Moteurs de recherche sur Internet

Confluence des télécommunications et des ordinateurs

Jacques Savoy
Institut d'informatique
Université de Neuchâtel

Internet et ses moteurs …

- Internet et le *Web*, c'est quoi? Comment ça marche?
- Moteurs de recherche (Google)
- Langues et thèmes populaires

Influence sur nos sociétés
Avec le risque de *surestimer* son impact à court terme et de sous-estimer son importance à long terme

Un peu d'histoire …

Télécommunication

1793 Invention du télégraphe optique (Claude Chappe)

Un peu d'histoire …

Télécommunication

1793 Invention du télégraphe optique (Claude Chappe)
1830 Monopole de l'Etat sur le réseau
1832 Le télégraphe électrique (Samuel Morse)
1840, dépôt du brevet
1844, Première ligne Washington-Baltimore

Un peu d'histoire …

Télécommunication

1793 Invention du télégraphe optique (Claude Chappe)
1830 Monopole de l'Etat sur le réseau
1832 Le télégraphe électrique (Samuel Morse)
1840, dépôt du brevet
1844, Première ligne Washington-Baltimore
1876 Graham Bell dépose le brevet du téléphone
1896 Marconi et la télégraphie sans fil (TSF)
1901 Première liaison transatlantique

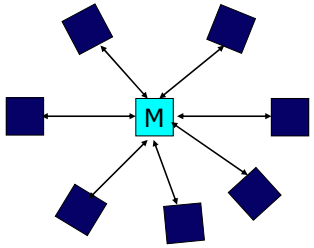
Un peu d'histoire …

Communications entre divers types d'ordinateurs

29 août 1949: Première bombe atomique soviétique
mil: Norad (couverture radar)
com: SABRE (réservations de siège)
1961: Faiblesse réseaux de communications

Un peu d'histoire …

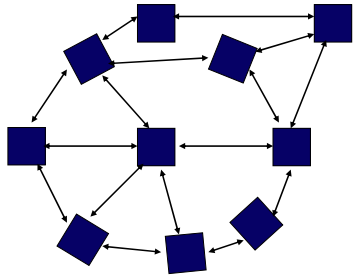
Pourquoi il ne faut pas un « grand maître »
→ Réseau décentralisé



Robustesse mais contrôle plus difficile

Un peu d'histoire …

Pas de « grand maître » (nous sommes tous égaux) dans un réseau sans structure préétablie



Un peu d'histoire …

Isolement des équipes / incapacité de partager les mêmes outils → on duplifie les efforts
"Aller sur la Lune" (rechercher des synergies)

1969: Début d'internet: interconnexion d'ordinateurs *hétérogènes* (UCLA, SRI, UCSB, Utah)
→ communication par paquets (≠ téléphone)
→ interface (traduction, standards – ISO)
1971: Courrier électronique (R. Tomlinson)
1973: Réseau *local* Ethernet (un bâtiment, PME/PMI)
Risque de prolifération des normes.

Un peu d'histoire …

La confrontation (communiquer entre ordinateurs)
A. les hommes d'affaires (téléphonie & IBM)
(organisation mondiale, spécification puis réalisation)
B. les militaires (DoD) + doctorants
(implémentation & code ouvert)

1975: Microsoft est fondé
Premier ordinateur personnel
1976: Apple voit le jour
France: Le fameux "22" à Anières → Transpac

Un peu d'histoire …

La confrontation (communiquer entre ordinateurs)
A. les hommes d'affaires (téléphonie & IBM)
B. les militaires (DoD) + doctorants

1981: Expérience du premier Minitel en France
Succès du Minitel dès 1984.
1983: Communication par paquets (TCP/IP) aux USA
Logiciel libre (Unix)
Pourquoi le minitel n'est pas sorti de France ?

Un peu d'histoire …

Et le WEB, WWW (World Wide Web)

1989: Tim-Bernes-Lee (CERN)
1990: Les URL: www.societe.ch
environ 200 sites
1993: MOSAIC un navigateur *simple*
une des clés du succès!
interface graphique
pointer-cliquer (retour arrière)

Netscape, Internet Explorer, Safari, FireFox

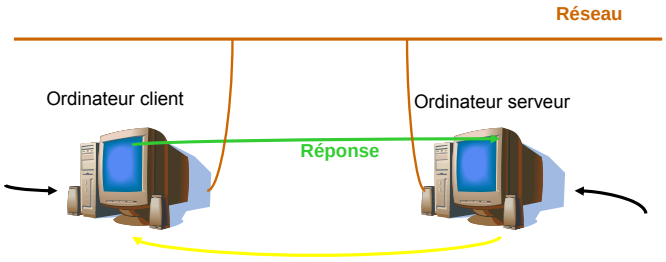
Pourquoi un tel succès ?

Raisons 1 & 2 pour les usagers
Raisons 3 & 4 pour les fournisseurs d'information

1. Simplicité d'emploi ("*retour arrière*")
2. Adressage : Comment spécifier n'importe quelle page / document sur n'importe quel ordinateur dans le monde? (URL)
   www.societe.ch/bienvenue.html
3. La simplicité des protocoles d'échange entre ordinateurs (HTTP)
   "question – réponse" ou "client-serveur"

Pourquoi un tel succès ?

Echange entre ordinateurs sur la base
"question / réponse" ou "client – serveur"



Pourquoi un tel succès ?

4. Spécifier la division logique d'une page / document sans se préoccuper sur quel ordinateur la page sera visualisée (HTML)

```
<html>
  <head>
    <title>My First HTML Page
  </head>
  <body>
    <p>Hello World!
  </body>
</html>
```

Et plus près de nous …

1994: Yahoo! (deux étudiants), annuaire du *Web*
Microsoft lance MSN
2 700 sites

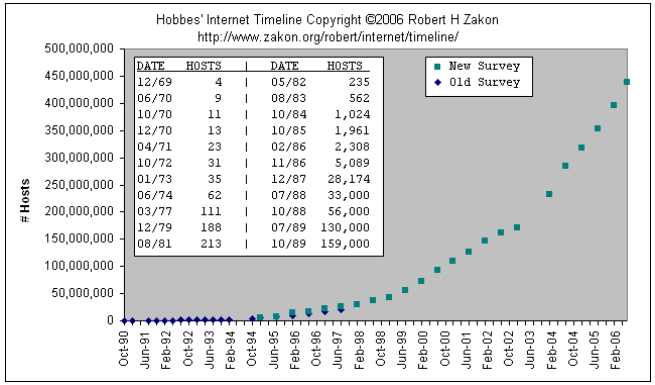
1995: AltaVista est lancé (Digital Computer, Dell)
23 500 sites

1998: Google est fondé (deux étudiants)
AOL rachète Netscape
2 000 000 sites

2007: 76 184 000 sites
Google gain 4,2 MM 5'600 personnes
UBS gain 9,4 MM 69'500 personnes

Internet

Progression très rapide (même avec la bulle spéculative)



Leçons de l'histoire

Succès des *start-up* ("jeunes pousses")

1. La décentralisation (pas de contrôle)
2. L'importance des normes (standards) marché mondial vs. niche
3. Partage des ressources / connaissance
4. La diffusion (gratuité) pour atteindre un marché mondial

Leçons de l'histoire

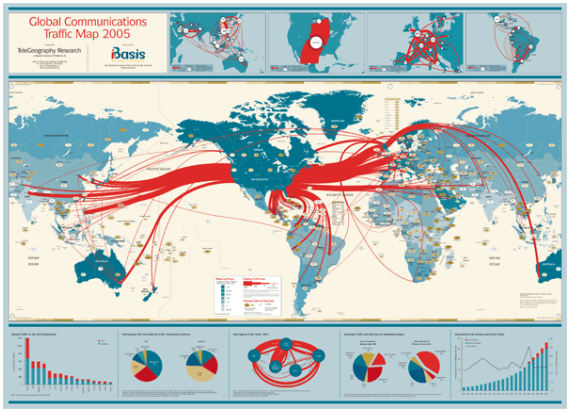
Leçons de l'histoire

Avenir

1. Diffusion du savoir
2. Préservation des données électroniques
3. Nouvelles directions (large corpus / données) pour les sciences
4. Dématérialisation (numérisation)

Internet

en 2005



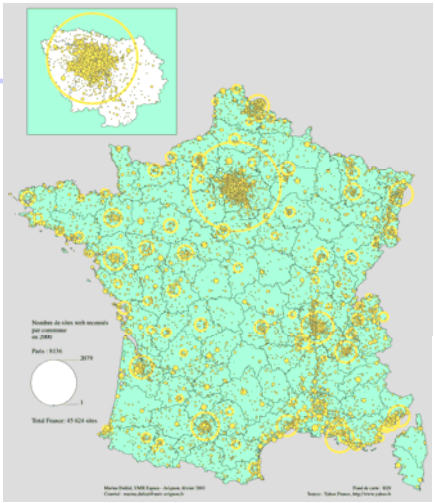
Internet

Réseau académique SWITCH

+ Swisscom
+ SunRise
+ CFF
+ …

Internet

Mais la répartition (sites recensés) n'est pas uniformément distribué sur le territoire



Les sites qui ont marqué …

Les moteurs de recherche
AltaVista, Yahoo, Google

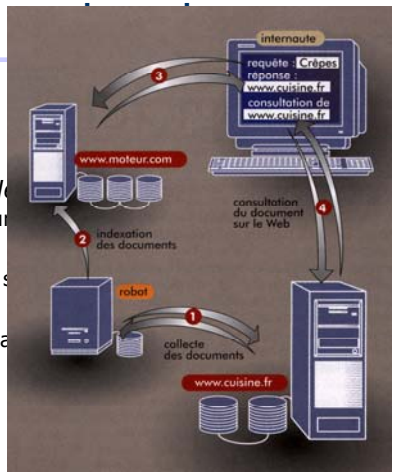
Les achats	Amazon.com
	Dell.com
	EasyJet.com
Enchères	eBay.com
Nouvelles	CNN.com
	SwissQuote.ch
Musique	Napster.com
& Vidéo	YouTube.com
	iTunes.com

Moteur de recherche

Moteur de

Trois composantes

1. L'aspirateur (*crawl*
   retrouver les pages su
2. L'indexeur
   représenter les pages
3. Le guichetier
   rechercher ce que l'usa
   de références



Moteur de recherche

L'aspirateur (*crawler*, *robot*)
connaître les sites
- annonce directe
- par les liens des autres sites
- mais sélection des sites
- visite de manière régulière les sites (différences entre *Le Monde et* UniNE)

Moteur de recherche

L'indexeur connaître le *vrai* contenu d'une page mais comment ?
- les mots présents sur la page avec une importance plus grande si :
  1. mots fréquents
  2. mots dans le titre / en gras
  3. mots peu fréquents dans les autres sites

Est-ce qu'une simple statistique sur les mots permet d'en prédire le sens ?

Moteur de recherche

6 x cubains
5 x nombre, floride, côtes
4 x réfugiés
3 x parvenus
2 x garde, atteint, année, pays
1 x utilisées, unis, années, économie, américaine, américains, tendance, embarcations, bateaux, indiqué, responsable, importante, dégradation, légalement, décédés, record, voyage, frêles, mer, illégalement, résidence, agit, cubaine, augmentation, titre, fuyant, fui, miami, jamais, furent, whitlock, embarquer, atteignant, bateau, exode, entraîné, remarqué

Moteur de recherche

ATS, 1er janvier 1994

Nombre record de réfugiés cubains parvenus en Floride en 1993.

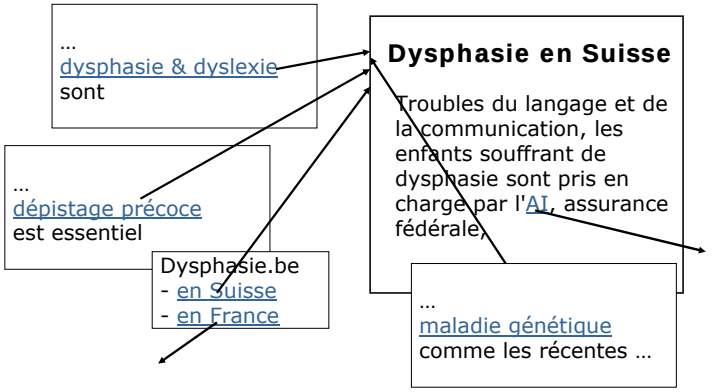
Miami, 1er jan (ats/afp) Plus de 3500 réfugiés cubains sont parvenus sur les côtes de Floride en 1993, un nombre jamais atteint depuis 1980, ont indiqué samedi les garde-côtes américains. L'année dernière, 3656 Cubains ont atteint les côtes de Floride en bateau, soit 43% de plus qu'en 1992, année durant laquelle ils furent au nombre de 2557, selon Chris Whitlock, un responsable des garde-côtes. Le nombre de réfugiés décédés durant le voyage n'est pas connu.
…

Moteur de recherche

L'indexeur connaître le *vrai* contenu d'une page mais comment ?

- les mots présents sur la page avec une importance plus grande si :
  1. mots fréquents
  2. mots dans le titre / en gras
  3. mots peu fréquents dans les autres sites
- les mots utilisés dans les hyperliens (depuis les autres pages vers la vôtre)

Moteur de recherche



Moteur de recherche

L'indexeur tiendra compte de la qualité des pages qui font référence à votre site.
Ainsi, une référence provenant du journal *Le Monde* aura plus d'impact qu'une référence venant du site de la "*Défense des castors*".

La qualité d'un site se mesure par la valeur *PageRank* de sa page d'accueil (valeur entre 0 et 10).

Moteur de recherche

Dans la page 782
mot retenu {France, Suisse, train, CFF, SNCF}

Dans l'index

mot	page	page	page	page	page	page	
France	34	345	543	567	**782**		
roi	12	34	64	567	678	987	999
Suisse	78	123	657	**782**	987	1034	
…							

Moteur de recherche

3. Le guichetier fait appel aux index pour trouver la réponse (temps de 0,5 sec.)
Estimation 3'000 à 5'000 PC en parallèle

Moteur de recherche

Requête {roi de France}

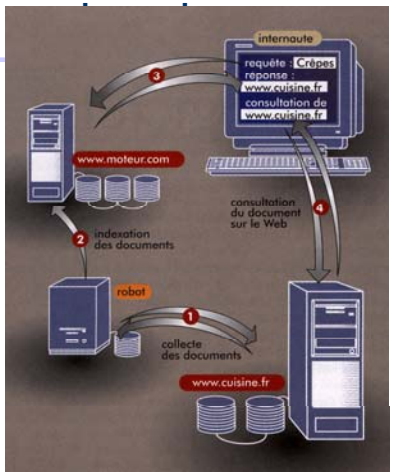
Dans l'index

mot	page	page	page	page	page	page	
France	**34**	345	543	**567**	782		
roi	12	**34**	64	**567**	678	987	999
…							

Réponse
page **567**
page **34**

Moteur de

Les moteurs sont utilisés par plus de 85 % des internautes pour dépister de l'information… mais ils nous révèlent d'autres facettes de nos sociétés …



Revenons sur *PageRank*

La valeur *PageRank* (Google) élevée si
1. beaucoup de sites pointent vers vous
2. des sites ayant un PageRank élevé pointent vers vous
   Un peu comme dans la vie réelle (être connu et avoir des recommandations de personnes célèbres)

PageRank

La valeur *PageRank* élevée
8 pour CreditSuisse.ch
7 pour swatch.ch, ubs.ch ou swiss.ch
6 pour nestle.ch, roche.ch.

Les administrations publiques
7 www.admin.ch, www.vd.ch
6 autres cantons

PageRank

7 pour LeTemps.ch, LeMatin.ch, Tribune de Genève, TSR, RSR
6 Agence Télégraphique Suisse, autres quotidiens

5 pour Xamax
4 pour Lausanne-Sport, de Gottéron ou de Genève Servette

Mais
9 avec Google.ch
10 avec Serono.ch

PageRank dans le monde

10 pour Google.com, Adobe.com, Apple.com

En France
8 la Bibliothèque nationale **de** France
7 le château **de** Versailles, l'Élysée, **La** Poste, **La** Tribune, **La** Recherche

Pour la Suisse francophone,
8 l'École polytechnique **de** Lausanne ou l'Université **de** Genève

PageRank dans le monde

1. .com
   PR = 6 www.novartis.**ch**
   PR = 8 www.norvatis.**com**
2. Chiffres d'affaires élevés
3. Marques connues (nescafé, chanel n° 5)
4. Firmes travaillant dans la haute technologie
5. Entreprises cotées au NASDAQ
6. Firmes américaines ?

Les requêtes populaires

1997 : Etats-Unis
« divertissements, loisirs » (20 %)
« sexe » (17 %)
« personne, lieux, chose » (7 %)
1999 : Etats-Unis
« commerce, voyage, emploi » (24,5 %)
« personne, lieux, chose » (20 %)
2002 : Etats-Unis
« personne, lieux, chose » (49 %)
« commerce, voyage, emploi » (12,5 %)

Les requêtes populaires

2006 : France (Yahoo.fr)
« Plus belle la vie »
« FFF »
« Shakira »
TV : « Smallville » (7e), « Lost » (10e), « Star Academy » (12e)
Sport : « PSG » (5e), « Zidane » (5e), « Ronaldinho » (6e), « AS Saint-Etienne » (8e), « Zidane Materazzi » (17e)
Pas de trace marquante : personnalités ?, CPE, le référendum sur la nouvelle constitution européenne, les Jeux olympiques d'hiver de Turin ou la grippe aviaire

Les requêtes populaires

2006 : Allemange
- « Wetter »
- « Routenplaner »
- « Erotik »
- « Telefonbuch »
- « chat »

2006 : Italie
- « Meteo »
- « Chat »
- « Oroscopo »
- « Giochi »
- « Tarocchi »

2006 : Angleterre
- « Heather Mills McCartney »
- « Pete Burns »
- « Big Brother »
- « The Ordinary Boys »
- « World Cup »

Les requêtes populaires

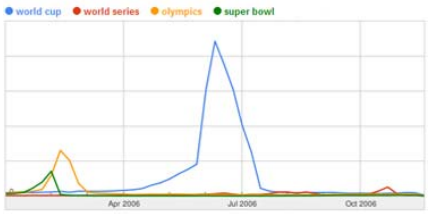
2006 : Canada
- « NHL »
- « Fifa/World cup »
- « American Idol »

2006 : Etats-Unis
- « Britney Spears »
- « WWE »
- « Shakira »

2007 : Etats-Unis – Avenir
- « MySpace »
- « YouTube »
- « iTunes »
- « Wikipedia »
- « orkut »

Les requêtes populaires

- La popularité de certains événements s'avère souvent passagère
- Mais dans certains cas, le phénomène se répète (comme la requête «Tour de France» en juillet)



Les langues sur le Web

Quelques faits (www.ethnologue.com)
- 6 800 langues dans le monde, dont
  - 2 197 en Asie
  - 2 092 en Afrique
  - 1 310 dans le Pacifique
  - 1 002 en Amérique
  - 230 en Europe.
- 600 d'entre elles sont écrites

Les langues sur le Web

80 % de la population mondiale parle
75 langues différentes
40 % de la population mondiale parle
8 langues différentes

75 langues sont parlées par + 10 M de personnes
20 langues sont parlées par + 50 M de personnes
8 langues sont parlées par + 100 M de personnes.

Les langues sur le Web

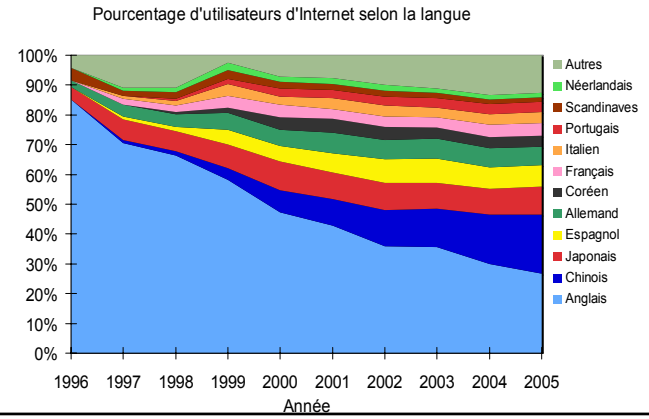
Identifiez ces langues !

1. Strč prst skrz krk
2. Mitä sinä teet?
3. Mam swoją książkę
4. Nem fáj a fogad?
5. Er du ikke en riktig nordmann?
6. Добре дошли в България!
7. Fortuna caeca est
8. نه ارسعيد
9. 정보검색시스템

Les langues sur le Web

Les langues sur le Web

Les langues sur le Web



Les langues sur le Web

1996 :	Anglais	80 %	et 47 millions
2005 :	Anglais	31,7 %	et 986 millions
	Chinois	16,5 %	
	Japonais	8,8 %	
	Espagnol	7,2 %	
	Allemand	6,4 %	
	Français	4,1 %	
	Coréen	3,6 %	
	Italien	3,6 %	

Les langues sur le Web

Mais les langues chinoise et japonaise utilisent des idéogrammes !

我不是中国人

人	homme / être humain
大	grand
大人	grand + homme = adulte
囚	prisonnier
国人	pays + homme = concitoyen

Les langues sur le Web

1.	der	de	di	the
2.	die	la	e	of
3.	und	le	il	to
4.	in	l	la	a
5.	den	les	che	and
6.	von	et	a	in
7.	das	des	un	s
8.	mit	d	per	that
9.	im	en	l	for
10.	zu	du	del	is

Les dix mots les plus fréquents

16 % de l'allemand ou l'italien

23,5% du français, 21,6% de l'anglais

Internet et ses moteurs …

- Internet et le Web, c'est quoi? Comment ça marche?
- Moteurs de recherche (Google)
- Langues et thèmes populaires

Et place aux questions !