Recherche d'Information sur le Web

Philippe Mulhem

Philippe.Mulhem@imag.fr

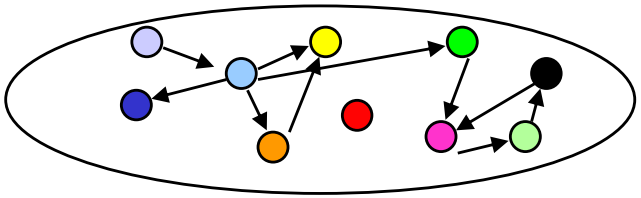
https://rimiashs.imag.fr

Plan

1. Introduction – Le web
2. Recherche de documents sur le Web par interrogation
3. Méta-moteurs
4. Conclusion

1. Introduction – Le Web

- Le Web : Un graphe



  - Pour la modélisation : un graphe
    - nœuds => les documents
    - arcs => les liens
  - Pour la présentation : un système hypertexte
    - affiche un nœud, avec les zones sensibles (ancres) qui sont les points de départ des liens vers d'autres documents
    - lors d'un clic swur un ancre, le système affiche le nœud qui est la cible du lien dont l'ancre est la source.
  - Avantages
    - Généralité : tout type de structure, tout types de liens
    - Grande souplesse et simplicité d'utilisation (des clics souris)
  - Inconvénients
    - Structure non explicite : perte de repères (séquentialité, hiérarchie)
    - Désorientation dans la navigation

1. Introduction – Le Web

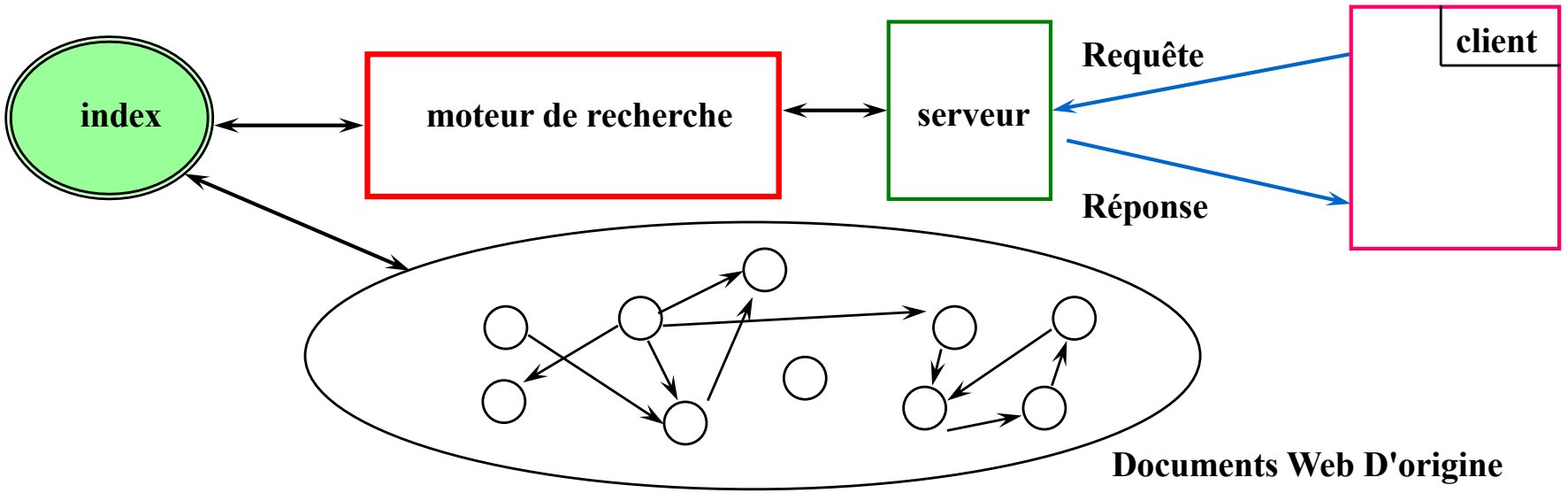
- Recherche de documents sur le Web
  - Constat
    - les sujets abordés quasi-illimités => on est presque certain de trouver des informations sur ce que l'on cherche.
  - Le problème n'est alors plus de savoir si un document qui nous intéresse existe, mais de pouvoir y accéder.

1. Introduction – Le Web

- Recherche de documents sur le Web
  - Pour accéder à des nœuds du graphe qui nous intéressent => recherche basée sur l'interrogation
  - Similaire à la Recherche d'Information classique :
    - Indexation des documents (pages Web)
    - Cette représentation des documents est utilisée comme base de la recherche par un utilisateur.

2. Recherche de documents sur le Web par interrogation

- Principe
  - celui des systèmes de recherche d'informations
    - on indexe les pages Web (analyse du contenu)
    - on stocke les index et les adresses (URL) des pages Web
    - lors d'un requête, on fait correspondre le résultat de l'analyse de la requête avec les représentations des documents.



2. Recherche de documents sur le Web par interrogation

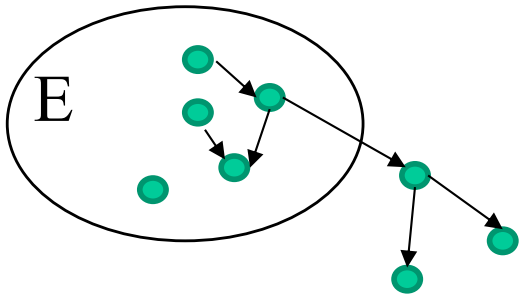
- Caractéristiques générales de l'approche
  - Indexation automatique robuste : hétérogénéité (langues, types de données)
  - Indexation multi-niveaux (contenu et structure)
    - Richesse des critères d'accès
  - Traitement d'un maximum de pages Web (couverture)
    - Grande capacité en stockage de masse (centaines de TeraOctets)
  - Rafraîchissement rapide des pages Web (Tous les 2/3 jours).
  - Temps de réponse très court (qualité de l'interaction)
    - Puissance de calcul : traitement d'un grand nombre de requêtes simultanées, et d'index très volumineux
    - Correspondance requête/document rapide à calculer (recherche de modèles de RI simples, voire simplification de modèles simples!)

exemple : Google : Modèle booléen pondéré++, Environ 55 milliards de pages, 5,6 milliards de requêtes par jour

2. Recherche de documents sur le Web par interrogation

- Indexation classique par "robots" (1)
  - Deux phases (spécifique au Web)
    - Découverte dynamique du corpus
    - Indexation des pages trouvées
  - Exemple: principe d'un robot utilisant un ensemble d'URL de départ E

tant que E non vide
accéder à une page p d'URL e de E
indexer le contenu de p
E = E ∪ cibles (p)



Raffinements éventuel sur :

» nb de pages prédéfini

» pages de catégories/contenu prédéfini (filtrage)

» pages nouvelles

» etc.

2. Recherche de documents sur le Web par interrogation

- Indexation classique par "robots" (2)
  - faire "le tour" du Web est long (estimation ancienne : un mois pour le robot Google)
  - l'ensemble de départ E est donné manuellement (seed pages)
  - Le niveau d'analyse du contenu de pages Web varie beaucoup
    - certains systèmes tronquent les mots (Lycos historiquement), d'autres non (Altavista, OpenText Index, Google)
    - certains systèmes ne gardent qu'une partie de l'index d'un document (Lycos stockait les 100 meilleurs termes pour une page Web) pour limiter le nombre de termes d'indexation

2. Recherche de documents sur le Web par interrogation

- Couverture estimée (documents indexés)
  - Selon http://www.worldwidewebsize.com/ (mars 2025)
    - Google : +/- [40, 60] milliards
    - Bing : +/- 4 milliards

2. Recherche de documents sur le Web par interrogation

- Indexation du contenu
  - Choix du modèle
    - Modèle de RI
  - Choix de la pondération
    - Basé sur la fréquence terme et documentaire
  - Choix du terme d'indexation, filtrage termes vides
    - Anti-dictionnaire par langue
    - Filtrage par la fréquence
  - Définition du vocabulaire
    - Troncature ou non
  - Divers
    - Position des mots
    - accents

2. Recherche de documents sur le Web par interrogation

- Sélection des termes d'indexation - Google
  - Pas d'anti-dictionnaire explicite, pas d'utilisation de fréquence documentaire
  - Fréquence documentaire dans le corpus, exemples :

	02/09	03/10	02/11	03/12	02/13	03/14	03/15	02/16	02/17	02/18
a ->	18M	12 M	23M	12M	25 M	13M	25M	25,3M	25,3M	25,3M
le ->	2M	2M	3M	25M	7,5M	2,2M	8M	7,8M	7,2M	8,0M
its ->	2M	2M	3,5M	6,7M	5,8M	1,3M	5,8M	5,1M	5,1M	5,3M
jour->	246K	200 K	300K	984K	862K	230K	830K	758K	681K	744K

	02/19	03/21	02/22	02/23	03/25
a ->	25,3M	25,3M	5,3M	25,3M	25,3M
le ->	15,5M	11,8M	3,3M	20,5M	14,9M
its ->	7,9M	5,7M	2,6M	15,8M	18,3M
jour->	1,7M	1,1M	2,0M	1,99M	2,5M

2. Recherche de documents sur le Web par interrogation

- Variantes du modèle booléen pondéré
  - La requête est une formule logique de descripteur avec les opérateurs ET, OU, NON
  - Les documents sont indexés par un ensemble de descripteurs
  - Correspondance :
    - l'opérateur ET est interprété comme la présence des deux termes dans le document, le OU comme la présence de l'un ou de l'autre (non exclusif) et le NON comme l'absence du terme dans le document.

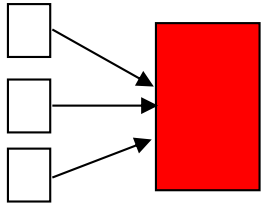
Google à proposé en premier la notion de popularité de page (PageRank) en 1998

2. Recherche de documents sur le Web par interrogation

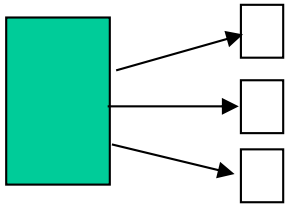
- Modèle booléen avec recherche plein texte
  - La requête est une formule logique de descripteur avec les opérateurs ET, OU, NON avec des opérateurs de proximité.
  - Les documents sont indexés par un ensemble de descripteurs avec la conservation de la distance entre les termes.
  - Correspondance
    - Les opérateurs logiques : modèle booléen avec adjacence de termes.
    - Selon l'opérateur utilisé, un document n'est sélectionné que si les deux termes sont à une certaine distance dans le document.

2. Recherche de documents sur le Web par interrogation

- HITS – Kleinberg 98
  - Objectif : Etant donnée une requête (topic), trouver
    - Les bonnes sources de contenu (autorités)

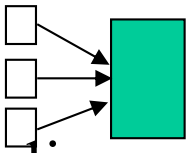


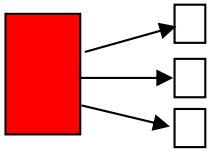
    - Les bonnes sources de liens (hubs)



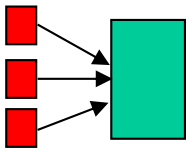
2. Recherche de documents sur le Web par interrogation

- HITS – Kleinberg 98
  - Intuitions
    - L'autorité vient des liens entrants
    - Le fait d'être un bon Hub vient des liens sortants

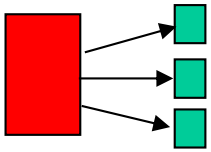




    - Une meilleure autorité vient de liens entrant de bon hubs



    - Un meilleur Hub vient des liens entrants de bonnes autorités



2. Recherche de documents sur le Web par interrogation

- HITS – Kleinberg 98
  - Initialisation (avec N pages) :

$$\sum_i AUTH[i]^2 = \sum_i HUB[i]^2 = 1 \quad \Rightarrow \text{pour une page V : } AUTH[V] = HUB[V] = \frac{1}{\sqrt{N}}$$

  - Calcul itératif jusqu'à convergence
    1. Pour chaque page V :

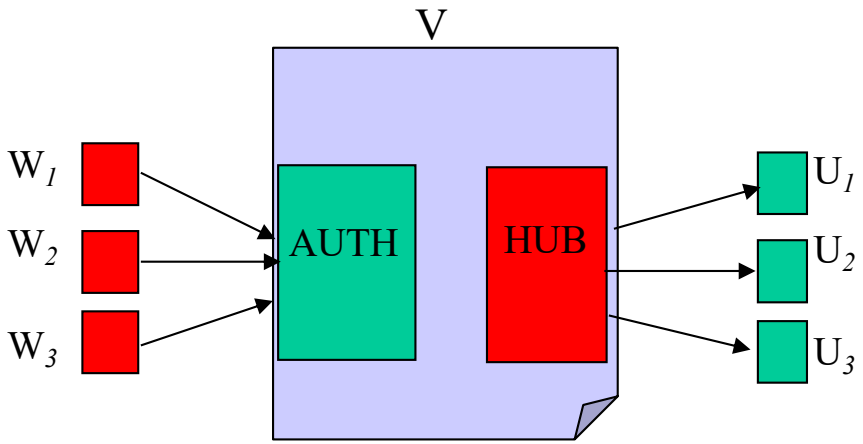
$$AUTH_{NN}[V] = \sum_{Lien(W_i,V)} HUB[W_i]$$

$$HUB_{NN}[V] = \sum_{Lien(V,U_i)} AUTH[U_i]$$

    2. Normaliser HUB et AUTH :

$$AUTH[V] = \frac{AUTH_{NN}[V]}{\sqrt{\sum_i AUTH_{NN}[i]^2}}$$

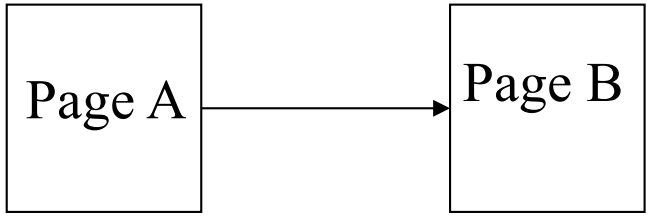
$$HUB[V] = \frac{HUB_{NN}[V]}{\sqrt{\sum_i HUB_{NN}[i]^2}}$$



2. Recherche de documents sur le Web par interrogation

- HITS – Exemple
  - Initialisation (étape 0) :

$$AUTH[A] = HUB[B] = \frac{1}{\sqrt{2}} = 0.71$$

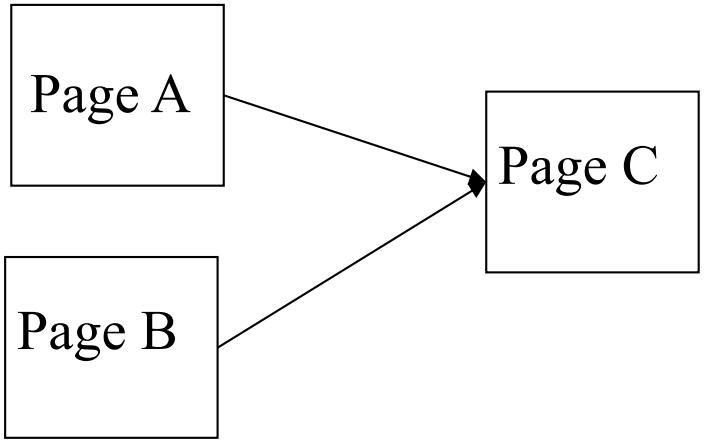


Lien(A,B)

	AUTH[A]	HUB[A]	AUTH[B]	HUB[B]	$\sqrt{\sum AUTH[U]^2}$	$\sqrt{\sum HUB[U]^2}$
0	0.71	0.71	0.71	0.71		
1.1 (NN)	0	0.71	0.71	0	0.71	0.71
1.2	0	1	1	0		
2.1 (NN)	0	1	1	0	1	1
2.2	0	1	1	0		

2. Recherche de documents sur le Web par interrogation

- HITS – Exemple
  - Calculs de HITS



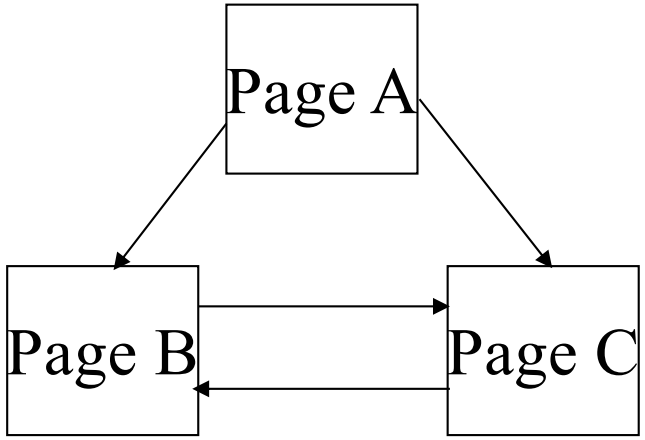
Lien(A,C)
Lien(B,C)

$$\frac{1}{\sqrt{3}} = 0.58$$

	AUTH[A]	HUB[A]	AUTH[B]	HUB[B]	AUTH[C]	HUB[C]	$\sqrt{\sum A[U]^2}$	$\sqrt{\sum H[U]^2}$
0	0.58	0.58	0.58	0.58	0.58	0.58		
1.1	0	0.58	0	0.58	1.15	0	1.15	0.82
1.2	0	0.71	0	0.71	1	0		
2.1	0	1	0	1	1.41	0	1.41	1.41
2.2	0	0.71	0	0.71	1	0		

2. Recherche de documents sur le Web par interrogation

- HITS – Exemple
  - Calculs de HITS



Lien(A,B)
Lien(A,C)
Lien(B,C)
Lien(C,B)

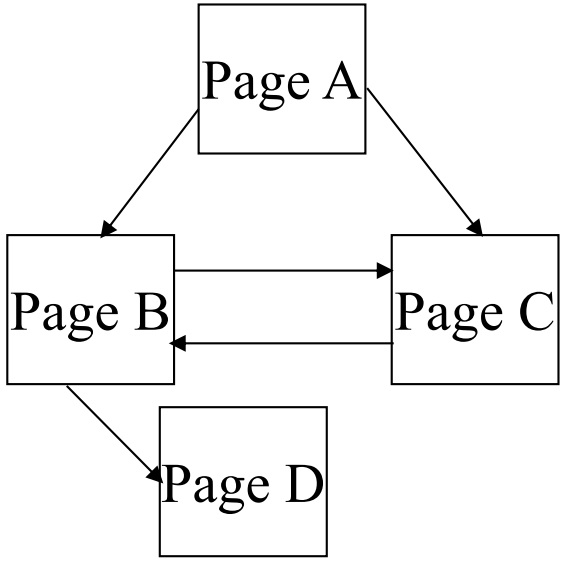
$\frac{1}{\sqrt{3}} = 0.58$

	AUTH[A]	HUB[A]	AUTH[B]	HUB[B]	AUTH[C]	HUB[C]
0	0.58	0.58	0.58	0.58	0.58	0.58
1.1	0	1,15	1,15	0,58	1,15	0,58
1.2	0	0,82	0,71	0,41	0,71	0,41
2.1	0	1,41	1,22	0,71	1,22	0,71
2.2	0	0,82	0,71	0,41	0,71	0,41

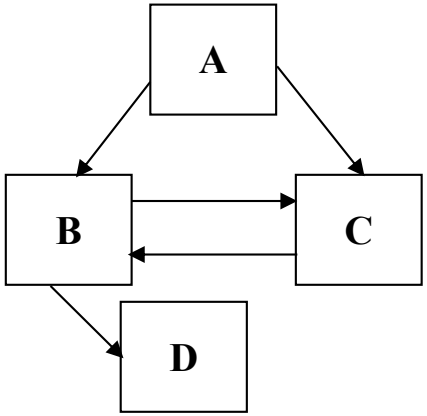
2. Recherche de documents sur le Web par interrogation

- HITS – Exemple
  - Calculs de HITS

Note : on s'arrête quand la moyenne des différences est inférieure au seuil 0.02.



	AUTH[A]	HUB[A]	AUTH[B]	HUB[B]	AUTH[C]	HUB[C]	AUTH[D]	HUB[D]
0								
1.1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
1.2								
2.1								
2.2								
…								



AUTH[A]	HUB[A]	AUTH[B]	HUB[B]	AUTH[C]	HUB[C]	AUTH[D]	HUB[D]
0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
0,00	1,00	1,00	1,00	1,00	0,50	0,50	0,00
0,00	0,67	0,67	0,67	0,67	0,33	0,33	0,00
0,00	1,33	1,00	1,00	1,33	0,67	0,67	0,00
0,00	0,74	0,56	0,56	0,74	0,37	0,37	0,00
0,00	1,30	1,11	1,11	1,30	0,56	0,56	0,00
0,00	0,72	0,62	0,62	0,72	0,31	0,31	0,00
0,00	1,34	1,03	1,03	1,34	0,62	0,62	0,00
0,00	0,74	0,57	0,57	0,74	0,34	0,34	0,00
0,00	1,32	1,09	1,09	1,32	0,57	0,57	0,00
0,00	0,73	0,60	0,60	0,73	0,32	0,32	0,00

2. Recherche de documents sur le Web par interrogation

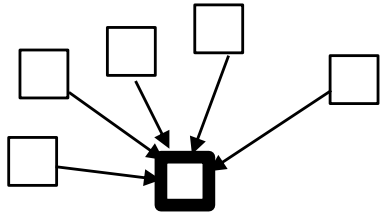
- HITS – Kleinberg 98
  - Conclusion
    - Nécessite un calcul pour chaque requête
    - Calcul long (même si convergence après 20 boucles pour 200 pages web)
    - Facile à « brouiller » (spam) pour augmenter artificiellement les valeurs d'autorité et de hub
      - Avec des liens générés automatiquement

2. Recherche de documents sur le Web par interrogation

- Google – L'approche Pagerank
  - Pagerank est une méthode utilisée déterminer la popularité des pages Web en dehors de toute requête
  - Le résultat de Pagerank est utilisé dans le calcul de correspondance avec une requête
  - Graphiquement :

Faible popularité

Popularité Moyenne



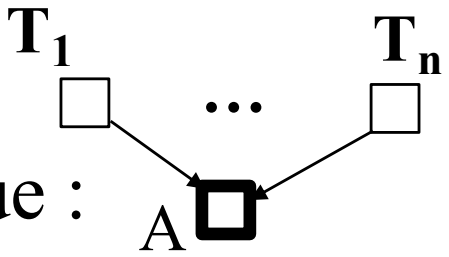
Grande popularité

2. Recherche de documents sur le Web par interrogation

- Google – L'approche Pagerank
  - Un vote, par toutes les autres pages du web, de la popularité d'une page web.
  - Un lien sur une page compte comme un vote « pour » la popularité de la page
  - Pas de lien se comporte comme une « abstention » sur la popularité de la page

2. Recherche de documents sur le Web par interrogation

- Google – L'approche Pagerank
  - Le « Pagerank » PR d'une page A telle que : est défini comme :



$$PR(A) = (1-d) + d * (\frac{PR(T_1)}{C(T_1)} + ... + \frac{PR(T_n)}{C(T_n)})$$

avec

- d : facteur d'équilibrage dans [0,1]
- T_i : les pages du web qui pointent sur A
- $C(T_i)$: le nombre de liens sortants de la page T_i

• PR va de (1-d) à $+\infty$

  - Problème : comment calculer PR(A)???

2. Recherche de documents sur le Web par interrogation

- Google – L'approche Pagerank
  - $PR(T_i)$: chaque page a une notion de sa propre importance
  - $C(T_i)$: chaque page disperse sont vote de manière égale à tous ses liens sortants
  - $PR(T_i)/ C(T_i)$: la valeur de vote de T_i à A
  - d(…) : tous les votes sont ajoutés, mais pour éviter que les autres pages aient trop d'influence on équilibre par d (d=0.85 couramment)
  - (1-d) : importance minimale d'importance d'une page, même si elle n'a pas de lien entrant. En fait indique que même si une page n'est pas la cible d'un lien on peut y accéder directement par son URL.

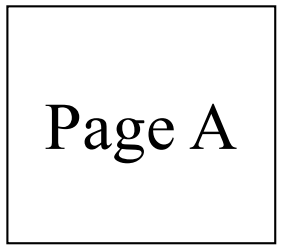
2. Recherche de documents sur le Web par interrogation

- Google – Calcul de Pagerank
  - Pour calculer PR(A) on doit connaître PR(T_1), mais si A pointe sur T_1 alors il faut connaître PR(A), etc.
  - Comment faire?

→ On itère avec des valeurs initiales fixées pour les PR, par exemple 1.

2. Recherche de documents sur le Web par interrogation

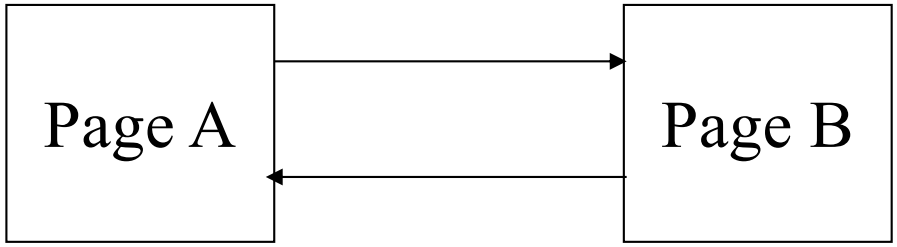
- Google – Calculs de Pagerank
  - Exemple



  - Avec d=0.85 et initialement PR(A) = 1
  - PR(A)=(1-d) = 0.15

2. Recherche de documents sur le Web par interrogation

- Google – Calculs de Pagerank
  - Exemple



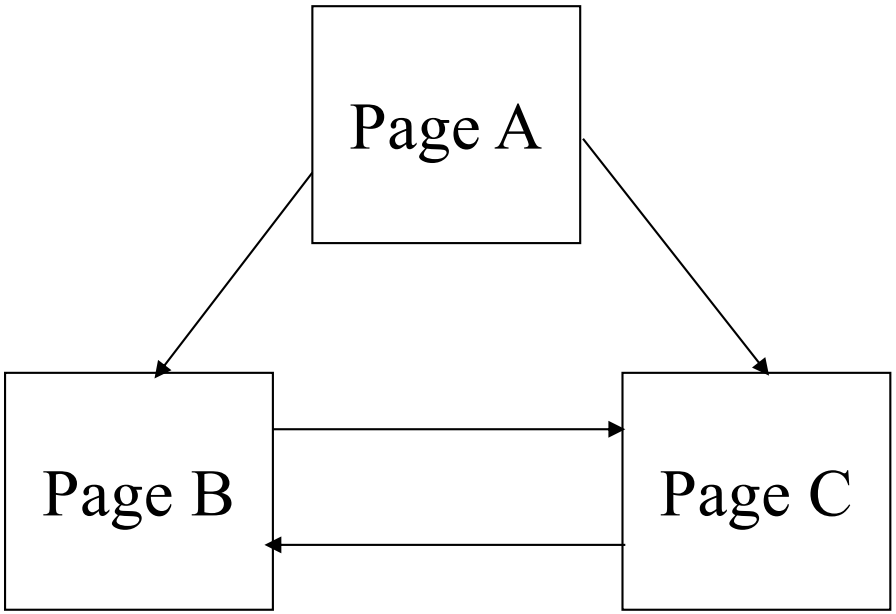
  - Avec d=0.85 et initialement PR(A) = 1 et PR(B)=1

  - PR(A)=(1-d)+d(PR(B)/1) = 0.15+0.85*1 =1
  - PR(B)=(1-d)+d(PR(A)/1) = 0.15+0.85*1 =1

  - Les valeurs ne changent pas, les valeurs initiales étaient correctes!

2. Recherche de documents sur le Web par interrogation

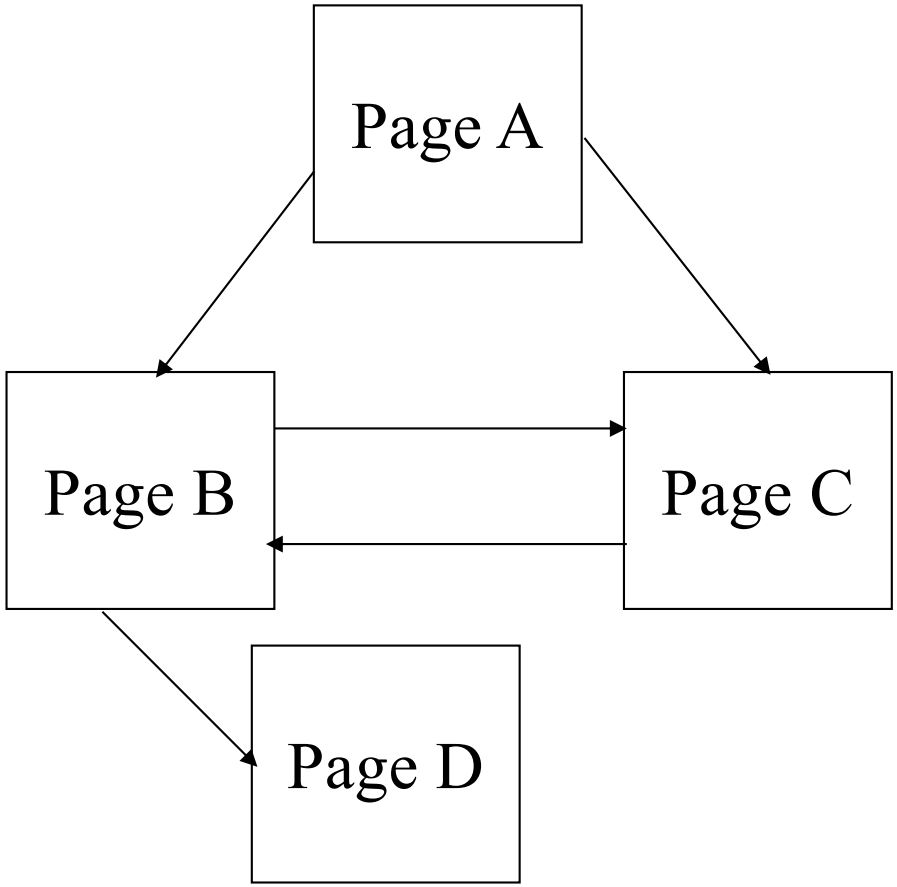
- Google – Calculs de Pagerank
  - Avec d=0.85 et initialement PR(A) = 1, PR(B) = 1 et PR(C)=1



	PR(A)	PR(B)	PR(C)
1	1	1	1
2	0,15	1,425	1,425
3	0,15	1,425	1,425

2. Recherche de documents sur le Web par interrogation

- Google – Calculs de Pagerank

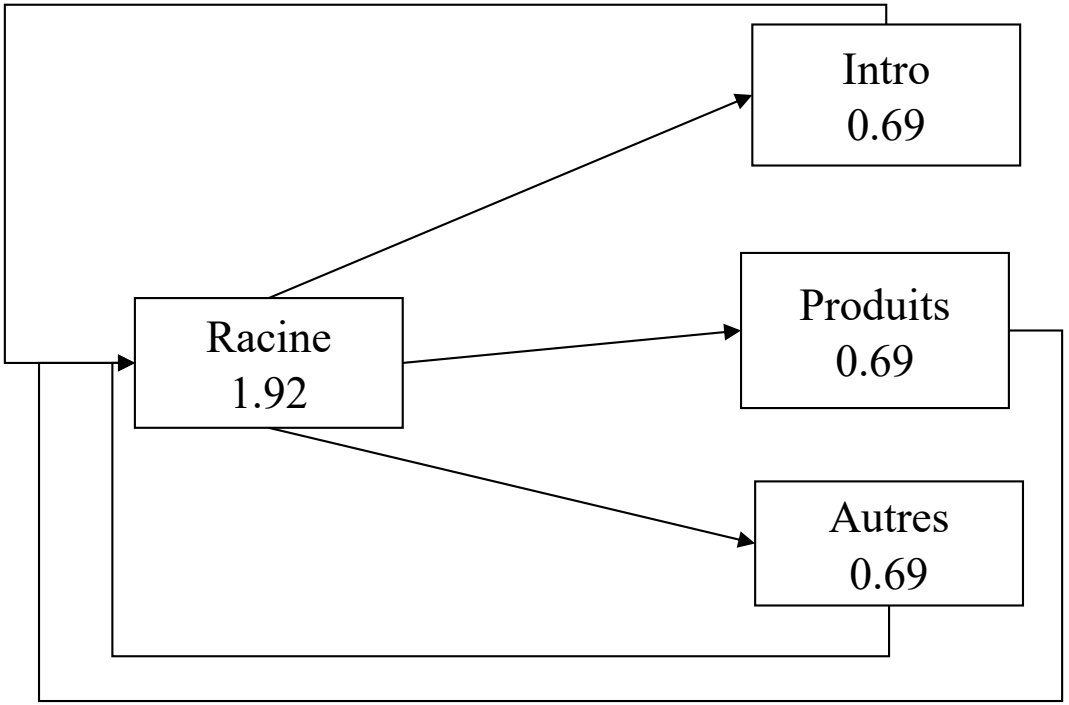


	PR(A)	PR(B)	PR(C)	PR(D)
1	1	1	1	1
2	0.15	1.425	1	0.575
3	0,15	1,06	0,82	0,76
4	0,15	0,91	0,67	0,60
5	0,15	0,78	0,60	0,54
6	0,15	0,72	0,55	0,48
7	0,15	0,68	0,52	0,46
8	0,15	0,66	0,50	0,44
9				

Note : on s'arrête quand la moyenne des différences est inférieure au seuil 0.02.

2. Recherche de documents sur le Web par interrogation

- Google – Calculs de Pagerank
  - Exemple (www.iprcom.com/papers/pagerank plus actif)
    - Hiérarchie simple avec retour



2. Recherche de documents sur le Web par interrogation

- Google – Calculs de Pagerank
  - Pas de données publiques sur l'intégration avec les pondérations des termes lors de la correspondance (*Learning to Rank*)

2. Recherche de documents sur le Web par interrogation

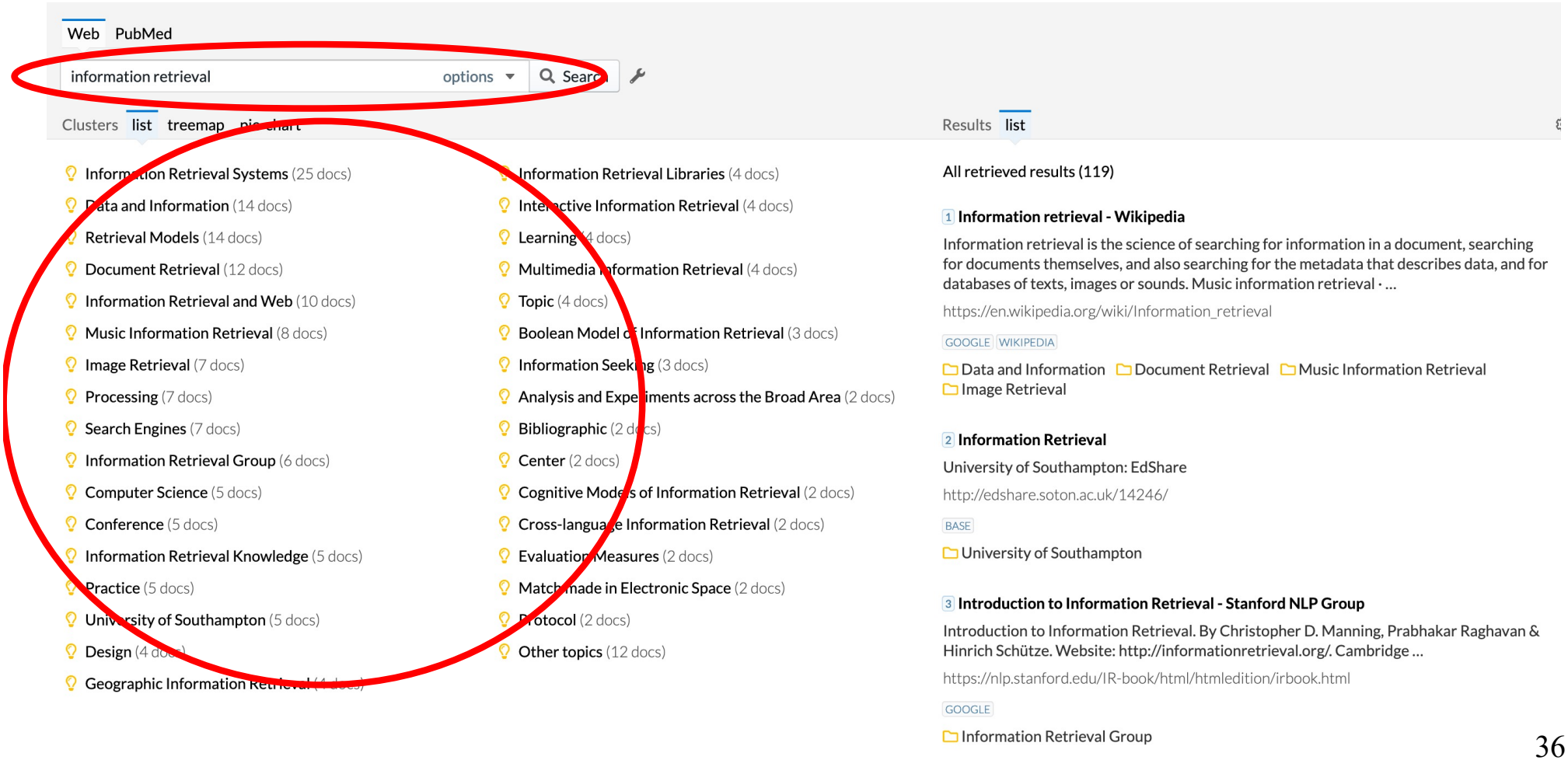
- Google – Calculs de Pagerank
  - Conclusion
    + Principe très intéressant et qui a prouvé son utilité
    + Difficile à spammer
    - Complexité de calcul sur les milliards de pages du web
    - Défavorise la nouveauté
    - Pas de liens typés : quels sens donner aux liens??

2. Recherche de documents sur le Web par interrogation

- Google – Calculs de Pagerank
  - Intégration avec recherche d'information, par exemple en additionnant le score cosinus et PageRank (cf. exercice)

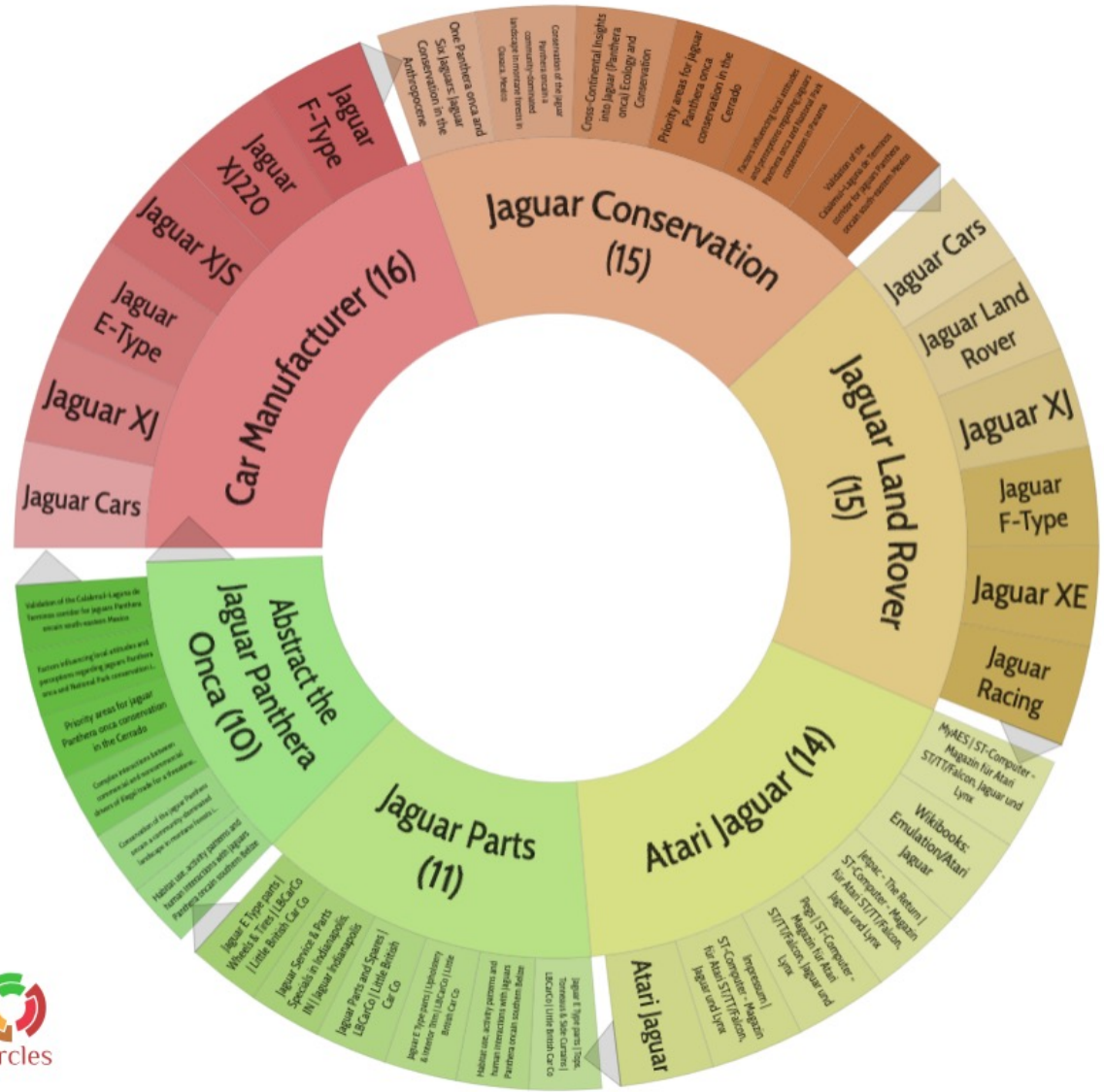
2. Recherche de documents sur le Web par interrogation

- Présentation des résultats
- https://search.carrot2.org regroupe les éléments de réponse en « clusters »



2. Recherche de documents sur le Web par interrogation

- https://search.carrot2.org, graphique pour « jaguar »



4. Conclusion

- On a étudié les éléments liés à la recherche d'information sur le Web
  - Fonctionnement
  - Hits/Pagerank