Projet - M2 BioInfo - Web sémantique

4 février 2020

Contexte Général

On s'intéresse ici à l'extraction de connaissances à partir de données structurées et non structurées. L'application de la méthode retenue se fera sur un corpus de textes en entrée (au format XML pour les données structurées et texte pour les résumés) à partir desquels on souhaite obtenir des associations entre concepts MeSH ou entre termes biomédicaux. Ces associations seront ensuite réutilisées pour constituer une ontologie au format standard OWL (hiérarchies de concepts, relations, hiérarchies de relations, domaines et co-domaines ...etc).

Modalité de rendu

Chacune des étapes (mots clés, constitution des corpus, annotation des résumes et titres avec Annotator de BioPortal (http ://bioportal.bioontology.org/annotator#)) extraction des règles d'association (avec Close ou Apriori ou autre algo de fouille sous Weka) etc...) seront détaillées dans un rapport PDF.

Les fichiers de règles d'associations ainsi que l'ontologie développée au format OWL seront également à remettre.

Le travail peut se faire en binômes.

Dates de remise des archives :

- ✓ **22 mars 2020** au plus tard (lien FileX ou Dropbox actif 3 semaines) pour la partie fouille de données.
- ✓ **24 mai 2020** au plus tard pour la partie fouille de textes et ontologie.

Constitution des corpus XML

Les fichiers XML devant être traités sont des extractions au format XML MEDLINE, de la banque de données des articles scientifiques en santé de la NLM (National Library of Medicine : http://www.ncbi.nlm.nih.gov/pubmed).
Une ensemble de mots clés vous est proposé en annexe, mais vous pouvez proposer 1 mot clé se rapportant à votre domaine d'étude. NB : les fichiers XML résultant des requêtes peuvent être de taille conséquente. Il est recommandé de prévoir une limite supérieure du nombre d'articles traités (par exemple au plus 1 000 articles pour les premiers tests que vous ferez).

Fouille de données

À partir de la collection XML, la fouille de données consistera à extraire des règles d'association entre MeSH Descriptor et entre couples (MeSH Descriptor/Qualifier). Vous pouvez utiliser l'algorithme que vous souhaitez. Des étapes préalables de prétraitement des données sera nécessaires.

Fouille de textes

L'indexation manuelle de MEDLINE peut être complétée par une indexation "automatique". Vous pouvez utiliser l'Annotator de BioPortal

(http ://bioportal.bioontology.org/annotator#)

L'Annotator ne fonctionnant que sur des textes de 300 mots, l'indexation automatique portera sur les résumés des notices de MEDLINE (champs Abstract) + des titres. Pour annoter avec Annotator, placer le texte à annoter puis sélectionner la ou les sources d'indexation (les ontologies). L'annotation résultante peut ensuite être récupérée au format TXT, CSV, XML. L'objectif est d'extraire des règles d'association entre ces concepts du corpus, puis d'en déduire une représentation ontologique.

À partir des sorties de l'Annotator, extraire des associations :

1. entre concepts candidats
2. entre concepts candidats et MeSH Descriptors initialement inclus dans l'index PubMeb
3. entre concepts candidats et couples MeSH Descriptor/Qualifier initialement inclus dans l'index PubMed

Représentation des connaissances

À partir des connaissances extraites ainsi que des connaissances connues du domaine, proposez une représentation ontologique des concepts et des relations.

Annexes :

Format des requêtes PubMed :

http ://www.ncbi.nlm.nih.gov/pubmed ?term=[terme]%5BMeSH%5D
Liste des éléments XML de MEDLINE :http://www.nlm.nih.gov/bsd/licensee/elements_alphabeti
Description des XML Element de MEDLINE http://www.nlm.nih.gov/bsd/licensee/elements_desc

Quelques mots clés MeSH

micro rna ; escherichia coli ; adverse drug events ; protein ligand ; hepatitis e virus ; nosocomial infections ;