

# 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
  - ▶ ...
- ▶ 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

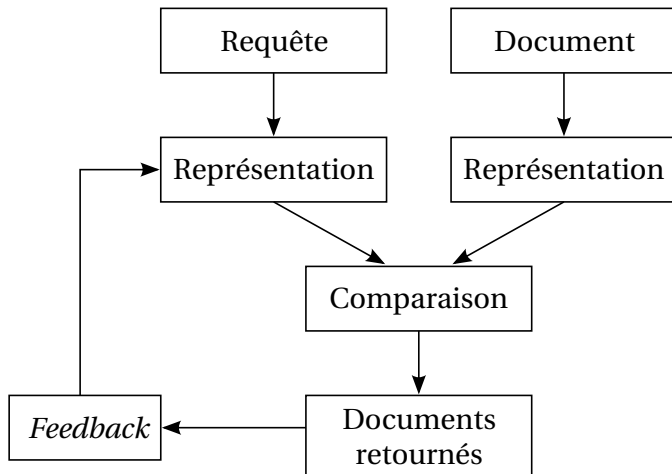
Introduction

Rappels des notions de RI

Les difficultés liées aux langues

Références

# 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<sup>+</sup>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  $D$ ,  $\text{avgdl}$  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

Références

# 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.
- ▶ Bien que les traitements “basiques” développés pour les langues européennes sont en partie ré-utilisable pour d'autres langues, certaines nécessitent des traitements spécifiques.

# Langues Européennes

## ► **Stemming** (racinisation)

- Porter [Por80] et Krovetz [Kro93] pour l'anglais.
- Snowball<sup>1</sup> : extension de l'algorithme de Porter à 15 langues.
- Les méthodes de stemming permettent souvent une amélioration de la performance mais pas toujours (les moteurs de recherche actuels ne l'utilisent pas).

## ► **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'affixes.
- *hungerstreiks* (de) est composé de *hunger* (faim), *strieks* (grève) et peut aussi s'écrire en deux mots séparés.

---

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



# References I



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.



Jian-Yun Nie.

*Cross-Language Information Retrieval*.

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



Jay M. Ponte and W. Bruce Croft.

A language modeling approach to information retrieval.

In *SIGIR*, pages 275–281. ACM, 1998.

# References II



Martin F Porter.

An algorithm for suffix stripping.

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



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 III

-  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.