Comparaison de méthodes d'extraction de mots-clés non supervisées

Alaric Tabariès[1] et David Reymond[2]

IMSIC, Université de Toulon, Toulon, France
alaric-tabaries@etud.univ-tln.fr
IMSIC, Université de Toulon, Toulon, France
david.reymond@univ-tln.fr

Résumé Avec l'émergence de l'accès libre et gratuit aux données scientifiques, la volumétrie d'informations accessibles par le chercheur augmente de manière exponentielle. Cette nouvelle dynamique informationnelle rend le processus de veille documentaire, essentiel à la recherche scientifique, tant complexe que chronophage. C'est dans ce contexte que l'extraction d'information se pose en tant que service support au prétraitement de la sélection documentaire. En effet, les mots-clés, qui représentent les sujets principaux traités dans un document, sont particulièrement utiles pour distinguer les ressources intéressantes dans un ensemble imposant de documents. Cependant, très peu en sont pourvus. L'extraction automatique de mots-clés permet de remédier à ce problème et montre d'ores et déjà des résultats satisfaisants sur des corpus de référence. Il a cependant été établi que les disciplines scientifiques dont relèvent les documents influent sur les performances des méthodes d'extractions. Dans cet article, nous ciblons en premier lieu le degré qualitatif du résumé et sa suffisance pour avoir recours à des méthodes extractives en vue de mettre à disposition d'une communauté scientifique variée un outil d'extraction automatique adapté.

Mots-clés: Recherche d'information · Extraction de l'information · Mots-clés · Source d'information

1 Introduction

Avec l'émergence de l'accès libre et gratuit aux données scientifiques, la volumétrie d'informations accessibles par le chercheur augmente de manière exponentielle : à ce jour, plus de 1,8 millions de documents scientifiques sont accessibles sur la plateforme d'archivage internationale arXiv soit près de 12% de plus qu'en 2019. Des initiatives françaises existent également : les plateformes HAL et ISTEX connaissent un essor similaire avec, respectivement, 2,5 et 23 millions de documents référencés. C'est dans ce contexte que l'extraction d'information se pose en tant que service support au pré-traitement de la sélection documentaire. En effet, les mots-clés, qui représentent les sujets principaux traités dans un document, sont particulièrement utiles pour distinguer les ressources intéressantes. Cependant, très peu en sont pourvus : nous avons mesuré qu'environ 30% des références sur la plateforme d'archivage HAL (déc. 2020) en possèdent. L'extraction automatique de mots-clés, permet de générer des mots mots-clés issus de l'auteur même du texte, et en ce sens est moins subjective qu'un lecteur/annotateur. Les différentes techniques extractives montrent d'ores et déjà des résultats satisfaisants sur des corpus de référence (Hasan & Ng, 2014). Il a cependant été établi que les disciplines scientifiques dont relèvent les documents influent sur les performances des méthodes d'extractions (Bougouin et al., 2014). Dans cet article, nous ciblons en premier lieu le degré qualitatif du résumé et sa suffisance pour avoir recours à des méthodes extractives en vue de mettre à disposition d'une communauté scientifique variée un outil d'extraction automatique adapté. Ainsi, nous nous interrogeons sur la fréquence d'apparition de mots-clés auteur dans les résumés des articles.

2 Expérimentation

2.1 Contexte

En préfiguration à cette expérimentation, nous avons réalisé une expérimentation sur un jeu de données d'une centaine de documents relevant des disciplines des sciences humaines et sociales dont l'objectif était de comparer les performances de différentes méthodes d'extraction non supervisées (Tabariès, 2020). À l'aide d'un jeu de données élaboré par des étudiants en linguistique qui extrayaient de manière manuelle des mots-clés des résumés pour qualifier un ensemble d'articles, nous avons comparé les résultats avec des outils d'extraction automatique. Nous avons constaté que le degré de recouvrement (la propension à extraire des mots comme l'ont fait les étudiants) des méthodes variait selon les disciplines étudiées, toutefois, le corpus de documents n'était pas assez conséquent et la qualification externe par les béotiens au regard du résumé ne nous permettent pas de poser de véritables conclusions.

Nous présentons ci-après le début d'une continuité de cette expérimentation portant sur un échantillon plus important de références HAL dont les mots-clés sont ceux des auteurs des documents.

2.2 Méthode

Nous interrogeons la plateforme HAL afin de collecter les articles annotés par, à la fois, un résumé et des mots-clés. Nous extrayons par la suite le jeu de données, composé de 4 500 articles scientifiques marqués par les mots-clés auteur, à l'aide d'un échantillonage aléatoire simple sur la collection d'articles constituée. Nous constituons également un second échantillon comportant 12 500 articles caractérisés par, à la fois, un résumé et des mots-clés et dont le texte complet est renseigné.

En moyenne, dans ces jeux de données, un article est caractérisé par 5 mots-clés, dont 2 composés, ainsi qu'un résumé de 120 mots.

Nous réalisons ensuite un traitement qui consiste à nettoyer les données textuelles en supprimant les mots vides avant d'effectuer une lemmatisation des termes à l'aide du lemmatiseur intégré à la librairie Spacy, paramétré selon le modèle *fr_core_news_sm*. Il est important de noter que nous décomposons les mots-clés composés, à titre d'exemple, le mot-clé composé "*analyse de l'image*" sera donc considéré comme deux mots-clés "*analyse*" et "*image*". Nous étudions alors la relation de présence des mots-clés renseignés par l'auteur dans le résumé d'un document. La relation de présence est analysée selon un premier angle d'étude, que l'on peut qualifier de binaire, qui consiste à définir si un mot est présent ou non. Nous analysons ensuite la similarité sémantique entre ces métadonnées en se basant sur le modèle français de wordnet (Sagot, 2017).

2.3 Résultats (en cours)

Sur le premier échantillon étudié, on constate qu'en moyenne, 38% des mots-clés sont présents dans le résumé d'un document. On constate également une disparité tant au niveau du type de document qu'au niveau de la discipline étudiée. Enfin, l'étude de la similarité sémantique telle que nous l'avons appliquée n'étend que très peu les résultats obtenus et nous paraît, par conséquent, peu pertinente à ce stade.

La figure 1 présente les taux de présence des mots-clés dans les résumés obtenus selon les différents types de documents (dans l'ordre, articles, thèses, commentaires, et mémoires) dans le premier échantillon. On constate que les mots-clés annotés aux articles sont moins présents (0.31%) dans le résumé que pour d'autres types de production.

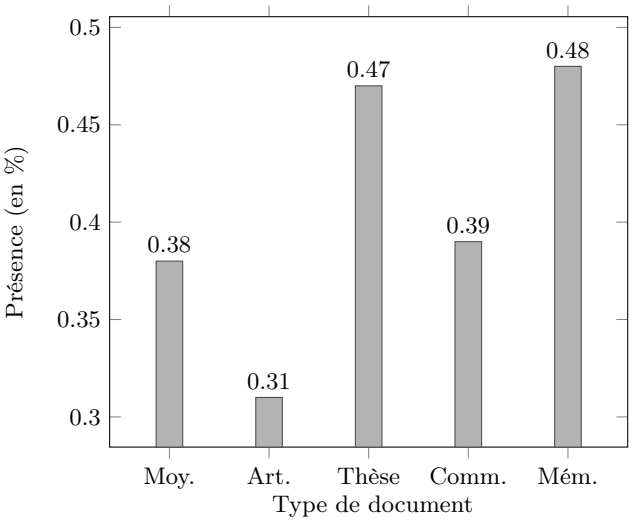

FIGURE 1: Taux de présence des mots-clés dans les résumés en fonction du type de document dans le premier échantillon

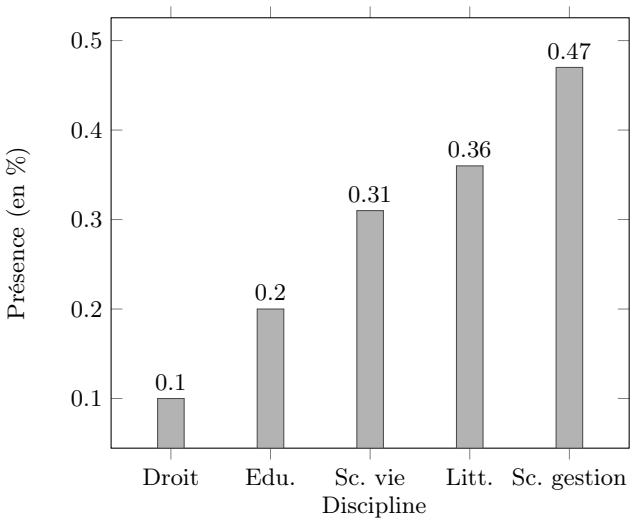

FIGURE 2: Taux de présence des mots-clés dans les résumés d'articles en fonction de la discipline dans le premier échantillon

La figure 2 présente le taux de présence des mots-clés dans les résumés obtenus selon différentes disciplines dans le premier échantillon. On constate un écart très important entre les disciplines étudiées. Pour des raisons pratiques, nous affichons les extrêmes suffisamment représentés dans l'échantillon pour illustrer le propos.

En étudiant le second échantillon, on constate que, en moyenne, 54% des mots-clés annotés par les auteurs sont présents dans les résumés des articles contre 83% dans le texte complet des articles.

3 Conclusion

La première partie de cette expérimentation, c'est-à-dire l'étude de la relation de présence des mots-clés renseignés par l'auteur dans le résumé d'un document est indispensable afin de déterminer la pertinence du résumé seul comme source de l'extraction de mots-clés et établir un niveau de référence en cas d'extension au texte plein. Les premiers résultats montrent qu'une part significative (38 % pour un échantillon généraliste, 54% pour un échantillon plus spécifique) des mots-clés auteur, bien que disparate selon les disciplines, est présente dans le résumé. Les résultats obtenus sont en accord avec le récent article de Lu et al. (2020) dans lequel les auteurs retrouvent sur une étude empirique près de 57% des mots-clés auteur dans le titre et le résumé. De plus, l'écart notable dans le taux de présence entre les différents échantillons étudiés tend à montrer qu'un nombre important de documents sont peu (voire mal) renseignés par le chercheur ce qui souligne en conséquent l'importance de l'accompagnement des chercheurs à la science ouverte au travers d'outils élaborés en ce sens (Reymond & Galliano, 2019). Il est cependant important de poursuivre cette expérimentation en définissant les modalités d'extraction les plus pertinentes selon les disciplines. En extension, nous tenterons alors de rapprocher les termes automatiquement extraits à des descripteurs issus de vocabulaires contrôlés pour tenter d'apposer une dimension qualitative supplémentaire à cette automatisation de procédés documentaires.

Références

Bougouin, A., Boudin, F., & Daille, B. (2014, juillet). Influence des domaines de spécialité dans l'extraction de termes-clés. In *Traitement Automatique des Langues Naturelles (TALN)* (pp. 13–24). Marseille, France.

Hasan, K. S., & Ng, V. (2014, juin). Automatic Keyphrase Extraction : A Survey of the State of the Art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)* (pp. 1262–1273). Baltimore, Maryland : Association for Computational Linguistics. doi: https://doi.org/10.3115/v1/P14-1119

Lu, W., Liu, Z., Huang, Y., Bu, Y., Li, X., & Cheng, Q. (2020, novembre). How do authors select keywords ? A preliminary study of author keyword selection behavior. *Journal of Informetrics*, *14*(4), 101066. Consulté

le 2021-01-15, sur http://www.sciencedirect.com/science/article/pii/S1751157720300134 doi: https://doi.org/10.1016/j.joi.2020.101066

Reymond, D., & Galliano, C. (2019, novembre). *Cartographie de l'expertise des chercheurs de l'Université de Toulon* (Intern report). Université de toulon. Consulté le 2021-01-21, sur https://hal.archives-ouvertes.fr/hal-02643329

Sagot, B. (2017, juin). Représentation de l'information sémantique lexicale : le modèle wordnet et son application au français. *Revue française de linguistique appliquee*, *Vol. XXII*(1), 131–146. Consulté le 2020-12-17, sur https://www.cairn.info/revue-francaise-de-linguistique-appliquee-2017-1-page-131.htm?contenu=resume (Publisher : Publications linguistiques)

Tabariès, A. (2020). Comparaison de méthodes d'extraction de mots-clés non supervisées pour les disciplines des sciences humaines et sociales. In *Communications des apprenti-e-s chercheur-euse-s 2020* (Vol. 7, p. 15). Consulté sur https://jep-taln2020.loria.fr/articles-acceptes/la-conference-sur-hal/