Notion de conversation dans les communications interpersonnelles instantanées sur IP

Alexandre Bouchacourt*, Luigi Lancieri**

*France Telecom R&D 42 Rue des coutures 14000 Caen
alexandre.bouchacourt@orange-ftgroup.com
**France Telecom R&D 42 Rue des coutures 14000 Caen
luigi.lancieri@orange-ftgroup.com

Résumé. Dans cet article nous étudions la contribution des techniques de fouille de données à l'amélioration des services de communications instantanées sur IP tel que la messagerie instantanée (IM) et la téléphonie sur IP (ToIP).

Dans cet article nous étudions les aspects temporels de traces d'activité de messagerie instantanée. Nous souhaitons pour ce faire détecter les conversations, en d'autres mots le début et la fin d'échanges de messages cohérents. Dans ce qui suit nous assimilons une conversation à un ensemble de messages consécutifs échangés entre deux interlocuteurs.

Nous partons du constat que bien souvent en IM on ne dispose pas d'information sur la durée des conversations (i.e. qu'on ne sait pas quand une conversation entre deux utilisateurs débute et quand elle se termine) car chaque message est daté indépendamment des autres.

Nous avons pour objectif de trouver une méthode permettant de positionner ces conversations dans le temps. Le matériau sur lequel nous nous appuyons est un corpus IPDR (Internet Protocol Detail Record). Le format IPDR enregistre des traces d'activité au niveau session (le contenu des conversations texte ou voix n'est pas accessible). De nombreuses informations peuvent en être extraites comme les identifiants des utilisateurs, des dates ou encore des tailles de messages. Le corpus que nous étudions représente 6 mois d'activité professionnelle et nous considérons les échanges de 778 couples d'utilisateurs.

Nous avons abordé la question de la segmentation des conversations à l'aide de 2 méthodes statistiques différentes et qui donnent des résultats assez proches.

Nous raisonnons d'abord sur les temps entre deux messages consécutifs (ou inter-temps) et sur la taille des messages. Nous avons ainsi calculé la distribution des inter-temps et tracé en parallèle la taille moyenne de ces inter-temps (comme taille du 1^{er} ou du 2^{nd} message, ou comme moyenne de ces deux tailles). On observe que la taille des messages augmente pour des inter-temps compris entre 0 et 2 minutes et qu'ensuite elle décroit. Nous l'expliquons par la probabilité qu'au-delà d'un inter-temps de 2 minutes les messages correspondent à des conversations distinctes.

Nous raisonnons ensuite sur la taille des conversations. En prenant un seuil d'inter-temps en deçà duquel on reste dans la conversation et au-delà duquel on en sort on peut extraire les conversations. Suivant le seuil d'inter-temps choisi elles ne seront pas toutes constituées du même nombre de messages. Nous traçons donc la taille moyenne (en nombre de messages) des conversations extraites en fonctions du seuil d'inter-temps choisi. La courbe est bien entendu croissante. On observe qu'entre 0 et 3 minutes de seuil d'inter-temps la taille des

conversations augmente très vite. Entre 3 et 7 minutes la croissance est plus lente. Au-delà de 10 minutes elle est très lente et pratiquement linéaire. Nous pensons donc qu'à partir de 3 minutes la probabilité que la conversation soit terminée augmente. Nous pensons aussi qu'au-delà d'un inter-temps de 10 minutes la probabilité que les deux messages appartiennent à la même conversation est faible.

Nous introduisons enfin une notion "d'accélération" qui rend compte du rythme intrinsèque à une conversation. Nous posons ainsi un indicateur d'accélération égal au quotient des deux inter-temps précédemment calculés. Pour différentes classes d'inter-temps considérées il est intéressant d'observer les distributions des accélérations. Ces dernières sont d'autant plus centrées autour de la valeur 1 (rythme inchangé) que l'on considère des inter-temps faibles.

En remarquant que les débuts et fins de conversations sont caractérisés respectivement par une très forte accélération et une très forte diminution du rythme on définit un autre critère d'extraction des conversations. En prenant comme mesure de comparaison la taille des conversations extraites, on obtient une similitude maximale avec l'extraction par seuil d'inter-temps pour une valeur d'inter-temps de 6 minutes. Cette valeur correspond au milieu de l'intervalle 2-10min défini précédemment et nous semble un bon compromis pour décider de l'appartenance d'un message à telle ou telle conversation.

Les hypothèses que nous formulons méritent bien entendu une validation expérimentale. Nous souhaitons de plus exploiter ces résultats dans un contexte d'usages multicanaux. Nous pensons notamment que les usages de l'IM en entreprise ne sont pas décorrélés des usages de la téléphonie, et nous aimerions quantifier et qualifier les phénomènes de transitions de média (passage d'un média à un autre dans une même conversation).

Références

[1]. 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.

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

[3]. 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)

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

[5] Alexandre Bouchacourt, Luigi Lancieri, "Usages Analysis in Instant Interpersonal Communications over IP, European Conference on Universal Multiservice Networks 2007, Toulouse, France

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

Summary

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.