Interactions Homme-Machine Évoluées

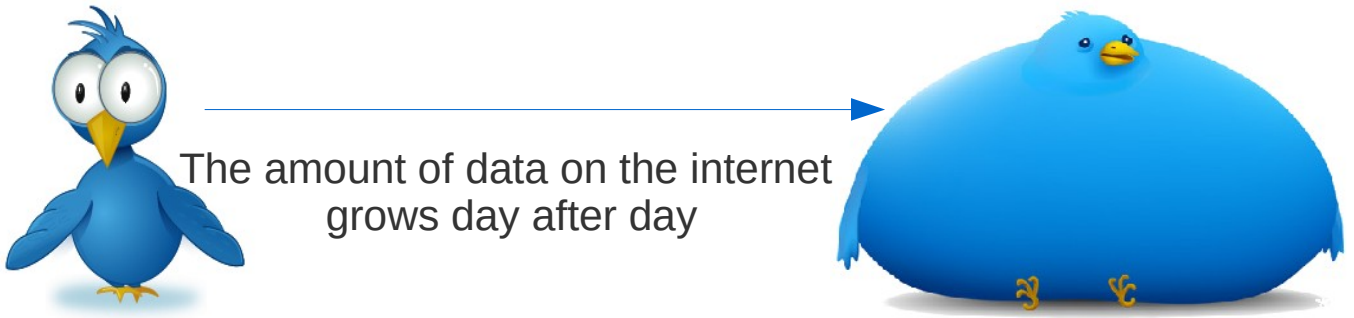
Recherche d'Informations personnalisée sur Internet

Alexandre Pauchet
INSA Rouen – Département ASI
BO.B.RC.18, pauchet@insa-rouen.fr

Définitions

- **RI** : système qui permet de retrouver des documents pertinents à une requête utilisateur, à partir d'une base de documents volumineuse
- **Document** : texte, morceau de texte, page web, image, vidéo, ... En RI, on appelle document toute unité qui peut constituer une réponse à une requête d'utilisateur.
- **Requête** : une requête est une formalisation du besoin d'information d'un utilisateur. Elle consiste souvent en une conjonction de mots-clefs.
- **Pertinence** : mesure permettant d'évaluer l'adéquation d'un document par rapport à un besoin d'information (requête).

Internet comme source d'information



- Contenu en constante **augmentation**
- Contenu **hétérogène** (textes, vidéos, images, …)
- Contenu **dynamique** (nouvelles pages, pages dynamiques : blogs/nouvelles/forums, …)
- Contenu **structuré peu utilisé** (métadonnées, hypermédias, sémantique de balisage, ...)

Internet = Source Ouverte

- **Publication simple et gratuite** : contenus non contrôlés (pages personnelles, blogs, wiki, forums)
  - Informations fausses
  - Informations vérifiées puis modifiées
  - Informations validées par des parties prenantes
  - La popularité n'est pas un gage de vérité

=> La pertinence d'une information doit être évaluée en fonction d'un besoin

Recherche d'informations (RI) sur Internet

- Satisfaction utilisateur difficile à mesurer

  => utilisation de mesures de pertinence
- La qualité des pages étant très différente, la pertinence doit en dépendre
- **Interface utilisateur de recherche importante**
  - Rapidité
  - Taille de l'index
  - Robustesse aux erreurs (approximations, mauvaises formulations, ambiguïtés)
  - "Services" offerts (cf. Google...)

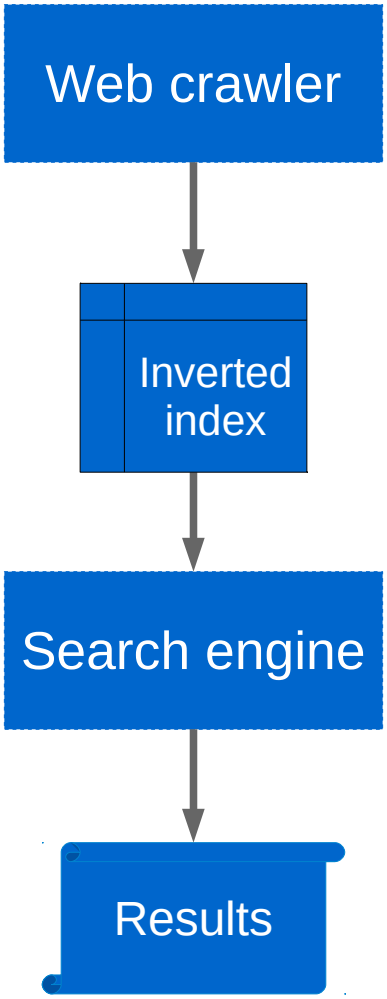
Évolution de la RI sur Internet

- **Première génération** (~ 1995 : Altavista, Excite, Lycos, etc.) :
  - Indexation à partir des informations de la page
- **Seconde génération** (~ 1998 : Google puis autres) :
  - Indexation à partir de la structure du Web (liens entrants, textes des liens, etc.)
- **Troisième génération** (en développement) :
  - Répondre au besoin utilisateur
  - Analyse sémantique et contextuelle
  - Aide à l'utilisateur : IHM, multilinguisme, correction et complétion orthographique, suggestion de requêtes, ...

IHM et RI

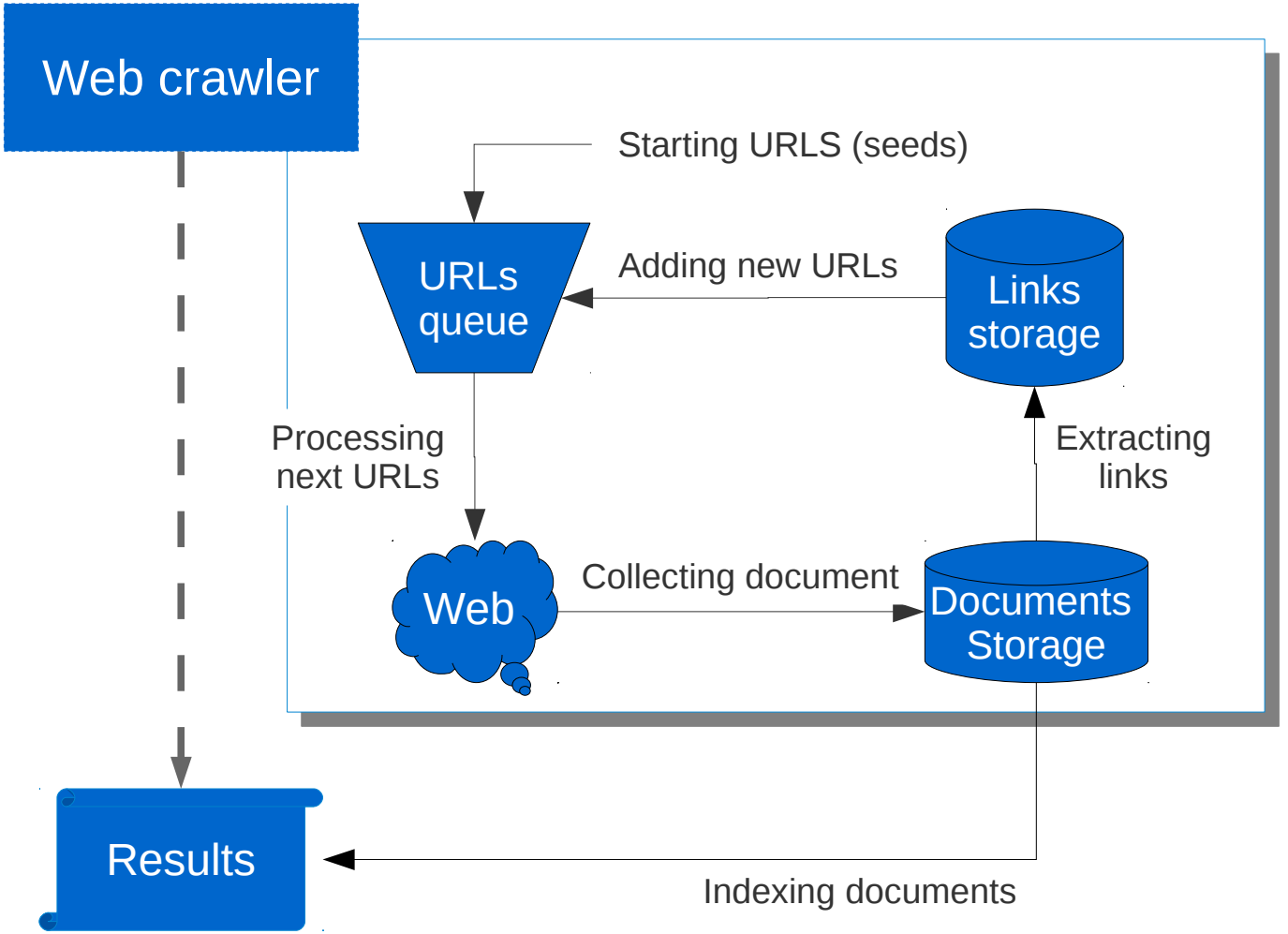
- Représentation des documents (index)
- Représentation des besoins utilisateurs
  - Besoins long terme (profil, domaines d'intérêt)
  - Besoins court terme (recherche contextuelle)
- Interaction utilisateur ↔ système de RI "pauvres"
  - Peu (pas) d'accès au contexte
  - Requête par mots-clefs
  - Index des documents non personnalisé
  - Retours de pertinence difficiles à capter

RI : approche classique



- Des robots (crawlers) parcourent les sites et indexent leur contenu
  - Fréquence dépend des moteurs
  - Une page peut demander à ne pas pas être indexée (robots.txt), ou de ne pas suivre les liens ("nofollow")
- L'index est un résumé à un instant T des pages du Web
- Les résultats d'un moteur de recherche sont classés par mesure de pertinence par rapport à une requête utilisateur

Web crawler


Web crawler
Starting URLS (seeds)
URLs queue
Adding new URLs
Links storage
Processing next URLs
Extracting links
Web
Collecting document
Documents Storage
Results
Indexing documents

Inverted index

- **Definition**: index data structure used to store a set of documents or elements, mapping them from their content such as words or numbers.

- **Example**:

- "crêpes" | {farine, œuf, lait}
- "génoise" | {œuf, sucre, farine}
- "caramel" | {sucre, beurre}
- "flan"| {œuf, lait, sucre}

- "farine" | {crêpes,génoise}
- "œuf" | {crêpes, génoise, flan}
- "lait" | {crêpes, flan}
- "sucre" | {génoise, caramel, flan}
- "beurre" | {caramel}

État de la RI sur le Web

Représentation du document :
Termes ou groupes de termes
[Salton & Yang, 1973]

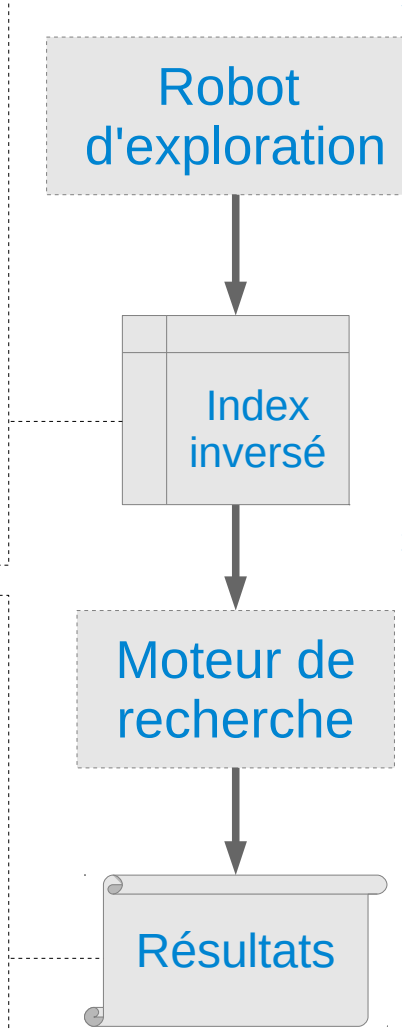
TF.IDF
[Salton & McGill, 1983]

Automatique ou semi-automatique
[Desmontils & Jacquin, 2002]

Google, Bing & Yahoo !

Formulation, représentation, correspondance
[Maisonnasse, 2008]

Modèle booléen, vectoriel, probabiliste, ...
[Salton, 1969], [Salton, 1971],
[Nottelmann & Fuhr, 2003]



Un robot d'exploration visite et collecte les pages du Web qui sont ensuite indexées

Un moteur de recherche fournit les pages Web indexées répondant à une requête donnée

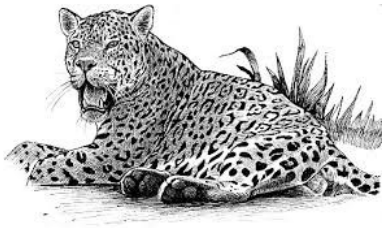
Modèle booléen

- Premier modèle de RI, basé sur la théorie des ensembles.
- Un document est représenté par un ensemble de termes
  - Exemple : d1(t1,t2,t5); d2(t1,t3,t5,t6); d3(t1,t2,t3,t4,t5)
- Requête = ensemble de mots et d'opérateurs booléens (∧, ∨, ¬)
  - Exemple : q = t1 ∧ (t2 ∨ ¬t3)
- Appariement document/requête **exact** basé sur la présence ou l'absence des termes de la requête dans les documents.
  - Exemple : Appariement(q,d1)=1 ; Appariement(q,d2)=0
- Inconvénients
  - La sélection d'un document est une décision binaire
  - Pas d'ordre pour les documents sélectionnés
  - Formulation de la requête difficile pour beaucoup d'utilisateurs
  - Problème de volumétrie : nombreux documents retournés

Modèle vectoriel

- Proposé par Salton dans le système SMART (Salton, G. 1970).
- Principe : documents et requêtes sont représentés sous la forme de vecteurs dans l'espace vectoriel des termes de la collection de documents.
  - Document *j* : $d_j = (w_{1j}, w_{2j}, ..., w_{Mj})$
  - Requête : $q = (w_{1q}, w_{2q}, ..., w_{Mq})$
  - Avec w_{ij}: poids du terme t_i dans le document d_j (ex : tf*idf)
- Une collection de n documents et M termes distincts peut être représentée sous forme d'une matrice *n*M* et la requête sous la forme d'un vecteur
- La pertinence est une mesure de similarité entre vecteurs (ex : cosinus)
- Avantages:
  - La pondération améliore les résultats de recherche
  - La mesure de similarité permet d'ordonner les documents selon leur pertinence
- Inconvénients :
  - Ne tient pas compte de l'ordre des mots (sac de mots)
  - La représentation vectorielle suppose l'indépendance entre termes

Limites actuelles de la RI sur le Web



Jaguar

Robot d'exploration

Index inversé

Corpus de document dont la taille s'approche de celle du Web

Moteur de recherche

Ambiguïté des requêtes
Contexte de recherche

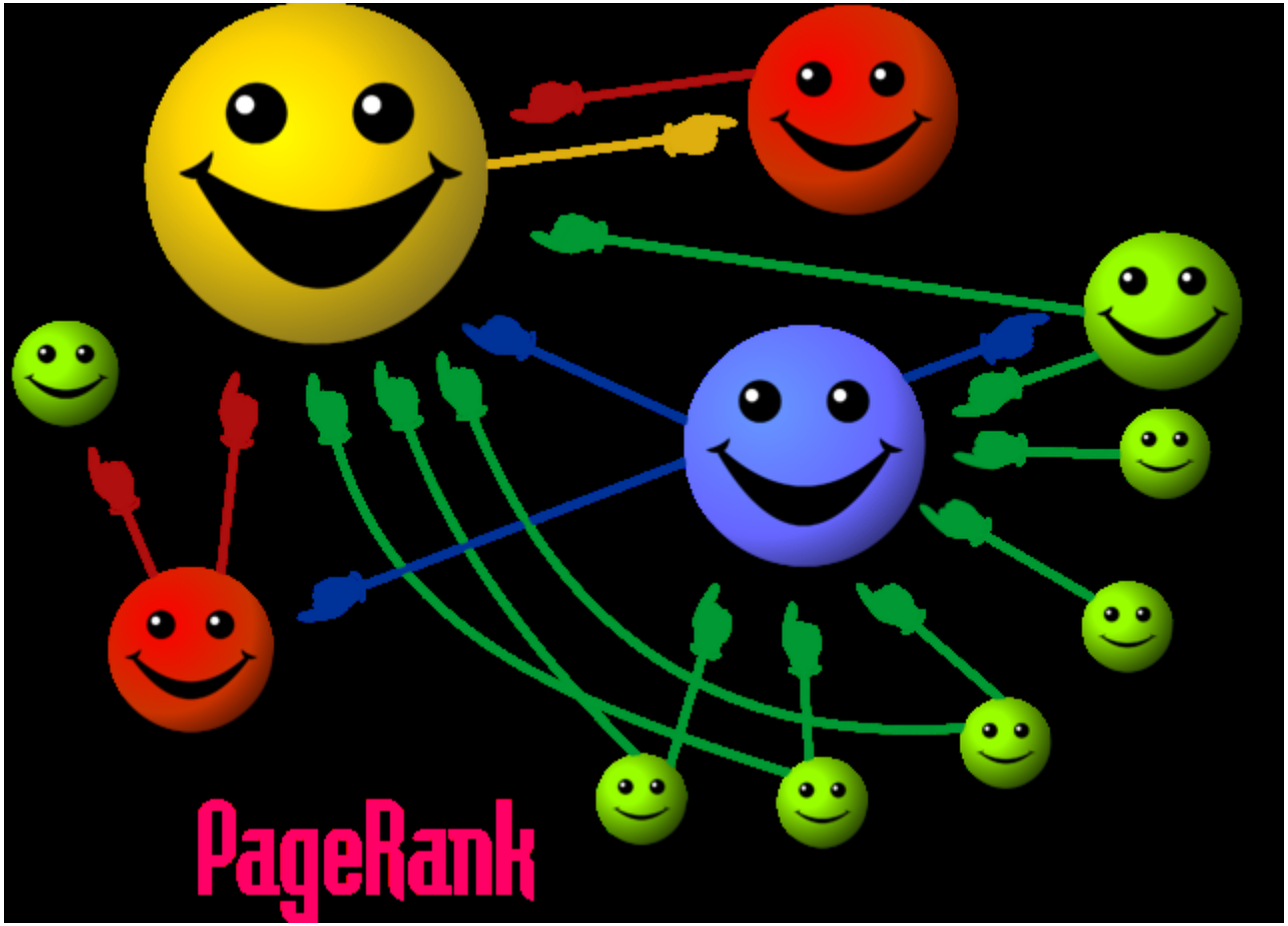
Résultats

Chat

Indexation

- Analyse (parsing) des pages/documents
  - Dépend du type de document
  - Conversion en ***Hits*** = pondération selon :
    - Nombres d'occurrences de chaque mot
    - Position dans le document
    - Taille relative de la fonte et casse du mot
- Seul le Web de surface est indexé
  - Web "profond" : sites non liés, contenu à accès limité, contenu refusant l'indexation, pages dynamiques
  - Web profond = 500 fois le Web de surface selon (BrightPlanet 2001)

Google : PageRank (Page et al., 1998)



http://en.wikipedia.org/wiki/File:PageRank-hi-res.png

Une page est importante si elle est pointée par d'autres pages importantes

PageRank ~ probabilité d'arriver au hasard sur une page

Formulation du PageRank

- 3 facteurs déterminent le PageRank d'une page P :
  - Le nombre de liens pointant vers P
  - Le nombre de liens contenus par les pages qui ont un lien vers P
  - Le PageRank des pages qui ont un lien vers P
- Le PageRank est la somme des PageRanks des pages ayant un lien vers Pi, pondérée par le nombre total de liens sortants :

$$r(P_i) = \sum_{P_j \in \mathscr{B}_{P_i}} \frac{r(P_j)}{|P_j|}$$

- Le calcul est **itératif**

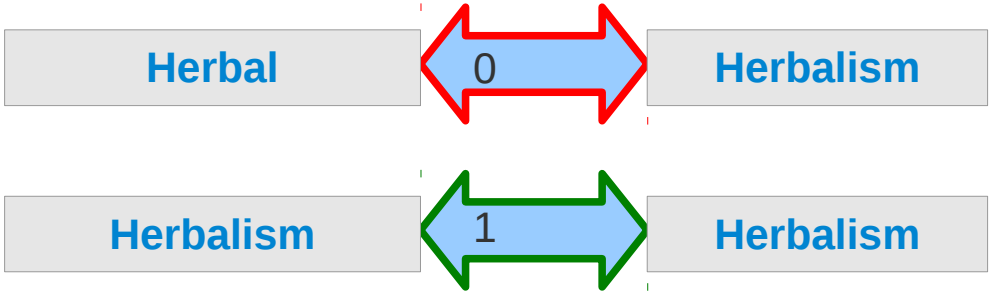
La RI chez Google

- Indexation des documents
  - Index inversé (mot(s) → document(s))
  - Importance des mots pondérée selon leur type (titre, ancre, URL, fonte, …) et leur localisation dans la page
  - Équivalent à un TF*IDF
- Mesure de similarité requête/documents (produit scalaire des 2 vecteurs)
- Combinaison du score de RI avec le PageRank pour ordonner l'affichage des résultats

Exemple de mesures de pertinence

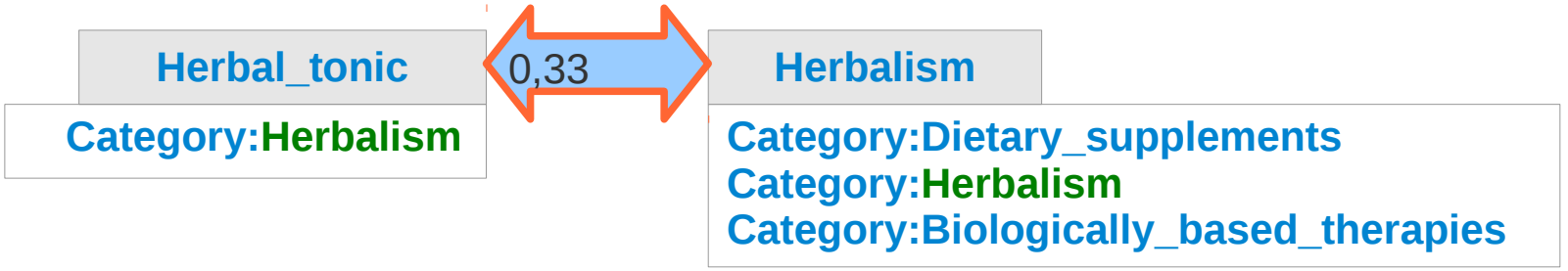
Mesure terminologique (ici cosinus)

0 ou 1



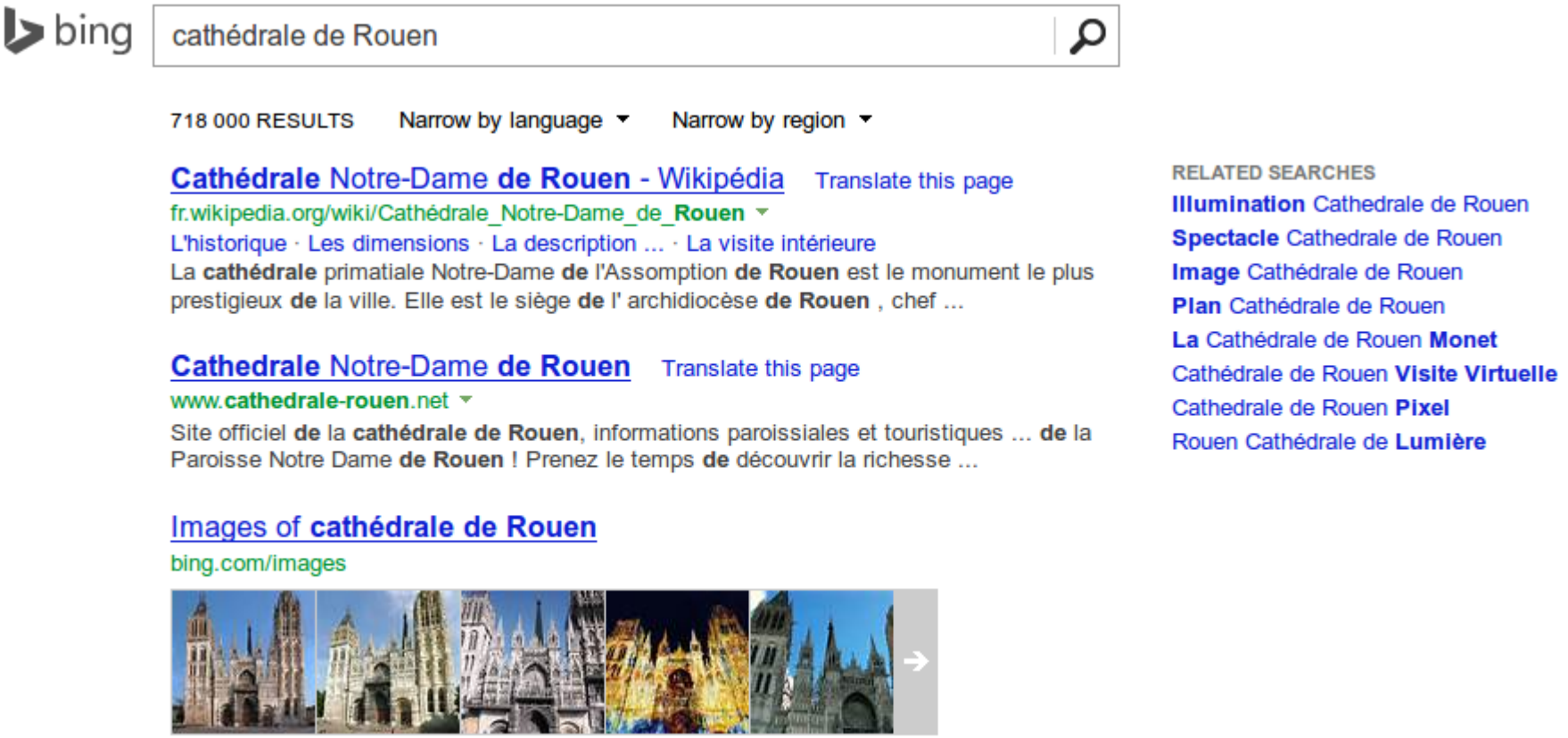
Mesure conceptuelle

Entre 0 et 1



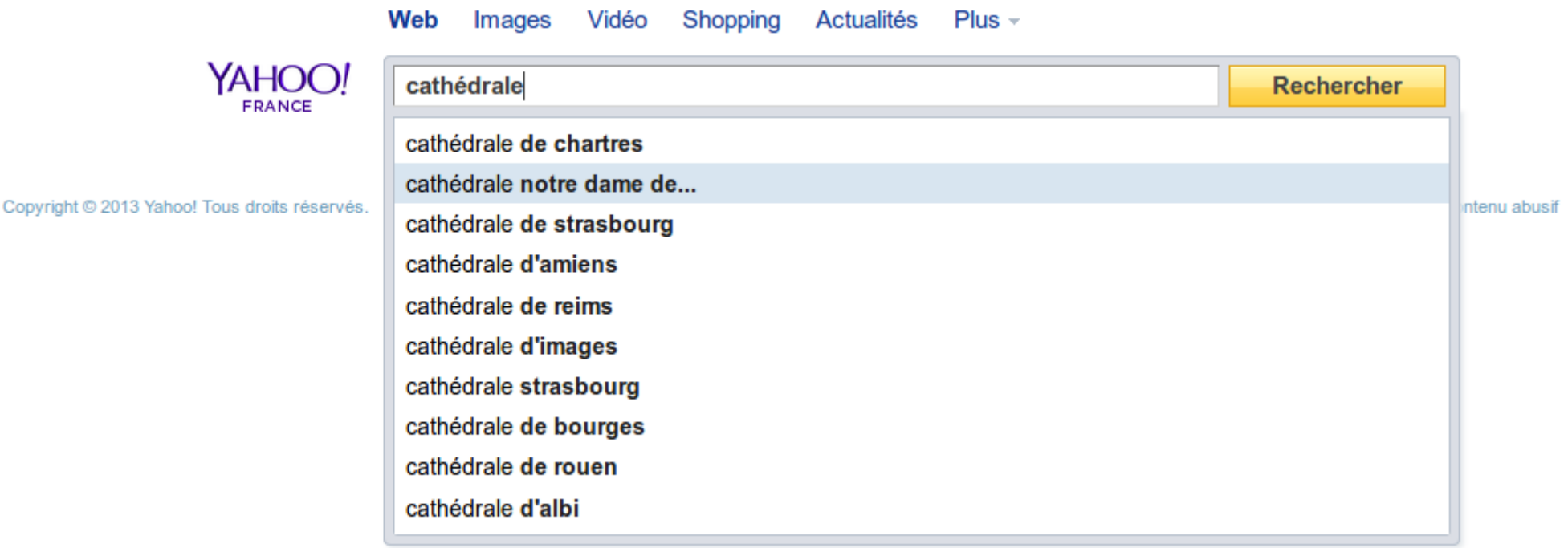
IHM basique

- IHM minimaliste : 1 champs texte, 1 page de résultats (snippet) avec des extraits
- Recherche avancée souvent proposée



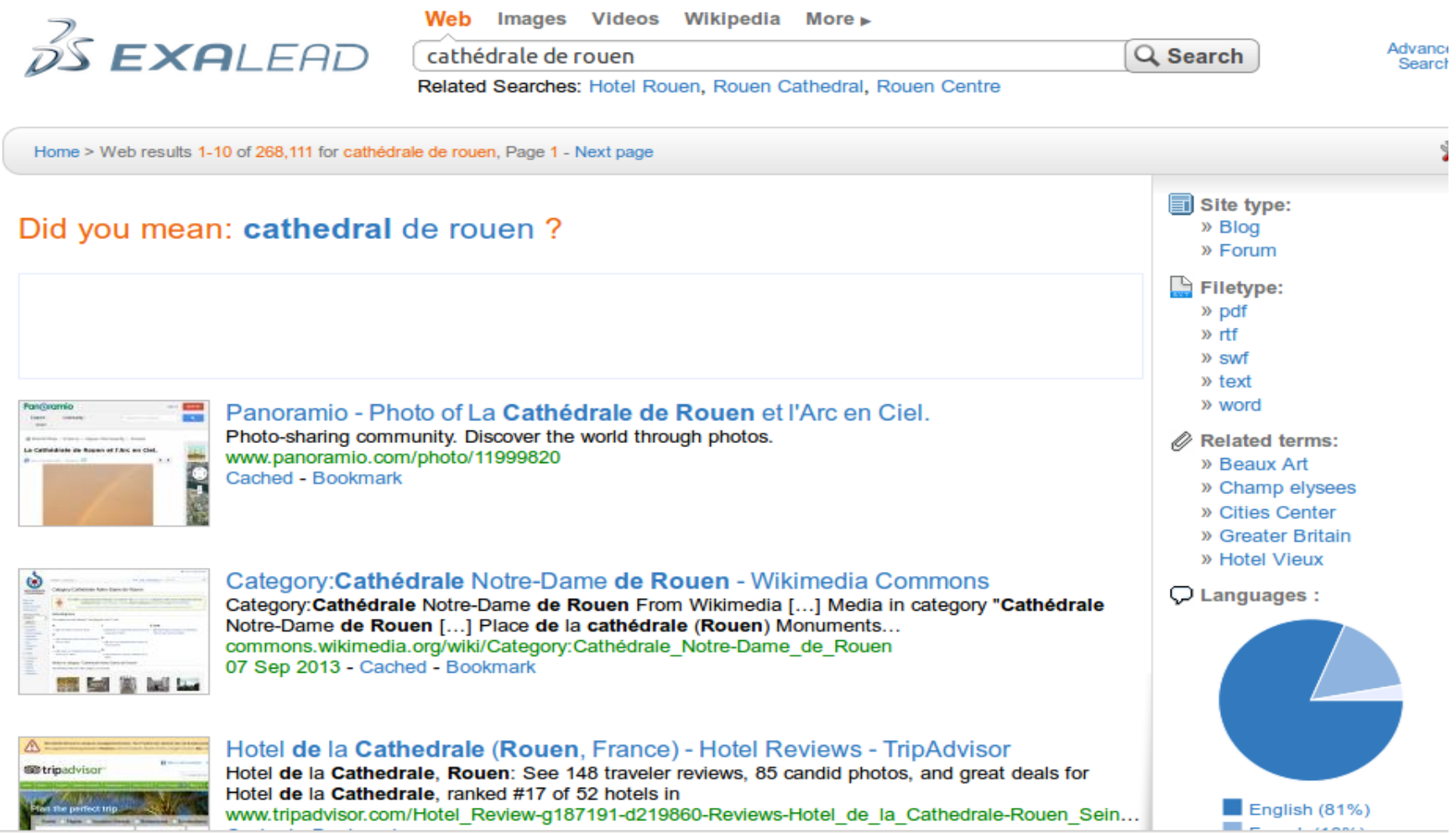
IHM : aide à la saisie

- Correction orthographique
- Complétion automatique selon les recherches les plus courantes



IHM : aide à l'analyse des résultats

- Aperçu des résultats
- Liens sponsorisés
- Termes et annuaires
- Choix de la langue

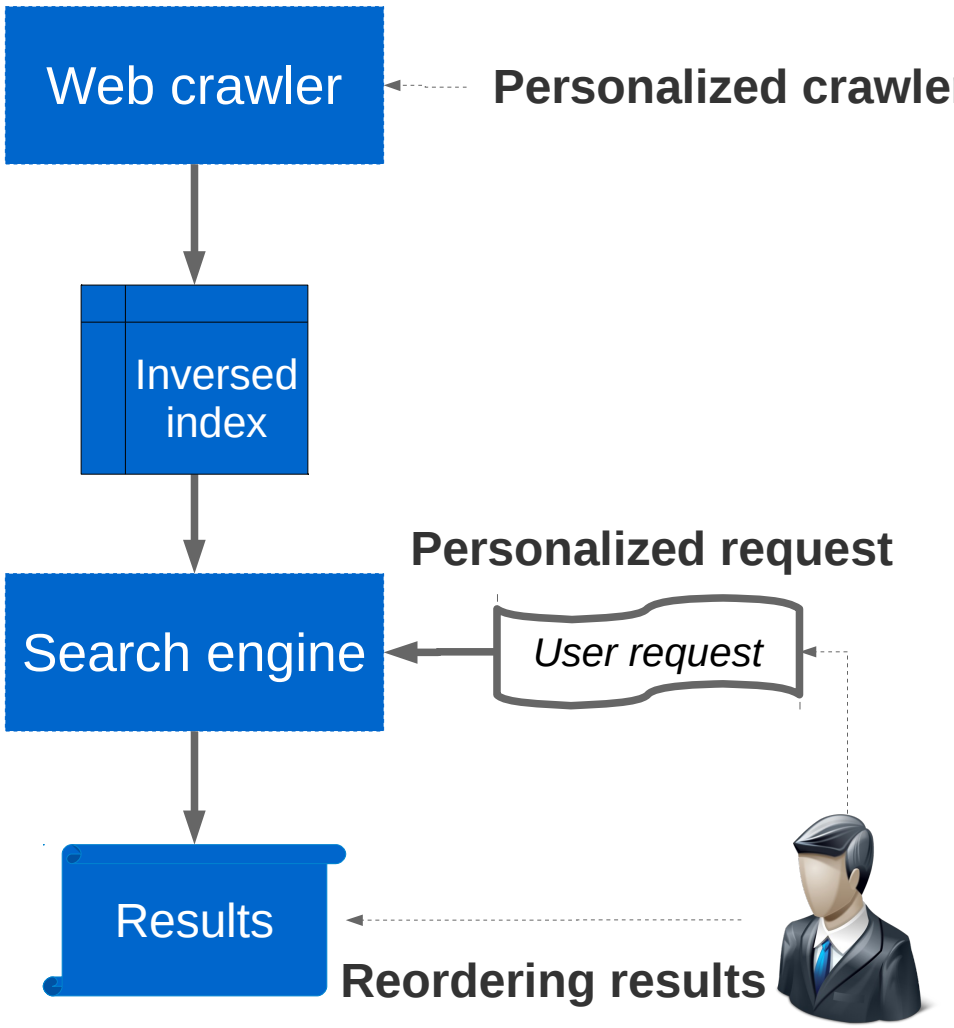


Inconvénients des approches classiques

- Problèmes liés au PageRank
  - Google : 82 % des requêtes en France
  - Une page est bien référencée si elle est populaire ; une page est populaire si elle est bien référencée
  - Émergence difficile de nouvelles pages
  - Le référencement est une activité lucrative
- La qualité du contenu n'est pas prise en compte
- Ambiguïtés ? Synonymies ?
- Contextes de recherche ?
- Généralisation (concepts) ?

⇒ Personnalisation !

Personnalisation : approches existantes



Statistical model

C. Aggarwal, F. Al-Garawi, and P. Yu. Intelligent crawling on the world wide web with arbitrary predicates. 2001.

Reinforcement learning

S. Chakrabarti, K. Punera, and M. Subramanyam. Accelerated focused crawling through online relevance feedback. 2002.

Multi-agent : genetic model

F. Menczer and R. Belew. Adaptive retrieval agents : Internalizing local context and scaling up to the web. 2000.

Multi-agent : biological model

F. Gasparetti and A. Micarelli. Swarm intelligence : Agents for adaptive web search. In ECAI, 2004.

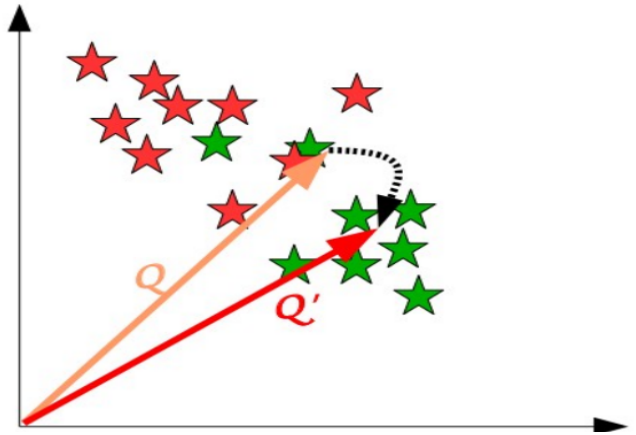
Interaction Homme – Système de RI

- Exemple : R="animal", D1="cheval"
  - D1 ne sera pas retourné s'il ne contient pas "animal"
  - Comment retourner les documents pertinents même s'ils ne contiennent aucun des termes de la requête ?
- Méthodes permettant d'améliorer le rappel :
  - Extension de requête (query expansion)
  - Retour de pertinence (relevance feedback)
  - Désambiguïsation de la requête : concepts, détection de synonymie, "dialogue" H-M pour validation, …

Query Expansion

- Modification de la requête en utilisant des ressources extérieures
  - Utilisation de dictionnaires, thésaurus, ontologies, …
  - Augmenter le rappel
  - La précision va baisser mécaniquement
- Exemples :
  - Hôpital → médical (thématique)
  - Taux d'intérêt → intéressant (grammatical)
  - Brillant → luisant (synonymie)
  - Cheval → animal (généralisation)
  - Animal → cheval, chien, ... (spécialisation)

Relevance Feedback



documents non pertinents

documents pertinents

Objectif : déplacer (implicitement) le vecteur de la requête pour le rapprocher des documents pertinents
→ Formule de Rocchio

$$\vec{Q}' = \alpha \vec{Q} + \beta \vec{P} + \gamma \vec{N} \vec{P}$$

moyenne des vecteurs des documents non pertinents

valeur négative (ex : *-0,25*)

moyenne des vecteurs des documents pertinents

valeur positive (ex : *0.5*)

vecteur requête initial

valeur positive supérieure aux autres (ex : 1)

nouveau vecteur requête

Tiré de http://www2.ifi.auf.org/personnel/Alain.Boucher/cours/recherche_information/RI-03.pdf

Fortement inspiré de

Cours

- Cours de Xavier Tannier (http://perso.limsi.fr/amax/enseignement/iri/M2PRO_IRI_6_RIWeb.pdf)
- http://math.univ-lyon1.fr/homes-www/malbos/lib/exe/fetch.php?media=ens:algapp12:algapiichapiiisec9.pdf
- http://www.iro.umontreal.ca/~nie/IFT6255/Introduction.html
- https://www.irit.fr/~Mohand.Boughanem/Enseignements_RI.php