RankMerging: Apprentissage supervisé de classements pour la prédiction de liens dans les grands réseaux sociaux

Lionel Tabourier*,**, Anne-Sophie Libert***, Renaud Lambiotte***

*Sorbonne Universités, UPMC Univ. Paris 06, UMR 7606, LIP6
**CNRS, UMR 7606, LIP6, Paris, France
***naXys, Université de Namur, Namur, Belgique

Résumé. Trouver les liens manquants dans un grand réseau social est une tâche difficile, car ces réseaux sont peu denses, et les liens peuvent correspondre à des environnements structurels variés. Dans cet article, nous décrivons *RankMerging*, une méthode d'apprentissage supervisé simple pour combiner l'information obtenue par différentes méthodes de classement. Afin d'illustrer son intérêt, nous l'appliquons à un réseau d'utilisateurs de téléphones portables, pour montrer comment un opérateur peut détecter des liens entre les clients de ses concurrents. Nous montrons que *RankMerging* surpasse les méthodes à disposition pour prédire un nombre variable de liens dans un grand graphe épars.

1 Introduction et contexte

Le problème de la prédiction de liens tel qu'il est formulé dans l'article de Liben-Nowell et al. (2007) peut être compris comme une tâche de classification binaire. Des outils classiques d'apprentissage tels que les arbres de classification, les SVM ou les réseaux de neurones ont été utilisés pour le résoudre sur des réseaux biologiques et de collaboration (Pujari et al. (2012)). Cependant, ces méthodes ne permettent pas à l'utilisateur de faire varier le nombre de prédictions selon ses besoins. Pour ce faire, il est possible de calculer un score pour chaque paire de nœuds, corrélé à la probabilité d'existence d'un lien entre ces nœuds, on obtient alors un classement, et l'utilisateur effectue la prédiction en sélectionnant les T paires les mieux classées. Le score peut être basé sur la structure connue du réseau, mais également sur d'autres sources d'information : par exemple les attributs des nœuds, la dynamique des contacts ou la localisation géographique (Scellato et al. (2011)). Pour combiner les informations capturées par différents scores, on utilise des méthodologies d'apprentissage de classements. Parmi les méthodes à disposition, certaines sont non-supervisées et peuvent être vues comme des méthodes de consensus, telles que celles décrites dans Dwork et al. (2001). Il existe également des méthodes supervisées : une solution consiste à se ramener à un problème de classification en effectuant une *transformation deux-à-deux*, plutôt que de considérer des éléments à ordonner, on examine des couples d'éléments dont on cherche à dire lequel doit être classé audessus de l'autre (Herbrich et al. (1999)). Malheureusement, cette méthode n'est pas adaptée à de grands réseaux où le nombre d'éléments à classer est élevé. Plus généralement, la plupart des méthodes d'apprentissage de classements ont été créées pour des tâches de recherche d'information, où l'on souhaite une précision élevée sur un petit nombre d'éléments (e.g., Burges