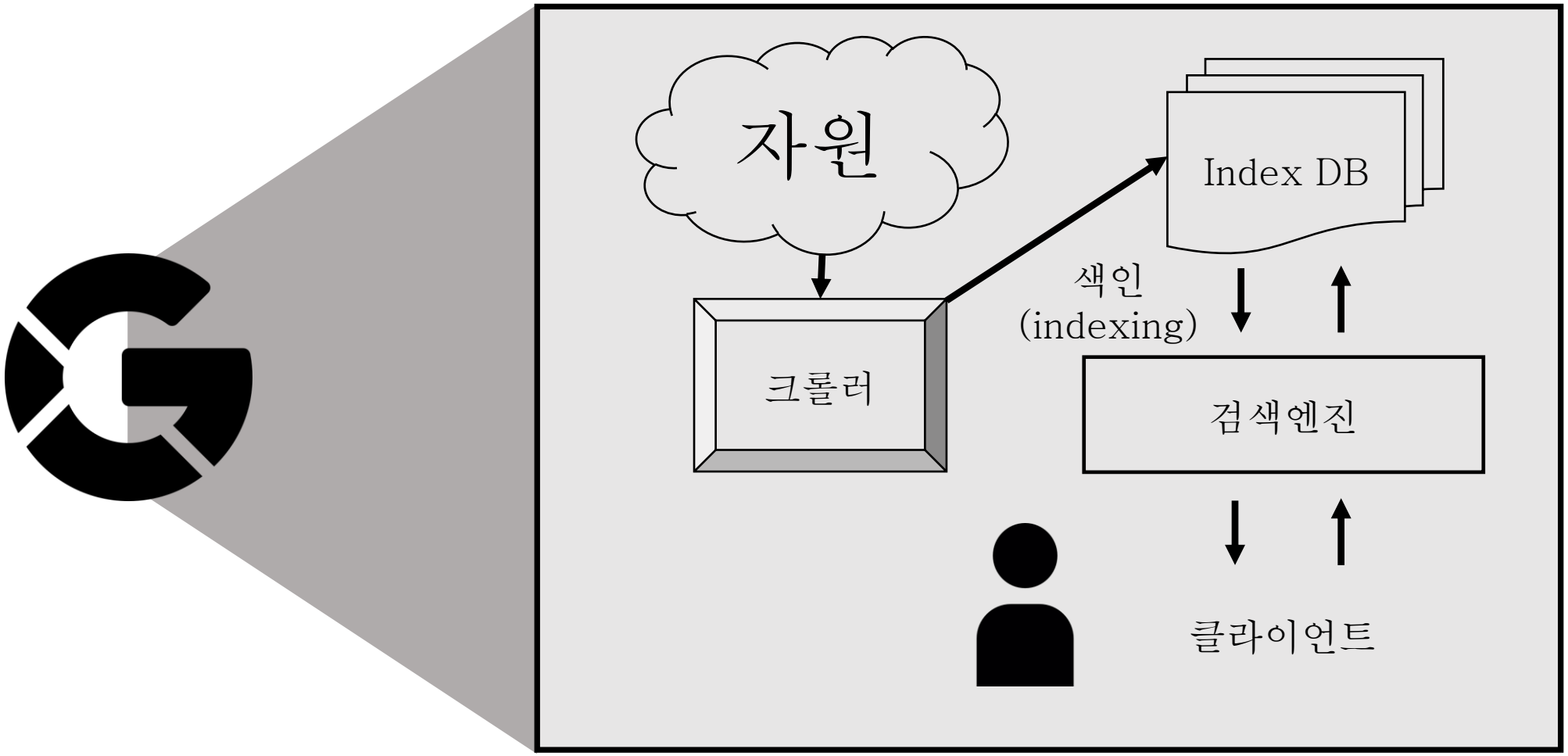
검색엔진과 크롤러

조대인

검색 엔진



* 크롤러 기반 검색엔진 입니다.

검색 엔진의 동작

1. 형태소 분석

무궁화꽃이 피었습니다. ⟶ 무궁화 + 꽃

본래는 '무궁화 꽃'이라 적어야 맞지만 복합 명사의 형태로 적을 때에도 동일한 검색결과를 내야함

신조어의 경우 (ex. 레게노, 억텐 등) 사전에 없는 단어도 올바르게 인식해야함

* 실상 형태소 분석이 검색엔진 구현의 제일 난관이라고 볼 수 있다.

2. 색인

색인은 단어와 그 주소를 Key와 Value의 자료구조로 묶어서 저장한다.

'구름'이라는 단어가 있다면 '구름'에 대한 문서를 나타내는 주소를 다음과 같이 나열할 수 있다.

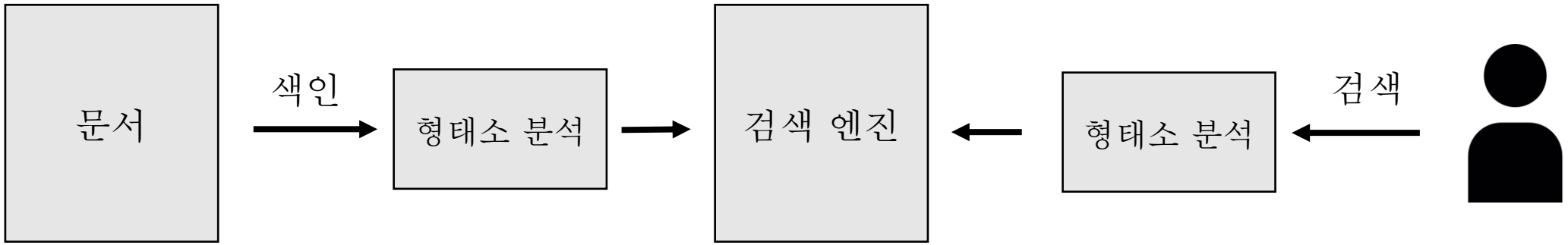
구름 : 100201, 101020, 120320, … , 458220

올바르게 색인되었다면 보다 빠르게 해당 단어가 들어간 문서를 찾을 수 있다.

3. 검색

질의(Query)를 분석해 원하는 결과를 보여준다.

질의를 분석할 때도 형태소 분석기를 사용한다.



https://needjarvis.tistory.com/167?category=704321

일부 검색엔진의 특수한 기능

구글 검색엔진에서의 검색 연산자

논리 연산자 : and, or, not

범위 검색 : 1..100 (1부터 100까지의 범위를 뜻함)

소셜미디어 검색 : @facebook

단어 제외 : - (-기호 뒤에 나오는 단어를 제외함)

정확히 일치 : “”

Cache: site: related: filetype:

강력한 검색기능으로 인한 피해 사례

검색엔진 구글 '해킹악용' 논란

인터넷 검색엔진 구글(Google)이 웹사이트 해킹 도구로 악용돼 논란이 일고 있다.

워싱턴포스트(WP)는 10일 "구글 사이트에서 간단한 검색명만 입력하면 주민등록번호 은행계좌번호 병원진료기록 학교성적표 등 개인정보는 물론 미 해군 잠수함의 현 위치까지 파악할 수 있다"며 이같이 보도했다.

실제 인터넷에서 구글을 이용,개인신상 정보를 얻어내기란 의외로 간단하다.

검색창에 'xls''cc''ssn' 등을 입력한 뒤 검색어를 조합하면 신용카드번호 주민등록번호 등을 손쉽게 해킹할 수 있다.

검색어로 전체를 뜻하는 'total'을 쳐넣으면 개인의 재무정보도 빼내올 수 있다.

1만여개의 구글 컴퓨터는 2주일에 한 차례씩 전세계 30억개의 웹사이트 및 서버를 돌며 새로운 정보를 '복사(crawl)'하고 있어,정보 수집능력이 엄청나다.

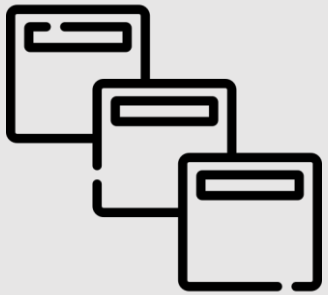
인터넷에 한 번이라도 올라간 적이 있는 정보는 삭제되더라도 구글 컴퓨터에는 남아 있어 인터넷을 떠돌게 된다.

보안업체인 INS의 에드워즈 스코우디스 컨설턴트는 "문제는 합법적인 검색엔진 구글을 이용한 정보수집이 불법이 아니라는 데 있다"고 말했다.

https://www.hankyung.com/international/article/2004021094551

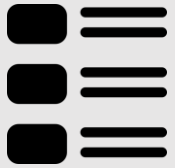
크롤러

1. 시드에서 URL을 큐에 저장한다.



2. URL을 재귀적으로 방문하며 큐에 저장한다.

3. 크롤링하면서 얻은 텍스트와 메타데이터를 데이터베이스에 저장한다.



4. 이를 시스템에서 인덱싱(색인)한다.

크롤링의 문제

Scale : 웹이 점점 방대해지는 문제

Content Selection Tradeoffs : 어떤 컨텐츠가 의미 있는지에 관한 문제

Social Obligations : 크롤링 하는 웹사이트에 부담을 주는 문제

Adversaries : 웹사이트 제공자가 크롤러에 악의적인 콘텐츠를 주는 문제

-> 이를 해결하기 위한 크롤링 방법이 필요

Robots.txt

*Robots.txt에 대한 오해
Robots.txt는 사이트 크롤링을 원천 차단하는 기능이 없다.

Robots.txt의 기능
1. (검색엔진) 크롤러에게 엑세스 할 수 있는 파일이 무엇인지 알려준다.
2. 동영상, 음성 파일 등이 검색 결과에 표시되는 것을 방지할 수 있다.
3. 중요하지 않은 리소스 파일을 차단할 수 있다.

특이사항
- Robots.txt를 해석하지 않는 크롤러도 있고, 또 크롤러마다 해석하는 문법이 각자 다르다.
- Robots.txt를 통해 접근을 차단한 경우에는 웹상에 다른 곳에 해당 URL이 있는 경우 크롤러가 색인을 찾아 생성할 수 있다. 따라서 완전히 차단하고 싶을 경우에는 메타태그에 noindex를 사용하거나 비밀번호를 사용해야 한다.
- 또 크롤러가 Robots.txt를 해석하기 위해서는 가장 상위폴더에 Robots.txt가 위치해야 한다.
- 유해한 크롤러의 경우 Robots.txt를 해석하지 않을 수도 있다.

출처: 구글 검색 센터

Robots.txt의 예시

```
User-agent: Googlebot          #이름이 Googlebot인 사용자 에이전트는 /nogooglebot/ 으로
Disallow: /nogooglebot/        시작하는 문서를 크롤링 할 수 없습니다.

 User-agent: *
Allow: /                       #그 외 모든 사용자는 전체 사이트를 크롤링 할 수 있습니다.

Sitemap: http://www.example.com/sitemap.xml

                               #사이트맵 (크롤러에게 해당 사이트에 있는 페이지, 동영상 등의
                               정보를 제공하는 파일)은 해당 경로에 있습니다.
```

알려진 크롤러

Google bot

Bing bot

Yandex bot

Baidu Spider

이러한 크롤러들은 Robots.txt에 명시된 내용을 준수하며 타사의 검색 데이터를 수집하지 않는다.

악성 크롤러

보통 사이트 관리자는 Robots.txt 문서를 통해 사이트 크롤링에 관한 권한을 명시함.

악성 크롤러는 이 규칙을 무시하고 민감한 정보까지 저장하는 크롤러이다.

또한, 악성 크롤러는 대상 사이트에 걸리는 부하를 신경 쓰지 않기도 하며 크롤링한 정보를

개인 또는 단체의 이익을 위해 사용하기도 한다.

스크래핑? 크롤링?

크롤링으로 인한 법정 분쟁 사례

잡코리아, 무단 크롤링한 사람인HR에 승소…손해배상금 등 4억5000만원 받는다

양사간 다툼은 9년 전인 2008년 사람인이 잡코리아에 등록된 기업 채용공고를 무단 크롤링해 게재한 것이 발단이 됐다.

2010년 잡코리아는 사람인HR을 상대로 법원에 채용정보 복제 금지 가처분을 신청했고, 2011년 서울중앙지방법원은 사람인HR이 잡코리아의 채용정보를 무단으로 게재하지 말라고 강제조정 결정을 내렸다.

법원의 강제조정 후 사람인HR은 재발 방지를 약속하는 등 사건이 일단락되는 듯 보였으나, 이후에도 사람인은 검색로봇을 이용해 같은 방식으로 채용정보를 무단 복제해 자신이 운영하는 웹사이트에 게재했다.

→ IP를 차단했음에도 VPN을 이용해 계속 스크래핑을 시도했다고 함.

서울고법은 잡코리아의 주장을 받아들여 DB 권리를 침해 받았다고 판단했다.

재판부는 "원고는 DB에 해당하는 원고 웹사이트를 제작하기 위해 인적·물적으로 상당한 투자를 했고, 그 소재의 갱신·검증 또는 보충을 위해 인적·물적으로 상당한 투자를 한 자이므로 원고 사이트에 대한 DB 제작자에 해당한다"며 "피고의 이 사건 게재 행위에 의해 저작권법 제93조 제2항, 제1항에서 정하고 있는 원고의 DB 제작자 권리가 침해됐다고 봄이 타당하다"고 밝혔다.

또한 재판부는 "피고인 사람인HR은 잡코리아 웹사이트의 채용 정보를 모두 폐기할 의무가 있다"며 "조정조서 위반으로 인한 간접강제금 2억원과 DB 권리 침해로 인한 손해배상금 2억5000만원을 합해 총 4억5000만원을 잡코리아에 지급하라"고 판결했다.

http://it.chosun.com/site/data/html_dir/2017/09/27/2017092785016.html

크롤링으로 인한 법정 분쟁 사례

法 "여기어때, 야놀자 정보 무단수집 맞다"...前 대표 '유죄'

서울중앙지법 형사5단독 신민석 판사는 11일 정보통신망 이용촉진 및 정보보호 등에 관한 법률 위반(정보통신망침해 등)과 저작권법 위반, 컴퓨터 등 장애 업무방해 혐의로 기소된 심 전 대표에게 징역 1년6개월에 집행유예 2년, 사회봉사 160시간을 선고했다.

위드이노베이션은 2016년 1월부터 10월 초까지 야놀자 제휴점수 등 정보를 취합하기 위한 크롤링(수집) 프로그램을 개발했다.

이를 이용해 야놀자의 모바일앱 용 API 서버에 총 1천594만여회 이상 침입한 혐의를 받았다. 이 같은 수법으로 야놀자의 제휴숙박업소 업체명, 주소, 원래 금액, 할인 금액 등 정보를 264여회 걸쳐 무단 복제한 것으로 조사됐다.

신민석 판사는 "여기어때는 야놀자와의 경쟁관계에서 우위를 점하기 위해 상당 기간 크롤링 프로그램을 이용해 서버에 침입, 숙박업소에 관한 각종 정보를 복제했다"면서 "이에 야놀자는 경쟁력 저하, 비밀 유출 등 상당한 피해를 입었을 가능성이 있지만, 여기어때는 피해 회복을 위한 노력을 하지 않았다"고 지적했다.

https://zdnet.co.kr/view/?no=20200211153634