Module
Recherche d'information (RI)

USTHB - Master 2 – SII – 2022-2023

Prof. KECHID.

1

Programme du module

2

Chapitre 1
Introduction à la RI

3

1. Introduction

La Recherche d'information (RI) est une branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information « salton 1968 »

1.1. Terminologie

- recherche d'information,
- informatique documentaire
- information retrieval
- textual information retrieval
- document retrieval

1.2. Utilité

Ce domaine est utile par tout le monde

4

1.3. Domaines d'application

- Web
- Bibliothèques numériques «digital library»
- Entreprises
- Nos propres PC

1.4. Domaines de recherche

- Recherche adhoc
- Classification /catégorisation (clustering)
- Question-réponses (Query answering)
- Filtrage d'information (filtering/recommendation)
- Méta-moteurs (data-fusion,Meta-search)
- Résumé automatique (Summarization)
- Croisement de langues (cross language)
- Fouille de textes (Text mining)

…

5

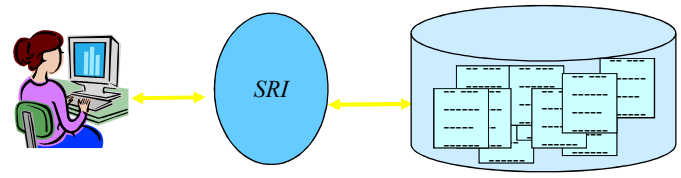
1.5. Quelques systèmes de RI connus

Google yahoo! Ask Microsoft Bing LYCOS WEBCRAWLER

6

2. Définition

Un Système de Recherche d'Information (SRI) est un programme (ensemble de programmes) informatique qui a pour but de sélectionner des **informations pertinentes** répondant à des **besoins des utilisateurs.**



Utilisateur SRI Collection de documents

Dans cette définition il y a 3 notions clès à savoir :

- ✓ L'information
- ✓ Le besoin de l'utilisateur
- ✓ la pertinence

7

- ❑ **L'information :**

  Peut être un texte libre, texte structuré, document, une page web, une image, une vidéo …etc. Dans ce cours nous traitons seulement les documents textuels.

- ❑ **Le besoin de l'utilisateur**

  Connu généralement par le mot « requête », qui exprime le besoin d'information d'un utilisateur. Une requête peut avoir différentes formes selon le modèle utilisé. Souvent elle exprimée par une liste de mots-clés

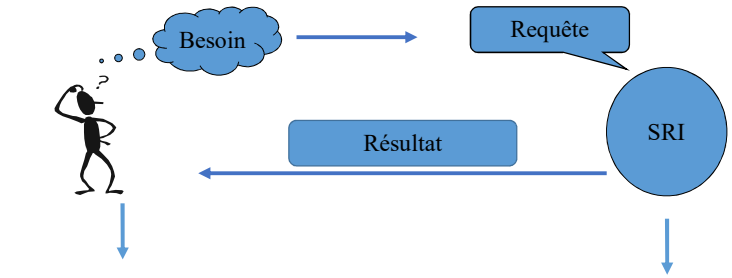
- ❑ **La pertinence**

  C'est une relation de correspondance entre un document et une requête (besoin en information), selon l'utilisateur ou le système. On distingue deux types de pertinences.

  - ➢ Pertinence système (similarité calculée par le système entre une document et une requête)
  - ➢ Pertinence utilisateur (satisfaction de l'utilisateur par le document)

8

2.1. Pertinence utilisateur vs. Pertinence système



Pertinence utilisateur

- **Pertinence subjective (*Subjective rel.*)**
- **Pertinence cognitive (*Cognitive rel.*)**

Pertinence système(system-based relevance)

- **Pertinence algorithmique (*Algorithmic relevance*)**
- **Pertinence objective (*Objective relevance*)**

9

2.2. Problématique de la pertinence

- La pertinence est multidimensionnelle
  - dépend de plusieurs paramètres : l'utilisateur, besoin en information, situations des utilisateurs, …
- La pertinence est graduelle
  - un document A peut être plus pertinent que B
- La pertinence est dynamique
  - peut changer dans le temps, selon l'état de connaissance de l'utilisateur au moment de la recherche
- La pertinence est difficile à automatiser « remplacer l'utilisateur par un système »

10

3. Approche générale de la RI

La vision simple de l'approche de la RI textuelle est de trouver les documents ayant les mêmes mots que la requête :

- La requête est une liste de mots clés
- Le document est une liste de mots clés
- Comparer les mots de chaque document à ceux de la requête
- Sélectionner les documents qui contiennent les mots de la requête.

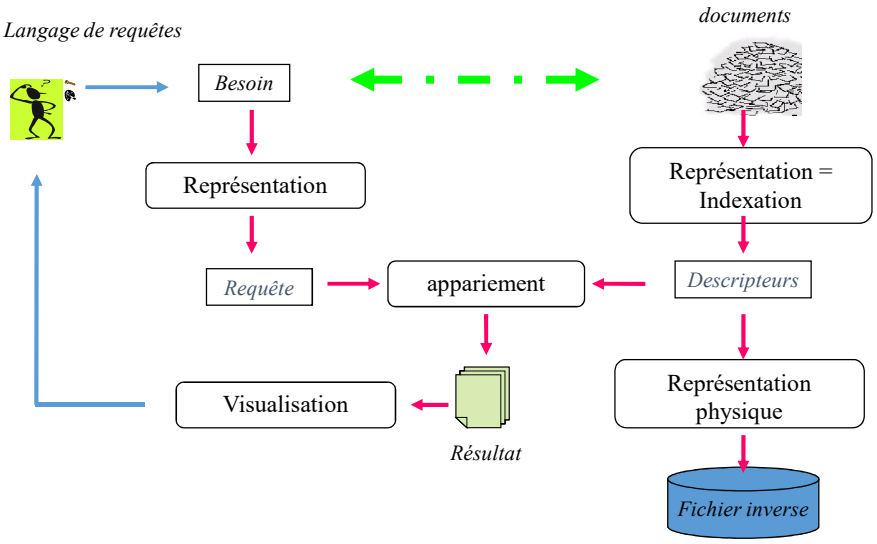
11

4. Un SRI n'est pas une BD

	BD	RI
Données	Attributs-valeurs	Texte libre
Champs	Sémantique claire	Pas de champs (texte libre)
Requête	Définie (algèbre relationnelle, SQL)	(texte, langage naturel, booléen)
Comparaison	Exacte(résultats corrects)	Imprécise (mesure une pertinence)

12

5. Processus de RI



13

5.1. Problématiques :

- *Comment représenter (indexer) un document ?*
- *Comment représenter (indexer) une requête ?*
- *Comment mesurer la pertinence (similarité ou appariement) entre un document est une requête ?*

14

6. Travaux de recherche en RI

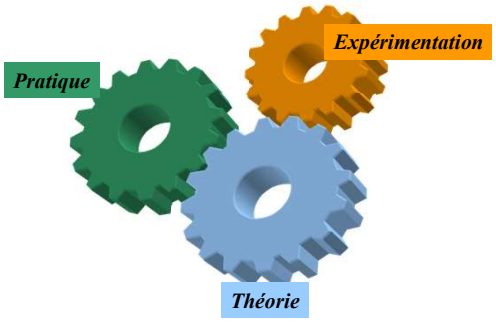
- **Proposer des solutions :**
  modèles, techniques, approches, outils pour répondre à ces problèmes

- **Quels supports théoriques ?**
  Souvent basés sur des théories mathématiques : probabilités, statistiques, ensembles, algèbre, logique floue, analyse de données, …

- **Quel processus pour la validation ?**
  Pour évaluer une approche ou un système de RI, il faut y avoir des donnée d'évaluation (environnement de tests, datatsets, benchmarks).

15

- En RI on a besoin de :
  - ✓ La théorie
  - ✓ La pratique
  - ✓ L'expérimentation



16

7. Conclusion

- La RI est un domaine en pleine expansion de plus en plus important car :
  - ✓ les masses d'information n'arrêtent pas d'augmenter
  - ✓ les demandes d'information (utilisateurs) n'arrêtent pas d'augmenter

17

Références bibliographiques

- **Ouvrages en ligne :**
  - Van Rijsbergen (1977) Information Retrieval, *Butterworths*
  - Baeza-Yates and Ribeiro-Neto, eds. (1999) Modern Information Retrieval *Addison-Wesley*
  - Recherche d'information 2008 : état des lieux et perspectives (M. Boughanem et J. Savoy)
- **Sites des cours :**
  - http://www-labs.iro.umontreal.ca/~nie/IR-book/
  - http://www-labs.iro.umontreal.ca/~nie/IFT6255/
  - http://nlp.stanford.edu/IR-book/information-retrieval.html

18

Conférences et Journaux

- **Conférences**
  - ACM SIGIR : Special interest group on Information Retrieval
  - CIKM : Conference on Information and Knowledge Management
  - ECIR : European Conference on Information Retrieval Research, University of Sunderland, U.K.
  - RIAO : Coupling approaches, coupling media and coupling languages for information retrieval
  - CORIA : Conférence Francophone en Recherche d'Information et Applications
- **Journaux**
  - JASIST : Journal of the American Society for Information Science and Technology
  - IP&M : Information Processing & Management
  - IJODL : International Journal on Digital Libraries
  - JDOC : Journal of Documentation
  - JIR : Journal of Information Retrieval
  - ACM-TOIS : Transactions on Information Systems

19

Fin.

20