A connectionist approach for evaluating the complexity of interaction in the World Wide Web
The case of News groups

luigi.lancieri@cnet.francetelecom.fr
France Telecom – Centre national d'études des télécommunications (CNET)

The increasing success of Internet is a great opportunity for business activities but also for interpersonal communications. Several researchers studied the interesting field of social interaction [ROS, HEY] through the net. One of the big difficulties to have a good view of human interaction is to define suitable tools and metrics to represent these characteristics. The goal of this paper is to going ahead in this field and more specifically to evaluate the complexity of the knowledge interaction on collaborative groups.

One of the first difficulties to achieve such studies is to find data correlated with the different level of interaction. It is easy to find in HTTP server's or caches log files data that inform on the activities of users: for example retrieving the top 10 more frequented site or extracting the thematic or behavioral profile of users[LAN99]. What is very difficult however is to have a view, in addition of what users get from the network, what they "give" to the network. These 2 actions (get and give) are the basis of the interaction. The evolution of the network is directly driven by human knowledge interaction. As in direct human interaction people are more or less easily influenced by the information they receive (consciously or not). In a group of people some are more or less leader on giving to the group. The knowledge of the group is a complex result of all this kind of interactions. Even the quantitative interaction (global flow of data) in the global network found its roots in the local human's interactions.

As we said before, it is difficult to study this kind of phenomena because of a lack of information related to these interactions. It is true in the field of sociology but also regarding the network. Regarding the evaluation of what users gives to the network, one of the most interesting service to study is the Usenet news groups. The study of collaborative working group like Usenet news can also be interesting in smaller electronic working groups as in company's Intranet.

In this context, we present a method to evaluate different forms of influence within and between discussion groups. This method involves a basic linear associative neural network which is an interesting tool to learn and show long term cross influence between groups.

Method

We measure, first, the activities of each participant of the group. For that, we compute in a period of time the quantity of information (messages) each subscribers sent to a given news group. The statistic study of the resulting frequency distribution allows to extract, in the same way as the principle of Zipf law [BRE, MAN, ALM], the characteristics of the interaction on the group (see below).

The main hypothesis made in this study is that a person that gives information to the group is supposed to receive also information. This is not always true because one can send a mail to a group one time and then never consult the group again. However, as in Zipf law this kind of situation is automatically under considered, this limits the risk of errors. At the contrary, people that are active in a group are by force "listening" the group. The main reason of this is that, as in real groups, people talking without first listening what others have to says are rapidly excluded from the group. The computation of the interaction characterization of the group is very easy, as the following plots show. Zipf state that series of data following his law should be a straight line after log(x)/ log(y) transformation.

$$y = \frac{1}{P^a} \qquad \log(y) = -a\log(p)$$

0-7695-0619-4/00/$10.00 (C) 2000 IEEE

The ICC (Interaction Complexity coefficient) is the factor "a": the slope of this characteristic. To compute this coefficient, we must, first rank users according to the number of messages they sent. It is interesting to note that this slope is directly correlated with the level of (knowledge) influences in the group. This is more evident if we consider extreme case of influences. The following figures show two cases of influences. Fig 1 describes the case in which the leadership is distributed on one or few persons in the group, most of the people listen (sending of few messages: questions for example) and few speak (send messages or answer to questions.). This case can be compared to a classroom in which, even if students speak, the teacher has the leadership in knowledge diffusion. In the second case (fig 2) the influence is widely distributed among the group, most of the subscribers send a comparable amount of messages. This is comparable to some expert forums in which most of the participants speak and share knowledge.

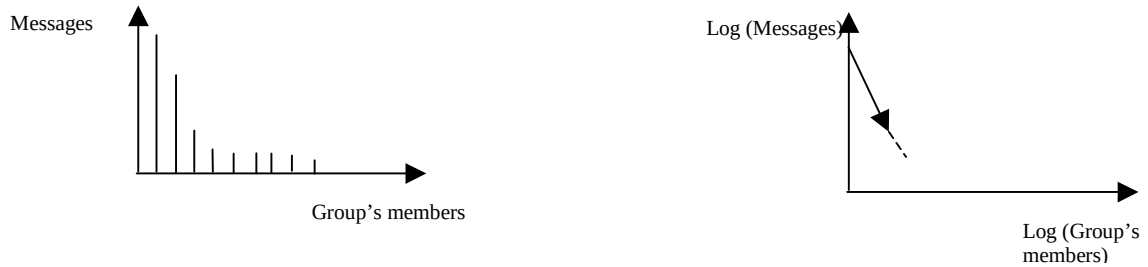

Fig 1:Low level of interaction

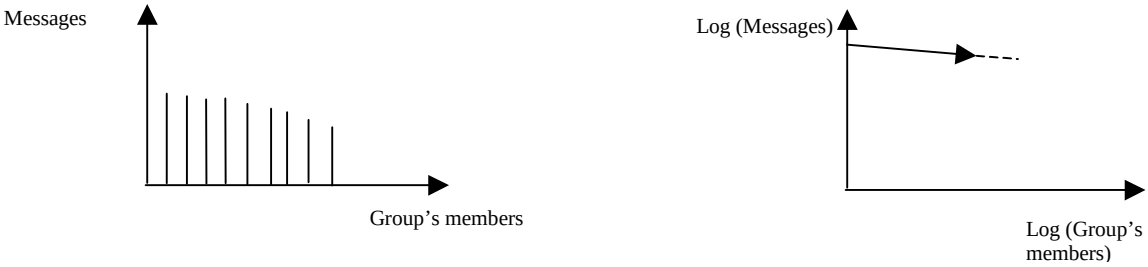

Fig 2: High level of interaction

In application of this brief introduction, the following plot shows the differences between the characteristics of 2 news groups (alt.os.linux and intel.inbusiness). We see that the slopes of the 2 curves are different (respectively 0,6 vs 2,5). The most vertical slope corresponds to the intel.inbusiness group where most of the messages come from limited sources.

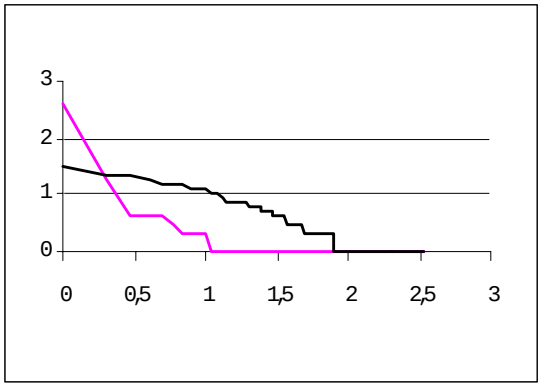

Fig3: Comparison between 2 news groups

A more complete study should be interesting to achieve with a larger set of data (only 500 messages per group here). In addition, it also possible to compute, for each group, the time evolution of the ICC. The resulting plot allows following the «life» of groups, comparing groups and predicting their evolutions. The variability of ICC can inform us on the dynamics of the influence inside groups. Some of them could evolve from a monolithic influence to a plural one.

Evaluation of cross influence

We try now to compare the ICC of large amount of groups. It is interesting to observe that the law defining the level of interaction within a group (ICC) is the same as the one defining distribution of ICC between groups. The following plot shows the Zipf distribution of ICC (y axis in log scale) among 640 (x axis in log scale) new groups (alt.binaries.cracks,…, fr.petites-annonces.immobilier).

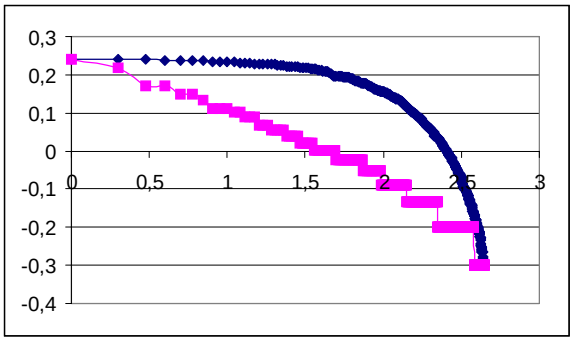

Fig4. Distribution of ICC for 640 groups

We represented on the same plot the ICC (line) and a random distribution (curve) in order to confirm that ICC is valid and is not coincidental parameter. The fact that internal and external ICC follow the same law tend to confirm the idea that both interactions are linked by a self similarity (fractal) relation. We call ICCg a global parameter involving a class of groups (a group of groups). It can help for example a manager to follows “social” dynamics of groups or communities of interests.

To investigate more deeply external influence of the groups, it would be interesting to see if specific people participate in several groups. We can image two extreme cases where all people participating in a group A also participate in a group B, may be with a different ICC or where people participate only in one group . The first case denotes a high and a second a null level of cross influence. The following table shows an example of cross influence between 6 news groups. The value indicates how much messages were sent to a group from another group users. For example, 8 messages were sent to the Linux group from members of the humor group.

	games	Humor	linux	Intel	java	amiga
games	-	5	2	1	1	0
humor	-	-	8	1	3	5
linux	-	-	-	1	1	1
intel	-	-	-	-	0	0
java	-	-	-	-	-	2
amiga	-	-	-	-	-	-
total	9	22	13	3	7	8

The total indicate the number of messages sent to a group from members of other groups. For example, 22 messages were sent to the humor group from members of other groups. This table is no more than an example to explain the method. For several reasons, interpretation should be handle with care due to the limited amount of data colected here. Each group have around 1000 messages on a 2 week period. To be reliable such computation should be done on a regularly collected data and statistic basis.

This statistical long term analysis can be imbedded in a linear neural network where each cell is a news group and where links between groups are regularly increased according to the level of cross relations (number of messages sent to a group from members of others groups). The value of the links are normalized and stored in a connection matrix. This matrix will be very helpful to evaluate the level of indirect influence. Since direct influence between group A and B is reported by the value of the link AB, the indirect influence is a little bit more complex. Let suppose we have 3 groups A, B and C, the indirect influence from A to C is the part of AB influence transmitted to C. We see that the global external influence (GEI) from A to C is higher than the sum of direct influence (DI) from A to B and the direct influence from B to C.

$$GEI(AC) \geq DI(AB) + DI(BC)$$

Perron and Frobenius [BER] have used the graph theory to solve this kind of problems. They have showed that the « iterated powers » of the matrix C associated to the graph (the same that the connectivity matrix) associates to each top a value which corresponds to the « influence » of this top in the graph. These values can be computed rapidly with analytical numeric methods.

$$p^{i}(k) = \sum_{j} Cij.\, p^{j}(k-1)$$

This formula computes the order *k « iterated power »* of the node (neuron or news group here) *Nj*. The order 1 « iterated power » corresponds to the sum of the matrix C lines. Increasing the value of *k* allows to take into acount indirect influence. By computing, for example, the k=2 iterated power for the group j (i.e the j component of the vector p) we involve the h groups directly connected to group j and the m groups directly connected to the previous h groups. In order to determine the overall best influence (GEI) and to classify the nodes, it is interesting to compute the « relative influence » of a specific *i* node by the following ratio.

$$IRi = \frac{p^{i}(k)}{\sum_{i} p^{i}(k)}$$

Perron and Frobenius have showed that for all the nodes, this ratio converges to the Eagan vector of the C matrix. These nodes are called the leader nodes.

Related work

There are several studies related to people's interactions in the Web. In this field, the Usenet groups are very interesting because it is a good reflect of social exchange. J.S Donath [DON98]studied the notion of identity and the impact of identity deception in the context of Usenet groups. She also developed [DON95] a tool that allows to create new Web interactions. This tool allows a Web surfer to discover who is looking on the same Web page and to communicate with him. On the domain of interactions in social groups D. Plewczynsky [PLE97] studied the Landau theory of social clustering. He proposes a cellular automata model of social influence. He highlights that the influence of a group of individual on a given person is proportional to 3 main factors : The strength of the members of the group, their social distance from the individual and their numbers.

In the field of rating and filtering information, the Usenet was the subject of several studies. H. Sorensen et al [SOR96] developed a profiling system called PSUN which extracts the user profile from written articles. The profile can be used to filter new article on a semantic basis. F. Kilander [KIL96] made a comparison of a dozen of news filtering tools that allow to give a good idea of the state of the art in this field. We also perform study [LAN98] in the domains of characterizations of Web information and interactions. We presented a general model to compare thematic and behavioral profile of Web User by using the easiness allowed by Web caches technology.

We show in our last paper that behavioral profile [LAN99] of specific Web user also follows, as the internal group and external group influence a Zipf law. The use of this law in general macroscopic networking model is very well known. V. Paxon et al [PAX95] paper show that packet arrivals of NNTP protocol (news group) best fit with self similar distribution rather than the generaly accepted Poisson distribution.

Conclusion

The method and the metrics we described can be used to understand and manage of large amount of collaborative groups like Web community of interest or company forums. Some of these groups are self built and free other are more supervised but there is a lack of tools to easily follow their activity.

The metrics we present associated to thematic metrics can be used in a company to reduce the amount of groups. Indeed the redundancy of groups is very difficult to follows as the numbers of groups grows. In the Usenet hierarchy, for example there is more than 20 groups on the Linux subject (except language differences) .

Acknowledgment

I Wish to thanks both of my colleagues B. Morin and J.P.Foucault for their help to collect and process news group data from the Web.

References

[BRE] On the implication of Zipf ‘s law for Web caching; L. Breslau, P. Cao, L.Fan, G.Phillips, S. Shenker

[MAN] Les objets Fractals; B. Mandelbrot; Edition Flamarion

[ALM] Characterizing reference locality in the WWW; V. Almeida, A. Bestravos, M. Crovella, A. de Oliveira

[HEY] The social superorganism and its global brain; F Heylighen; Principia Cybernetica Web; http://pespmc1.vub.ac.be/SUPORGLI.html

[LAN98] Luigi Lancieri : Distributed Multimedia modeling ; In proceedings of IJCNN98 (IEEE International Joint Conference on Neural Network)

[LAN99] Luigi Lancieri ; Description of Internet User Behavior ; In Proceedings of IJCNN99 (IEEE International Joint Conference on Neural Network)

[ROS97] Joël de Rosnay ; L’homme symbiotique ; Editions du Seuil

[BER] La théorie des Graphes (Graph theory) Claude BérgéDunod

[PAX95] Vern Paxson, Sally Floyd; Wide area traffic: the failure of Poisson modeling; university of Berkeley California July 1995.

[KIL96] Frederik Kilander; A brief comparison of News Filtering Software; Stockholm University; June 96

[PLE97] Dariusz Plewczynski; Landau theory of social Clustering; Institute of Social study Warsaw November 1997.

[SOR96] H. Sorensen, M.Mc Elligott; PSUN:a profiling system for Usenet News; Cork university college.

[DON95] J. Donath; sociable information space; IEEE workshop on community networking; June 1995

[DON98] J. Donath; Identity and deception in the virtual communities; In M.Smith and P.Kollock (eds); 1998