MODELING THE RELATIONSHIPS BETWEEN THE USERS DB AND THE WEB-LOG FILE OF A LARGE VIRTUAL COMMUNITY

Sergio M. Savaresi, Simone Garatti, Sergio Bittanti
Dipartimento di Elettronica e Informazione, Politecnico di Milano, Piazza L. da Vinci, 32, 20133 Milano, ITALY.

Abstract: In this paper the analysis and modeling 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, characterized by a unique nickname. Each user has its own "profile", which is filled during the registration procedure, on a voluntary basis. Two data-sets are used: the database of the users (nickname and profile), and the database of their web navigation sessions. The latter has been obtained from the log-file of the servers hosting the community web-site. This work is constituted by three main parts: 1) analysis and clustering of the users DB; 2) analysis and clustering of the navigation sessions; 3) correlation of users clusters and navigation sessions clusters. This analysis provides a complete and full-rounded picture of the virtual community users. *Copyright © 2003 IFAC*

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

1. INTRODUCTION AND PROBLEM STATEMENT

This paper deals with the analysis and modeling of a large data-set related to a very popular Italian virtual community which is constituted by more than 550.000 registered users. Each user is characterized by a unique nickname and has an own "profile" which is filled (choosing among a list of items) during the registration, on a voluntary basis (the profile can be left completely blank or can be only partially filled).

The analysis is made using two different data-sets:

- the database (DB) of the users (nicknames and profiles);
- 1-week log-file of the servers hosting the community web site.

Needless to say, these two data sets are extremely different: they deliver complementary pieces of information, and they must be processed and analyzed using 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 relationships between these two heterogeneous data sets. This is inherently a very challenging task, and – to the best of our knowledge – this is one of the first attempts documented in the Data-Modeling literature to "merge" and to find relationships between Users DB and web-navigation behaviors of a very large Virtual Community ([7,8]).

The search for relationships between half-million Users and millions of page-views cannot be faced directly from the raw data sets. The basic idea and methodological approach proposed in this work is the following:

- the Users DB has been analyzed and clustered into a small number (12) of clusters; each class represents a "prototype" of User (Section 2);
- the log-file of the web server has been 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 Users 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 Users and navigation sessions (Section 4). Note that this can be done since the page-views registered in the log-file contain the nickname of the User, stored in a *cookie*. This allows the linking between the Users DB and the log-file.

This not-trivial analysis and modeling – although preliminary – provides a very general and full-rounded picture of the virtual community users.

2. ANALYSIS AND CLUSTERING OF THE USERS DB

The bulk of the Users database has a simple structure, which is condensed into a single table, where each row is given by:

– the nickname (*primary-key* of the table)

– 12 fields, describing the “profile” of the user. For each field the user can select among a finite (well-defined) set of items. In the data-base only the numeric code of the selected item is stored.

Each user is, thus, characterized by a sequence of 12 values (plus the nickname) which are:

– *categorical*, since each field value represents a category, not a quantity (e.g. *job*=2 means that the user is a student)

– *non-ordinal* (e.g. saying that a student is greater than a clerk has non sense).

According to the indications expressed by the management of the virtual community, the analysis of this Data-Base has been done by focusing on 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 (i.e. the user can leave undefined one or more fields in his profile).

As first step, the entire data-set of the Users 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.

Using this data-set, preliminarily, the amount of users leaving undefined a specific field has been computed for each field. The result is displayed in Fig.1.

It is interesting to observe that:

- The gender is – by far – the most “filled” field. This confirms that this kind of Virtual Community is mainly seen as a mean for meeting (dating…) people.
- Age, language, sexual orientation, and “alone with” are the less voted fields.

The second step of the analysis made on the matrix *M* was the search of hidden relationships (also known as *Association Rules* – [8]) between the 12 fields of the profile. This analysis has been done by computing a sort of normalized “correlation” (or “dependence”) index $\Gamma(h|k)$ $(h,k=1,2,...,12)$. $\Gamma(h|k)$ has been computed as follows.

First, the *average mutual information* $I(h,k)$ between fields *h* and *k* ([8]) has been computed. It is defined as:

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

where *i* and *j* take all the possible values for the fields *h* and *k*, respectively. *p(E)* is the sample probability of the event *E*; it has been computed by exhaustive search in *M*. Using $I(h,k)$, the correlation index $\Gamma(h|k)$ hence can be computed as:

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

Note that $\Gamma(h|k)\in[0,1]$. $\Gamma(h|k)$ measures the dependence between *h* and *k*. More precisely, $\Gamma(h|k)$ measures the information level on *h* which can be obtained from the knowledge of *k*. For example, if *h* and *k* are independent, then $\Gamma(h|k)=0$ since the knowledge of *k* gives no information on *h*; on the contrary, if *k*=*h* then $\Gamma(h|k)=1$ since *k* describes completely *h*.

Note that $\Gamma(h|k)$ is not symmetric (in general $\Gamma(h|k)\neq\Gamma(k|h)$). As a matter of fact, the information on *h* given by *k* may be different from the information on *k* given by *h*.

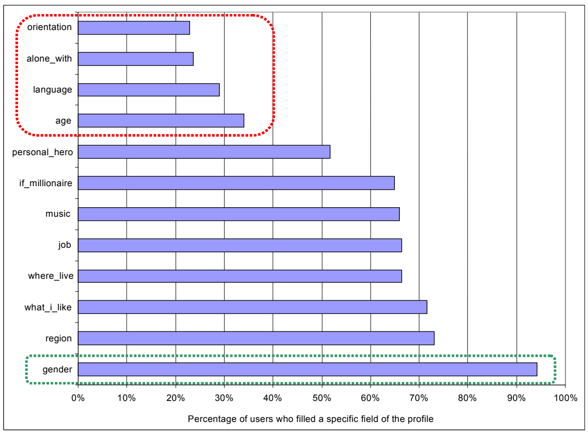


Fig.1. Willingness of the Users to fill a specific field of the profile

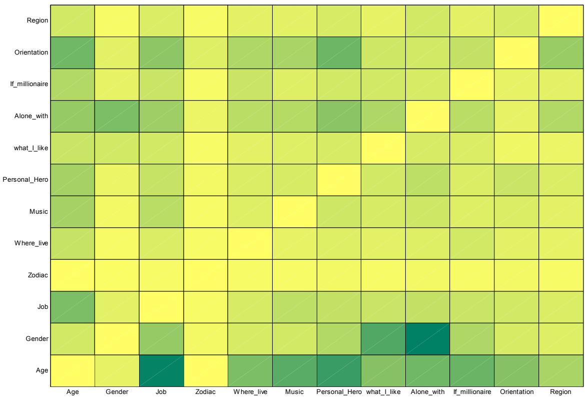


Fig.2. Association rules between the 12 fields of the profile (dark = strong correlation).

The results of this field-correlation analysis are condensed in Fig.2, where $\Gamma(h|k)$ is plotted as follows:

- each cell has been coloured proportionally to the value of $\Gamma(h|k)$, where *h* is the field on the row and k is the field on the column; the darker the cell is, the more $\Gamma(h|k)$ is close to 1 (hence a dark cell means strong correlation);

- 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 analysis of the correlation plot in Fig.2 reveals many interesting things. Among others:

- Age depends on most of the other fields (e.g. from the choice of the "Personal Hero", the age of the users can be easily predicted), 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 correlation is between age and job.
- The gender is strongly dependent on fields: "Job", "Personal Hero", "What I like" and "Alone With".
- "Region" seems to be independent of other fields, showing that it represents a piece of information which cannot be predicted from other fields.

The map in Fig.2 delivers many interesting association rules between the 12 fields. It represents, per-se, an interesting result.

The third step of the Users DB analysis was the clustering of the data-set into a limited number of clusters. According to the marketing goals of the web site provider and to the results of the correlation analysis, the clustering has been done using 3 fields only: gender, age, and geographic region. These fields are the most filled by users and are significant from a socio-demographic point of view.

Since data were categorical and non-ordinal, the most appealing clustering procedure, in this case, was a hierarchy of univariate decision on the 3 chosen fields ([8]). Thus, it has been built a *classification tree* such that each internal node specifies a subset member-ship test on a singular field (for example a test on Age could be: "verify if age is greater than 35, less than 35 or undefined"). Then, each row-vector **u** in M (representing a user) descends a unique path from the root node to a leaf node depending on how the values of individual components of **u** match the tests of internal nodes. The set of users reaching the same leaf node is a cluster of the data-set and the characteristics of each cluster can be obtained by following the path connecting the root-node of the classification tree to the corresponding leaf node.

Fig.3 displays the size and the characteristics of the 12 clusters obtained by applying the algorithm above to M. This partition is a simple socio-demographic partition, actionable for target marketing.

3. SESSIONIZATION, ANALYSIS AND CLUSTERING OF A 1-WEEK WEB-LOG FILE

The second data set analyzed in this work is the log-file of the servers hosting the Virtual Community web-site, collected during a week of January 2002.

It is a standard log-file delivered by an Apache 1.3 web server ([1]). In Fig.4 a small sample of this huge file is shown.

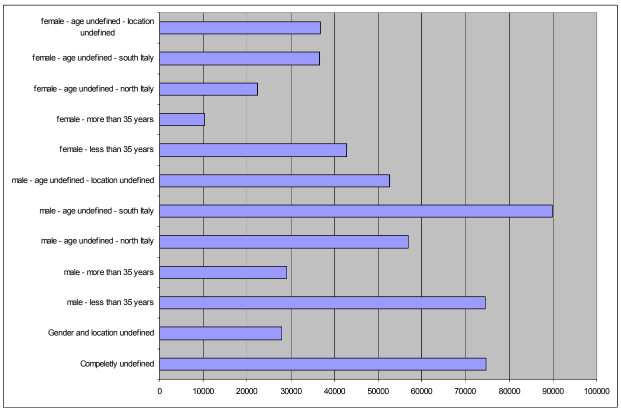
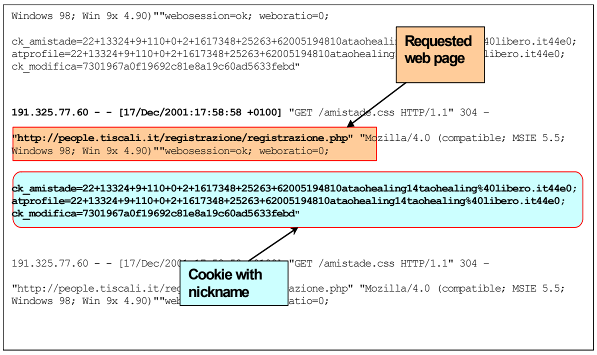


Fig.3. Partition of the whole Users DB (550.000 registered Users) into 12 clusters.

Each item (row) of the log-file represents a single "page-view" of a navigation session. The log-file contains, among other, the following data (see Fig.4):

- 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 indication of the nickname of the user.



```
Windows 98; Win 9x 4.90)""webosession=ok; weboratio=0;

ck_amistade=22+13324+9+110+0+2+1617348+25263+62005194810ataoheali 40libero.it44e0;
atprofile=22+13324+9+110+0+2+1617348+25263+62005194810ataohealing libero.it44e0;
ck_modifica=7301967a0f19692c81e8a19c60ad5633febd"


191.325.77.60 - - [17/Dec/2001:17:58:58 +0100] "GET /amistade.css HTTP/1.1" 304 -

"http://people.tiscali.it/registrazione/registrazione.php" "Mozilla/4.0 (compatible; MSIE 5.5;
Windows 98; Win 9x 4.90)""webosession=ok; weboratio=0;

ck_amistade=22+13324+9+110+0+2+1617348+25263+62005194810ataohealing14taohealing%40libero.it44e0;
atprofile=22+13324+9+110+0+2+1617348+25263+62005194810ataohealing14taohealing%40libero.it44e0;
ck_modifica=7301967a0f19692c81e8a19c60ad5633febd"


191.325.77.60 - - [17/Dec/200                "GET /amistade.css HTTP/1.1" 304 -

"http://people.tiscali.it/re              azione.php" "Mozilla/4.0 (compatible; MSIE 5.5;
Windows 98; Win 9x 4.90)""web             atio=0;
```

Fig.4. Sample of the raw log-file delivered by an Apache web server.

The log-file analyzed in this work is quite huge (about 2.7 GBytes). It is referred to one week (from 00:00 of Monday, to 24:00 of Sunday) of January 2002. The treatment of such a log-file has required some non-trivial pre-processing. After pre-processing, 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 Fig.8). 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 any).

All the (thousands of) different URLs registered by the web-server have been manually grouped into 30 sets, in order to be easily managed and interpreted.

Using this single-table Data-Base, a preliminary analysis has been done by computing the sample distribution of the page views on the 30 sets of URLs has been computed. This distribution results to be very 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 pages out of the Virtual Community web site. This can be explained by the fact that, in a fashion similar to the famous www.geocities.com Virtual Community, the User can put in his/her profile the link to his/her personal home page, which often is hosted on different domains. This confirms that personal web-pages are intensely visited during web navigation.

Starting from the raw data-base extracted from the 1-week log-file, the next step has been the “sessionization” of the page views. Sessionizing a log-file is known to be a tricky and subtle task, which requires some heuristics and a-priori assumptions (see e.g. [2]).

The navigation sessions have been stored into a real-valued matrix *S*. It is a 30×460.000 matrix, where the element $S_{i,j}$ is the number of seconds spent on the the *i*-th URL in the *j*-th session of the week. The matrix *S* has been built as follows:

- 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 (two URLs requested by the same IP address, separated by more than 15 minutes are assumed to belong to different sessions);
- the last page of a session has been assigned a nominal visiting time of 30 seconds (all other visiting times can be computed as the time difference between two subsequent page-views made by the same host).

In contrast with the matrix *M* of user profiles, each elements $S_{i,j}$ of *S* assumes quantitative and ordinal values (in fact, $S_{i,j}$ is the time spent by a user on a web page). Therefore, each row of *S* (i.e. each session) can be seen as a vector in a Euclidean space of dimensionality equals to 30 and the Euclidean metric can be used as distance between two sessions.

In this particular setting, using iteratively bisecting divisive partitioning algorithm (i.e. the data-set is divided in 2 clusters maximizing the distance between clusters itself and minimizing the distance between items in each cluster, see [9]) proved to be particularly suitable for the clusterization of matrix *S*. To be more precise, the bisection of the clusters was done, according to the analysis developed in [11,12], 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 1 and 2.

K-means is probably the most celebrated and widely used clustering technique; hence it is the best representative of the class of iterative centroid-based divisive algorithms ([10,13]). PDDP is a recently proposed technique ([4,5]). It is representative of the non-iterative techniques based upon the Singular Value Decomposition (SVD) of a matrix built from the data-set ([3,6]).

Table 1: PDDP clustering algorithm.

Step 1.	Compute the centroid *w* of *S*.
Step 2.	Compute the auxiliary matrix $\tilde{S}$ as: $\tilde{S} = S - we$, where *e* is a *N*-dimensional row vector of ones, namely $e = [1,1,1,1,1,\dots 1]$.
Step 3.	Compute the Singular Value Decompositions (SVD) of $\tilde{S}$: $\tilde{S} = U \Sigma V^T$, where Σ is a diagonal $p \times N$ matrix, and *U* and *V* are ortonormal unitary square matrices having dimension $p \times p$ and $N \times N$, respectively (see [7] for an exhaustive description of SVD).
Step 4.	Take the first column vector of *U*, say $u = U_1$, and divide $S = [x_1, x_2, \dots, x_N]$ into two sub-clusters S_L and S_R, according to the following rule: $\begin{cases} x_i \in S_L & if \quad u^T(x_i - w) \le 0 \\ x_i \in S_R & if \quad u^T(x_i - w) > 0 \end{cases}.$

Table 2: Bisecting K-means.

Step 1.	(Initialization). Randomly select a point, say $c_L \in \Re^p$; then compute the centroid *w* of *S*, and compute $c_R \in \Re^p$ as $c_R = w - (c_L - w)$.
Step 2.	Divide $S = [x_1, x_2, \dots, x_N]$ into two sub-clusters S_L and S_R, according to the following rule: $\begin{cases} x_i \in S_L & if \quad \lVert x_i - c_L \rVert \le \lVert x_i - c_R \rVert \\ x_i \in S_R & if \quad \lVert x_i - c_L \rVert > \lVert x_i - c_R \rVert \end{cases}$
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.

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 ([11,12]) 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 PDDP result.

The PDDP+K-means algorithms has been first applied to *S*; after the first bi-sectioning step, the algorithm has been iterated, each step bisecting one the clusters obtained in the step before. The decision on the cluster to split has been made heuristically, by direct inspection of the actual clusters.

The final result is a 8-cluster partition. The details on the whole partition (taxonomy of the all clustering steps) of *S* are displayed in Fig.5. The final clusters along with a brief characterization of them are displayed in Fig.6.

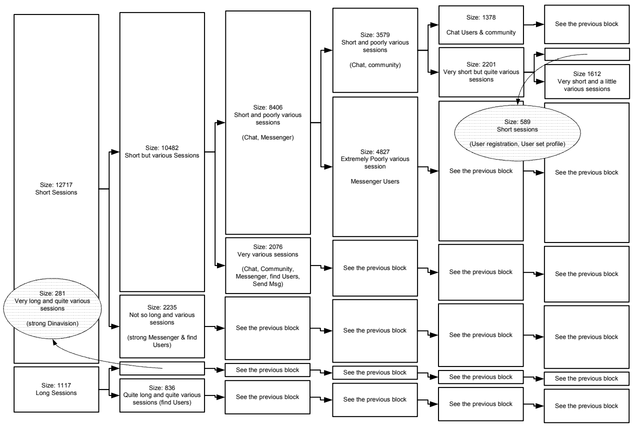

Size: 12717
Short Sessions
Size: 1117
Long Sessions
Size: 281
Very long and quite various sessions
(strong Dinavision)
Size: 10482
Short but various Sessions
Size: 2235
Not so long and various sessions
(strong Messenger & find Users)
Size: 836
Quite long and quite various sessions (find Users)
Size: 8406
Short and poorly various sessions
(Chat, Messenger)
Size: 2076
Very various sessions
(Chat, Community, Messenger, find Users, Send Msg)
Size: 3579
Short and poorly various sessions
(Chat, community)
Size: 4827
Extremely Poorly various session
Messenger Users
Size: 1378
Chat Users & community
Size: 2201
Very short but quite various sessions
Size: 589
Short sessions
(User registration, User set profile)
Size 1612
Very short and a little various sessions
See the previous block

Fig.5. Complete partition-tree of the session matrix *S*.

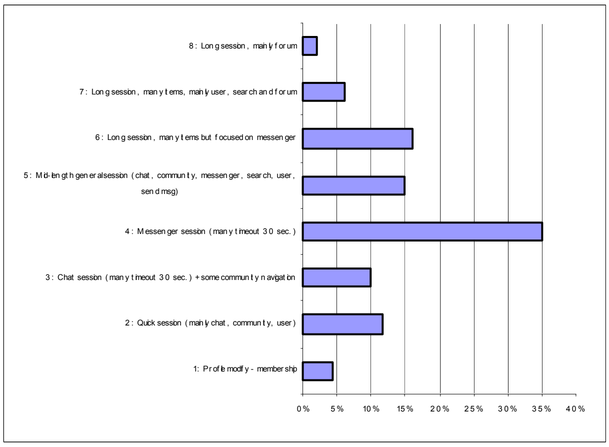


Fig.6.Leaves of the partition of the session matrix *S*.

As expected, the partition made by PDDP+K-means shows the most typical navigation sessions. Note, in particular, the relevance of messenger-based or chat-based sessions, and the navigations spent out of the Virtual Community domains.

4. ESTABLISHING RELATIONSHIPS BETWEEN USERS AND SESSIONS

The last step of the analysis presented in this work is the search of the main relationships between the Users DB and the navigation log-file. As already said in the Introduction, this task cannot be faced directly from the raw data-sets. The basic idea was to pre-process and reduce the Users 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.

To perform the correlation analysis between Users and sessions, a matrix *C'* of dimension 12×8 has been built as follows:

- According to his/her profile, each user 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 User has been extracted from the sessions-matrix *S*. Each session then has been classified into one of the 8 session clusters.
- A row vector of size 8 is built for each user; this vector represents the sample probability distribution (its sum is normalized to 1) of the sessions made by that user during the week.
- All the rows of the Users belonging to the *i*-th cluster have been summed. The result has been normalized and represents the *i*-th row of the matrix *C'*.

Thus, *C'(i,j)* represents the average frequency of performing a session in the *j*-th session cluster by a user in the *i*-th user cluster. The plot of *C'* is in Fig.7. The colour (darkness) of each cell of Fig.7 is proportional to the value of *C'(i,j)*. The coding of *i* and *j* is that displayed in Fig. 4 and Fig.13.

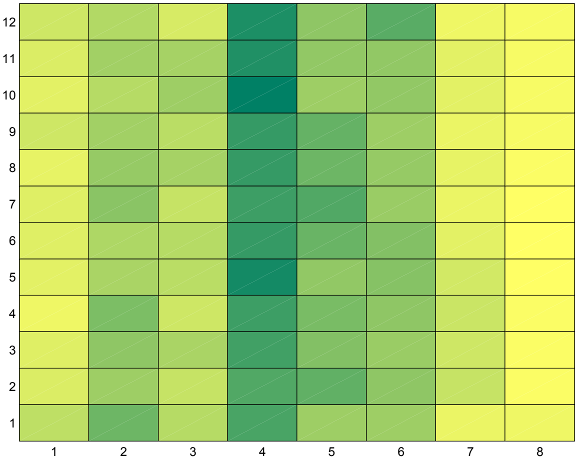


Fig.7. Correlation between the 12 clusters of Users and the 8 clusters of sessions.

As it appears, each cluster of users performs Messenger-based sessions in mainly (column 4). This is an interesting result since it highlights that all the prototypes of users behave similarly in average. However, this predominance of Messenger session hides the difference between user clusters.

To avoid this, a new correlation matrix *C* has been computed from *C'* by dividing (scaling) each column of *C'* by the average value of the column. The plot of *C* is in Fig.8. Again, the colour (darkness) of each cell of Fig.8 is proportional to the value of *C*(*i*,*j*), where *i* is the User cluster (row) and *j* is the session cluster (column).

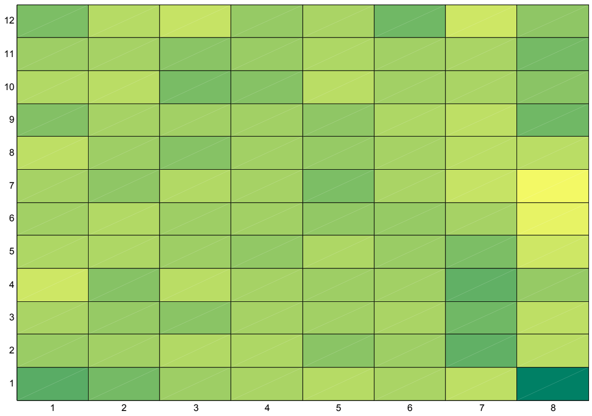


Fig.8. Scaled correlation between the 12 clusters of Users and the 8 clusters of sessions.

By analysing the results displayed in Fig.8, 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.9 and, among others things, it can be seen that:

- males seems to be very related to long and various sessions;
- females seems to be primarily interested to sessions with forum or chat;
- long sessions focused on the messenger seem very correlated with Users who left the gender blank.

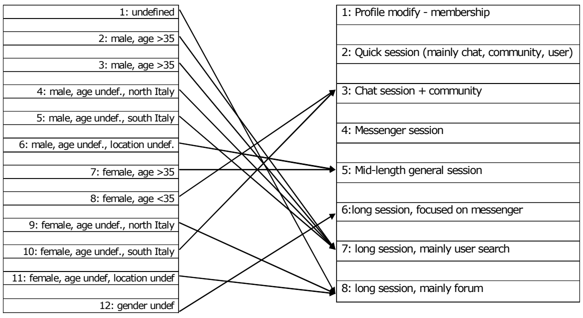

1: undefined
2: male, age >35
3: male, age >35
4: male, age undef., north Italy
5: male, age undef., south Italy
6: male, age undef., location undef.
7: female, age >35
8: female, age <35
9: female, age undef., north Italy
10: female, age undef., south Italy
11: female, age undef, location undef
12: gender undef
1: Profile modify - membership
2: Quick session (mainly chat, community, user)
3: Chat session + community
4: Messenger session
5: Mid-length general session
6:long session, focused on messenger
7: long session, mainly user search
8: long session, mainly forum

Fig.9. Main association rules between Users and sessions

5. CONCLUSIONS

In this paper a case study of Data-Modeling 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.

ACKNOWLEDGMENTS

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

REFERENCES

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

[2] 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*.

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

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

[5] 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.

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

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

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

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

[10] 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.

[11] 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.

[12] 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.

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