Usages Analysis in Instant Interpersonal Communications over IP

Alexandre Bouchacourt, Luigi Lancieri
France Telecom R&D
42 Rue des coutures 14000 Caen France
{alexandre.bouchacourt, luigi.lancieri}@orange-ftgroup.com

Abstract

This paper investigates the contribution of data mining techniques in order to optimize IP based real time services such as Instant Messaging or Telephony over IP. The main difficulty is here to handle temporal constraints and to interpret the corresponding raw data into users' behaviour high level knowledge. As an example, we focus on IM and try to find a light mean to detect conversations consistent from the human point of view (i.e sum of several basic exchanges).

1. Introduction

The study of WEB usages is quite common nowadays and offers a better understanding of users' behaviors. This is usually done by extracting and analyzing traces produced on servers. Now that telecommunications are shifting to Internet Protocol (IP) similar investigations can be carried out in that field, easier than before since data networks offer easy access to a large quantity of traces from the users' activity. Several applications already benefit from such transparent user feedback. Let us take for example the case of collaborative filtering or CRM (Customer Relationship Management) used in online trade. Recently, new opportunities appear with multimedia and real-time services. Telephony over IP (ToIP) and Instant Messaging (IM) are mainly concerned since most recent communication tools include voice and text media.

From a conceptual point of view, traces can be seen as the human memory with semantic and episodic mechanisms. Indeed, traces result from the human activity but also influence it. As the human memory is a result and influences the activity of a person, traces store and influence the interactions between individuals. This is true even if computer based devices are involved in the interactions. This reciprocal influence can be observed in many services based on the exploitation of traces. We think that such mechanisms are keys in collective intelligences (see [9] for more details on conceptual framework and practical examples of traces analysis in data networks) .

If the mining of users' traces is a promising scope of research it also has its complexity. The main goal of traces analysis which is to map raw data on high level knowledge is not always easy, especially if we need to take into account the time line of the users' activity or its dynamicity. These constraints are omnipresent in real-time services.

In this paper, we investigate temporal aspects of traces feedback. In order to do that we wish to detect conversations, in other words the beginning and end of coherent human exchanges. A conversation is composed of sequential messages between two users. Voice conversations are easy to detect. Dates (beginning and end) are immediately accessible thanks to the protocol which gives non ambiguous information (hang up, intermediate session details, etc). Unfortunately this is usually much harder as for IM where the basic element is a message. Actually, sessions and conversations are most of the time different since IM sessions often match one only message. In that case we have no direct mean to date the beginning and end of conversations, since they could be composed of several sessions. Thus, we suggest a method to detect conversations in IM. Several criteria are discussed, based on time between consecutive messages (TbM), "accelerations" and the sizes of messages.

We first describe some related works and then detail our approach. Finally we discuss some results and possible applications.

2. Related works

The analysis of traces is a major field in science. Actually, many phenomena from physics to computer science cannot be observed as they happen, due to a short observation window for instance. However these phenomena sometimes produce traces whose collection and analysis help study them.

In computer science and more especially in data networks traces have been used from ages in order to monitor the behaviour of devices or to locate possible failure. Some advanced works in these domains have led to model used for example to understand network bottleneck [11 ,12]. From usage perspective, the WEB mining has met a great success. Not only has it enabled to understand the users' behaviours better but it has also proved useful to suggest new services adapted to users and context. Several studies [1, 2, 9, 10] address that issue mainly from the semantic point of view and more rarely in relation with episodic aspects [3]..

Different methodologies have been proposed in order to extract higher level knowledge from raw traces. Outside general statistical approaches, other methods based on the identification of leading structure in the data (Principal component analysis, neural networks, etc) were used in order to identify imbedded models. For example, some authors suggest the use of ILP (Inductive Logic Programming) to produce general rules (new knowledge) from examples and observations (users' logs)[4, 5, 6]. Such extracted knowledge can help us to have a better understanding of usages but also to enable new services which could adapt to users' behaviour in real time.

Instant messaging is a quite recent media and has not been much covered yet. Most studies focus on social aspects and for example analyse interactions in an IM population and try to reconstitute human networks [7, 13, 15]. Electronic mail has received much more attention since it is everywhere (see [14] for a voice mail related study). Text mining tools are used to classify mails. However email and IM are not the same. The first is an asynchronous media whereas the second is supposed to be synchronous (like telephony is). Therefore the pace in instant messaging is most interesting. Some works investigate that issue and classify users according to the rhythm of their conversations [8]. The detection of conversations (when they begin and finish) when the only available dates are those of messages taken individually is not addressed though.

3. Methodology

The material on which this study is carried out is an IPDR (Internet Protocol Detail Record)[16] corpus. The corresponding normalized format is often used for billing. Thanks to an XML structure it gathers information about instant communications over IP such as dates, identities, various sizes, error codes and so on. However it does not register the content of voice or IM text conversations. The following schema is a simplified SIP (Session Initiation Protocol) architecture that collects IPDR traces (we did not represent SIP proxies that are responsible for forwarding SIP signalization):

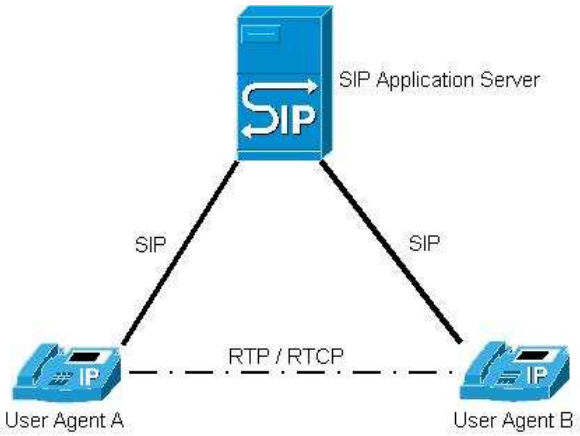


Figure 1: a simple SIP architecture

In that simple case, if User A and User B wish to establish a call, they first exchange SIP signalisation with the SIP application server on which an accounting module can be set up. Once the call is established users exchange audio data via Real Time Protocol and Real Time Control Protocol (RTP/RTCP). In a wider context SIP also manages the integration of voice and text oriented application (e.g. telephony and instant messaging)

In the following pages we try to find a criterion to identify conversations. We define a conversation as a temporally coherent collection of messages exchanged by two speakers. We first describe and justify a new intermediary format, and introduce definitions in the meantime. Then we discuss notions like time between messages, length of the longest conversation and acceleration to make our hypothesis more accurate.

3.1. IM traces and formats.

These IPDR are made of records. A record is a collection of traces that corresponds to an event (the

sending of an IM from user A to user B for instance). Records are sorted by date so that IM and telephony records follow each other within the same file. Since we focus on IM, we first filter the corpus by keeping only IM records.

Each IM record is comprised of two identities, namely the sender and the receiver of the message, its date and its size in bytes (more information are available but none is relavant for the purpose of this article). Moreover we want to argue about conversations (between two speakers). In that end we produce as many files as there are couples exchanging through IM, according to the following format:

 # UserAUserB.log

$Date_1$; $Size_1$; Way_1

...

$Date_N$; $Size_N$; Way_N

$Date_i$: date of the i^{th} message
$Size_i$: size of the i^{th} message in bytes.
Way_i: "+" if the message was sent from UserA to UserB, "-" otherwise.
N stands for the number of IM exchanged by UserA and UserB over the length of the study.

From then on we define the time between two consecutive messages as follows:

Def 1: *let $Date_i$ and $Date_{i+1}$ be two consecutive dates. The corresponding time between these messages is given by $TbM_i = Date_{i+1} - Date_i$*

To make things easier, we only consider cases for which Way_i and Way_{i+1} are different. In other words we define an exchange as a set of two consecutive messages with a change of roles. In the following we study the influence of the size of these messages in the user's interaction. We study 3 cases of inter-messages size (S) corresponding to the size of the 1^{st} message (i) , that of the 2^{nd} message (i+1) and the mean of the two sizes. Consequently a new format from which we will extract conversations is a sequential file of messages inter-time (TbM) associated with an inter- message size (S) (3 cases)

3.2. Time between messages

We first calculate the distribution of these TbM following 2 temporal perspectives. First, from a global point of view, we define 16 time classes from 0 to 15 minutes with a minute granularity (see fig 2). Then from a more detailed perspective, 18 classes from 0 seconds to 300 seconds with 5 seconds granularity. For each of these temporal perspectives, we finally calculate the corresponding sizes (S).

In the following figures the ordinate gives in the meantime the frequency divided by 100 for the TbM distributions and directly the sizes (S) in bytes for the curves. The 3 cases are represented in figures 2 and 3: (1) the size of the first message, (2) the average of the two sizes and the (3) size of the second message.

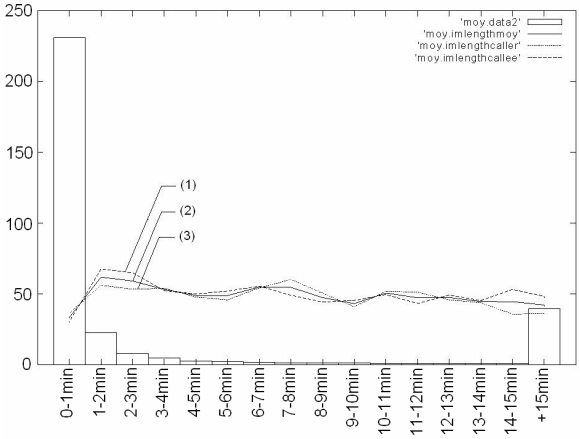


Figure 2: TbM distribution from 0 to 15 minutes compared with S evolution.

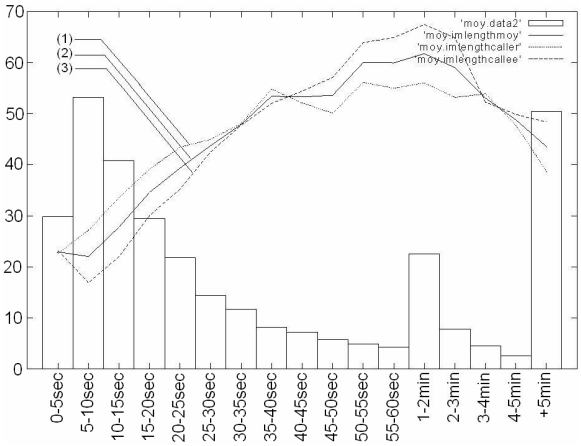


Figure 3: TbM distribution from 0 to 5 minutes compared with S evolution.

Distributions show that most exchanges are done in less than 1 minute (figure 2), and many take between 5 and 15 seconds (figure 3). Let us also make two general remarks:

- a small TbM implies that the second message (i+1) is small (since the typing speed is limited). However a small second message

does not necessarily mean that it was sent quickly (the user may want to delay).

- a long second message implies that the TbM be long (once again du to typing speed limit).

These a priori remarks are clearly confirmed by the previous figures. What is really interesting to notice is that passed 1-2 minutes the size of messages does not increase but on the contrary decreases (clear in figure 3). If the observation had been isolated we would have invoked temporization but since it appears on average we could explain it by the belonging of the second message to a new conversation (the first message of a conversation being usually a short introduction message). Evidently, it is possible to find 2 consecutive messages between two users separated from less than 2 minutes semantically unrelated (i.e. 2 distinct conversations) but the common sense makes us to think that this is a rare case. Consequently we advance the following hypothesis:

Hyp 1: *If a TbM is lower than 2 minutes then the messages belongs to the same conversation with a maximal probability that decreases as the inter-time rises.*

If the observations leading to the hypothesis 1 seem to identify the TbM where the probability of being in a conversation is maximum, we may want to identify the TbM where this probability is minimal.

3.3. Length of conversation

In the following, we propose to check the length (number of messages) of the longest conversation detected for each couple. Therefore we define a TbM threshold (TbMT) as follows:

Def 2: *If TbM < TbMT then both messages belong to the same conversation.*
If TbM >= TbMT then the 2nd message is part of a new conversation.

By varying this threshold we obtain several values as for the length of the longest conversation.

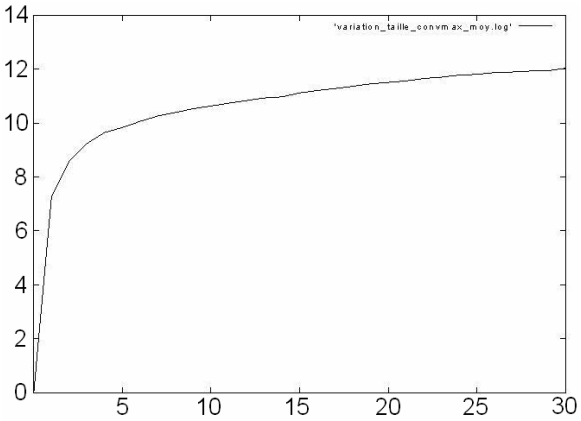


Figure 4: variation of the lenght of the greatest conversation averaged on all users

This curve not only confirms that the length we search increases as the threshold but it also highlight a plateau around 10 minutes. After 12 or 13 minutes that length keeps on increasing quite linearly. Therefore, we put the following hypothesis forward:

Hyp 2: *If the TbM is lower than 10 minutes then the messages may belong to the same conversation but with a minimal probability that increases as the inter time decreases. Otherwise the 2nd message is likely to belong to a distinct conversation.*

3.4. Pace

In order to check the strength of those hypotheses we thought about another partially independent criterion. Instead of arguing about time we would like to work with a sort of "acceleration" indicator to depict the pace of conversations. We define an acceleration indicator as follows:

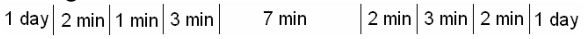
Def 3: *let TbM_i and TbM_{i+1} be two consecutive TbM. The corresponding acceleration indicator is given by $Acc_i = TbM_{i+1} / TbM_i$*

If Acc << 1 then $TbM_{i+1} << TbM_i$, which means the rhythm of exchanges quickens.
If Acc >> 1 then $TbM_{i+1} >> TbM_i$, which means the rhythm slows down.
If Acc ~ 1 then $TbM_{i+1} \sim TbM_i$, which means the rhythm remains almost constant.

A conversation will start with an Acc value very small (the pace quickens) and finish with a greater value (the pace slows down). Consecutive messages of a same conversation will induce values near 1. We take in account theses remarks and define a pace threshold Acct as follows:

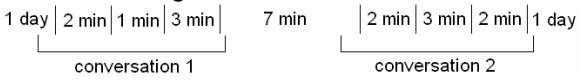
Def 4:
Acc <= 1 / Acct *is equivalent to Acc << 1*
Acc >= Acct $_t$ *is equivalent to Acc >> 1*
1 / Acct *<Acc<* Acct *is equivalent to Acc ~ 1*

By varying this parameter we detect several configurations of conversations. What follows illustrates the two methods of extraction (the first with inter times, the second with accelerations). The hypothetical conversation we study is comprised of 8 messages as follows:

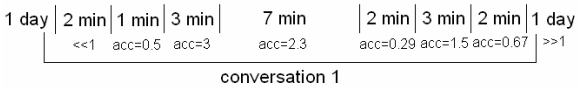


Each vertical bar stands for a message and the inter time between two consecutive messages is specified. The two 1 day values that bind the conversation are exaggerated on purpose.

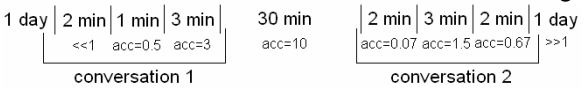
With an inter time threshold of 5 minutes for instance the first method gives:



With an acceleration threshold of 10, the second method gives:



Let us notice that if instead of 7 minutes we had had 30 minutes or more, then with the same acceleration threshold the extraction would have been the following:



According to the pace of conversations and the values of parameters for each method, extractions do not always coincide.

Once again if we argue about the length of the longest conversation to compare extractions on average we find that methods coincide best for TbMT = 6 minutes and Acct = 50.

If we assume that both methods are able to extract conversations accurately, we are tempted to venture that they will best coincide if their respective parameters are well chosen. We notice that this value (6 min) is near the middle of the boundaries proposed in the hypotheses 1 and 2, respectively (2 and 10 minutes).

There is no proof indeed but the 6 minute value is a too coherent one to be ignored considering the previous results. Therefore we propose this final hypothesis:

Hyp 3: *If the TbM is lower than 6 minutes then the messages* ***are likely*** *to belong to the same conversation, otherwise the 2nd message* ***is likely*** *to belong to a distinct conversation.*

4. Discussion

In this paper we proposed several statistical oriented methods in order to clarify to which conversation an instant message belongs. It is all the more a simple approach since it relies on temporal information and does not require high end semantic analysis. Though the hypotheses we advanced are yet to be validated, they most probably enable us to detect conversations quite accurately. In further studies, we will compare our present hypotheses with the real usage of IM in order to have factual conversation duration. Thus, this real duration will be also statistical value, since obviously the duration of a conversation depends on a lot of factors. The key question is actually not to be able to identify the precise duration of a conversation in the absolute but to evaluate this duration taking into account the context of the exchanges (message length, etc). From this point of view, we think that our study gave some usable perspective.

From a practical point of view we could certainly benefit from a real time detection of conversations to understand and react better to IM usages, by proposing for instance different services depending on whether IM users are active or idle. So during a quick paced or long message text conversation users could be suggested to automatically switch to telephony. Another example, when a user has just finished a conversation he is probably most receptive to commercials for premium services (since a user is not necessarily available too long after a conversation and since pop up windows in the midst of a conversation would be too intrusive).

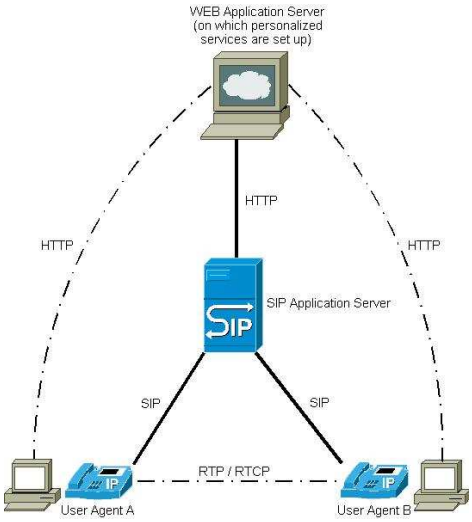


Figure 5: a simple SIP architecture for advanced services

The previous SIP architecture can be simply modified by adding a web application server on which new services can be deployed. By analyzing in real time signalization that goes through the SIP application server services could react to a change of behaviour.

Information that need to be stored could be externalized in a database.

In the future we would like to use these results to compare IM and telephony usages. In particular we would like to see if the use of IM precedes the use of telephony or if in some cases usages are simultaneous. Actually the only available metric with telephony is the conversation, at least at a protocol level. Comparisons will be all the more relevant that we use a similar metric.

5. References

[1]. Mobasher, B.; Cooley, R.; Srivastava, J, "Creating adaptive Web sites through usage-based clustering of URLs", Knowledge and Data Engineering Exchange, 1999. (KDEX '99) Proceedings. 1999 Workshop on 7 Nov. 1999 Page(s):19 – 25

[2]. Mobasher, B.; Honghua Dai; Tao Luo; Nakagawa, M, "Using sequential and non-sequential patterns in predictive Web usage mining tasks" Data Mining, 2002. ICDM 2002. Proceedings. 2002 IEEE International Conference on 9-12 Dec. 2002 Page(s):669 – 672

[3]. Masseglia, F.; Teisseire, M.; Poncelet,P. , "Real time Web usage mining: a heuristic based distributed miner" Web Information Systems Engineering, 2001. Proceedings of the Second International Conference on Volume 1, 3-6 Dec. 2001 Page(s):288 - 297 vol.1

[4]. Zelezny F., Zidek J., Stepankova O.:. In: Bustard D., Weiru L., Sterritt R., "A Learning System for Decision Support in Telecommunications" (eds.): Soft-Ware 2002 Computing in an Imperfect World, 1st Int. Conference, Belfast, Northern Ireland, 4/2002. Springer-Verlag 2002, ISBN 3-540-43481-X.

[5]. Zelezny F., Miksovsky P., Stepankova O., Zidek J.:. In: Brazdil P., Jorge A. "KDD and telecommunications" (eds.): Workshop on data mining, decision support, meta-learning and ILP: forum for practical problem presentation and prospective solutions (workshop at PKDD 2000). Lyon, France, 9/2000 Univ. of Porto.

[6]. Zelezny F., Miksovsky P., Stepankova O., Zidek J. In: Cussens J., Frisch A, "ILP for automated telephony". (eds.): Proceedings of the work-in-progress track of the 10th International Conference on Inductive Logic Programming (ILP 2000), Imperial College, London, UK, 6/2000.

[7]. Resig, J, Dawara, S, et al., "Extracting Social Networks from Instant Messaging Populations" KDD 04 Link Discovery Workshop

[8]. Ellen Isaacs, Candace Kamm, Diane J. Schiano, Alan Walendowski, & Steve Whittaker "Characterizing Instant Messaging from Recorded Logs" AT&T Labs, 75 Willow Road, Menlo Park, CA 94025 Conference on Human Factors in Computing Systems, Minneapolis, Minnesota, April 20-25, 2002 (CHI 2002)

[9] Luigi Lancieri, Interactions humaines dans les réseaux;; Book (french), Hermes ed. ISBN 2-7462-1108-4

[10] Luigi Lancieri, Nicolas Durand, Internet User Behavior: Compared Study of the Access Traces and Application to the Discovery of Communities, In IEEE international Journal, Transaction in Systems, Man and Cybernetics (T-SMCA) January 2006

[11] M. Crovella and A. Bestravos, "Self-similarity in WorldWideWeb traffic: Evidence and possible causes," *IEEE ACM Trans. Netw.*, vol. 5, no. 6,pp. 835–846, Dec. 1997.

[12] V. Paxson, "Fast approximation of self-similar network traffic," Univ. California, Berkeley, Tech. Rep. LBL-36750, Apr. 1995.

[13] Bryant, J. A., Sanders-Jackson, A., & Smallwood, A. M. K. (2006). IMing, text messaging, and adolescent social networks. Journal of Computer-Mediated Communication, 11(2), http://jcmc.indiana.edu/vol11/issue2/bryant.html

[14] Leysia Palen and Marilyn Salzman (2002) Voice-Mail Diary Studies for Naturalistic Data Capture under Mobile Conditions. In Proceedings of the 2002 ACM Conference on

Computer Supported Cooperative Work (CSCW '02), New Orleans, Louisiana.

[15] Grinter, Rebecca and Leysia Palen (2002). Instant Messaging in Teen Life. In Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work (CSCW '02), New Orleans, Louisiana.

[16] IPDR (Internet Protocol Detail Record) web site ipdr.org