Proposition de recherche doctorale

Modèles vectoriels de documents pour la fouille de textes bio-médicaux : Application à l'identification de relations gènes-maladies

Mots clés :

- **Directeur de thèse** : Mohamed NADIF
- **Co-encadrant(s)** :
- **Unité de recherche** : Laboratoire d'Informatique PAris DEscartes
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Divers

Résumé du projet de recherche (Langue 1)

{{Contexte}} Le volume des publications bio-médicales augmente à un rythme sans précédent. A titre d'exemple, la base PubMed contient 20 millions d'articles et ce nombre augmente d'environ 50 000 par mois . Il est maintenant reconnu que des outils avancés de text mining sont nécessaires pour extraire automatiquement les riches informations bio-médicales contenues dans ces nombreuses publications. Des techniques de traitement du langage et de fouille de textes ont donc été appliquées à des tâches aussi diverses que la reconnaissance d'entités nommées, l'annotation fonctionnelle des gènes, les interactions entre protéines, la construction automatique d'ontologies. Krallinger [4] fournit une synthèse récente de ce domaine en pleine évolution. Au delà de leurs différences, les applications évoquées ci-dessus visent fondamentalement à utiliser des descriptions textuelles pour : -# regrouper entre elle des entités de même nature (par exemple, effectuer un clustering de gènes), -# identifier des relations entre des entités de nature différente (par exemple, mettre en relation un ensemble de gènes et de protéines). L'identification de groupes de gènes est un problème qui est étudié depuis les débuts de la bio-informatique. Plus récemment, la recherche médicale a mis en évidence l'intérêt qu'il y a également à pouvoir regrouper automatiquement des maladies que l'on avait cru jusqu'ici sans rapport les unes avec les autres. De plus, les groupes de maladies et des groupes de gènes ainsi identifiés peuvent entretenir entre eux des relations identifiables en utilisant des techniques d'analyse exploratoire. La constitution de groupes de gènes ou de maladies fait appel à des techniques de clustering tandis que l'identification des relations entre groupes peut s'appuyer soit sur des techniques de co-clustering [1] soit sur des techniques relevant de l'analyse de graphes et de réseaux [2]. Cependant, quelle que soit la technique utilisée, la qualité des résultats obtenus dépend en grande partie de la richesse des représentations informatisées initiales. La recherche médicale en génomique utilise depuis longtemps de nombreuses bases de données structurées qui peuvent être interrogées en utilisant des langages de requêtes sophistiqués pour retrouver des informations sur les entités à étudier (gènes, protéines, SNPs, maladies génétiques, etc.) Les méthodes permettant d'intégrer ces bases non textuelles dans des processus de text mining ont encore été peu étudiées. Il s'agit d'un enjeu important qui sera au coeur des recherches menées dans le cadre de cette thèse dont le but est de construire, de valider et d'exploiter des representations multiples d'entités comme les gènes et les maladies. {{Travail à effectuer}} Le premier axe du travail de thèse consistera à représenter les entités à étudier (gènes, maladies) sous forme de vecteurs dans différents espaces. Ces espaces pourront être dérivés de corpus d'articles (représentation d'un gène dans l'espace des mots contenus dans les articles) mais également de bases de données génomiques (par exemple, représentation d'un gène dans l'espace des séquences de nucléotides). La qualité des différentes représentations vectorielles obtenues sera évaluée sur la base de leur capacité à apparier des groupes de maladies et des groupes de gènes apparentés. Différentes combinaisons de représentations vectorielles et de mesures de similarité seront expérimentées et comparées en termes de rappel et de précision. Le scénario applicatif suivant servira de fil conducteur aux études menées : à partir d'un ensemble de gènes connus pour être liés à l'asthme on identifiera des maladies pouvant être liées à l'asthme. Pour construire les représentations des gènes et des maladies, on combinera les informations génétiques contenues dans la base GenBank avec les informations issues de l'analyse du texte d'articles contenus dans Pubmed et des descripteurs MESH correspondant à la catégorie “maladie”. La seconde phase des travaux exploitera les multi-représentions les plus performantes identifiées lors de l'étape précédente, le but étant de créer des matrices de similarité utilisées dans deux contextes : -# clustering de gènes et clustering de maladies, en mettant en oeuvre des méthodes d'ensemble (ensemble methods) [5], -# co-clustering de gènes et de maladies, -# analyse de réseaux regroupant gènes et maladies, en utilisant des mesures de centralité, notamment les mesures d'autorité (hub and authority) [3]. L'évaluation des résultats obtenus durant cette phase portera sur la qualité des regroupements d'entités similaires (clusters de gènes, clusters de maladies) ainsi que sur la qualité des mises en relation d'entités de nature différente (gènes potentiellement liés à des maladies).

Résumé du projet de recherche (Langue 2)

Un des défis principaux du travail envisagé consiste à identifier les bons points d'articulation entre les informations textuelles et non textuelles ainsi que les méthodes (fusion initiale des sources, combinaison des résultats a-posteriori via des méthodes d'ensemble) à employer pour mettre en oeuvre cette articulation. Une seconde retombée attendue est le développement de méthodes innovantes pour la qualification des relations gènes-maladies qui pourront être identifiées. Les techniques d'évaluation automatisées proposées dans le domaine de la {Literature Based Discovery} sont encore embryonnaires [4] et il y a là matière à apporter une contribution significative. Enfin, les techniques d'analyse de réseau et les techniques de co-clustering commencent à être utilisées en fouille de texte bio-médical mais sont encore sous-exploitées dans ce domaine.

Informations complémentaires (Langue 2)

Ce sujet s’adresse à des étudiants souhaitant contribuer au domaine de la fouille de données dans le domaine biomédical. Cette recherche se déroulera au sein de l’équipe GFD (Gestion et Fouille de données) sous la direction de Mohamed Nadif et en co-direction avec François Role nouveau membre de l'équipe GFD. Ce travail d'évaluation sera mené en étroite collaboration avec Dr Florence Demenais-Diao de l'UMR 946 Unité mixte de recherche 946 - Variabilité génétique et maladies humaines. Le sujet de cette thèse nécessite des connaissances dans le domaine de l’apprentissage non supervisé et supervisé. Un bagage en mathématiques appliquées et une compétence avérée en programmation seront très appréciés. {{Références}} [1] Dhillon I, Mallela S and Modha DS. Information-Theoretic Co-clustering. KDD'03, pp. 89-98, 2003. [2] Hossain MS, Gresock J, Edmonds Y et al. Connecting the Dots between PubMed Abstracts. PLoS One, (1), 2012. [3] Kleinberg J. Authoritative sources in a hyperlinked environment. Journal of the ACM 46 (5): 604--632, 1999. [4] Kostof R., Validating discovery in literature-based discovery. Biomed Inform. 2007 Aug;40(4):448-50 [5] Krallinger M., Leitner F., Valencia A. Analysis of Biological Processes and Diseases using Text Mining Approaches. Methods Mol Biol, 341-382, 2010. [6] Strehl A., Ghosh J., Cluster ensembles – a knowledge reuse framework for combining multiple partitions, Journal on Machine Learning Research (JMLR) 2002