Recherche d'information

Introduction

Université de Biskra- Département d'informatique

2ème Master SIOD

2019-2020

Introduction

- Les techniques de la recherche d'information (abrégée en RI ou IR en anglais pour « Information retrieval ») sont directement issues des sciences de l'information et plus précisément de la bibliothéconomie.
- La problématique majeure de cette discipline qui est ancienne et antérieure à l'apparition des ordinateurs a toujours été de permettre un accès rapide aux documents. Ces accès nécessitent plusieurs intermédiaires et de gros moyens.
- Il faut en effet établir un classement des documents existants et sélectionner pour chaque document un jeu de mots clés représentatif. Cette description synthétique par mots clés appelés, « index « , suppose du documentaliste une connaissance suffisante de chaque ouvrage pour pouvoir en traduire le contenu

3

Introduction

- On peut aujourd'hui dire que la recherche d'information (RI) s'est développée et est devenue un champ transdisciplinaire.
- La recherche d'information multimédia, par exemple, qui combine des données de différents types (texte, image, audio, vidéo) présente un panorama des différentes modélisations et interrogations possibles des documents multimédia.
- La notion de document dans le contexte de la RI a connu également une extension au domaine des bases de données dédiées à un accès local, ou bien mises en réseau reliées par des liens hypertextes comme sur la toile du Web.

4

Définitions

- 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]
- Terminologie
  - Recherche d'information, Informatique documentaire
  - Information Retrieval / Textual Information Retrieval / Document Retrieval

5

Définitions

- **Recherche d'Information (RI)**

  Ensemble d'outils et techniques qui permettent de retrouver les documents contenant l'information pertinente à un besoin,

- **Un Système de Recherche d'Information (SRI)**

  Un système de recherche d'information (RI) est un système qui permet de retrouver les documents pertinents à une requête d'utilisateur, à partir d'une base de documents volumineuse.

- Trois notions clés: **documents**, **requête**, **pertinence**.

6

Définitions

- **Requête** : exprime le besoin d'information d'un utilisateur

- **Document** : toute unité qui peut constituer une réponse à une requête,

Un document peut être un texte, un morceau de texte, une page Web, une image, une bande vidéo, etc,

- **Base de documents** : ensemble des documents disponibles

- **Pertinence** : De façon générale, dans document pertinent, l'utilisateur doit pouvoir trouver les informations dont il a besoin.

Sur cette notion le système doit juger si un document doit être donné à l'utilisateur comme réponse ou non

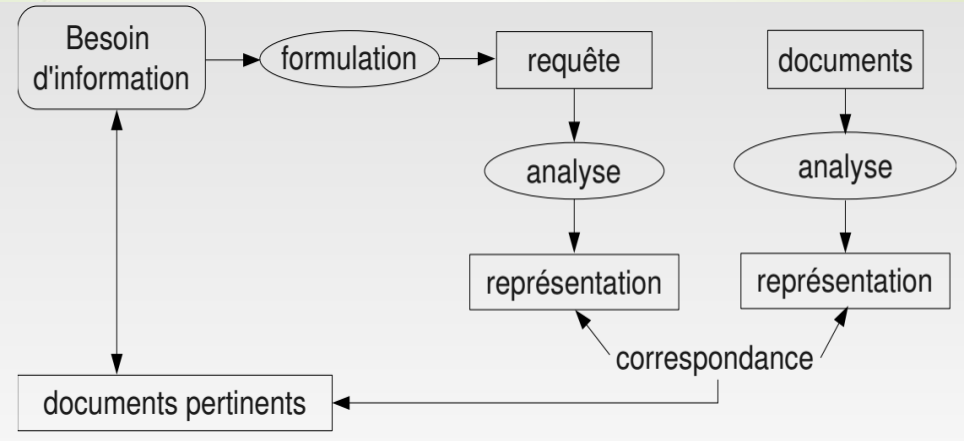
7

Exemples d'applications

- Outils de recherche dans les mails, dans les fichiers, ...
- Systèmes de RI documentaires,
- Systèmes de RI pour les bases de documents d'une entreprise,
- Systèmes de RI sur le Web tels que google, bing ,,,etc.

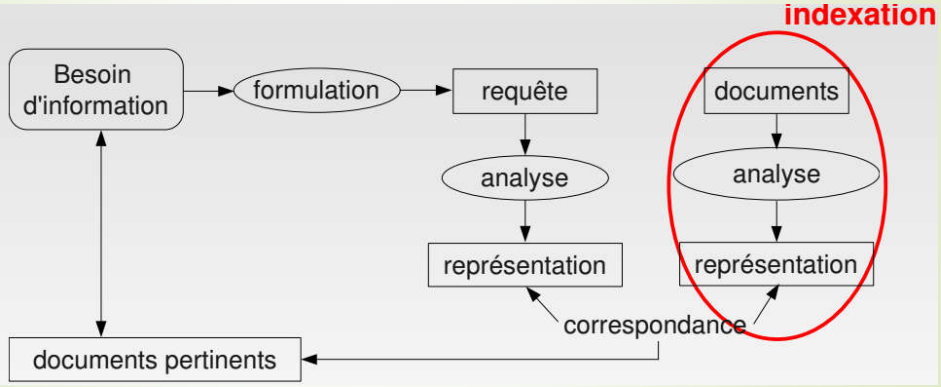
8

Approche classique de la RI



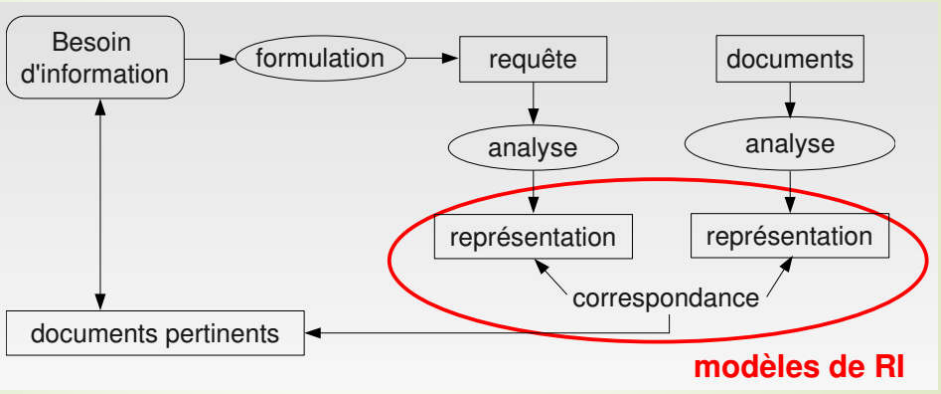
9

Approche classique de la RI



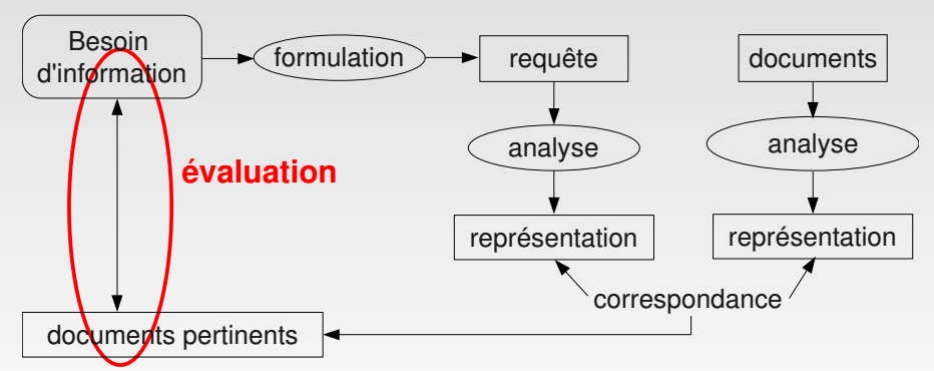
10

Approche classique de la RI



11

Approche classique de la RI



12

Problématiques de RI

- Représentation de l'information
  - –Comment construire une représentation à partir de documents ?
  - –Qu'est ce qu'une «bonne» représentation ?
  - –Quelle organisation physique pour les index ?
- Représentation des besoins
  - –Comment exprimer le besoin (langage de requêtes) ?
  - –Comment représenter le besoin ?
- Mise en correspondance des Représentation

13

Historique

- années 1940-50s
  - automatisation des bibliothèques
  - notion de **pertinence**
  - *Calvin N. Mooers* introduit en 1948 le terme « recherche d'information » (*information retrieval*)
  - La première conférence sur le thème a lieu en 1958 (*International Conference on Scientific Information*)
  - *Luhn* décrit un modèle statistique pour la recherche d'information (KWIC)

14

Historique

- années 1960-70s
  - *Maron* et *Kuhns* définissent un modèle de recherche d'information probabiliste
  - le projet d'évaluation CRANFIELD définit les mesures d'évaluation
  - premier livre de *Gerard Salton* sur le système SMART
  - développement des modèles booléens et vectoriels pour la recherche d'information

15

Historique

- années 1980s
  - grandes bases de données de documents
  - les premiers PCs intègrent la recherche d'information
  - introduction du TALN en recherche d'information
  - développement du domaine en France

16

Historique

- années 1990s
  - baisse du coût des disques => stockage d'information
  - WESTLAW premier système de recherche d'information à grande échelle qui utilise un modèle de recherche probabiliste
  - recherche sur des fichiers sur Internet
  - évaluations TREC
  - systèmes de recommandations
  - catégorisation et classification de textes
  - essor des modèles probabilistes (Okapi)
  - introduction des modèles de langues à la fin des années 90s

Historique

- Années 2000s
  - analyse de liens pour la recherche d'informations sur le web (google)
  - extraction d'information
  - réponses à des questions (TREC QA track)
  - indexation et recherche d'informations multimédia (image, vidéo, audio et musique)
  - recherche d'information multilingue (CLEF, NTCIR, DARPA, Tides)
  - résumé automatique de documents