Sébastien Forner
Sébastien Péron
Ruben Zamblé-bi

GOOGLE

La recherche de données

30 Mars 2006

Master A.S.S – 2006

Résumé

Les moteurs de recherche sur Internet existaient bien avant l'arrivée de Google sur le marché. Et pourtant, qui aurait pu croire que ce simple jeu de mot serait à l'origine du moteur de recherche le plus puissant au monde ! Google a su se distinguer grâce à une offre de services diversifiée et de très haute qualité. Le fonctionnement de Google est un aspect qui reste méconnu du grand public. Ce document a pour but d'éclaircir la situation en présentant une étude détaillée et claire du fonctionnement du moteur de recherche avec notamment, une explication de l'algorithme *PageRank* à la base de l'impulsion novatrice, qui a porté Google au rang de numéro 1, et du moteur d'indexation des documents ainsi que son architecture et les structures composant le moteur de recherche. Mais Google ne se résume pas qu'à cela, c'est aussi un innovant « business model » reposant essentiellement sur la vente d'espaces publicitaires dont l'enchérissement grandissant garanti le profit et l'expansion de Google. Quoi de mieux pour un publicitaire que de voir son annonce spécialement présentée pour un client déjà intéressé ? C'est ce que propose Google en analysant les requêtes de ses utilisateurs et en proposant aux annonceurs de mieux cibler leur public et donc de mieux placer leur publicité. C'est en cela que repose la force de Google.

Abstract

Search engines on the Internet had existed well before the arrival of Google on the market. And yet WHO could have believed that this mere pun would be at the origin of the most powerful search engine in the world! Google is known to be distinguished thanks to a diversified offer of services and very high quality. The functioning of Google is an aspect which remains widely unknown by its users. The purpose of this document is to clear up the situation by presenting a detailed and clear study of the search engine with in particular, an explanation of the *PageRank* (algorithm at the base of the innovations) which carried Google to the Number 1 rank, and of the indexing system of the documents as well as its architecture and the structures composing the search engine. But Google is not just limited to that, it is also one innovating "business model" based on the sale of advertising space whose growing rise guaranteed the profit and expansion of Google. What could be better for an advertising executive than to see his advert especially presented for an already interested customer? It is what Google offers by analyzing the requests of its users and proposing the advertisers to target their public better and thus to insert their publicity better. It is there that Google's force lies.

Table des matières

I. Introduction

Fondée en 1998 par Lawrence E. Page et Sergey Brin, la société « Google » est issue de la naissance du moteur de recherche portant le même nom [(1)]. Véritable leader incontestable, Google a su s'imposer face aux pionniers de l'époque (Alatavista, Lycos, Hotbot…) jusqu'à devenir aujourd'hui cette énorme machine introduite en bourse en mai 2004. La simple utilisation d'un moteur de recherche a laissé place à toute une panoplie de services et de logiciels ayant pour thématique principale la recherche, la classification et le référencement de documents.

Ce document décrit le large éventail de service que propose Google au delà du moteur de recherche qui l'a fait connaître. On y trouvera également son fonctionnement et sa méthode d'indexation des sites web qui ont fait de lui le plus performant des moteurs de recherche à l'heure actuelle.

L'objectif principal de cette étude est de montrer comment une simple startup a su s'imposer comme un géant de l'industrie grâce un sens de l'innovation présent depuis la technologie mise en place jusqu'à son modèle de croissance économique.

[1] *Le moteur Google était à l'origine un projet de recherche appelé « Backrub » (1996)*

II. Le principe de recherche selon « Google »

La principale caractéristique du moteur de recherche Google est de sélectionner les résultats en évaluant l'importance de chaque page web répertoriée avec des méthodes mathématiques basées sur plus de 500 millions de variables et de 2 milliards de termes. Cette technologie appelée « *PageRank* » contrôle le contenu des pages web mais également d'autres sites ayant un lien avec les pages analysées.

De plus, ce moteur de recherche est réputé pour sa rapidité d'exécution ainsi que la grandeur de ses données archivées [(2)], et également pour sa sobriété. Au fil des années, il a consolidé son succès grâce à un important réseau de serveurs d'indexation et à la qualité et à la pertinence des recherches retournées.

1. Sobriété et Valorisation des mots

Le site de recherche ne dispose que d'une page web minimaliste: pas de javaScript, de Flash ni de bandeau publicitaire clignotant, mais un unique champ de saisie dédié à la requête. Cette simplicité a permis l'utilisation du service aux internautes qui surfaient encore en bas débit, majoritaires à l'époque du lancement de Google.

Google met en place un système de valorisation des mots et de vente d'espaces de publicité associés qui lui permettent de se rémunérer. Ce système est basé sur une valeur par mot selon sa demande. Plus le mot sera demandé, plus il sera payé cher par clic. Ensuite Google affiche les publicités se rapprochant le plus de la demande saisie. Le but étant de « profiler » les utilisateurs et de leur présenter une offre très ciblée. Cette source de revenu représente 99% des revenus de la société.

2. Le système de classement « PageRank »

Le principe de fonctionnement de Google, qui a fait son succès, tourne autour d'une invention de ses créateurs, le « PageRank ». On peut le définir comme l'indice de popularité d'une page web, calculé selon un algorithme très sophistiqué, élaboré par Google. Ce calcul est basé sur l'étude des liens entre les pages web.

2.1. Principe de fonctionnement

L'algorithme PageRank (PR) fait partie des critères utilisés pour déterminer le positionnement des pages dans Google. Ainsi, pour deux pages au contenu comparable, celle ayant le meilleur PageRank sera souvent classée devant, surtout pour les requêtes très concurrentielles. Le PageRank d'une page est d'autant plus grand que, d'une part, de nombreuses autres pages font un lien vers elle, et d'autre part, que chacune de ses pages faisant un lien aient elles aussi un PageRank élevé. Le PageRank réel d'une page n'est connu que de Google, mais on peut en connaître une approximation évaluée de 0 à 10 notamment grâce à des outils développés par Google tels que *GoogleToolbar*.

[2] *Google aurait indexé plus de 8 milliards de pages web, 1 milliard d'images. Le temps de réponse moyen d'une requête est de l'ordre de 0.29 secondes.*

2.2. Expression Mathématique

Un article [3] rédigé par les deux fondateurs de Google aborde l'aspect théorique et mathématique du PageRank. Cependant, il s'agit d'une version initiale qui a évoluée mais la base reste la même (à noter que le secret du PageRank est bien gardé car il constitue un élément clé de la puissance de recherche de Google).

Soient A1, A2, ..., An , n pages pointant vers une page B. Notons PR(Ak) le PageRank de la page Ak, N(Ak) le nombre de liens sortants présents sur la page Ak, et d un facteur compris entre 0 et 1, fixé en général à 0,85. Alors le PageRank de la page B se calcule à partir du PageRank de toutes les pages Ak de la manière suivante :

```
PR(B) = (1-d) + d x ( PR(A1) / N(A1) + ... + PR(An) / N(An) )
```

Les premières constatations que l'on peut tirer de cette expression sont que le PageRank d'une page B ne dépend que de 3 facteurs : *(1)* le nombre de pages Ak faisant un lien vers B, *(2)* le PageRank de chaque page Ak, *(3)* le nombre de liens sortants de chaque page Ak. A contrario, il ne dépend donc pas des éléments suivants : *(1)* le trafic des sites faisant un lien vers B, *(2)* le nombre de clics sur les liens vers B dans les pages Ak, *(3)* le nombre de clics sur les liens vers B dans les pages de résultats sur Google.

2.3. Dérive liée au PageRank

La compréhension du PageRank a donné naissance au « *bombardement Google* » (*Google bombing*), technique visant à fausser en quelques sortes le classement d'une page web dans le moteur de Google. Cette technique exploite une caractéristique du moteur PageRank qui accorde un certain poids au texte avec un hyperlien vers une page. Ce poids peut varier en fonction de l'algorithme utilisé par Google. Si plusieurs sites utilisent le même texte pour pointer sur la même cible, Google additionne ce poids et il devient possible de faire apparaître la page cible dans les résultats d'une recherche sur le texte contenu dans les liens pointant vers elle.

Par exemple, si de nombreux sites font un lien vers A avec le mot-clé K, alors le site A sera très bien placé pour une recherche sur K, même si ce mot K est absent de tout le site A. Cette technique a été utilisée à maintes reprises pour des fins ludiques, politiques et économiques [4] par des internautes, par exemple :

« *weapons of mass destruction* » (armes de destruction massive) renvoyait une page d'erreur typique de domaine du système d'exploitation Windows critiquant les raisons d'entrée en guerre des États-Unis contre l'Irak en 2003.

« *french military victories* » (victoires militaires françaises) renvoie une page d'erreur à la Google initiée par les partisans du gouvernement de George Bush suite à la menace de veto

[3] *The Anatomy of a Large-Scale Hypertextual Web Search Engine, Sergey Brin et Lawrence Page, www-db.stanford.edu/~backrub/google.html*

[4] *La société BMW a procédé de cette façon pour obtenir un meilleur référencement par le moteur de Google ce qui a provoqué la colère de Google qui l'a éliminée de son fichier d'index en janvier 2006.*

français au Conseil de sécurité des Nations Unies de l'avant-guerre en Irak de 2003. La page d'erreur suggère de rechercher french military defeats.

« *miserable failure* » (échec lamentable) renvoie la biographie officielle de George Bush ! Il y a quelque temps encore, « *great president* » renvoyait à une fausse biographie du président américain.

« *gros balourd* » renvoie vers une brève biographie de Jean-Pierre Raffarin, 1er ministre du gouvernement français en 2004.

« *Nicolas Sarkozy* » renvoie vers le site officiel du film Iznogoud et Iznogoud renvoie vers la biographie officielle de Nicolas Sarkozy.

La société Google ne remet pas en cause ses algorithmes car ce phénomène reflète l'opinion publique et n'endommage pas la qualité globale de ses services. Google considère cela comme un effet de mode qui va retomber dans l'oubli.

2.4. Conclusion sur le PageRank

Le PageRank reste un algorithme complexe et assez mal connu, surtout qu'il est en partie caché par ses auteurs, pour des raisons évidentes de concurrence industrielle. Ce principe a immédiatement été un succès, car il a permis des résultats plus pertinents que les autres moteurs de recherche qui se contentaient de comptabiliser les mots-clés insérés dans les pages des sites.

3. La « Google Dance »

Ce terme désigne le processus de mise à jour des pages indexées par le moteur Google. Une fois par mois, Google procède au calcul du PageRank de chacune des pages indexées. Cette période de mise à jour (aussi appelée *Google Update*) dure plusieurs jours, pour différentes raisons, premièrement parce que le calcul du PageRank d'une page A dépend de celui de toutes les pages liées à cette page A. Il faut donc laisser le temps à l'algorithme itératif de converger. Et ensuite parce que, une fois que les nouveaux PageRank sont calculés, il faut les distribuer sur les milliers de serveurs qui composent Google, et qui donnent les réponses aux requêtes des internautes. Pendant ce temps, il est possible de voir le PageRank ou le classement d'une page dans les réponses augmenter ou diminuer. On parle alors de « danse ».

4. Le processus d'indexation dans Google

Afin de mettre en place son système d'indexation, Google dispose de Bots, sorte de programmes autonomes. Il existe plusieurs versions de GoogleBot. La première, le G*oogleBot classique,* sert à indexer les pages web pour les inclure dans l'index de Google : il visite les pages à une fréquence qui dépend de plusieurs facteurs. La seconde, le

GoogleBot Mediapartner, est dédiée à l'analyse des pages afin de cibler au mieux les annonces à afficher.

4.1. Les GoogleBot

Google a mis en place un logiciel de type *crawler*, dénommé *GoogleBot*. Il s'agit d'un robot d'indexation des pages web. Le volume des données à traiter étant considérable, ce robot est programmé pour fonctionner sur des centaines de serveurs, avec des adresses IP différentes, pour assurer la mise à jour des quelques milliards de pages déjà indexées plus les millions de nouvelles pages à ajouter.

Le principe de fonctionnement est simple : quand le GoogleBot lit une page pour l'indexer, il rajoute à sa liste de pages à visiter toutes celles liées à la page en cours de traitement. Théoriquement, il devrait donc être capable de connaître la plupart des pages du web, c'est-à-dire toutes celles qui ne sont pas orphelines (une page est dite orpheline si aucune autre ne pointe vers elle). La fréquence de visite de GoogleBot sur une page web dépend de son PageRank : plus il est grand, plus il l'indexera souvent. D'un passage à l'autre, GoogleBot peut détecter une page devenue inexistante.

Dans cette "famille des GoogleBot" on distingue deux sortes de robots. Premièrement, le *Fresh Crawler*, dont l'adresse IP commence par 64.68.82., correspond au robot qui indexe les nouvelles pages trouvées par Google ; une fois visitées par ce robot, les pages apparaissent dans Google seulement quelques jours après. Ensuite, le *Deep Crawler* (ou *Full Crawler*), dont l'adresse IP commence par 216.239.46., correspond au robot qui effectue une indexation massive de tous les documents connus de Google, en général pendant environ une semaine, juste après la Google Dance.

Le Fresh Crawler n'indexe que les documents aux formats HTML et texte (formats MIME text/html et text/plain). A chaque mot ou phrase est en effet associé son type, basé sur le langage HTML. C'est ainsi qu'un mot contenu dans le titre sera jugé plus important que dans le corps du texte. Une échelle de valeurs classe les types de mots (titre de la page, titre de paragraphe H1 à H6, gras, italique, etc.). Ce prétraitement, associé à d'autres critères dont celui du PageRank, permet de fournir les résultats les plus pertinents en premier.

En ce qui concerne le Deep Crawler, il indexe d'autres types de documents (PDF, PostScript, Word, Excel, PowerPoint...). Il a pour objectif de faire une indexation massive de chaque site qu'il visite.

4.2. Les GoogleBot Mediapartner

Mediapartner parcourt et indexe les sites affichant des liens publicitaires textuels fournis par Google, afin de pouvoir adapter le contenu de ces publicités à celui de la page les affichant. Ce bot est totalement indépendant des robots permettant la génération de l'index de recherche de Google. Le choix de la fréquence des visites de Mediapartner sur une page appartient à Google.

5. La gestion des liens publicitaires dans Google

Google propose de diffuser des publicités *AdSense*. Il s'agit de liens publicitaires textuels tels que ceux qui apparaissent sur la droite des pages de résultats de Google (les *AdWords*). L'intérêt majeur de ce système est qu'il détermine de manière automatique le contexte de la page, et choisit en conséquence quelle publicité afficher. Ainsi, les taux de clics sont bien meilleurs quand la pertinence est au rendez-vous.

5.1. Principe de Google AdWords

Le système des Google AdWords est le système publicitaire de Google. Les AdWords sont des emplacements publicitaires dans les pages de résultats de Google contenant des liens textes vers les annonceurs. Les annonces sont choisies en fonction des mots tapés par l'internaute : si un annonceur a acheté un des mots tapés par l'internaute, alors sa publicité apparaîtra. Cependant, Google limitant à 8 le nombre de liens publicitaires par page, un système d'enchères permet de départager les annonceurs ayant acheté les mêmes mots. Le prix à payer par clic dépend de la concurrence.

5.2. Principe de Google AdSense

Les AdSense sont des liens publicitaires textuels tels que ceux qui apparaissent sur la droite dans les pages de résultats de Google (les AdWords). Le choix des annonces à afficher sur une page donnée est fait de manière totalement automatique par le système AdSense, qui détermine le contexte de chaque page.

Le fonctionnement du système des AdSense nécessite une inscription et une acceptation du site (sur lequel les AdSense doit être mise en place) par l'équipe de Google [5]. Après quoi, on peut commencer à afficher des publicités. Il suffit d'insérer un code JavaScript, identique pour toutes les pages. Contrairement aux AdWords qui n'impliquent que l'annonceur et Google, la mise en place des AdSense, permet le versement d'un pourcentage des gains à l'affilié (celui qui diffuse la publicité sur son site) de la part de Google.

[5] *Service accessible depuis l'adresse : https://www.google.com/adsense/*

III. La structure déployée

Google est une société de hautes technologies. Cependant, il est extrêmement difficile d'expliquer de façon claire et détaillée tous les rouages de son fonctionnement, les chiffres réels concernant la complexité de l'architecture étant gardés secrets, notamment pour ne pas permettre de calculer facilement l'investissement nécessaire pour concurrencer Google. Nous allons dans cette partie nous attacher à éclaircir cette situation.

Le premier terme à définir pour décrire la structure de Google est le *Googleplex* que l'on désigne comme une boite à outils logiciels pour les ingénieurs et développeurs de Google.

1. Le GooglePlex

Nous allons proposer ici plusieurs illustrations de la structure de Google. Voici une première approche globale [6] présentée par le schéma ci-dessous :

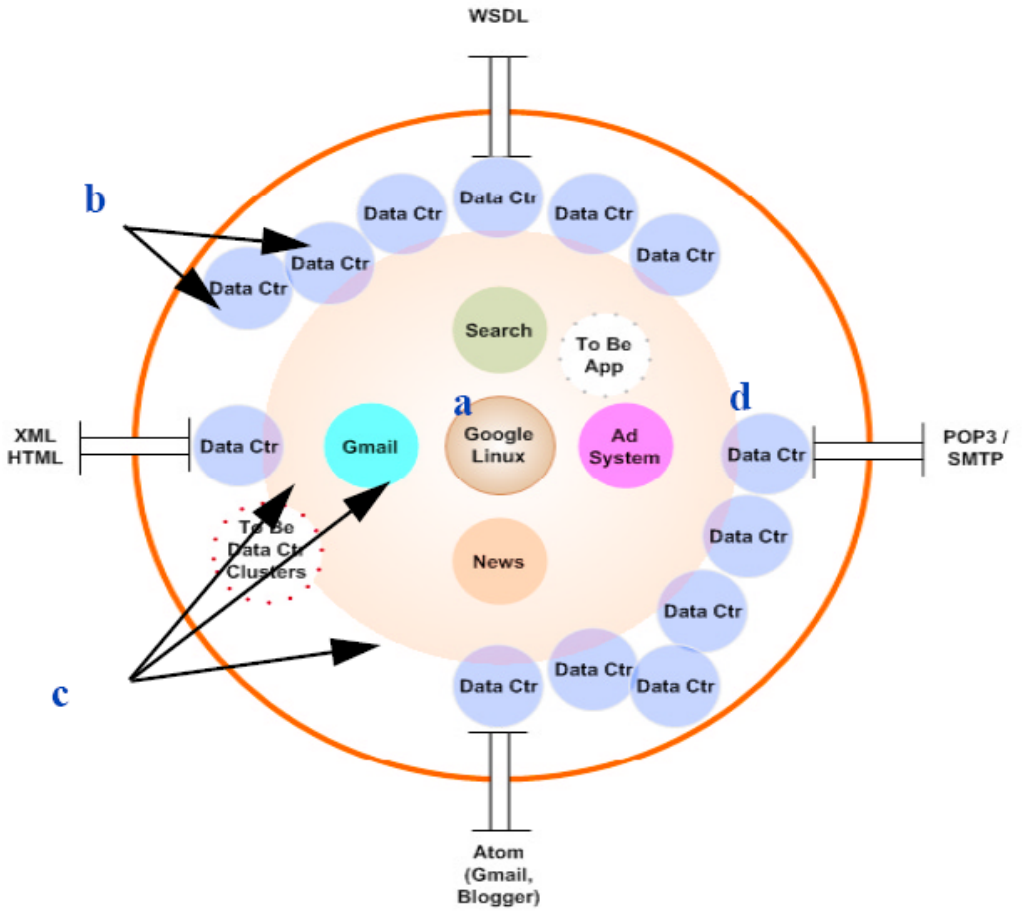


Fig. 1- Structure du réseau Google

[6] *A : Noyau linux B : Données indexées C : Une architecture technique identique à chaque niveau D : Le réseau Internet*

Actuellement, Google repose sur un système linux qui lui a permis d'étendre la taille des fichiers traités ainsi que d'autres fonctions permettant d'accélérer globalement la vitesse du système. Une multitude d'applications sont développées (telles que Gmail, Google Search ou Google News) et disponibles sur le réseau Internet qui permet un lien direct avec les données traitées.

On a maintenant une vision globale des différents niveaux d'interaction de Google avec le réseau Internet et les données qu'il utilise. Cette première approche ainsi faite, nous devons maintenant définir les différentes activités qui développent cette structure.

On distingue deux types d'activités dans le fonctionnement de Google. La première, l'*ingénierie logicielle,* est focalisée sur le Page Rank et d'autres applications. La seconde, l'*ingénierie « matérielle »*, garantie un faible coût de mise en place et de maintenance. Google se base sur la fusion de ces deux activités pour produire et développer ces applications (figure 2).

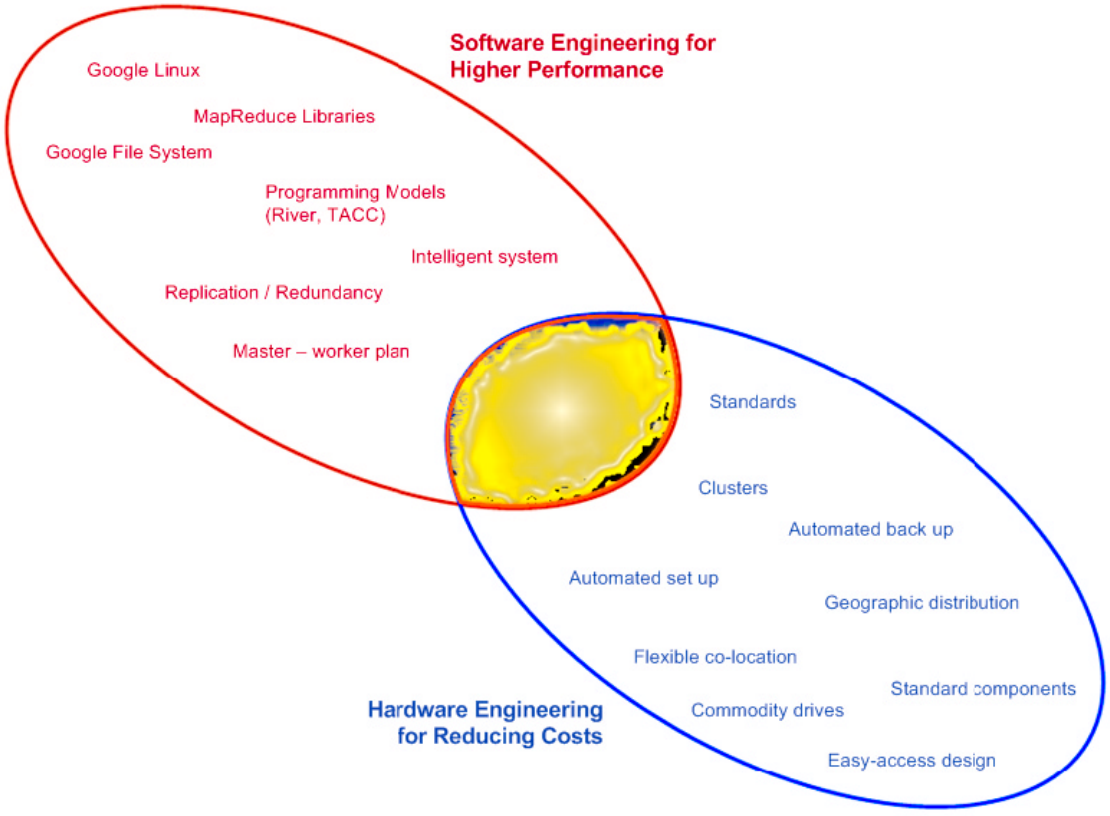


Fig. 2- Activités Google

Les logiciels ont un besoin crucial en matériels et infrastructures de réseau fiables et robustes. « Hardware et Software » sont très fortement liés pour Google. Les ingénieurs doivent faire des avancées significatives sur le hardware. Ainsi lorsqu'une avancée est faite, les ingénieurs software l'utilisent pour développer les fonctions des logiciels.

2. Les serveurs et les centres de données

2.1. Les serveurs

On distingue plusieurs types de serveurs utilisés, chacun est assigné à une tâche spécifique :

- *Les "Google Web Servers"*

Ils coordonnent l'exécution des requêtes envoyées par les utilisateurs et formatent le résultat dans une page html. L'exécution consiste à envoyer les requêtes au serveur, regrouper les résultats, calculer leur rang, extraire un sommaire pour chaque hit (en utilisant le serveur de documents) en demandant des suggestions au serveur d'orthographe, et finalement récupérer une liste de publicités fournie par le serveur de publicité.

- *Les "Data-gathering Servers"*

Les serveurs de rassemblement de données sont dédiés à la recherche de nouvelles données sur le web. Ils mettent à jour l'index et les documents des bases de données et appliquent les algorithmes Google pour assigner un rang aux pages.

- *Les "Index Servers"*

Chaque serveur d'index contient une série d'index. Ils retournent une liste d'identifiants de documents (« docId »), telle que les documents correspondent à un certain docId contenu dans les mots de la requête. Ces serveurs ont besoin de peu d'espace disque.

- *Le "Document Servers"*

Les serveurs de documents stockent les documents [7]. Chaque document est stocké sur une douzaine de serveur de document. Lorsque que l'on effectue une recherche, un serveur de document retourne un sommaire pour le document basé sur les mots de la requête.

- *Les "Ad Servers"*

Les serveurs de publicité sont en charge des publicités offertes par les services Adwords et Adsense.

[7] *Ces documents pour être accédés par les utilisateurs grâce à la visualisation « en cache » présente sur le résultat de la recherche renvoyée par Google*

- *Les "Spelling Servers"*

Les serveurs d'orthographe proposent des suggestions sur l'orthographe des requêtes.

2.2. Les Data Centers

Google utilise un ensemble de centre de données (« *Data Centers* ») regroupant les serveurs utilisés pour la recherche et l'indexation et ses diverses applications. On connaît à l'heure actuelle l'emplacement de plusieurs centres de données mais, bien entendu il n'est pas exclu que Google garde secret l'existence d'autres centres. On peut en donner une liste non exhaustive :

Nom, emplacement, description	Nom de domaine	Adresse IP
Santa Clara (Californie, Etats-Unis), hébergé par Exodus Communications, en service depuis 1998	www-ex.google.com	216.239.33.104
San Jose (Californie, Etats-Unis), hébergé par Global Crossing, en service depuis début 2000. Hors-service depuis octobre 2003	www-sj.google.com	216.239.35.104
Herndon (Virginie, Etats-Unis)	www-va.google.com	216.239.37.104
Washington DC (Etats-Unis)	www-dc.google.com	216.239.39.104
Virgnie (Etats-Unis)	www-fi.google.com	216.239.41.104
Sterling (Virginie, Etats-Unis), hébergé par Exodus Communications	www-ab.google.com	216.239.51.104
Santa Clara (Californie, Etats-Unis), hébergé par Exodus Communications	www-in.google.com	216.239.53.104
Zurich (Suisse) depuis juin 2002. Hors-service depuis novembre 2003.	www-zu.google.com	216.239.55.104
Palo Alto (Virginie, Etats-Unis) depuis janvier 2003. Par Cable & Wireless.	www-cw.google.com	216.239.57.104
Dublin (Irlande), depuis le 1er août 2003	www-gv.google.com	216.239.59.104
Connu depuis septembre 2003. Irlande	www-kr.google.com	66.102.11.104
Connu depuis octobre 2003. Semble être situé à Santa-Clara (Californie, Etats-Unis)	www-mc.google.com	66.102.7.104
Connu depuis novembre 2003. Semble être situé à Dublin, le QG de Google en Europe.	www-lm.google.com	66.102.9.104

Le trafic est à son maximum environ une fois par mois quand Google met à jour sa base. Des dizaines de tera-octets doivent alors transiter sur les réseaux internes. Lors de la Google Dance, les centres de données reçoivent les uns après les autres la nouvelle version de l'index (contenant l'ensemble des pages indexées par Google). Ceci a pour conséquence que pendant cette phase (qui dure en général 2 ou 3 jours), la réponse à une requête faite sur Google peut provenir d'un centre de données mis à jour ou d'un centre de données contenant les anciennes données.

Intégrés à Googleplex, voici la représentation résultante [8].

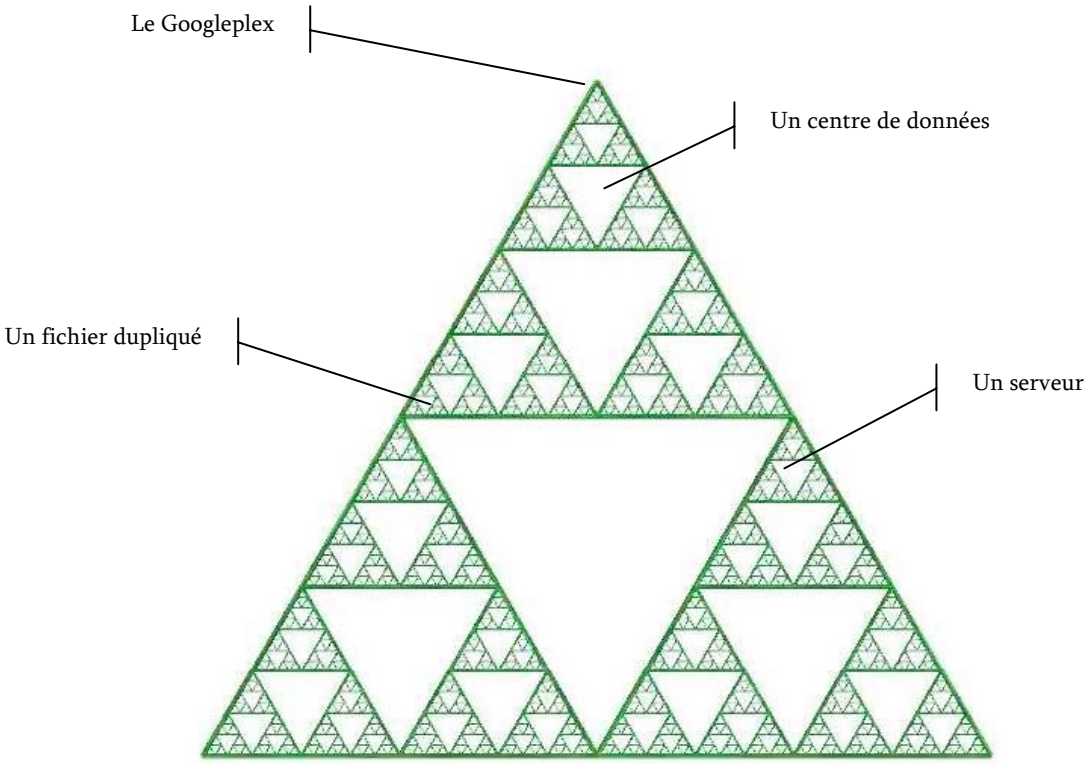


Fig. 3- Schématisation du Googleplex

Cette structure pyramidale inclue les centres de données de tailles plus ou moins importantes qui incluent eux-mêmes des serveurs. On obtient alors cette structure avec comme élément de base un fichier dupliqué encapsulé dans un serveur, un centre de données et enfin le Googleplex.

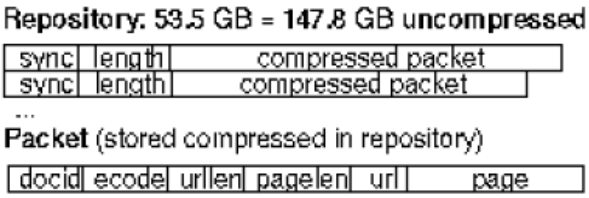
Dans cette approche, on souligne une architecture dite en « boites de pizza » imbriquées qui se base sur un réseau de centre de données répartis aux Etats-Unis et dans le reste du monde. Ces centres de données sont conçus pour être tolérants aux pannes.

[8] *Cette illustration est un triangle de Sierpinkski. Elle a été choisie car la structure de base utilisée, le triangle équilatéral, exprime la stabilité de l'approche Google sur son système*

3. Architecture d'indexation

Google accorde beaucoup d'importance au temps de réponse de chaque requête. Pour ne pas excéder 0,5 seconde, Google déploie ses centres de données dans le monde entier afin de rapprocher les serveurs des utilisateurs. Les centres de données de Google utilisent du matériel usuel que l'on peut se procurer dans un magasin classique d'informatique, et le plus intéressant est qu'il ne faut que 72 heures pour modifier leurs services *on-line* contrairement aux autres fournisseurs de services et applications web qui nécessitent une semaine voire un mois pour une même modification.

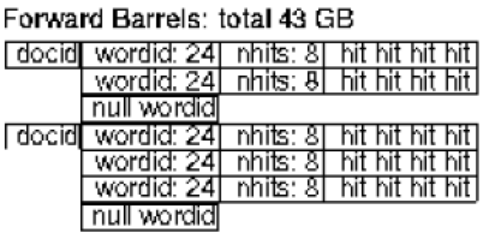
Le but de cette partie est de montrer les différentes étapes permettant le processus global d'indexation des données jusqu'au traitement des requêtes. Dans Google, le téléchargement des pages web est effectué par plusieurs « Crawler ». Au niveau de l'architecture matérielle, on retrouve des serveurs d'URL qui envoient des listes d'URLs récupérées par les Crawlers puis transmises au serveur de stockage qui compressent et stockent les pages web dans un entrepôt de données.



L'index de Google est découpé en petits bouts afin qu'il puisse être stocké sur chaque machine. Chacun de ces bouts est appelé un « *shard* ». La répartition des documents se base entre autres sur le PageRank. Chaque shard est dupliqué pour être sur plusieurs machines (il y a d'autant plus de duplications que le PageRank est élevé).

Toutes les pages web ont un numéro d'identifiant associé, le « *docID* » qui est assigné lorsqu'un nouvelle Url est analysé à partir d'une page web.

La fonction d'indexation est exécutée par l'« *indexer* » et le « *sorter* ». Chaque document est converti en une série d'occurrences de mots appelés « hits ». Les hits enregistrent le mot, sa position dans le document et une approximation sur sa police. L'indexeur dispose les hits dans une série de conteneurs, créant alors des index partiellement rangés. Enfin il analyse les liens contenus dans les pages web et stocke les informations importantes les concernant dans un fichier « *anchor* ». Ces fichiers contiennent les informations minimales permettant de déterminer d'où et vers quoi pointe chaque lien et son texte.



Le « *resolver* » d'URL lit les fichiers anchors et les URLs relatives en URLs absolues pour ensuite générer des docID. Il génère une base de donnée de liens (ensemble de pair de docID). Cette base de données est ensuite utilisée pour le calcul des PagesRanks de tous les documents.

Le « sorter » prend les conteneurs, rangés par docID, et les réarrange par *wordID* pour générer l'index inversé. Il produit alors une liste de wordID. Un programme appelé « *DumpLexicon* » prend la liste de wordID et le lexique produit par l'indexeur et génère alors un nouveau lexique qui pourra être utilisé par l'agent de recherche [9] (« *searcher* ») lors d'une requête.

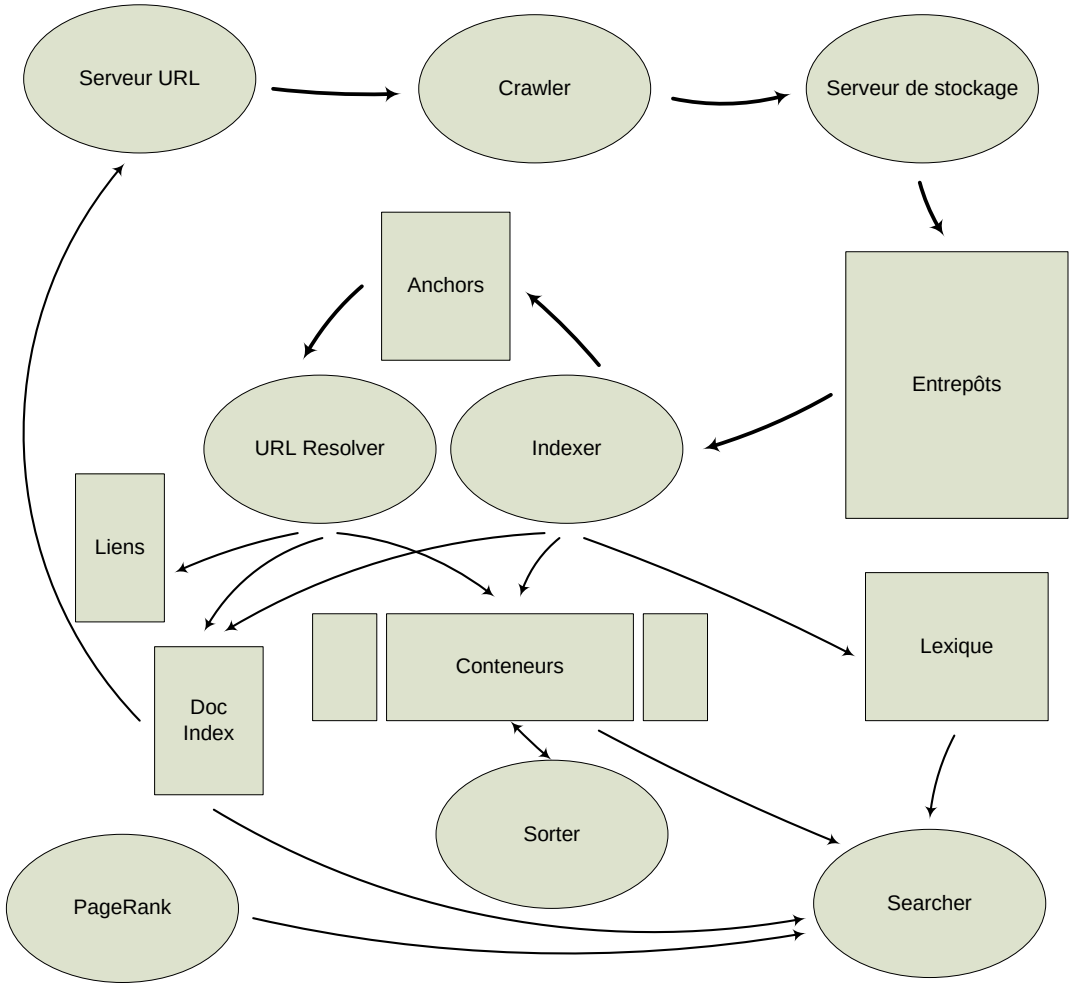


Fig. 6- Schématisation des interactions

[9] *Un agent est affecté à chaque recherche effectuée par un utilisateur de Google*

IV. La diversité de l'offre « Google »

Les applications et services de Google se diversifient et se multiplient au point de changer notre façon d'utiliser un ordinateur. Les services proposés couvrent à peu près tous les besoins des utilisateurs à la recherche d'informations et de produits, aussi bien pour les particuliers que les entreprises : images, livres, vidéo, news …. En regardant de plus près la thématique de ces services, on remarque clairement qu'ils indiquent bien la stratégie de Google d'indexer et de répertorier l'information.

Selon la source ou le genre d'informations recherchées, le moteur de recherche Google décline une version adaptée spécifiquement à la requête (livres, images, vidéos…). Cette partie présente les principaux services de recherche de données.

1. Google Web

Ce service correspond au moteur de recherche proprement dit. Accessible depuis le site d'accueil par n'importe quel internaute, le moteur dispose de nombreuses options et subtilités permettant d'améliorer la recherche (comme l'utilisation d'opérateurs logiques ou certains mots clés). Néanmoins, l'accès à la recherche ne se fait pas exclusivement à travers ce site. Il existe d'autres recours prévu par Google pour démocratiser d'avantage l'utilisation de son moteur. Google permet notamment aux propriétaires de sites d'inclure le service de recherche Google au sein de leurs propres sites (*Google Free Search*). L'utilisateur dispose également d'outils développés (tels que *Google Deskbar*) par Google permettant d'utiliser les services Google depuis un poste de travail ordinaire.

2. Google DeskTop

Avec l'accroissement du volume de données présentes sur l'ensemble des ordinateurs, il devient nécessaire d'utiliser un outil de recherche pour indexer et organiser ses fichiers. Google DeskTop dispose de la puissance du moteur de Google. Il indexe les fichiers et effectue des recherches sur l'index créé et stocké en local. Il passe en revue plusieurs types de fichiers bureautiques comme Word, Excel et PowerPoint, mais également les messages mails et leurs contenus (pour les principaux clients mails). Ce service utilise la même interface de recherche que Google Web et présente les résultats de recherche de la même manière.

3. Google Mini & Search Appliance

Google Mini est une déclinaison du moteur de recherche adaptée aux PME et TPE. Il permet d'indexer les informations de l'entreprise pour atteindre deux objectifs selon que le service est proposé aux employés (en interne) pour effectuer des recherches sur l'intranet, ou aux visiteurs du site Web de l'entreprise. Pour ce type de service, il est possible de personnaliser les feuilles de styles des pages de recherche pour les mettre aux couleurs de

l'entreprise. Google Mini se présente sous la forme d'un boîtier au format rack et est limité à l'indexation de 100 000 documents.

Un peu plus évolué que Google Mini, Google Search Appliance est conçu pour répondre aux besoins des grosses PME. Il permet d'accéder aux informations contenues dans un site web, un intranet et aussi dans les bases de données et les serveurs de fichiers. Le moteur d'indexation traite environ 220 types de fichiers, en particuliers les fichiers HTML, MS Office, PDF, PostScript,….. Google Search Appliance est décliné en trois modèles s'adressant à des entreprises de différentes tailles permettant d'indexer respectivement 1.5, 3 et jusqu'à 15 millions de documents.

4. Google Images

Cette variante du moteur de recherche permet de retrouver des photographies et des illustrations parmi une base comprenant plus d'un milliard de documents. Le principe de recherche est le même que la recherche de documents classique, incluant la gestion des opérateurs. D'autres caractéristiques types sont prises en compte pour affiner la recherche, telles que la taille des images, le format recherché (JPEG, PNG, GIF), monochrome ou couleurs …

5. Google Search Book

Ce service de recherche sur le contenu des livres est l'un des plus controversés de Google et a pour objet, selon les termes de l'éditeur, « d'organiser l'information mondiale ». Google numérise les livres d'éditeurs et les bibliothèques. Initialement nommé « *Google Print* », il fut renommé afin de lui éviter une connotation de viol des droits d'auteurs. Ce service permet de rechercher un livre selon son auteur, le nom de son éditeur, sa date de publication, son numéro ISBN [10], ou encore un thème ou des mots présents dans son contenu. Lorsqu'un livre est libre de droits, et que son contenu a été numérisé, il est possible de le consulter dans son intégralité, en revanche, lorsqu'il s'agit d'œuvre soumises aux droits d'auteur, on ne peut consulter que des extraits ou seulement la page de garde.

6. Google Earth

Ce service a pour vocation de permettre à n'importe qui de visualiser la Terre entière à différentes altitudes selon un niveau de Zoom que l'on peut faire varier. Pour obtenir un tel résultat, Google utilise la technologie de *Keyhole* (un logiciel de cartographie aérienne) enrichie d'un certains nombres d'informations qui permettent maintenant de repérer les commerces, les sites touristiques et bien d'autres. Google propose également des versions professionnelles de Google Earth, liées à la géolocalisation et à la planification d'itinéraires.

[10] Numéro de dépôt légal permettant de commander un livre auprès des libraires

7. Google News

Google News permet de rechercher et de lire des articles issus de plus de 500 sources d'informations. Textes et photos sont actualisés plusieurs fois par jour afin d'être toujours en phase avec les évènements. Le moteur compile les titres de l'actualité des différentes publications et les classe dans différentes catégories (Economie, Sport, Science, Politique,…). Ce service permet de définir une page personnalisée avec un choix de présentation et des catégories des news à afficher. Bien entendu Google News est également un moteur de recherche qui permet d'effectuer des recherches sur tous les articles indexés par Google News dans les 30 derniers jours.

V. Une approche du Datamining : Google vous surveille

1. Google et les entreprises

1.1. Google s'invite dans les TPE et PME

Mi janvier 2005 Google lance un nouveau service, mais contrairement à ses habitudes celui ci n'est pas gratuit. La raison est simple : l'outil est destiné aux entreprises, ce qui est une première chez Google. En effet celui-ci pourra offrir aux employés et clients les mêmes possibilités de recherche que sur Google.com. Ainsi l'Intranet de l'entreprise ou son site web public sera associé à Google qui est associé à l'idée de puissance et de facilité de recherche. La promesse est grande car le principal but est de ne plus perdre de documents stratégiques et surtout de les retrouver directement et donc de bénéficier d'un gain de temps.

La machine de base moyennant 2995 euros permet d'indexer jusqu'à 100 000 documents et est compatible avec 220 formats de documents (HTML, PDF, Microsoft Office...). Elle supporte 60 requêtes à la minute, le tout stocké dans format de boîtier rack 1U.

1.2. Google Appliance

Un peu plus évolués que Google Mini, les boîtiers Google Search Appliance sont conçus pour répondre aux besoins des grosses PME et des grands comptes. Basé sur un système d'exploitation Linux, le boîtier repose sur un Pentium III 1GHz, 2 Go de mémoire vive et 80 Go de disque dur. Ses possibilités d'accès sont accrues car il permet d'accéder aux informations contenues dans les bases de données (IBM, DB2, Microsoft SQL Server, MySQL, Oracle et Sybase) ainsi que les serveurs de fichiers.

Google Search Appliance est décliné en trois versions s'adressant à des entreprises de différentes tailles:

- GB-1001 indexe jusqu'à 1,5 millions de documents
- GB-5005 en indexe jusqu'à 3 millions
- GB-8008 15 millions.

Ils sont fournis avec des licences valides pendant 2 ans garantissant le matériel, le logiciel, les mises à jour du produit, l'assistance et l'assurance de remplacement du produit en cas de panne.

Le service assure la gestion d'au moins 150 requêtes par minute sur une base de plusieurs millions de documents, le tout affichant des performances relativement rapides. De plus ces produits bénéficient de l'adjonction de deux fonctions de recherche à son socle logiciel : un mécanisme d'authentification des utilisateurs pour contribuer à la protection des documents sensibles ainsi qu'une fonction d'administration visant à définir les sources prioritaires lors de la mise en oeuvre du

processus d'indexation. Ce dernier ne génère pas de ralentissement du réseau car il se fait petit à petit en tâche de fond et peut être accéléré lors des heures creuses.

2. Google chez les particuliers

2.1. Google Desktop Search

Avec Google Desktop Search, on peut disposer de la puissance du moteur de recherche sur son bureau. En effet il indexe les fichiers et effectue des recherches sur l'index crée et stocké en local. Hormis l'indexation des fichiers bureautiques les plus courants, il indexe aussi les messages et leurs contenus provenant de Outlook, Outlook Express, Thunderbird. Google Toolbar, un élément intégrable à Google Desktop, a été développé de manière à pouvoir utiliser ce potentiel de recherche très facilement. Google espère démocratiser et légitimer cet outil auprès des utilisateurs. Pour cela, il propose des fonctions avancées telles que le correcteur orthographique, le blocage des fenêtres indésirables et le traducteur, un système de recherche amélioré avec une suggestion en temps réel pour ne citer que les plus pratiques.

2.2. L'espion qui m'aimait

Maintenant nous allons réfléchir à ce qui se passe lorsque nous naviguons à travers le web et que notre poste est équipé d'une GoogleBar. Dans sa course effrénée à la performance, Google use de toutes les possibilités pour améliorer ses résultats de recherche. Connaître nos moindres faits et gestes sur la toile fait évidemment partie des voies explorées.

2.2.1. Opportunités amenées par la GoogleBar

A chaque fois qu'un utilisateur arrive sur une nouvelle page, la barre Google indique le PageRank de cette page. Pour obtenir cette information, la Google ToolBar a du envoyer une requête à l'un des Datacenter de Google. Cette requête contient l'adresse IP du poste de départ (ce qui permet de savoir dans quel pays l'utilisateur réside), ainsi que l'adresse de la page visitée.

Le Datacenter interrogé n'a plus qu'à constituer une base de données de toutes les requêtes, en y rajoutant les dates et les heures de chacune d'elles pour pouvoir connaître les sites préférés d'un utilisateur, le temps passé sur chaque page, le parcours à l'intérieur de chaque site ainsi le parcours de sites en sites, l'utilisation des résultats proposés par Google (Est-ce que l'utilisateur clique toujours sur le premier site proposé ? Combien de sites proposés l'utilisateur explore t- il suite à une recherche?).

Tous ces renseignements permettent à Google de connaître très précisément les préférences et habitudes de navigation des utilisateurs. Nul doute qu'ils font désormais partie des indices permettant de classer les sites dans les résultats des recherches.

2.2.2. Un atout indéniable

Si on estime que plusieurs dizaines de millions d'internautes ont installé la Google ToolBar sur leur poste, on peut donc dire que Google est en mesure de connaître les habitudes de navigation d'une immense proportion des utilisateurs du web. Cet échantillon est largement suffisant pour se faire une idée très précise des sites qui sont appréciés ou boudés par les internautes.

Cette source d'information présente au moins deux avantages inédits et très intéressants par rapport aux autres sources dont peut disposer Google. D'abord, elle est infalsifiable : car aucun webmaster ne peut truquer la fréquentation réelle de son site, ensuite, elle indique le parcours des internautes et le temps passé sur chaque site.

2.2.3. Exemples d'applications possibles

Estimer la qualité d'un site

Un moyen simple consiste à regarder combien de temps un utilisateur passe sur un site, combien de pages sont consultées et combien de liens proposés par le site sont suivis. A titre d'exemple, une page à laquelle les internautes consacrent en moyenne 5 secondes et qu'ils quittent sans suivre les liens qu'elle propose est probablement une page qui les a déçu.

Vérifier la qualité des résultats de Google

Il suffit de comparer l'ordre des sites proposés par Google avec l'ordre des "qualités" estimées au point précédent. Il est aussi très intéressant de regarder quels sites sont choisis dans la liste que propose Google et combien de temps l'utilisateur passe sur ces sites. On peut supposer que si l'utilisateur quitte ces sites au bout de quelques secondes, c'est que il les estime "non-pertinents" par rapport à la recherche mise en oeuvre.

Identifier les "spammeurs"

Etudions les fréquentations des sites qui occupent souvent la première place des résultats. Si cette fréquentation est très rapide (l'internaute ne reste que quelques secondes sur le site), alors que ces sites sont censés être de bonne qualité, il y a de forte chance pour qu'il s'agisse de "spammeurs" qui ont gagné leurs premières places de façon "malhonnête".

Nous venons de démontrer que le système traditionnel de notation des sites est sans doute devenu ou va devenir très accessoire dans le classement des résultats de Google. En effet la GoogleBar devrait permettre une amélioration des résultats délivrés par Google car les spammeurs devraient pouvoir être éradiqués.

3. Vers une publicité intelligente

3.1. Un profilage complet de l'utilisateur pour une publicité plus rentable

Lorsque Google aura suffisamment de profils d'utilisateurs et de capacités de calcul pour analyser ces masses de données, il pourra alors pleinement tirer partie du recoupement des horaires de connexion, des recherches effectuées, des sites visités, du carnet d'adresses des utilisateurs (celui de Gmail), des produits achetés….. Il pourra ainsi faire un portrait plus vrai que nature des habitudes de consommation des utilisateurs, à tel point qu'il sera capable de devancer leurs désirs...

De plus, si les utilisateurs disposent de GoogleEarth, Google pourrait aller jusqu'au géomarketing [(11)] et donc fournir dans ces liens publicitaires les adresses des revendeurs les plus proches.

Le but est de vendre à des annonceurs potentiels des liens publicitaires ciblés dont le taux de transformation en achat sera important. Actuellement, le problème est que la publicité a un taux de transformation faible car elle est de plus en plus rejetée par une partie grandissante de la population qui la considère comme une sorte de pollution. Grâce aux liens sponsorisés, Google n'a pas trop souffert de ce problème mais se retrouve face à un plafonnement des prix à cause des prix atteints aux enchères sur certains mots-clés très demandés.

3.2. Confirmation d'une solution ?

La publicité ciblée est vue comme une solution car elle permettra aux annonceurs de limiter leur diffusion et à Google de justifier une augmentation des tarifs. Notons que les liens sponsorisés représentent près de la moitié des revenues de la publicité en ligne.

La publicité en ligne et le datamining ont donc de beaux jours devant eux et Google des raisons de penser que son modèle d'affaire est bien plus rentable que la vente de logiciel aux particuliers.

[11] *A condition de stocker les vues favorites dans un marque-pages accessible par Google*

3.3. Google et la législation

Que dit la loi sur le traitement des données personnelles ? Elle signale que les utilisateurs d'un service en ligne doivent être informés de ce qui se passe exactement avec leurs données personnelles et doivent avoir la possibilité de s'y opposer. Les logiciels sérieux respectent ces dispositions en publiant des chartes mais les internautes les lisent-ils vraiment ? C'est pour cela que Google a signé le *Safe Harbor* le 15 octobre dernier pour pouvoir exporter des données européennes vers les Etats-Unis tout en s'engageant à respecter les principes de transparence et d'opt-in [12].

Il ne faut pas oublier que nous ne connaissons de Google que la face émergée de l'iceberg puisque, dés qu'un de ses secrets est dévoilé ou démasqué, il est instantanément repris par les webmasters, entre autres, pour essayer d'en tirer profit. Peut-être ne connaissons nous que les techniques utilisées auparavant et donc tout reste à découvrir.

[12] *L'internaute doit donner son accord explicite.*

VI. Le TrustRank en guerre contre le spamdexing

Le TrustRank (indice de confiance) est rapidement devenu un sujet à la mode lorsque la communauté des spécialistes de Google et des webmasters s'est aperçue que Google venait de déposer cette marque en mars 2005. Le vocable avait été introduit une année plus tôt par un article publié à Standford exposant une méthode pour combattre le spamdexing [13]. Certains se sont alors avancés à prédire la mort prochaine du PageRank. En fait, cette vision est excessive puisque TrustRank et PageRank apparaissent largement complémentaires.

1. Principes du trustRank

Le but du TrustRank est d'essayer de détecter les cas de spamdexing. Seulement sa détection est quelque peu subjective, et même si elle semble facile pour un humain, ce n'est pas forcément le cas pour un ordinateur. Les sociétés qui exploitent les moteurs de recherche emploient des salariés à visiter constamment le web à la recherche des contrevenants pour écarter le spamdexing. Lorsqu'une page suspecte est identifiée, le robot d'indexation cesse de la visiter, et la page sort de l'index. Ce processus est particulièrement lent et onéreux, mais il est essentiel à la crédibilité du moteur de recherche : sans recourir à la suppression des pages incriminées, la qualité des résultats de recherche serait significativement dégradée.
L'identification du spamdexing par l'algorithmique pure étant difficile, le système reposera partiellement sur une intervention humaine. L'algorithme commence par sélectionner un petit échantillon de pages dont le statut au regard du spamdexing doit être fixé. Un expert examine cet échantillon pour indiquer à l'algorithme quelles pages relèvent du spamdexing. Puis, l'algorithme identifie les autres pages qui sont probablement de bonnes pages du fait de leurs relations avec les bonnes pages de l'échantillon.

2. Outils et techniques utilisés

2.1. Vision du web

Nous modélisons le web comme un graphe $G = (V, E)$ consistant en un ensemble V de N pages (sommets du graphe) et un ensemble E de liens orientés (arcs du graphe) qui connectent les pages. Dans la réalité, une page p peut avoir plusieurs hyperliens vers une autre page q. Les liens d'une page sur elle-même ne sont pas pris en compte. La figure 7 présente un graphe simple de 7 pages et 8 liens.
Chaque page a des liens entrant (*inlinks*), et des liens sortant (*outlinks*). Le nombre de liens entrant d'une page p est son "taux entrant", alors que le nombre de liens sortant sera désigné "taux sortant". Par exemple, le taux entrant de la page 3 est de 2 alors que son taux sortant est de 1.
Les pages n'ayant pas de lien entrant sont des "pages sans référence". Les pages sans lien sortant sont des "pages ne référençant pas". Les pages à la fois sans référence et ne

[13] *Ensemble de techniques permettant de fausser le classement de Google.*

référençant pas sont des "pages isolées". La page 1 est une page sans référence, tandis que la page 7 ne référence pas.

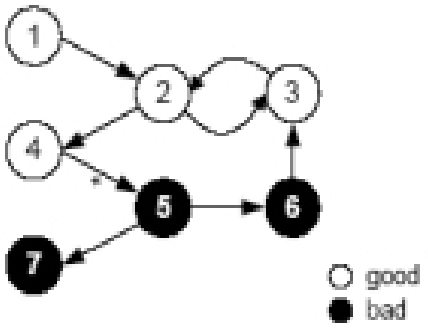


Fig. 7- Représentation du web selon un graphe

2.2. Sélection de l'échantillon

Cette première étape consiste à choisir des pages sur le web pour pouvoir ensuite les soumettre à l'Oracle [14] (intervention humaine). De plus il faut faire en sorte de limiter la taille de l'échantillon car l'appel à l'Oracle coûte relativement cher. Le nombre de pages confiées à l'Oracle sera noté L.

2.2.1. Aléatoire

Manière la plus simpliste consistant à prendre des pages aléatoirement à travers le web.

2.2.2. PageRank inversé

Comme la confiance est réduite par les bonnes pages de l'échantillon source, une des approches est de donner la préférence aux pages permettant d'atteindre le plus grand nombre de pages, celles comprenant de nombreux liens sortants. Dans la figure précédente, l'échantillon approprié si `L = 2` serait `S = {2,5}` puisque les pages 2 et 5 offrent le plus grand nombre de liens (2 chacune).

Le raisonnement peut être étendu pour une meilleure couverture du graphe (du web). Nous pouvons construire l'échantillon à partir des pages qui pointent vers de nombreuses pages qui elles mêmes pointent à nouveaux vers de nombreuses pages et ainsi de suite. Cette approche nous conduit à un modèle proche du PageRank, à la différence que le critère à optimiser repose sur le nombre de liens sortants au lieu du nombre de liens entrants. En conséquence, le calcul de l'éligibilité d'une page peut être traité en appliquant le PageRank. Or, les liens étant inversés, nous appelons cette version du PageRank le PageRank inversé.

[14] *Au sens de principe algorithmique.*

2.2.3. PageRank élevé

Une bonne remarque serait de dire qu'il est certainement intéressant de s'assurer de la qualité des pages qui apparaissent en premier dans les résultats de recherche. Prenons par exemple 4 pages p, q, r, et s dont le contenu correspond de façon équivalente à une recherche donnée. Si le moteur de recherche utilise le PageRank pour ordonner ses résultats, la page présentant le meilleur PageRank, disons p, sera placée en tête, puis la page q si son PageRank arrive en deuxième, et ainsi de suite. Comme il est vraisemblable que l'utilisateur s'intéresse d'abord aux pages p et q plutôt qu'à r et s (qui ne seront peut-être même pas vues si elles sont sur les pages de résultats suivantes), il semble plus utile d'obtenir une bonne précision sur les pages p et q. Si la page p utilise le spamdexing, l'utilisateur devrait plutôt visiter q à la place.

Une seconde heuristique pour établir l'échantillon de départ serait de donner la préférence aux pages de fort PageRank. Comme les pages de fort PageRank ont de bonnes chances de pointer sur des pages de fort PageRank également, les scores de confiance vont se propager aux pages qui ont de fortes probabilités d'être en début de résultats de recherche. En sélectionnant l'échantillon de départ avec le PageRank élevé nous identifierions certainement la qualité d'un plus petit nombre de pages qu'avec le PageRank inversé,mais cette qualification toucherait les pages dont il est particulièrement important de connaître le niveau de confiance.

2.3. Appel de l'Oracle et propagation de la confiance

2.3.1. Partitionnement de l'échantillon

Une fois l'échantillon de départ constitué, il est transmis à l'Oracle qui inspecte les pages web désignées pour pouvoir les séparer en deux sous-ensembles : les bonnes pages et les mauvaises pages. A présent, les pages vont recevoir une note comprise entre 0 et 1. Toutes les pages désignées comme bonne par l'Oracle obtienne un 1, les mauvaises pages un 0 et enfin les pages ne faisant pas partie de l'échantillon de départ obtiennent 0,5 car on ne peut pas dire pour l'instant si elles sont bonnes ou non. C'est ce que l'on appelle la fonction de confiance ignorante.

2.3.2. « Isolation approximative» des bonnes pages

Pour déterminer les bonnes pages sans invoquer l'Oracle sur tout le web, on se repose sur un constat empirique essentiel appelé "isolation approximative" des bonnes pages : les bonnes pages pointent rarement vers les mauvaises. Cette propriété est assez intuitive car les pages utilisant le spamdexing sont réalisées pour tromper les moteurs de recherche, et non pour présenter de l'information utile. Il en résulte que les personnes créant des bonnes pages n'ont que peu de raisons de les lier à des mauvais contenus.
Cependant, les créateurs de bonnes pages peuvent être trompés et abusés, ainsi, on trouve parfois des bonnes pages pointant vers des mauvaises (lien de la page 4 vers la page 5 sur la figure 7).
Précisons que la réciproque du principe d'isolation approximative n'existe pas car dans la pratique les mauvaises pages sont souvent reliées aux bonnes pages.

2.3.3. Confiance à M étapes

Mélangeons maintenant les notions de confiance ignorante et d'isolation approximative afin de propager la confiance à travers le web. Pour cela nous allons nous baser sur un exemple en prenant le graphe de la figure 7 et en partant du principe que l'échantillon de départ contient les pages 1, 3 et 6.
Nous obtenons donc : [1 ½ 1 ½ ½ 0 ½]

Propageons la confiance durant M=3 étapes :

M = 1 -> [1 1 1 ½ ½ 0 ½]
M = 2 -> [1 1 1 1 ½ 0 ½]
M = 3 -> [1 1 1 1 1 0 ½]

Finalement, en analysant les résultats, on s'aperçoit que nous sommes confrontés à un problème. En effet, la page 5 qui est sensée être une mauvaise page se retrouve avec une confiance de 1 ce qui veut dire qu'elle aurait obtenu un nouveau statut de bonne page ce qui est une erreur. En fait, plus la distance à notre échantillon de bonnes pages est grande, plus la probabilité de tomber sur une bonne page baisse. Nous allons donc essayer de diminuer la confiance que nous accordons à une page en fonction du nombre d'étapes.

2.3.4. Amortissement de la confiance

Il y a bien des façons d'envisager l'amortissement de la confiance, ainsi nous allons en décrire deux possibles.

- Confiance atténuée

La figure 8 illustre la première idée appelée « confiance atténuée ». Comme la page 2 est à une étape de l'échantillon de bonnes pages grâce à la page 1, nous lui affectons un score de confiance amorti de β, avec β < 1. Comme la page 3 est à un lien de la page 2 dont le score est β, nous affectons à cette page 3 un score amorti de β x β. Nous avons également besoin de définir l'affectation de la confiance dans le cas de liens entrants multiples. Par exemple, supposons que la page 1 pointe également vers la page 3. Nous pourrions alors affecter à la page 3 le score de confiance le plus élevé c'est à dire β, ou alors un score moyen, c'est-à-dire (β + β x β) / 2.

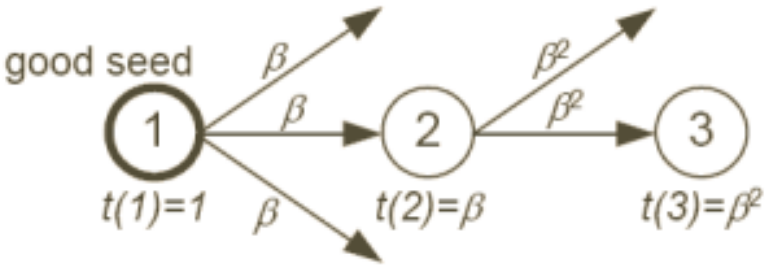


Fig. 8- Schématisation de la confiance atténuée

- Confiance fractionnée

La seconde méthode proposée pour l'amortissement de la confiance est appelée confiance fractionnée et repose sur le constat suivant : le soin apporté à ajouter des liens sur une page est souvent inversement proportionnel au nombre de liens présents sur la page. En d'autres termes, si une bonne page n'a qu'une poignée de liens sortants, alors il est vraisemblable que les pages ciblées soient bonnes également. Au contraire une bonne page contenant des centaines de liens sortant a une probabilité plus élevée de pointer vers des mauvaises pages.

Cette observation nous conduit à fractionner la confiance lors de la propagation aux autres pages : si une page p a niveau de confiance de 1 et pointe vers X pages, alors chacune des X pages va recevoir une fraction 1 / X de la confiance de p. Dans ce cas la confiance d'une page sera la somme des fractions reçues de tous ses liens entrant. Intuitivement, plus une page accumule de crédit provenant des pages environnantes, plus il y a de chance qu'elle soit bonne. (Nous pouvons bien sûr normaliser la somme pour que la confiance reste dans l'intervalle [0,1]). La figure illustre ce fractionnement de la confiance. La page 1, élément de l'échantillon de bonnes pages, a 2 liens sortants et distribue donc la moitié de son score de confiance à chacune de ses cibles. De façon similaire, la page 3 distribue le tiers de son score de confiance. Le score de la page 3 sera donc de 1 / 2 + 1 / 3 = 5 / 6.

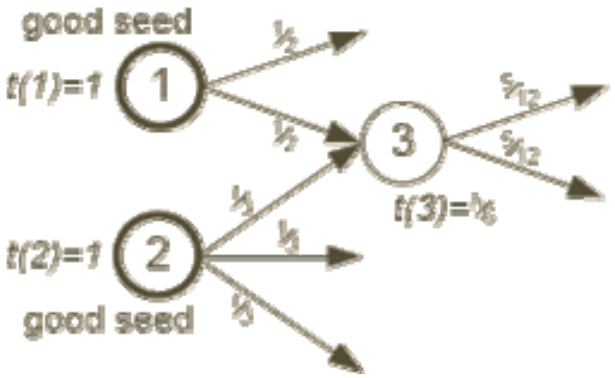


Fig. 9- Schématisation de la confiance fractionnée

3. Résultats et analyse sur le web

D'après un article publié en mars 2004 par 2 chercheurs de l'Université de Stanford et un ingénieur de Yahoo! Inc., les techniques utilisées ci-dessus pour des tests sur « l'ensemble » du web rapportent des confirmations sur le classement du PageRank ainsi que sur la qualité de ses résultats retournés. Toujours d'après eux, « le TrustRank est un outil honnête de détection du spamdexing » car il supprime la majeure partie du spamdexing présent parmi les sites les mieux classés. En conséquence, ils affirment que, contrairement au PageRank, le TrustRank garantit que les sites du haut du panier sont de bons sites.

Tout cela pourrait expliquer pourquoi Google s'est emparé du TrustRank. Mais n'oublions pas que Google n'a jamais fait de communiqué explicitant l'utilisation du TrustRank et que nous ne pouvons pas savoir clairement s'ils l'utilisent et surtout de quelle manière ils l'utilisent. En effet, comme nous l'avons vu auparavant, le TrustRank peut être utilisé avec plusieurs variantes pour ses différentes fonctions constituantes.

Il est à noter que durant la période comprise entre les mois d'Avril et Juillet 2005, il s'est produit quelques « bizarreries » dans les pages indexées et retournées par les recherches sur Google. Certains sites ont complètement disparu et d'autres se sont retrouvés sur les cimes des classements sans pour autant avoir fait d'optimisations spécifiques. Ce phénomène amplement analysé sur les forums spécialisés a reçu comme doux nom: *l'effet Bourbon*. Il semblerait que durant cette période Google ait fait des modifications sur ses algorithmes et les ait testés sur le web. Mais à l'heure actuelle, on ne sait toujours pas si cela a été un premier essai grandeur nature du TrustRank.

VII. Conclusion

On constate actuellement une omniprésence de Google dans le monde de l'Internet. Google s'est imposé de lui-même comme référence dans le domaine des moteurs de recherche grâce à la structuration robuste de son architecture et à un constant développement de ses applications. Ses fondateurs, simples étudiants d'université à l'époque de sa création, ont su analyser les réels besoins des internautes ; des applications simples, rapides et efficaces, voilà ce que l'on attend d'Internet !

L'interface claire, sans artifice, associée à l'invention du Pagerank a permis, dès le début d'Internet, et malgré de bas débits, de proposer un service de recherche efficace. Aujourd'hui Google a ouvert la voie à une nouvelle vision du service sur Internet en se basant sur un modèle économique révolutionnaire. Peu nombreux sont ceux qui, au départ, auraient parié sur l'audacieuse et formidable réussite de Google. Et pourtant aujourd'hui, Google suscite l'admiration et l'envie chez beaucoup d'entre eux.

Google est une entreprise de référence tournée vers l'avenir et elle travaille chaque jour pour le démontrer. Son principal défi a été annoncé avec le venue de Google Search Book et le défi de la numérisation et l'indexation de millions de documents qui remplissent nos bibliothèques. La numérisation de la culture est un tournant de notre avenir c'était donc dans la parfaite logique de Google d'en faire partie.

Anticiper nos besoins pour mieux y répondre ça semble difficile et pourtant c'est ce que Google applique au quotidien pour le plaisir de tous les internautes.

VIII. Bibliographie

- Ouvrages

Google Hacks, 100 Industrial Strengh Tips & Tools – Tara Calishain & Rael Dornfest

PageRank Uncovered - Chris Ridings and Mike Shishigin

- Presse

PC Expert – N°161

- Internet

http://www.dicodunet.com
http://www.lesmoteursderecherche.com
http://fr.wikipedia.org
http://www.googleraide.net
http://www.webrankinfo.com
http://www.precisement.org/blog/article.php3?id_article=7
http://www.wcom.fr/informations/google/8008.htm
http://www.01net.com
http://www.zdnet.fr/
http://www.generation-nt.com
http://www.webmaster-hub.com/
http://www.ultra-fluide.com
http://www.infonortics.com
http://www-db.stanford.edu