

Interactions Homme-Machine Évoluées

Recherche d'Informations personnalisée sur Internet

Alexandre Pauchet

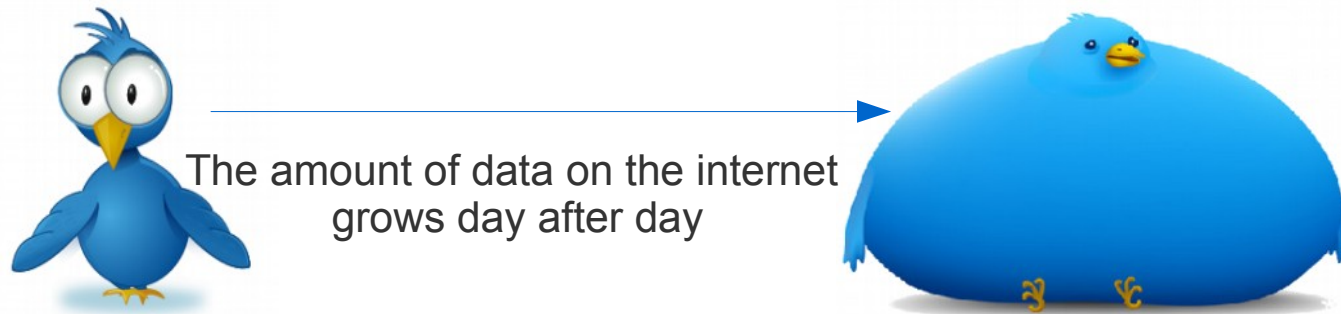
INSA Rouen – Département ASI

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

- Métadonnées peu utilisées
- Publication simple et gratuite => Contenus non contrôlés (pages personnelles, blogs, wiki, forums)
 - Informations fausses
 - Informations vérifiées puis modifiées
 - Informations validées par des parties prenantes
 - 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 importante**
 - Rapidité
 - Taille de l'index
 - Robustesse aux erreurs (approximations, mauvaises formulations, ambiguïtés)
 - "Services" offerts (cf. Google...)

Évolution de la RI sur Internet

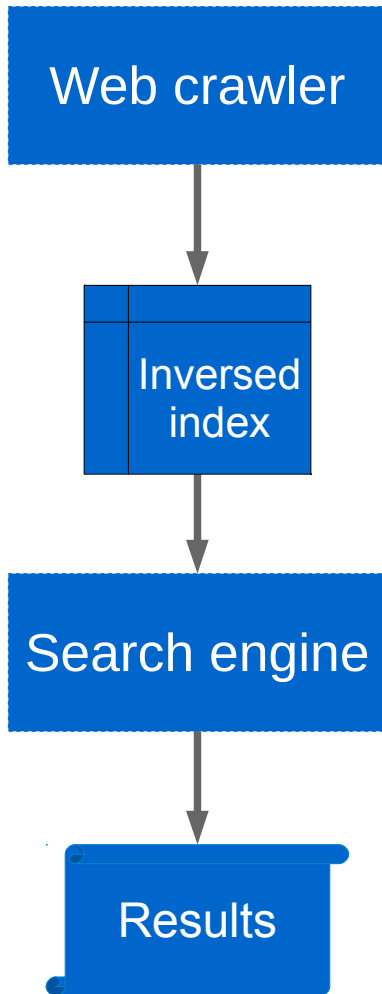
- Première génération (~ 1995 : Altavista, Excite, Lycos, etc.) :
 - Indexation à partir des informations de la page
- Seconde génération (~ 1998 : Google puis autres) :
 - Indexation à partir de la structure du Web (liens entrants, textes des liens, etc.)
- Troisième génération (en développement) :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

IHM et RI

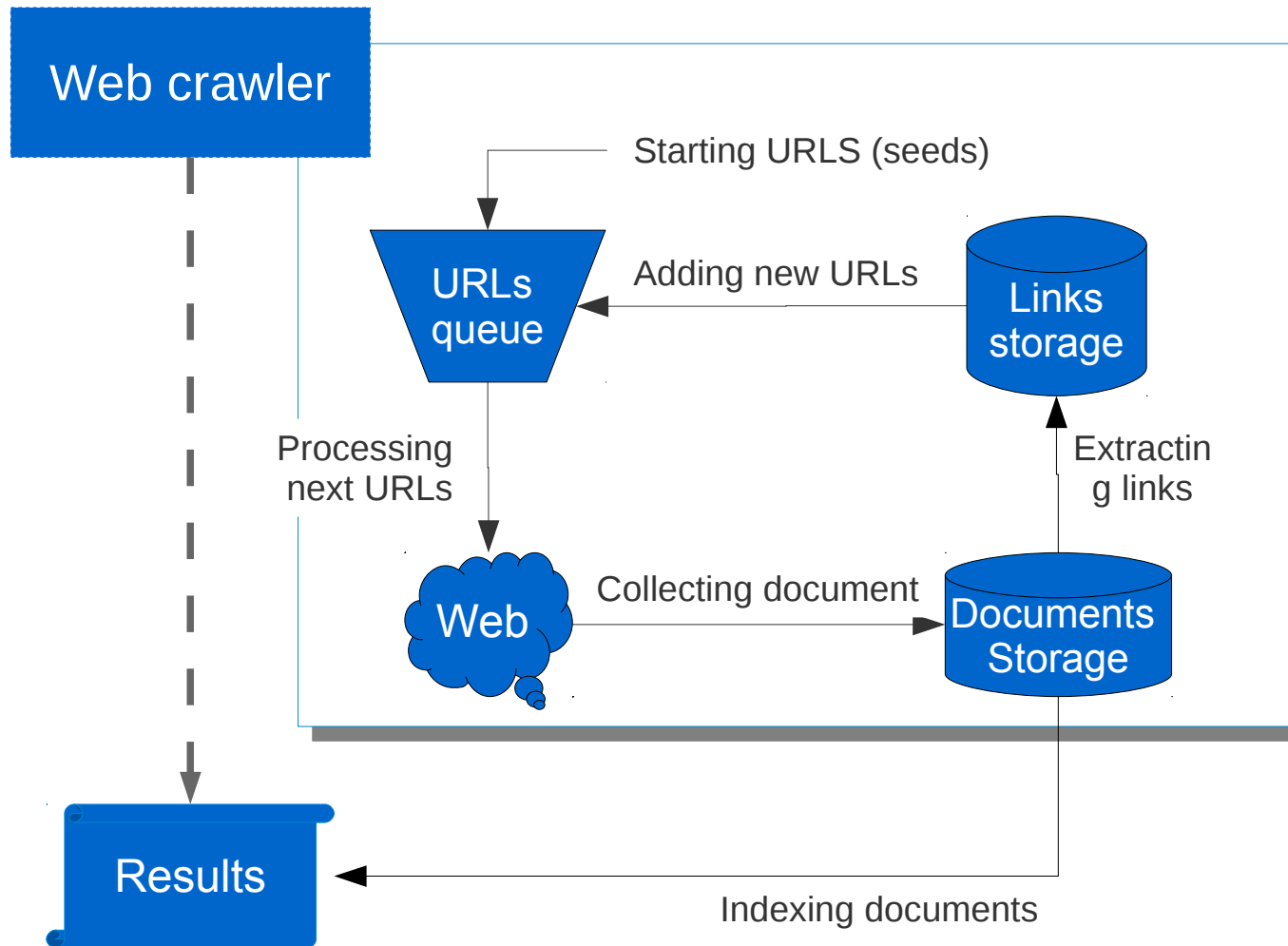
- Représentation des documents (index)
- Représentation des besoins utilisateurs
 - Besoins long terme (profil, domaines d'intérêt)
 - Besoins court terme (recherche contextuelle)
- Interaction utilisateur ↔ système de RI "pauvres"
 - Peu (pas) d'accès au contexte
 - Requête par mots-clefs
 - Index des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



Crawler



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

[Maisonasse, 2008]

Modèle booléen, vectoriel,
probabiliste

[Salton, 1969], [Salton, 1971],
[Nottelmann & Fuhr, 2003]

Robot
d'exploration

Index
inversé

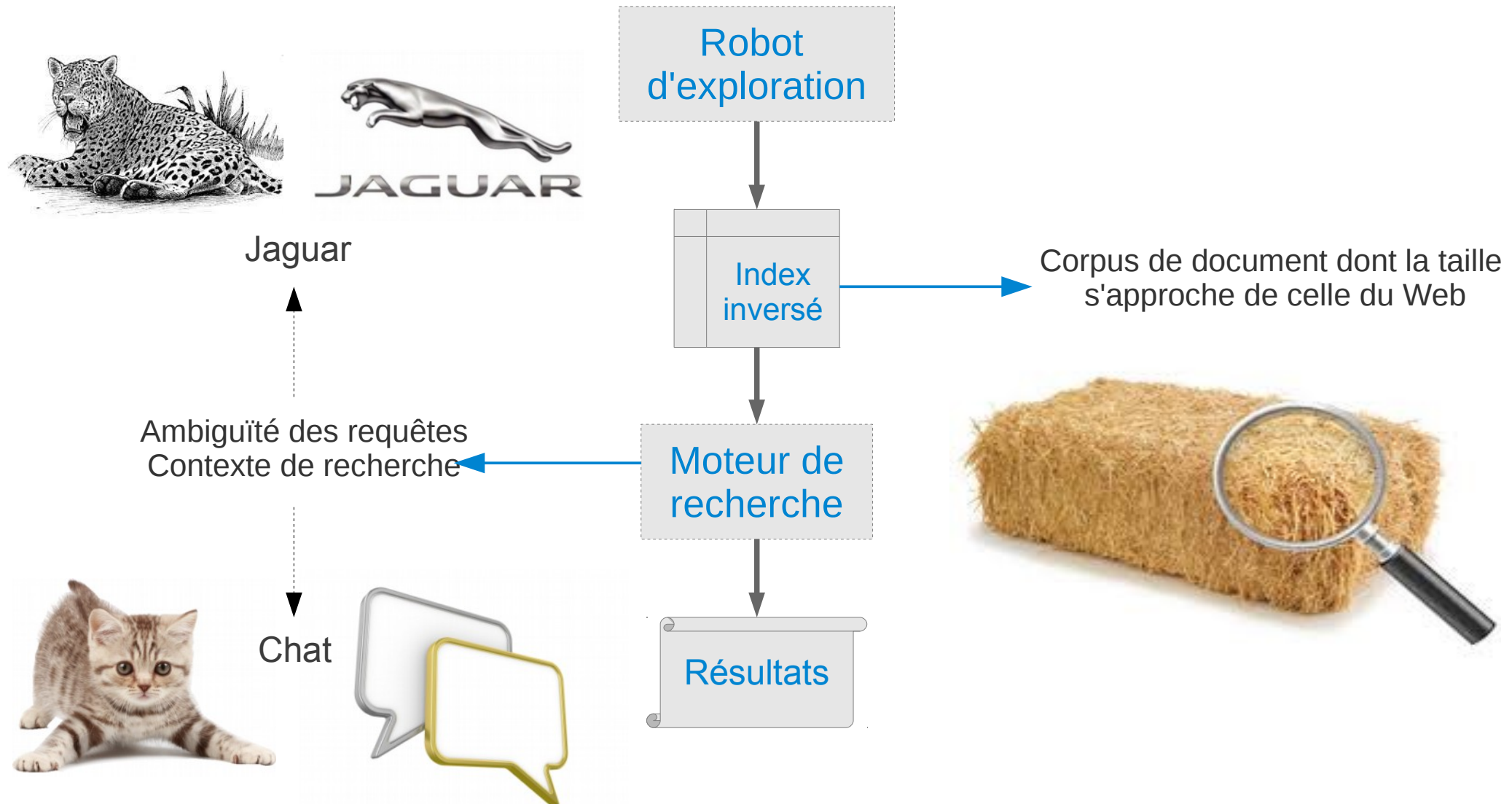
Moteur de
recherche

Résultats

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

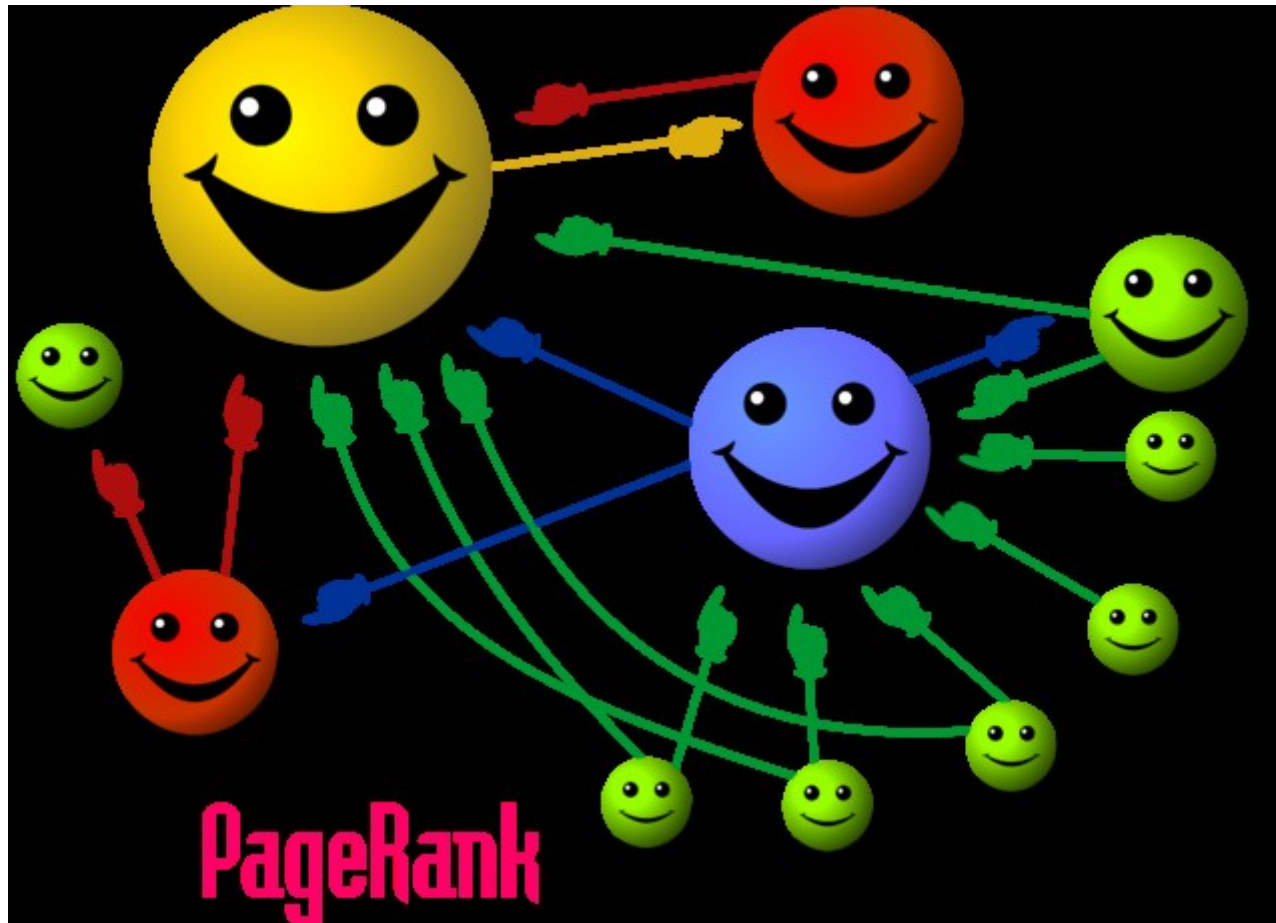
Limites de la RI sur le Web



Indexation

- Analyse (parsing) des pages/documents
 - Dépend du type de document
 - Conversion en *Hits*
 - 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 P_i , pondérée par le nombre total de liens sortants :

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

- Le calcul est itératif

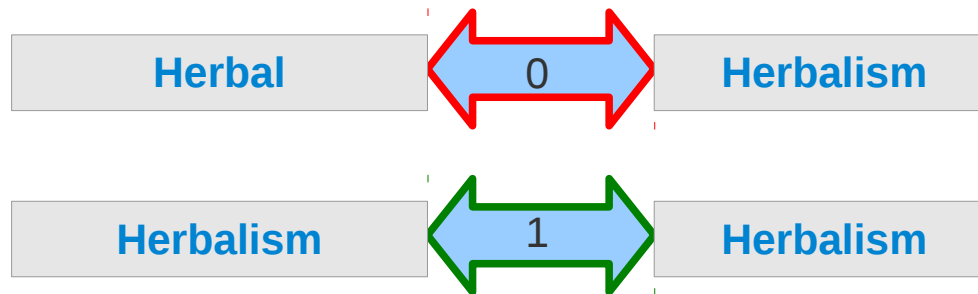
La RI chez Google

- Indexation des documents
 - Index inversé (mot(s) \rightarrow document(s))
 - Importance des mots pondérée en fonction des types (titre, ancre, URL, fonte, ...) et de la 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

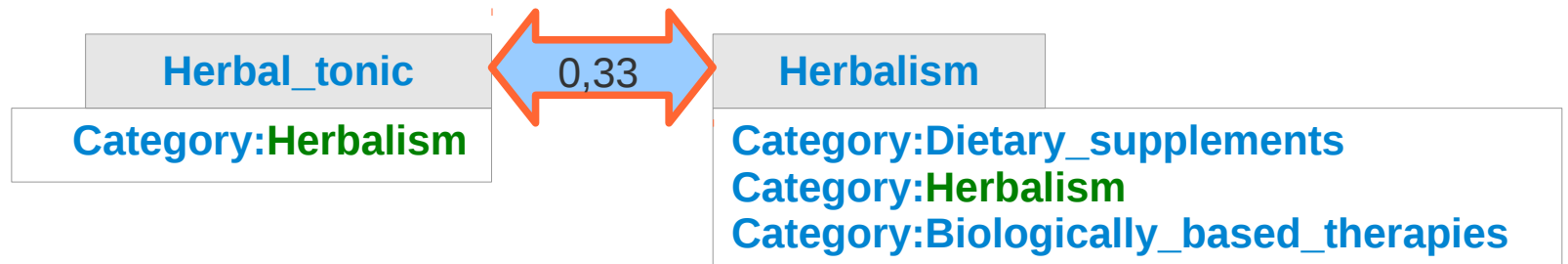
Mesure terminologique (ici cosinus)

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



718 000 RESULTS Narrow by language ▼ Narrow by region ▼

[Cathédrale Notre-Dame de Rouen - Wikipédia](#) Translate this page

fr.wikipedia.org/wiki/Cathédrale_Notre-Dame_de_Rouen ▼

L'historique · Les dimensions · La description ... · La visite intérieure

La **cathédrale** primatiale Notre-Dame de l'Assomption de **Rouen** est le monument le plus prestigieux de la ville. Elle est le siège de l'archidiocèse de **Rouen**, chef ...

[Cathedrale Notre-Dame de Rouen](#) Translate this page

www.cathedrale-rouen.net ▼

Site officiel de la **cathédrale de Rouen**, informations paroissiales et touristiques ... de la Paroisse Notre Dame de **Rouen** ! Prenez le temps de découvrir la richesse ...

[Images of cathédrale de Rouen](#)

bing.com/images



RELATED SEARCHES

[Illumination Cathedrale de Rouen](#)

[Spectacle Cathedrale de Rouen](#)

[Image Cathédrale de Rouen](#)

[Plan Cathédrale de Rouen](#)

[La Cathédrale de Rouen Monet](#)

[Cathédrale de Rouen Visite Virtuelle](#)

[Cathedrale de Rouen Pixel](#)

[Rouen Cathédrale de Lumière](#)

IHM : aide à la saisie

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

YAHOO!
FRANCE

Web Images Vidéo Shopping Actualités Plus ▾

cathédrale|

Rechercher

cathédrale de chartres

cathédrale notre dame de...

cathédrale de strasbourg

cathédrale d'amiens

cathédrale de reims

cathédrale d'images

cathédrale strasbourg

cathédrale de bourges

cathédrale de rouen

cathédrale d'albi

IHM : aide à l'analyse des résultats

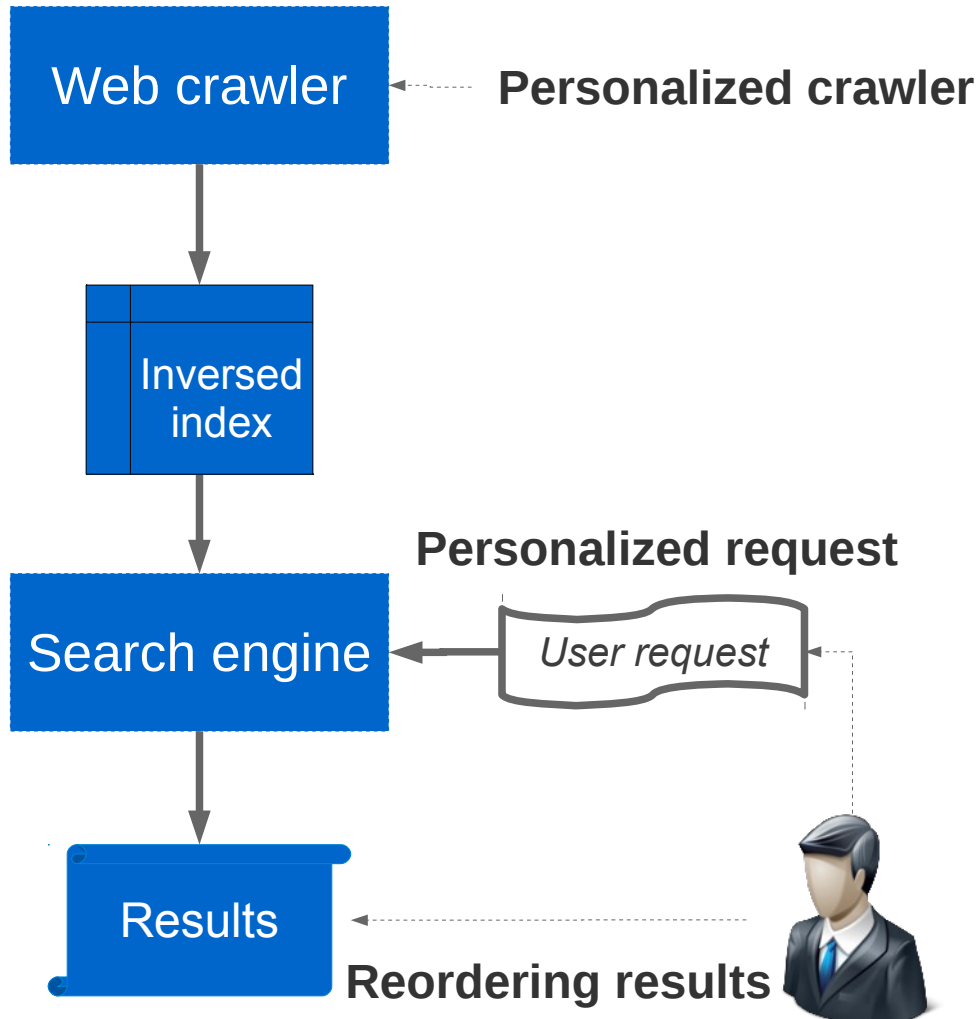
- Aperçu des résultats
- Liens sponsorisés
- Termes et annuaires
- Choix de la langue

The screenshot shows the Exalead search engine interface. At the top, the Exalead logo is on the left, and navigation links for 'Web', 'Images', 'Videos', 'Wikipedia', and 'More' are in the center. A search bar contains the text 'cathédrale de rouen' and a 'Search' button. To the right of the search bar, there is a link for 'Advanced Search'. Below the search bar, 'Related Searches' are listed: 'Hotel Rouen', 'Rouen Cathedral', and 'Rouen Centre'. A breadcrumb trail at the top of the results area reads 'Home > Web results 1-10 of 268,111 for cathédrale de rouen, Page 1 - Next page'. A suggestion box asks 'Did you mean: cathedral de rouen ?'. The main results area displays three items: 1) A Panoramio photo of 'La Cathédrale de Rouen et l'Arc en Ciel' with a link to 'www.panoramio.com/photo/11999820'. 2) A Wikimedia Commons category page for 'Cathédrale Notre-Dame de Rouen' with a link to 'commons.wikimedia.org/wiki/Category:Cathédrale_Notre-Dame_de_Rouen'. 3) A TripAdvisor review for 'Hotel de la Cathedrale (Rouen, France)' with a link to 'www.tripadvisor.com/Hotel_Review-g187191-d219860-Reviews-Hotel_de_la_Cathedrale-Rouen_Sein...'. On the right side, there is a sidebar with filters: 'Site type' (Blog, Forum), 'Filetype' (pdf, rtf, swf, text, word), 'Related terms' (Beaux Art, Champ elysees, Cities Center, Greater Britain, Hotel Vieux), and 'Languages' (English (81%), French (14.9%), ...). A pie chart is partially visible at the bottom of the sidebar.

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 ?
- 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.

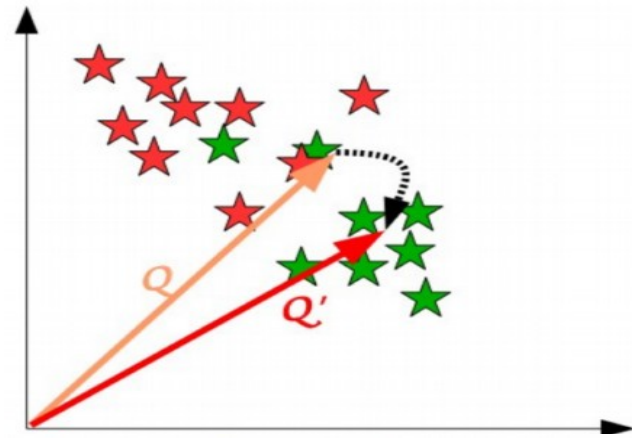
Multi-agent : biological model

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 :
 - Retour de pertinence (relevance feedback)
 - Extension de requête (query expansion)
 - Désambiguïsation de la requête : concepts, détection de synonymie, "dialogue" H-M pour validation, ...

Relevance Feedback



- ★ documents non pertinents
- ★ documents pertinents

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

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

→ 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

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)

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>