

Démonter le moteur ... de recherche

Jian-Yun Nie
DIRO

<http://www.iro.umontreal.ca/~nie/>
nie@iro.umontreal.ca

Moteur de recherche –

Search Engine journal <http://www.searchenginejournal.com/24-eye-popping-seo-statistics/42665/>

- 93% d'expériences en ligne commencent par un engin de recherche
- Plus 100 milliards de recherches globales / mois
- [MarketingCharts](#) montre que plus de 39% de consommateurs sont amenés par les engins de recherche
- Recherche (Search) est responsable de 25% d'achats d'appareils en ligne aux E.S. en 2010.
- 1.5 milliard de visiteurs en ligne en Q2 2012 (comScore)
- 70% de liens que les utilisateurs cliquent sont des résultats

Résultats organiques

Google

Search About 103,000,000 results (0.33 seconds)

Web
Images
Maps
Videos
News
More


Brossard, QC
Change location

The web
Pages from Canada
More search tools

Ad related to organic search results ⓘ

[Find The Best SEO Company](#)
www.freeinternetmarketingquote.com/
Compare SEO Companies, Get 5 Free **Search** Engine Optimization Quotes

[Organic search - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Organic_search
Organic search results are listings on search engine results pages that appear because of their relevance to the search terms, as opposed to their being ...

[Organic vs. Paid Search Results: Organic Wins 94% of Time ...](#)
searchenginewatch.com/.../Organic-vs.-Paid-Search-Results-...
 by Danny Goodwin - in 433 Google+ circles
23 Aug 2012 – **Search** engine users overwhelmingly click on **organic results** on Google and Bing by a margin of 94 percent to 6 percent. That's according to ...

[Definition: Organic Search](#)
webdesign.about.com/od/seo/g/bldeforganicsea.htm
Most **search** engines offer two types of **search results** to their customers: paid **results** (typically at the top or on the side) and **organic** or **natural results**. While paid ...

[New research: Organic search results and their impact on search ...](#)
adwords.blogspot.com/.../new-research-organic-search-results-and.ht...
27 Mar 2012 – The Google Research team has a new study out today that examines

Ads ⓘ

[SEO Experts from \\$10/hr.](#)
www.odesk.com/
Browse 1,000's of SEO Experts. Find the Best Global Talent Today!
608 people +1'd or follow oDesk

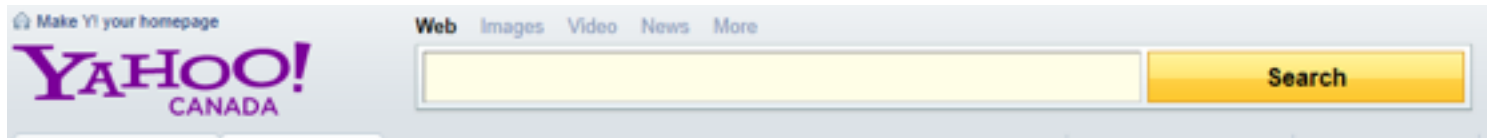
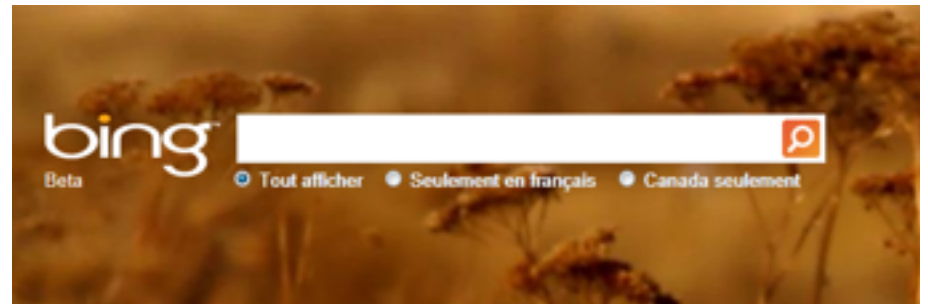
[Need a SEO Expert?](#)
www.freelancer.ca/Outsource
Post A Free Project Today!
Over 126,000 SEO Experts

[SEO Services - \\$190/month](#)
www.wmarketing.com/
1 (877) 539 6766
Higher Rankings in 1-4 weeks.
See Your Backlinks Improve Today!

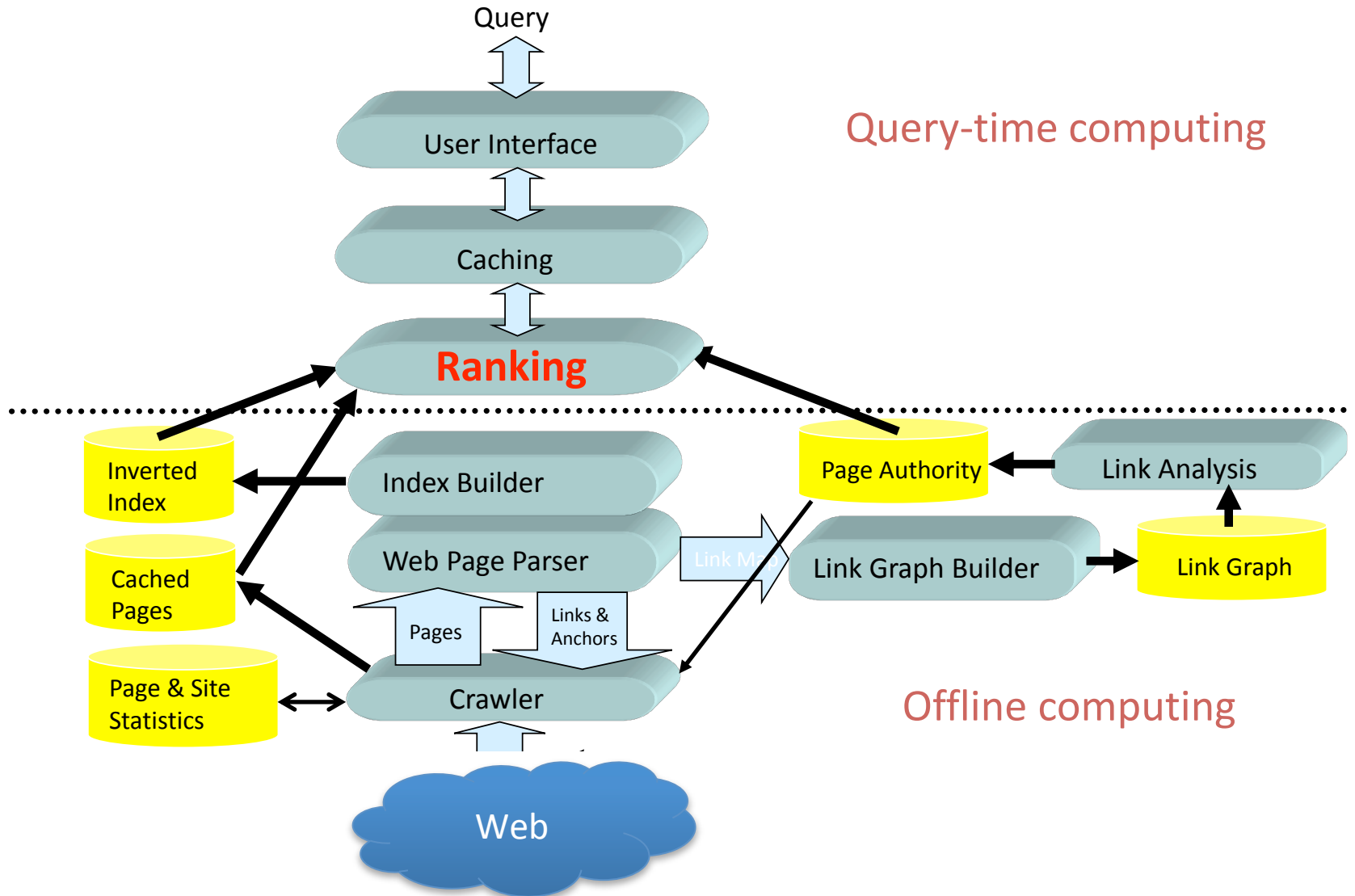
[Free SEO Score](#)
www.seoengine.com/
Automated Website Analysis
See Changes Before Anyone!

See your ad here >

Les faces visibles des moteurs de

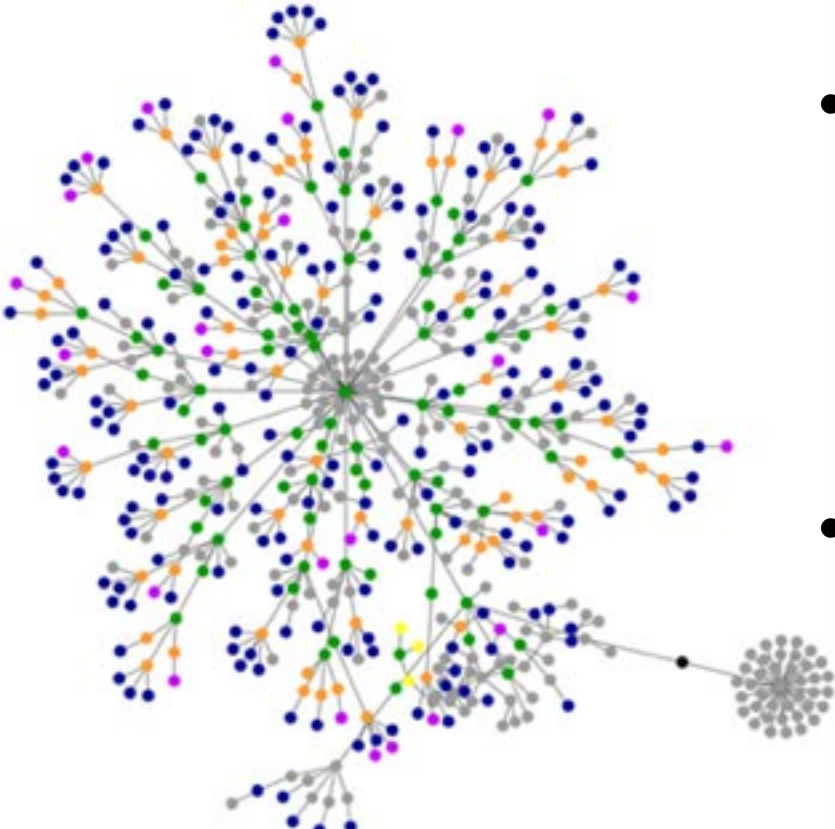


La face cachée



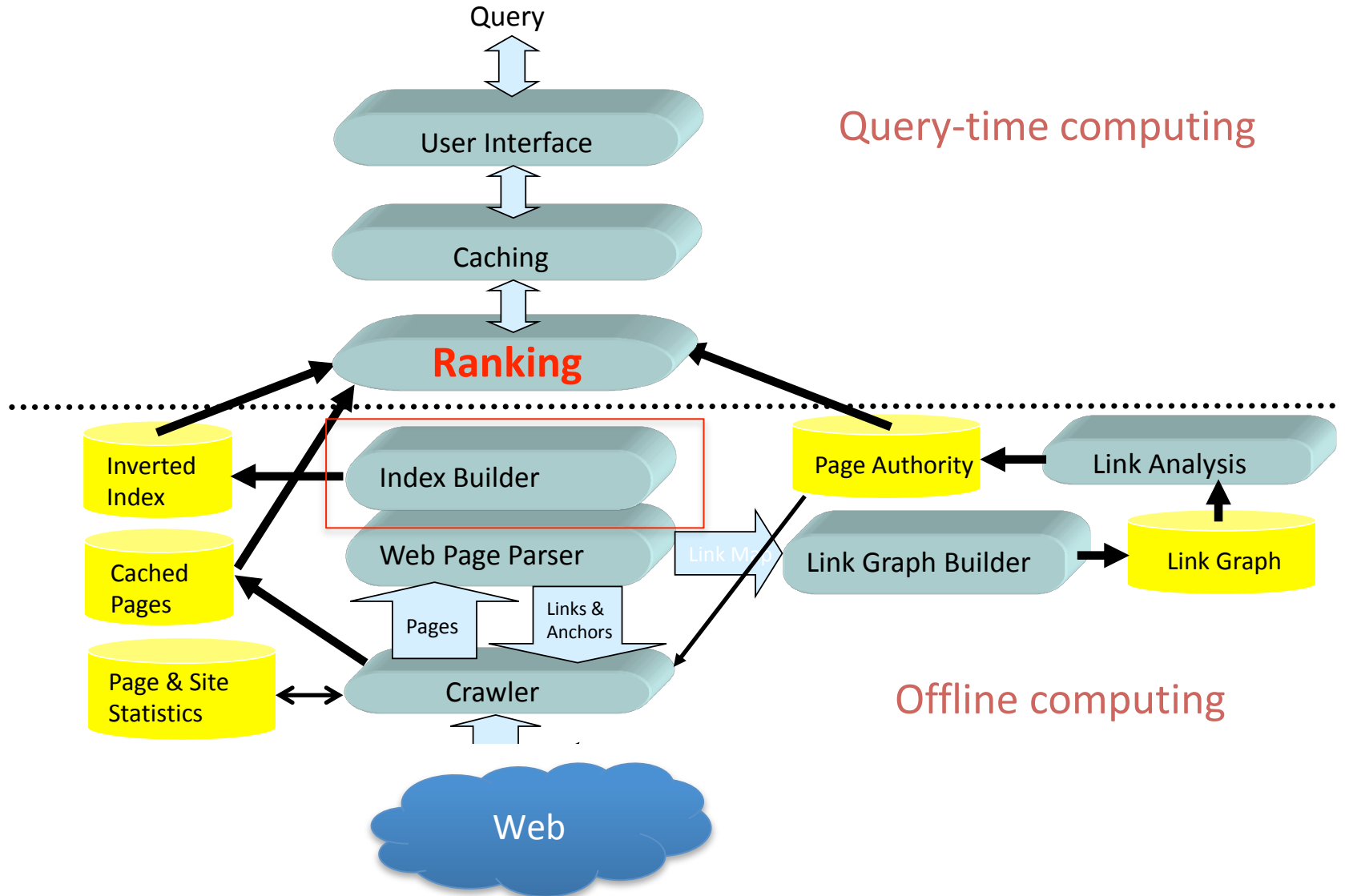
Collecter des documents - principe

- Documents “seeds”
- Explorer des hyperliens pour découvrir d'autres sites/pages



- Stratégies
 - Largeur-d'abord (couverture équilibrée)
 - Prioriser des sites importants
 - ...
- Taille: (Google)
 - 1 billion de pages (2008)
 - Indexe ~1 peta-octet (2012)

Ouvrons le capot



Indexer les documents

La recherche en recherche d'information est très active.



Tokenisation

La, recherche, en, recherche, d, information, est, très, active



Filtrage de mots outils

recherche, recherche, information, active



Standardisation des mots

recherch, recherch, informat, actif



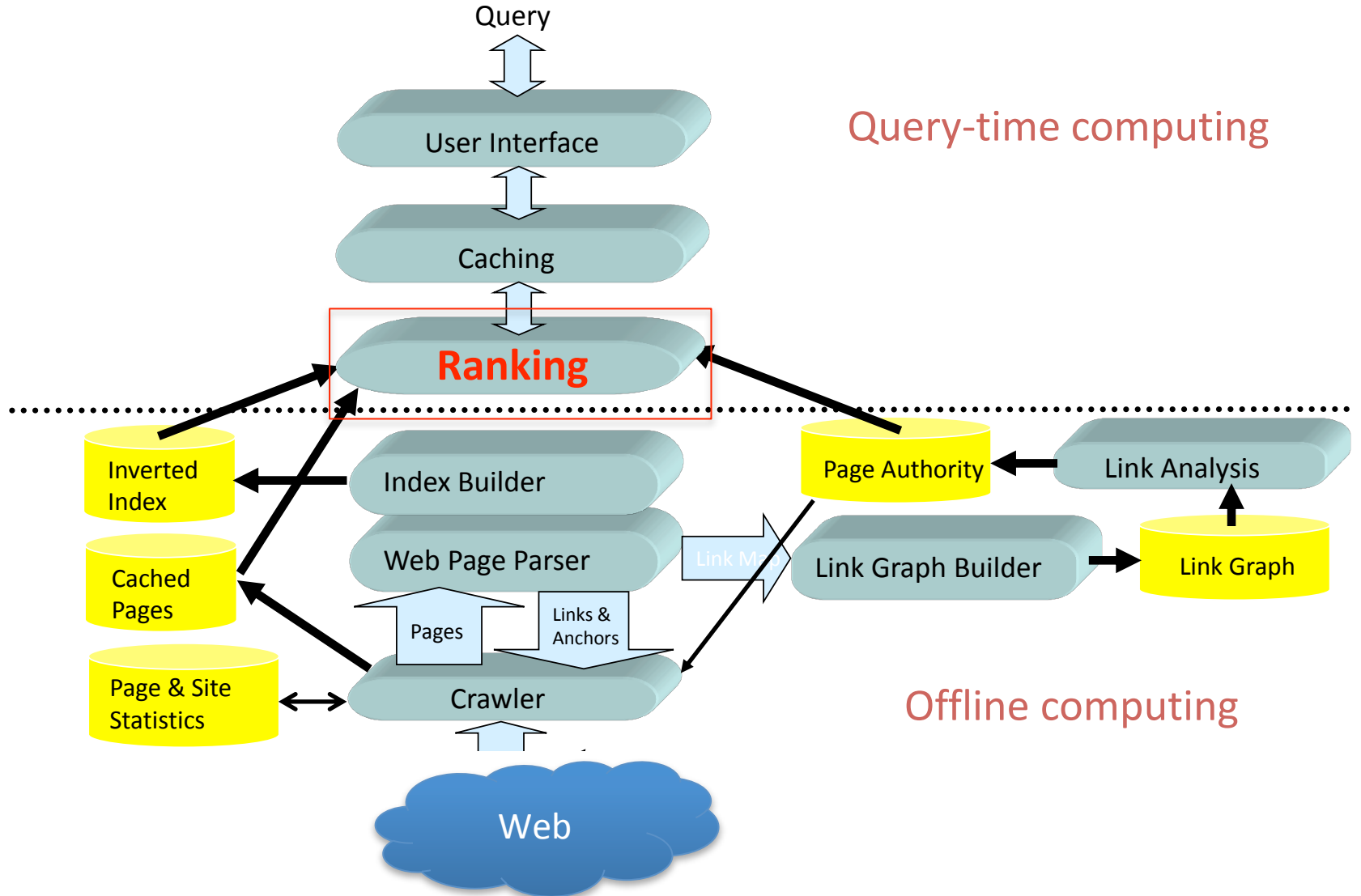
Créer index

recherch → {1:2 (2,4)}

informat → {1:1 (6)} index inversé

actif → {1:1 (9)}

Ouvrons le capot



Recherche / *ranking*

- Recherche booléenne

recherche AND information OR navigation

recherch → 1, 2, 4, 8, 10

informat → 1, 3, 8

navig → 5, 6



recherch AND informat → 1, 8

recherch AND informat OR navig → 1,5,6,8

Modèle vectoriel

- Chaque mot retenu définit une dimension ($\sim 100K - 1 M$)
- Espace vectoriel

$$\langle t_1, t_2, t_3, \dots, t_n \rangle$$

- Document

$$D = \langle a_1, a_2, a_3, \dots, a_n \rangle$$

$a_i =$ poids de t_i dans D

- Requête

$$Q = \langle b_1, b_2, b_3, \dots, b_n \rangle$$

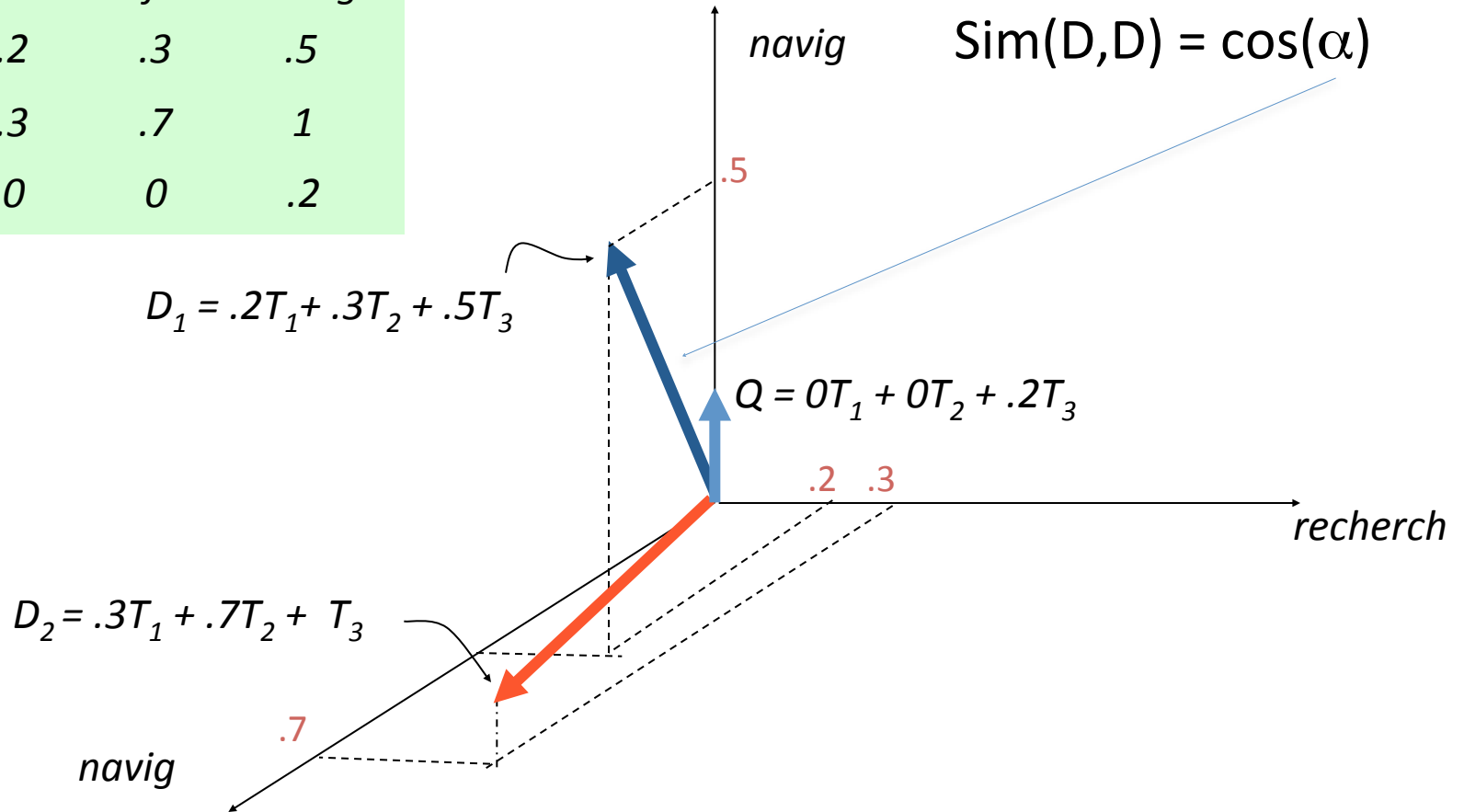
$b_i =$ poids de t_i dans Q

- $R(D, Q) = \text{Sim}(D, Q)$

Illustration

Exemple:

	<i>recherch</i>	<i>inform</i>	<i>navig</i>
$D_1 =$.2	.3	.5
$D_2 =$.3	.7	1
$Q =$	0	0	.2



Pondération des termes – $tf*idf$

tf – term frequency

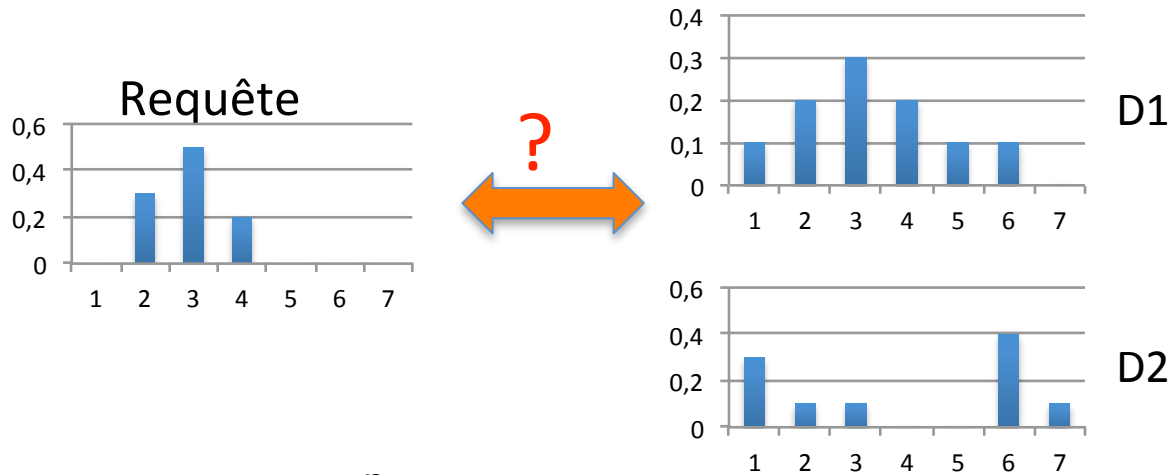
idf – inverse document frequency

Intuition:

- Plus $tf(t, D)$ est élevée, plus t est important
- Plus t est distribué uniformément dans différents documents, moins il est important

$$tfidf(t, D) = tf(t, D) \log \frac{N}{df(t)}$$

Modèle de langue statistique



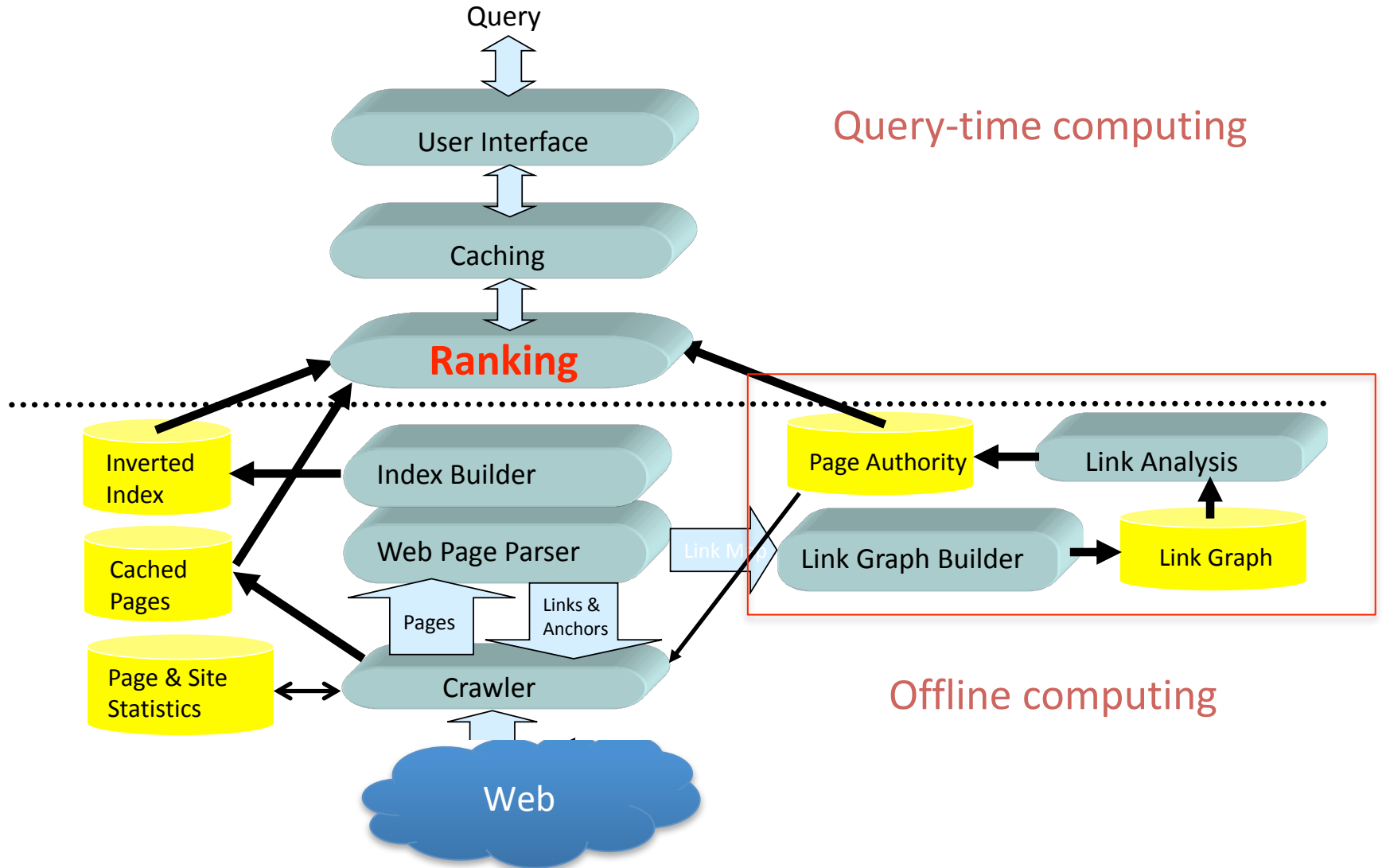
$$Score(Q, D) = \sum_{i=1}^n P(q_i | \theta_Q) * \log P(q_i | \theta_D)$$

Modèle de requête

Modèle de document

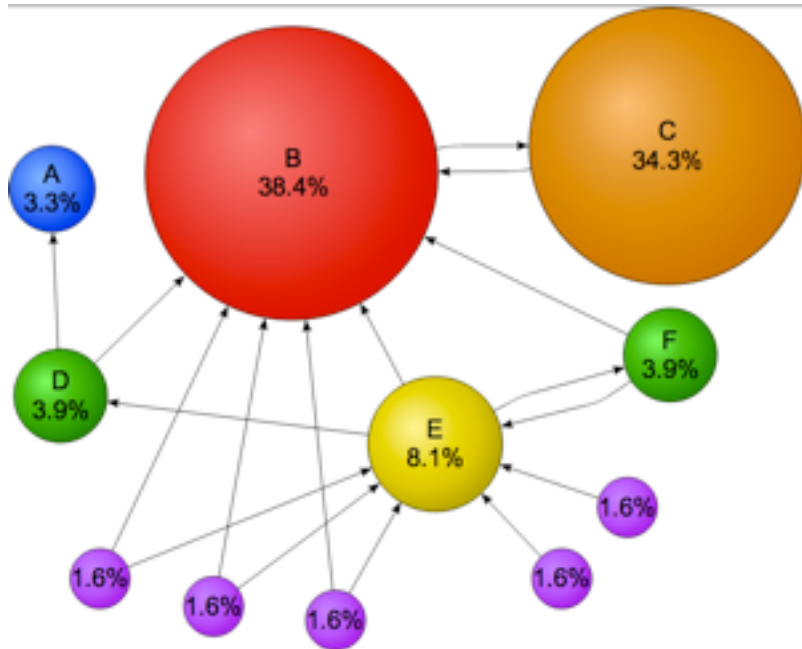
$$P(q_i | M_Q) = \frac{tf(q_i, Q)}{|Q|}$$

Ouvrons le capot



PageRank

- Un score « d'autorité » pour une page Web **voté** par les autres pages
 - Plus il y a des liens vers une page, plus cette page a d'autorité.
 - Plus les liens viennent des pages importantes, plus ce vote a d'importance



wikipedia.com

$$PR(p_i) = \frac{1-d}{N}$$

$d = 0.85$ damping factor

Anchor text

- Lien hypertexte venant d'une autre page (~une annotation)

The image shows a screenshot of a Wikipedia article titled "Moteur de recherche". Several callout boxes with blue borders and arrows point to specific anchor text elements on the page:

- ... [moteur de recherche](#) ...
- ... [engin de recherche](#) ...
- ... voir sur [wikipedia](#) ...
- ... [recherche d'information](#) ...
- ... [dépistage](#) ...
- ... [cliquer ici](#) ...

The article content includes a search bar at the top right, a navigation menu (Article, Discussion, Lire, Modifier, Afficher l'histoire), and a main heading "Moteur de recherche". A warning box states: "Cet article ne cite pas suffisamment ses sources. Si vous disposez d'ouvrages ou d'articles de référence ou si vous connaissez des sites web de qualité traitant du thème abordé ici, merci de compléter l'article en donnant les références utiles à sa vérifiabilité et en les liant à la section « Notes et références »." Below this, the text defines a search engine as a web application for finding resources, and mentions "robots" and "crawlers".

Surprises parfois

activer

About 43,900,000 results (0.38 seconds)

Tip: [Search for English results only](#). You can specify your search language in [Preferences](#)

[activer - Wiktionary](#)

en.wiktionary.org/wiki/activer

activer. Definition from Wiktionary, the free dictionary. Jump to: navigation, search ... gerund, en activant, en ayant **activé**. present participle, activant ...

[Pronunciation](#) - [Verb](#) - [Conjugation](#) - [Anagrams](#)

[Actinver :::](#)

www.actinver.com/ - [Translate this page](#)

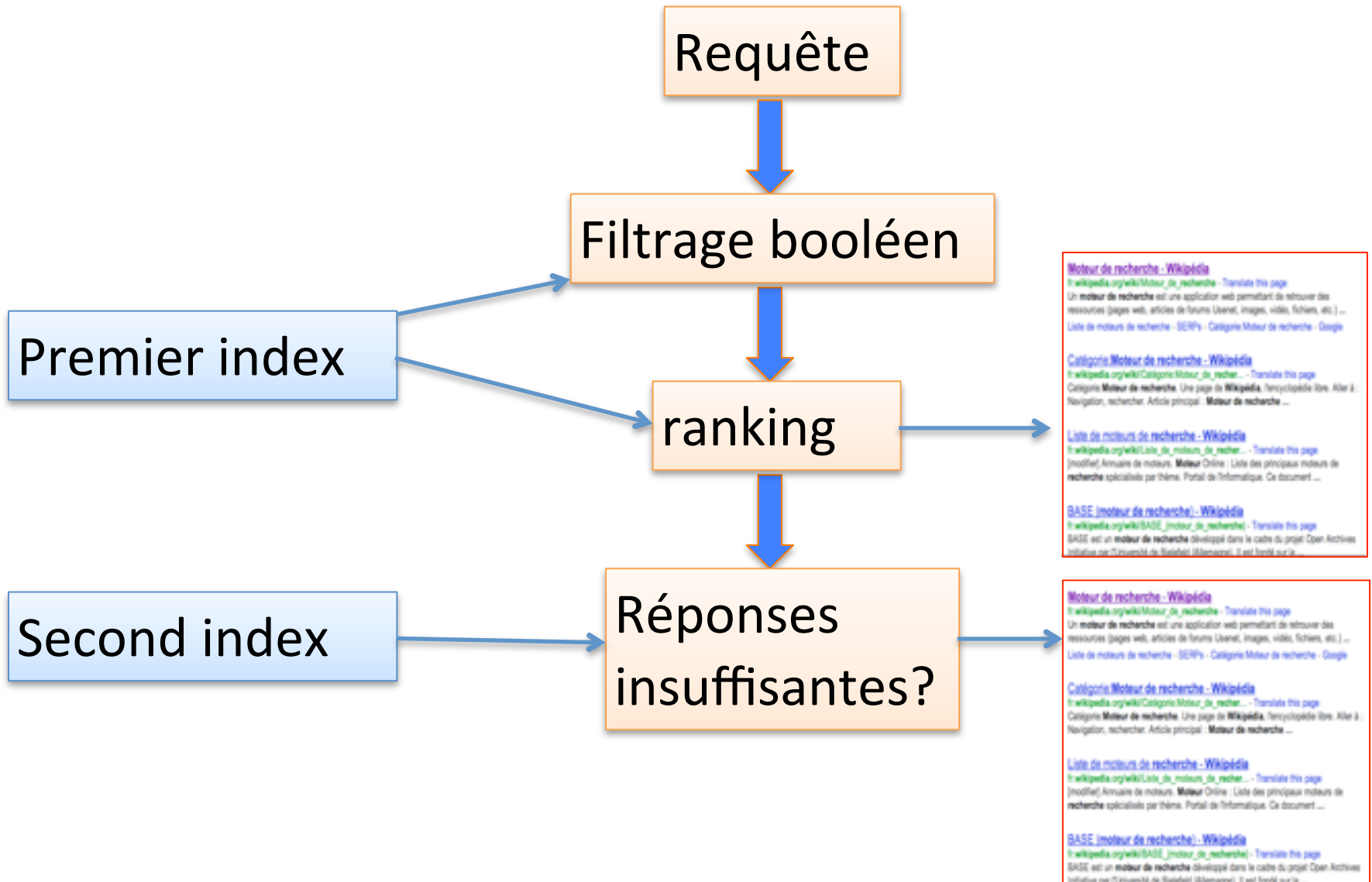
Empresa dedicada al manejo, promoción y administración de sociedades de inversión. Ofrece información sobre precios, rendimientos, preguntas frecuentes y ...

[Fondos de Inversión](#) - [Casa de Bolsa](#) - [Bursanet](#) - [Sucursales](#)

Combiner tout: *Learning to rank*

- Extraire des (milliers de) caractéristiques (features) pour Q-D
 - Poids $tf*idf$ (dans le titre, corps, anchor text, ...)
 - Score SIM(D,Q), Score modèle de langue, score BM25
 - ...
 - PageRank, popularité
 - Nombre de cliques
 - ... heuristiques
- Apprendre une fonction de *ranking* sur un ensemble d'exemples $\{(Q_i, D_j, score_{ij})\}$ afin d'ordonner les résultats le mieux possible

En pratique



Évaluation

- Évaluation par des organismes
 - Click-Through Rate (CTR) = taux de clique si présenté à l'utilisateur
- Évaluation de la qualité par des évaluateurs humains
 - Précision = documents pertinents retrouvé / retrouvés
 - Rappel = documents pertinents retrouvés / pertinents
 - Mean Average Precision (MAP) ~ Moyenne des précisions sur toutes les positions de documents pertinents
 - NDCG@k – Normalized Discounted Cumulative Gain

v_i = valeur du résultat i

$$NDCG@k = \sum_k^{-1} \frac{2^{v_i} - 1}{\log(i+1)}$$

5	(parfait)
4	(excellent)
3	(très bon)
2	(bon)
1	(correct)
0	(mauvais)

Premier problème –

- Mot comme index de base
 - système d'informatique → système, informatique
 - pomme de terre → pomme, terre (?)
 - Les mots ne sont pas indépendants dans une phrase (requête)
- Des idées
 - Regrouper des syntagmes ('pomme_de_terre')
 - Dictionnaire
 - séquences de mots (n-grammes)
 - Proximité pour une plus grande flexibilité

Proximité

- Utilisée par Google: Les documents contenant des mots de la requêtes à proximité (petite distance) sont favorisés

recherche d'information

Environ 433 000 000 résultats (0,19 secondes)

[Recherche d'information - Wikipédia](#)

fr.wikipedia.org/wiki/Recherche_d'information

Abrégée en RI ou IR (Information Retrieval en anglais), la **recherche d'information** est le domaine qui étudie la manière de répondre pertinemment à une ...

Catégorie Recherche ... - Système de recherche ... - Modèles cognitifs de la ...

[Les 6 étapes d'un projet de recherche d'information \(1996-2011 ...](#)

www.ets.umontreal.ca/etrouve/projet/index.htm

15 janv. 2011 – Les étapes présentées ci-dessous forment un tout. Dans une situation concrète de résolution de problème d'information, elles peuvent être ...

[Recherche de l'information](#)

cep.cyberscol.qc.ca > ... > Guides > Pédagogie de projet et ses composantes

La **recherche de l'information** se nomme parfois cueillette de données ou collecte d'information. C'est une étape importante de l'élaboration d'un projet et la ...

[La recherche d'information en classe](#)

www.csafluentes.qc.ca/rmi/

LA RECHERCHE D'INFORMATION EN CLASSE, dans le cadre de la démarche scientifique - Réalisation et crédits.

[Trousse de recherche efficace dans internet - Cégep@distance ...](#)

cefd.prosement.qc.ca/cours/trousse/introduction/

Guide méthodologique pour apprendre à **rechercher de l'information** sur internet et à l'analyser.

[UQAM | CIRIEC | Accueil](#)

www.ciriec.uqam.ca/

Le CIRIEC-Canada, Centre interdisciplinaire de **recherche et d'information** sur les entreprises collectives, est une association scientifique qui s'intéresse à ...

[Introduction à la recherche d'information dans InfoSphère](#)

www.bibliotheques.uqam.ca/infoSphere/sciences...?commencer2.html

Introduction à la **recherche d'information** >>>: Tester ses connaissances ... pertinentes et de les utiliser dans le cadre particulier d'une recherche précise.

[IRIS – Accueil](#)

www.iris-recherche.qc.ca/

L'actualité vue par IRIS Le fil twitter de IRIS Le fil RSS du blogue Le fil RSS des publications. L'IRIS est un institut de **recherche** sans but lucratif indépendant ...

[Images correspondant à recherche d'information -](#)

Signaler des images inappropriées



Proximité

recherche d'information

Environ 433 000 000 résultats (0,19 secondes)

[Recherche d'information - Wikipédia](#)
fr.wikipedia.org/wiki/Recherche_d'information

[Recherche d'information - Wikipédia](#)

fr.wikipedia.org/wiki/Recherche_d'information

Abrégée en RI ou IR (Information Retrieval en anglais), la **recherche d'information** est le domaine qui étudie la manière de répondre pertinemment à une ...

Catégorie:Recherche ... - Système de recherche ... - Modèles cognitifs de la ...

www.csafluents.qc.ca/rmi/

LA RECHERCHE D'INFORMATION EN CLASSE, dans le cadre de la démarche scientifique · Réalisation et crédits.

[Trousse de recherche efficace dans internet - Cégep@distance ...](#)

ccfd.crosemont.qc.ca/cours/trousse/introduction/

Guide méthodologique pour apprendre à **rechercher de l'information** sur internet et à l'analyser.

[UQAM | CIRIEC | Accueil](#)

www.ciriec.uqam.ca/

Le CIRIEC-Canada, Centre interdisciplinaire de **recherche et d'information** sur les entreprises collectives, est une association scientifique qui s'intéresse à ...

[Introduction à la recherche d'information dans InfoSphère](#)

www.bibliotheques.uqam.ca/infosphere/sciences.../commencer2.html

Introduction à la **recherche d'information** >>> Tester ses connaissances ... pertinentes et de les utiliser dans le cadre particulier d'une recherche précise.

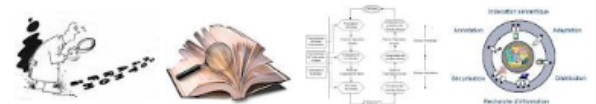
[IRIS – Accueil](#)

www.iris-recherche.qc.ca/

L'actualité vue par l'IRIS Le fil twitter de l'IRIS Le fil RSS du blogue Le fil RSS des publications. L'IRIS est un institut de **recherche** sans but lucratif indépendant ...

[Images correspondant à recherche d'information -](#)

Signaler des images inappropriées



Proximité

recherche d'information

Environ 433 000 000 résultats (0,19 secondes)

[Recherche d'information - Wikipédia](#)
fr.wikipedia.org/wiki/Recherche_d'information

[Recherche d'information - Wikipédia](#)

fr.wikipedia.org/wiki/Recherche_d'information

Abrégée en RI ou IR (Information Retrieval en anglais), la **recherche d'information** est le domaine qui étudie la manière de répondre pertinemment à une ...

Catégorie:Recherche ... - Système de recherche ... - Modèles cognitifs de la ...

www.csafluents.qc.ca/rmi/

LA RECHERCHE D'INFORMATION EN CLASSE, dans le cadre de la démarche scientifique · Réalisation et crédits.

[Trousse de recherche efficace dans internet - Cégep@distance ...](#)

ccfd.crosemont.qc.ca/cours/trousse/introduction/

Guide méthodologique pour apprendre à **rechercher de l'information** sur internet et à l'analyser.

[UQAM | CIRIEC | Accueil](#)

www.ciriec.uqam.ca/

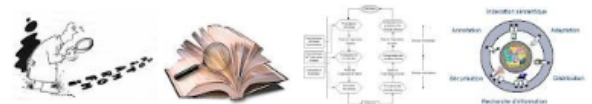
Le CIRIEC-Canada, Centre interdisciplinaire de **recherche et d'information** sur les

[IRIS – Accueil](#)

www.iris-recherche.qc.ca/

L'actualité vue par l'IRIS Le fil twitter de l'IRIS Le fil RSS du blogue Le fil RSS des publications. L'IRIS est un institut de **recherche sans but lucratif indépendant** ...

signaler des images inappropriées



Intégrer le critère de proximité

- Si les mots de requête apparaissent à proximité dans un document, booster son score selon la distance
- Requête = recherche information médicale
recherche-information, information-médicale, recherche-médicale

e.g.

$\text{tf}(\text{recherche}, D) + \lambda \text{prox}(\text{recherche}, \text{mots-contexte}, D)$

ou

utiliser une mesure de proximité comme feature additionnelle (*learning-to-rank*)

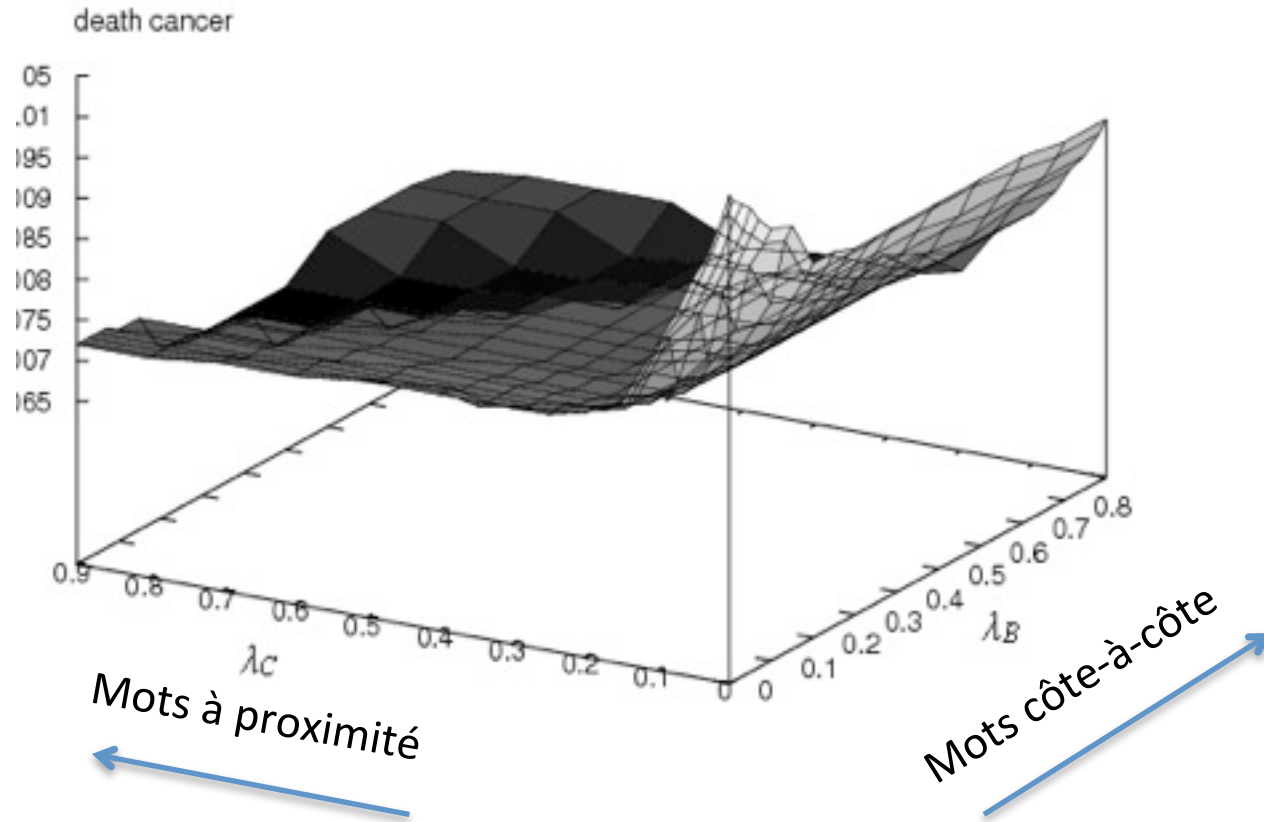
Dépendances variables

(Shi et Nie, CIKM 2010, AIRS 2010)

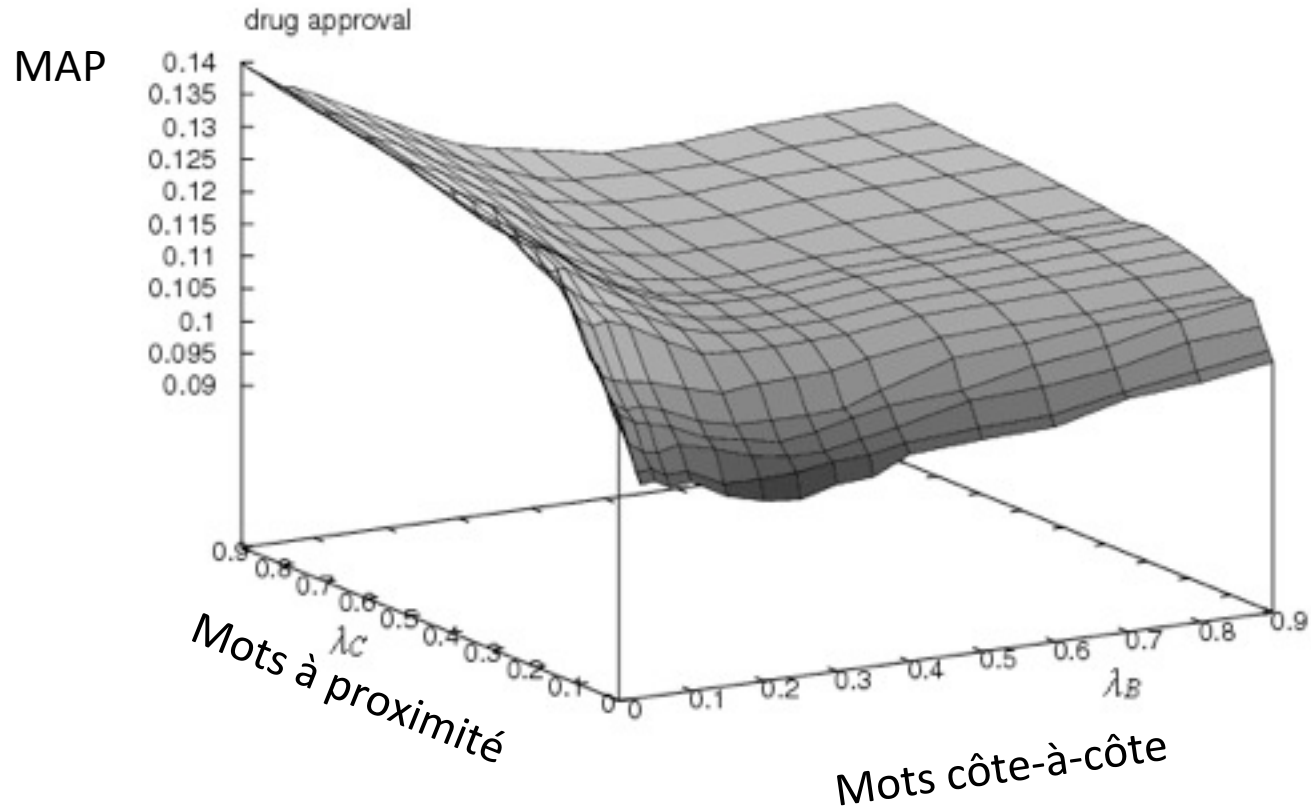
- Regrouper des mots: utile pour certains, mais inutile voire nuisible pour d'autres
 - Black Monday
 - Pomme de terre
 - Université de Montréal
 - Prolog input ?
 - death due to cancer ?
- Types de dépendance
 - Syntagme (côte-à-côte dans l'ordre)
 - Co-occurrence (dans proximité)

Death cancer – séparer

MAP

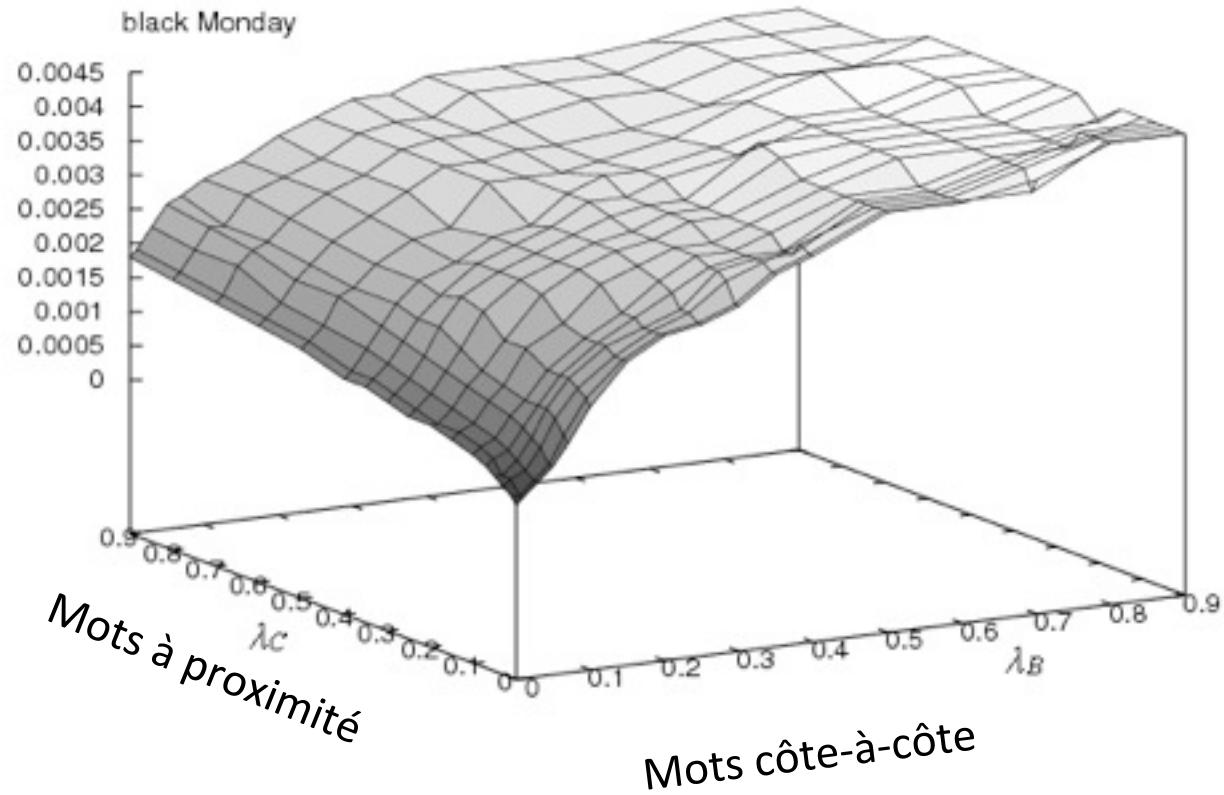


Drug approval – proximité



Black Monday – côte-à-côte et dans l'ordre

MAP



Deuxième problème – synonymes ou mots reliés

- On peut décrire une chose de multiples façons

Requête		Document
– recherche d'information	~	recherche documentaire
– computer		~ PC



- Comment retrouver des documents pertinents avec des mots différents?
 - ➔ Expansion: ajouter des mots reliés
 - Dans le document: Expansion de document

Expansion

computer game

Environ 808 000 000 résultats (0,20 secondes)

[PC Games. Computer Games - GameSpot.com](#)

www.gamespot.com/pc/index.html - Traduire cette page

PC - GameSpot is your source for PC and online **game** reviews, cheats, news, downloads, videos, previews, and walkthroughs.

[All PC Games, List of All PC ...](#)

PC Games - GameSpot is your source for the most ...

[Top PC Games, Best PC Video ...](#)

Top PC Video Games - GameSpot bring you the top reviewed PC ...

[Autres résultats sur gamespot.com »](#)

[PC game - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/PC_game - Traduire cette page

A **PC game**, also known as a **computer game**, is a video game played on a personal computer, rather than on a video game console or arcade machine.

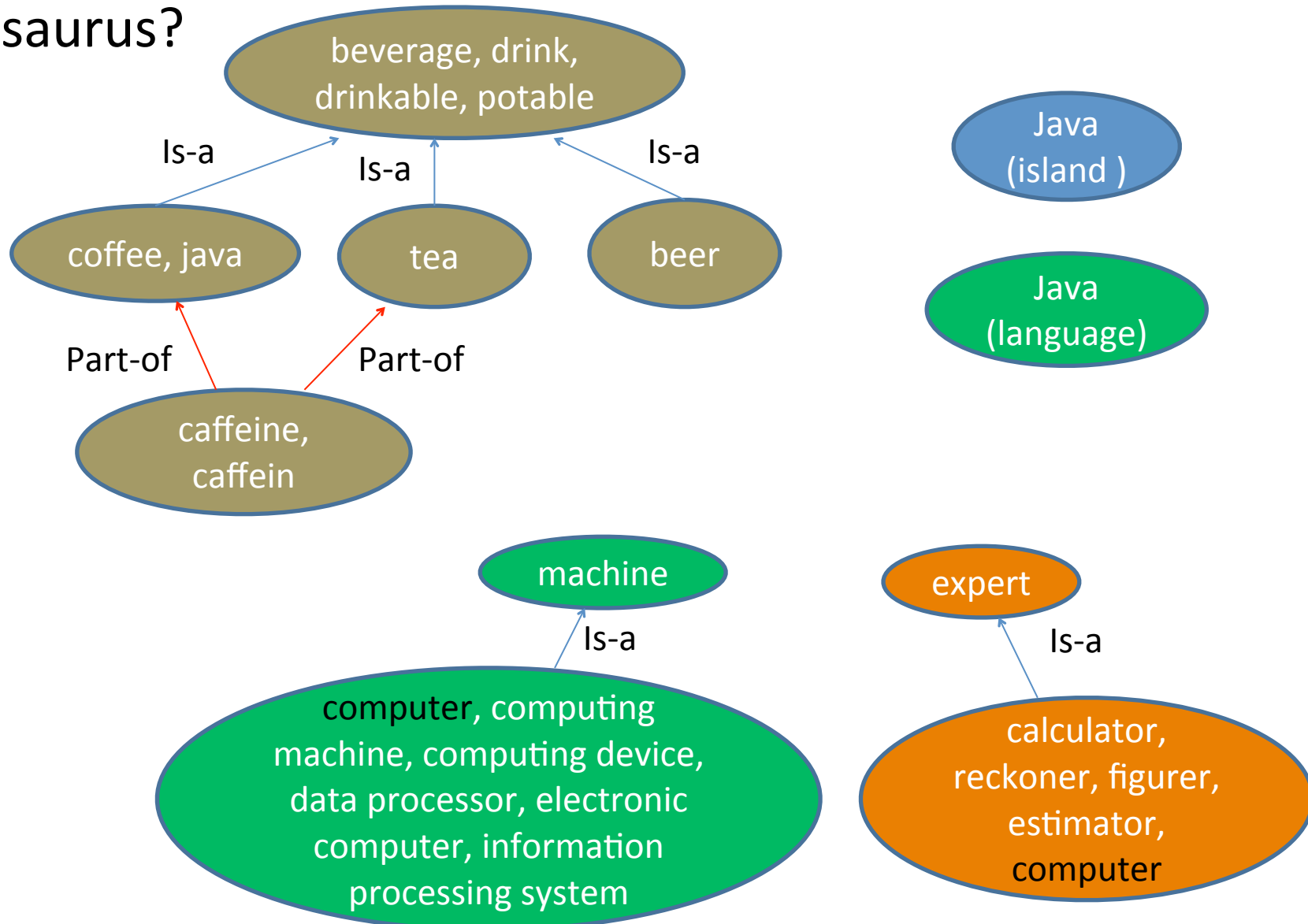
[Video game - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Video_game - Traduire cette page

In common use a "PC **game**" refers to a form of media that involves a player interacting with a IBM PC compatible personal computer connected to a video ...

Quels mots ajouter?

- Thésaurus?



Quels mots ajouter?

- Analyse de co-occurrences
 - Si deux mots apparaissent ensemble souvent, ils sont sémantiquement reliés.

L'**escouade Marteau**, bras armé de l'**Unité permanente anticorruption (UPAC)**, a procédé à l'**arrestation de 11 personnes**, jeudi, au terme d'une opération qui visait à démanteler un vaste stratagème de **collusion** qui aurait été échafaudé par **neuf entreprises de construction** de Saint-Jean-sur-Richelieu et des environs.

Escouade Marteau → anticorruption
→ construction

$$P(t | s) = \frac{c(t, s)}{\sum_{t_i} c(t_i, s)}$$

- Méthode prouvée utile
- Mais grande ambiguïté
 - Java → coffee
 - Java → île
 - Java → langage

Déterminer les mots reliés selon contexte

(Bai, Nie, et al. CIKM'05, EMNLP'06, SIGIR'07)

- Requêtes multi-mots:
 - java program sort
 - java hotel
 - java taste
- Remplacer la relation (java --> t) par (java, program --> t)
 - $P(s_1, s_2 \rightarrow t)$: le contexte (fenêtre) de s_1 et s_2 , est-ce que t apparaît?

$$P(t | s_1, s_2) = \frac{c(t, s_1, s_2)}{\sum_{t_i} c(t_i, s_1, s_2)}$$

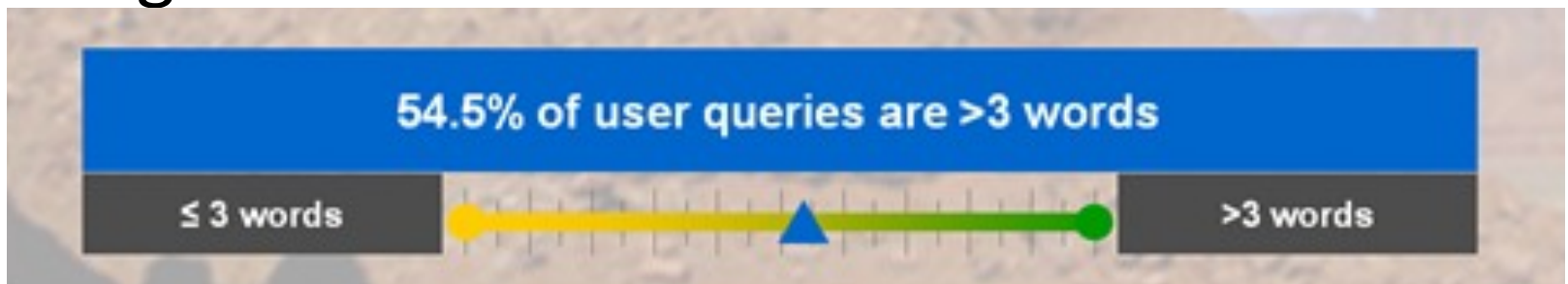
- Une approche très performante

Un mot n'apparaît souvent pas seul

- Hitwise, 2011:

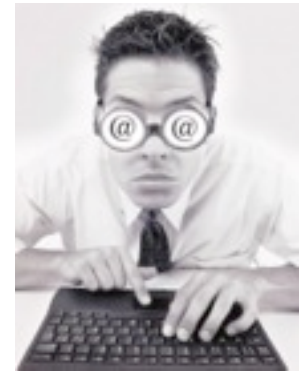
- 1 word – 26.45%
- 2 words – 23.66%
- 3 words – 19.34%
- 4 words – 13.17%
- 5 words – 7.69%
- 6 words – 4.12%
- 7 words – 2.26%

- Google 2012



Troisième problème –

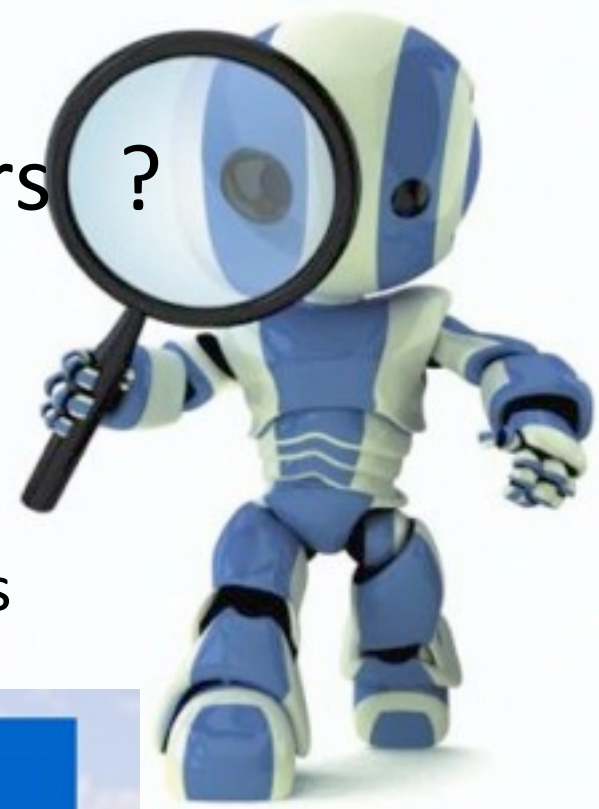
- Les utilisateurs
 - **Casse-tête** des moteurs de recherche: des intentions/expressions de recherche variées, imprévues et imprévisibles
 - **Amis** des moteurs de recherche: Ils enseignent aux moteurs de recherche comment faire (requête-cliques)
- Logs d'utilisateurs (*query logs*)
 - Stockage de toutes les interactions des utilisateurs



Query logs

[10/09 10:05:57] Query: furniture shopping [1-10]
[10/09 10:06:13] Click: [Webresult][q=furniture shopping][3]
<http://www.acybermall.com/>
[10/09 10:23:00] Query: hold everything [1-10]
[10/09 10:24:05] Query: hold everything catalog [1-10]
[10/09 10:24:06] Query: [Web]hold everything [11-20]
[10/09 10:24:06] Query: hold everything catalog [1-10]
[10/09 10:24:21] Query: [Web]hold everything catalog [11-20]
[10/09 10:24:41] Query: [Web]ethan allen [1-10]
[10/09 10:24:44] Click: [Webresult][q=ethan allen][1]
<http://navigation.realnames.com/resolver.dll>
[10/09 10:24:45] Click: [Webresult][q=ethan allen][1]
<http://navigation.realnames.com/resolver.dll>
[10/09 10:30:36] Query: tv media stand [1-10]
[10/09 10:30:50] Click: [Webresult][q=tv media stand][10]
http://www.gerpie.com/electronics/swivel_tv_stand.htm
[10/09 10:33:40] Query: tv furniture [1-10]
[10/09 10:34:04] Query: [Web]tv furniture [11-20]
[10/09 10:34:24] Click: [Webresult][q=tv furniture][17]
http://www.furnitureontheWeb.com/noframe/products/p_et11nf.htm

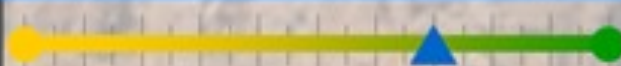
Que cherchent les utilisateurs ?



- Une simple statistique?
 - Les requêtes les plus populaires
 - Les expressions très variées dans les requêtes

70% of queries have no exact-matched keywords

Exact match



No exact match

- Synonymes
- Acronymes, ...
- Erreurs (>600 épellations différentes pour

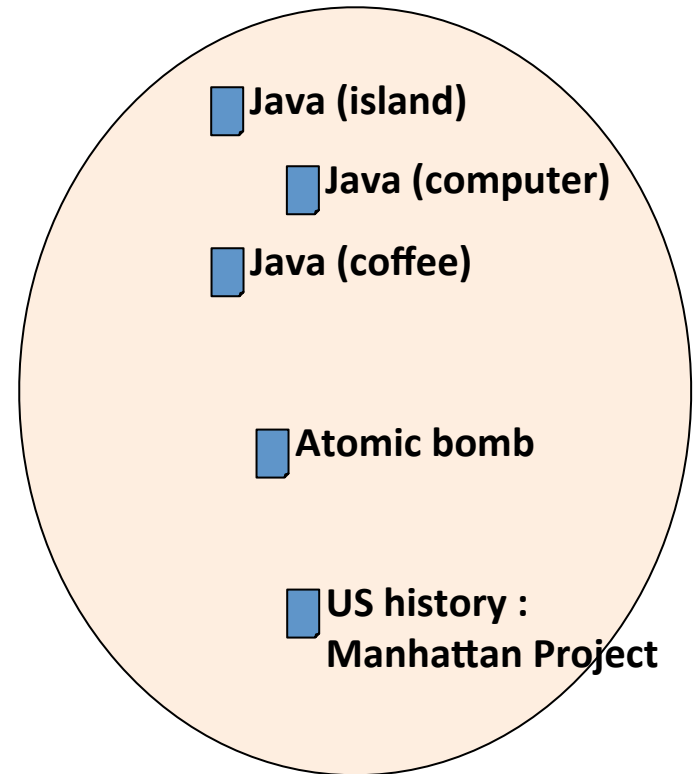
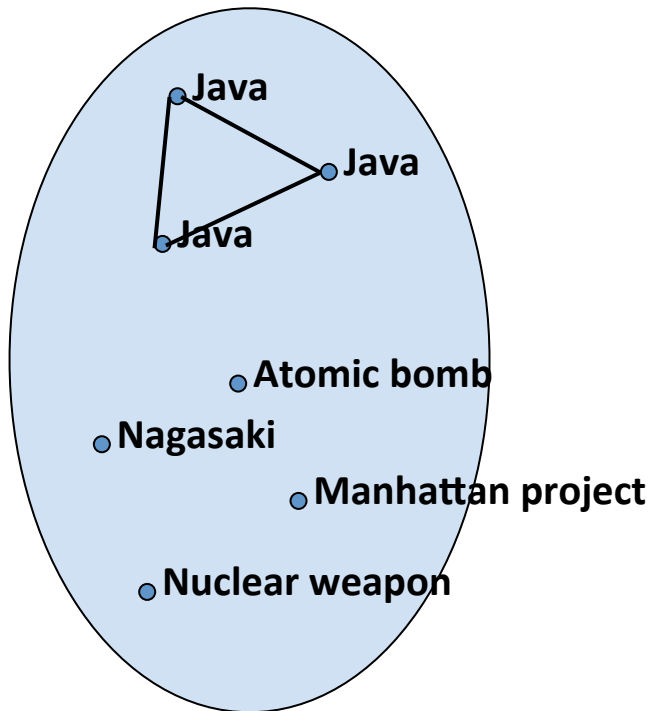
Britney Spears?

488941	britney spears	29	britent spears	9	brinttany spears
40134	brittany spears	29	brittnany spears	9	britanay spears
36315	brittney spears	29	britttany spears	9	britinany spears
24342	britany spears	29	btiney spears	9	britn spears
7331	britny spears	26	birttney spears	9	britnew spears
6633	briteny spears	26	breitney spears	9	britneyn spears
2696	britteny spears	26	brinity spears	9	britrney spears
1807	briney spears	26	britenay spears	9	brtiny spears
1635	brittny spears	26	britneyt spears	9	brtittney spears
1479	brintey spears	26	brittan spears	9	brtny spears
1479	britanny spears	26	brittne spears	9	brytny spears
1338	britiny spears	26	btittany spears	9	rbitney spears
1211	britnet spears	24	beitney spears	8	birtiny spears
1096	britiney spears	24	birteny spears	8	bithney spears
991	britaney spears	24	brightney spears	8	brattany spears
991	britnay spears	24	brintiny spears	8	breitny spears
811	brithney spears	24	britanty spears	8	breteny spears
811	brtiney spears	24	britenny spears	8	brightny spears
664	birtney spears	24	britini spears	8	brintay spears
664	brintney spears	24	britnwy spears	8	brinttey spears
664	briteney spears	24	brittni spears	8	briotney spears
601	bitney spears	24	brittnie spears	8	britanys spears
601	brinty spears	21	biritney spears	8	britley spears
544	brittaney spears	21	birtany spears	8	britneyb spears
544	brittnay spears	21	biteny spears	8	britrney spears
364	britey spears	21	bratney spears	8	brinty spears
364	brittiny spears	21	britani spears	8	brittner spears

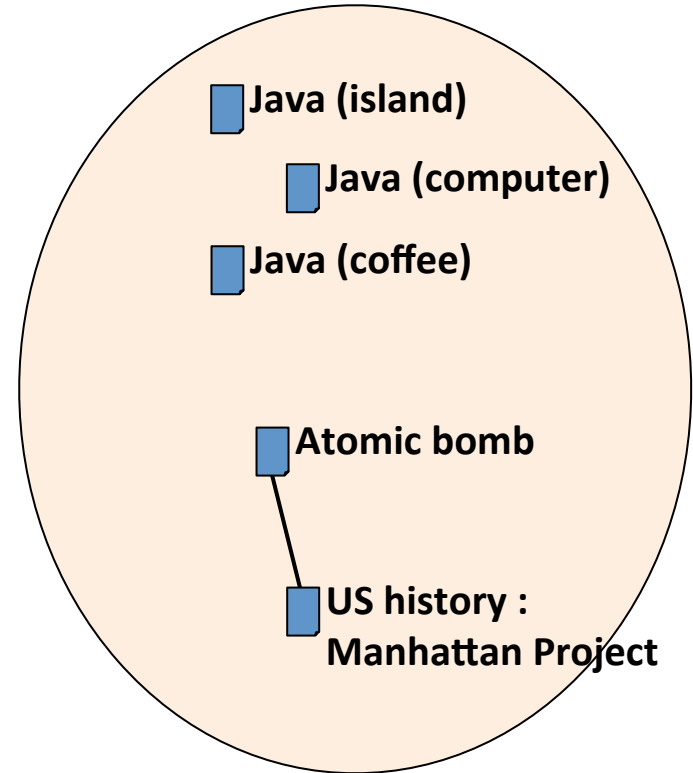
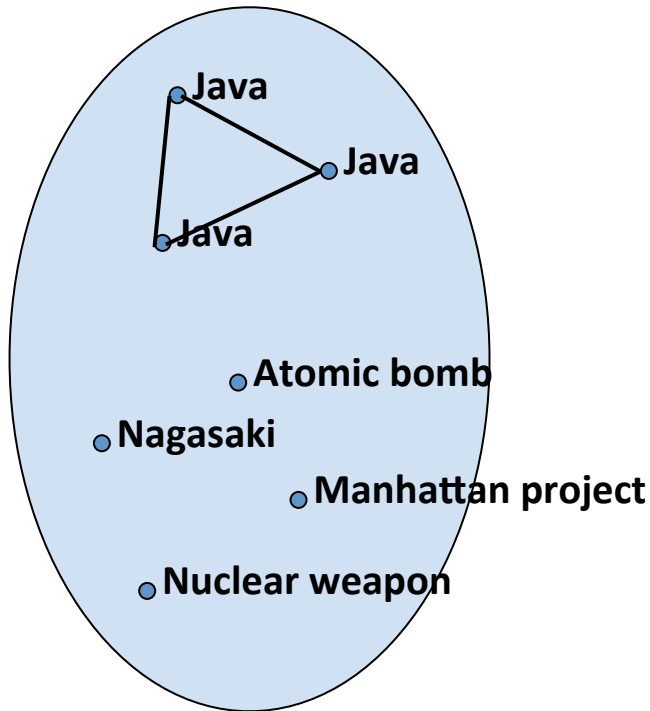
Exploiter les logs pour connaître les

- Est-ce que 2 requêtes différentes cherchent la même chose?
 - Deux requêtes sont reliées si elles utilisent des mots identiques ou similaires
 - Deux requêtes sont reliées si elles ont amené à cliquer les mêmes documents (*co-click*)
- ➔ combiner les 2 critères pour estimer la similarité des requêtes
- ➔ Regroupements les requêtes (clusters) ~

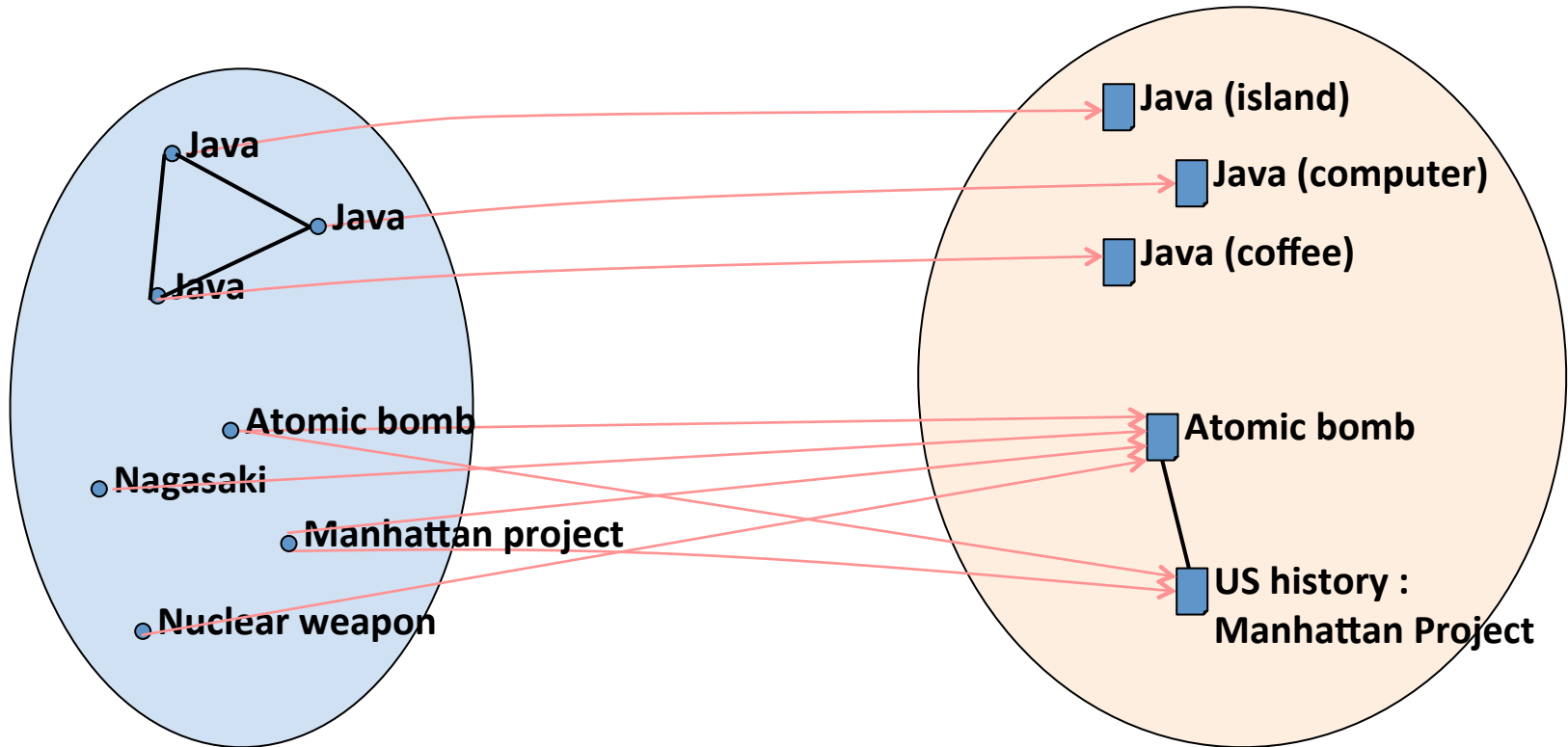
Espace de requêtes \neq Espace de documents



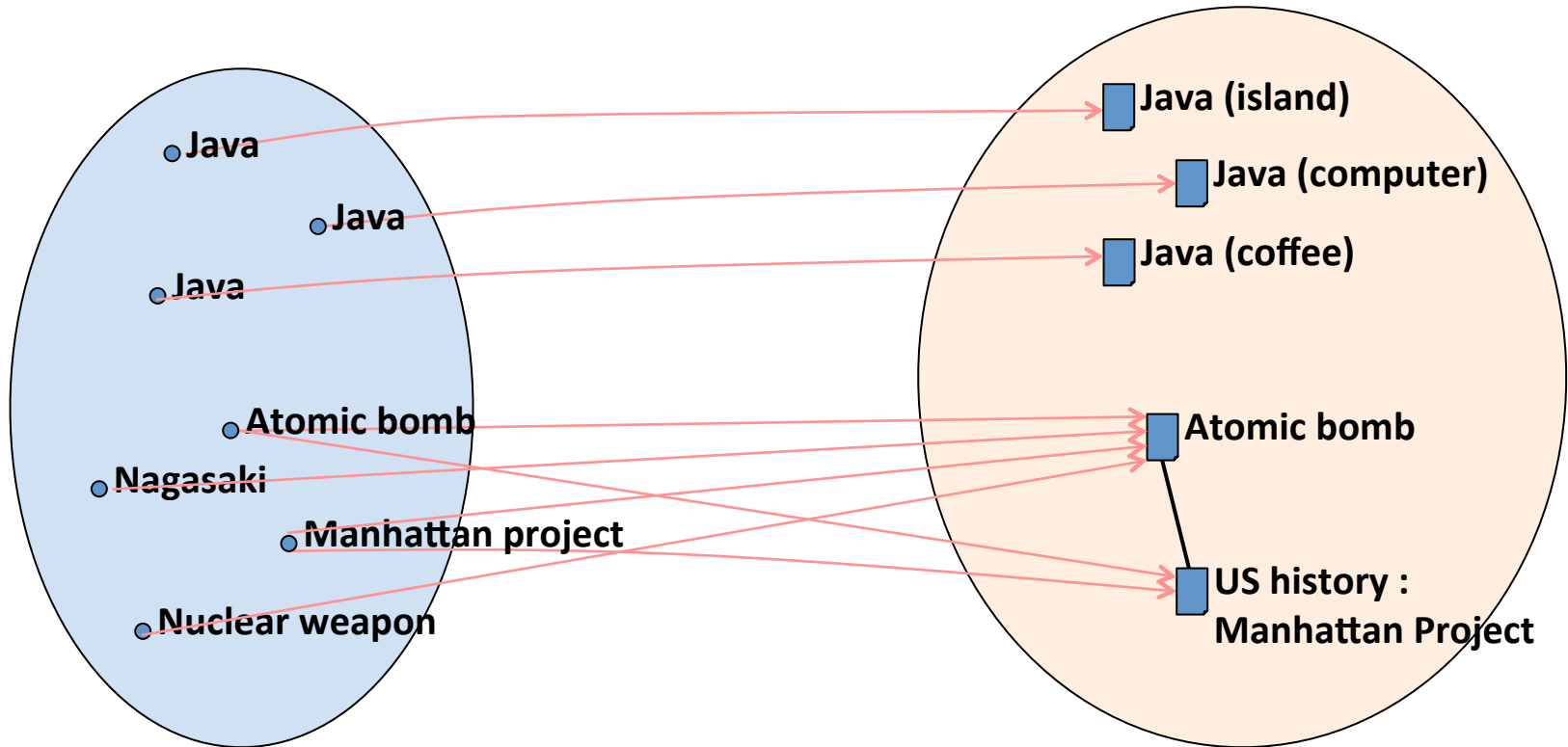
Espace de requêtes \neq Espace de documents



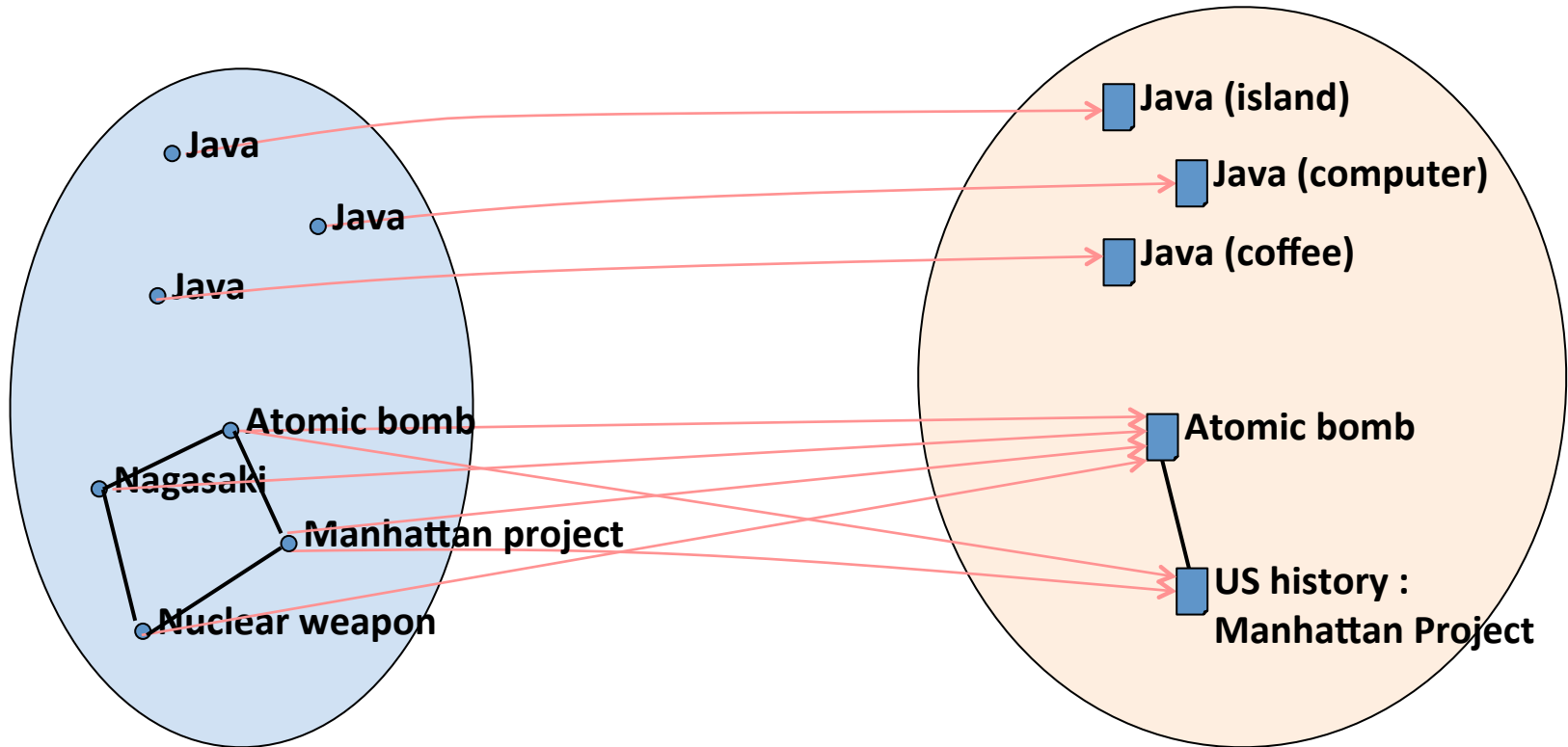
Espace de requêtes \neq Espace de documents



Espace de requêtes \neq Espace de documents

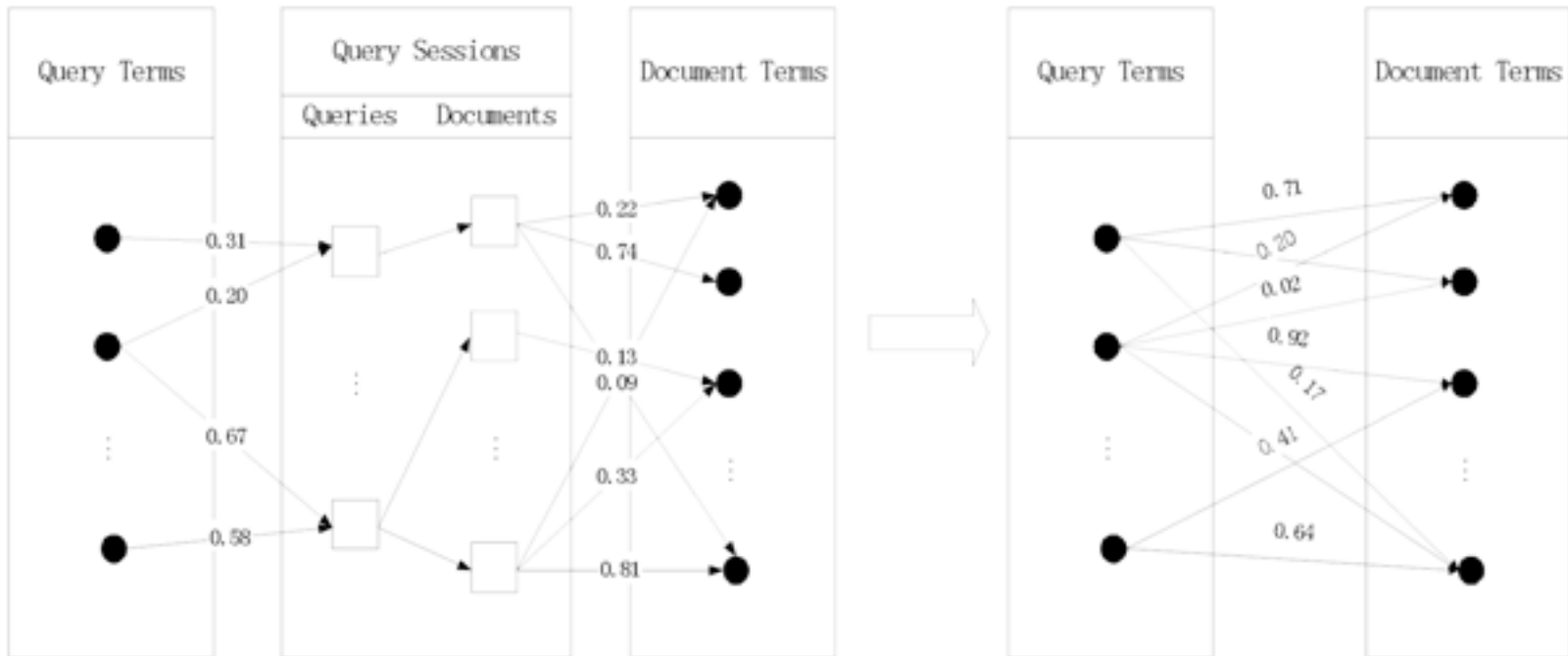


Espace de requêtes \neq Espace de documents



Aller plus loin avec les logs

- apprendre à relier les mots



- Utiliser les relations créées pour faire l'expansion de requête et de documents

État de l'art

- Générer un graphe bipartite (Q-D)
- Random walk
- Méthode largement utilisée dans les moteurs de recherche

État de l'art

- Générer un graphe bipartite (Q-D)
- Random wa
- Méthode la
de recherche
- D'autres uti
– Suggestion
requêtes



alain air

296 000 000 RÉSULTAT(S) Affiner par langue ▼ Affiner par pays ▼

Affichage des résultats pour **alan air** également.
Voulez-vous voir les résultats uniquement pour **alain air** ?

[Alan Air Media Services | About Cumbria's...](#) Traduire cette page
www.alanair.co.uk ▼
About **Alan Air**, Cumbria's leading PR public relations company.

[Alan Air Media Services | Journalism...](#) Traduire cette page
www.alanair.co.uk/journalism.htm ▼
Contact **Alan Air** for all your Journalism requirements in Carlisle, Cumbria

[Alain de Maricourt - Wikipédia](#)
fr.wikipedia.org/wiki/Alain_de_Maricourt ▼
Biographie · Maricourt et les ... · Voir aussi · Notes et références
Pour montrer l'avantage de l'armée de l'air sur celle de terre dans ce genre de missions s'effectuant dans les deux éléments, **Alain de Maricourt**, avec un léger brin ...

Surmonter la barrière de langue – recherche d'information tranalinguistique

- Requête en français, documents en anglais, chinois, japonais, ...
- Pourquoi faire ça?
 - On n'a pas toujours les informations pertinentes dans sa langue (informations locales)
 - Recherche exhaustive (examen d'une demande de brevet)
 - Informations indépendantes de langue
 - ...

Chercher un vol

vol de montréal paris

About 20,400,000 results (0.66 seconds)

Translated foreign pages



Translated results for **vol de montréal paris** - My language: [French](#) ▼

Language	Translated query	
English ✕	flight Montreal paris - Edit	19,700,000 results
Spanish ✕	vuelo Montreal paris - Edit	609,000 results

[Add language](#) ▼ - [Automatically select languages to search](#)

[Montréal à Paris \(France\) Vol](#)

www.cheapflights.ca/flights-to-Paris/Montreal/

Translated from: English

Vols pas chers à destination **de Paris** à partir **de Montréal**. Rechercher et comparer **Montréal (QC) à Paris (France) vols** avec Cheapflights.ca.

+ [Show original text](#)

Chercher des images

image d'avion



About 18,800,000 results (0.33 seconds)



La traduction est difficile, très difficile!



Exit

Attention: sol glissant



Oeufs de canard confis



龍牌

松花皮蛋

Preserved Duck Eggs
canard D'oeuf De Confiture

無鉛工藝

配料：鴨蛋、水、碱、食鹽、茶葉
Ingredients: Duck egg, Water,
Sodium carbonate, Salt, Tea.
Ingrédients: Oeufs du canard, L'eau,
Carbonate de sodium, Sel, Feuilles de thé.

Lost in translation !



Quelques exemples en RIT

- Vol Montréal Paris
 - Flight Montreal Paris

vol de montréal paris

About 20,400,000 results (0.66 seconds)

Translated foreign pages



Translated results for **vol de montréal paris** - My language: [French](#) ▼

Language	Translated query	
English ✕	flight Montreal paris - Edit	19,700,000 results
Spanish ✕	vuelo Montreal paris - Edit	609,000 results

[Add language](#) ▼ - [Automatically select languages to search](#)

[Montréal à Paris \(France\) Vol](#)

www.cheapflights.ca/flights-to-Paris/Montreal/

Translated from: English

Vols pas chers à destination **de Paris** à partir **de Montréal**. Rechercher et comparer **Montréal (QC) à Paris (France) vols** avec Cheapflights.ca.

+ [Show original text](#)

Quelques exemples en RIT

- Vol Montréal Paris
 - Flight Montreal Paris
- Vol de canards vers la floride
 - Flight of ducks?
 - Flying ducks

vol de canards vers la floride

Environ 94 100 000 résultats (0,19 secondes)

Pages en langue étrangère traduites



Résultats de recherche traduits pour : **vol de canards vers la floride** - Ma langue : [français](#) ▼

Langue

Recherche traduite

[anglais](#) ▼

ducks flying to florida - [Modifier](#)

[Ajouter une langue](#) ▼ - [Sélectionner automatiquement les langues pour la recherche](#)

[WEC243/UW287: Canards de la Floride](#)

[edis.ifas.ufl.edu/uw287](#)

Langue du texte original : anglais

Des représentants de chacune de ces **canards** peuvent être vus en **Floride** et sont décrits ci-dessous. Pour chaque...Tache blanche dans l'aile visible en **vol** peut être caché au repos...

[+ Afficher le texte original](#)

[Canard d'identification-Que le canard? - Florida Fish and Hunt](#)

[www.floridafishandhunt.com/Florida.../Florida.../duck-identification-...](#)

Langue du texte original : anglais

HOME FORUMS DE PÊCHE, **FLORIDE FLORIDE**, CHASSE, AVIS...Les pattes arrière et les ailes arrondies de ces **canards** volants lents rend l'air plus grand...

[+ Afficher le texte original](#)

[Vols pas chers à Duck Key, Floride - à partir de 151 € RT - TripAdvisor](#)

[www.tripadvisor.com/Flights-g34185-Duck_Key_Florida_Keys_Flor...](#)

Langue du texte original : anglais

Quelques exemples en RIT

- Vol Montréal Paris
 - Flight Montreal Paris
- Vol de canards vers la floride
 - Flight of ducks?
 - Flying ducks
- Vol de sirop d'érable
 - Flight of mapple syrup

vol de sirop d'érable

About 5,420,000 results (0.75 seconds)

Translated foreign pages



Translated results for **vol de sirop d'érable** - My language: [French](#) ▼

Language

Translated query

English ✕

flight of maple syrup - [Edit](#)

5,360,000 results

Spanish ✕

vuelo de jarabe de arce - [Edit](#)

57,200 results

[Add language](#) ▼ - [Automatically select languages to search](#)

[Foodies Ottawa - Forum - volants au sirop d'érable](#)

ottawafoodies.com/forum/3703

Translated from: English

J'ai volé avec du **sirop d'érable** en conserve au Royaume-Uni avant sans aucun problème dans les bagages enregistrés. Ne pouvait pas parler spécifiquement aux coutumes indiennes, mais je ne vois pas pourquoi il...

[+ Show original text](#)

[Est-ce que le sirop d'érable exploser sur un vol international plus de ...](#)

answers.yahoo.com/question/index?qid... - [United States](#)

Translated from: English

si oui, que dois-je faire pour éviter le gâchis? (Je dois le prendre avec moi pour...Non, le **sirop d'érable** ne va pas exploser dans la soute d'un avion commercial.

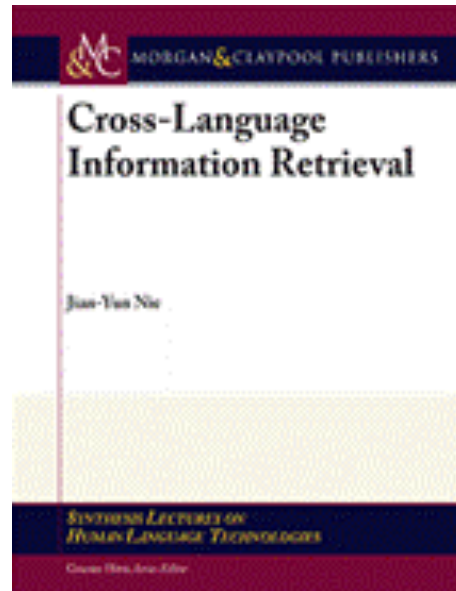
[+ Show original text](#)

La traduction de requête

- plus difficile que la TA, et différente

- Requêtes généralement courtes
- Pas de phrases complètes
- Nouveaux mots/expressions
- Est-ce que la traduction automatique est toujours l'outil idéal?
 - Dans certains cas, mais pas toujours
 - On désire avoir plusieurs traductions alternatives pour la RI (expansion)
 - La “traduction” peut inclure des mots reliés (traduire “fraud” par “illégal” est OK).
 - On doit choisir les traductions les plus *utiles* pour la recherche (“green food” en “aliment vert” ou “aliment bio”?)
 - Combiner la traduction et la suggestion d'une meilleure requête

Plus de détails sur la RI translinguistique



Une question de taille


- Google
 - 1998: 25 million de pages (10^7)
 - 2000: 1 milliard (10^9)
 - 2008: 1 billion (10^{12})
 - 2012: Index de 100 peta-octets de données (10^{15} ~
½ des documents imprimés de toute l'humanité)

Une question de taille

- Google
 - 1998: 25 million de pages (10^7)

The Google logo is displayed in its characteristic multi-colored font: 'G' in blue, 'o' in red, 'o' in yellow, 'g' in blue, 'l' in green, 'e' in red, and an exclamation point in blue.

Search the web using Google!

10 results  Google Search I'm feeling lucky

Index contains ~25 million pages (soon to be much bigger)

[About Google!](#)

[Stanford Search](#) [Linux Search](#)

Get Google! updates monthly!

[Subscribe](#) [Archive](#)

Copyright ©1997-8 Stanford University

Une question de taille

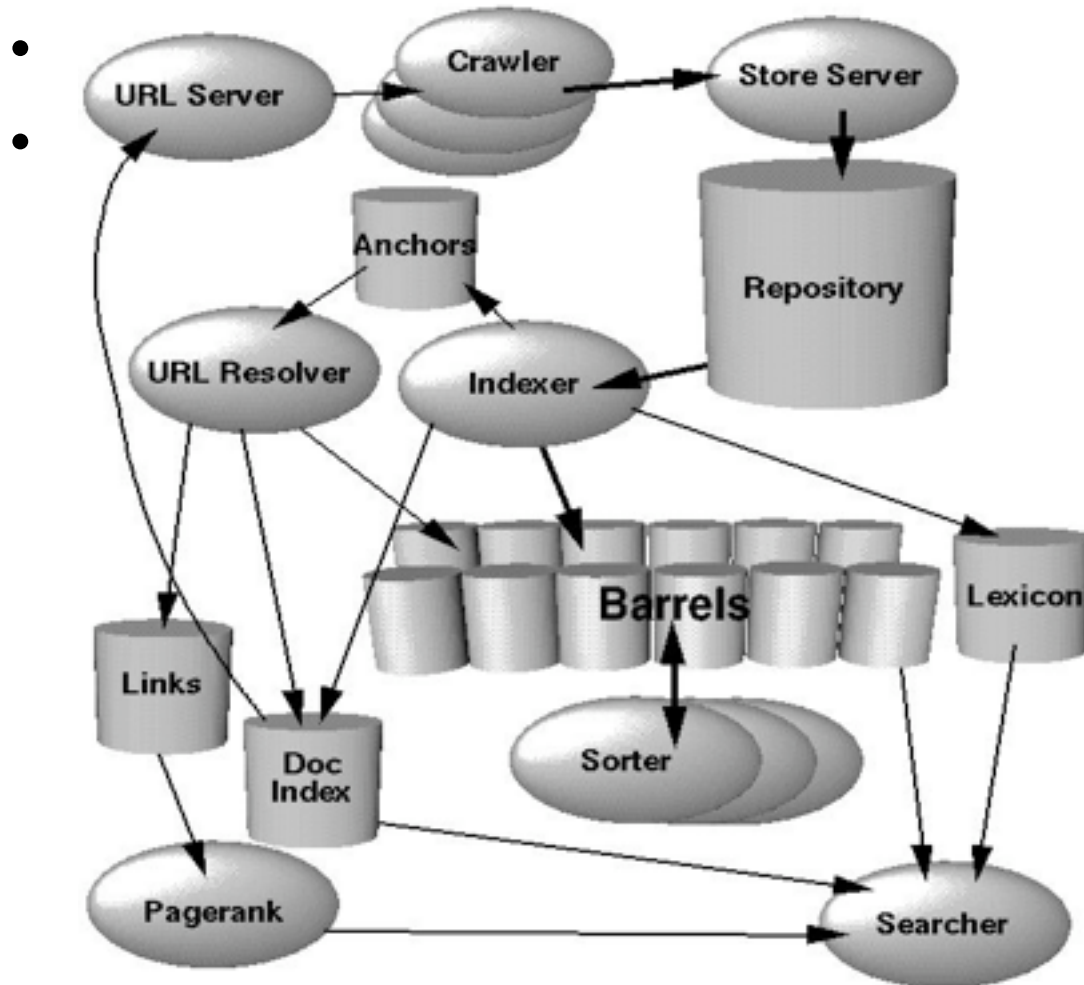
- Google
 - 1998: 25 million de pages (10^7)
 - 2000: 1 milliard (10^9)
 - 2008: 1 billion (10^{12})
 - 2012: Index de 100 peta-octets de données (10^{15} ~
½ des documents imprimés de toute l'humanité)

Rendre ceci possible

- Google 2010: 34,000 recherches /seconde, 2 millions /minute; 121 millions /heure; 3 milliards /jour; 88 milliards /mois
- *Cloud computing*, parallélisme massif
- Google: 900 000 serveurs (estimation en 2011 selon l'électricité consommée)

Rendre ceci possible

- Google 2010: 34,000 recherches /seconde, 2 millions /minute; 121 millions /heure; 3 milliards /jour; 88 milliards /mois



n 2011 selon

Rendre ceci possible

- Google 2010: 34,000 recherches /seconde, 2 millions /minute; 121 millions /heure; 3 milliards /jour; 88 milliards /mois
- *Cloud computing*, parallélisme massif
- Google: 900 000 serveurs (estimation en 2011 selon l'électricité consommée)

Rendre ceci possible

- Google 2010: 34,000 recherches /seconde, 2 millions /minute; 121 millions /heure; 3 milliards /jour; 88 milliards /mois
- *Cloud computing*, parallélisme massif
- Google: 900 000 serveurs (estimation en 2011 selon l'électricité consommée)



Vaut-il la peine de continuer à investir

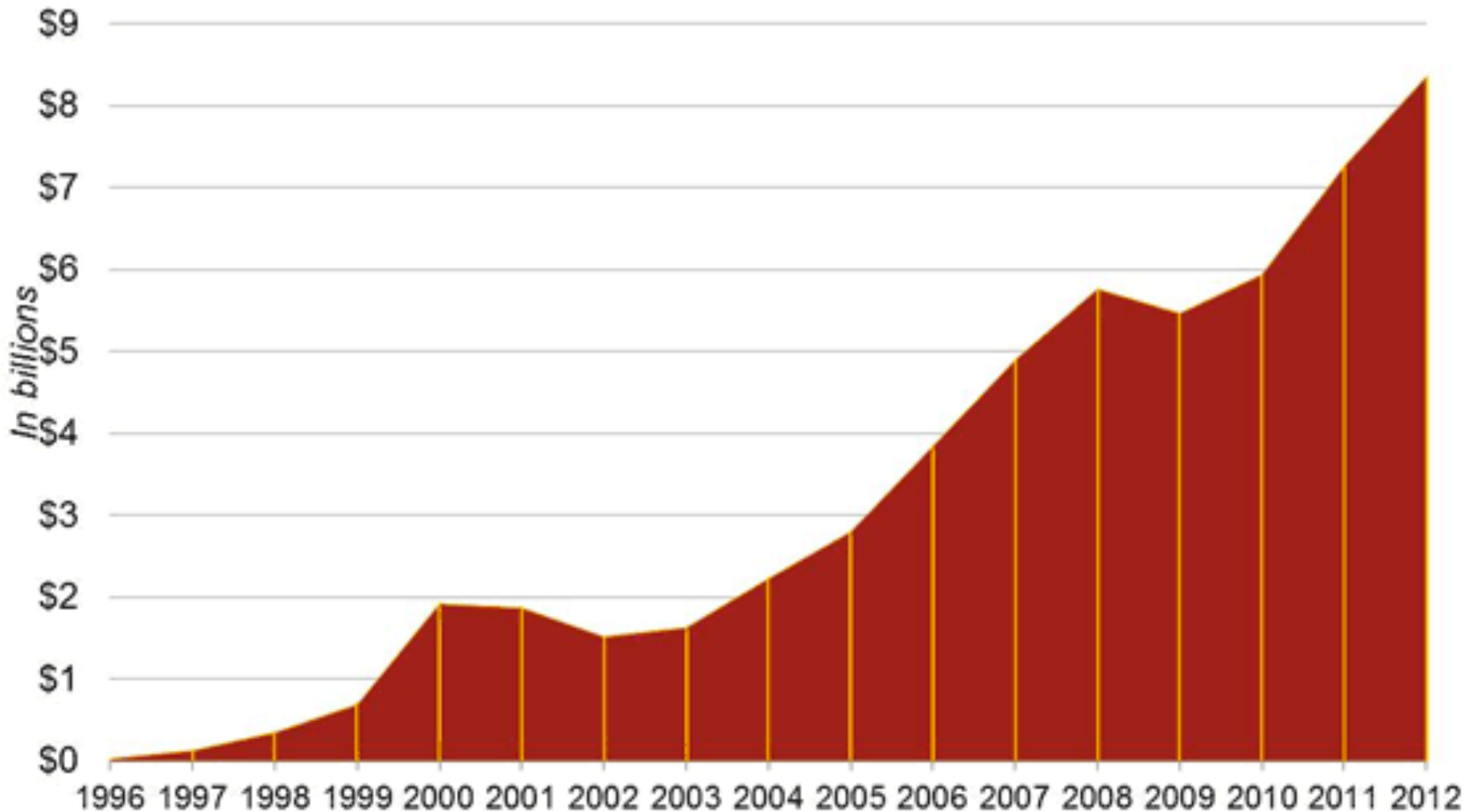
- La valeur de l'industrie de moteur de recherche est estimée à **16 milliard \$**.
- Google:
 - Revenue 2011: 37.9 milliard \$
 - Profit: 9.7 milliard \$
 - 96% du revenue: publicités

Vaut-il la peine de continuer à investir

- La valeur de l'industrie de moteur de recherche est estimée à **16 milliard \$**.
- Google:
 - Revenue 2011: 37.9 milliard \$
 - Profit: 9.7 milliard \$
 - 96% du revenue: publicités
- Q2 2012:
 - Pub on ligne: 8.4 milliard \$ (Interactive Advertising Bureau)
 - ... encore 95% de budgets pour les pub. pour les média traditionnels
 - Recherche (search) a un taux de conclusion (close rate) 14.6%, tandis que les méthodes *outbound* (email ou pub

Vaut-il la peine de continuer à investir

First Quarter Revenue Growth Trends, In billions — 1996-2012



Vaut-il la peine de continuer à investir

- La valeur de l'industrie de moteur de recherche est estimée à **16 milliard \$**.
- Google:
 - Revenue 2011: 37.9 milliard \$
 - Profit: 9.7 milliard \$
 - 96% du revenue: publicités
- Q2 2012:
 - Pub on ligne: 8.4 milliard \$ (Interactive Advertising Bureau)
 - ... encore 95% de budgets pour les pub. pour les média traditionnels
 - Recherche (search) a un taux de conclusion (close rate) 14.6%, tandis que les méthodes *outbound* (email ou pub

Vaut-il la peine de continuer à investir

- La valeur estimée
- Google
 - Recherche
 - Produits
 - 96% du revenue: publicités
- Q2 2012:
 - Pub on ligne: 8.4 milliard \$ (Interactive Advertising Bureau)
 - ... encore 95% de budgets pour les pub. pour les média traditionnels
 - Recherche (search) a un taux de conclusion (close rate) 14.6%, tandis que les méthodes *outbound* (email ou pub

Encore beaucoup de chemin à faire pour satisfaire les utilisateurs

Vaut-il la peine de continuer à investir

- La valeur estimée de Google
- Google encore beaucoup de chemin à faire pour satisfaire les utilisateurs
 - Recherche
 - Produits
 - 96% de satisfaction
- Q2 2010: L'avenir est devant nous!
 - Pub on ligne: 8.4 milliard \$ (Interactive Advertising Bureau)
 - ... encore 95% de budgets pour les pub. pour les médias traditionnels
 - Recherche (search) a un taux de conclusion (close rate) 14.6%, tandis que les méthodes *outbound* (email ou pub

Remerciement

- À tous mes étudiants et collaborateurs

