CIMSA 2003 - International Symposium on
Computational Intelligence for Measurement Systems and Applications
Lugano, Switzerland, 29-31 July 2003

Evaluating the Impact of the user profile dimension on its characterization effectiveness: Method based on the evaluation of user community organizations quality

Luigi Lancieri, Nicolas Durand
France Telecom R&D
42 Rue de coutures, 14000 Caen, France
{luigi.lancieri, nicola.durand}@francetelecom.com

***Abstract**: This paper studies the effect of the users' profile size (weighted set of keywords) on the efficiency of the measurement. The goal is not directly to identify the better dimension but to understand its influence on the quality of the result. We show that contrary to what common sense suggests, high dimension profile (high level of knowledge about users) reduces the precision of the measurement. To obtain this result, we develop a metric comparison method based on the evaluation of cluster organization quality. One of the main practical applications of this result is that: more than better precision, reduced profiles speed up computation time and reduce needed resources. We also discuss more theoretical possible conclusions and application of our work in the context of community characterization.*

I. INTRODUCTION

Information systems take an increasing position in our every day life. In this context, one of the biggest challenges is taking into account the human factor to optimize information management. This is necessary not only to provide the best information service to users but also because information system dynamics are highly influenced by human factors. This is obvious for example in the Internet where information availability and network bottlenecks are directly dependent on users' activity.

Taking into account the human factor in computer systems involves characterizing individuals in order to build efficient metrics. One of the questions is what do we need to measure? A possible answer to this question is to define one-dimensional metrics in order to have a quantitative view of the user activity (e.g. how many website visited per day, average size of each download, etc.). This approach was the basis of lots of paper that studied statistics based models of the human factors. For example some works shows that distribution laws involving human activity tend to be self-similar (fractal) [6][7]. An other answer more complex is to take into account the semantic of the user activity. The potential of such approach is higher since it is more descriptive and allows having a view on human motivation since the nature of the user description is the same that of the information he manipulated. In fact, since human produce it, this information reflects the operating mode of his own cognition. Such description can be used for example to identify thematic or behavior similarity between users or to identify communities of interest.

We consider a profile as a generalized descriptor composed with a set of weighted symbolic elements used to characterize an entity. In this study, the entity is a user or a community and the profile is a set of keywords associated to numerical weights. The figure 1 shows an example of a profile.

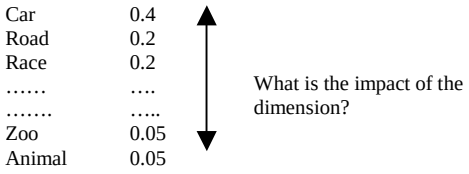

Fig 1: An example of profile

One of the main problems is that such profiles are highly multidimensional and permeable to noise. Furthermore, the part of the noise is very difficult to evaluate mainly because such metric is subjective. So, it is difficult to define a suitable precision for these characteristics. For example in textual document characterization, the typical dimension of a word vector is of the size of the vocabulary and tens of thousands of words are used routinely [13]. In image characterization, the typical dimension is 128^2 (16384). Intuitively, it can seems that high dimension (i.e high quantity of knowledge) profile should give a better precision but higher computational cost. This intuition may lead us to seek the highest profile dimension that stays compatible with the available computational capacity. As we will see, our study shows that contrary to the feeling; a high dimension profile does not give the better results. Furthermore, most of the time, low dimension profile gives excellent results with low computation time.

On the other hand, a motivation for dimension reduction is, independently to the precision, that some application as data visualization needs it to be usable since multidimensional dataset is visually understandable. Furthermore, a high dimensional space is sparse by nature

since the size of the sample needed to estimate a multivariable function grows exponentially with the number of variables (e.g. a word appearing frequently in a document may not appears in any of other document of the set).

More generally, the interest of the information space reduction was already shown in several fields. For example information retrieval techniques based on dimensionality reduction, are known to be very efficient [11]. Statistic analysis kind of method such as Latent Semantic Indexing (LSI) [11], Principal Component Analysis (PCA) [16] or Multidimensional Scaling (MDS) [17] are interesting when the relation among the variables are linear. For a more general set of data it is interesting to use Neuro-computational models as Kohonen Self-Organizing Map (SOFM) [12]. These methods are adaptative and self optimize the choice of the best space dimension. The problem is that they are not very descriptive and give a poor view on the underlying model linking, for example, the space dimension and the efficiency of the metrics.

The objective of the study is to evaluate the impact of the profile size (i.e number of keywords) in its descriptive effectiveness. In order to evaluate this effectiveness, we compared the result of a clustering based on statistically composed profile with a reference clustering based on human evaluation.

First we describe the general method underlying our experiment and we develop the principle of the evaluation of communities' organization quality measure. Then we present the numerical and associated models as results of the study. Before concluding, we discuss the limits and on some perspective of this experiment.

II. METHODOLOGY

The evaluation of the impact of the profile dimension on the effectiveness of the measurement is not an easy task. It needs to have a reference and a comparison method. We choose as reference the human point of view. About the comparison method, the difficulty is to compare 2 profiles of different sizes. Indeed a profile can be expressed as a point in a multidimensional vector space and the comparison of 2 profiles can only be done in the same vector space (i.e same profile dimension). A practical way to compare 2 profiles of different dimension is by comparing the corresponding effect in their use. We choose, as comparison method the organization of clusters of users based on their profile (see § III). Indeed the metric capacity of each user's profile influences the clusters organization. So we use the quality of the organization as representative of the profile metric quality. More the profiles are effective more the tested cluster organization is close to a reference one.

Initially, we have a set of 60 textual documents of homogeneous sizes (about 2 pages each) analyzed by a 6 people jury. Each member of the jury has provided a "subjective" evaluation for each document. This evaluation consists on giving a percentage (according to their feeling) for 20 normalized themes (keywords given by us). For example, we can estimate that a document deals mainly with sports (30 %), technology (10 %), etc. By carrying out an average of the individual evaluations, we obtained for each document a normalized profile composed of 20 weighted themes. In the second step, we consider a population of 50 virtual users consulting these documents. To simulate this consultation, we assign randomly 20 documents to each user. We calculate then the "virtual" profile of these users by identifying the most frequent keywords in the 20 consulted documents. In order to evaluate the best size of this statistical profile we cluster the users in communities for several size of their profile. So, we obtained several community organizations, each one corresponding to a specific users' profile size. This was carried out with a hierarchical agglomerative-clustering algorithm (HAC) [RON98]. This algorithm starts from a distance matrix between users. Then it dispatches in the same cluster users having lowest inter-user distance (i.e. highest profile similarity) and preserving highest inter-cluster distance. The third stage consists in clustering the users' in communities as in the previous stage but this time on the basis of the profile supplied by the jury evaluation. The results of this "real" categorization will be compared with the "virtual" one obtained previously. Stage 4 carries out this comparison between each of the virtual communities organization (for several profile size) and the real one. The figure 2 shows a synthesis of the experiment.

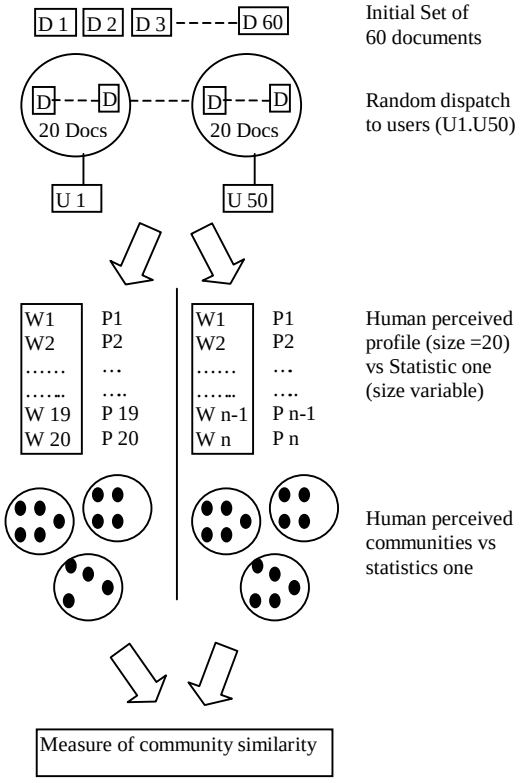

Fig 2 Synthesis of the experiment

III. COMPARING QUALITY OF COMMUNITIES' ORGANIZATION.

The postulate of our approach is that the description effectiveness of two sets of profiles (i.e real and virtual) is close if the resulting communities organization are close. Two close organizations involve close distribution of users in similar clusters separated with similar distances. In order to compare the organization of each virtual community with that of the referenced one, we use a measurement based on the Hausdorff distance named after Felix Hausdroff (1868-1942). This measurement is used in many applications in classification and imagery, for instance: face identification, object tracking and classification, comparing 2D images of the 3D world, etc.

The Hausdorff distance is the "maximum distance of a set to the nearest point in the other set". More formally, Hausdorff distance from set *A* to set *B* is a maximum function, is defined as

$$h(A,B)=\max_{a\in A}\left\{\min_{b\in B} d(a,b)\right\}$$

where *a* and *b* are points of sets *A* and *B* respectively, and *d(a,b)* is any metric between these points ; for example, the Euclidian distance between *a* and *b*.

In order to compare two organizations, Karonski and Palka [1] proposed a Hausdorff based distance using the Marczewski and Steinhaus similarity measure [2]. This distance makes it possible to compare all couples of partitions of the same unit, even if the partitions contain a different number of groups. If *A* and *B* are two partitions (i.e. two sets of clusters), the distance between *A* and *B* is defined in the following way:

$$D(A,B)=\frac{1}{2}\left(h(A,B)+h(B,A)\right)$$

$$D(A,B)=\frac{1}{2}\left(\max_{a\in A}\left\{\min_{b\in B} d(a,b)\right\}+\max_{b\in B}\left\{\min_{a\in A} d(a,b)\right\}\right)$$

If *a* and *b* are two clusters respectively from partition *A* and the partition *B*, the distance between these two clusters is defined in the following way:

$$d(a,b)=1-\frac{|a\Delta b|}{|a\cup b|} \quad où \quad a\Delta b=a\cup b-a\cap b$$

The symmetrical difference (noted Δ), of two clusters corresponds to the set of the objects belonging to only one of both clusters (see Figure 3).

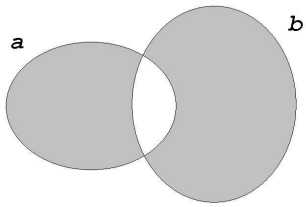

Fig 3: Symmetrical difference between two sets *a* and *b*

The distance *D(A,B)* evaluates the similarity between the two partitions in the worst case. With each cluster of the first partition, we associate the cluster of the second partition that is closest for it within the meaning of *d*, and we take into account that in the measurement only the associated couple of clusters whose distance is largest. In order to obtain a distance, i.e. a symmetrical measurement, we repeat the process by exchanging the role of *A* and *B*. The distance *D* is then the average of the two selected *d* distances.

IV. RESULTS.

We present two main results from which we tried to build a simple model underlying the variation. First, we studied the impact of the profile dimension in the number of obtained clusters and consequently on the compared classification quality.

Figure 4 shows that the number of clusters *(Ncl)* moves in a quadratic way according to the size of the profile *(S)*, with *k1 and k2* constant:

$$Ncl=k1.S^{2}+k2$$

This model materialized by the layout in dotted lines represents the approached curve with a value of k1=1/150 and of k2=5. We see that the number of groups, independent from the number of user, increases very quickly with the dimension of the profile. and tends to the maximum population of user (here 50).

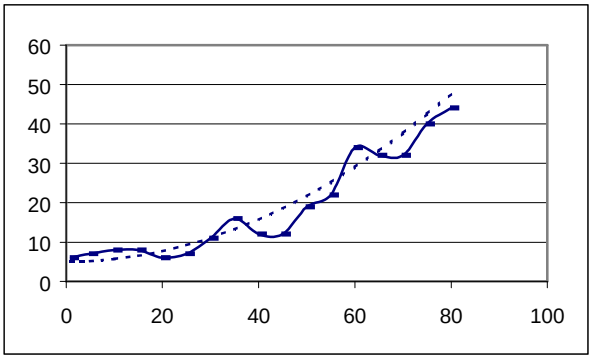

Fig 4: Numbers of clusters (Y-coordinate) according to the size of profiles (X-coordinate).

In other words, as shown in Figure 5, whatever the number of users, a too high profile dimension result on associating only one user per cluster, which is a poor organization! We see that not only is a high dimension profile are computer resources consuming to build and to use but also it is far to be effective for clustering purpose. The practical consequence is that clustering users on community needs a rather low profile dimension in order for it to be useful.

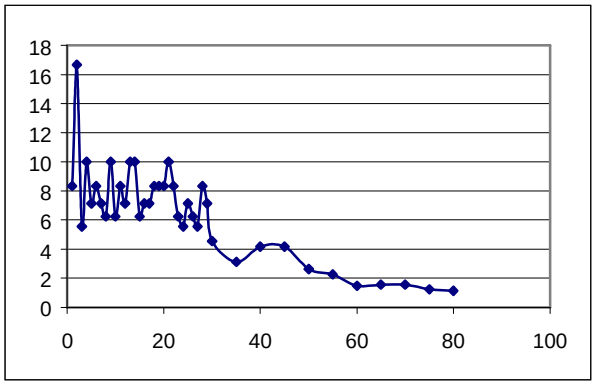

Fig 5: Average numbers of users per cluster.

To evaluate the quality of the virtual organizations we compare them with the real one that we consider as the best. The Figure 6 shows the evolution of this similarity according to the size of the users' profile. In this case one also notes a quadratic evolution with an optimum of similarity for sizes of profile ranking between 6 and 13 elements. The following formula models the evolution of the similarity (*Sim*) according to the size (*S*) of the virtual users' profile with k3, k4, k5 constant.

$$Sim = k3.(S - k4)^2 + k5$$

The layout in dotted line represents the approached curve with a value k4 corresponding to the model optimal size of profiles (here 11), of k3=0.27 and k5=1/400.

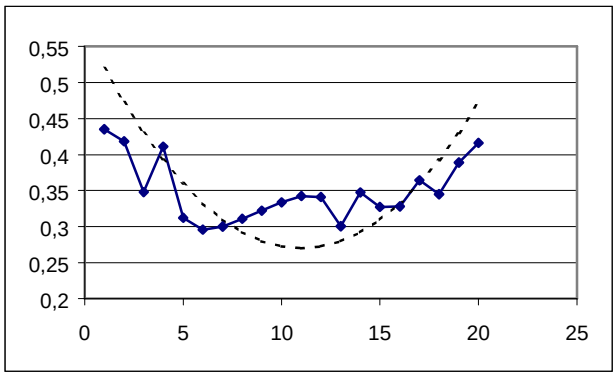

Fig 6: Clustering quality according to the size of the users' profile

We can see that there is a limited range of profile size that gives results close to the real segmentation. This result is interesting because it makes it possible to consider reduced computer resources and processing times with however good characterization qualities.

V. DISCUSSION

It is clear that out of a certain dimension the size of the profile rapidly reduces the precision of the measurement. According to our experiment, some works dealing with the comparative evaluation of classical dimension reduction methods [13][14] often remark that the best 5 or 6 dimensions are highly most significant than others and sufficient to obtain good results. So, an interesting question not solved in this paper could be to define a model capable to give in a general case the best dimension. The problem is not easy. For example it is important to evaluate the impact of the size of the "real profile" that was fixed to 20 in our study in order to make vary only the "virtual profile". It is also interesting to evaluate the impact of the "real profile" subjectivity. The table 1 shows that sometime the individual perception can be sparse. The table shows for each user the correlation coefficient to others users in the set of document evaluation.

Table 1: Inter user correlation coefficient

U1	U2	U3	U4	U5	U6	
1,00	0,75	0,76	0,74	0,59	0,79	U1
	1,00	0,75	0,72	0,44	0,70	U2
		1,00	0,72	0,53	0,75	U3
			1,00	0,49	0,71	U4
				1,00	0,44	U5
					1,00	U6

Even if several improvements can be done in order to get more general results, this study shows that a characterization that is too detailed, in the same manner as one that is too limited harms the quality of the community segmentation. Knowing that the typical dimension size is around ten and that it is not worth manipulating hundreds of dimension profiles is yet a practical usable result.

The other interesting question is to understand the basic conceptual principle underlying these models. Paradoxically, by reducing the size of the profile, i.e. loosing information we obtain better results. This kind of phenomenon can be also observed in the cases of over learning or in the everyday life: too much information is prejudicial. Some work on artificial forgetting (implicit selective loss of information) showed that this process could be controlled and used for optimization purposes [4]. It was shown, for example, that forgetting processes drove by collective intelligence is able to automatically reduce the information space on topics that are of main interest to a community. By keeping the most popular objects downloaded by the community and implicitly deleting others this approach can be interesting for information search purpose [10]. In fact the natural forgetting process in a human brain also save "processing"

capacity in order to concentrate brain activity on the essential task. From this point of view and contrary to the common point of view that the forgetting effect can be positive.

The method of evaluation of cluster organization is also interesting to characterize communities. Such method could be useful also to evaluate the community evolution over a period or to compute the compared community behavior. Such community-oriented metrics can be useful to have a better view of the cooperation dynamics and can be interesting associated with social approach.

Another more prospective aspect of this study can be discussed from the social point of view. The question could be expressed as: what is the effect of the level of individual knowledge of the others on the community constitution. Our study suggests (see Figure 5) that, on a knowledge basis only, the more information on the individuals is available the less individuals tends to regroup themselves. We could say from a theoretical point of view that, if people group themselves on the basis of affinity (i.e "hope" of maximum shared interest) the increasing knowledge of others reduce the ratio of potential shared interest since all individuals are different. It is difficult to get a definitive conclusion on this matter since the knowledge aspect is bound to be the only factor of influence in social groups (affective, power relation, etc). This view may not seem realistic in the real world but it is not so unrealistic in the information world. For example lets imagine that people "hear" about (low level of information) an interesting news forum on Internet. At the beginning, people get connected and exchange mail in the forum (the community of people). After a period, people could estimate that the forum is well known and that it is no longer worth visiting. A lot of news forums died for this kind of mechanism. The case of new forums is interesting because it is a simplified social case with less affective or physic interaction.

.

REFERENCES

[1] *Mr. Karonski and Z Palka. "One standard Marczewski-Steinhaus outdistances between hypergraphs". Zastosowania Matematyki Applicationes Mathematicae, 16(1):47-57, 1977.*

[2] *E Marczewski and H. Steinhaus. "One has certain outdistances of sets and the corresponding distance of functions ". Colloquium Mathematicum, 6:319-327, 1958.*

[3] *P. Ronkainen. "Attribute Similarity and Event Similarity Sequence in Dated Mining "Technical C-1998-42 Carryforward, University of Helsinki, October 1998.*

[4] *L Lancieri " Memory and forgetting: Two complementary mechanisms to characterize the various actors of the Internet in their interactions "; PhD Thesis (computer and cognitive science) University of Caen (France) 2000*

[5] *A. Balachandran and G. M. Voelker and P. Bahl and P. Venkat Rangan. "Characterizing User Behavior and Network Performance in a Public Wireless LAN". In Proc. of SIGMETRICS, June 2002.*

[6] *M. Crovella and A. Bestravos. "Self-Similarity in World Wide Web Traffic: Evidence and possibles causes". In Proceedings of the IEEE ACM Transactions on Networking, December 1997.*

[7] *N. Durand and L. Lancieri. "Study of the Regularity of the Users' Internet Accesses". In Proceedings of IDEAL 2002, Manchester, August 2002.*

[8] *L. Lancieri. "Description of Internet User Behavior". In Proceedings of the IEEE International Joint Conference on Neural Network (IJCNN'99), Washington, 1999.*

[9] *W. Leland and M. Taqqu and W. Willinger and D. Wilson. "On the Self-Similar Nature of Ethernet Traffic". In Proceedings of ACM SIGCOMM'93, pages 183-193, San Francisco, 1993.*

[10] *Luigi Lancieri, Nicolas Berthier Bonnel, Ludovic Stumme; To exploit the collective intelligence thanks to the Co-operative replication; In proceedings of International Conferences on Info-tech & Info-net ICII2001-Beijing Oct.29 - Nov.1, 2001.*

[11] *S.T. Dumais. Using LSI for information filtering: TREC-3 experiments. In Proc. of the Third Text REtrieval Conference (TREC-3), National Institute of Standards and Technology, 1995.*

[12] *T. Kohonen. Self-Organization and Associated Memory. Springer-Verlag, 1988.*

[13] *Chris Ding, Xiaofeng He, Hongyuan Zha, Horst Simon Adaptive Dimension Reduction for Clustering High Dimensional Data.. Proc. 2nd IEEE Int'l Conf. Data Mining, pp.147-154, Dec. 2002. Maebashi, Japan.*

[14] *Shiping Huang, Matthew O. Ward, and Elke A. Rundensteiner, Exploration of Dimensionality Reduction for Text Visualization (April 2003).. Technical report Computer Science Department Worcester Polytechnic Institute WPI-CS-TR-03-14*

[15] *George Karypis and Eui-Hong (Sam) Han Concept Indexing: A Fast Dimensionality Reduction Algorithm with Applications to Document Retrieval & Categorization, in proc of CIKM 2000*

[16] *J. Jolliffe. Principal Component Analysis. Springer Verlag, 1986.*

[17] *M. Davison. Multidimensional scaling. John Wiley & Sons, 1983.*