Cybergeo Networks

Cahier des Charges pour une Application Autonome

P.O. Chasset, H. Commenges, C. Cottineau, J. Raimbault

Abstract

Nous développons une estimation des tâches et de la charge de travail (ceci n'est pas un cahier des charges informatique au sens classique) pour la création d'une première version d'un logiciel libre d'exploration bibliographique à partir du prototype développé à l'occasion de l'anniversaire de *Cybergeo* [Chasset et al., 2016].

1 Introduction

L'exploitation des nouvelles sources de données et des traitements associés a radicalement changé la nature de la réflexivité des connaissances, avec par exemple des possibilités de cartographie dynamique [Chavalarias and Cointet, 2013] et d'exploration interactive de domaines scientifiques [Chen, 2010], d'anticipation des fronts de recherche émergents [Shibata et al., 2008], ou de l'établissement de mesures prédictives du devenir d'un travail scientifique [Newman, 2013, Sarigöl et al., 2014]. Le prototype *CybergeoNetworks* [Chasset et al., 2016], développé à l'occasion des 20ans de la revue *Cybergeo*, est une application interactive en ligne offrant des possibilités d'exploration de différents aspects bibliométriques, et se réclamant de deux principes fondamentaux : d'une part, les analyses bibliométriques quantitatives pures sont néfastes à la science (voir les études empiriques comme [Alberto and Giuseppe, 2015]) et une approche hybride intégrant qualitatif et quantitatif est nécessaire ; d'autre part, ces nouvelles approches ont rapidement subit la prédation des géants privés de l'édition scientifique qui cherchent à valoriser pour leur profit cette "valeur ajoutée": il est alors essentiel pour les éditeurs libres de s'armer d'outils libres offrant au moins des fonctionnalités similaires, dans l'idéal des fonctions plus avancées[1].

Afin de poursuivre l'esprit du prototype, la mise à disposition d'un outil libre flexible et simple offrant au départ les mêmes fonctionnalités est nécessaire. Nous détaillons ici les étapes nécessaires pour passer du prototype à une version beta de l'application.

2 Objectifs

Court terme Obtenir une version beta du logiciel satisfaisant les contraintes suivantes :

1. Pour toute revue utilisant le système Lodel, fonctionnalités *front* équivalentes à celles de la version de démonstration sur Cybergeo[2]
2. Logiciel qui pourra de déployer de façon autonome et sans besoin de compétences avancées pour une revue donnée, sur un serveur ayant accès à la base de production
3. Architecture *back-end* permettant la mise à jour en continu (dans la limite des ressources disponibles) des données bibliographiques utilisées par l'application

[1] le développement libre est en général à la pointe de l'innovation dans la plupart des domaines

[2] `http://shiny.parisgeo.cnrs.fr/CybergeoNetworks`

Long terme Développements futurs : ajouts de fonctionnalités, études de scalabilité, système de gestion administrateur etc.

3 Contraintes Générales

Parmi d'autres, les contraintes suivantes retiennent particulièrement notre attention :

- Hétérogénéité des languages : nécessité de maitriser de manière avancée R/shiny, Java, Python.
- Licence : l'application devra être Open Source et Libre (non négociable), dès la récupération du code du prototype.
- Légalité de la collecte des données : le prototype utilise les bibliothèques `TorPool` [Raimbault, 2016b] et `ScholarAPI` [Raimbault, 2016a], dont une utilisation abusive peut être interprétée comme un contournement des conditions d'utilisation du fournisseur de données ; il conviendra de vérifier la position légale de l'application sur ce point. L'utilisation d'autres sources de données n'est guère envisageable, cette approche permettant d'étudier des revues non référencées par les bases propriétaires fournissant des API.

Par la suite, l'estimation des temps de travail est donnée en heures ou jours (1jour = 7h) de travail d'un ingénieur compétent, et calibrée sur les temps effectifs de développement du prototype ; les temps de parallelisation des tâches/organisation sont négligés.

4 Tâches Préliminaires

Il reviendra à l'équipe du prototype d'assurer les tâches préliminaires suivantes pour une reprise en main réaliste par des développeurs extérieurs :

- Nettoyage du code, factorisation, niveau de commentaires raisonnable **[ETA 1j - tous]**
- Mise en cohérence de l'architecture : autonomisation, isolation et spécification fonctionnelle pour les différents blocs **[ETA 4h - tous]** :
  - Collecte des données brutes
  - Pré-traitements des données/analyse statistiques pour une comestibilité directe par l'application shiny
  - Application shiny
- Bugs mineurs de l'application shiny, ajustements cosmétiques, suppression des fonctionnalités non-automatisables (e.g. visualisation du réseau sémantique) **[ETA 2h - tous]**
- Automatisation de certains traitements statistiques préliminaires (e.g. estimation des paramètres optimaux du réseau sémantiques) **[ETA 4h - JR]**
- Intégration de l'analyse thématique LDA des textes complets dans l'application shiny **[ETA 4h - POC]**
- Etude de faisabilité de l'automatisation du géocodage des articles, solutions alternatives (*question : l'onglet des cartes est-il toujours pertinent pour des revues qui ne font pas de géographie ?*) **[ETA 4h - CC]**

- Etude de faisabilité de l'automatisation de la construction du thésaurus des mots-clés, solution alternatives **[ETA 4h - HC]**
- Note sur les performances de l'application (complexité des différents traitements, vitesse maximale/optimale de collecte des données à estimer) **[ETA 1j - tous]**
- Note sur l'architecture générale et des différents modules si nécessaire, **[ETA 1j - tous]**
- Guide de navigation détaillant avec des exemples l'intérêt et le fonctionnement des différentes analyses produites par l'application, **[ETA 4h - CC]**

5 Cahier des Charges

5.1 Court Terme - Version Beta fonctionnelle

Etapes de développement nécessaire pour satisfaire les objectifs sans modification majeure du code du prototype, en supposant le travail préliminaire par l'équipe du prototype réalisé.
[ETA total - 7j]

Prise en main

Prise en main du code, de l'architecture et de la documentation **[ETA 2j]**

Collecte des données

- Rendre générique et propre l'interface avec la base de production (nom des tables et champs comme paramètres ; *sur ce point tant que Lodel ne fournit pas d'API propre, il risque d'y avoir une étape d'installation peu évidente dans l'installation du logiciel qui consistera en la définition de ces paramètres*) **[ETA 1j]**
- Démon de collecte et de mise à jour **[ETA 4h]**

Pré-traitements

- Envelopper les différents scripts de pré-traitement dans une sous-application **[ETA 4h]**
- Démon de prétraitement **[ETA 2h]**

Application Shiny

- Rendre générique et modifiable facilement (ex. : fichier de configuration externe) le texte de description de la revue et les popups issues du guide de navigation **[ETA 2h + 2h]**
- Adapter les interfaces selon les modifications faites sur les prétraitements **[ETA 4h]**

Application

- Envelope globale de l'application **[ETA 4h]**
- Programme d'installation "clés-en-main" pour l'utilisateur novice **[ETA 1j]**

5.2 Long Terme

Nous donnons ici des pistes pour les développements futurs, les objectifs sont plus flous et non chiffrés.

- Professionnalisation du code (la version beta ci-dessus aura toujours un code "artisanal")
- Elaboration d'un outil de gestion administrateur des différentes options de l'application
- Généricisation du type de base de production; écriture d'une API
- Nouvelles fonctionnalités
- Scalabilité : vers une application multi-revues, étude de faisabilité "grandes-données"
- Sur le très long terme : réécriture complète de l'application de manière intégrée (un seul language) ; point à débattre car la nature hétéroclite de l'architecture fait la force de l'application pour l'instant (sous-optimisations)

Chiffrage des coûts

- Pour l'étape court terme : 7 jours temps plein ; coût minimisé par l'emploi de précaires : doctorants **500€ HC**, post-docs **600€ HC** ; ingénieurs **1500€ HC**
- Pour le long terme, attention à ne pas se faire enfumer par des "bureaux d'études" privés qui vendent du vent - serait-il possible de rester en interne au CNRS ou organisme public ?

References

[Alberto and Giuseppe, 2015] Alberto, B. and Giuseppe, D. N. (2015). Do they agree? bibliometric evaluation vs informed peer review in the italian research assessment exercise. *arXiv preprint arXiv:1505.00115*.

[Chasset et al., 2016] Chasset, P.-O., Commenges, C. C. H., and Raimbault, J. (2016). cybergeo20 v1.0, doi 10.5281/zenodo.53905.

[Chavalarias and Cointet, 2013] Chavalarias, D. and Cointet, J.-P. (2013). Phylomemetic patterns in science evolution—the rise and fall of scientific fields. *Plos One*, 8(2):e54847.

[Chen, 2010] Chen, C. (2010). Citespace: Visualizing patterns and trends in scientific literature. *Retrieved January*, 27:2010.

[Newman, 2013] Newman, M. E. J. (2013). Prediction of highly cited papers. *ArXiv e-prints*.

[Raimbault, 2016a] Raimbault, J. (2016a). Scholarapi v1.0, doi 10.5281/zenodo.53763.

[Raimbault, 2016b] Raimbault, J. (2016b). Torpool v1.0, doi 10.5281/zenodo.53739.

[Sarigöl et al., 2014] Sarigöl, E., Pfitzner, R., Scholtes, I., Garas, A., and Schweitzer, F. (2014). Predicting Scientific Success Based on Coauthorship Networks. *ArXiv e-prints*.

[Shibata et al., 2008] Shibata, N., Kajikawa, Y., Takeda, Y., and Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation*, 28(11):758–775.