État de l'art des méthodes d'extraction automatique de termes-clés

Adrien Bougouin, LINA - UMR CNRS 6241
adrien.bougouin@univ-nantes.fr

20e conférence TALN 15e édition RECITAL 17 au 21 Juin 2013 LES SABLES D'OLONNE www.taln2013.org

Extraction de termes-clés

L'**extraction de termes-clés** consiste à sélectionner les locutions (**termes candidats**) les plus représentatives d'un document.

Deux catégories de méthodes :
- non-supervisées
- supervisées

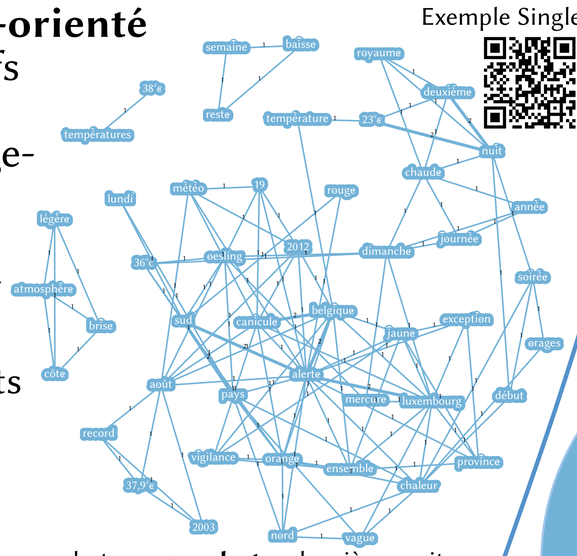
Diverses applications :
- indexation automatique
- résumé automatique
- classification de document

non-supervisées

Méthodes à base de graphe

- document = **graphe non-orienté**
- noeuds = noms et adjectifs
- liens = co-occurrences
- **mots ordonnés** avec PageRank
- **termes-clés =**
  - k meilleurs mots + concaténation si possible (**TextRank [5]**)
  - *k* meilleurs termes-candidats en fonction de la somme du score PageRank de leurs mots (**SingleRank [11]**)

Exemple SingleRank



Termes-clés :
alerte orange ; alerte jaune ; alerte rouge ; **alerte** ; deuxième nuit ; **août 2012** ; août 2003 ; vigilance orange ; légère brise ; **luxembourg**

Méthodes par regroupement

Utilisation de **groupes sémantiques** pour couvrir au mieux tous les aspects du document.

Matsuo et Ishizuka [4]

1. Regroupement des mots fréquents
2. Estimation de la probabilité de co-occurrence d'un terme candidat avec chaque groupe
3. Sélection des termes candidats qui co-occurrent plus que selon toute probabilité avec les groupes

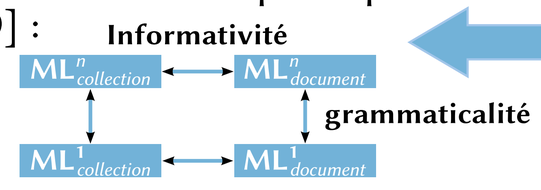
KeyCluster [3]

1. Regroupement des mots
2. Identification du centroïde de chaque groupe
3. Sélection des termes candidats qui contiennent un ou plusieurs centroïdes

Méthodes statistiques

TF-IDF [2] et **Likey [6] :**
Un terme candidat est un terme-clé si les mots qu'il contient sont :
- **fréquents** dans le document
- **spécifiques** au document

Calcul de la **divergence Kullback-Leibler** entre trois **modèles de langue** (ML) pour sélectionner les termes candidats qui respectent deux propriétés [9] :



Document à analyser

1

Pré-traitements :
- segmentation en phrases
- segmentation en mots
- POS tagging

2

Extraction de termes candidats

3 3

Collection

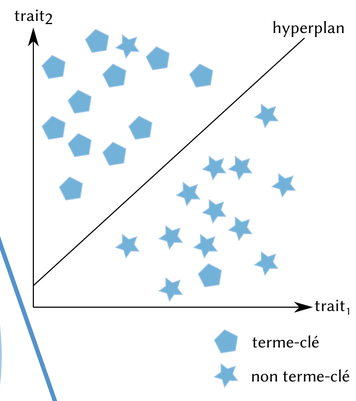
+ annotations

supervisées

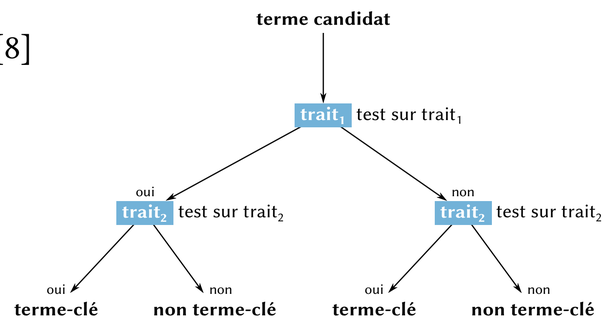
Classifieurs

Modèle probabiliste :
- classifieur naïf bayésien [12]
- modèle d'entropie maximale [8]

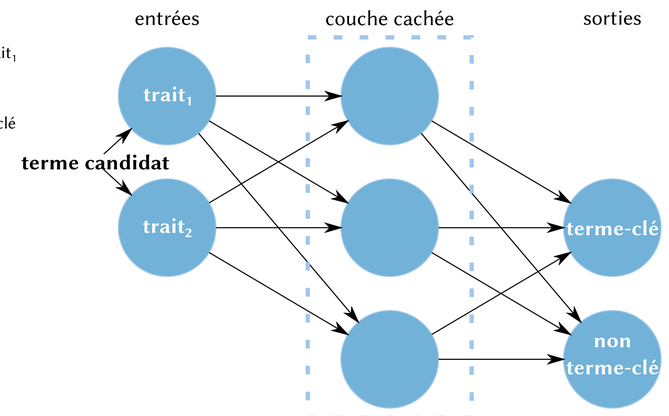
Séparateur à Vaste Marge [1]



Arbre de décision [10]



Réseau de neurones [7]



entraînés à partir de

Traits

- fréquence (TF)
- inverse de la fréquence documentaire (IDF)
- position de la première occurrence
- position de la dernière occurrence
- partie du discours (nom, adjectif, etc.)
- catégorie syntagmatique (syntagme nominal, syntagme verbal, etc.)
- taille (en nombre de mots)
- entité nommée (personne, lieu, pays, organisme, etc.)
- structure du document (résumé, introduction, ...), etc.

Conclusion

De nombreuses méthodes.

⟶ Quelques points communs :
- pré-traitements
- extraction des termes candidats

⟶ Diverses approches :
- usage de groupes sémantiques
- usage d'un graphe
- entraînement de classifieurs, etc.

Références

[1] Eichler, K. et Neumann, G. : DFKI KeyWE : Ranking Keyphrases Extracted from Scientific Articles.
[2] Jones K.S. : A Statistical Interpretation of Term Specificity and its Application in Retrieval.
[3] Liu Z., Li P., Zheng Y. et Sun M. : Clustering to Find Exemplar Terms for Keyphrase Extraction.
[4] Matsuo Y. et Ishizuka M. : Keyword Extraction from a Single Document Using Word Co-occurrence Statistical Information.
[5] Mihalcea R. et Tarau P. : TextRank : Bringing Order Into Texts.
[6] Paukkeri M.S et Honkela T. : Likey : Unsupervised Language-Independent Keyphrase Extraction.
[7] Sarkar K., Nasipuri M. et Ghose S. : A New Approach to Keyphrase Extraction Using Neural Networks.
[8] Sujian L., Houfeng W., Shiwen Y. et Chengsheng X. : News-Oriented Keyword Indexing with Maximum Entropy Principle.
[9] Tomokiyo T. et Hurst, M. : A Language Model Approach to Keyphrase Extraction.
[10] Turney P.D. : Learning Algorithms for Keyphrase Extraction.
[11] Wan X. et Xiao J. : Single Document Keyphrase Extraction Using Neighborhood Knowledge.
[12] Witten I.H., Paynter G.W., Frank E., Gutwin C. et Nevill-Manning C.G. : KEA : Practical Automatic Keyphrase Extraction.