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
  - Recherche d'Information (rappels)
  -
- 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

Le processus de recherche d'information

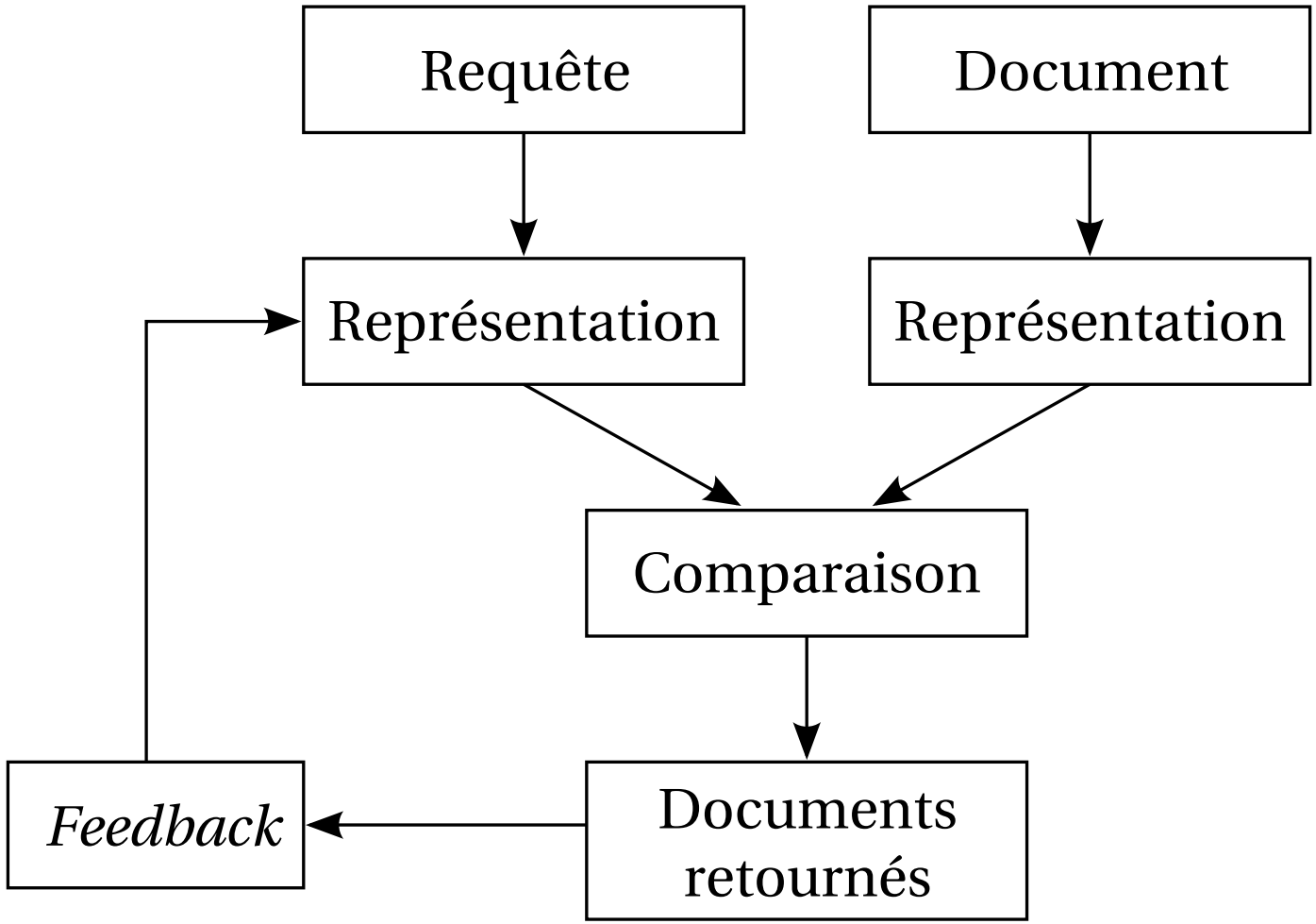

FIGURE : Processus de recherche d'information (Figure 1.1 de [Nie10]).

Modèles de RI

- 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*).
- Modèles de RI classiques
  - Modèle booléen
  - Modèle vectoriel, i.a. [SWY75]
  - Modèle probabiliste, i.a. [RJ76]
  - Modèle de langue, i.a. [PC98]

Évaluation I

- Plusieurs mesures ont été proposées pour évaluer l'efficacité des systèmes de RI.
- Les mesures de base sont la précision et le rappel.

$$\text{precision} = \frac{\#\text{ documents pertinents retrouvés}}{\#\text{ documents retrouvés}} \quad (1)$$

$$\text{rappel} = \frac{\#\text{ documents pertinents retrouvés}}{\#\text{ documents pertinents dans la collection}} \quad (2)$$

Évaluation II

- Une autre mesure largement utilisée est la *Mean Average Precision* (MAP)

$$\text{MAP} = \frac{1}{M} \sum_{j=1}^{M} \left(\frac{1}{N_j} \sum_{i=1}^{N_j} pr(d_{ij}) \right) \tag{3}$$

$$pr(d_{ij}) = \begin{cases} \frac{r_{ni}}{n_i} & \text{si } n_i < \text{MAX} \\ 0 & \text{autrement} \end{cases} \tag{4}$$

Plan

References

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.

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.

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