비정형 빅데이터 분석의 응용과 실습

Week-06. Search Engine - Part 1

서중원 2020.10.23

구글은 어떻게 만들까?

Google 검색 I'm Feeling Lucky

Google 제공 서비스: English

검색엔진

검색엔진에서의 고려사항

- Scalability
  - 콘텐츠적인 측면에서의 확장성
  - 사용자적인 측면에서의 확장성
- High Quality Results
  - 관련된 콘텐츠인가?
  - 스팸인가?
- Dynamics
  - 하루에 생성되는 웹사이트의 수: *547,200
  - 추가적으로 기존의 웹사이트의 콘텐츠도 업데이트

* https://siteefy.com/how-many-websites-are-there/

검색엔진의 구성요소

1초 안에 양질의 검색결과가 나오기 위해서는?

- Crawling
  - Focused Crawling: 우선순위를 정해 크롤링
  - Deep Crawling: 페이지 안에.. 링크 안에.. 페이지 안에..
- Indexing
  - 웹 페이지를 등록하는 작업
  - 분산처리, Map Reduce?
- Ranking
  - 콘텐츠간의 순위를 매기는 작업
  - 어떤 결과가 더 좋은 결과인가?

검색엔진의 구성요소

크롤링과 스크래핑

- Web Scraping
  - 데이터를 추출(extracting)하는 행위
- Web Crawling
  - 반복적으로 링크를 찾고 데이터를 저장하는 행위
  - 링크를 찾는 과정 또는 데이터를 추출 하는 과정에서 웹 스크래핑 과정이 포함됨

검색엔진의 구성요소

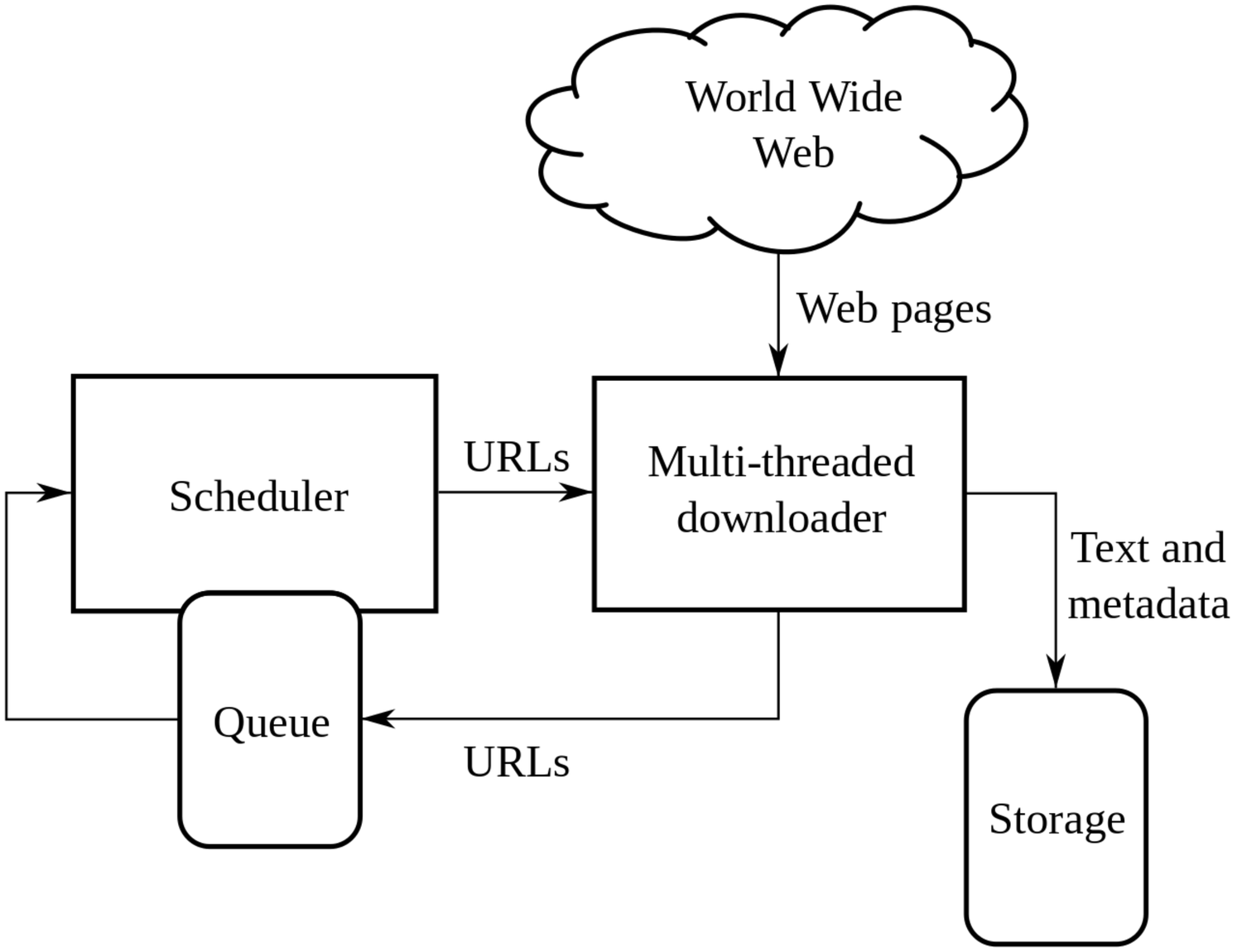
Web Crawling

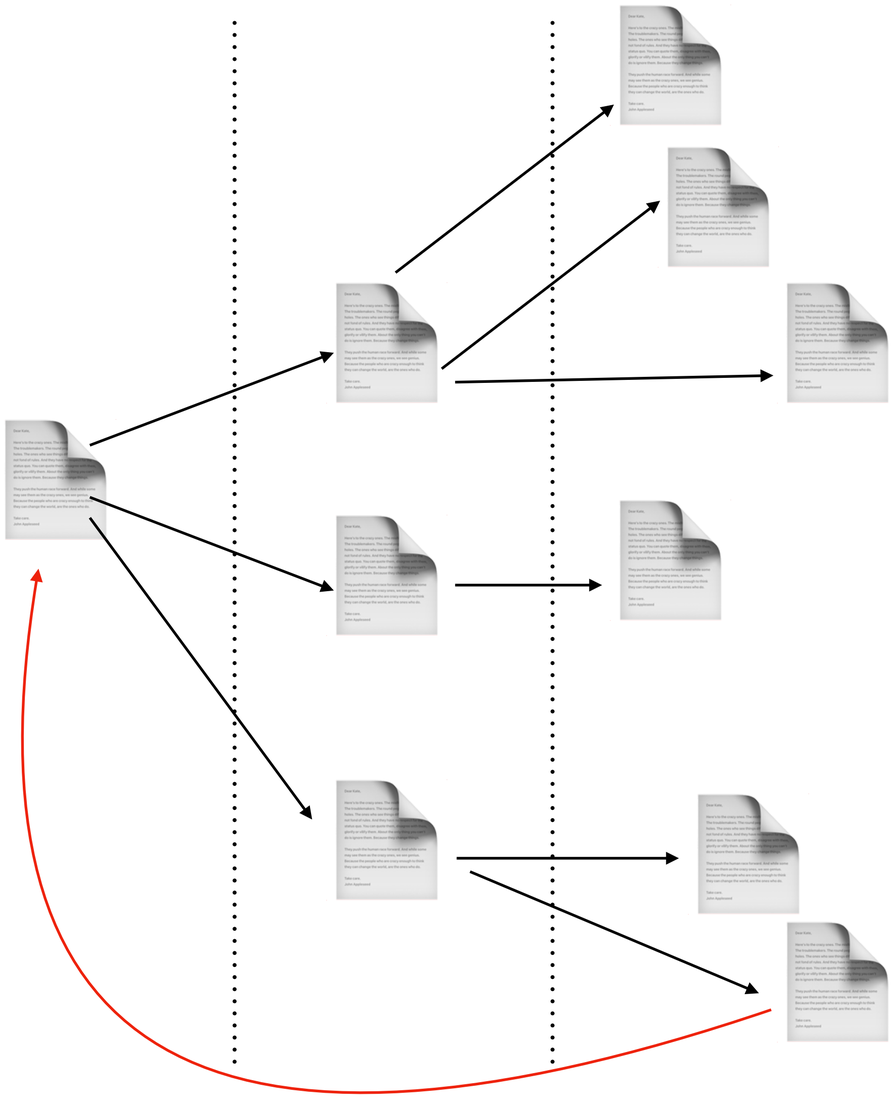
- 웹 페이지를 자동으로 찾고 다운받는 작업
- 웹이라는 것은 거대하고 지속적으로 성장
- 웹이라는 것은 검색엔진 제공자의 통제하에 있지 않음
- 웹 페이지는 지속적으로 변화함
- 크롤러는 다른 타입의 데이터를 활용하기도 해야함

검색엔진의 구성요소

Web Crawling

- Seeds라고도 불리우는 몇몇 페이지를 기준으로 시작
- Seeds는 URL 요청 대기줄에 추가됨
- 크롤러는 페이지들을 URL 요청 대기줄에서 하나씩 빼와서 읽기 시작함
- 다운받아진 페이지는 페이지 내의 링크를 추출하기 위해 파싱
- 추출된 링크들을 URL 요청 대기줄에 추가
- 더 이상 새로운 URL이 없거나, 디스크의 용량이 가득 찼을 때까지 실행





검색엔진의 구성요소

Web Crawling

- 요청을 주고 받는 과정은 많은 시간을 필요로 함
  - 한 페이지당 1초 정도 소요되지만, 검색엔진이 되기 위해서는 엄청난 양의 웹페이지를 인덱싱 하고 있어야 함
- 효율 적으로 크롤링을 진행하기 위해, 멀티 쓰레딩 방식으로 동시에 요청을 보냄
- 이러한 방식은 특정 웹사이트를 마비 시킬 위험이 있음
- 피해를 주지 않기 위해, 동일한 웹사이트에 대한 요청은 인위적인 지연시간을 삽입

크롤링

링크 분석

- 링크는 웹에서 가장 중요한 성분 중 하나
- 이동을 위해서도, 검색을 위해서도 중요

```
<a href="https://yonsei.ac.kr">Yonsei University</a>
```

목적지의 주소 — Anchor text

- Anchor Text와 링크 둘 다 Search Engine에서 사용됨

링크의 중요성

예제

Jungwon's Blog

Coding Blog

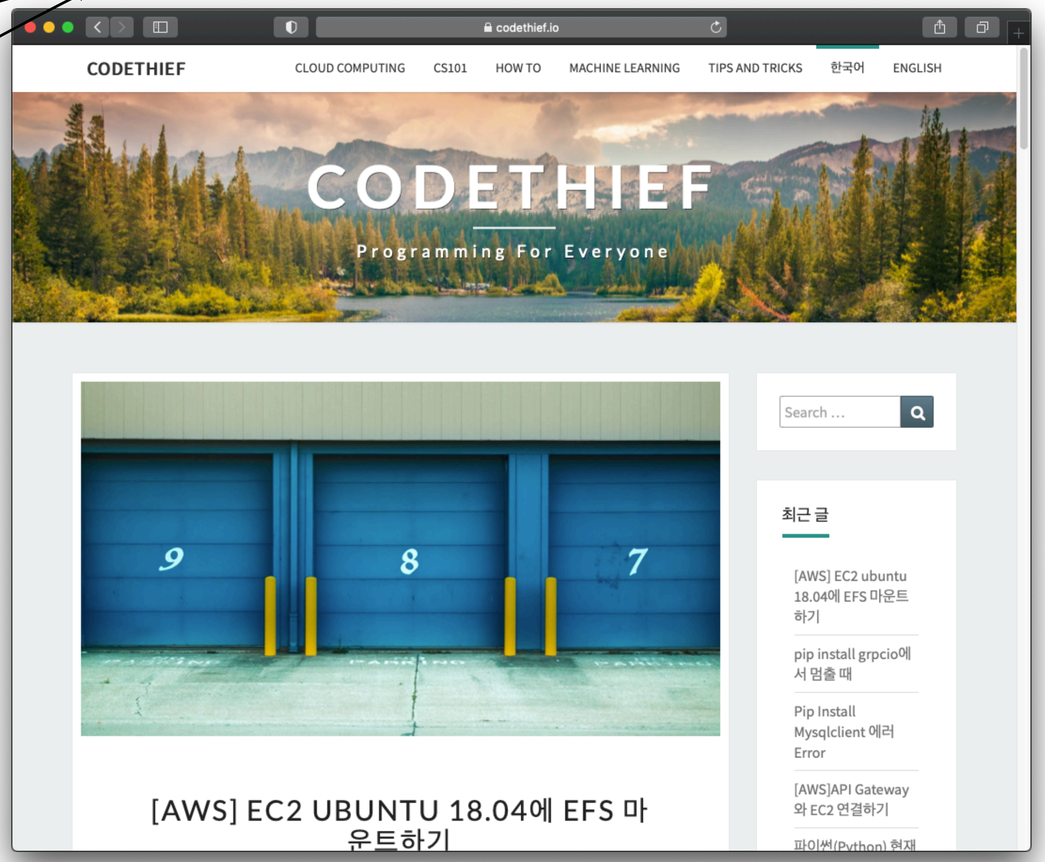
Best Blog

CODETHIEF

```
<a href="https://codethief.io">Jungwon's Blog</a>
```

```
<a href="https://codethief.io">Coding Blog</a>
```

```
<a href="https://codethief.io">Best Blog</a>
```



Fielded Document Representation

문서를 저장할 때, 필드를 구별하여 저장 하는 방식

Title

- Yonsei Big Data 2020

Meta

- Yonsei, University, Bigdata, Machine Learning, 2020

Headings:

- Yonsei Big Data Course 2020 in Sinchon Campuse

Body:

- Yonsei GSI opens new course called “Big Data Analytics Programming” …

Anchors:

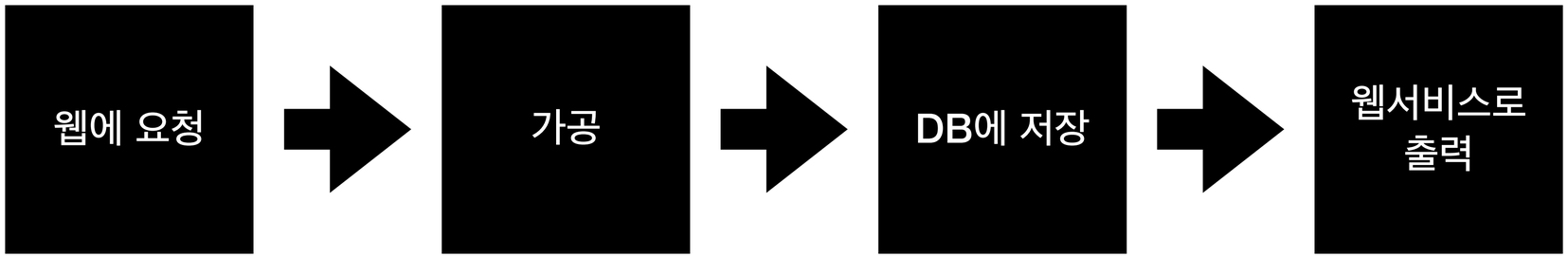
- Coding Blog
- Jungwon’s Blog
- Best Blog

References

Krisztian Balog, DAT630, University of Stavanger, October 23, 2017, https://speakerdeck.com/kbalog/2017-web-search

오늘의 실습!

데이터 확보-저장-출력



E.O.D