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