Prise en compte de la structure des documents pour une indexation performante

Pascal Cuxac, Nicolas Kieffer

INIST-CNRS
2, allée du parc de Brabois
54519 Vandoeuvre lès Nancy
pascal.cuxac@inist.fr, nicolas.kieffer@inist.fr

1 Introduction

L'indexation d'un document a des intérêts multiples : répondre à une requête, faire une classification, comparer des textes... (Gupta et Lehal, 2017).

Nous présentons Skeeft, une méthode automatique d'indexation en langage libre, sans ressource appliquée à du texte intégral (des articles scientifiques). Ce choix permet de s'affranchir du domaine scientifique et traiter un corpus multi-thématiques sans difficulté. L'analyse du texte intégral permet l'extraction de termes pertinents et une meilleure pondération. La structure du document est souvent utilisée afin de cibler des zones d'extractions ou pondérer les termes (You et al., 2013). Notre approche est très différente car nous ne hiérarchisons pas les différentes parties mais nous les mettons en concurrence : il suffit d'identifier les parties du document sans avoir à identifier leur "rôle" (introduction, méthodologie...).

2 Méthodologie

Nous assimilons un document à une classification dont chaque partie est une classe composée de termes. A partir de là le procédé se déroule en 4 grandes étapes :

— extraction des termes pour chaque partie identifiée (toute méthode est utilisable ici) ;
— application d'une sélection de variables : un terme est une variable, la partie ou il apparaît est sa classe d'appartenance (Lamirel et al., 2015) ;
— pondération des termes sélectionnés pour chaque partie ;
— filtrage final, fusion des résultats et affichage.

Le lecteur se rapportera à l'article de Lamirel et al. (2015) pour plus de détails.

3 Résultats expérimentaux

Skeeft est testée sur des documents multi-disciplinaires du réservoir ISTEX [1], indexés manuellement par des experts et comparée aux méthodes suivantes : TopicRank (Bougouin et

1. https ://api.istex.fr/documentation/

Boudin, 2014), Keyterm (Lopez et Romary, 2010), KPMiner (El-Beltagy et Rafea, 2009), SingleRank (Wan et Xiao, 2008), Kea (http ://www.nzdl.org/Kea) et Termostat (Drouin, 2003).

Méthode	Rappel	Précision	F-mesure
TopicRank	0.11	0.18	0.14
Keyterm	**0.25**	0.21	0.23
SingleRank	0.06	0.09	0.07
Kea	0.14	0.21	0.17
KPMiner	0.17	0.22	0.19
Termostat	0.24	0.30	**0.27**
Teeft	0.23	0.20	0.21
Skeeft	0.21	**0.32**	**0.25**

TAB. 1 – *Comparaison des performances de Skeeft sur un corpus test indexé manuellement*

Les résultats présentés dans le tableau ci-dessus montrent de bonnes performances de Skeeft en terme de F-mesure puisque seul Termostat donne des résultats légèrement meilleurs.
Ces travaux sont financés par le projet ISTEX - programme ANR-10-IDEX-0004-12.

Références

Bougouin, A. et F. Boudin (2014). Topicrank : ordonnancement de sujets pour l'extraction automatique de termes-clés. *TAL 55*(5), 45–69.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology 9*(1), 99–117.

El-Beltagy, S. et A. Rafea (2009). Kp-miner : a keyphrase extraction system for english and arabic documents. *Inf Syst 34*(1), 132–144.

Gupta, V. et S. Lehal (2017). Keyword extraction :a review. *Int.j.eng.appl.sci. 2*(4), 215–220.

Lamirel, J.-C., P. Cuxac, A. Chivukula, et K. Hajlaoui (2015). Optimizing text classification through feature selection based on quality metric. *J. of Intel. Inf. Syst. 45*(3), 379–396.

Lopez, P. et L. Romary (2010). Humb : Automatic key term extraction from scientific articles in grobid. *In Proc. of the 5th Int. workshop on semantic evaluation, ACL*, 248–251.

Wan, X. et J. Xiao (2008). Single document keyphrase extraction using neighborhood knowledge. *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, 855–860.

You, W., D. Fontaine, et J.-P. Barthès (2013). An automatic keyphrase extraction system for scientific documents. *Knowledge and Information Systems 34*(3), 691–724.

Summary

We present Skeeft, a method that improves terms extraction , taking into account the structure of the document. We consider a document to be a classification; a feature selection method is then used to select specific terms.