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

Spécialité : Informatique
Laboratoire : LINA
Équipe : TALN

Extraction automatique de termes-clés à partir de sujets

Directrice :
Béatrice DAILLE

Adrien BOUGOUIN
E-mail : adrien.bougouin@univ-nantes.fr

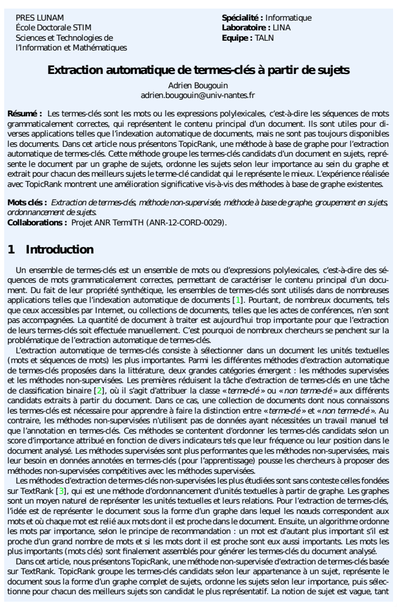
Encadrant :
Florian BOUDIN

Problématique

- Les termes-clés sont des unités textuelles capables de synthétiser le contenu d'un document
- Quelle est la nature linguistique des termes-clés ?
- Comment distinguer les termes-clés des autres unités textuelles de même nature linguistique ?

Solution proposée

1. Extraction des groupes nominaux du document
2. Groupement des groupes nominaux en sujets (concepts)
3. Ordonnancement par importance des sujets
4. Sélection d'un terme-clé pour chaque sujet



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

Sujets

- sujets	*s1*
- meilleurs sujets	
- termes-clés	*s2*
- termes-clés candidats	
- document	*s3*
- ...	

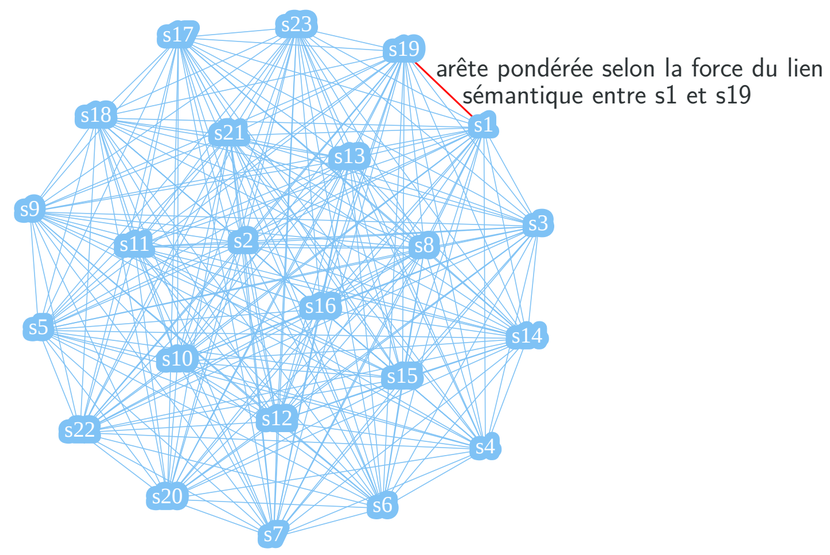
(1) Extraction des plus longues séquences de noms, noms propres et adjectifs

(2) Groupement lexical des groupes nominaux (s'ils partagent $^1/_4$ de leurs mots)

- sujets
- meilleurs sujets
- termes-clés
- termes-clés candidats
- document
- ...

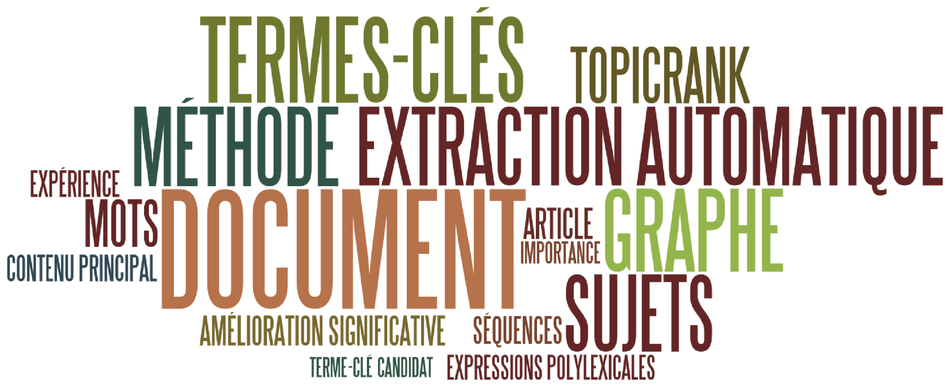
Groupes nominaux

Modélisation du lien sémantique entre chaque sujet



(3) Ordonnancement des sujets, selon TextRank [1]

(4) Pour les k meilleurs sujets : extraction du groupe nominal apparaissant en premier dans le document



Termes-clés

Résultats

- Comparaison de TopicRank avec TextRank [1]
- Application à 100 articles journalistiques (WikiNews)
- Extraction de 10 termes-clés par document

Méthode	Précision	Rappel	F-mesure
TextRank	9,3	8,3	8,6
TopicRank	**35,0**	**37,5**	**35,6**

Perspectives

- Usage de connaissances linguistiques pour le groupement en sujets
- Exploration de diverses stratégies de selection du terme-clé le plus représentatif d'un sujet

Références

[1] Rada Mihalcea and Paul Tarau.
TextRank: Bringing Order Into Texts.
In Dekang Lin and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence (ANR-12-CORD-0029).