

Data-Mining of a Large Virtual Community: Relationships Between the Users DB and the Web-Log File

S.M. Savaresi[†]

Simone Garatti[†]

Sergio Bittanti[†]

Luca La Brocca*

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, 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 Data-Base of the Users, and 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.

1. Introduction and problem statement.

This paper deals with the analysis and Data-Mining of a large data-set related to a very popular Italian Virtual Community. This Virtual Community is constituted by more than 500.000 registered users, characterized by a unique nickname. Each user has its own profile, which is filled during the registration, on a voluntary basis.

The analysis is made using two different data-sets:

- the Data-Base (DB) of the Users;
- 1-week log-file of the servers hosting the Community web-site.

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 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-Mining literature to merge and to find relationships between Users DB and web-navigation behaviors of a very large Virtual Community ([4, 5]).

The search for relationships between Users and 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 into 8 clusters using an unsupervised bisecting divisive clustering approach; each cluster represents a “navigation behavior” (Section 3);
- thanks to the 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 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 Users DB and the log-file.

This analysis 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 DB has a simple structure, which is condensed into a single table; 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 set of items. In the data-base only the numeric code of the item selected by the user is stored. The User profile is “entertainment-oriented”.

According to the indications expressed by the Management of the *Tiscali Virtual Community Division*, the analysis of this Data-Base has been done 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 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,

[†] Politecnico di Milano, Milano, ITALY.

* Tiscali S.p.A., Cagliari, ITALY.

Paper 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 Davide Romieri and Paolo Prestinari for enlightening discussions on Virtual Communities.

the following preliminary analysis has been done:

- the amount of Users having all “1” (all fields have been left undefined) has been computed.
- the amount of Users having “1” in a specific field has been computed for each field.

The results are displayed in Fig.1 and in Fig.2, respectively. It is interesting to observe that:

- A small number of Users (less than 15%) leaves the profile completely blank. This reveals that only a small part of subscribers are pure visitors who are not interested in establishing relationships.
- 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.

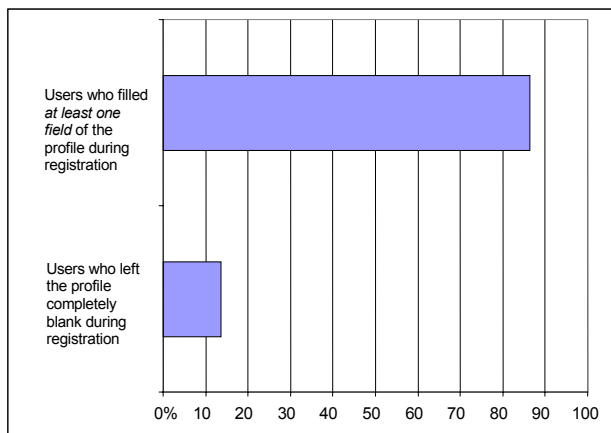


Fig. 1. Blank vs. filled profiles.

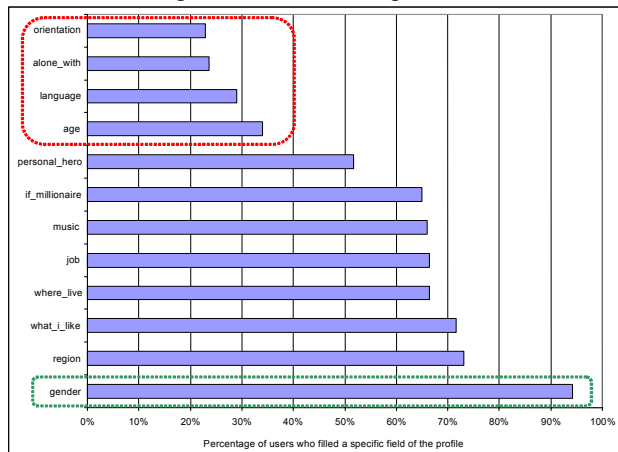


Fig. 2. Willingness to fill a specific field of the profile

The second step of the analysis made on the matrix M was the search of hidden relationships (also known as *Association Rules* in Data-Mining – [5]) between the 12 fields of the profile. This analysis has been done by computing a sort of normalized “correlation” (or “dependence”) index $\Gamma(h|k)$ ($h, k = 1, 2, \dots, 12$). $\Gamma(h|k)$ has been computed as follows.

First, the *average mutual information* $I(h,k)$ between fields h and k ([5]) 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. $p(E)$, the sample probability of the event E , 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. As a matter of fact, the information on h given by k may be different from the information on k given by h .

The results of this field-correlation analysis are condensed in Fig.3, where $\Gamma(h|k)$ is plotted as follows:

- each cell is 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 (dark 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.

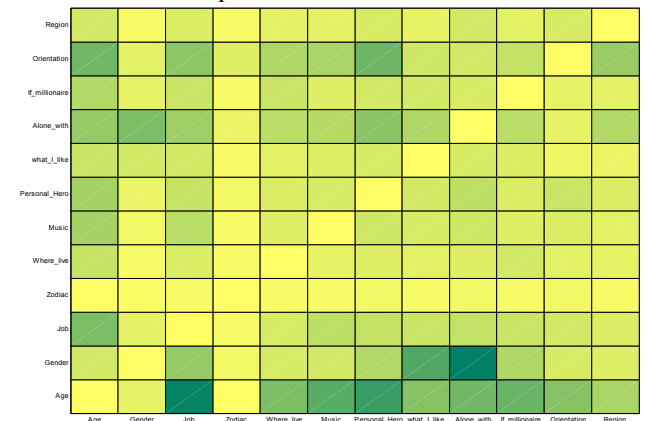


Fig. 3. Association rules between the 12 fields of the profile (dark = strong correlation).

The analysis of the correlation plot in Fig.3 reveals many interesting things. Among others:

- Most of the fields are strictly correlated with the age of the User (e.g. from the choice of the “Personal Hero” the age of the Users can be easily predicted). This is somehow expected and suggests that the age is a good field for clustering. Note that the strongest correlation

- 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).

The matrix S then has been clustered using a bisecting divisive partitioning algorithm ([6]). According to the analysis developed in [2, 7, 8], the bisection of the clusters was 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, the PDDP algorithm is here briefly recalled.

PDDP: the 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 $e = [1, 1, 1, 1, \dots, 1]$.
- 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 orthonormal unitary square matrices having dimension $p \times p$ and $N \times N$, respectively (see [3]).
- 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:

$$\begin{cases} x_i \in S_L & \text{if } u^T(x_i - w) \leq 0 \\ x_i \in S_R & \text{if } u^T(x_i - w) > 0 \end{cases}$$

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. PDDP is a recently proposed technique ([2]). It is representative of the non-iterative techniques based upon the Singular Value Decomposition (SVD) of a matrix built from the data-set.

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. It has been proven 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 decision on the cluster to split has been made heuristically, by direct inspection of the actual clusters. The final result is a 12-cluster partition.

The same procedure has been applied to the smaller matrix S_0 ($S_0 \subset S$). S_0 is made by the subset (about 3%) of navigation sessions made by registered logged-in Users. This partition is made by 8 clusters. The details on the whole partition-tree (taxonomy) of S_0 are displayed in Fig.6. The “leaves” of this partition are displayed in Fig.7. This partition will be used in the next Section, where Users

and sessions will be correlated.

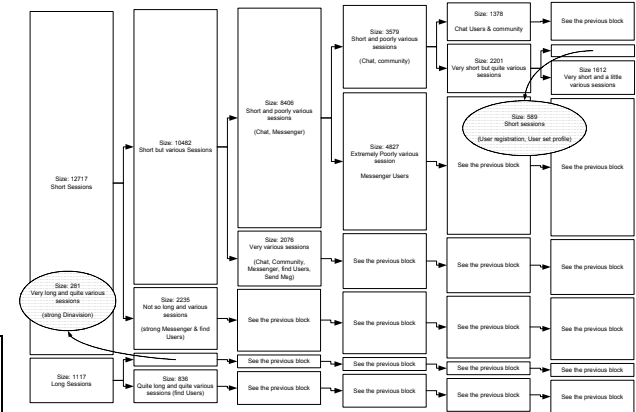


Fig.6. Complete partition-tree of the session matrix S_0 (sessions made by logged-in Users).

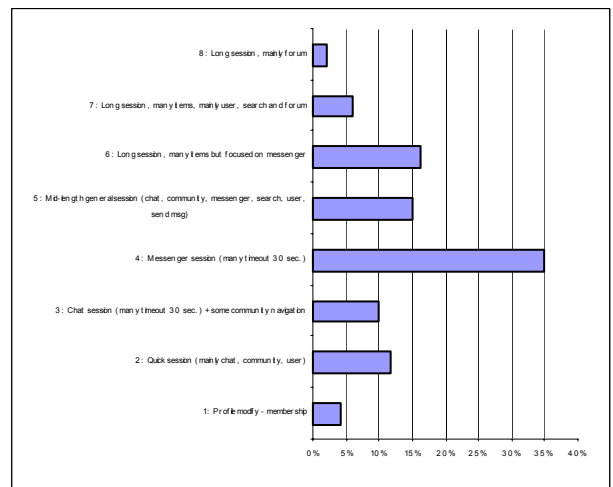


Fig.7. Leaves of the partition of the session matrix S_0 (sessions made by logged-in Users).

By inspecting the clustering results in Figs.6-7 the following remarks are due:

- 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 onto not-Tiscali domains.
- Note 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 than the average session time.

4. 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.

Regarding the 12 clusters of Users and the 8 clusters of navigation sessions, the following remarks are due:

- 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. It is interesting to 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. For example, the 3.000 “active” users registered in the log-file have a much higher willingness to fill the profile during registration.

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

- According to his/her profile, each of the 3000 Users has been classified into the i -th of the 12 Users cluster (hence it has been associated to the i -th row of the matrix C); the whole set of sessions made by such User has been extracted from the sessions-matrix S_0 . 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 types of 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 \tilde{C} .
- The matrix C has been computed from \tilde{C} by dividing (scaling) each column of \tilde{C} by the average value of the column.

The plot of C is in Fig.8. 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).

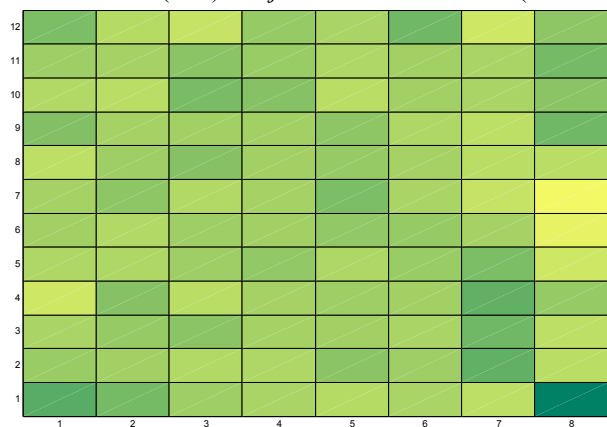


Fig.8. Correlation between the 12 clusters of Users and the 8 clusters of sessions (dark=strong correlation).

By analysing the results displayed in Fig.8, the map of the main association rules between clusters of Users and clusters of sessions can be built. This map is displayed in Fig.9. From this figure many interesting pieces of information can be drawn. Among others: 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.

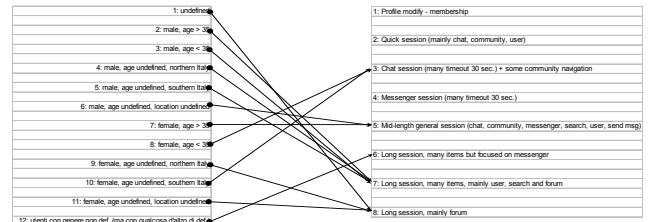


Fig.9. Main association rules between Users and sessions.

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.

References

- [1] 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*.
- [2] Boley, D.L. (1998). “Principal Direction Divisive Partitioning”. *Data Mining and Knowledge Discovery*, vol.2, n.4, pp. 325-344.
- [3] Golub, G.H, C.F. van Loan (1996). *Matrix Computations (3rd edition)*. The Johns Hopkins University Press.
- [4] Hagel J.III, Armstrong A.G. (1999). *Net Gain: Expanding Markets Through Virtual Communities*. Harvard Business School Press.
- [5] Hand D., Mannila H., Smyh P. (2001). *Principles of Data Mining*. MIT Press.
- [6] Jain, A.K, M.N. Murty, P.J. Flynn (1999). “Data Clustering: a Review”. *ACM Computing Surveys*, Vol.31, n.3, pp.264-323.
- [7] 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.
- [8] 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.