제4장 검색엔진

4.1 검색엔진 등장하다

- 우리는 매일 인터넷 검색을 하며 생활한다. 궁금한 게 있으면 언제든, 무엇이든 검색창에 검색한다 . 하루에도 몇 번씩 검색 서비스에 쿼리(Query)를 날린다. 이제 검색이 없는 세상은 상상할 수가 없다.

검색은 현대인의 분실에 꼭 필요한 정보를 찾아주는 가장 핵심적인 역할을 담당하고 있다. 또한 검색은 빠르고 언제든 필요한 문서를 순식간에 찾아낸다.

4.2 엄청난 돈을 벌어들이다

- 검색 기능만으로 이토록 엄청난 수익을 낼 수 있을까?

검색엔진이 시장의 흐름을 주도하면서 본격적으로 검색광고를 도입한다. 항상 동일한 광고가 노출되는 게 아닌 쿼리에 적합한 광고를 매번 다르게 보여주는 타깃 마케팅을 진행하고, 사용자 피드백을 기반으로 광고료를 산정 하는 CPC 방식 Cost Per Click을 도입.

대표적인 사용자의 피드백이 클릭이다. 사용자의 클릭에 따라 광고료를 매기기 때문에 이제 시스템은 더 정교해져야 했다. 그 결과 구글은 2024년 상반기 기준 시가총액 3,300조 원이 넘는 세계 4위기업 으로 성장했다

4.3 엄청난 문서를 수집하다

- 검색엔진이 인터넷에 있는 문서를 수집하여 검색에 적합하도록 보관하고 있는 것을 색인(Index)이라고 한다.

구글은 많은 문서를 대체 어디에 보관하고 있을까

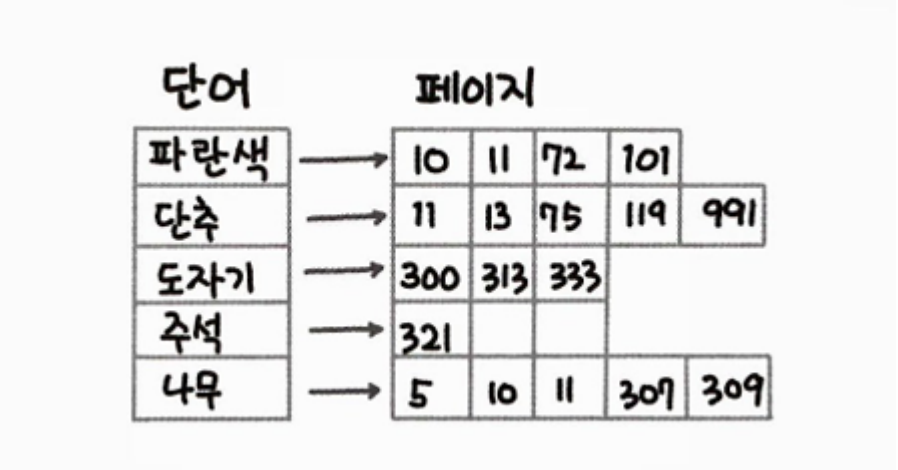
저렴한 컴퓨터 수백, 수천 대에 나눠서 저장하는 방식을 택했다. 이를 위해 구글 파일 시스템 Google File System, GFS이라는 효율적인 분산 파일 시스템을 만들어냈고 덕분에 아무리 큰 파일도 여러 대의 서버에 나누어 저렴한 비용으로 저장할 수 있게 됐다

문서를 수집하려면 웹사이트를 구석구석 돌아다녀야 하는데, 웹 문서를 갈고리처럼 긁어온다고 해서 크롤러(Crawler)라고 한다.

검색의 첫 번째 작업을 담당하는 크롤러는 매우 정교해야 한다. 한번 방문한 사이트는 일정 시간 재방문 하지 않는 정책이 필요하고. 스케줄러가 이런 정책을관리한다.

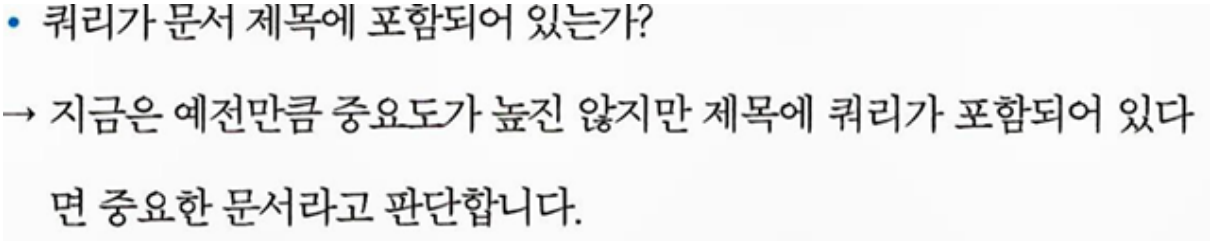
4.4 검색엔진은 어떻게 검색할까

-책의 색인처럼 항목을 먼저 정 리해두는 과정이 필요한데, 이 과정을 색인 구축 과정이라고 한다.



4.5 랭킹, 수십 조 가치의 줄 세우기 기술

- 검색 랭킹을 매기기 위해서는 어떻게 점수를 정해야 합리적일까? 검색엔진이 문서의 점수를 정하는 과정은 저마다 다른 특성을 지닌 문서를 비교하는 과정과 유사하다. 단순히 한 두 가지 조건이 아니라 종합적인 면을 모두 고려 해야한다. 구글의 경우 약 200여 가지의 랭킹 조건을 이용한다고 알려졌다.



- 긴 문서인가?

→ 짧은 글보다는 긴 글이 품질 점수가 높습니다.

- 문서 로딩이 빠른가?

→ 빠른 문서는 더 좋은 경험을 줍니다.

- 사이트에 접속할 수 없는 상황이 자주 발생하는가?

→ 빠른 로딩과 함께 사이트의 안정성은 매우 중요합니다. 자주 다운되는 사이트에 있는 문서라면 아무리 내용이 좋아도 문서를 보기가 어렵겠지요.

- 모바일에서 잘 보이는가?

→ 이제 모바일 인터넷 트래픽이 PC를 앞질렀습니다. 문서가 모바일에서 잘 보이느냐는 매우 중요한 사항입니다.

- 문서 내에 쿼리가 많이 포함되어 있는가?

→ 딱 한 번 나오는 것보다는 여러 번 반복해서 나오는 게 좋겠죠. 이 개념은 유사도를 판별하는 TF-IDF 알고리즘의 바탕이기도 합니다. 이후에 다시 설명하겠습니다.

- 동일한 사이트에 중복으로 나오는 콘텐츠인가?

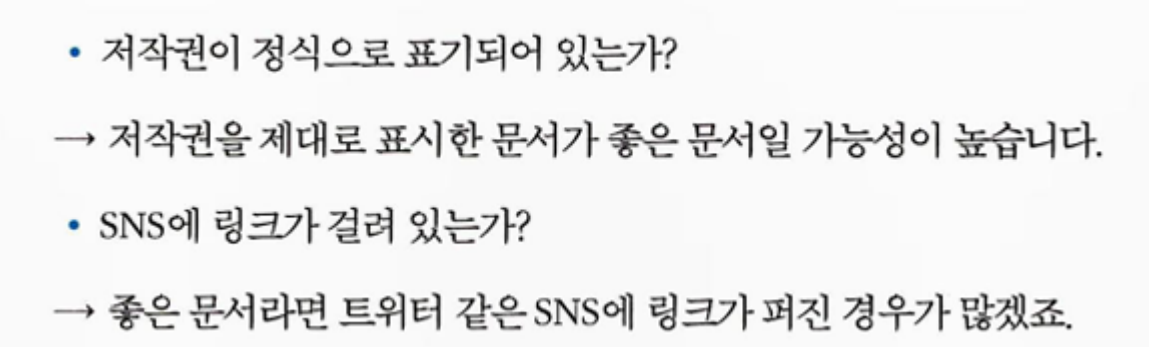
→ 긴 문서로 만들기 위해 불필요하게 내용을 반복하는 경우가 있습니다. 당연히 감점 요인입니다.

- 다른 문서에서 복사한 내용인가?

→ 흔히 불펌이라고 하죠. 당연히 불펌한 문서는 점수가 낮고, 원본이 더 높은 점수를 받아야겠죠. 이 부분도 판단하여 점수에 포함합니다.

- 본문 내용의 수준이 지적인 내용인가, 욕설로 가득한 내용인가?

→ 글의 품위도 평가 기준으로 삼습니다.



4.6 품질 좋은 문서를 찾아서

- 품질이 좋은 문서란 어떤 문서일까

품질이 좋은 문서란 검색 쿼리 에 관계없이 항상 좋은 문서를 말한다. 또한 권위 있는 사이트의 문서는 좋은 품질의 문서로 볼 수 있다

4.7 페이지 랭크, 구글의 역사가 탄생하다

-유명한 사이트가 많이 가리킬수록 문서의 점수가 올라가는 알고리즘으로, 좋은 논문은 인용 횟수가 많다는 아이디어에서 출발했고, 여기에 에르되시 수가 낮을수록 권위가 높아진 것 처럼 권위 있는 사이트에 가중치를 높였다. 이 알고리즘의 이름은 페이지 랭크(Page Rank)입니다

권위 있는 사이트가 많이 참조할수록 순위가 올라가는 구조 덕분에 권위가 없는, 이른바 스팸 사이트는 아무리 링크를 늘려봐야 순위에 오를 수 없게 됐다. 점차 검색 엔진에서 스팸 이 사라지기 시작했고, 높은 검색 품질을 무기로 구글은 검색 시장을 하나둘 점령하기 시작했다.

4.8 쿼리에 딱 맞는 문서 찾는 법

-어떻게 쿼리에 딱 맞는 문서를 불러올 수 있을까

사용자가 입력한 쿼리가 문서 어디쯤에 위치하느냐가 중요하다 제목에 위치하는지 또는 본문에 위치하는지, 아무래도 제목이 훨씬 더 중요하므로 제목에 위치하는 경우 더 높은 점수를 준다

또한 순서대로 매칭되었는지도 중요하다 검색 분야에서는 이를 근접도Proximity라고 하며, 단어와 단어 사이의 간격이 좁을수록 더 유사한 문서라고 판단하고 높은 점수를 줍니다.

4.9 검색엔진 최적화, 창과 방패의 싸움

-구글이나 네이버의 검색 결과에서 상위를 차지하면 엄청난 트래픽을 가져올 수 있다

그래서 검색엔진 최적화Search Engine Optimization, SEO를 시도하는 업체들은 여러 가지 실험을 해보면서 랭킹을 높이기 위해 끊임없이 도전한다

구글에는 200여 가지의 랭킹 조건이 있는데, 검색엔진 최적화는 이들 조건 사이에서 바늘구멍 같은 빈틈을 찾아 랭킹을 올리기 위해 끊임없이 노력한다

4.10 점점 더 똑똑해지는 구글 검색의 진화

-원래 검색엔진의 역할은 쿼리에 정확하게 매칭하는 문서를 찾아주는 것이다

구글을 비롯한 요즘의 검색엔진은 충분히 이 역할을 해내고 있다. 하지 만 사람들은 점점 더 똑똑한 검색 엔진을 원하고 있다. 이제는 쿼리의 맥락을 파악하여 적절한 문서를 제시해주는 수준에 이르렀다

검색은 점점 더 문서, 비디오, 이미지 같은 형식의 경계, 언어의 경계를 무너트리고 복잡한 질문을 이해하며 방대한 정보를 바탕으로 추천까지 하는 진정한 개인 비서의 역할을 톡톡히 해내고 있다. 앞으로도 검색은 인공지능 시대의 핵심 서비스로 계속해서 자리매김할 것이다