

Recherche d'information cross-lingue

Applications multilingues - module X9IT100

Florian Boudin

Département informatique, Université de Nantes

Révision 1 du 30 juillet 2012

Préface

- ▶ Volume horaire (2h40)
- ▶ Notions abordées dans ce cours
 - ▶ Rappels des notions de RI
 - ▶ Les difficultés liées aux langues
 - ▶ La RI cross-langue
- ▶ Ce cours est basé sur le livre *Cross-Language Information Retrieval* de Jian-Yun Nie [Nie10].

Introduction

- ▶ La **Recherche d'Information** (RI) fait partie intégrante de notre vie quotidienne.
- ▶ Dans la plupart des cas, nous recherchons des documents rédigés dans notre langue maternelle, en général celle utilisée dans la requête.
- ▶ **Cependant...**
 - ▶ L'information pertinente n'est pas toujours disponible dans notre langue maternelle.
 - ▶ Le web offre une mine d'information riche et multilingue à laquelle nous souhaitons avoir accès.
- ▶ Émergence de la problématique de la RI cross-lingue

Plan

Introduction

Rappels des notions de RI

Les difficultés liées aux langues

La RI cross-lingue

Le processus de recherche d'information

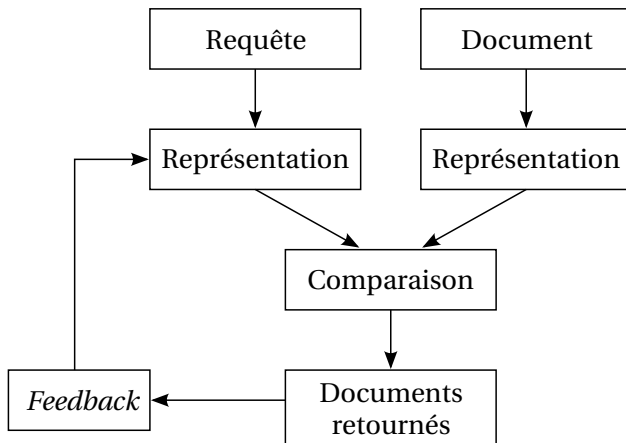


FIGURE : Processus de recherche d'information (Figure 1.1 de [Nie10]).

Modèles de RI I

- ▶ Un modèle définit la représentation des documents et des requêtes ainsi que la fonction de pondération.
- ▶ La plupart des modèles sont construits sur la notion de *terme*, qui peut être un mot (e.g. *computer*), un stem (e.g. *comput*) ou une expression multimots (e.g. *computer system*).
- ▶ **Modèle booléen**
 - ▶ Les documents sont représentés par une conjonction de termes, e.g. $D = t_1 \wedge t_2 \wedge t_3$ qui signifie que les termes t_1 , t_2 et t_3 apparaissent dans D .
 - ▶ Une requête est représentée par une expression booléenne, e.g. $Q = (t_1 \wedge t_2) \vee t_3$.
 - ▶ Un document est considéré comme pertinent si et seulement si il y a l'implication logique $D \rightarrow P$.

Modèles de RI II

► **Modèle vectoriel** [SWY75, SM84]

- Les documents et requêtes sont représentés par des vecteurs dans un espace vectoriel composé de tous les termes.
- Dans chaque vecteur, chaque élément (d_i ou q_i , $1 \leq i \leq n$) représente le poids du terme.

espace vectoriel : (t_1, t_2, \dots, t_n)

document : (d_1, d_2, \dots, d_n)

requête : (q_1, q_2, \dots, q_n)

- Les poids des termes peuvent être binaires, i.e. 1 pour la présence et 0 l'absence, ou calculés avec $tf \times idf$.
- La pertinence des documents est habituellement calculée avec une mesure de similarité cosinus.

Modèles de RI III

► **Modèle probabiliste [RJ76]**

- Le score de pertinence d'un document D par rapport à une requête Q est estimé à partir de $P(\text{pertinence}|D, Q)$.
- Okapi BM25 [RWHB⁺92] est le modèle le plus utilisé.

$$\text{score}(D, Q) = \sum_{i=1}^n \text{idf}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})}$$

- où $f(q_i, D)$ est la fréquence du terme q_i dans D , avgdl est la longueur moyenne des documents, $k_1 \in [1.2, 2.0]$ et $b = 0.75$.

Modèles de RI IV

- ▶ **Modèle de langue** [PC98]

- ▶ Utiliser $P(D|Q)$ pour estimer le score de pertinence d'un document D par rapport à une requête Q .
- ▶ Avec le théorème de Bayes, nous avons :

$$P(D|Q) = \frac{P(Q|D)P(D)}{P(Q)} \propto P(Q|D)P(D)$$

- ▶ En considérant les termes de la requêtes comme indépendants :

$$P(Q|D) = \sum_{t_i \in Q} P(t_i|D)$$

- ▶ $P(t_i|D)$ est estimée par un modèle de langue du document.

Évaluation

$$\text{precision} = \frac{\# \text{ documents pertinents retrouvés}}{\# \text{ documents retrouvés}}$$

$$\text{rappel} = \frac{\# \text{ documents pertinents retrouvés}}{\# \text{ documents pertinents dans la collection}}$$

$$\text{MAP} = \overbrace{\frac{1}{M} \sum_{j=1}^M}^{\forall \text{ requêtes}} \underbrace{\left(\frac{1}{N_j} \sum_{i=1}^{N_j} pr(d_{ij}) \right)}_{\substack{\text{moyenne des précisions} \\ \text{aux rangs des documents} \\ \text{pertinents}}}$$

Plan

Introduction

Rappels des notions de RI

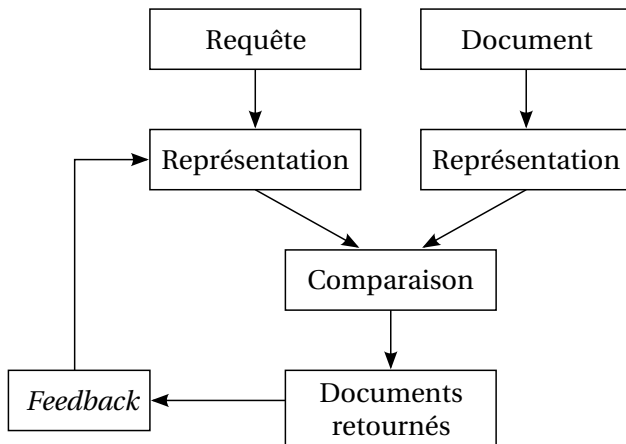
Les difficultés liées aux langues

La RI cross-lingue

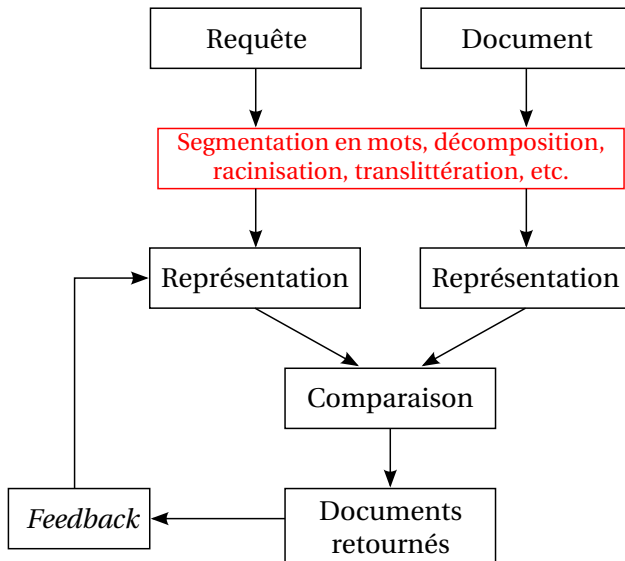
Introduction

- ▶ Les travaux en RI ont longtemps porté uniquement sur les langues européennes.
- ▶ Cette situation a changée avec l'avènement du web et la disponibilité de grandes collections de documents dans de nombreuses langues.
- ▶ Les traitements “basiques” développés pour les langues européennes sont en partie ré-utilisable pour d'autres langues, mais certaines nécessitent des traitements spécifiques.

Le processus de recherche d'information



Le processus de recherche d'information



Langues européennes

Stemming (racinisation)

- ▶ Porter [Por80] et Krovetz [Kro93] pour l'anglais.
- ▶ Snowball¹ : extension de l'algorithme de Porter à 15 langues.
 - ▶ **Langues romanes** (fr, es, pt, it, ro)
e.g. contradictoires → contradictoir (fr)
 - ▶ **Langues germaniques** (de, nl)
e.g. aufeinanderschlagen → aufeinanderschlug (de)
 - ▶ **Langues scandinaves** (se, no, da)
e.g. klostergården → klostergård (se)
 - ▶ **Autres langues** (ru, fi)
e.g. edeltäjälleen → edeltäj (fi)
- ▶ Permet souvent une meilleure précision mais les moteurs de recherche actuels ne l'utilisent pas.

1. <http://snowball.tartarus.org/>

Langues européennes

Decompounding (décomposition)

- ▶ Dans les langues agglutinantes (e.g. Allemand, Néerlandais, Finnois), les mots se forment à partir d'une racine lexicale à laquelle on peut ajouter un certain nombre d'uffixes.
e.g. *hungerstreiks* (de) est composé de *hunger* (faim), *strieks* (grève) et peut aussi s'écrire en deux mots séparés.
- ▶ De multiples expressions d'un même concept peut engendrer des *mismatches* entre les documents et la requête.
- ▶ Le processus de *decompounding* correspond à la détection des mots constituants.
 - ▶ Difficulté liée à l'ambiguïté des mots, par exemple *hungerstreiks* contient les mots suivants : *erst*, *hung*, *hunger*, *hungers*, *hungerst*, *reik*, *reiks*, *streik*, *streiks*

Langues asiatiques

Découpage en mots

- ▶ Le Chinois, Japonais et Coréen (*CJK languages*) partagent un héritage commun du aux liens culturels et linguistiques entre ces pays.
- ▶ Utilisation d'idéogrammes (cn), de kanjis/kanas (jp) ou de hanjas/hangeul (ko).
- ▶ Une caractéristique des textes chinois et japonais est l'absence d'espaces pour délimiter les mots.

e.g. わたしはフランス人です。 (jp)

→ わたし は フランス人 です
watashi wa furansujin desu

Langues asiatiques

L'ambiguïté du découpage en mots

公	路	局	正	在	治	理	解	放	大	道	路	面	积	水	问	题	。
					治	理											
						理	解										
							解	放									
								放	大								
									大	道							
										道	路						
											路	面					
												面	积				
													积	水			

Aménager

Comprendre

Libération

Elargir

Avenue

Route

couche de surface (road surface)

Superficie

Accumulation

FIGURE : Exemples de découpage en mots [Li06].

Langues asiatiques

Plusieurs types d'ambiguïtés [Wan13]

► Ambiguïté de segmentation

e.g. 薄熙来自

- 薄 / 熙来 / 自 (Bo / Xilai / à partir de)
- 薄 / 熙 / 来自 (Bo / Xi / vient de)
- 薄 / 熙 / 来 / 自 (Bo / Xi / vient / depuis)

► Ambiguïté de catégorisation

e.g. 白雪

- 白雪 neige blanche (nom)
- 白雪 Bai Xue (nom propre de personne)

Autres langues

- ▶ L'Arabe
- ▶ Langues d'Inde

Plan

Introduction

Rappels des notions de RI

Les difficultés liées aux langues

La RI cross-lingue

Introduction I

- ▶ La principale difficulté en RI cross-lingue et multilingue réside dans la représentation des documents et des requêtes.
- ▶ Comment comparer des représentations construites à partir d'informations disponibles dans différentes langues ?

fr Un Boeing 777 d'Asiana s'écrase à l'atterrissage à San Francisco
en Boeing 777 from Seoul crashes on landing at San Francisco airport
jp サンフランシスコ国際空港で6日、ボーイング777型機が着陸に失敗し、炎上した

→ Comment réussir à trouver les informations ci-dessus avec la requête “crash d'avion à San Francisco” ?

Introduction II

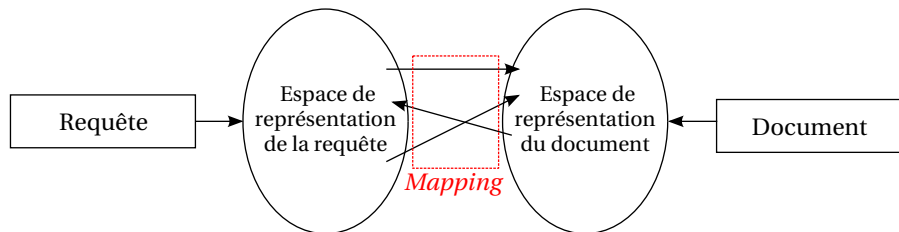


FIGURE : *Mapping* entre les représentations (Figure 1.2 de [Nie10]).

- Utiliser un module de **traduction automatique**.
 1. Mapping de la représentation du document dans celle de la requête : approche par traduction de documents [OH97].
 2. Mapping de la représentation de la requête dans celle du document : approche par traduction de requêtes.
 3. Mapping des représentations de la requête et du document dans un troisième espace (i.e. langue pivot) [RDS99, KK06].

Traduction du document vs de la requête

- ▶ Les méthodes par traduction de la requête sont plus flexibles.
 - ▶ L'utilisateur peut choisir les langues des documents retrouvés.
 - ▶ Dans le cas où il est capable de comprendre la traduction de la requête, il peut la corriger.

References I



Kazuaki Kishida and Noriko Kando.

A hybrid approach to query and document translation using a pivot language for cross-language information retrieval.

In Proceedings of the 6th international conference on Cross-Language Evaluation Forum : accessing Multilingual Information Repositories, CLEF'05, pages 93–101, Berlin, Heidelberg, 2006. Springer-Verlag.



Robert Krovetz.

Viewing morphology as an inference process.

In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '93, pages 191–202, New York, NY, USA, 1993. ACM.

References II



Yiping Li.

Étude des problèmes spécifiques de l'intégration du chinois dans un système de traitement automatique pour les langues européennes.

PhD thesis, 2006.



Jian-Yun Nie.

Cross-Language Information Retrieval.

Synthesis Lectures on Human Language Technologies.
Morgan & Claypool Publishers, 2010.

References III



Douglas W Oard and Paul G Hackett.

Document translation for cross-language text retrieval at the university of maryland.

In *Information Technology : The Sixth Text REtrieval Conference (TREC-6)*, pages 687–696. US Dept. of Commerce, Technology Administration, National Institute of Standards and Technology, 1997.



Jay M. Ponte and W. Bruce Croft.

A language modeling approach to information retrieval.

In *SIGIR*, pages 275–281. ACM, 1998.



Martin F Porter.

An algorithm for suffix stripping.

Program : electronic library and information systems, 14(3) :130–137, 1980.

References IV



Miguel E Ruiz, Anne Diekema, and Páraic Sheridan.
Cindor conceptual interlingua document retrieval : Trec-8
evaluation.
In *TREC*, 1999.




Stephen E Robertson and K Sparck Jones.
Relevance weighting of search terms.
Journal of the American Society for Information science,
27(3) :129–146, 1976.




Stephen E. Robertson, Steve Walker, Micheline
Hancock-Beaulieu, Aarron Gull, and Marianna Lau.
Okapi at trec.
In *TREC*, pages 21–30, 1992.

References V

 Gerard Salton and Michael McGill.
Introduction to Modern Information Retrieval.
McGraw-Hill Book Company, 1984.

 Gerard Salton, A. Wong, and C. S. Yang.
A vector space model for automatic indexing.
Commun. ACM, 18(11) :613–620, 1975.

 Zhen Wang.
Une approche mixte morpho-syntaxique et statistique pour la reconnaissance d'entités nommées en langue chinoise.
In Actes de la 15e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'2013), pages 231–243, Sables d'Olonne, France, 2013.