THE ANATOMY OF A LARGE-SCALE HYPERTEXTUAL WEB SEARCH ENGINE

Sergey Brin and Lawrence Page
{sergey, page}@cs.stanford.edu
Computer Science Department, Stanford University, Stanford, CA 94305

M2 LFC BU Shiqing
Université de Lille

PLAN

- Introduction
- Caractéristiques du système
- Anatomie du système
- Résultats
- Conclusion

INTRODUCTION

- Google, un prototype de moteur de recherche à grande échelle, utilise beaucoup la structure présente dans l'hypertexte. Il est conçu pour analyser et indexer efficacement le Web et produire des résultats de recherche beaucoup plus satisfaisants que les systèmes existants.
- Google, dont l'orthographe est googol, ou 10^{100}, correspond bien à l'objectif de créer des moteurs de recherche à très grande échelle.
- La structure supplémentaire présente dans l'hypertexte est beaucoup utilisée pour fournir des résultats de recherche de meilleure qualité.
- La question : comment construire un système pratique à grande échelle pouvant exploiter les informations supplémentaires présentes dans l'hypertexte ?

- **Moteurs de recherche Web – Mise à l'échelle : 1994 - 2000**
- En 1994, le World Wide Web Worm a eu un index de 110,000 pages Web et documents accessibles. Et en mars et en avril, il a reçu en moyenne environ 1500 requêtes par jour.
- En novembre 1997, Altavista a affirmé qu'il traitait environ 20 millions de requêtes par jour.

- **Google : Mise à l'échelle avec le Web**
- Conçu pour bien s'adapter aux données extrêmement volumineux. Il utilise efficacement l'espace de stockage pour stocker l'index. Ses structures de données sont optimisées pour un accès rapide et efficace.

- **Objectifs de conception**
- Améliorer la qualité de recherche
- Développer et construire un système pour que la plupart de gens puissent faire la recherche académique sur le web à grande échelle

CARACTÉRISTIQUES DU SYSTÈME

- PageRank – mettre les pages Web en ordre
- Utiliser la structure de liens du Web pour calculer un classement de qualité pour chaque page Web

- Texte d'ancre :
- 1. Fournir souvent des descriptions plus précises des pages Web que les pages elles-mêmes.
- 2. Pouvoir exister pour des documents qui ne peuvent pas être indexés par un moteur de recherche textuel : images, programmes et bases de données

CARACTÉRISTIQUES DU SYSTÈME

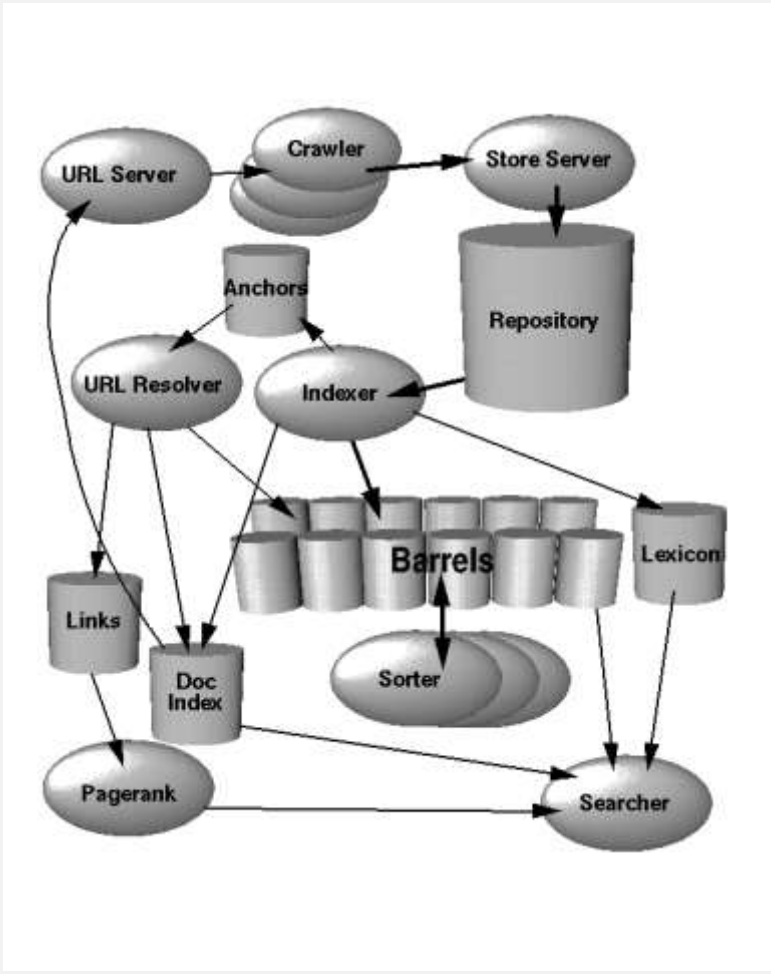
- D’autres:
- Contenir des informations de localisation pour tous les hits et utiliser donc largement la proximité dans la recherche

- Garder la trace de certains détails de la présentation visuelle tels que la taille de la police des mots

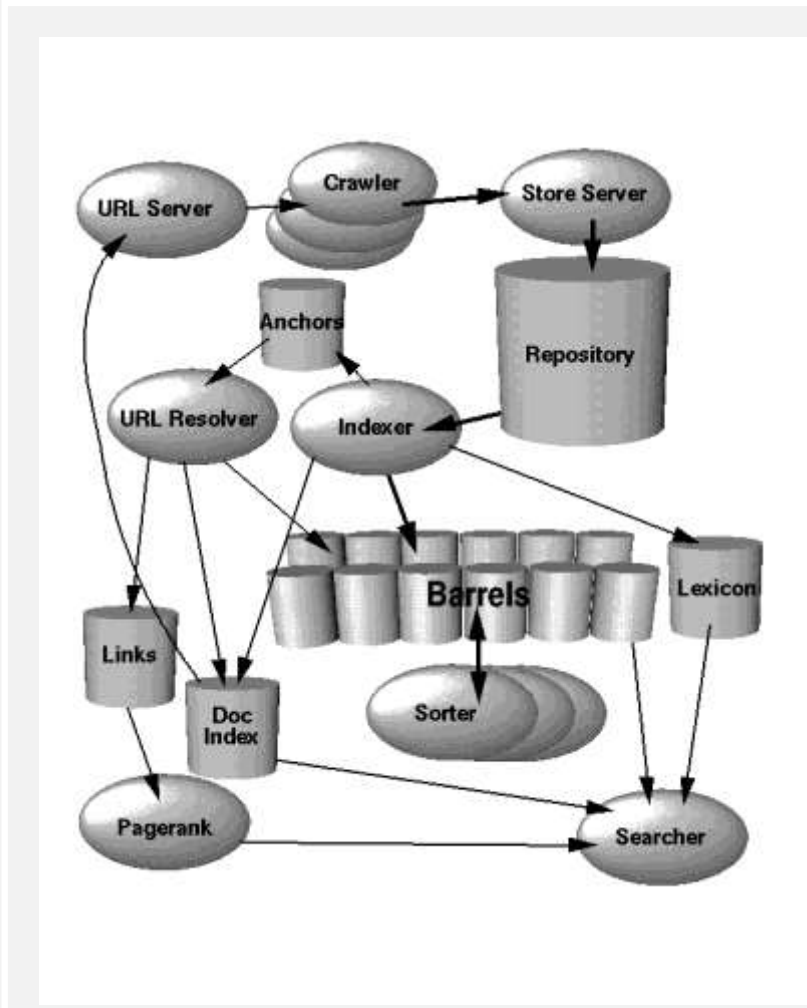
- Le code HTML brut complet des pages est disponible dans un référentiel

ANATOMIE DU SYSTÈME

- La plupart de Google est implémenté en C ou C ++ pour l'efficacité et peut fonctionner sous Solaris ou Linux.
- Serveur URL : envoyer des listes à récupérer aux robots d'exploration
- Serveur de stockage : compresser et stocker les pages Web dans un référentiel



- Indexeur : lire le référentiel, décompresser et analyser les documents, distribuer les résultats dans les « barils », analyser tous les liens et stocker des informations importantes dans un fichier d'ancres
- Résolveur d'URL : lire le fichier d'ancres et convertir les URL relatives en URL absolues et en DocID, placer le texte d'ancre dans l'index direct
- Trieur : prendre les barils et les recourir par wordID, générer les wordID dans l'index inversé



Les structures de données majeures

- Les structures de données de Google sont optimisées pour qu'une grande collection de documents puisse être explorée, indexée et recherchée avec peu de coût. Google est conçu pour éviter les recherches de disque chaque fois que possible, ce qui a eu une influence considérable sur la conception des structures de données.

- BigFiles : des fichiers virtuels couvrant plusieurs systèmes de fichiers et adressables par les entiers 64 bits.
- Référentiel : contient le code HTML complet de chaque page Web.
- Index du document : garde les informations sur chaque document.

Les structures de données majeures

- Index du document : garde les informations sur chaque document.
- Lexique : se présente sous différentes formes : l'un des changements importants par rapport aux systèmes précédents est que le lexique peut tenir dans la mémoire pour un prix raisonnable.
- Liste de hits : correspond à une liste d'occurrences d'un mot particulier dans un document particulier, y compris des informations sur la position, la police et la capitalisation.
- L'index direct : est déjà partiellement trié, stocké dans un certain nombre de barils (nous avons utilisé 64)
- L'index inversé : est constitué de mêmes barils que l'index direct, sauf qu'ils ont été traités par le trieur.

Indexer le Web

Analyse - Tout analyseur conçu pour s'exécuter sur le Web entier doit gérer un grand nombre d'erreurs possibles.

Indexation de documents dans les barils - Une fois chaque document analysé, il est codé dans un nombre des barils.

Tri - Afin de générer l'index inversé, le trieur prend chacun des barils suivants et les trie par wordID pour produire un baril inversé pour les titre et hits d’ancre et un baril inversé en texte intégral.

Evaluation de requête du Google

- 1. Analyser la requête.
- 2. Convertir les mots en wordID.
- 3. Rechercher le début de la doclist dans le baril court pour chaque mot.
- 4. Parcourir les doclists jusqu'à ce qu'un document correspondant à tous les termes de recherche.
- 5. Calculer le rang de ce document pour la requête.
- 6. Si on est dans les barils courts et à la fin d'une doclist, rechercher le début de la doclist dans le baril complet pour chaque mot et passer à l'étape 4.
- 7. Si on n'est pas à la fin d'une doclist, passer à l'étape 4.
- 8. Trier les documents qui ont correspondu par rang et renvoyer le k.

RÉSULTATS

La mesure la plus importante d'un moteur de recherche est la qualité de ses résultats de recherche. Bien qu'une évaluation complète de l'utilisateur dépasse le cadre de cet article, leur propre expérience de Google a montré qu'elle produisait de meilleurs résultats que les principaux moteurs de recherche commerciaux pour la plupart des recherches.

CONCLUSION

Google est conçu pour être un moteur de recherche évolutif. L'objectif principal est de fournir des résultats de recherche de haute qualité sur le World Wide Web en croissance rapide. Google utilise un certain nombre de techniques pour améliorer la qualité de la recherche, notamment le classement des pages, le texte d'ancre et les informations de proximité. De plus, Google est une architecture complète permettant de regrouper, d’indexer et d’exécuter des requêtes de recherche sur des pages Web.

Merci!