# Semantic Organization of Data Networks

Luigi Lancieri
*France Telecom R&D*
*Luigi.lancieri@francetelecom.com*

## Abstract

*The goal of this article is to study the principle of self organization of p2p data networks on the basis of users' activity. Basically, the idea is to link new discovered nodes according to the similarity between the previously accessed contents of each node. We studied the feasibility and the influence of several parameters on such organisation. This allows us to evaluate the efficiency and the stability of the association. The study shows that a semantic organization is more effective than a random organization*

## 1. Introduction

As interpersonal vector of communication, data networks as Internet tend to supplant traditional media like the telephone or the postal mail. One of the major changes in communication introduced by the evolution of data networks is certainly linked to the interactions between users and more generally to the human factor. Actually, data networks and users are mutually influenced. The human influence can be demonstrated by the variable level of latency times when accessing to "popular" and consequently saturated resources. It is also obvious that the technology changes our behaviour. According to a study carried out in 2003 by Taylor Nelson/Sofres for MSN, 59 % of the users utilize more frequently the e-mail than the traditional postal mail. (See [18] for a philosophical and social approach of these questions)

Since it is clear that human factor and data networks are mutually influenced, it can be attractive to use data describing uses, topics of interest or users' behaviour in order to organize and optimize the access to the information. To be effective, these data must be more descriptive than basic quantitative information and need to take into account the semantic of exchanges. For example, the evolution during a period of syntactic elements like the URLs or the keywords contained in the exchanged documents can reveal the needs or the behaviour of users. Indeed, such data are closer to human cognition mechanisms than quantitative information (e.g. quantity of requests, objects size distribution, etc).

In this context, our paper reports a study aiming the validation of the semantic organization of data networks on the basis of the content of exchanged data flows. Initially, we would like to validate the hypothesis that semantic organization is more effective than a random organization. Such approach can be used on ad hoc or overlay data networks like p2p or more generally in agent based technologies. The basic situation corresponds to nodes having a certain level of autonomy and deciding to choose one or several other nodes to connect with. In our case the criteria of decision are linked to the semantic of nodes activity. As a selection process, this strategy is used to make possible or to restrict connections between nodes based on users' information needs. In a wide data network, it is not optimal (nor feasible in some case) for each node to scan the content of all new nodes because that will be time and resources consuming. The semantic organisation allows a node to restrict the scope of investigation based on thematic similarity. The reduced set of nodes can be then scanned with a higher guaranty that it contains interesting data. Even though we took the example of P2P networks, our study has a more general goal.

This article is organized as follows: In the first part, we show that the sharing of data between connected nodes is higher if the association is made on criteria of semantic proximity. In the second step, we evaluate the granularity impact of used semantic data. Then we study the stability of resulting networks. Finally, before discussing some applications and limits of our approach, we presented a survey of related works.

## 2. Relation between semantic distance and real interactions

The basic architecture underlying our experiment is as showed in the following figure.

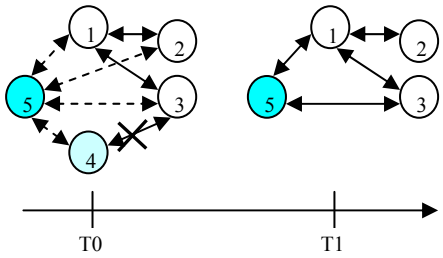


**Figure 1: Architecture of nodes interactions**

The figure 1 (t0) shows 3 nodes (1,2 and 3) exchanging data in a stable way whereas the node 4 is leaving from the group and the node 5 is coming in the group (ephemeral organisation). Our concern is to evaluate node 5 priority of connection toward other nodes. Potentially it could connect itself to all nodes but, if this is easy here with only 4 other nodes, we should imagine a situation with hundreds of nodes as it could occur in some case. In this situation the time and the resources needed to explore all the nodes are not supportable. In our context the node 5 only needs the thematic profile (a sum up of knowledge) of the other nodes in order to decide a priority of connexion. After a while, the result could be to connect only to the nodes 1 and 3 (figure 1 t1).

Our experiment aims at exploring the feasibility of such approach. We have real activity traces of 245 nodes over a period of 17 months. In order to have a realistic view of nodes behaviour, we simulate their activity with real user's web consultation. For practical easiness, the browsing traces and consulted documents are centralized and collected from an intranet proxy-cache. From the rough traces, we preserved only the consultations dealing with textual documents (HTML, txt, pdf, Doc., etc). This correspond to 451 935 requests from 245 users. We then fetched and analyzed the corresponding documents in order to extract for each one the 10 most frequent keywords. Finally, the source data of our study are the following: A node identifier, a temporal indicator of data accesses, an identifier of the accessed document (URL) and finally, the words extracted from the document.

In order to validate our approach, the first question that we could ask ourselves could be: is there a correlation between a keywords profile and real consultations? More simply, up to what level, 2 nodes having a similar profile consult the same documents. This is far to be obvious since a same set of keywords may corresponds to a lots of different documents. This can be shown easily by questioning a search engine with a set of keywords. In our experiment each profile is a vector of 84800 words valued according to their occurrences. In this study we preserved the original profile size in order to facilitate the demonstration but the impact of the profile size on the efficiency of the measure was explored in a previous study [16] and showed that it was possible and even desirable to reduce the size. In basic cases the size of the knowledge sum up of each node is about 50 kilo Bytes.

The following curve shows a correlation between the profile vs. the real consultations of 70,1 %. In this figure, each point corresponds, in a log vs. log (log) reference, to the value D (semantic distance in y-coordinate) and R (redundancy of consultation in x-coordinate) for 2 nodes.

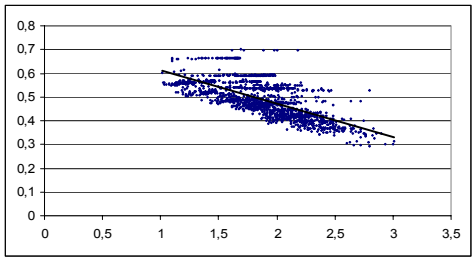


**Figure 2: Correlation between semantic distance and consultation redundancy.**

As shown in the followings formula, the Euclidean distance computes the semantic distance between 2 nodes. In this formula, *P2i* is the occurrence value of the keywords profile *i* of the user *2*. The redundancy of the consulted URLs between the nodes 1*(u1)* and *2 (u2)* is calculated in the second formula. R thus corresponds to the ratio between the numbers of shared consulted URLs by the 2 users divided by the number of URLs consulted by at least one of the 2 nodes over the reference period.

$$D = \sqrt{\sum_{i=1}^{i=n} \left(p1_i - p2_i\right)^2} \qquad R = \frac{u1 \cap u2}{u1 + u2 - (u1 \cap u2)} = \frac{u1 \cap u2}{u1 \cup u2}$$

The figure 2 shows that when 2 users browse the Web, the consulted documents tend to be the same when the 2 users have a close profile. This means that the thematic profile is a reasonable indicator describing the activity of each user taking into account the fact that each user is unaware of the consultation of the others. If this is true for users this is a fortiori true for network nodes that can be better controlled than human behaviour.

## 3. Keywords organizational capacity

However, in order to be useful within a framework of networks organization it is necessary to show that the correlation, studied in the previous section, remains positive after the process of organization for futures

consultation accesses. Thus, having used the traces of the first period to organize the network, we must check on the following period that the members of an associated network (i.e. nodes having close profiles) consult more often the same documents than a random nodes community.

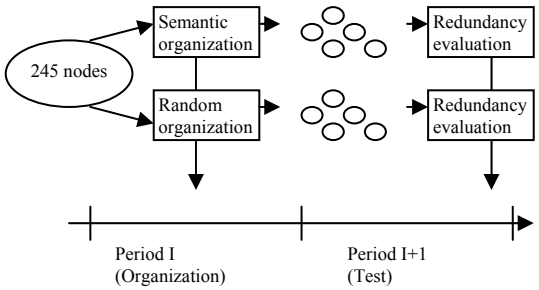


**Figure 3: Experimental method**

Thus, referring to the preceding figure, the identification of groups realized over the period I is validated over the period I+1, for all the test periods. We first need to modify the preceding formula (R) that is well adapted to evaluate the redundancy between 2 nodes but difficult to use in a more populated group. The new access redundancy rate is thus calculated as follows in a period for the entire group accesses:

$$Tr = 100.\frac{\sum_{d=1}^{d=n}(accesse(d) > 1)}{\sum accesses}$$

This formula expressed as a percentage, evaluates the ratio between the sums of all accesses to documents (d) which were reached more than one time divided by the sum of all accesses. This metric is rather traditional. It is used, for example, to evaluate proxy-caches efficiency (Hit rate) and computes the probability that a user request can be satisfied by the cache (previous stored accesses). Higher this value is, faster will be users' accesses since the downloaded objects have a higher probability to be local.

While the preceding study (figure 2) was done within all the period of 17 month, we compute now the profile of each node within a given limited period. Then, the nodes profiles are provided to a clustering process named CLUTO [15] that identify the most semantically homogeneous groups on the basis of the proximity of the 245 profiles. Then, we randomly select groups with a population equal to semantic groups. Finally, we check within the following period that for each group, the consulted data are indeed shared better (i.e. more accesses redundancy) than the accesses made by the random groups. We have remade this experiment for 7 different period durations (3, 5, 7, 9, 11, 13 and 15 days) to check the impact of this parameter. For example, the profile within 2 days represents the most frequent words included in the consulted documents during these 2 days. In all the tests, the number of groups was fixed to 5 in order to make clear the study but the impact of the number of group was previously studied [16]. The following table shows, for the 7 reference periods, the redundancy (Tr) average and standard deviation for the 5 organized groups compared with the random groups.

**Table 1: Redundant access rate compared between a semantic and a random regrouping.**

| | **Semantic** | | **Random** | |
|---|---|---|---|---|
| period (days) | Average (%) | std | Average (%) | std |
| 3 | 40,2 | 15,0 | 33,8 | 11,3 |
| 5 | 45,3 | 15,8 | 34,5 | 15,3 |
| 7 | 47,4 | 16,1 | 39,4 | 14,3 |
| 9 | 48,6 | 14,4 | 41,2 | 12,3 |
| 11 | 51,4 | 13,7 | 45,1 | 10,5 |
| 13 | 52,7 | 13,7 | 47,5 | 7,9 |
| 15 | 52,9 | 13,4 | 47,9 | 7,1 |

For a given duration, the first value represents the average within 30 successive periods for the 5 groups (average of 150 values of the redundancy rate). For the random groups we carried out the same test (random segmentation of 245 nodes into 5 groups sized as in semantic groups) but for a more statistical realism, we remade this average for 20 random regrouping (i.e. average of 3000 values of rate of redundancy).

This table permit several remarks. First of all, the higher level of information shared involves that the semantic organization is more effective than the random groups. We have always, at least, an improvement of 10 % and at best 30 % between the 2 modes of regrouping. It is clear however that the efficiency is not the same one for all the period sizes. For the 2 kinds of organization (random, semantic) the performance rises according to the period duration while the gain in performance for the semantic organization decreases according to the period duration. This is not surprising because the number of accesses grows according to the period duration, which results on a higher probability of accesses redundancy, even for a random regrouping. In our experiment a period size of 5 days seems to give the best gain between semantic and random regrouping. We also observe that the dispersion variations (std) between the 2 modes of segmentation strongly rises according to the period size with a minimum of std corresponding to the maximum efficiency (5 days).

If it is clear that the data related to the semantics of accesses are significant to organize data networks, we could wonder in what way the granularity of these data influences the organization. To evaluate this point, we tested the organizational capacity of profiles based on URL compared to keywords. The nodes profiles thus consist respectively in a vector of URLs valued by its level of consultation and a vector of keywords as in the previous case. It appears that indeed the granularity of each type of data has a consequent impact on the organization efficiency.

We see for example that with keywords we have 39 groups (13 % of all clusters) that have only one user whereas there is 261 groups (93 %) with a URL based segmentation. We also see that the average number of nodes per groups is very limited, almost 1, with the URLs that make this kind of segmentation not very usable.

## 4. Strategy for networks organisation

We could wonder if the stability of such organisation is enough. Indeed, profiles changing too rapidly could cause oscillations with frequent successive connections and disconnections. In order to evaluate this point, we investigate the organisation capacity of several method of profile computation depending on the nodes behaviour and on chronological criteria. The chronological criteria involve to compare long-term (e.g one month of activity) and short-term nodes profiles (e.g one day). Evidently, a long-term profile that integrates in a unique set the most frequent keywords within one month of activity is more stable than a short-term profile. This kind of profile is used as a thematic reference since it can filter and eliminate short-range variations. The criteria based on nodes behaviour is also used as a reference. Here, a group of nodes is seen as unique node, and a profile becomes an aggregation of most frequent keyword consulted by the group. Each of these references (group and long term profile), can be used to drive the organisation of the network. In order to evaluate the results of these modes of organisation, we present, here after, 2 experiments based on 150 nodes activity.

As we said, one of the criteria to select a group member is the level of the node proximity compared to the existing group. The idea is to compute the profile of the entire potential group and to discard nodes that have, for example, a distance less than the average. The following figure shows for our 150 nodes the repartition of the distance classes. The figure 4 shows long-term profile of nodes according to the group long-term profile and, the figure 4 shows the short-term node according to the short-term group profile. We can observe that the repartition based on long term profile allows having shorter distances and consequently more stable groups.

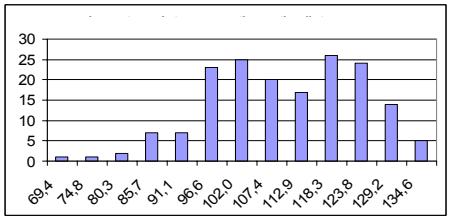


**Figure 4: Repartition of distances between long term nodes profile**.

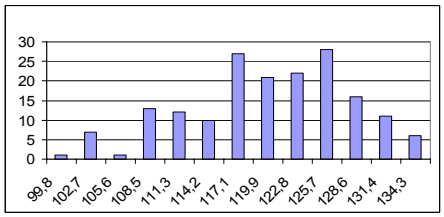


**Figure 5: Repartition of distances between short term nodes profile.**

As an example we could use the average of the distance as a tuning value in order to take a decision on the acceptation or not of one node. Actually, this value is a key parameter influencing the network dynamic.

**Table 2: Accepted nodes in the network depending on distance tuning parameter**

| % max distance (dist. Value) | 50 % (70) | 60 % (84) | 70 % (98) | 80 % (112) |
|---|---|---|---|---|
| Short term | 0 | 0 | 0 | 33 |
| Long term | 2 | 7 | 44 | 89 |

The above table shows for 4 values of the maximum distance percentage the quantity of nodes accepted in the network for the short term and long term node profile. The maximum of distances for both profiles is to 140 and the minimum is zero involving that two profiles are identical. Wee see that the majority of the nodes are concentrated on high distances if we use short-term profile, This involve that the probability to obtain several nodes having the same level of low similarity is higher with short-term profile than with long term. Having several possible choices for each connexion decision, could increase the risk of having too frequent oscillations that will perturb the architecture. It is clear that the condition for having stable groups is to take into account long-term profiles. However, it is not obvious that one month is the better period. A balance should be done between a good

stability (long term profile) and realism (short-term profile better describe present activity).

This is illustrated by the following figure that shows the repartition of distances when comparing short-term according to long-term profile. The x-axis represents the class of value of the Euclidean distances between the short-term and the long-term profile of one node. The y-axis represents the number of nodes having the corresponding distance. For example, we have 20 nodes (on 150) that have between 67.3 and 77.9 between their short-term and their long-term profile. A short distance can be interpreted as a level of stability in interest topics that we could expressed as: "In this period I am mainly interested in what I am usually interested".

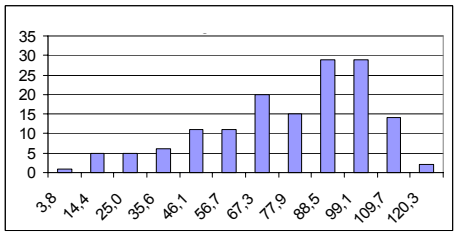


**Figure 6: Repartition of distances between long term and short nodes profile.**

We see that only 25 % (lower than the average) of the nodes have a coherent thematic activity over the time. This criterion can be analyzed as a measure of individual behaviour and can be used as complementary to decide if a node should be added to a group, in order to favouring stable organisation of nodes. An important point regarding p2p networks is that it is very difficult to predict their stability, since nodes may appear and disappear without notice. The following moving average indicator measures within the time the level of stability of the nodes association.

$$S_i = \frac{1}{2}(S_{i-1} + \frac{New_i}{Total_i})$$

The S value within the period i, computes the average between, the rate of new nodes and the value of S within the prior period. The evaluation of this indicator can help to compare networks behaviour and take a decision since a too high rate means that resources may have a very low level of availability.

## 5. State of the art

As an interdisciplinary work, our study relates to topics which are usually addressed by different communities. Indeed the capture and the management of user profiles are traditionally studied by the KDD community (Knowledge Discovery from Data) [3] which has a good experience on data-mining and linguistics basic techniques (lemmatisation, etc.) that allow to build effective profiles with limited resources [4]. These kinds of techniques are also used in the context of contents indexation and information research [5] but also in collaborative filtering techniques [6].

Even if they were used till these past years in a centralized context (e.g. request to a search engine) this kind of techniques were applied more recently within distributed environments like resources localization in P2P networks, cache architectures or CDN. The researchers considered a passive approach which consists in locating distant contents [1] or a more active approach which aims at moving contents close to users or to modify connections between users and data sources [2]. For example, P. Rochat et al. [7] suggest taking into account the cultural influence of users to optimize proxy-caches architectures performances. The basic postulate is that there is a strong link between words contained in consulted URLs and user's topics of interests. Wang et al [19] analyzed the role of social groups in ephemeral group management. They proposed a model and experimentation showing the relation between social groups and computing device. Ahmed et al [20] proposed a smart meeting room with pervasing technology. They proposed a method aiming at managing participant's context information, knowledge usability and ephemeral group communication.

Following the same goal, J. Gwertzman [8] takes into account geographical criteria as well as the user navigation path in order to push information in proxy-caches. This approach is near that of prefetching which aims at anticipating the move of contents in proximity of users. Prefetching techniques can be based on analyses of URLs ([9]) or of keywords [17][10] contained in previously consulted documents in order to foresee the future activity. Algorithms based on categorization [11], training [12], decision trees [13], or Markov hidden chains [14] can be used in order to obtain decision rules to identify users' future activity.

## 6. Discussion

This study showed us that a semantic descriptor is relevant and allows an effective organization of ephemeral data networks. Indeed, it is clear that nodes linked on semantic proximity criteria, have a stronger probability to be provided and to share contents in an effective way. We showed also that the granularity of the descriptor had a high influence and that the vector of valued keywords was more effective than URLs.

However, these positive results should be balanced since, in our study we assumed that the user's behaviour regarding textual documents could be transposed with non textual one. In other words, we could wonder if close keywords profiles mean high redundancy on non textual objects (e.g. video, music, etc.)? This was not verified in our work and would be interesting to study in order to validate our concept in a more general way.

Taking into account the fact that we wish to approach the semantic organization of data networks from a general point of view applicable to various technologies we did not studied deployment and scaling problems. One other topic to consider is the computation time. Nevertheless, since real time organisation is out of the scope this computation delay is not really a big problem.

Within the framework of P2P networks, it is possible to optimize contents research by giving a better priority to nodes semantically close instead of random, geographical or accessibility criteria. In this context, we developed the concept of active mirrors [18]. These devices automatically reuse the previous download of users in order to propose interesting contents to a large community. This approach deals with implicit collaborative filtering and collective intelligence. The first work showed that the active mirror allows boosting the research of interesting documents (compared to a regular search engine). This study allows to automatically interconnecting active mirrors in order to share interesting data in a larger way.

## 7. References


[1] Edith Cohen, Amos Fiat, and Haim Kaplan. Associative Search in Peer to Peer Networks: Harnessing Latent Semantics. Proc. IEEE INFOCOM 2003, Apr. 2003.

[2] K. Sripanidkulchai, B. Maggs, and H. Zhang. Efficient Content Location Using Interest-Based Locality in Peer-to-Peer Systems. Proc. IEEE INFOCOM 2003, Apr.2003.

[3] Knowledge Discovery & Data Mining fundation http://www.kdd.org

[4] S.T. Dumais. Using LSI for information filtering: TREC-3 experiments. In Proc. of the Third Text REtrieval Conference (TREC-3), 1995.

[5] Hu, W / Chen, Y / Schmalz, M S / Ritter, G X An overview of the World Wide Web search technologies, In the proceedings of 5 th World Multi-conference SCI2001.

[6] A. Moukas. Amalthaea: Information Discovery and Filtering Using a Multi-Agent Evolving Ecosystem. Int. Journal of Applied Artificial Intelligence,: 1997.

[7] Rochat, P ; Thompson, St (1999) Proxy Caching based on object location considering semantic usage, in proceedings of Web caching workshop

[8] Gwertzman, James (1995); Autonomous replication; Senior Thesis Harvard university;

[9] Y. Aumann, O. Etzioni, R. Feldman, M. Perkowitz, T. Shmiel. Predicting Event Sequences: Data Mining for Prefetching Web-pages. In Proc. of the Int.Conference on Knowledge Discovery in Databases (KDD'98), 1998.

[10] B. D. Davison , Predicting Web Actions from HTML Content: Proceedings of the The Thirteenth ACM Conference on Hypertext and Hypermedia (HT'02)

[11] S. H. Kim, J. Y. Kim, and J. W. Hong. A Statistical, Batch, Proxy-Side Web Prefetching Scheme for Efficient Internet Bandwidth Usage. In Proc. of the Network+Interop Engineers Conference, Las Vegas, May 2000.

[12] Q. Yang, H. H. Zhang, and T. Li. Mining Web Logs for Prediction Models in WWW Caching and Prefetching. In the 7 th ACM SIGKDD Int. Conference on knowledge Discovery and Data Mining KDD'01, USA, August 2001.

[13] T. S. Loon and V. Bharghavan. Alleviating the Latency and Bandwidth Problems in WWW Browsing. In Proceedings of the USENIX Symposium on Internet Technologies and Systems (USITS'97), December 1997

[14] R. R. Sarukkai. Link prediction and path analysis using Markov chains. In Proc. of the 9 th World Wide Web Conference, 2000.

[15] Cluto Clustering software package http://www-users.cs.umn.edu/ ~karypis/cluto/

[16] L. Lancieri, N. Durand, Evaluating the Impact of the user profile dimension on its characterization effectiveness. In proc. of Int. Symposium on Computational Intelligence for Measurement systems and applications. (CIMSA2003)

[17] L. Lancieri, N. Durand Activity forecast of the Internet users based on the collective intelligence, the IASTED International Conference on Artificial Intelligence and Application (AIA 2004); Innsbruck, Austria.

[18] L. Lancieri, Reusing Implicit Cooperation, A novel approach to knowledge management;, In tripleC (Cognition, Cooperation, Communication) International Journal, 2004 pp 28-46; ISSN 1726-670X

[19] Bin Wang; Bodily, J.; Gupta, S.K.S.; Supporting persistent social groups in ubiquitous computing environments using context-aware ephemeral group service;. Proceedings of the 2 nd Conference PerCom 2004.

[20] Ahmed, S.; Sharmin, M.; Ahamed, S.I.; A Smart Meeting Room with Pervasive Computing Technologies;. Sixth International Conference SNPD/SAWN 2005