TopicRank en domaines de spécialité : participation du LINA à DEFT 2016

Adrien Bougouin Florian Boudin Beatrice Daille

Laboratoire LINA,
Université de Nantes

DEFT 2016, Paris, France

Introduction

- Participation du LINA à la campagne DEFT 2016
  - Système fondé sur TopicRank, une méthode d'extraction de termes-clés à base de graphe
- Originalités de la méthode proposée :
  - Les termes-clés de l'ensemble d'entraînement sont exploités pour construire une représentation du domaine
  - Les termes-clés du domaine et ceux extraits du document sont ordonnés par renforcement mutuel

Plan

Méthode proposée I

Description de TopicRank en 5 étapes

1. **Sélection des mots-clés candidats** : plus longues séquences de noms et d'adjectifs en tant que mots-clés candidats :

$$\text{mots\_cles\_candidat} = (NOM|ADJ)+ \tag{1}$$

2. **Groupement des candidats similaires en sujets** : deux candidats c_i et c_j sont jugés similaires lorsqu'ils partagent au moins un quart de leurs mots racinisés :

$$\text{sim}(c_i, c_j) = \frac{|\text{Porter}(c_i) \cap \text{Porter}(c_j)|}{|\text{Porter}(c_i) \cup \text{Porter}(c_j)|} \tag{2}$$

$$\forall c_i, c_j \in CANDIDATS, c_j \in \text{sujet}(c_i) \Rightarrow \text{sim}(c_i, c_j) \geq \frac{1}{4} \tag{3}$$

Le groupement est réalisé avec un algorithme de groupement hiérarchique agglomératif.

Méthode proposée II

Description de TopicRank en 5 étapes

3. **Construction du graphe** : le document est représenté par un graphe complet $G = (N, A \subseteq N \times N)$ où les nœuds N sont les sujets. Chaque sujet $n \in N$ est connecté aux autres par une arête pondérée $a \in A$ selon la force du lien sémantique entre les sujets :

$$\text{poids}(n_i, n_j) = \sum_{c_i \in n_i} \sum_{c_j \in n_j} \text{distance}(c_i, c_j) \tag{4}$$

$$\text{distance}(c_i, c_j) = \sum_{p_i \in \text{positions}(c_i)} \sum_{p_j \in \text{positions}(c_j)} \frac{1}{|p_i - p_j|} \tag{5}$$

Plus faible est la distance entre les mots-clés candidats de deux sujets dans le document, plus élevé est le poids de l'arête entre les deux sujets.

Méthode proposée III

Description de TopicRank en 5 étapes

4. **Ordonnancement des sujets** : les sujets sont ordonnés par importance selon le principe de recommandation. Plus un sujet est fortement connecté à un grand nombre de sujets, plus il gagne d'importance, et plus les sujets avec lesquels il est fortement connecté sont importants, plus l'importance qu'il gagne est forte :

$$\text{importance}(n_i) = (1-\lambda) + \lambda \times \sum_{n_j \in A(n_i)} \frac{\text{poids}(n_i, n_j) \times \text{importance}(n_j)}{\sum_{n_k \in A(n_j)} \text{poids}(n_j, n_k)} \quad (6)$$

   Où λ est un facteur de lissage fixé à 0,85.

5. **Extraction des mots-clés** : un unique terme-clé (celui qui apparaît en premier dans le document) est extrait pour chacun des k plus importants sujets.

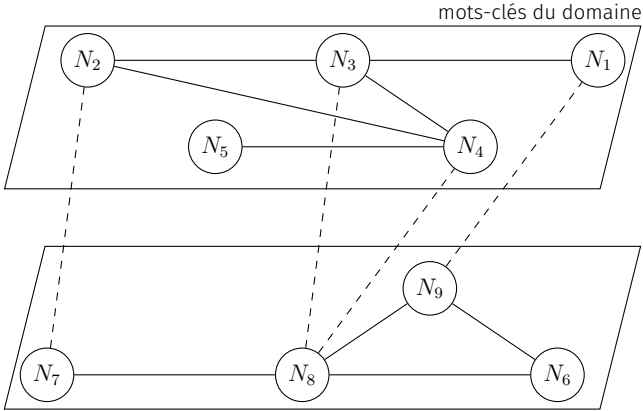
Méthode proposée I

Modifications apportées à TopicRank

- La construction du graphe étend le graphe de sujet en l'unifiant à un graphe des mots-clés de référence du domaine
- L'ordonnancement est conjoint entre les sujets du document et les mots-clés du domaine
- La sélection des mots-clés ajoute la possibilité de puiser dans le graphe du domaine

Méthode proposée II

Modifications apportées à TopicRank



Légende :

——— : $A_{interne}$

- - - - : $A_{externe}$

Méthode proposée III

Modifications apportées à TopicRank

- Deux sujets ou deux mots-clés du domaine sont connectés lorsqu'ils apparaissent dans le même contexte et leur arête est pondérée par le nombre de fois que cela se produit
  - Pour les sujets, le contexte est une phrase du document
  - Pour les mots-clés du domaine, le contexte est l'ensemble des mots-clés du domaine d'un document d'apprentissage

Méthode proposée

Ordonnancement conjoint des sujets et des mots-clés du domaine

$$\text{importance}(s_i) = (1 - \lambda_s)\ R_{\textit{externe}}(s_i) + \lambda_s\ R_{\textit{interne}}(s_i) \tag{7}$$

$$\text{importance}(m_i) = (1 - \lambda_m)\ R_{\textit{externe}}(m_i) + \lambda_m\ R_{\textit{interne}}(m_i) \tag{8}$$

$$R_{\textit{interne}}(n_i) = \sum_{n_j \in A_{\textit{interne}}(n_i)} \frac{\text{poids}(n_i, n_j)\text{importance}(n_j)}{\sum_{n_k \in A_{\textit{interne}}(n_j)} \text{poids}(n_j, n_k)} \tag{9}$$

$$R_{\textit{externe}}(n_i) = \sum_{n_j \in A_{\textit{externe}}(n_i)} \frac{\text{importance}(n_j)}{|A_{\textit{externe}}(n_j)|} \tag{10}$$

Où λ_s et λ_m sont deux facteurs de lissage définis empiriquement pour l'ordonnancement par importance des sujets et des mots-clés du domaine, respectivement.

Plan

Paramètres expérimentaux

- Trois runs/variantes de notre méthode
  - M1 méthode proposée
  - V1.1 Extraction seule (mots-clés extraits du document)
  - V1.2 Assignement seul (mots-clés extraits du domaine)
- Les trois méthodes extraient les 10 meilleurs mots-clés

Résultats I

Méthode	Archéologie			Chimie		
	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
M1	49,86	31,16	37,28	20,87	17,45	18,11
V1.1	43,63	26,63	32,17	15,77	13,10	13,60
V1.2	**53,77**	**33,46**	**40,11**	**21,15**	**17,54**	**18,28**

Méthode	Linguistique			Sciences de l'information		
	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
M1	22,23	24,87	23,24	20,61	20,65	20,21
V1.1	13,77	15,56	14,47	15,67	15,87	15,39
V1.2	**23,16**	**25,85**	**24,19**	**21,93**	**21,83**	**21,45**

Résultats II

	Archéologie		Chimie		Linguistique		Sciences de l'information	
Rang	Équipe	F-mesure	Équipe	F-mesure	Équipe	F-mesure	Équipe	F-mesure
1	EXENSA	45,59	EXENSA	21,46	EBSIUM	31,75	EBSIUM	28,98,
2	LIMSI	43,26	EBSIUM	21,07	EXENSA	26,30	EXENSA	23,86
3	**LINA**	**40,11**	**LINA**	**18,28**	**LINA**	**24,19**	**LINA**	**21,45**
4	EBSIUM	34,96	LIPN	15,31	LIPN	19,07	LIPN	15,34
5	LIPN	30,75	LIMSI	15,29	LIMSI	15,63	LIMSI	12,49

Plan

Discussion

- Méthode fondé sur TopicRank et ajoutant la possibilité d'assigner des mots-clés du domaine
- Utilisation des mots-clés de l'ensemble d'entraînement en lieu et place des thésaurus
- Résultats intéressants, avec la variante effectuant l'assignement seul plus performante