

# Démonter le moteur ... de recherche

Jian-Yun Nie  
DIRO

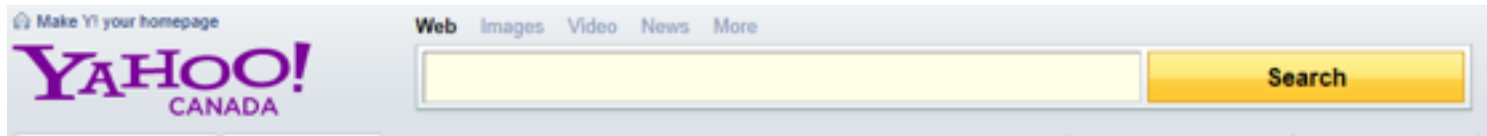
<http://www.iro.umontreal.ca/~nie/>  
[nie@iro.umontreal.ca](mailto:nie@iro.umontreal.ca)

# Moteur de recherche –

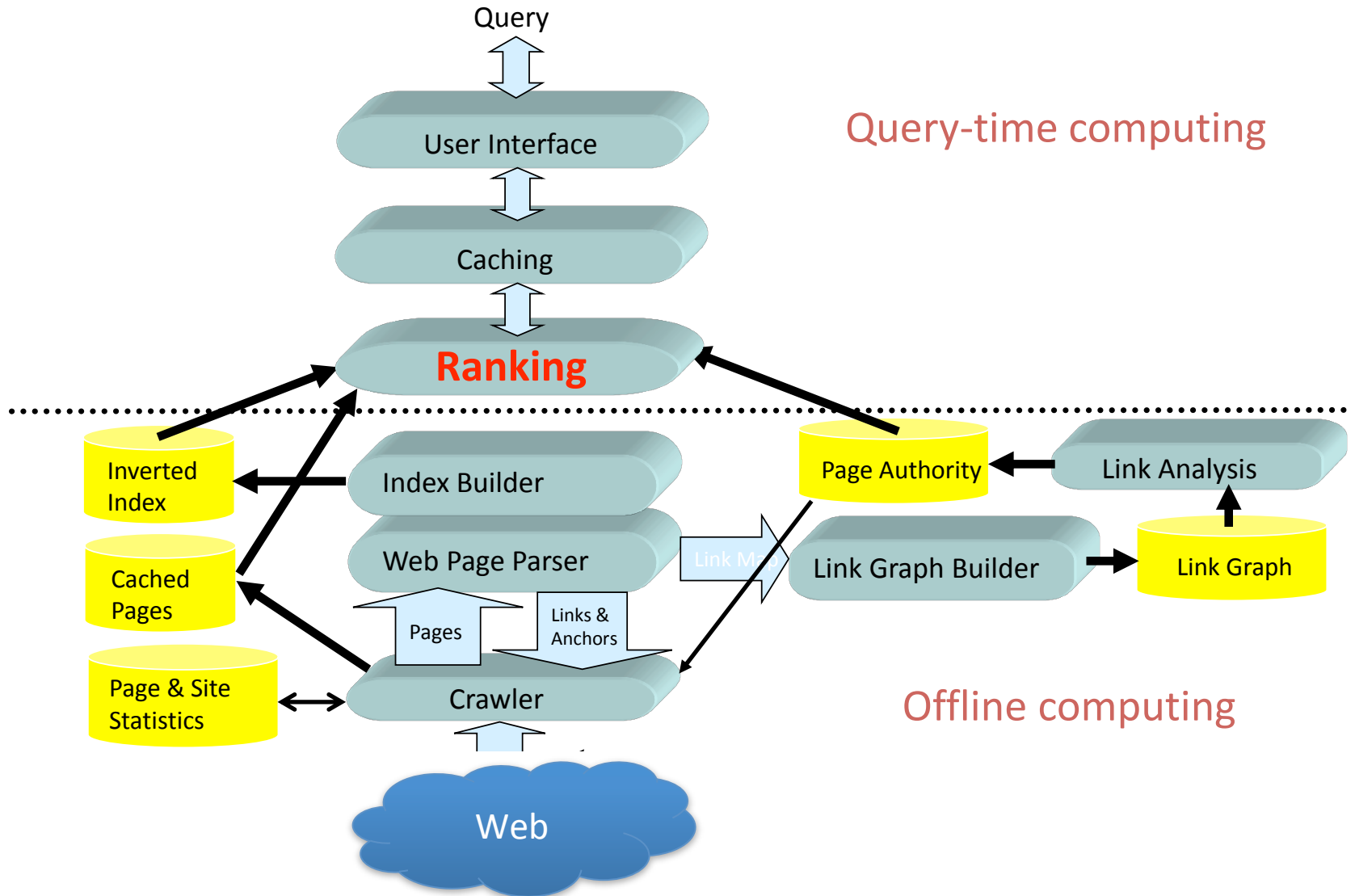
Search Engine journal <http://www.searchenginejournal.com/24-eye-popping-seo-statistics/42665/>

- 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

# Les faces visibles des moteurs de

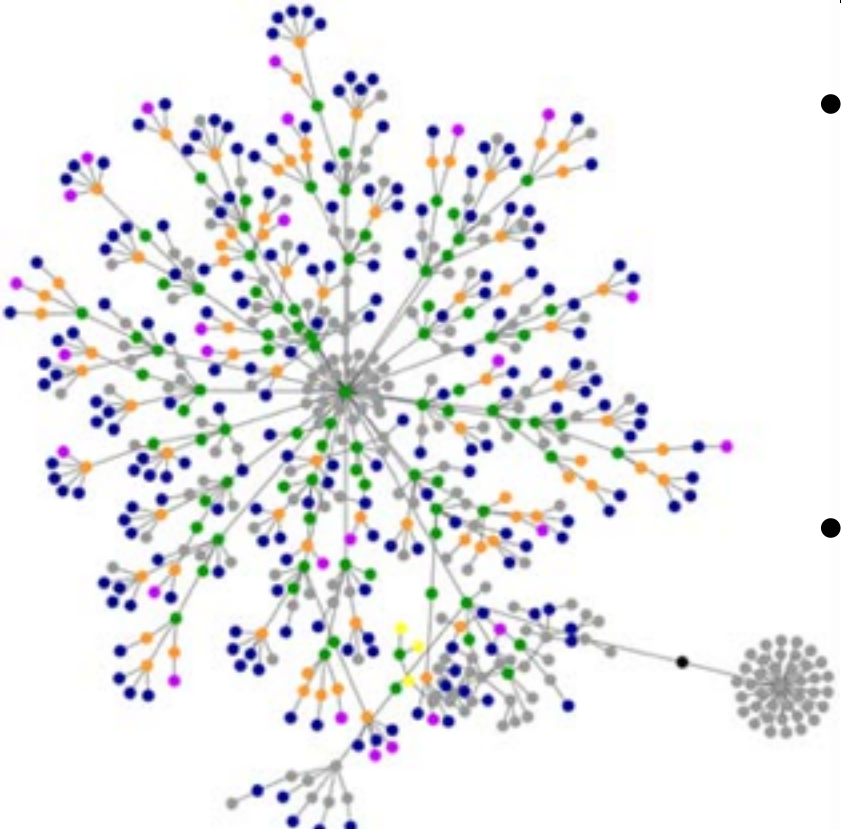


# La face cachée



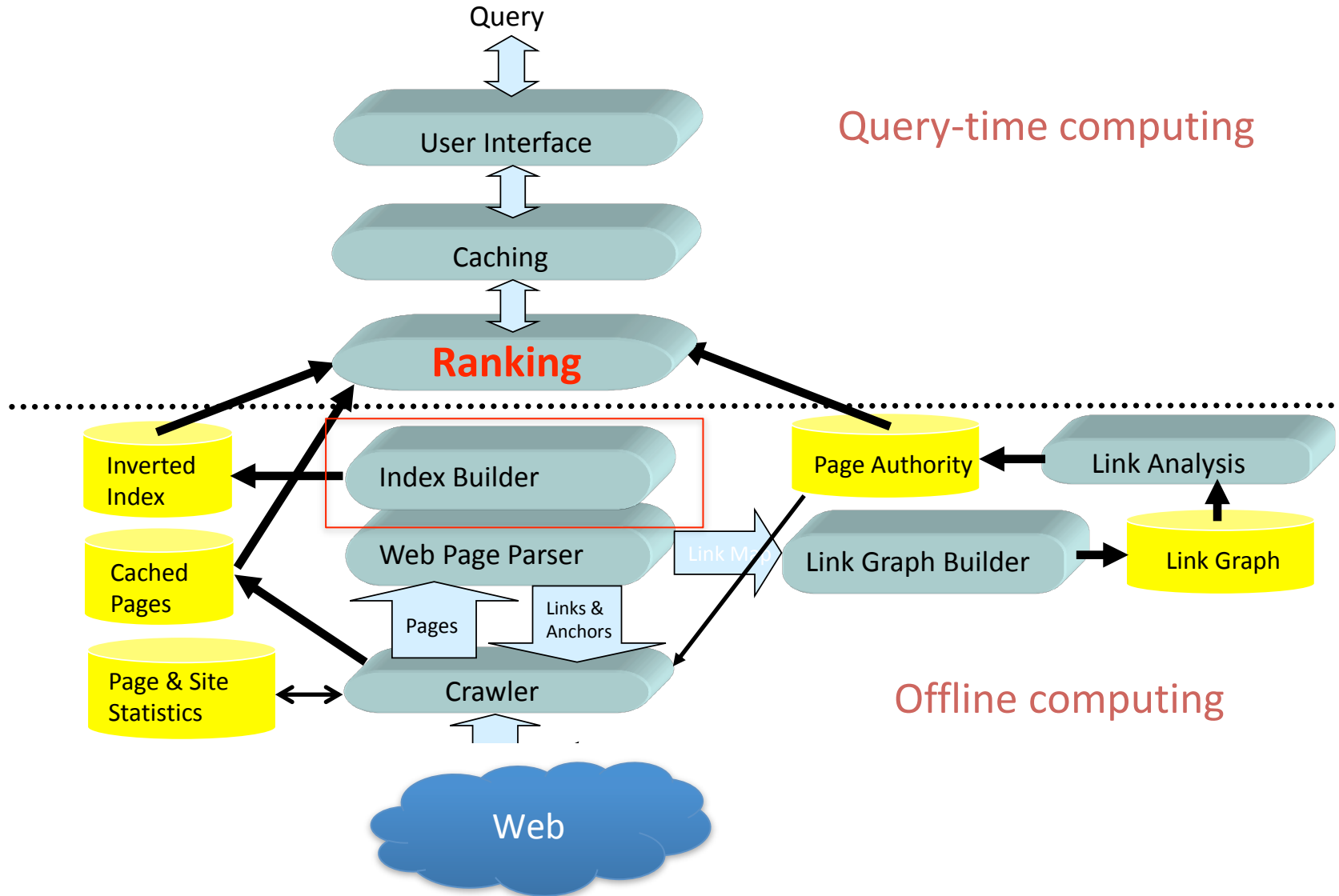
# Collecter des documents - principe

- 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



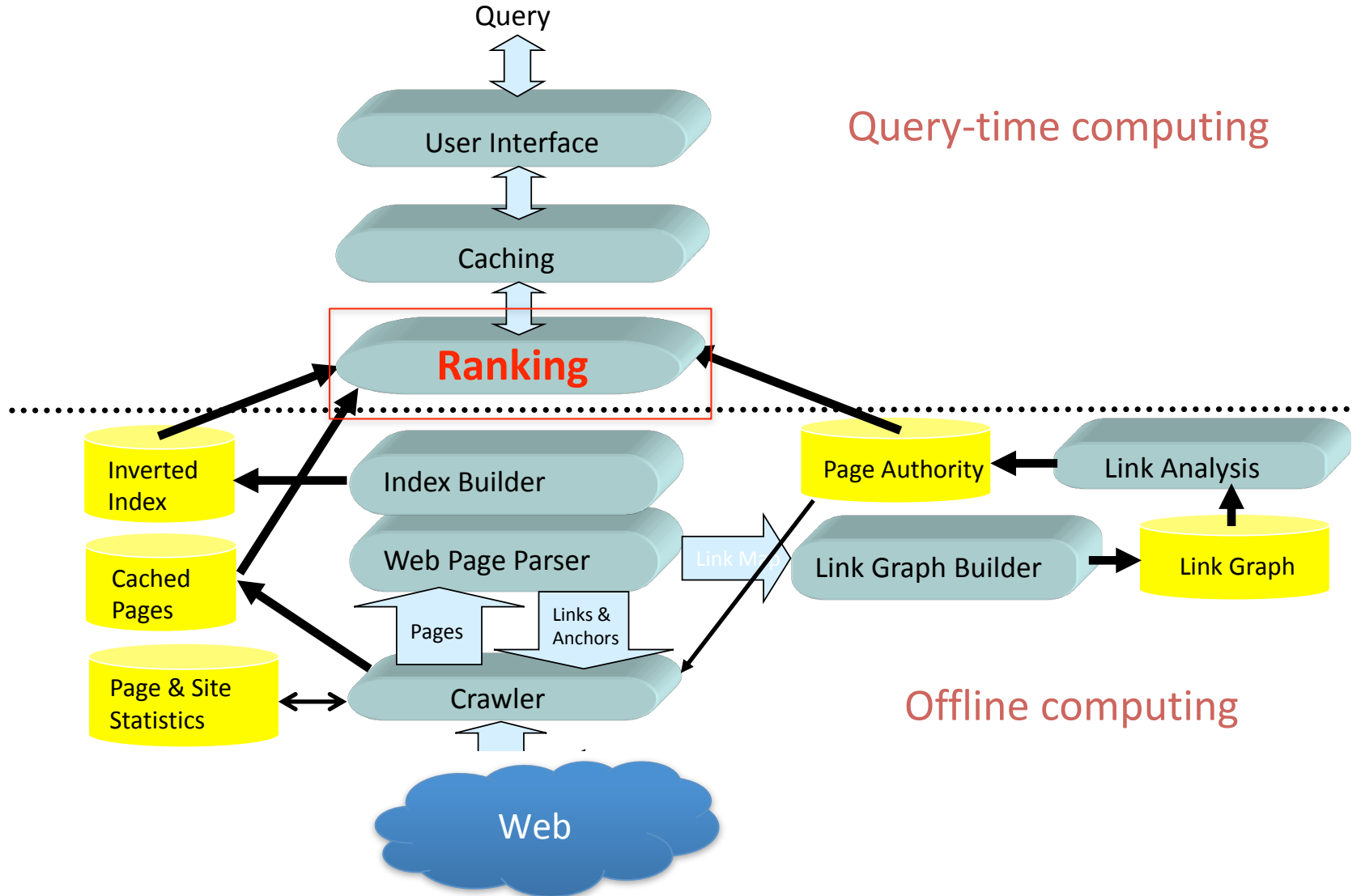
Créer index

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

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

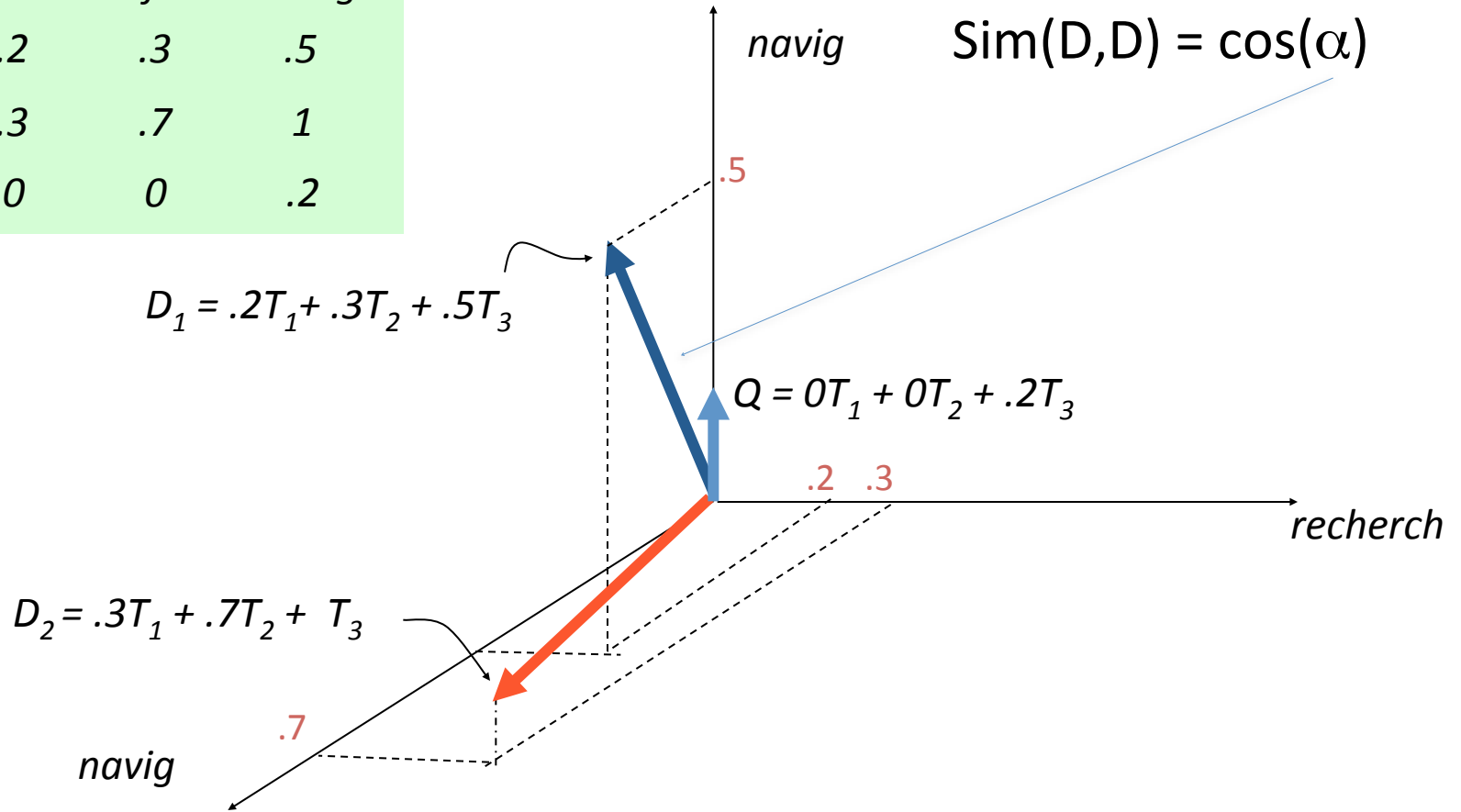
$b_i =$  poids de  $t_i$  dans  $Q$

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

# 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

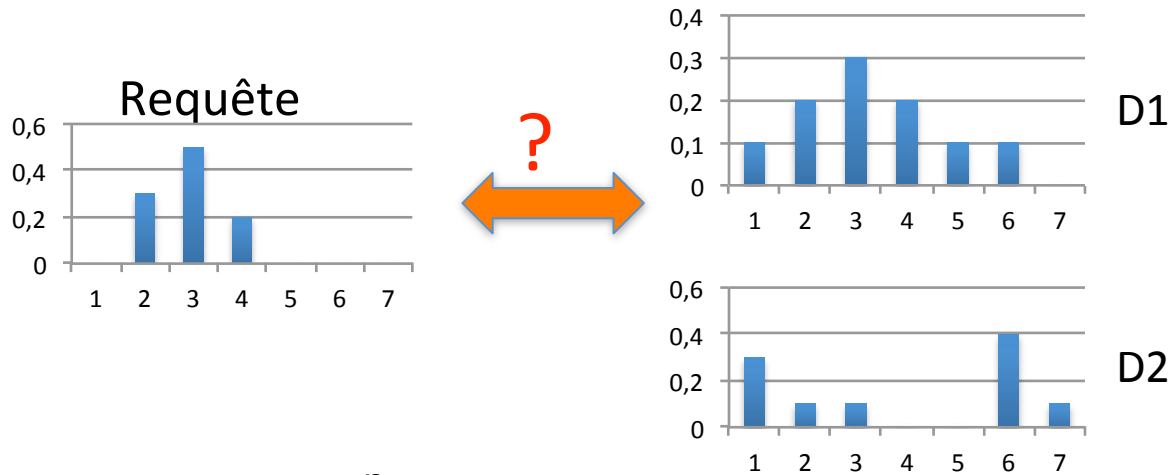
$idf$  – inverse document frequency

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



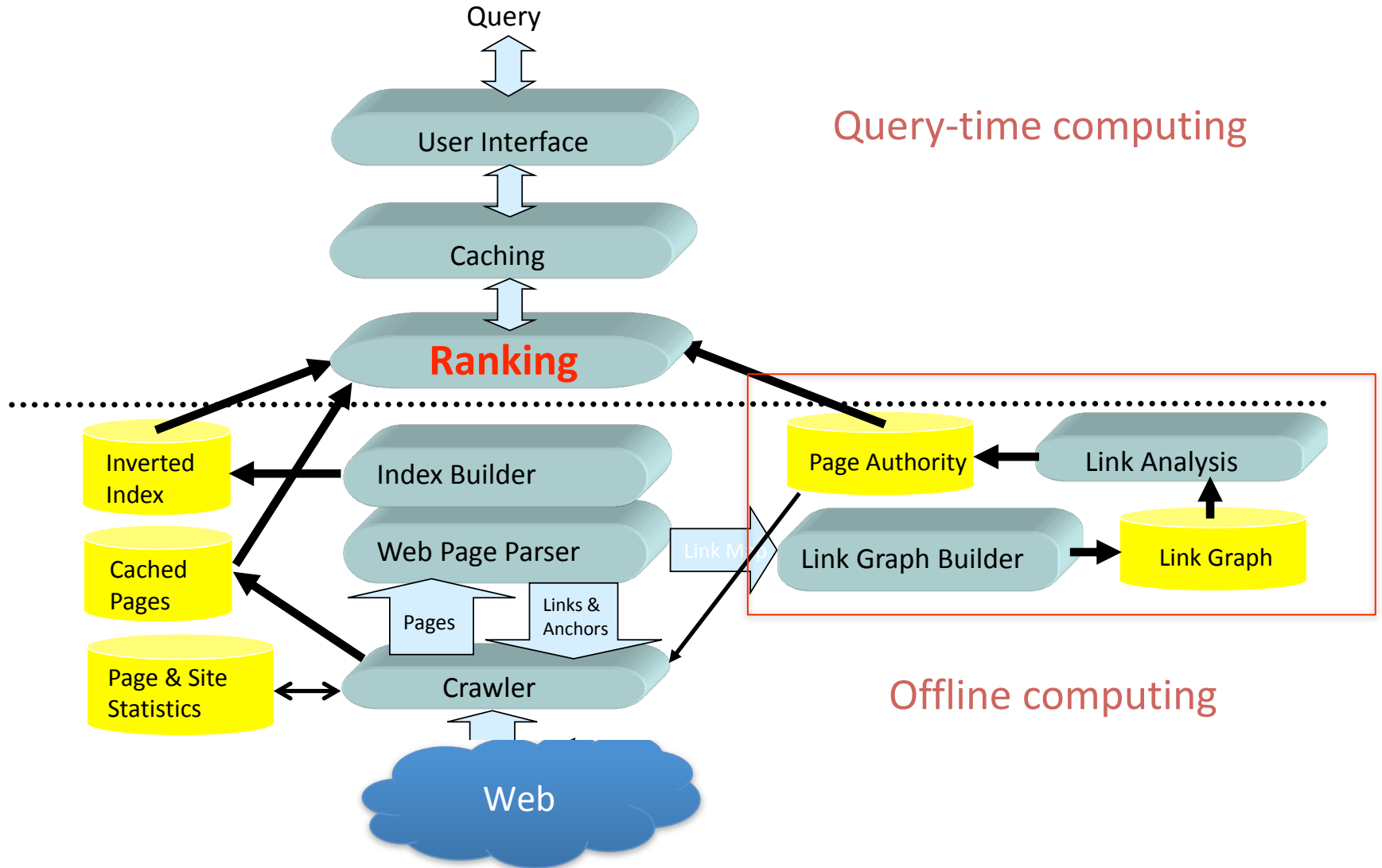
$$\text{Score}(Q, D) = \sum_{i=1}^n P(q_i | \theta_Q) * \log P(q_i | \theta_D)$$

Modèle de requête

Modèle de document

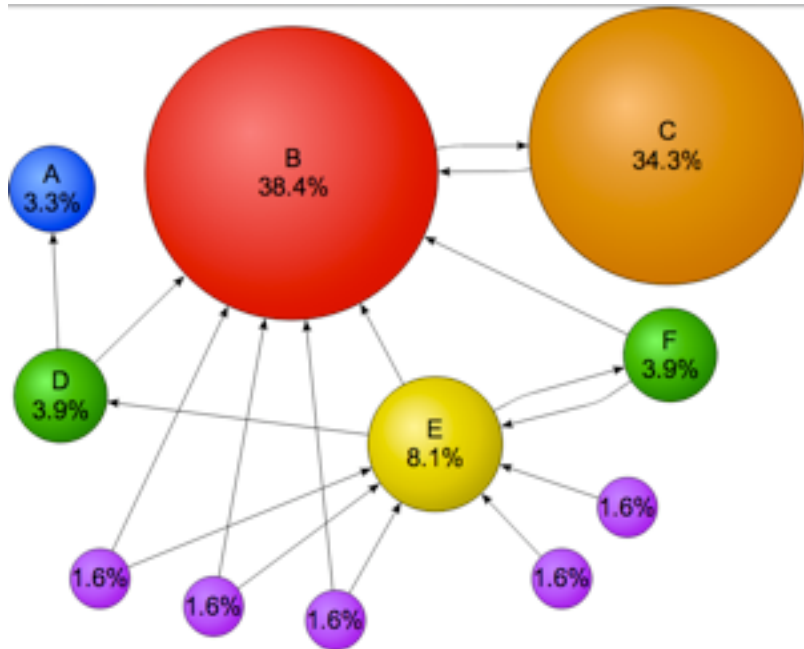
$$P(q_i | M_Q) = \frac{tf(q_i, Q)}{|Q|}$$

# 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



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

$d = 0.85$  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](#) ...

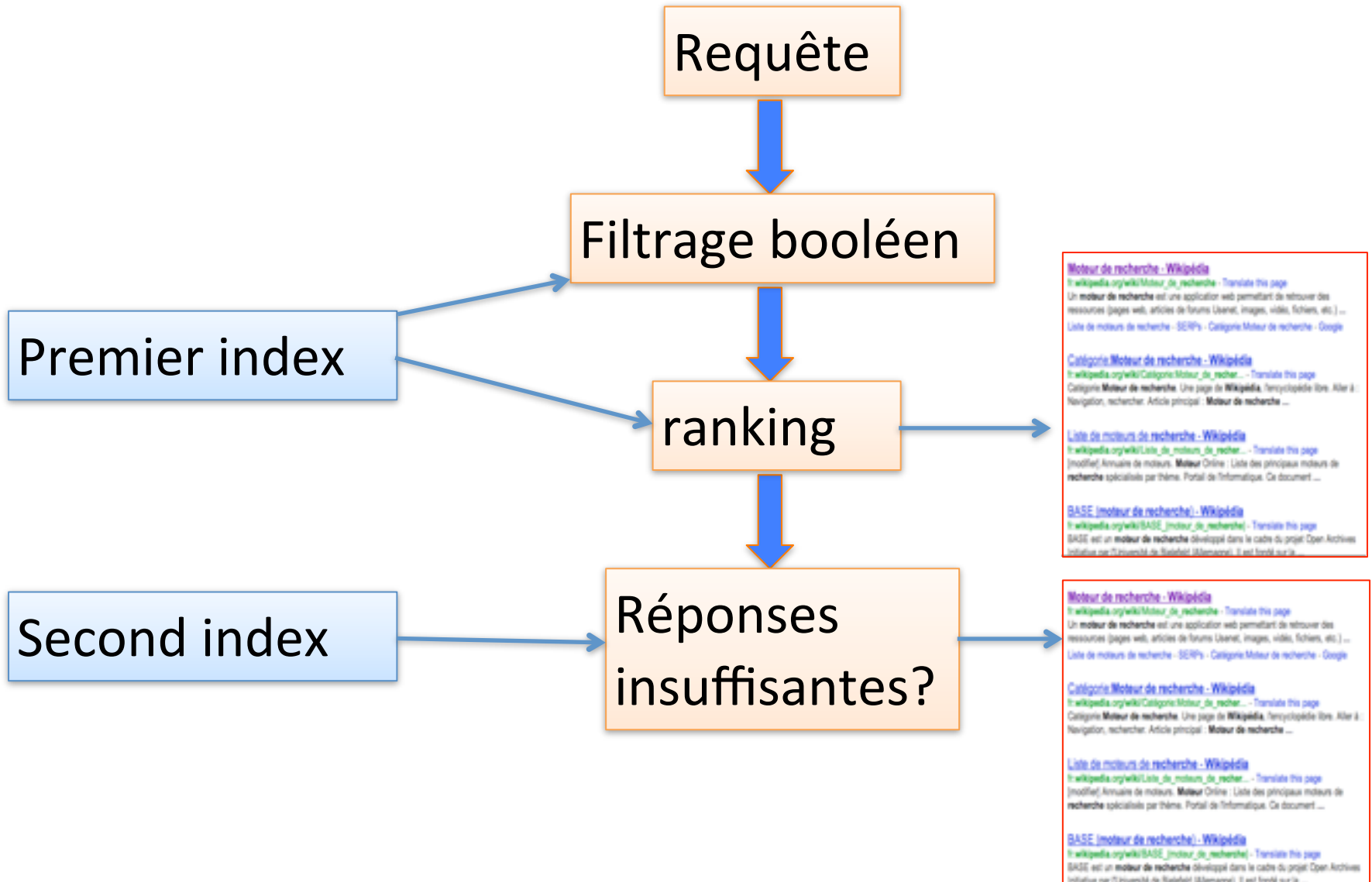
The article content includes a warning box: "Cet article ne cite pas suffisamment ses sources. Si vous disposez d'ouvrages ou d'articles de référence ou si vous connaissez des sites web de qualité traitant du thème abordé ici, merci de compléter l'article en donnant les références utiles à sa vérifiabilité et en les liant à la section « Notes et références »." Below this, the text defines a search engine: "Un **moteur de recherche** est une application web permettant de retrouver des ressources (pages web, articles de forums Usenet, images, vidéo, fichiers, etc.) associées à des mots quelconques. Certains sites web offrent un **moteur de recherche** comme principale fonctionnalité ; on appelle alors **moteur de recherche** le site lui-même (Google Video par exemple est un moteur de recherche vidéo). Instrument de recherche sur le web constitué de « robots », encore appelés bots, spiders, crawlers ou agents qui parcourent les sites à intervalles réguliers et de façon automatique (sans intervention humaine, ce qui les distingue des annuaires) pour découvrir de nouvelles adresses (URL). Ils suivent les liens hypertextes (qui relient les pages les unes aux autres) rencontrés sur chaque page atteinte. Chaque page identifiée est alors indexée dans une base de données, accessible ensuite par les internautes à partir de mots-clés. C'est par abus de langage qu'on appelle également moteurs de recherche des sites web proposant des annuaires de sites web : dans ce cas, ce sont des instruments de recherche élaborés par des personnes qui répertorient et classifient des sites web jugés dignes d'intérêt, et non des robots d'indexation — on peut citer par exemple *Voilà* et *Yahoo!*, etc."



# 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](https://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](http://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](http://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/](http://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/](http://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/](http://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](http://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/](http://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



# 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)

# Deuxième problème – synonymes ou mots reliés

- On peut décrire une chose de multiples façons

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

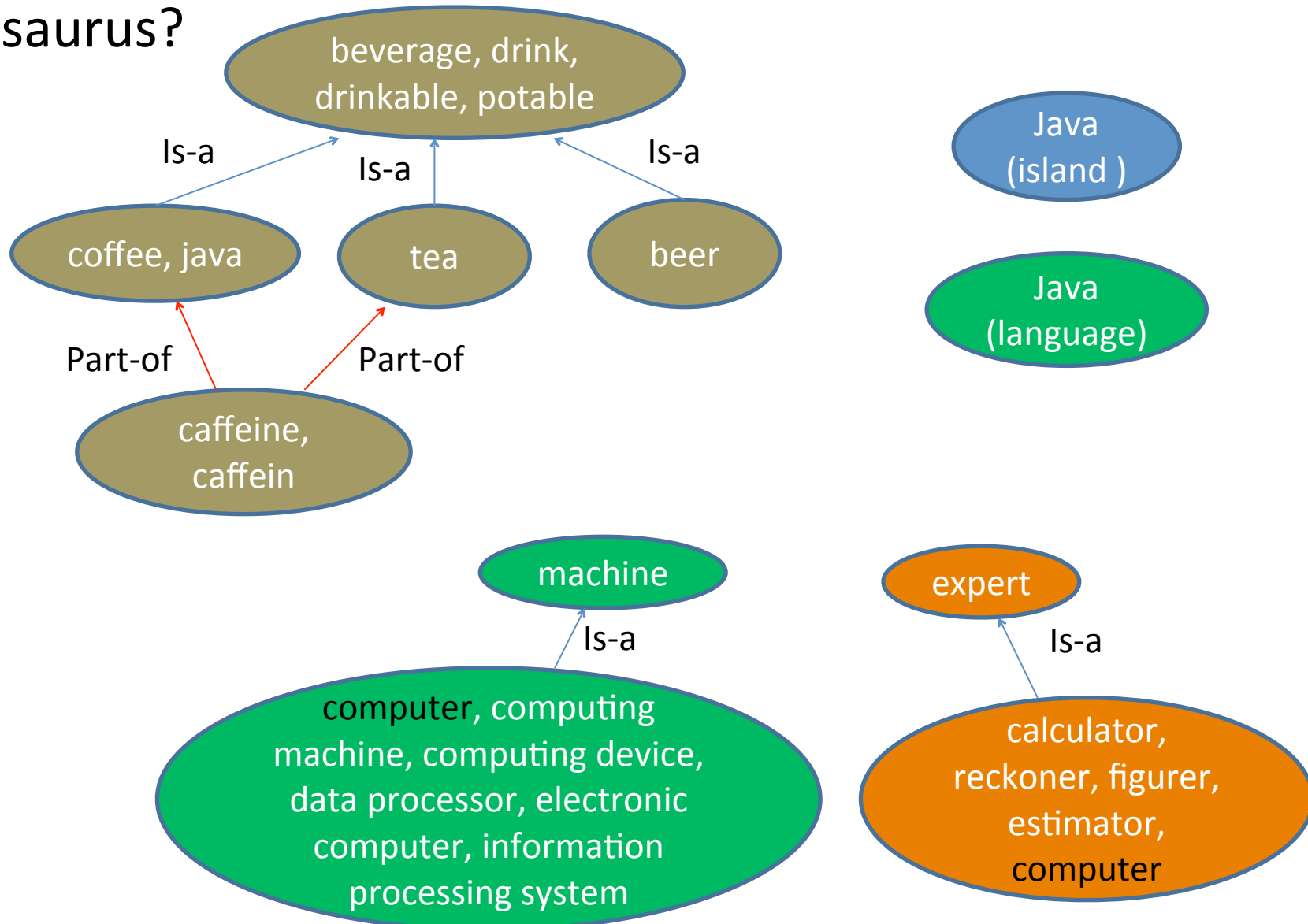


- Comment retrouver des documents pertinents avec des mots différents?
  - ➔ Expansion: ajouter des mots reliés
    - Dans le document: Expansion de document



# 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

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

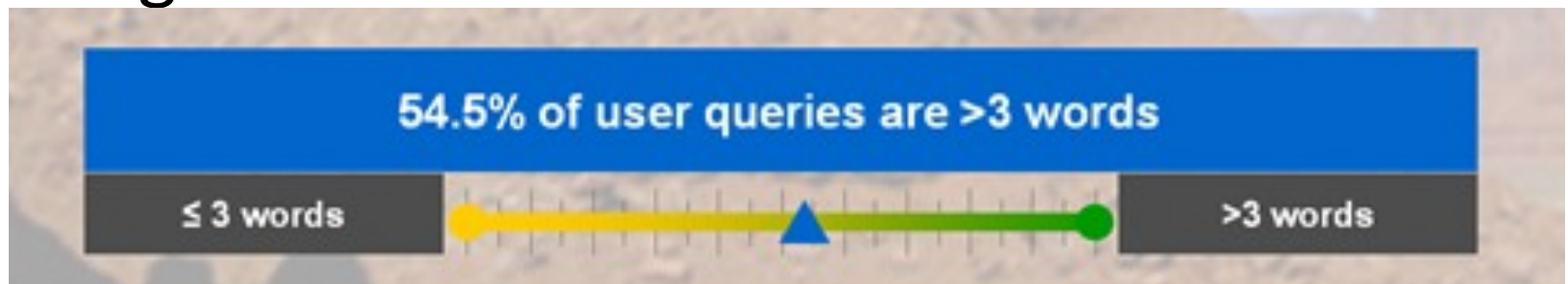
- Méthode prouvée utile
- Mais grande ambiguïté
  - Java → coffee
  - Java → île
  - Java → langage

# 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](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](http://www.furnitureontheWeb.com/noframe/products/p_et11nf.htm)

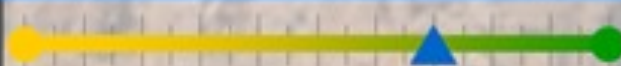
# Que cherchent les utilisateurs ?



- 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	britney 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	brittyny spears	21	britani spears	8	brittner spears

# Exploiter les logs pour connaître les

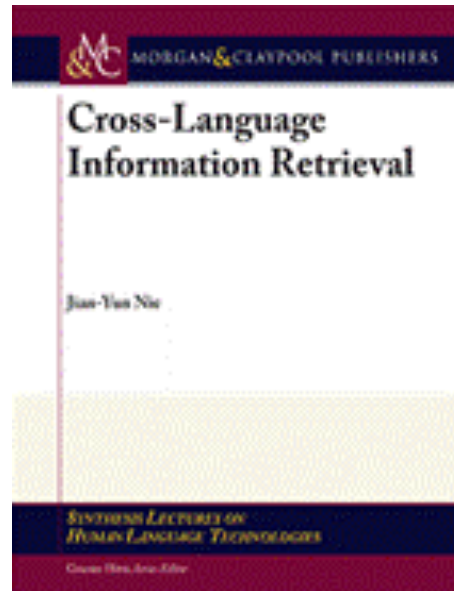
- Est-ce que 2 requêtes différentes cherchent la même chose?
  - Deux requêtes sont reliées si elles utilisent des mots identiques ou similaires
  - Deux requêtes sont reliées si elles ont amené à cliquer les mêmes documents (*co-click*)
- ➔ combiner les 2 critères pour estimer la similarité des requêtes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



# 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
  - ...

# Plus de détails sur la RI translinguistique



# Une question de taille

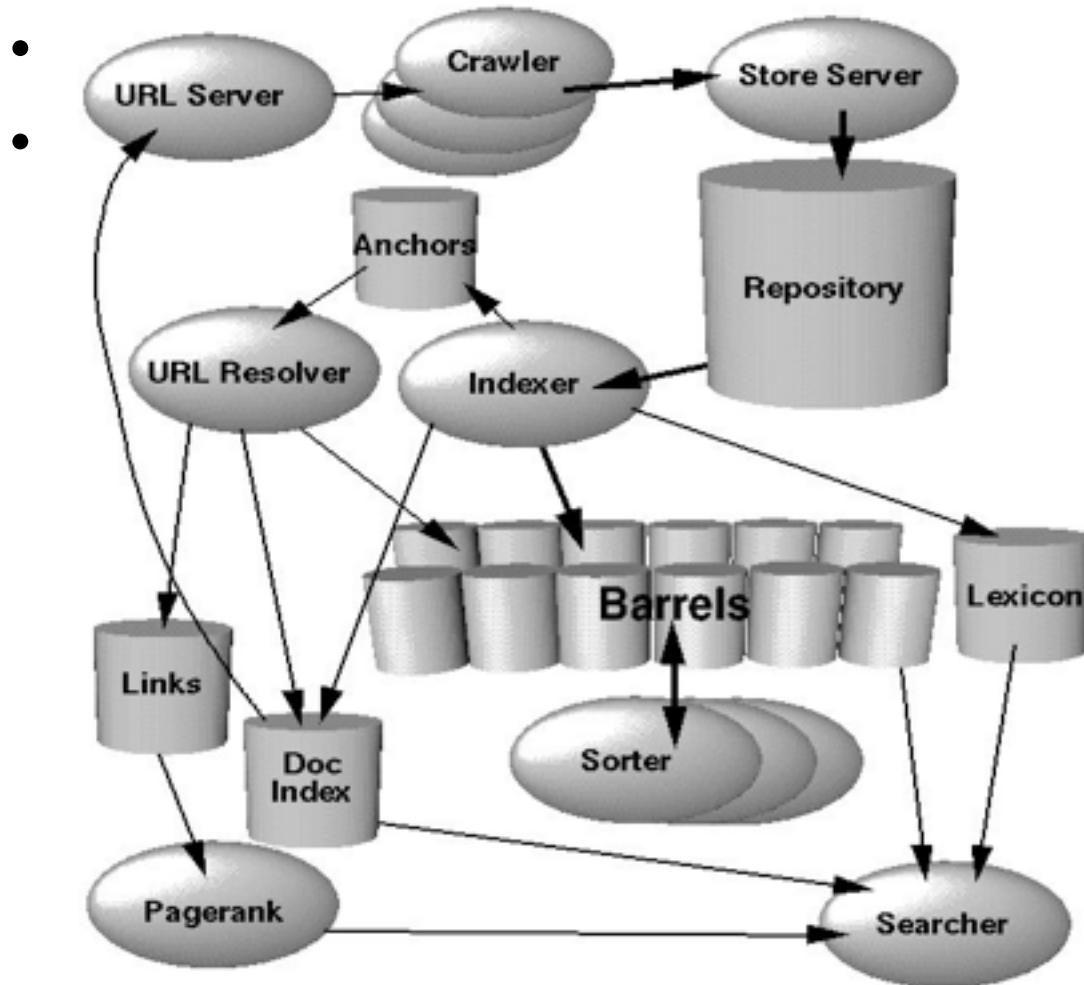
- Google
  - 1998: 25 million de pages ( $10^7$ )
  - 2000: 1 milliard ( $10^9$ )
  - 2008: 1 billion ( $10^{12}$ )
  - 2012: Index de 100 peta-octets de données ( $10^{15}$  ~  
½ des documents imprimés de toute l'humanité)

# 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)



# 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

- 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  
– Recherche  
– Produits  
– 96% de la valeur de Google
- Q2 2010  
– 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

L'avenir est devant nous!