The Anatomy of a Large-Scale Hypertextual Web Search Engine

http://www-db.stanford.edu/~backrub/google.html

Sergey Brin and Lawrence Page

Google™

Contents

Google™

Introduction

- 스탠포드 **Ph.D** 학생인 **Larry Page** 와 **Sergey Brin** 이**1998**년 설립
- **2500**만 달러를 공동 출자하여 **1999**년 **6**월 상장 **(Kleiner Perkins Caufield & Byers** 와 **Sequoia Capital**도 출자**)**
- **www.google.com**사이트 서비스를 제공하기 시작하였고, 또한 정보 제공자에게 동일 브랜드의 웹 검색 솔루션을 제공하기 시작

Google™

Introduction (Cont'd)

- **CEO - Eric Schmidt 박사**
- 공동 창립자 겸 사장/제품담당

Larry Page

- 공동 창립자 겸 사장/기술담당

Sergey Brin

Google™

Web Search Engines

1. **Web Search Engines -- Scaling Up: 1994 - 2000**

2. **Google: Scaling with the Web**

    1. 빠른 속도의 크롤링 기술
    2. 저장 공간의 효율적 활용
    3. 인덱싱 시스템의 발전
    4. 빠른 질의어 처리

3. **Design Goals**

    1. 검색 품질 개선 : 기하급수적인 문서 수 증가 - 사람들은 여전히 수십 개의 검색 결과만 보려 함 - 높은 정확률 필요
    2. 관련 학술 연구 : 대용량 웹 데이터 상에서 많은 새로운 연구 활동이 이뤄질 수 있도록 기반 구축

Google™

System Features

1. Page Rank: Bringing Order to the Web

$$PR(A) = (1-d) + d\left(\frac{PR(T1)}{C(T1)} + \frac{PR(T2)}{C(T2)} + .. + \frac{PR(Tn)}{C(Tn)}\right)$$

PR(A) = 페이지 A의 PageRank

$T_1, T_2 ... T_n$ = 페이지 A를 가리키는 다른 페이지

C(A) = 페이지 A에서 밖으로 나가는 링크 개수

d = damping factor (0.85)

Google™

Features (Cont'd)

2. Anchor Text

- 검색 엔진들이 linked text를 link를 담고 있는 해당 페이지와 연관시킴. 구글은 링크가 가리키는 페이지를 연관시킴.

- 장점

① 앵커는 종종 그 링크가 담겨있는 페이지보다 그 링크가 가리키는 페이지에 대한 보다 정확한 설명을 담고 있는 경우가 많다.

② 일반적인 텍스트 검색 엔진이 인덱싱 할 수 없는 이미지나 프로그램, 데이타베이스로의 링크도 존재할 수 있으므로 앵커를 활용하면 실제로 크롤링 되지 않은 웹 페이지들까지 찾아낼 수 있다.

Google™

Features (Cont'd)

3. Other Features

- 모든 hit에 관한 모든 위치 정보 저장 및 검색 시 근접도를 광범위하게 활용한다.
- 단어의 폰트 크기와 같은 시각적인 세부 요소를 추적한다. 폰트 크기가 큰 단어나 볼드체로 된 단어는 그렇지 않은 단어에 비해 더 높은 가중치가 부여된다.
- HTML 문서를 저장하기 때문에 이를 이용할 수 있다.

Google™

Related Work

1. 정보 검색

- 일반적인 정보 검색 연구에서는 잘 통제되어 있는 동질적인 콜렉션 대상 - 과학 논문이나 서로 연관되는 뉴스 기사들
- TREC : 작은 크기의 잘 통제되어 있는 컬렉션을 이용 - 20GB
- 구글 개발 당시 : 2천 400만개의 웹 페이지 - 147GB
- 웹 상에서는 표준 벡터공간 모델을 사용할 수 없음.

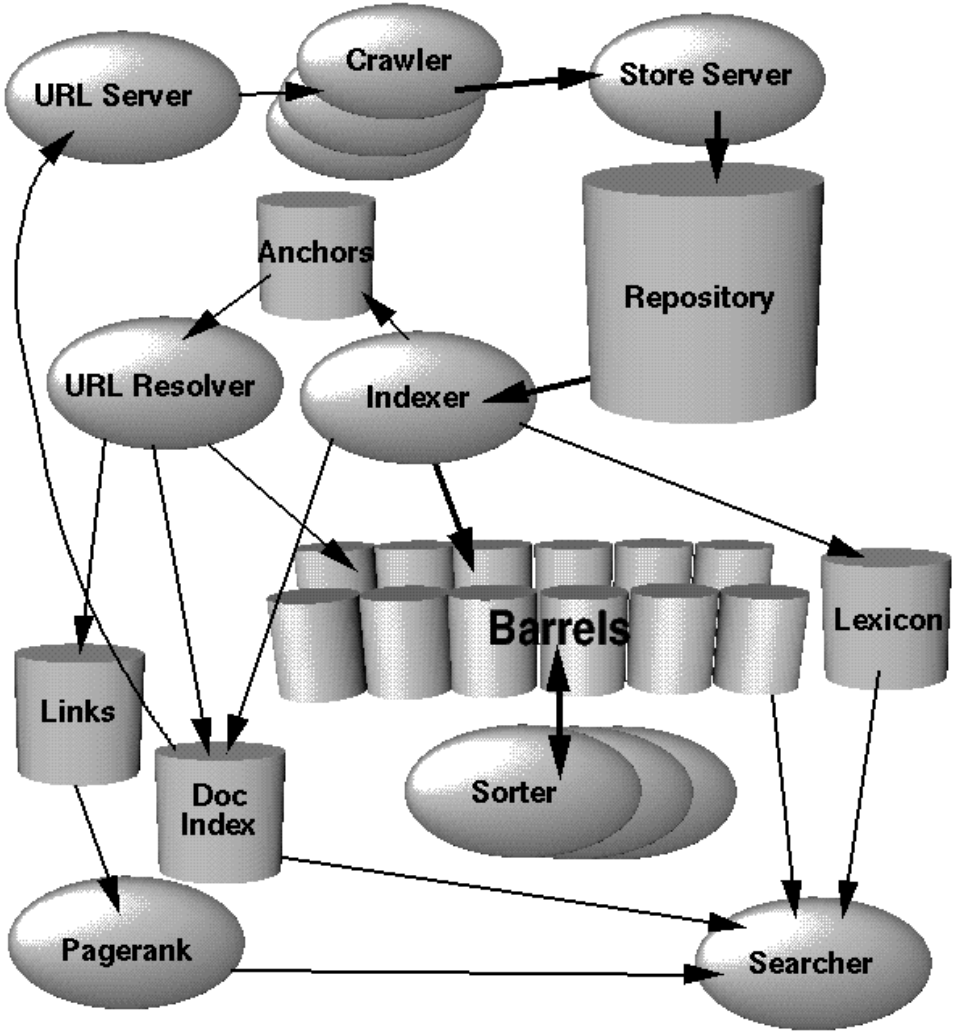
2. 웹 문서와 통제된 컬렉션의 차이점

- 웹 문서는 내부적으로 다양하고 외부적인 메타 정보 역시 다양
- 웹에서는 누가 무엇을 올려 놓을지 조절할 방법이 사실상 없다.

Google™

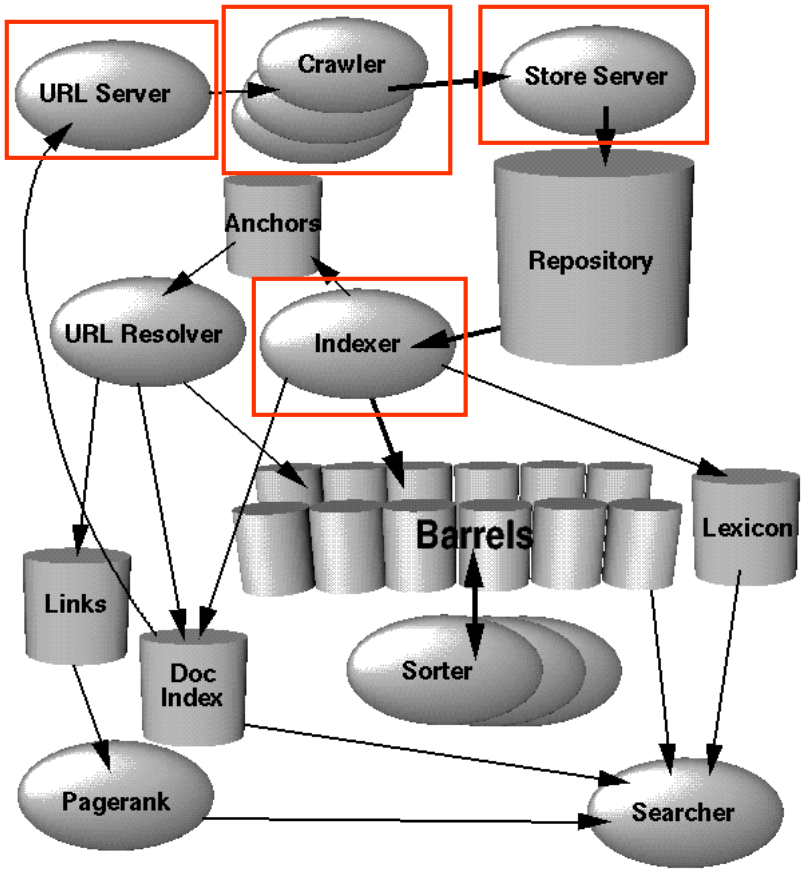
System Anatomy

1. Google Architecture



Google™

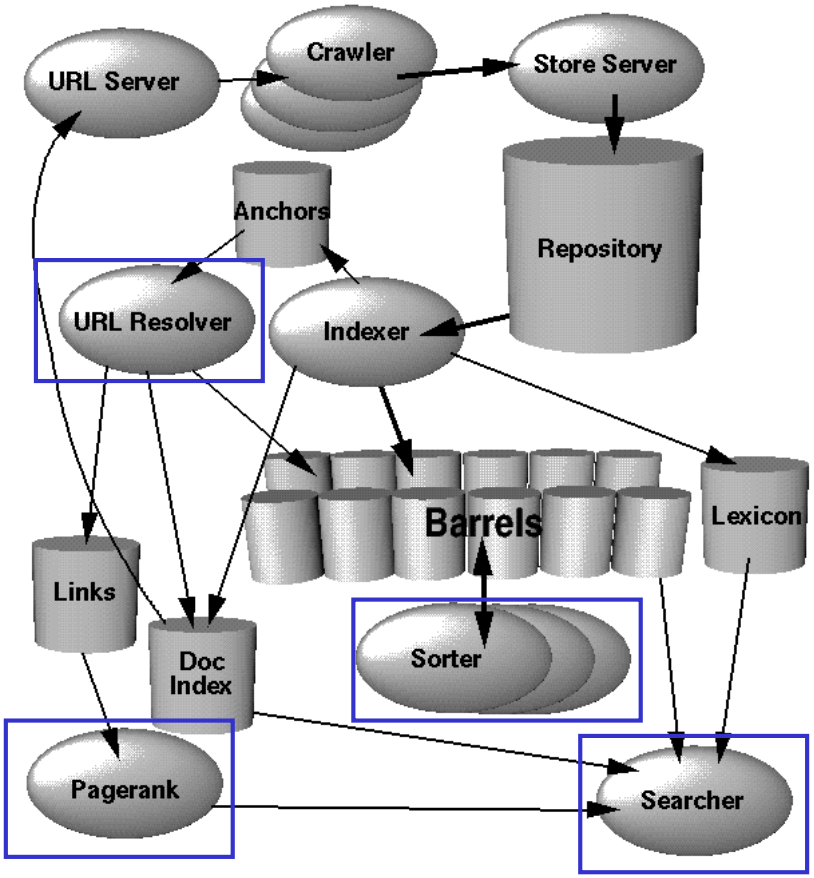
System Anatomy(Cont'd)



1. Crawler : 웹 크롤링
2. URL Server : URL 목록을 Crawler에 알려줌
3. Store Server : 웹 페이지를 압축하여 Repository에 저장 (docID로 관리)
4. Indexer : 웹 페이지 인덱싱, Anchor 파일,Doc Index 생성

Google™

System Anatomy(Cont'd)

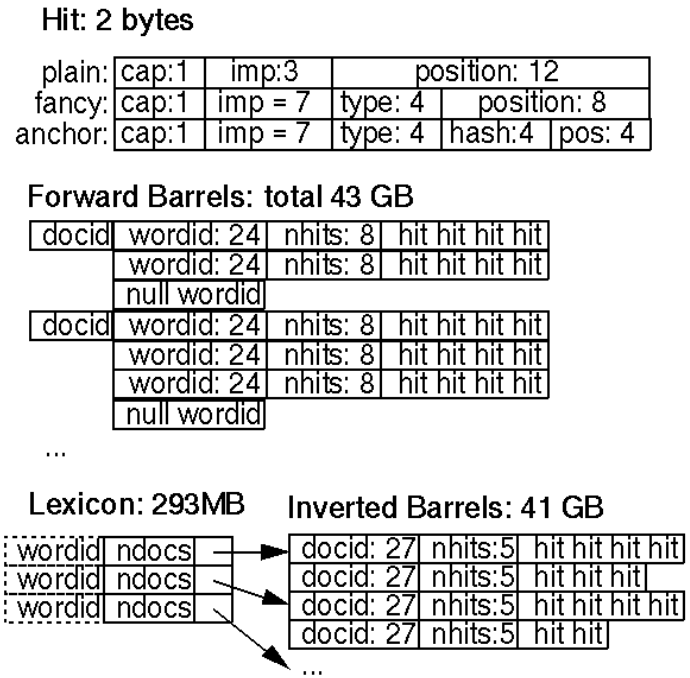


5. Sorter : 역파일 생성
6. URL Resolver : 상대 URL → 절대 URL → docID로 변환
7. PageRank : links를 통해 pagerank 결정
8. Searcher : 검색 및 결과를 사용자에게 보여줌

System Anatomy(Cont'd)

2. 주요 자료구조

1. BigFiles : 64 bit integer 가상 파일
2. Repository : 크롤링된 웹문서의 저장소, zlib을 이용하여 압축
3. Document Index : ISAM index, docID 순서
4. Lexicon : 단어 1400만개
5. Hit Lists : 단어의 위치 정보,
6. 폰트, 대문자 여부
7. Forward Index : docID로 정렬, wordID 포함
8. Inverted Index : wordID로 정렬



System Anatomy(Cont'd)

3. Crawling the Web

- URL Server로부터 URL 목록을 받는 여러 개의 crawler
- 개발 언어 : Python
- 크롤러 동작 시 300개의 연결을 유지하고 초당 100개의 웹문서 수집(초당 600K 의 데이터)
- DNS의 캐쉬를 통한 DNS lookup

4. Indexing the Web

- HTML parsing: lexical analysis by 'flex'
- Indexing Documents into Barrels
  - Convert word into wordID : Lexicon(Memory Hash Table)
- Sorting: Forward Barrel → Inverted Barrel

Google™

System Anatomy(Cont'd)

5. 검색 기법

- 랭킹 - hitlist(폰트,대문자,위치) + 앵커 텍스트 + 페이지 랭크
- 단일 단어 - hitlist 중요시 + 페이지 랭크
- 다수의 단어 – 단어들이 근접해 있는 경우를 가중치 부여

  Proximity + hitlist + 페이지 랭크
- 피드백 – 사용자의 피드백에 따라 랭킹 결정 함수의 인자를 수정할 수 있는 메커니즘

Google™

System Anatomy(Cont'd)

Google Web Server

Google User

1. The web server sends the query to the index servers. The content inside the index servers is similar to the index in the back of a book - it tells which pages contain the words that match the query.

Index Servers

2. The query travels to the doc servers, which actually retrieve the stored documents. Snippets are generated to describe each search result.

Doc Servers

3. The search results are returned to the user in a fraction of a second.

Results

Query: bill clinton

http://www.whitehouse.gov/
100.00% (no date) (0K)
http://www.whitehouse.gov/

- Office of the President
  99.67% (Dec 23 1996) (2K)
  http://www.whitehouse.gov/WH/EOP/OP/html/OP_Home.html
- Welcome To The White House
  99.98% (Nov 09 1997) (5K)
  http://www.whitehouse.gov/WH/Welcome.html
- Send Electronic Mail to the President
  99.86% (Jul 14 1997) (5K)
  http://www.whitehouse.gov/WH/Mail/html/Mail_President.html

mailto:president@whitehouse.gov
99.98%

- mailto:President@whitehouse.gov
  99.27%

The "Unofficial" Bill Clinton
94.06% (Nov 11 1997) (14K)
http://zpub.com/un/un-bc.html

- Bill Clinton Meets The Shrinks
  86.27% (Jun 29 1997) (63K)
  http://zpub.com/un/un-bc9.html

President Bill Clinton - The Dark Side
97.27% (Nov 10 1997) (15K)
http://www.realchange.org/clinton.htm
$3 Bill Clinton
94.73% (no date) (4K) http://www.gatewy.net/~tjohnson/clinton1.html

Google™

Performance

Storage Statistics	
Total Size of Fetched Pages	147.8 GB
Compressed Repository	53.5 GB
Short Inverted Index	4.1 GB
Full Inverted Index	37.2 GB
Lexicon	293 MB
Temporary Anchor Data (not in total)	6.6 GB
Document Index Incl. Variable Width Data	9.7 GB
Links Database	3.9 GB
Total Without Repository	**55.2 GB**
Total With Repository	**108.7 GB**

Web Page Statistics	
Number of Web Pages Fetched	24 million
Number of Urls Seen	76.5 million
Number of Email Addresses	1.7 million
Number of 404's	1.6 million

	Initial Query		Same Query Repeated (IO mostly cached)	
Query	**CPU Time (s)**	**Total Time (s)**	**CPU Time (s)**	**Total Time (s)**
al gore	0.09	2.13	0.06	0.06
vice president	1.77	3.84	1.66	1.80
hard disks	0.25	4.86	0.20	0.24
search engines	1.31	9.63	1.16	1.16

Google™

Conclusions

- **Primary Goal : high quality search results**
  - page rank, anchor text, and proximity information
- **Google is a complete architecture for gathering web pages, indexing them, and performing search queries over them.**

1. **Future Work : query caching, smart disk allocation, subindices**
2. **High Quality Search**
3. **Scalable Architecture**
4. **A Research Tool for I.R.**

Google™

Driving forces

1. 숫자로 본 구글

- 2004년 12월 기준 웹 페이지 색인 숫자는 8,058,044,651 개
- 클러스터 하나당 PC 2,000대
- 30개가 넘는 클러스터 존재
- 104개 언어 지원
- 클러스터당 1 페타 바이트(10^{15} Bytes = 10^{6} GB)
- 클러스터 내부에서 2Gbps로 자료 전송
- 클러스터당 매일 컴퓨터 두 대가 고장 난다.
- 2000년 2월 이후 한번도 심각한 장애가 생긴 적이 없다.
- 단일 클러스터 프로젝트중 최고 인력 투입 : 박사급 200명,기타 600명

Google™

Driving forces (Cont'd)

2. 하드웨어

- 일반 x86 CPU에 표준 IDE를 장착한 일반 PC 사용.
- 평균적으로 하루에 한 대 이상 고장.
- 모든 서버는 복사본이 50개 존재
- 전원선과 랜선만 있으면 자동으로 프로그램과 데이터를 내려 받음.
- 블레이드 서버가 아닌 일반 서버로 500W 전력 정도 소모.

3. 소프트웨어

- GFS(Google File System)을 통해 SCSI, RAID도 아닌 표준 IDE 디스크에 색인 자료를 저장- GFS는 범용 파일 시스템이 아니라 색인 저장을 위한 전용 파일 시스템, 블록 크기 64Mbytes.
- 디버깅을 위한 별도 환경을 갖추고 있음.
- 철자 교정기: 학습을 통해 끊임없이 철자를 학습하는 시스템 갖춤.

Google™

Driving forces (Cont'd)

4. 구글을 사용하는 이유

- 구글은 인터넷에 질서를 수립합니다.
- 구글은 사용자가 80억개 이상의 URL을 검색할 수 있게 해줍니다.
- 구글은 단지 사용자가 원하는 용어를 포함한 페이지만을 보내줍니다
- 구글은 검색어가 있는 페이지의 위치에 많은 중요성이 있다고 봅니다.
- 구글은 페이지의 내용을 미리 보여드립니다.
- 구글은 운이 좋을 것 같은 예감을 불러옵니다.
- 구글은 수집한 웹 문서들을 저장합니다.

→ 혁신적인 검색 기술과 깨끗한 디자인

Google™

Q & A

Google™