Détection de communautés sur un graphe biparti et application à la classification automatique des résultats d'une recherche web (système Kodex).

Emmanuel Navarro[†], Yannick Chudy[⋆], Bruno Gaume[⋆]

[†] : IRIT, CNRS-Université Toulouse 3, Toulouse

[⋆] : CLLE-ERSS, CNRS-Université Toulouse 2, Toulouse

14 janvier 2010

La liste ordonnée de documents est incontestablement la forme de présentation des résultats la plus couramment utilisée par les Systèmes de Recherche d'Information (SRI). Cette liste est naturelle au vue du paradigme usuel en recherche d'information : étant donné une certaine requête, on cherche à mesurer la pertinence de chacun des documents puis les meilleurs sont retournés. Cette mesure (pour une requête donnée) induit donc un ordre complet entre les documents : deux documents sont toujours comparables en terme de pertinence pour une requête donnée.

Mais quand un SRI reçoit une requête, cette requête est souvent polysémique au sens ou des communautés différentes peuvent entretenir des rapports différents avec elle, d'où des attentes différentes selon les intentions de l'utilisateur : par exemple pour la requête "*orange*", l'utilisateur peut chercher des documents concernant le fruit, ou bien des documents concernant la ville, la couleur ou encore l'entreprise. Dès que la requête s'avère ambigüe, une liste ordonnée ne permet que trois solutions :

- Une désambiguisation tacite : "*orange*" en tant qu'entreprise est la requête la plus courante, mais est-ce donc ce que cherche l'utilisateur ?
- Une désambiguisation profilée : "*orange*" en tant que ville d'après le profil de l'utilisateur, mais encore faut-il connaitre le profil de l'utilisateur.
- Un assortiment : une liste panachée de documents concernant des interprétations très différentes de la requête, mais comment (avec une simple liste) informer l'utilisateur de cet assortiment ?

Une autre solution serait de faire connaître à l'utilisateur l'existence de ces multiples points de vue sur sa requête relativement à la base documentaire interrogée, l'utilisateur pouvant ainsi en toute connaissance de cause préciser son choix selon les informations recherchées.

Ce problème d'insuffisance de la simple liste ordonnée a déjà été de nombreuses fois soulevé et de nombreuses solutions ont été imaginées afin d'aider l'utilisateur à naviguer dans les résultats d'une recherche. C'est en particulier l'objectif des méthodes de clustering des résultats d'une recherche web (*web search clustering*). Dans ce cadre nous proposons le système Kodex qui se veut être une réponse originale à cette question qu'est la classification (non supervisée) des résultats d'une recherche.

La principale originalité de notre proposition est d'utiliser une technique de *détection de communautés* pour faire ressortir la structure existante dans l'ensemble des résultats. En effet, notre approche se nourrit d'avancées relativement récentes concernant les grands graphes de terrain[1] : il a été montré que la plupart des grands réseaux construits à partir de données réelles (et donc en particulier les graphes modélisant des collections de documents) partagent des propriétées structurelles non triviales. En particulier une organisation en communautés de sommets plus fortement connectés entre eux qu'au reste du graphe. Ces avancées ont engendré d'importants travaux concernant la définition et la détection de ces communautés, les récents articles [Fortunato and Castellano, 2007] et [Porter et al., 2009] témoignent de l'effervescence des recherches autour de ce problème.

Notre système repose sur une méthode de détection de communautés dans un graphe biparti. L'idée est, pour une requête donnée, de modéliser l'ensemble des documents retournés (par un systeme de recherche d'information "classique") en un graphe biparti et de chercher les communautés qui se dégagent de cette structure. Dans ce graphe chaque document est connecté à un certain nombre de "labels" qui le décrivent.

Notre détection de communautés est une adaptation aux cas des graphes bipartis de la méthode *Walktrap* proposée par [Pons and Latapy, 2006]. Elle consiste à ramener le problème de détection de communautés dans un graphe à un problème de clustering hiérarchique d'un ensemble de points dans un espace vectoriel. Cette transformation (nous parlons de *géométrisation*) de la topologie du graphe dans un espace vectoriel pertinent est operée par la méthode stochastique *Prox*[2] initialement introduite par [Gaume, 2004].

Références

[Fortunato and Castellano, 2007] Fortunato, S. and Castellano, C. (2007). Community structure in graphs. *eprint arXiv :0712.2716.* Chapter of Springer's Encyclopedia of Complexity and System Science.

[Gaume, 2004] Gaume, B. (2004). Balades aléatoires dans les petits mondes lexicaux. *I3 Information Interaction Intelligence*, 4(2).

[Gaume, 2008] Gaume, B. (2008). Mapping the forms of meaning in small worlds. *International Journal of Intelligent Systems*, 23(7) :848–862.

[Pons and Latapy, 2006] Pons, P. and Latapy, M. (2006). Computing communities in large networks using random walks (long version). *Journal of Graph Algorithms and Applications (JGAA)*, 10(2) :191–218.

[Porter et al., 2009] Porter, M. A., Onnela, J., and Mucha, P. J. (2009). Communities in networks. *Notices of the American Mathematical Society*, 56(9).

[1] On parle de grands graphes de terrain, ou de réseaux de petits mondes hiérarchiques ou encore de *small world* ou de *complex networks* en anglais.

[2] L'article [Gaume, 2008] présente, en anglais, cette méthode. Notons qu'une démonstration de cette même méthode pour visualiser un dictionnaire de synonymes est accessible à l'adresse : `http://prox.irit.fr`.