Quelques travaux récents du Lattice en rapport avec Mastodons Aresos

Lattice

Contenu

- Participation au programme PoliInformatics
- Thèse d'Elisa Omodei

PoliInformatics

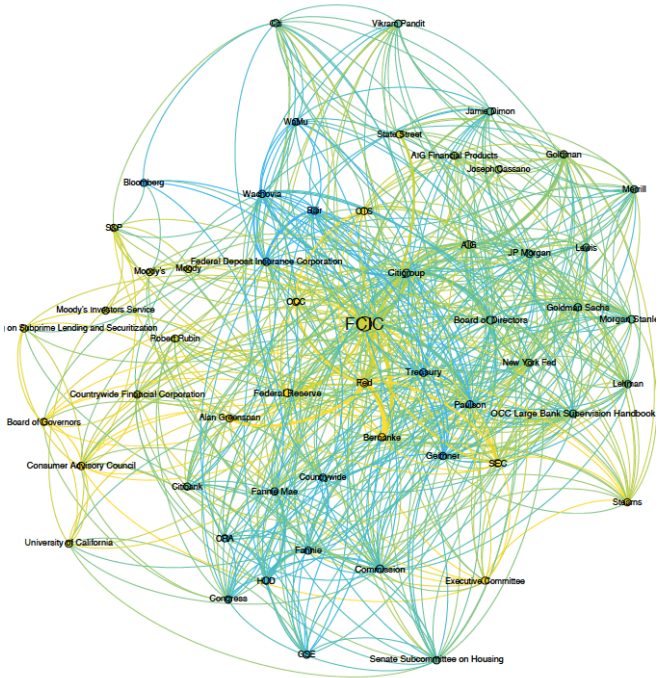
- "An interdisciplinary field that promotes diverse methodological approaches to the study of politics and government" (http://poliinformatics.org/)
- Principalement informatique et TAL pour les sciences sociales
- A partir de documents sur la crise financière de 2007-2008, répondre à des questions comme "*Who was the financial crisis?*" ou "*What was the financial crisis?*"
- Campagne exploratoire
  - Temps de développement très réduit (inscription tardive à la campagne)
  - Etude de faisabilité à partir d'outils standards

Travail effectué

- Travail en temps très limité (1 mois et demi)
- Analyse des entités nommées (Stanford NE Tagger)
- Regroupement des entités coréférentes (par un algorithme ad hoc)
- Création de graphes
  - Représentation des liens entre entités
  - Evolution des thèmes au cours du temps

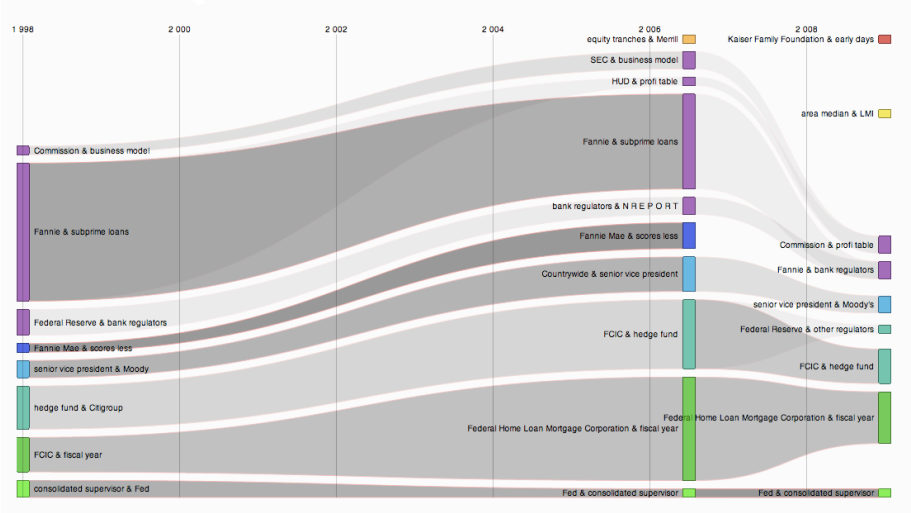
➢ Montrer le passage du texte à des représentations graphiques pertinentes pour l'exploration du corpus

Vue statique des liens entre entités



- Visualisation avec Gephi

Vue dynamique des thèmes abordés



- Représentation obtenue avec la plate-forme Cortext

Conclusion sur PoliInformatics

- Bilan
  - Des outils performants sur de grandes masses de données
  - Passage du texte à des représentations graphiques
  - Mais représentations peu utilisables en l'état (nécessité de prendre en compte des points de vue pertinents)
- Perspectives
  - Améliorer la stratégie de regroupement des entités (thèse en cours de Pablo Ruiz Fabo)
  - Mieux focaliser l'analyse (analyse des points de vue sur les raisons de la crise, *idem*)
  - Valider l'analyse avec l'aide d'experts du domaine

Publication

- Bourreau, Pierre; Poibeau, Thierry (2014). « Mapping the Economic Crisis: Some Preliminary Investigations » . arXiv:1406.4211, 06/2014.

- In this paper we describe our contribution to the PoliInformatics 2014 Challenge on the 2007-2008 financial crisis. We propose a state of the art technique to extract information from texts and provide different representations, giving first a static overview of the domain and then a dynamic representation of its main evolutions. We show that this strategy provides a practical solution to some recent theories in social sciences that are facing a lack of methods and tools to automatically extract information from natural language texts.

  (le soutien de Mastodons Aresos est mentionné)

Thèse d'Elisa Omodei

- Collaboration Lattice – ISC-PIF
- Soutenance prévue le 19 décembre 2014 (à 10h, à l'ISC)
- Thèse dirigée par J.P. Cointet (ISC) et T. Poibeau (Lattice)

- Sujet : méthodes pour la modélisation et l'analyse de grands corpus d'articles scientifiques
  - Dimension sociale : co-publications
  - Dimension sémantique : analyse des cooccurrences de mots clés dans les résumés d'articles
  - Analyse conjointe grâce à un graphe biparti fondé sur les deux modélisations précédentes
  - Analyse de l'évolution des graphes au cours du temps

Réseau de co-auteurs

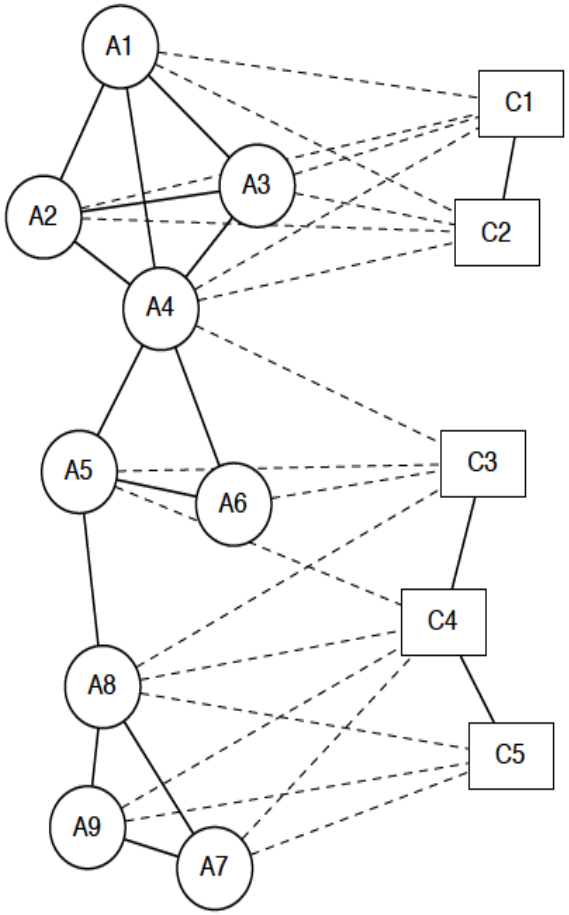
- Obtenu assez simplement à partir de l'analyse des auteurs d'articles
  - Problème pour les corpus de physique, avec des publications de plus de 100 (voire 1000) auteurs
  - Problème de l'ambiguïté des noms d'auteurs (surtout pour les noms asiatiques)

Réseau sémantique

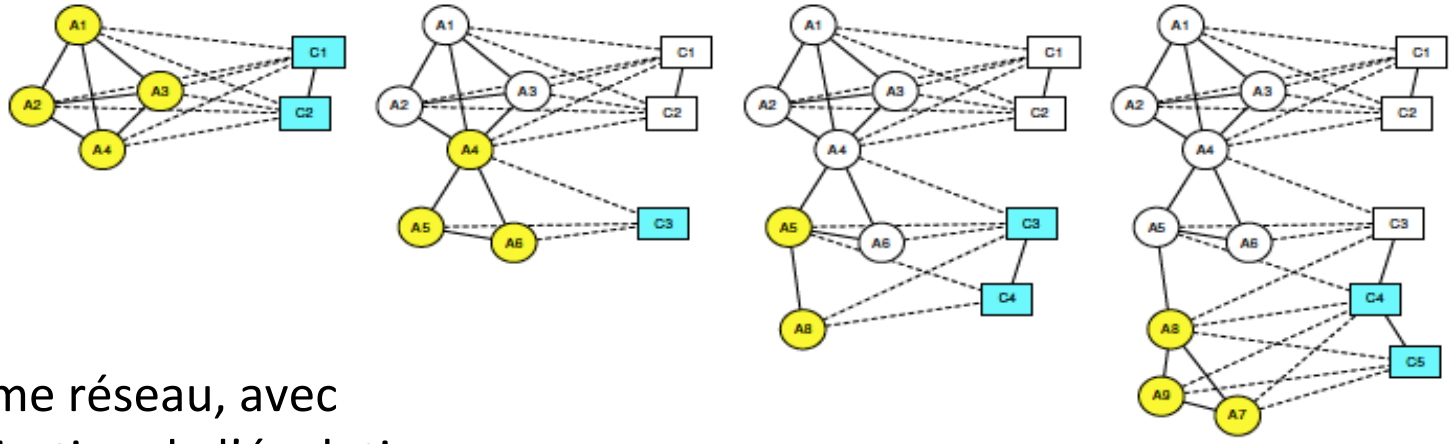
- Nécessité d'extraire les termes pertinents à partir des résumés d'articles
  - Utilisation d'une méthode hybride : patrons syntaxiques + critères statistiques (proche d'Acabit de B. Daille, 1996)
- Catégorisation des termes en fonction de leur valeur informationnelle
  - Analyse rhétorico-discursive des résumés (*text zoning*)
  - Catégorisation des mots clés en fonction de cette analyse
  - Analyse du corpus en fonction des méthodes utilisées (grâce à l'analyse des mots clés référant à des méthodes)

Réseau biparti

Réseau obtenu à partir de l'analyse de 4 articles : le premier a été produit par les auteur A1, A2, A3 et A4, et comprend les concepts C1 et C2 ; le deuxième article a pour auteur A4, A5, A6 avec le concept C3. Les auteurs A5 et A8 ont produit un article avec les concepts C3 et C4, enfin A7, A8 et A9 ont produit un article sur C4 et C5.



Dimension temporelle



Le même réseau, avec modélisation de l'évolution au cours du temps.

Principaux résultats

- Une modélisation originale
  - Recours à des techniques de TAL avancées
  - Choix des mesures pour l'analyse (clustering, évaluation du poids de différentes variables pertinentes)
  - Prise en compte du réseau sémantique complet, ou juste du graphe des méthodes
- Des résultats en termes d'analyse de l'évolution de domaines scientifiques
  - Comment sont introduites de nouvelles méthodes dans un champ scientifique ?
  - Qui collabore avec qui ? Peut-on prédire de futures collaborations possibles ou probables ? Etc.

Quelques publications liées à la thèse

- Elisa Omodei, Thierry Poibeau, Jean-Philippe Cointet, "Mapping the Natural Language Processing Domain: Experiments using the ACL Anthology", Proceedings of the 9th edition of the Language Resources and Evaluation Conference, 26-31 May, 2014, Reykjavik.
- Elisa Omodei, Yufan Guo, Jean-Philippe Cointet and Thierry Poibeau, "Social and Semantic Diversity: Socio-semantic Representation of a Scientific Corpus", Proceedings of the EACL 2014 workshop on "Language Technology for Cultural Heritage, Social Sciences, and Humanities", April 26th, 2014, Gothenburg.
- Elisa Omodei, Thierry Poibeau and Jean-Philippe Cointet, "A symmetric approach to understand the dynamics of scientific collaborations and knowledge production", Proceedings of the 4th French Conf. on "Modèle & Analyse de réseaux : Approches mathématique & informatiques (MARAMI 2013)", 2013.
- Elisa Omodei, Thierry Poibeau and Jean-Philippe Cointet, "Multi-Level Modeling of Quotation Families Morphogenesis", Proceedings of the 2012 ASE/IEEE International Conference on Social Computing, 2012.

Autres travaux en 2014

- un étiqueteur-chunker-reconnaisseur d'entités nommées en français appris par CRF (avec Y. Dupont, logiciel disponible)
- Travaux sur la recherche de patrons séquentiels d'étiquettes POS (DMNLP'14)
- travaux sur la reconnaissance de chaînes de co-références en français (1er corpus annoté disponible) par apprentissage (avec F. Landragin, article TAL, projet ANR déposé)
- Projet ANR proposé avec MediaLab ScPo (analyse de textes issus de sommets sur le changement climatique)

Pour 2015

- Une post-doc (financé par Labex EFL, partagé entre le Lattice et le LIPN) va travailler sur les *relations entre entités* (sans doute avec patrons séquentiels…)
- Un stage co encadré par Marco et Tim Van der Cruys (ANR jeune chercheur) sur RN et sémantique distributionnelle
- Demandes du Lattice à Aresos :
  - Un stage M2 sur l'implémentation des CRF
  - Un stage M2 sur les chaînes de co-réréfences ?