each sub-matrix (intra-similarity), and minimize the similarity among elements of different sub-matrices (inter-similarity).

Note that, in contrast to the matrix M of user profiles (Section 2), the data stored in *S* are quantitative and ordinal, except for the nickname column, since each $S_{i,j}$ (*i*=1…30) measures the time spent by a user on a web page. As a consequence, if we do not consider the last nickname column of *S* – which can in fact be seen as a label on data allowing the linking with the User-DB – each row of *S* (i.e. each session) is a vector in a Euclidean space with dimensionality equal to 30. The Euclidean metric can be used to measure the distance between two different sessions as well.

In this work, we resort to a bisecting divisive partitioning algorithm. In brief (see e.g. [15] and [16] for a more detailed discussion) these algorithms are first used to split the entire data-set in 2 clusters (maximizing intra-similarity and minimizing inter-similarity of the two obtained clusters), and then, the same bisecting procedure is iteratively applied, each time dividing a single cluster among those obtained in the previous step. Iterations end when the total amount of obtained clusters satisfies a certain stopping criterion.

According to the analysis developed in [18,19,21], the bisection of the clusters has been done using the cascade of the Principal Direction Divisive Partitioning (PDDP) algorithm and the bisecting K-means algorithm. For the sake of self-consistency of this paper, these two algorithms are here briefly recalled in Tables 5 and 6. In both cases, the input is a $N \times p$ matrix S where data are the rows of the matrix, and outputs are two matrices S_L and S_R. Both algorithms are based on the following quantity:

$w = \frac{1}{N}\sum_{i=1}^{N} x_i$, where x_i's are the rows of S.

w is nothing but the average of data samples and is called the centroid of S.5

K-means was first introduced in [17] and it is probably the best known and most widely used clustering technique; hence it is the best representative of the class of iterative centroid-based divisive algorithms ([18,23]). PDDP is a recently proposed technique ([6,7]). It is representative of the non-iterative techniques based upon the Singular Value Decomposition (SVD) of a matrix built from the data-set ([5,7]).

The main difference between K-means and PDDP is that K-means is based upon an **iterative** procedure, which, in general, provides different results for different initializations, whereas PDDP is a **one-shot** algorithm, which provides a unique solution. Is has been proven ([17,20,21]) that the best performance (in terms of quality of partition and of computational effort) can be obtained by applying PDDP, followed by K-means initialized with the centroids of the clusters obtained as a result of PDDP.

After the first bisection, the criterion to select the cluster to be split as well as to decide when halting the iterations of the PDDP+K-means bisecting algorithm has been chosen according to the solution suggested in [20]. This solution is based on the computation of a certain performance index for a given matrix of data (as the initial data set or one of the clusters obtaining after performing PDDP+k-means), measuring quantitatively how much is convenient to split the considered matrix through PDDP+k-means (this index in fact measures the separation degree of the two clusters which would be obtained after bisection – see [20] for details).

After an intensive computation of this index for each cluster at a given step, the cluster to be split is chosen as the one optimizing the performing index, or if the performance improvement is no better than a given level, then no further divisions are necessary so obtaining the final partition of the data set.

The PDDP+K-means algorithm has been applied to the session matrix S, and, then, the same procedure has been also applied to the smaller matrix S_0 ($S_0 \subset S$) constituted by the subset (about 3%) of navigation sessions performed by registered logged-in Users. These sessions (and their partition in clusters) will be used in the next Section, where the correlation analysis between Users and sessions will be discussed.

The obtained results are a 12-cluster partition for S and an 8-cluster partition for S_0. The details on the whole

partition (taxonomy of the all clustering steps) are displayed in Fig. 9 and 10 for S and S_0 respectively. The details of the leaves of the two partitions are given in Fig. 11 for S and in Fig. 12 for S_0; note that the brief characterization given for each cluster has been reconstructed by looking at the centroid (i.e. the average session) of the considered cluster.

By inspecting the clustering results in Figs.9-12 the following remarks are due:

- As expected, the partition made by PDDP+K-means shows the most typical navigation sessions for both S and S_0. Note, in particular, the relevance of messenger-based or chat-based sessions, and the navigations spent onto not-Tiscali domains.
- The most voted session clusters are mainly focused on a precise service provided by the community web-site, like chat, messaging, forums and profile modification. In these cases, the session is mainly made by a unique visited page (the page offering the service), as can be clearly spotted by looking at the centroids of the considered clusters. In fact, in this cases, the centroids components are approximately equal to 30 (the nominal time assigned for the last page view in a session – see subsection 3.2) for the URL devoted to the web-service and approximately equal to 0 for all the others URLs.
- As a main difference between the partition of S and the partition of S_0, note that in the first case the largest cluster is the "chat session" cluster while in the second case the "messenger session" is the most voted. This is somehow expected since both "chat" and "messenger" provide a similar service but "messenger" is available for logged-in users only.

Finally, it is perhaps worth noting that the percentage of logged-User sessions is 3%, whereas the percentage of logged-User page-views is 12%. This clearly shows that logged registered Users perform navigation sessions which are much longer and more various than the average session of all the users.

4. Establishing relationships between users and sessions

The last step of the analysis presented in this work has been the search of the main relationships between the User DB and the navigation log-file. As already stated in the Introduction, this task cannot be faced directly from the raw data-sets and the basic idea is to pre-process and reduce the User DB to 12 clusters, and the log-file to 8 "prototype" clusters of sessions. The correlation then is searched between clusters. In this way the complexity of the problem is enormously reduced, and the results can be more easily interpreted.

In Fig. 13 the 12 clusters of Users and the 8 clusters of navigation sessions are displayed.

Note that:

- The navigation sessions considered in this analysis are the subset S_0 (about the 3% of sessions, corresponding to the 12% of page-views) of sessions made by logged-in registered Users. These sessions contain the "signature" of the nickname of the User.
- The Users considered in this analysis are the subset of Users (about 3.000) who logged-in in the "*People*" web-site at least once during the analyzed week. Note that the distribution of these 3.000 users among the 12 clusters is remarkably different from the distribution of the whole set of

550.000 Users. This can be easily seen in Fig. 14, where the leaves of the two partitions (% of the total set of Users considered in that partition) are displayed. For example, it is apparent that the 3.000 "active" users registered in the log-file have a much higher willingness to fill the profile during registration.

The correlation analysis between Users and sessions has been performed by building a correlation matrix *C* whose dimensions are 12×8 (i.e. the number of user clusters and the number of session cluster). Each entry *C(i,j)* (i=1..12, j=1..8) represents the average frequency of performing a session in the *j*-th session cluster by a user in the *i*-th user cluster. The matrix *C* has been computed as follows:

- According to his/her profile, each of the 3000 users has been classified into one of the 12 user clusters (hence it has been associated to one row of the matrix *C*); the whole set of sessions made by such a user has been extracted from the sessions-matrix S_0. Then, each session then has been classified into one of the 8 session clusters.
- A row vector of size 8 has been built for each user; the *j*-th component of this vector represents the user sample probability of performing a session in the *j*-th session cluster during the week. The sum of all the components of the row vector is equal to 1.
- For *i*=1..12, all the row vectors computed for the users belonging to the *i*-th cluster have been summed and each component of the obtained vector has been divided by the cardinality of the *i*-th user cluster. The result is a probability vector averaged to all the users in the same *i*-th cluster. This vector represents the *i*-th row of the matrix *C*.

The plot of *C* is in Fig. 15. The colour (darkness) of each cell of Fig. 15 is proportional to the value of *C*(*i*,*j*), where *i* is the User cluster (row) and *j* is the session cluster (column). The coding of *i* and *j* is displayed in Fig.13.

As it appears, all columns are more or less uniformly coloured, revealing that all the users behave similarly on average. In particular they seem to be mainly interested in the sessions focused on "Messenger" (column 4).

On the other hand, it should be noted that each column presents some variations with respect to the nominal (average) colour. However, these variations are not much visible due to the predominance of the messenger session. This is especially true for columns 1,7 and 8 which are less performed by users.

To enhance the differences between user behaviours, a new correlation matrix $\widetilde{C}$ can be computed from *C* by dividing (scaling) each column of *C* by the average value of the column. The plot of $\widetilde{C}$ is in Fig. 16. Again, the colour (darkness) of each cell of Fig. 16 is proportional to the value of $\widetilde{C}(i, j)$, where *i* is the User cluster (row) and *j* is the session cluster (column).

By analysing the results displayed in Fig. 16, many interesting pieces of information can be drawn. For example the map of the main associations between clusters of Users and clusters of sessions can be built. This map is displayed in Fig. 17 and, among other things, it can be noticed that:

- Males seem to be more related to long and various sessions.

– Females seem to be primarily interested to sessions with forum or chat.

– Long sessions focused on forums seem correlated with Users who left the gender blank.

5. Conclusions

In this paper a case study of Data-Mining is presented: two heterogeneous and very large Data-Bases of a Virtual Community have been analyzed and correlated. The approach used for this analysis has been the preliminary pre-processing and independent clustering of the two Data-Bases, and then the correlation of the clusters only. This approach revealed well-suited to manage this kind of data, and a complete and easy to interpret picture of the Virtual Community Users has been built.

The main points of this analysis can be summarized as follows:

- There are substantial differences between active users (logged-in users) and non-active users, regarding both profiles and navigation sessions. In fact, the former are more willing to fill their profile (4% only of completely undefined profiles) and their navigation session are more focused on specific services provided by the community web-site. Non-active users, instead, tends to leave their profiles more undefined (14% of completely undefined profiles) and there is a higher number of navigation sessions focused on the whole web-site exploration and on personal web-site navigation.
- Both for the total amount of users and active users, sessions based on "messenger" and "chat" are the most voted sessions (in both cases clusters "messenger" and "chat" contain the 37-39% of the total amount of performed sessions). This reveals that the community web-site is mostly used as a means to interact with other people.
- Among active users, "messenger" based sessions are the most performed sessions (35% of sessions are in the "messenger" cluster). Yet, males seem more interested in establishing single-user relationships (search of single profile), while females prefer multi-user services like forum and chat.

Acknowledgments

This work has been supported by *Tiscali S.p.A.*, by the MIUR project "*New Methods for Identification and Adaptive Control for Industrial Systems"*, and by CNR-IEIIT. Thanks are also due to Davide Romieri and Paolo Prestinari of *n-Machines s.r.l* for enlightening discussions on Virtual Communities.

References

[1] Andersen E. B. (1990). *The statistical analysis of categorical data*. Springer-Verlag.

[2] Aulds C. (2000). *Linux Apache Web Server Administration*. Sybex.

[3] Berent B., Mobasher B., Spiliopoulou M., and Wiltshire J. (2001). "Measuring the Accuracy of Sessionizers for Web Usage Analysis". *Web Mining Workshop, at 1st SIAM International Conference on Data Mining*.

[4] Berent B., Spiliopoulou M. (2000). "Analysis of navigation behaviour in web sites integrating multiple information systems". *The VLDB Journal*. Vol. 9. pp. 56-75.

[5] Berry, M.W., Z. Drmac, E.R. Jessup (1999). "Matrices, Vector spaces, and Information Retrieval". *SIAM Review*, vol.41, pp.335-362.

[6] Boley, D.L. (1998). "Principal Direction Divisive Partitioning". *Data Mining and Knowledge Discovery*, vol.2, n.4, pp. 325-344.

[7] Boley, D.L., M. Gini, R. Gross, S. Han, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, J. Moore (2000). "Document Categorization and Query Generation on the World Wide Web Using WebACE". *AI Review*, vol.11, pp 365-391.

[8] Borges, J., Levene, M. (1999). "Data mining of user navigation patterns". *Proceedings of the Workshop on Web Usage Analysis and User Profiling (WEBKDD'99)*, San Diego, CA, pp. 31-36

[9] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., (1984). *Classification and Regression trees*. Wadsworth Statistical Press.

[10] Catledgw, L., Pitkow, J. (1995). "Characterizing browsing behaviors on the world wide web". *Computer Networks and ISDN Systems*. Vol. 27.

[11] Cooley, R., Mobshaer, B., Srivastava, J. (1999). "Data preparation for minimg world wide web browsing patterns". *Journal of Knowledge and Information Systems*, vol. 1, n. 1.

[12] Deerwester, S., S. Dumais, G. Furnas, R. Harshman (1990). "Indexing by latent semantic analysis". *J. Amer. Soc. Inform. Sci*, vol.41, pp.41-50.

[13] Golub, G.H, C.F. van Loan (1996). *Matrix Computations (3rd edition)*. The Johns Hopkins University Press.

[14] Hagel J.III, Armstrong A.G. (1999). *Net Gain: Expanding Markets Through Virtual Communities*. Harvard Business School Press.

[15] Hand D., Mannila H., Smyh P. (2001). *Principles of Data Mining*. MIT Press.

[16] Jain, A.K, M.N. Murty, P.J. Flynn (1999). "Data Clustering: a Review". *ACM Computing Surveys*, Vol.31, n.3, pp.264-323.

[17] MacQueen, J., (1967) "Some methods for classification and analysis of multivariate observations". In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* L.M. Le Cam, J. Neyman (eds.) University of California Press, Berkeley, pp. 281-297.

[18] Selim, S.Z., M.A. Ismail (1984). "K-means-type algorithms: a generalized convergence theorem and characterization of local optimality". *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.6, n.1, pp.81-86.

[19] Savaresi S.M., D.L. Boley (2001). "On the performance of bisecting K-means and PDDP". *1st SIAM Conference on Data Mining*, Chicago, IL, USA, paper n.5, pp.1-14.

[20] Savaresi S.M., D.L. Boley, S. Bittanti, G. Gazzaniga (2002). "Cluster selection in divisive clustering algorithms". *2nd SIAM International Conference on Data Mining*, Arlington, VI, USA, pp.299-314.

[21] Savaresi, S.M., D.L. Boley (2003). "A Comparative Analysis on the Bisecting K-Means and the PDDP Clustering Algorithms". *International Journal on Intelligent Data Analysis* (to appear).

[22] Spiliopoulou M. (1999) "The laborious way from data mining to web mining". *Int. Journal of Comp. Sys., Sci. & Eng.,* Vol. 14 pp. 113-126

[23] Steinbach, M., G. Karipis, V. Kumar (2000). "A comparison of Document Clustering Techniques". *Proceedings of World Text Mining Conference, KDD2000*, Boston.

Figure captions

Nickname	Age	Gender	Language	Job	Wherelive	Music	Hero	Whatlike	Alone	Ifmillionaire	Orientation	Region
Topogigio	9	1	1	1	1	1	1	1	1	1	1	10
Silvana01	6	3	23	1	1	6	1	9	10	13	1	10
Big.pig	1	2	1	1	1	1	1	1	1	1	1	1
Biferno	1	2	1	1	1	1	1	1	1	1	1	1
Ginkoman	5	2	23	24	9	1	37	4	24	13	2	15
Volture	4	2	23	2	31	6	8	47	21	9	1	14

Table 1. A sample of records from the User DB..

Field AGE	
Code	**Meaning**
1	Undefined
2	< 14 years old
3	> 14 and < 18 years old
4	> 19 and < 28 years old
5	> 28 and < 36 years old
6	> 36 and < 46 years old
7	> 46 and < 65 years old
8	> 65 and < 85 years old
9	> 85 years old

Table 2. Coding for the field “Age”

IP1	IP2	IP3	IP4	DATA	H	M	S	URL	NICK
62	11	34	85	11/Jan/2002	0	26	20	http://messenger.tiscali.it/tiscali-it/visualizza.php	zeppelin26
62	11	34	85	11/Jan/2002	0	26	13	http://messenger.tiscali.it/tiscali-it/visualizza.php	zeppelin26
62	11	34	77	08/Jan/2002	23	50	20	http://64.4.18.250/cgibin/linkrd?_lang=IT&lah=4	zeppelin26
62	11	34	77	08/Jan/2002	23	50	31	http://people.tiscali.it/client/client.php	zeppelin26
62	11	34	77	08/Jan/2002	23	50	21	http://people.tiscali.it/client/client.php	zeppelin26
62	11	34	77	08/Jan/2002	23	50	13	http://people.tiscali.it/client/client.php	zeppelin26
62	11	34	77	08/Jan/2002	23	50	40	http://people.tiscali.it/client/client.php	zeppelin26
62	11	168	188	11/Jan/2002	21	27	21	http://people.tiscali.it/modifica/index.php?step=2	zeta1962
62	11	168	188	11/Jan/2002	21	27	21	http://people.tiscali.it/modifica/index.php	zeta1962
62	11	168	188	11/Jan/2002	21	27	16	http://people.tiscali.it/modifica/index.php?step=2	zeta1962
62	11	161	73	13/Jan/2002	22	33	37	http://messenger.tiscali.it/minibrowser/menu.htm	zeta1962
62	11	161	73	13/Jan/2002	22	33	40	http://messenger.tiscali.it/minibrowser/menu.htm	zeta1962
62	11	161	73	13/Jan/2002	22	33	32	-	zeta1962
62	11	161	73	13/Jan/2002	22	33	33	http://messenger.tiscali.it/minibrowser/index.html	zeta1962
62	11	161	73	13/Jan/2002	22	33	37	http://messenger.tiscali.it/minibrowser/news.html	zeta1962
62	11	161	73	13/Jan/2002	21	6	12	http://messenger.tiscali.it/	zeta1962
62	11	161	73	13/Jan/2002	21	6	10	-	zeta1962
212	63	99	175	10/Jan/2002	0	58	34	http://chat.tiscali.it/	zevic
212	63	99	175	10/Jan/2002	0	52	31	http://people.tiscali.it/directory/directory_livello2.php	zevic
212	63	99	175	10/Jan/2002	0	48	14	http://messenger.tiscali.it/minibrowser/index.html	zevic
212	63	99	175	10/Jan/2002	0	54	3	http://people.tiscali.it/directory/link.php?url=http://	zevic
212	63	99	175	10/Jan/2002	0	59	40	http://community.tiscali.it/	zevic

ID	Domain
1	http://chat
2	http://community.tiscali.it
3	http://dinavision
4	http://messenger.tiscali.it
5	http://people.tiscali.it
6	http://people.tiscali.it/aminews
7	http://people.tiscali.it/bacheche
8	http://people.tiscali.it/classific
9	http://people.tiscali.it/client
10	http://people.tiscali.it/creami
11	http://people.tiscali.it/directory
12	http://people.tiscali.it/forum
13	http://people.tiscali.it/help
14	http://people.tiscali.it/login
15	http://people.tiscali.it/modifica/
16	http://people.tiscali.it/password
17	http://people.tiscali.it/registrazione
18	http://people.tiscali.it/registrazione/registrazionesito
19	http://people.tiscali.it/ricerca
20	http://people.tiscali.it/sendmsg
21	http://people.tiscali.it/stat
22	http://people.tiscali.it/svd
23	http://people.tiscali.it/tour
24	http://people.tiscali.it/utente/
25	Indefinito
26	Indirizzi Numerici
27	NO Tiscali
28	Siti Tiscali Estero
29	Varie People
30	Varie Tiscali

Table 3. Sample of the DB of the page-views (left); coding of the 30 sets of URLs (right).

URL1	URL2	URL3	URL4	URL5	...	URL28	URL29	URL30	NICK
20	56	0	30	0	...	0	0	66	0
105	0	12	46	30	...	124	0	0	0
234	14	30	0	0	...	0	0	0	0
0	0	30	0	0	...	0	0	0	345670
23	45	66	123	0	...	111	23	0	345670
10	30	40	0	45	...	0	0	45	123456
...	...	...	...	...	...	...	...	...	...

Table 4. Sample of the session matrix *S*

Table 5: PDDP clustering algorithm.

Step 1. Compute the centroid *w* of *S*.

Step 2. Compute the auxiliary matrix $\tilde{S}$ as:

$$\tilde{S} = S - ew,$$

where *e* is a *N*-dimensional vector of ones, namely $e = [1,1,1,1,1,...1]^T$.

Step 3. Compute the Singular Value Decompositions (SVD) of $\tilde{S}$:

$$\tilde{S} = U\Sigma V^T,$$

where Σ is a diagonal $N \times p$ matrix, and *U* and *V* are orthonormal unitary square matrices having dimension $N \times N$ and $p \times p$, respectively (see [12] for an exhaustive description of SVD).

Step 4. Take the first column vector of *V*, say $v = V_1$, and divide $S = [x_1, x_2, \ldots, x_N]^T$ into two sub-clusters S_L and S_R, according to the following rule:

$$\begin{cases} x_i \in S_L & if \quad v^T(x_i - w) \le 0 \\ x_i \in S_R & if \quad v^T(x_i - w) > 0 \end{cases}$$

Table 6: Bisecting K-means algorithm.

Step 1. (Initialization). Select two points in the data domain space, say $c_L, c_R \in \Re^p$.

Step 2. Divide $S = [x_1, x_2, \ldots, x_N]^T$ into two sub-clusters S_L and S_R, according to the following rule:

$$\begin{cases} x_i \in S_L & if \quad \|x_i - c_L\| \leq \|x_i - c_R\| \\ x_i \in S_R & if \quad \|x_i - c_L\| > \|x_i - c_R\| \end{cases}$$

Step 3. Compute the centroids of S_L and S_R, w_L and w_R.

Step 4. If $w_L = c_L$ and $w_R = c_R$, stop. Otherwise, let $c_L := w_L$, $c_R := w_R$ and go back to Step 2.

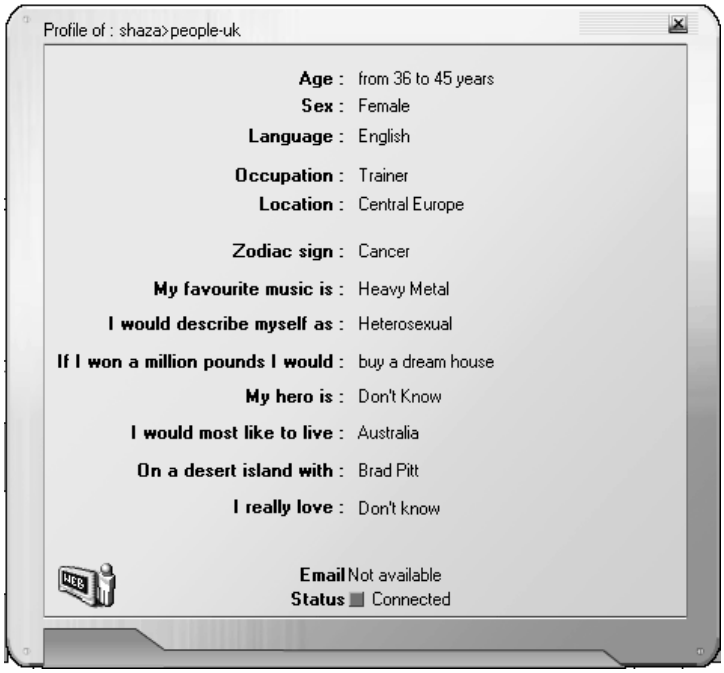

Fig.1. Example of the profile of a User.

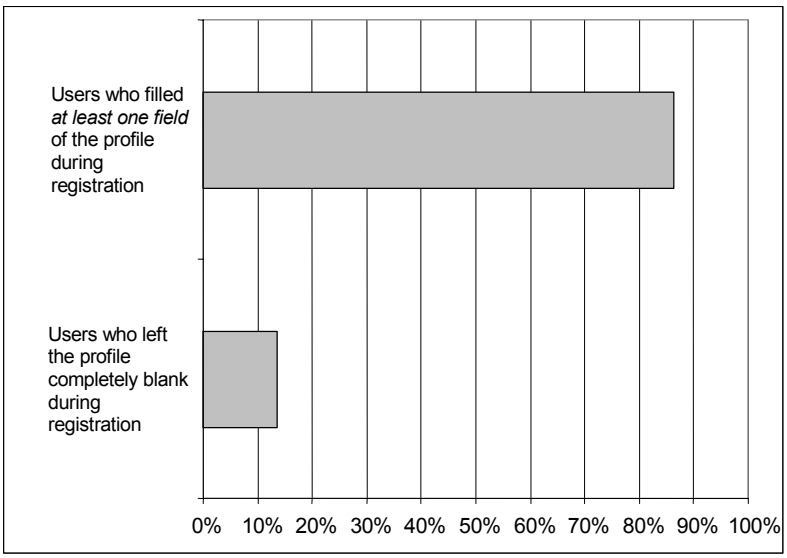

Fig.2. Blank vs. filled profiles.

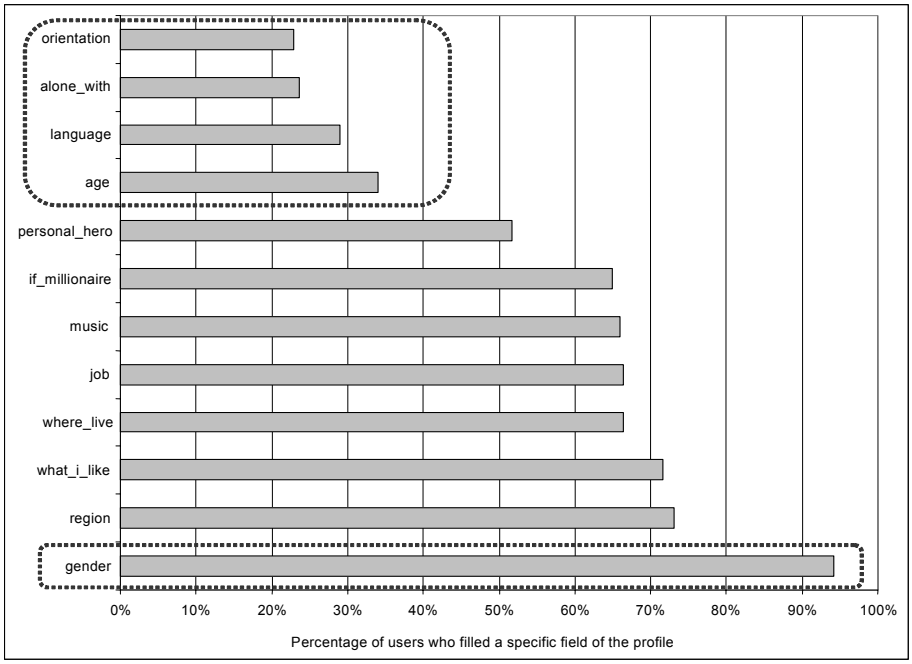

Fig.3. Willingness of the Users to fill a specific field of the profile

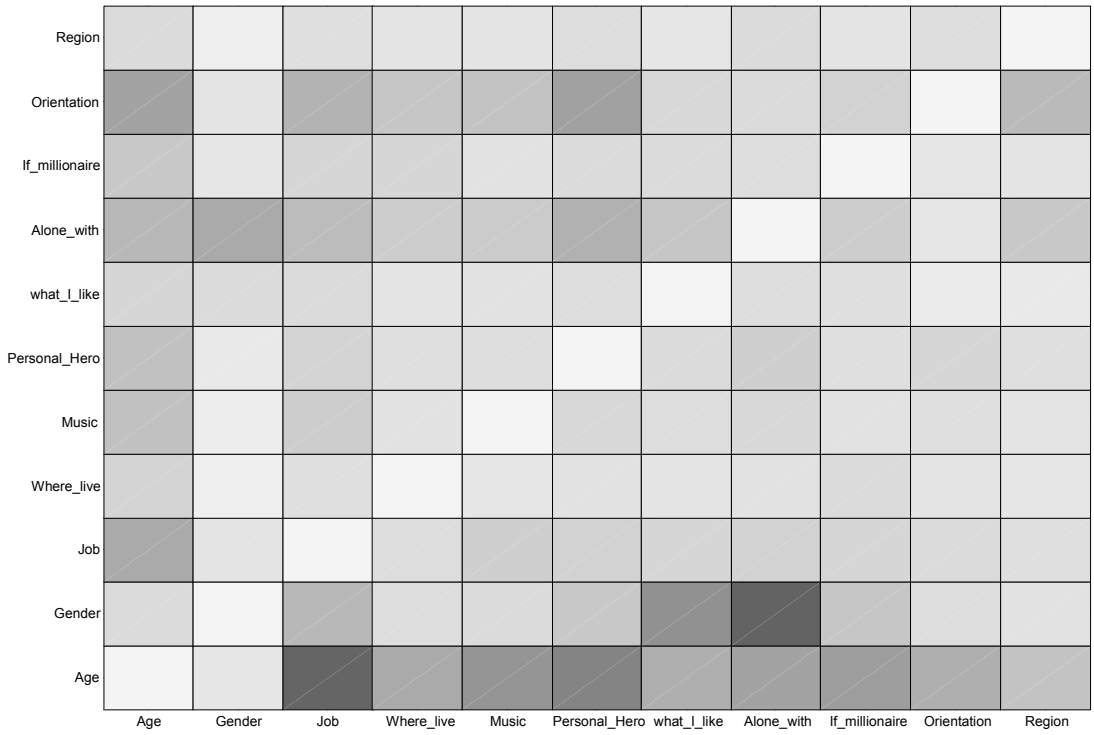

Fig.4. Association rules between the fields of the profile (dark = strong correlation).

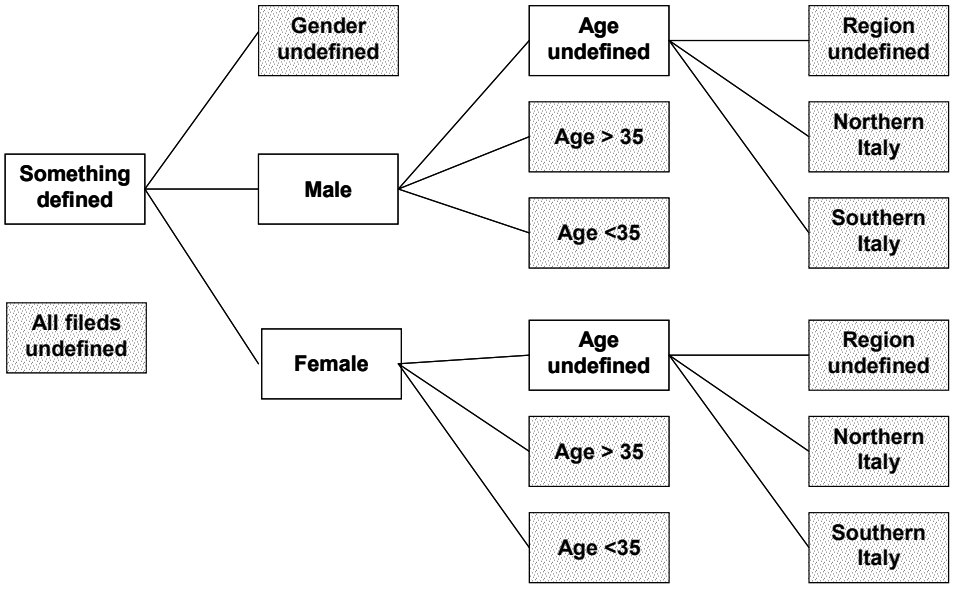

Fig.5. The classification tree (grey cells = leaf nodes)

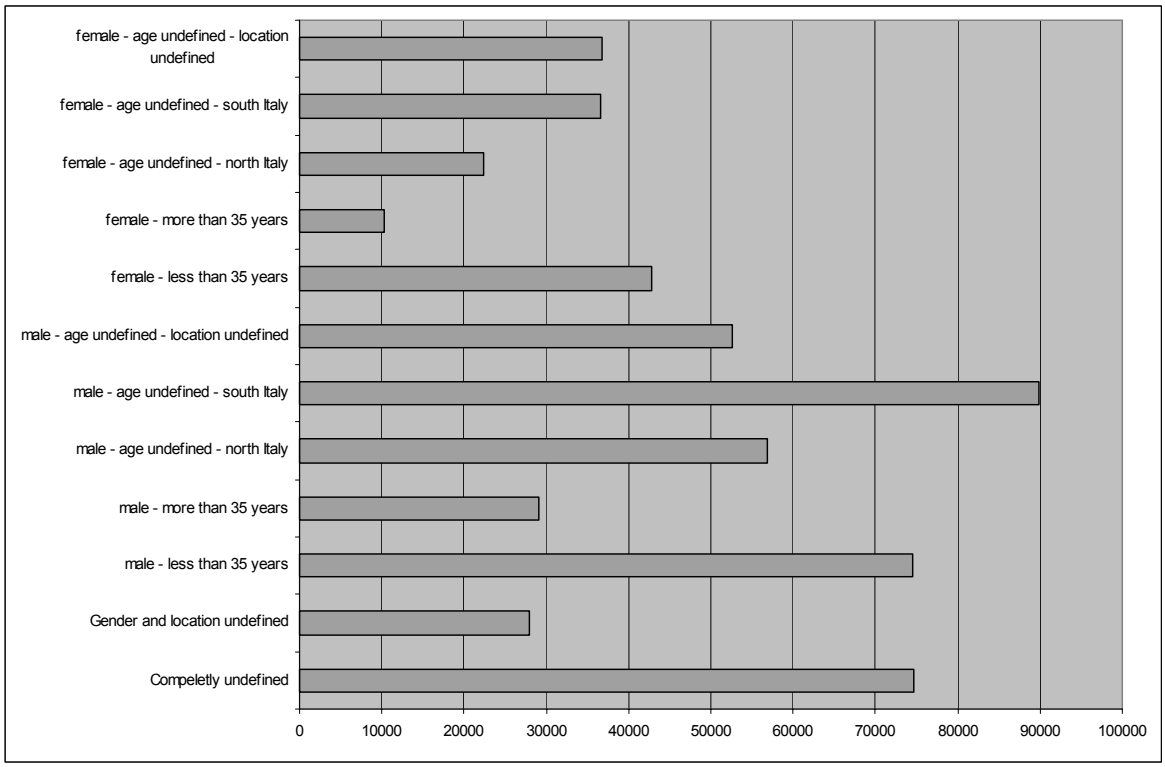

Fig.6. Partition of the whole User DB (550.000 registered Users) into 12 clusters.

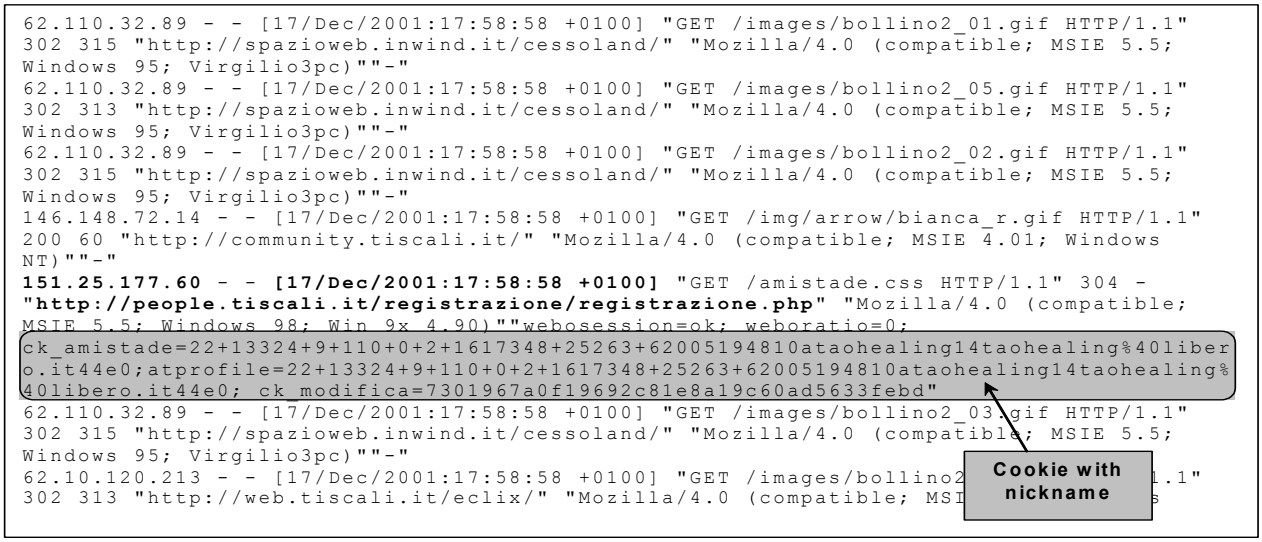

Fig.7. Sample of the raw log-file delivered by an Apache web server.

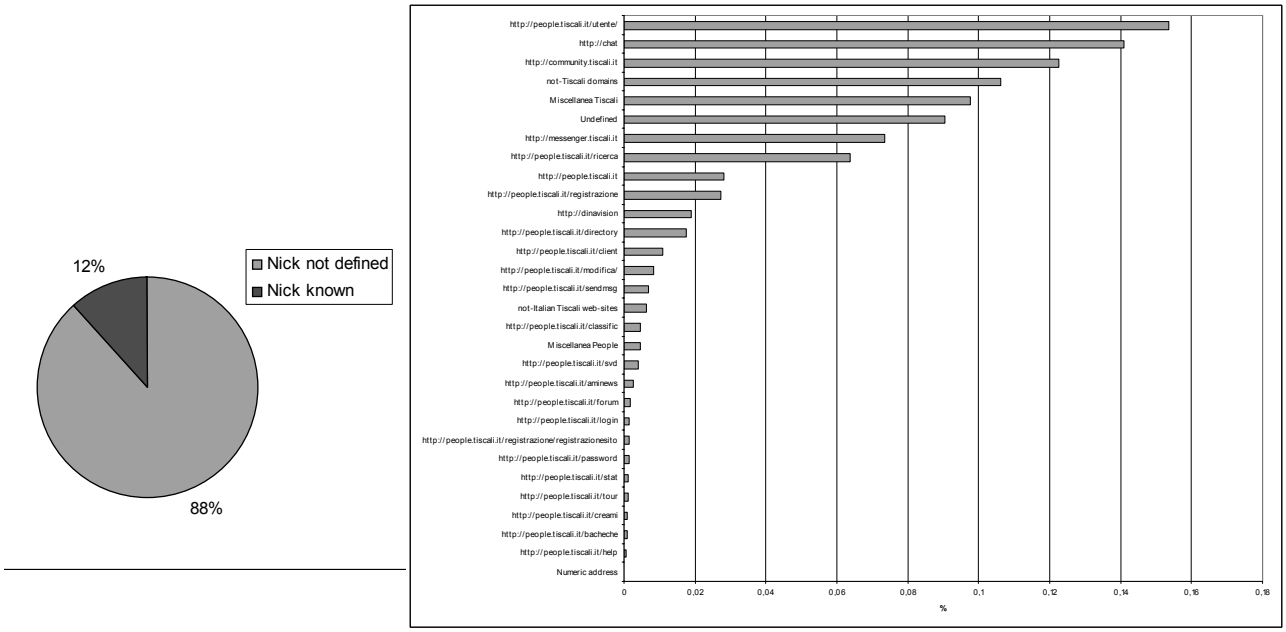

Fig.8. Distribution of page views of logged and not-logged Users (left); sample distribution of page-views in the 30 classes of web-pages (right).

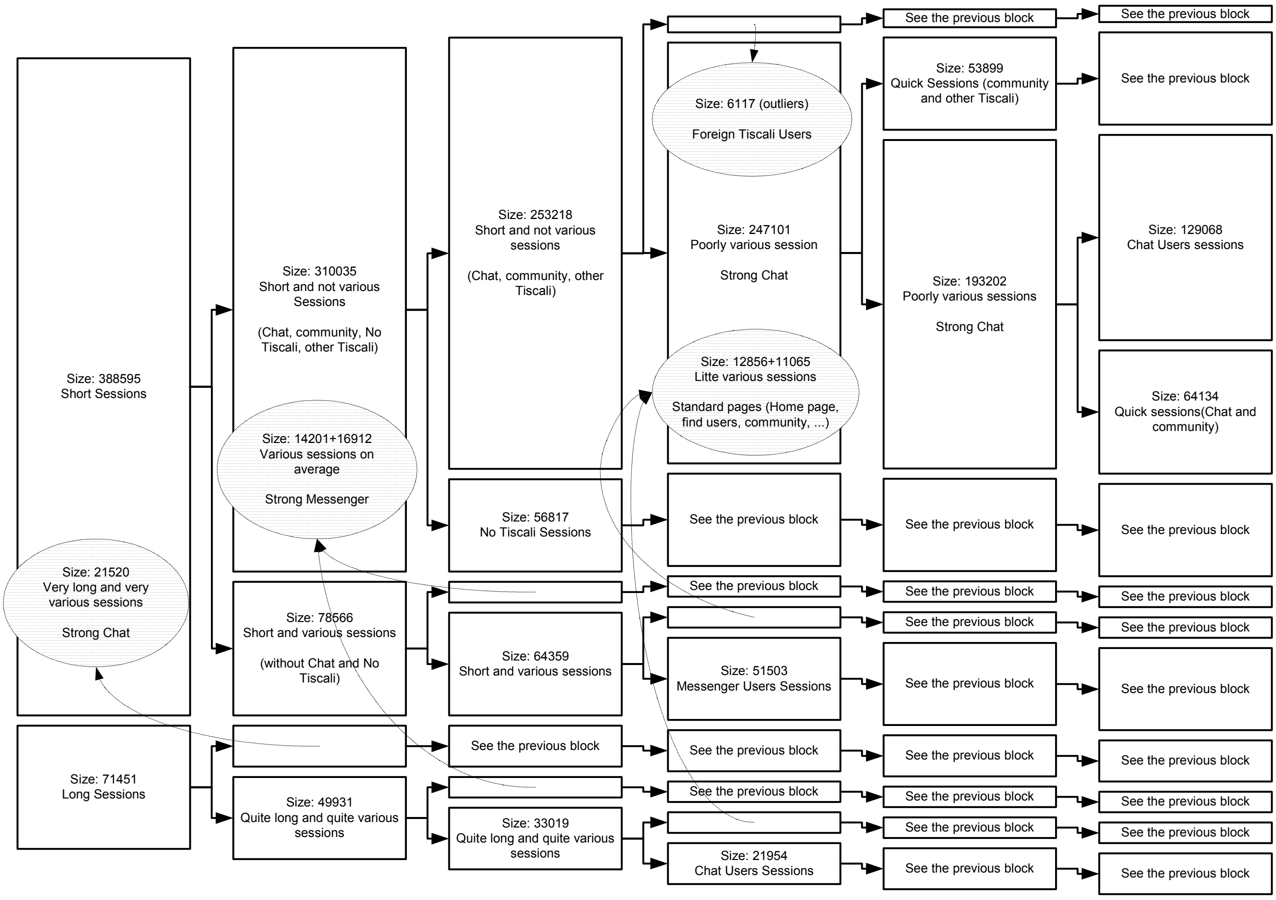

Fig.9. Complete partition-tree of the session matrix *S*.

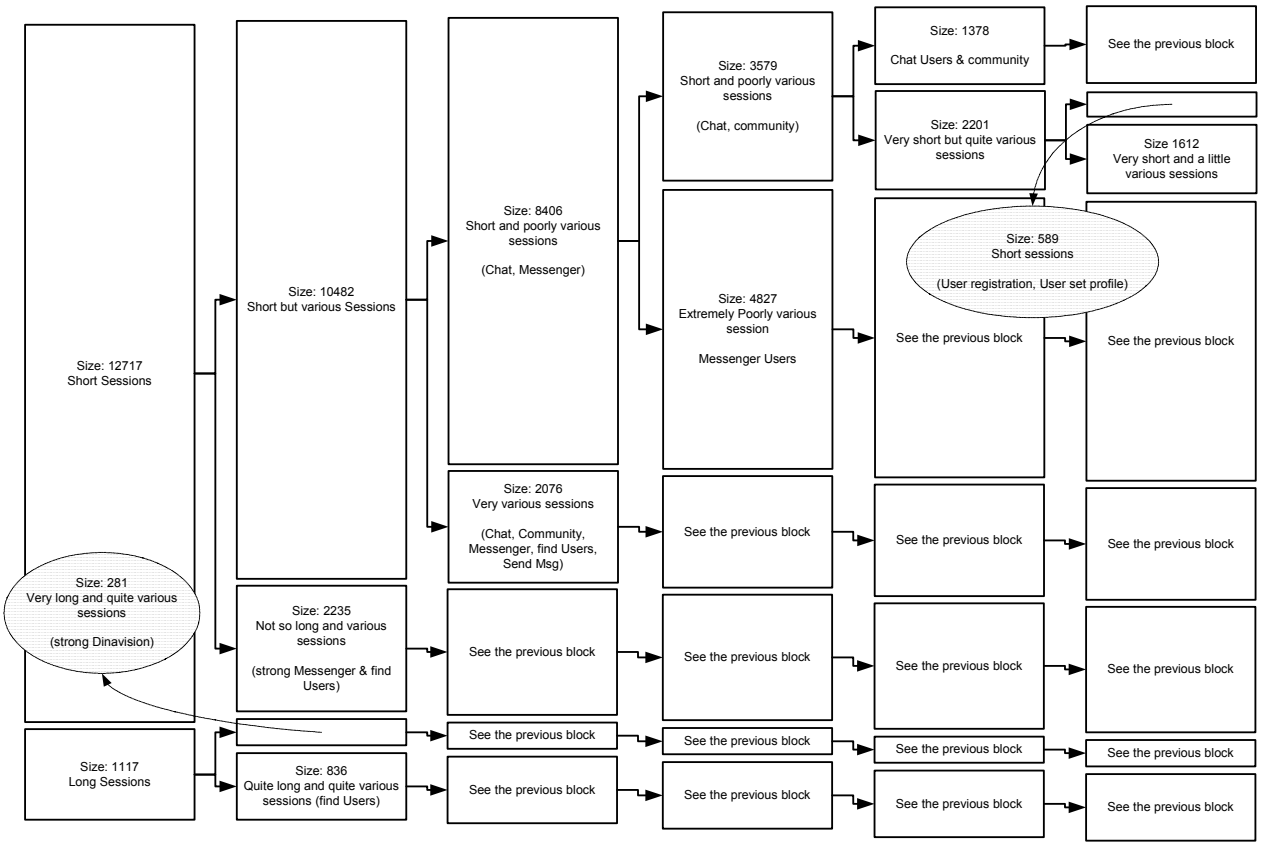

Fig.10. Complete partition-tree of the session matrix S_0 (sessions made by logged-in Users).

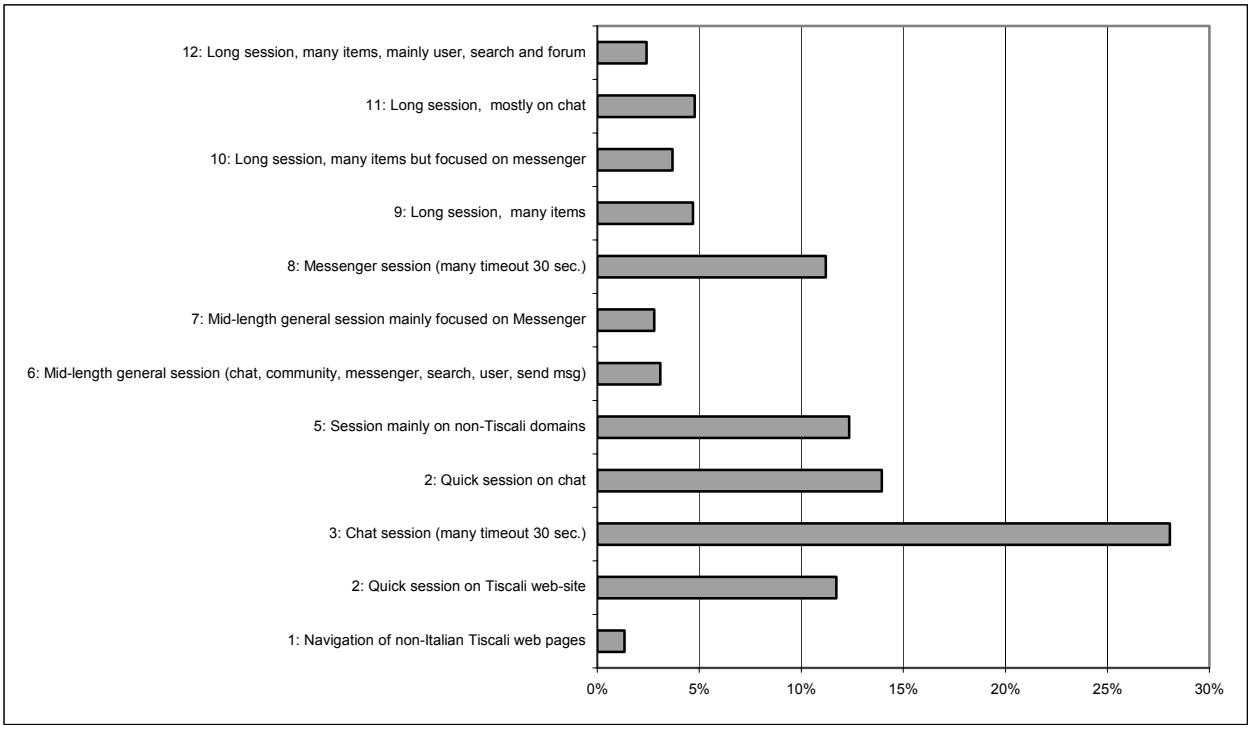

Fig.11. Leaves of the partition of the session matrix *S*.

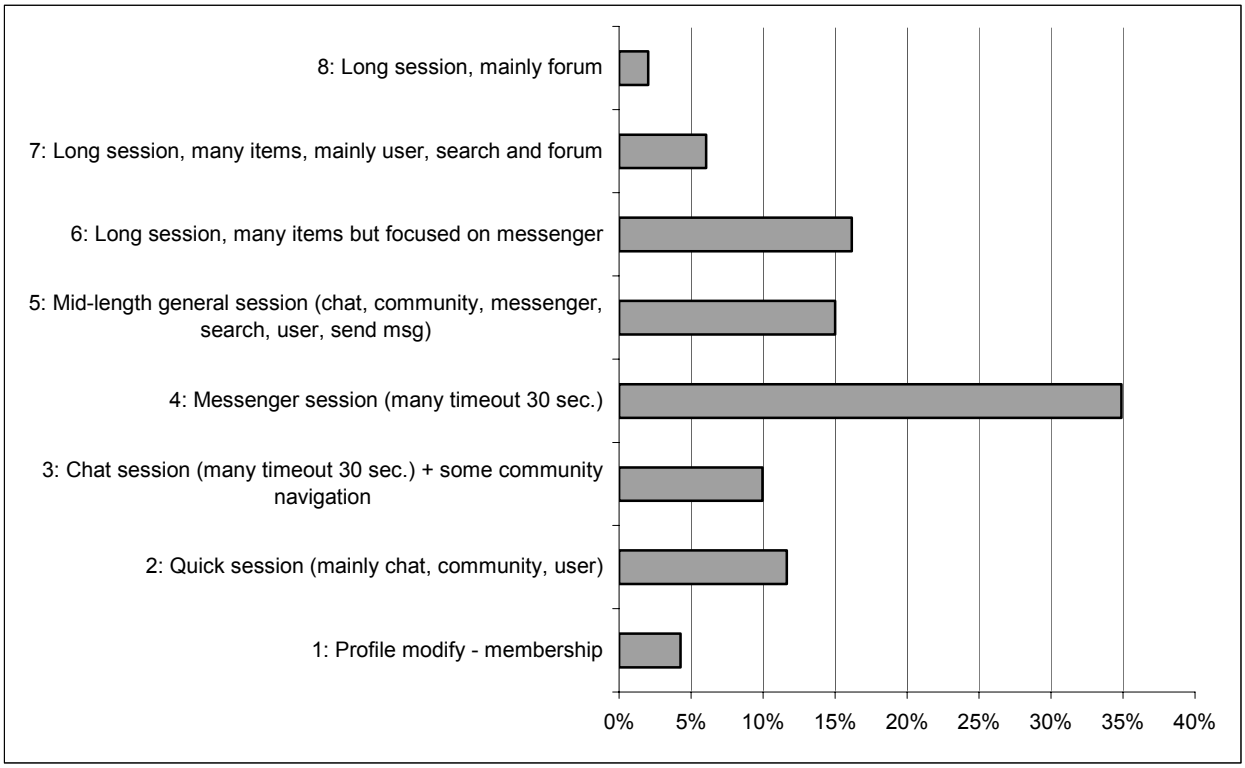

Fig.12.Leaves of the partition of the session matrix S_0 (sessions made by logged-in Users).

Partition of the web sessions DB (8 Clusters)
1: profile modified - membership
2: quick session (mainly chat, community, user)
3: chat session (many timeout 30 sec)
4: messenger session (many timeout 30 sec.)
5: mid-length general session (chat, messenger, search user)
6: long session, many items, focused on messenger
7: Long session, many items, mainly user, search and forum
8: Long session, mainly forum

Partition of the User DB (12 Clusters)
1: all undefined
2: male, age >35
3: male, age < 35
4: male, age undefined, north Italy
5: male, age undefined, south Italy
6: male, age and region undefined
7: female, age > 35
8: female, age < 35
9: female, age undefined, north Italy
10: female, age undefined, south Italy
11: female, age and region undefined
12: gender undefined

Fig.13. Clusters of Users vs. clusters of navigation sessions.

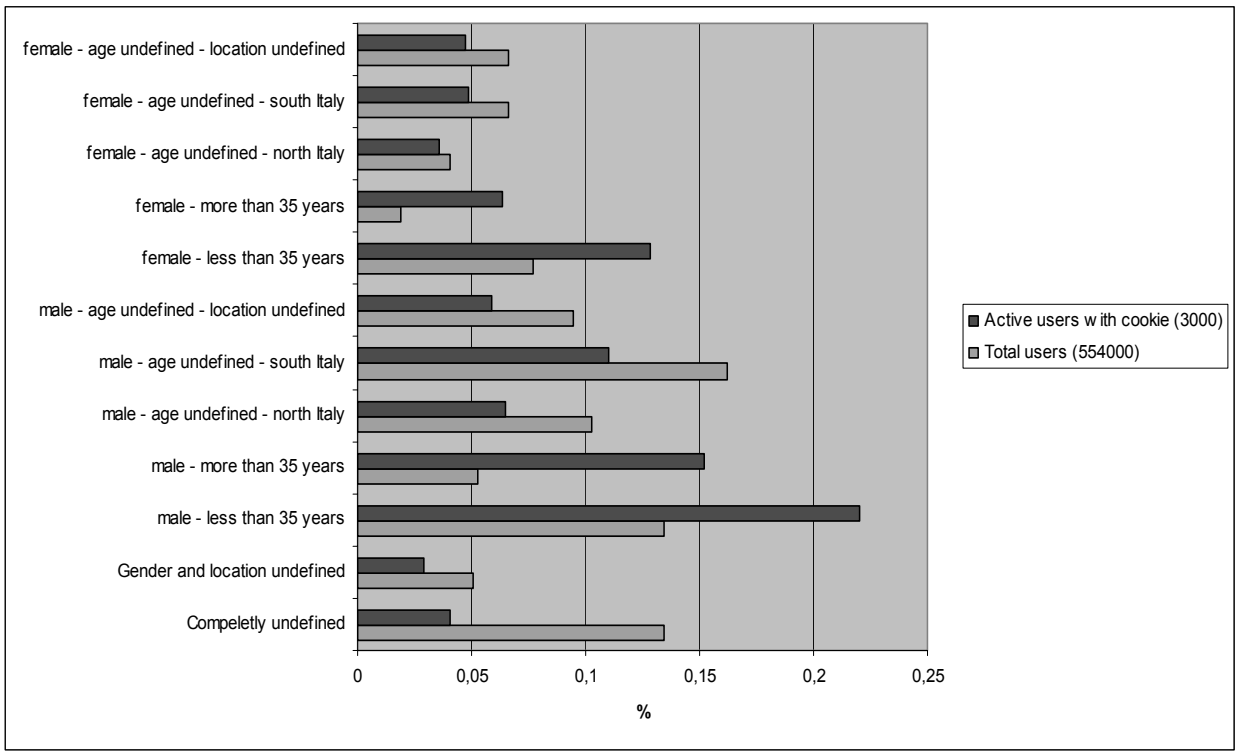

Fig.14. Clusters of Users: all Users vs. active Users in the week considered in this analysis.

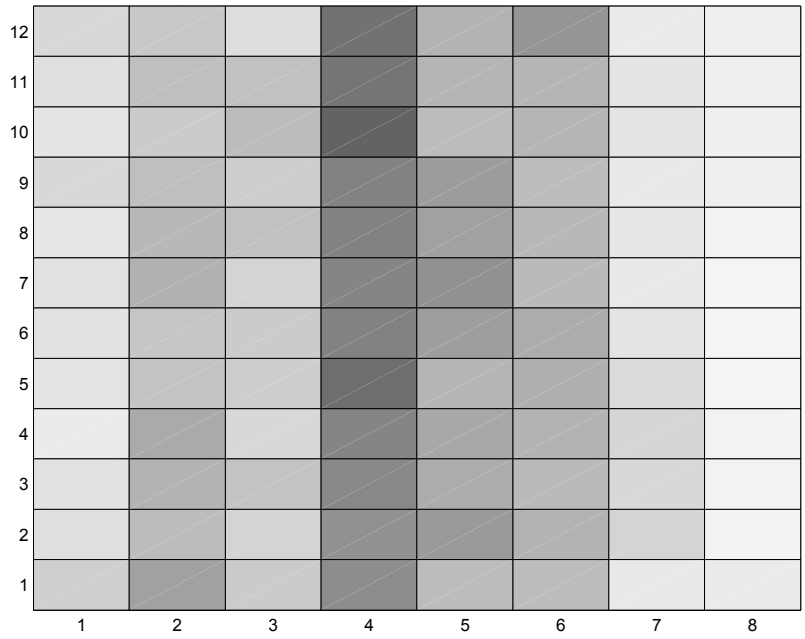

Fig.15. Correlation between the 12 clusters of Users and the 8 clusters of sessions.

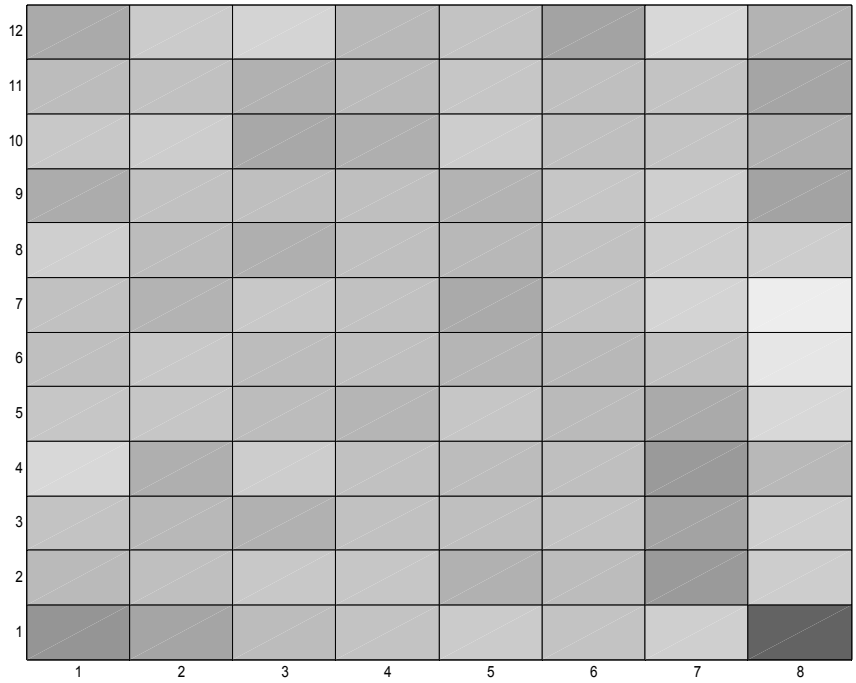

Fig.16. Correlation between the 12 clusters of Users and the 8 clusters of sessions (dark=strong correlation).

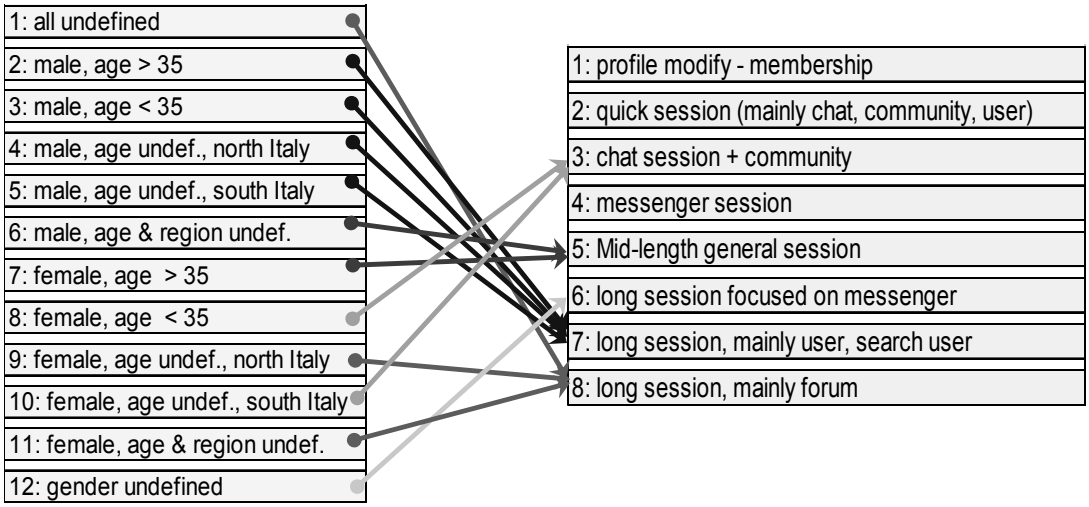

Fig.17. Main association rules between Users and sessions.