CHAPITRE IV : INDEXATION DANS LES MOTEURS DE RECHERCHES

4.1 INTRODUCTION :

Nous avons déjà expliqué qu'un moteur de recherche est d'abord un outil d'indexation, c'est-à-dire qu'il dispose d'une technologie de collecte de documents à distance sur les sites Web, via un outil que l'on appelle robot ou bot. Un robot d'indexation dispose de sa propre signature (comme chaque navigateur web).

L'indexation des ressources récupérées consiste à extraire les mots considérés comme significatifs du corpus à explorer. Les mots extraits sont enregistrés dans une base de données organisée comme un gigantesque dictionnaire ou, plus exactement, comme l'index terminologique d'un ouvrage, qui permet de retrouver rapidement dans quel chapitre de l'ouvrage se situe un terme significatif donné. Les termes non significatifs s'appellent des mots vides. Les termes significatifs sont associés à une valeur de poids. Ce poids correspond à une probabilité d'apparition du mot dans un document. Cette probabilité est indiquée sous la forme d'une "fréquence de terme".

4.2 LES ENJEUX D'UNE BONNE INDEXATION :

La taille de la toile ne cesse d'augmenter. Cette croissance touche à la fois le nombre de pages, mais aussi la taille moyenne de chaque page. Il est donc indispensable d'optimiser les crawlers pour qu'ils puissent indexer la Toile de manière efficace. Cette efficacité se mesure en fonction de trois facteurs, notamment :

1. L'indexation doit être **rapide** : c'est la condition essentielle pour qu'un robot puisse passer souvent et assurer que les pages figurant dans l'index aient une "fraîcheur suffisante"

2. L'indexation doit être **complète** possible : on sait qu'elle ne peut pas être exhaustive, car certaines portions du web ne sont pas reliées entre elles par des liens. Mais le robot doit être capable d'indexer une portion significative de la Toile.

4.3 LES METHODES TRADITIONNELLES DE CRAWL :

La méthode de crawl la plus communément utilisée (jusqu'à 2003) était l'"indexation par lot" ("batch crawling" en anglais. Elle se déroule en trois étapes :

1. D'abord, on détermine une série d'urls de départ à crawler (les " seed url "). La taille de ce fichier d'urls et la manière de les choisir a d'ailleurs une influence certaine sur le résultat final. Ensuite, on lance les robots d'indexation qui vont "aspirer" les pages, tout en récupérant les liens qu'elles contiennent vers d'autres pages.
2. Ces nouvelles url sont ajoutées à une "file d'attente" des liens qui restent à explorer. Lorsque cette file d'attente ne contient plus de nouvelles url, le processus s'arrête. Le crawl est terminé. Et l'on considère que la Toile entière a été indexée.

3. Les urls ainsi recueillies pendant le crawl serviront par ailleurs d'urls de départ au prochain cycle de crawl...

Critique :

L'indexation par lots possède des inconvénients importants , dès que l'on veut développer un moteur de recherche "grand public" comportant un très grand nombre de pages. D'abord les crawlers indexent l'ensemble des pages à chaque cycle, y compris celles qui ne changent jamais. Le processus est donc long, il dure dans la pratique plusieurs jours, jusqu'à une semaine complète. Or, c'est seulement à la fin du processus que l'on déclarera l'index "complet" . Le problème, c'est que certaines des pages sont déjà obsolètes à la fin du processus. D'autres auront disparu. De nouvelles pages auront été mises en ligne, et ne figurent pas dans l'index.

Ce n'est donc pas un moyen très efficace de gérer le problème de la fraîcheur des pages. Google, qui lançait un "full crawl" par mois, se retrouvait donc avec un index dont la fraîcheur posait problème. La solution adoptée pour compenser le problème consistait à lancer un robot différent "le fresh crawler" destiné à détecter de nouvelles pages apparues entre deux "full crawl".

De nombreux chercheurs ont donc réfléchi à des solutions alternatives ou à des perfectionnements de la méthode. La plupart des travaux ont abouti à la création d'agents d'indexation spécialisés, extrêmement efficaces pour "aspirer" et "surveiller" des portions limitées du web.

4.4 LE CALCUL DU PAGERANK :

(source : "The Google Pagerank Algorithm and How It Works", Ian Rogers) :

Le **PageRank** (en abréviation PR) est une formule mathématique utilisée par certains moteurs de recherches (dont Google) pour déterminer l'importance d'une page Web.

Elle se base sur un concept très simple : un lien émis par une page A vers une page B est assimilé à un « vote » de A pour B. C'est-à-dire que plus une page reçoit de « votes », plus cette page est considérée comme importante par Google, exactement comme le principe des élections.

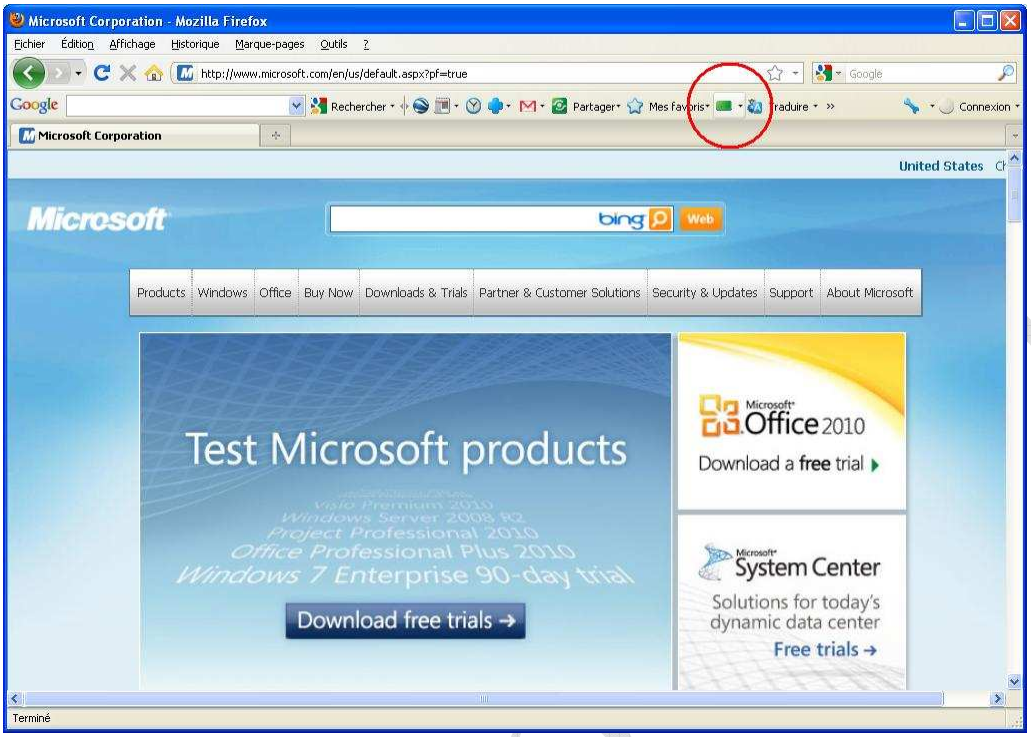
Toutefois, il faut préciser que pour le concept du PageRank les pages « électrices » n'ont pas toutes le même pouvoir de « vote ».. En effet, un vote émis par la page d'accueil d'un site majeur, comme Microsoft ou AFP, pèse beaucoup plus lourd qu'un vote émis par la page personnelle d'un internaute.

Remarque Importante : Le PageRank est une mesure de l'importance d'une page, et non d'un site entier. On entend souvent parler de « site de rang n », il s'agit d'un abus de langage décrivant le rang de la page d'accueil du site.

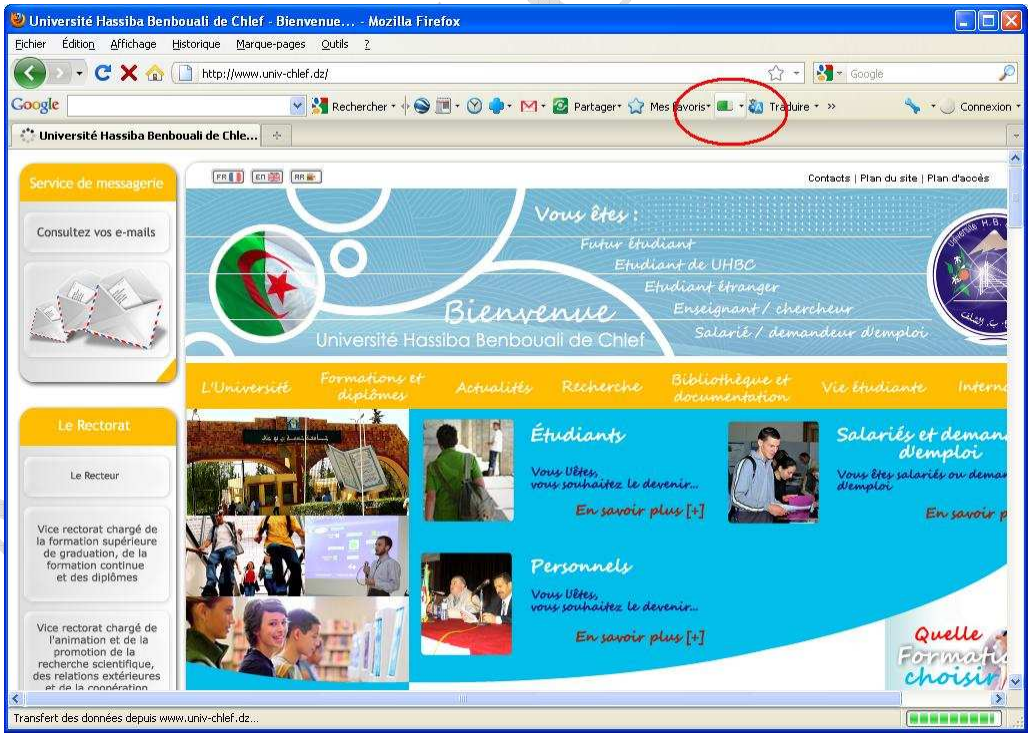
De même, l'**importance** d'une page est sans rapport aucun avec l'intérêt ou la pertinence de celle-ci, ces deux dernières notions étant totalement absentes de l'algorithme du PageRank. Elles interviennent néanmoins dans les pages de résultat de recherche.

Le Range Rank d'une page peut-être affiché sur certains moteurs de recherches (exemple : Page d'accueil de Microsoft : 9/10, page d'accueil de l'Université de Chlef : 6/10).

Page d'accueil de Microsoft affichant un PR de 9



Page d'accueil de l'Université de Chlef affichant un PR de 6



Le Page Rank d'une page A est défini par la formule suivante :

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)) \quad (1)$$

où :

- d est un paramètre d'ajustement (nombre compris entre 0 et 1.
- Les Ti sont les pages votantes pour la page A
- C(Ti) est le nombre de liens émis par la page Ti.

Comme le calcul réel d'un PageRank peut prendre des nombre très grands, on représente souvent le PR sur une échelle logarithmique (allant de 0 à 10), comme cela est affiché par Google-toolbar. Le tableau suivant donne quelques valeurs de correspondance entre la valeur PR en échelle logarithmique et la valeur réelle :

PR affiché (échelle logarithmique, à base 10 par exemple)	**PR réel**
PR = 0	0 <= PR <1
PR = 1	1 <= PR <10
PR = 2	10 <= PR <100
PR = 3	100 <= PR <1000
PR = 4	1000 <= PR <10000
…	…

On donc qu'à chaque niveau, le PR est 10 fois plus élevé que le niveau précédent.

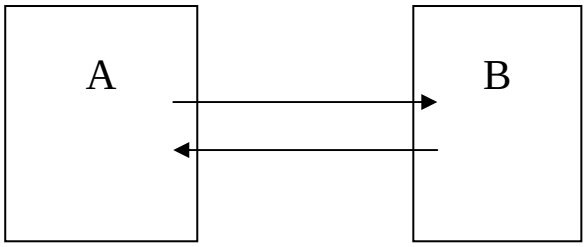
NB : Le calcul du PR d'une page peut évoluer (diminuer ou augmenter) au fur et à mesure que le moteur de recherches actualise son index. Cette évolution expliquerait pourquoi certaines pages voient leur PageRank diminuer au fil des indexations, alors que le nombre de liens entrant reste inchangé.

Calcul du PageRank :

Le calcul du PR (voir formule 1) est récursif : c'est-à-dire que le PageRank d'une page A dépend du PageRank des pages T1...Tn qui émettent un lien vers A, et ne peut donc pas être déterminé sans connaître le PR de ces dernières, et de toutes celles qui émettent un lien vers elles, et ainsi de suite...

Pour rendre ce calcul faisable, Google se base sur le principe suivant : Le PageRank peut être calculé en utilisant un simple algorithme itératif, et correspond au vecteur propre principal de la matrice normalisée des liens du Web.

Exemple : Deux pages A et B, qui pointe l'une vers l'autre.



Chaque page a un lien sortant, donc C(A) = C(B) = 1

Nous ne connaissons pas le PR des deux pages, donc il nous faut une valeur de départ : 1.0 par exemple.

PR(A) = (1 - d) + d(PR(B)/1)
PR(B) = (1 - d) + d(PR(A)/1)

Soit, avec un facteur d'amortissement d de 0.85 :

PR(A) = 0.15 + 0.85 * 1 = **1**
PR(B) = 0.15 + 0.85 * 1 = **1**

Prenons une valeur de départ différente : 0

Première itération
PR(A) = 0.15 + 0.85 * 0 = **0.15**
PR(B) = 0.15 + 0.85 * 0.15 = **0.2775**

Deuxième itération
PR(A) = 0.15 + 0.85 * 0.2775 = **0.385875**
PR(B) = 0.15 + 0.85 * 0.385875 = **0.47799375**

Troisième itération
PR(A) = 0.15 + 0.85 * 0.47799375 = **0.5562946875**
PR(B) = 0.15 + 0.85 * 0.5562946875 = **0.622850484375**

Nous remarquons que les valeurs augmentent à chaque itération.

Dans notre exemple, avec nos deux pages A et B, nous savons que le PR doit être égal à un, l'algorithme précise que le PR moyen de toutes les pages du Web est égal à 1.

Vérification : Essayons avec une valeur supérieure pour voir ce qui se passe : prenons une valeur 2.0 pour redémarrer notre expérience.

PR(A) = 0.15 + 0.85 * 2 = **1.85**

PR(B) = 0.15 + 0.85 * 1.85 = **1.7225**

Bon, cela baisse ! Essayons une fois de plus :

PR(A) = 0.15 + 0.85 * 1.7225 = **1.614125**
PR(B) = 0.15 + 0.85 * 1.614125 = **1.52200625**

Une troisième fois :

PR(A) = 0.15 + 0.85 * 1.52200625 = **1.4437053125**
PR(B) = 0.15 + 0.85 * 1.4437053125 = **1.377149515625**

Nos valeurs continuent à converger vers 1, c'est ce que nous attendions.

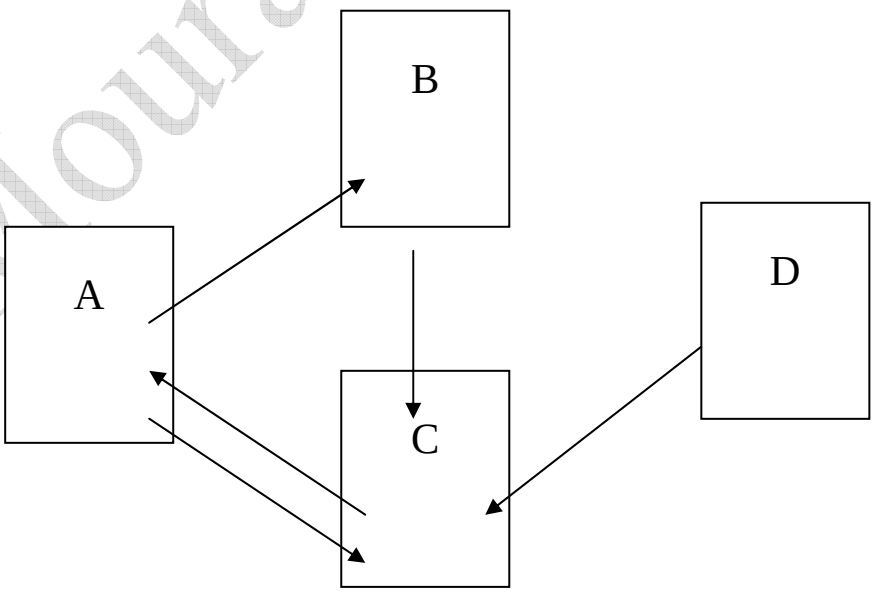
Résultat : Quelle que soit la valeur de départ prise pour le calcul du PR, la moyenne normalisée tendra vers 1.

L'exemple qui a précédé nous montre sur un cas smple, seulement 2 pages, combien d'itérations faut-il pour voir les résultats converger pour un grand nombre de pages ? . Ainsi, avec ses 100 milliards de pages dans sa base, Google nécessitera plusieurs milliards d'itérations.

C'est ici que le facteur d'amortissement joue son rôle. S'il est choisi trop élevé, le calcul demandera un nombre d'itérations énorme, alors que s'il est trop bas les valeurs ne convergeront pas véritablement, mais finiront par osciller autour de la valeur théorique vraie, un peu à la manière d'un pendule.

Il a été démontré qu'avec un facteur d'amortissement de 0.85, il nous faut une quarantaine d'itérations pour affiner le calcul du PageRank.

Autre exemple :



Dans cet exemple, nous avons un site comprenant quatre pages, dont une ne recevant aucun lien (la page D).
Le PR de cette page sera donc de 0.15, grâce au premier terme de la formule du PageRank (1 - d)
Bien qu'ayant un PR calculé, il est vraisemblable que cette page disparaîtra de l'index Google très vite, n'ayant aucun lien entrant.

Au bout d'une vingtaine d'itérations, les valeurs de PR pour nos pages convergent vers les valeurs suivantes :

Page A	1.49
Page B	0.78
Page C	1.58
Page D	0.15
Somme des PageRank	**4.0**
Moyenne	**1.0**

Nous voyons que dans notre exemple, la page C a le PR le plus élevé. C'était prévisible dès l'examen du graphique, comme elle reçoit un lien entrant des pages A,B et D, et n'en émet qu'un seul vers la page A.