

# 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 problématique de la RI cross-lingue
  - ▶ La méthodes de traduction en RI cross-lingue
  - ▶ Le besoin de méthodes de RI cross-lingue et multilingue
  - ▶ ...
- ▶ 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 problématique de la RI cross-lingue

La méthodes de traduction en RI cross-lingue

Le besoin de méthodes de RI cross-lingue et multilingue

# Le processus de recherche d'information

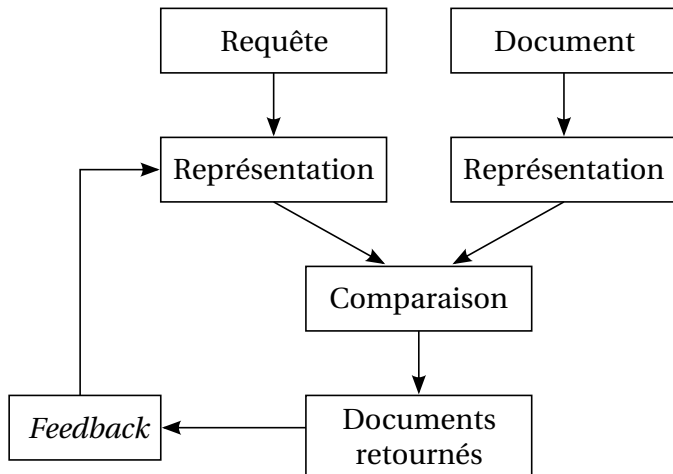


FIGURE : Processus de recherche d'information (Figure 1.1 de [Nie10]).

# 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

La problématique de la RI cross-lingue

La méthodes de traduction en RI cross-lingue

Le besoin de méthodes de RI cross-lingue et multilingue

# 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

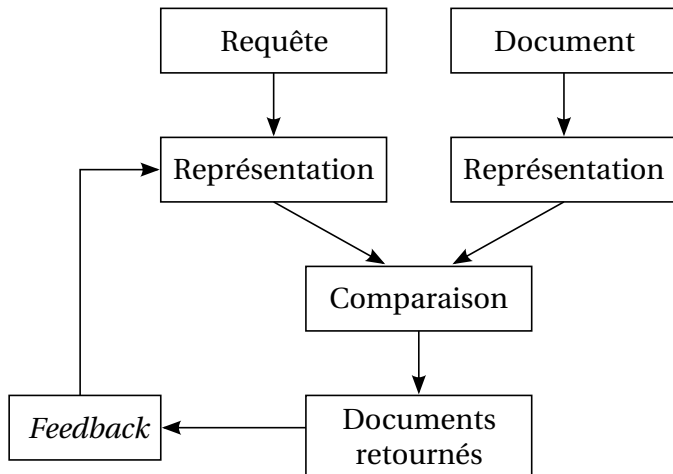


FIGURE : Processus de recherche d'information (Figure 1.1 de [Nie10]).

# Le processus de recherche d'information

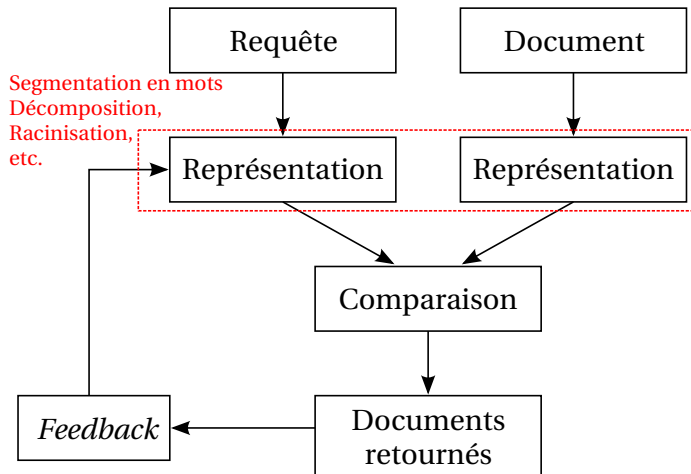


FIGURE : Processus de recherche d'information (Figure 1.1 de [Nie10]).

# 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.
  - ▶ **Langues romanes** (fr, es, pt, it, ro)  
e.g. contradictoires → contradictoir (fr)
  - ▶ **Langues germaniques** (de, nl)  
e.g. aufeinanderschließen → aufeinanderschluss (de)
  - ▶ **Langues scandinaves** (se, no, da)  
e.g. klostergården → klostergård (se)
  - ▶ **Autres langues** (ru, fi)  
e.g. edeltäjälleen → edeltäjä (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'affixes.  
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.

# 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									Aménager
		理	解									Comprendre
			解	放								Libération
				放	大							Élargir
					大	道						Avenue
						道	路					Route
							路	面				Couche de surface
								面	积			Superficie
									积	水		Accumulation

TABLE : Exemple illustrant la difficulté du découpage en mots [Li06].

# Langues asiatiques

## Plusieurs types d'ambiguïtés [Wan13]

### ► Ambiguïté de segmentation en 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)

e.g. 薄熙来自 (cn)

→ 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 (cn)

→ 白雪 (neige blanche, nom)

→ 白雪 (Bai Xue, nom propre de personne)

# Autres langues

- Problématiques de la langue Arabe

- Lettre qui change de forme en fonction de sa position.

e.g. Forme isolée ع (ayin)

→ عين (oeil)

→ بعد (après)

→ اصبع (doigt)

Position in word:	Isolated	Final	Medial	Initial
Glyph form:	ع	ع	ع	ع

- Les voyelles peuvent être omises

...

- Beaucoup d'autres langues (et de problèmes spécifiques!).

# Plan

Introduction

Rappels des notions de RI

Les difficultés liées aux langues

La problématique de la RI cross-lingue

La méthodes de traduction en RI cross-lingue

Le besoin de méthodes de RI cross-lingue et multilingue

# Introduction

- ▶ 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'un Boeing 777 à San Francisco” ?

# Introduction

- ▶ 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'un Boeing 777 à San Francisco” ?

# Mapping entre les représentations

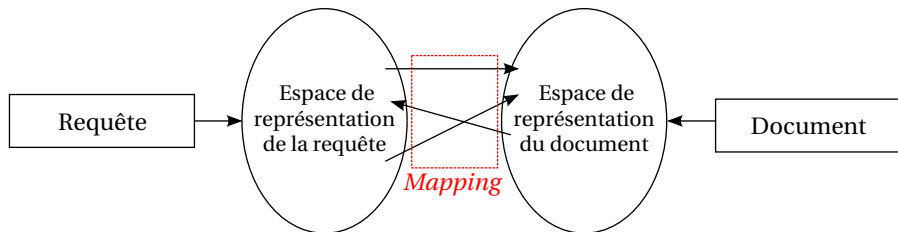


FIGURE : *Mapping* entre les représentations (Figure 1.2 de [Nie10]).

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].



# Le processus de RI cross-lingue

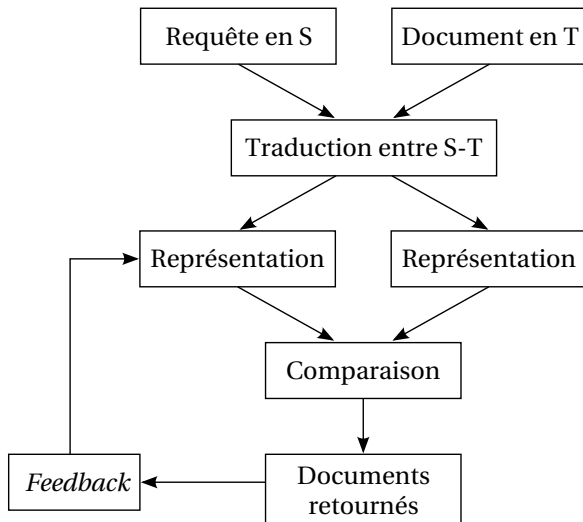


FIGURE : Processus de RI cross-lingue (Figure 1.3 de [Nie10]).

# Traduction du document ou de la requête ? I

- ▶ Les méthodes par traduction de la requête sont plus flexibles.
  - ▶ L'utilisateur peut choisir les langues des documents à chercher.
  - ▶ Dans le cas où il est capable de comprendre la requête traduite, il peut la corriger ou l'étendre.
- ▶ **Cependant**, la traduction automatique de la requête peut être ambiguë de part la taille très limitée du contexte.
  - ▶ Erreurs de segmentation pre-traduction  
任天堂 (nintendo) (jp) → 任 (responsabilité) 天 (ciel) 堂 (salle)
  - ▶ Erreurs de (sur-)traduction  
*perfume japanese band* (en) → parfum groupe japonais (fr)

# Traduction du document ou de la requête ? II

- ▶ La traduction du document est plus robuste mais les travaux précédents ne montrent pas de différence [McC99].
    - ▶ Les systèmes de traduction automatique n'utilisent qu'une partie du contexte riche du document (traduction phrase à phrase).
  - ▶ **De plus**, les documents doivent être au préalable traduits dans toutes les langues possibles.
    - ▶ Irréaliste avec les capacités de calcul et de stockage actuelles.
- La majorité des approches en RI cross-lingue utilisent donc la traduction de requête.

# Utilisation d'une langue pivot

- ▶ La traduction directe entre deux langues peut ne pas être possible dans le cas où il n'existe pas de ressources suffisantes.
- ▶ **Mais**, il y a peut-être des ressources disponibles entre ces langues et une troisième langue (e.g. l'Anglais).
- ▶ Deux approches sont possibles :
  1. Traduction de la requête et du document dans la langue pivot.
  2. Traduction de la requête ou du document dans la langue pivot puis dans la langue cible.

# Plan

Introduction

Rappels des notions de RI

Les difficultés liées aux langues

La problématique de la RI cross-lingue

La méthodes de traduction en RI cross-lingue

Le besoin de méthodes de RI cross-lingue et multilingue

# La traduction en RI cross-lingue I

- ▶ En plus des difficultés liées à la RI monolingue, la traduction automatique (TA) est la difficulté principale de la RI cross-lingue et multilingue.
- ▶ La TA peut être nécessaire dans deux étapes :
  1. Sachant une requête en langue A (source), si l'utilisateur recherche des documents en langue B (cible), les termes en langue A doivent être traduits en langue B.
  2. Une fois que l'ensemble des documents pertinents en langue B est retrouvé, l'utilisateur peut vouloir les traduire en langue A de façon à pouvoir les comprendre.

# La traduction en RI cross-lingue II

- ▶ Les deux étapes de traduction ci-dessus sont très différentes.
  - ▶ La seconde est une tâche classique de TA de document.
  - ▶ **Cependant**, la TA n'est pas nécessairement un outil approprié pour la première étape.
- ▶ **Une syntaxe moins stricte est nécessaire**
  - ▶ La tâche de traduction d'une requête en RI cross-lingue n'est pas de la rendre lisible par un humain mais de permettre au système de faire un *matching*.
  - La syntaxe/grammaire n'est donc pas très importante.
- ▶ **Un niveau d'ambiguïté plus important**
  - ▶ Les requêtes sont généralement très courtes (2-3 mots).

# La traduction en RI cross-lingue III

## ► Un effet d'expansion de requêtes

- Le but de la traduction de requête est de produire une représentation de cette dernière dans une autre langue.
- De manière à pouvoir *matcher* les termes de la requête avec ceux du document, il est préférable d'inclure toutes les alternatives de traduction.

## ► La pondération des termes

- Le poids assigné à chacun des termes de la requête reflète l'importance de ce terme lors du *matching*.
- En RI cross-lingue, le poids d'un terme doit refléter son importance mais également la qualité de la traduction.



# La traduction en RI cross-lingue IV

- ▶ Dans la littérature, en plus de la TA, les deux approches suivantes ont été largement testées et évaluées :
    1. Traduction basée sur dictionnaire : les traductions possibles des termes sont issues d'un dictionnaire bilingue [PHKJ01].
    2. Traduction basée sur un corpus parallèle : utilise les relations de traduction entre deux langues des termes [CyN00].
  - ▶ Il n'y a pas de séparation stricte entre les approches, e.g. certaines approches de TA utilisent des dictionnaires.
- En séparant les approches utilisant un système de TA, nous mettons en avant le fait que les systèmes de TA sont dans la plupart des cas utilisés comme une *black box*.

# Plan

Introduction

Rappels des notions de RI

Les difficultés liées aux langues

La problématique de la RI cross-lingue

La méthodes de traduction en RI cross-lingue

Le besoin de méthodes de RI cross-lingue et multilingue

# L'utilité de la RI cross-lingue et multilingue


- ▶ Bien que le besoin de méthodes de RI monolingue ne soit plus à démontrer, celui de méthodes cross-lingue et multilingue peut apparaître moins évident.
  - ▶ L'objectif en RI est d'identifier les informations pertinentes.
    - ▶ La forme de la description de l'information n'a que peu d'importance : un texte, une image, un tableau, etc.
    - ▶ Les informations ne sont utiles que si elles sont compréhensibles.
  - ▶ L'obstacle de la compréhension de documents dans une langue différente est la raison pour laquelle les moteurs de recherche actuels sont monolingues.
- **Cependant**, cet obstacle est en train de tomber avec les progrès faits dans les outils de TA.


# Rechercher des documents dans une autre langue I

- ▶ Indépendamment de la disponibilité des outils de TA, il y a plusieurs raisons de rechercher des documents sans égard pour la langue.
- ▶ L'information recherchée peut être sous une forme directement compréhensible (e.g. image). Les méthodes de recherche d'images sont en grande partie basées sur leurs descriptions textuelles.
- ▶ L'information recherchée peut ne pas exister dans la langue de l'utilisateur.
- ▶ Les documents peuvent mélanger plusieurs langues.



# References I

 Jiang Chen and Jian yun Nie.  
Parallel web text mining for cross-language ir.  
In *RIAO*, pages 62–77, 2000.

 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.

# References II



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.



Yiping Li.

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

PhD thesis, 2006.

# References III



J. Scott McCarley.

Should we translate the documents or the queries in cross-language information retrieval?

In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 208–214, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.



Jian-Yun Nie.

*Cross-Language Information Retrieval*.

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



# References IV



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.

# References V



Ari Pirkola, Turid Hedlund, Heikki Keskustalo, and Kalervo Järvelin.

Dictionary-based cross-language information retrieval : Problems, methods, and research findings.

*Inf. Retr.*, 4(3-4) :209–230, September 2001.



Martin F Porter.

An algorithm for suffix stripping.

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



Miguel E Ruiz, Anne Diekema, and Páraic Sheridan.

Cindor conceptual interlingua document retrieval : Trec-8 evaluation.

In *TREC*, 1999.

# References VI



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, Aaron Gull, and Marianna Lau.  
Okapi at trec.


In *TREC*, pages 21–30, 1992.




Gerard Salton and Michael McGill.

*Introduction to Modern Information Retrieval*.  
McGraw-Hill Book Company, 1984.

# References VII

 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.