PRES LUNAM
École Doctorale STIM
Sciences et Technologies de
l'Information et Mathématiques

Spécialité : Informatique
Laboratoire : LINA
Equipe : TALN

Extraction automatique de termes-clés à partir de sujets

Adrien Bougouin
adrien.bougouin@univ-nantes.fr

Résumé : Les termes-clés sont les mots ou les expressions polylexicales, c'est-à-dire les séquences de mots grammaticalement correctes, qui représentent le contenu principal d'un document. Ils sont utiles pour diverses applications telles que l'indexation automatique de documents, mais ne sont pas toujours disponibles les documents. Dans cet article nous présentons TopicRank, une méthode à base de graphe pour l'extraction automatique de termes-clés. Cette méthode groupe les termes-clés candidats d'un document en sujets, représente le document par un graphe de sujets, ordonne les sujets selon leur importance au sein du graphe et extrait pour chacun des meilleurs sujets le terme-clé candidat qui le représente le mieux. L'expérience réalisée avec TopicRank montrent une amélioration significative vis-à-vis des méthodes à base de graphe existentes.

Mots clés : *Extraction de termes-clés, méthode non-supervisée, méthode à base de graphe, groupement en sujets, ordonnancement de sujets.*
Collaborations : Projet ANR TermITH (ANR-12-CORD-0029).

1 Introduction

Un ensemble de termes-clés est un ensemble de mots ou d'expressions polylexicales, c'est-à-dire des séquences de mots grammaticalement correctes, permettant de caractériser le contenu principal d'un document. Du fait de leur propriété synthétique, les ensembles de termes-clés sont utilisés dans de nombreuses applications telles que l'indexation automatique de documents [1]. Pourtant, de nombreux documents, tels que ceux accessibles par Internet, ou collections de documents, telles que les actes de conférences, n'en sont pas accompagnées. La quantité de document à traiter est aujourd'hui trop importante pour que l'extraction de leurs termes-clés soit effectuée manuellement. C'est pourquoi de nombreux chercheurs se penchent sur la problématique de l'extraction automatique de termes-clés.

L'extraction automatique de termes-clés consiste à sélectionner dans un document les unités textuelles (mots et séquences de mots) les plus importantes. Parmi les différentes méthodes d'extraction automatique de termes-clés proposées dans la littérature, deux grandes catégories émergent : les méthodes supervisées et les méthodes non-supervisées. Les premières réduisent la tâche d'extraction de termes-clés en une tâche de classification binaire [2], où il s'agit d'attribuer la classe « *terme-clé* » ou « *non terme-clé* » aux différents candidats extraits à partir du document. Dans ce cas, une collection de documents dont nous connaissons les termes-clés est nécessaire pour apprendre à faire la distinction entre « *terme-clé* » et « *non terme-clé* ». Au contraire, les méthodes non-supervisées n'utilisent pas de données ayant nécessitées un travail manuel tel que l'annotation en termes-clés. Ces méthodes se contentent d'ordonner les termes-clés candidats selon un score d'importance attribué en fonction de divers indicateurs tels que leur fréquence ou leur position dans le document analysé. Les méthodes supervisées sont plus performantes que les méthodes non-supervisées, mais leur besoin en données annotées en termes-clés (pour l'apprentissage) pousse les chercheurs à proposer des méthodes non-supervisées compétitives avec les méthodes supervisées.

Les méthodes d'extraction de termes-clés non-supervisées les plus étudiées sont sans conteste celles fondées sur TextRank [3], qui est une méthode d'ordonnancement d'unités textuelles à partir de graphe. Les graphes sont un moyen naturel de représenter les unités textuelles et leurs relations. Pour l'extraction de termes-clés, l'idée est de représenter le document sous la forme d'un graphe dans lequel les nœuds correspondent aux mots et où chaque mot est relié aux mots dont il est proche dans le document. Ensuite, un algorithme ordonne les mots par importance, selon le principe de recommandation : un mot est d'autant plus important s'il est proche d'un grand nombre de mots et si les mots dont il est proche sont eux aussi importants. Les mots les plus importants (mots clés) sont finalement assemblés pour générer les termes-clés du document analysé.

Dans cet article, nous présentons TopicRank, une méthode non-supervisée d'extraction de termes-clés basée sur TextRank. TopicRank groupe les termes-clés candidats selon leur appartenance à un sujet, représente le document sous la forme d'un graphe complet de sujets, ordonne les sujets selon leur importance, puis sélectionne pour chacun des meilleurs sujets son candidat le plus représentatif. La notion de sujet est vague, tant

elle peut exprimer un thème ou un domaine général (par exemple, « le traitement automatique des langues ») ou plus spécifique (par exemple, « l'extraction automatique de termes-clés »). Ici, nous nous intéressons aux sujets les plus spécifiques, car ils caractérisent avec plus de précision le contenu d'un document.

L'article est structuré comme suit. Tout d'abord, nous décrivons le fonctionnement de TopicRank (section 2), puis nous présentons notre évaluation (section 3) et enfin, nous concluons et présentons les perspectives de ce travail (section 4).

2 Extraction de termes-clés avec TopicRank

TopicRank modélise un document sous la forme d'un graphe de sujets. Cette méthode se différencie des autres méthodes à base de graphe, car, plutôt que de chercher les mots importants du document, elle cherche ses sujets importants. Dans un premier temps, les sujets sont identifiés, dans un second temps ils sont ordonnés, puis les candidats les plus représentatifs des k meilleurs sujets sont extraits comme termes-clés.

2.1 Identification des sujets

La première étape de l'identification des sujets consiste à extraire les termes-clés candidats. Pour ce faire, nous suivons Wan et Xiao [4] et Hasan et Ng [5] en extrayant les plus longues séquences de noms, de noms propres et d'adjectifs.

La seconde étape de l'identification des sujets consiste à grouper les termes-clés candidats lorsqu'ils appartiennent au même sujet. Dans le soucis de proposer une méthode ne faisant pas l'usage de données supplémentaires, nous optons pour un groupement lexical des candidats. Ceux-ci sont groupés lorsqu'ils partagent un nombre suffisant (définition d'un seuil) de mots.

2.2 Ordonnancement des sujets

L'ordonnancement des sujets a pour objectif de trouver quels sont ceux qui ont le plus d'importance dans le document analysé. À l'instar de TextRank [3], l'importance des sujets est déterminée à partir d'un graphe.

Les sujets du document analysé composent les nœuds d'un graphe complet. Chaque nœud est lié aux autres par une relation qui représente la force du lien sémantique entre eux. Plus les candidats de deux sujets sont proches dans le document analysé, plus le lien sémantique entre ces deux sujets est fort.

Une fois le graphe construit, l'algorithme d'ordonnancement de TextRank est utilisé pour identifier quels sont les sujets les plus importants du document. Cet ordonnancement se fonde sur le principe de recommandation, ou de vote, c'est-à-dire un sujet est d'autant plus important qu'il est fortement lié à un grand nombre de sujets et que les sujets avec lesquels il est fortement lié sont importants.

2.3 Sélection des termes-clés

La sélection des termes-clés est la dernière étape de TopicRank. Elle consiste à chercher les termes-clés candidats qui représentent le mieux les sujets importants qui sont abordés dans le document. Dans le but de ne pas extraire de termes-clés redondants, un seul candidat est sélectionné par sujet. Ainsi, lorsque k termes-clés doivent être extraits, nous nous assurons de couvrir exactement k sujets disctincts.

La difficulté de ce principe de sélection réside dans la capacité à trouver parmi plusieurs termes-clés candidats d'un même sujet celui qui le représente le mieux. Dans ce travail, nous faisons l'hypothèse qu'un sujet est tout d'abord introduit par sa forme la plus appropriée. Nous sélectionnons donc, pour chaque sujet, le candidat qui apparaît en premier dans le document analysé.

3 Évaluation

Pour valider notre approche, nous réalisons une expérience d'extraction automatique de termes-clés à partir d'une collection de documents de test dont nous connaissons déjà les termes-clés. L'objectif est de proposer la méthode capable d'extraire le plus de termes-clés en commun avec ces termes-clés de référence, associés pour chaque document.

Méthode	WikiNews		
	P	R	F
TF-IDF	33,9	35,9	34,3
TextRank	9,3	8,3	8,6
SingleRank	19,4	20,7	19,7
TopicRank	**35,0**	**37,5**	**35,6**†

TABLE 1 – Résultats de l'extraction de 10 termes-clés avec TF-IDF, TextRank, SingleRank et TopicRank. † indique une amélioration significative de TopicRank vis-à-vis de TextRank et SingleRank, à 0,001 pour le t-test de Student.

3.1 Collection de test

Pour notre expérience, nous utilisons la collection de documents Wikinews [1]. Wikinews est une collection de 100 articles journalistiques français extraits du site Web WikiNews [2] entre les mois de mai et décembre 2012, puis annotés en termes-clés avec l'aide de trois étudiants pour chaque article.

3.2 Mesures d'évaluation

Nous mesurons la performance de chaque méthode d'extraction automatique de termes-clés en termes de précision (P), rappel (R) et F-mesure (F) lorsque 10 termes-clés sont extraits, c'est-à-dire les 10 termes-clés candidats ayant le meilleur score d'importance. La précision mesure la capacité d'une méthode à extraire un minimum de termes-clés erronés, le rappel mesure sa capacité à fournir le plus de termes-clés corrects et la F-mesure est le compromis entre la précision et le rappel.

3.3 Méthodes de référence pour l'extraction de termes-clés

Dans nos expériences, nous comparons TopicRank avec trois autres méthodes non-supervisées d'extraction automatique de termes-clés. Nous choisissons TextRank et SingleRank, les deux méthodes qui sont la fondation des méthodes à base de graphe, ainsi que la pondération TF-IDF, qui consiste à donner un score d'importance élevé aux termes-clés dont les mots sont fréquents et spécifiques au document analysé [3].

3.4 Comparaison de TopicRank avec l'existant

Le tableau 1 montre la performance de TopicRank comparée à celle des trois méthodes de référence. Comparée aux méthodes à base de graphe existantes, TopicRank donne des résultats significativement meilleurs, et confirme ainsi l'importance de grouper les termes-clés candidats représentant le même sujet afin de rassembler des informations (complémentaires) utiles à l'algorithme d'ordonnancement par importance. De plus, il est important de noter le gain de TopicRank vis-à-vis de la méthode TF-IDF, car cette dernière fait l'usage de statistiques extraites de documents supplémentaires, alors que TopicRank utilise uniquement le document à analyser.

4 Conclusion et perspectives

Dans ce travail, nous proposons une méthode à base de graphe pour l'extraction non-supervisée de termes-clés. Cette méthode groupe les termes-clés candidats par sujets, détermine quels sont ceux les plus importants, puis extrait le terme-clé candidat qui représente le mieux chacun des sujets les plus importants. Cette nouvelle méthode offre plusieurs avantages vis-à-vis des précédentes méthodes à base de graphe. Dans un premier temps, le groupement des termes-clés potentiels en sujets distincts permet le rassemblement d'informations utiles qui sont éparpillés avec les autres méthodes. Dans un second temps, le choix d'un seul terme-clé pour représenter l'un des sujets les plus importants permet d'extraire un ensemble ne contenant pas de termes-clés redondants – pour k termes-clés extraits, exactement k sujets sont couverts.

1. https://github.com/adrien-bougouin/WikinewsKeyphraseCorpus
2. http://fr.wikinews.org/
3. Il est important de noter que la notion de spécificité calculée par TF-IDF requière une collection de documents (non annotée). Ceci, rend donc TF-IDF difficile à battre avec une méthode tirant uniquement profit des informations contenues dans le document.

Plusieurs perspectives émergent de ce travail. Tout d'abord, le groupement qui est effectué est un peu naïf, car il ne prend pas en compte l'ambiguïté sémantique des mots et les relations synonymiques que certain entretiennent. Il est donc envisagé d'effectuer un groupement de meilleure qualité, à partir de connaissances linguistiques. Enfin, la stratégie de sélection qui consiste à extraire, pour chaque sujet, le candidat qui apparait en premier dans le document doit être confrontée à d'autres stratégies (sélectionner le candidat le plus fréquent, le candidat le plus similaire aux autres candidats du sujet, etc.). Il est même envisageable de chercher la solution la plus optimale au moyen de technique de Recherche Opérationnelle.

Références

[1] Olena MEDELYAN et Ian H. WITTEN : Domain-Independent Automatic Keyphrase Indexing with Small Training Sets. *Journal of the American Society for Information Science and Technology*, 59(7):1026–1040, may 2008. ISSN 1532-2882. URL http://dx.doi.org/10.1002/asi.v59:7.

[2] Ian H. WITTEN, Gordon W. PAYNTER, Eibe FRANK, Carl GUTWIN et Craig G. Nevill MANNING : KEA : Practical Automatic Keyphrase Extraction. *In Proceedings of the 4th ACM Conference on Digital Libraries*, pages 254–255, New York, NY, USA, 1999. ACM. ISBN 1-58113-145-3. URL http://doi.acm.org/10.1145/313238.313437.

[3] Rada MIHALCEA et Paul TARAU : TextRank : Bringing Order Into Texts. *In* Dekang LIN et Dekai WU, éditeurs : *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[4] Xiaojun WAN et Jianguo XIAO : Single Document Keyphrase Extraction Using Neighborhood Knowledge. *In Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, pages 855–860. AAAI Press, 2008. ISBN 978-1-57735-368-3. URL http://dl.acm.org/citation.cfm?id=1620163.1620205.

[5] Kazi Saidul HASAN et Vincent NG : Conundrums in Unsupervised Keyphrase Extraction : Making Sense of the State-of-the-Art. *In Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, pages 365–373, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1944566.1944608.