4장 검색

1. 검색엔진의 본질과 역사적 변화

현대의 검색엔진은 과거의 메모장을 넘어 인간의 기억력을 보조하는 **'뇌의 방(뇌수의 분실)'** 역할을 수행하며, 방대한 인터넷 공간에서 필요한 문서를 순식간에 골라줍니다.

- **디렉토리 서비스(과거):** 1990년대 초반 야후!(Yahoo!)가 대표적이었으며, 사람이 직접 웹사이트를 주제별로 분류했습니다. 하지만 웹사이트 수가 기하급수적으로 늘어나면서 사람이 수동으로 정리하는 데 한계가 발생했습니다.
- **검색 알고리즘(현재):** 구글과 같은 기업이 등장하며 컴퓨터 알고리즘이 자동으로 정보를 찾는 시대가 열렸습니다. 구글은 현재 전 세계 검색 시장의 90% 이상을 점유하는 기업으로 성장했습니다.

2. 검색엔진의 3단계 작동 원리

검색엔진은 크게 **수집, 색인, 랭킹**이라는 세 가지 과정을 거쳐 사용자에게 결과를 보여줍니다.

① 수집 (Crawling)

전 세계 웹사이트를 돌아다니며 정보를 긁어모으는 단계입니다.

- **전문 용어 - 크롤러(Crawler):** 웹페이지의 링크를 타고 다니며 문서를 수집하는 소프트웨어 로봇입니다. 거미줄(Web)을 타고 다닌다고 하여 '스파이더'라고도 불립니다.
- **부가 설명:** 크롤러는 방문한 사이트의 URL을 '큐(Queue)'라는 목록에 저장하고, 새로운 링크를 계속 찾아내며 수집을 반복합니다.

② 색인 (Indexing)

수집한 방대한 데이터를 검색하기 좋게 정리하여 보관하는 단계입니다.

- **전문 용어 - 색인(Index):** 책의 맨 뒤에 있는 '찾아보기'와 같습니다. 특정 키워드가 몇 페이지(어느 웹사이트)에 있는지 미리 정리해두는 작업입니다.
- **부가 설명:** 구글은 현재 수백 조 개의 문서를 색인하고 있으며, 이를 일반 PC 수천 대에 분산 저장하는 기술을 사용합니다.

③ 랭킹 (Ranking)

수천만 개의 검색 결과 중 어떤 것을 가장 먼저 보여줄지 순위를 매기는 과정입니다.

- **전문 용어 - 페이지랭크(PageRank):** 얼마나 권위 있는 사이트로부터 링크(인용)를 많이 받았느냐에 따라 해당 문서의 점수를 매기는 알고리즘입니다.

- **전문 용어 - 댐핑 팩터(Damping Factor):** 사용자가 링크를 타고 이동하다가 어느 순간 클릭을 멈추고 사이트를 나갈 확률을 계산에 반영한 수치입니다(보통 0.85로 설정).

3. 검색 결과의 품질을 결정하는 기술

단순한 키워드 매칭을 넘어 사용자가 정말 만족할 만한 정보를 찾기 위한 고도의 계산식이 사용됩니다.

- **TF-IDF (단어 빈도-역문서 빈도):**

◦ **TF(Term Frequency):** 특정 단어가 문서 내에 얼마나 자주 등장하는지 측정합니다.

◦ **IDF(Inverse Document Frequency):** 해당 단어가 다른 문서들에는 얼마나 희귀하게 등장하는지 측정합니다. 흔한 단어(예: '의', '가')보다 희귀한 단어가 포함된 문서에 더 높은 점수를 줍니다.

- **BM25 (Best Matching 25):**

◦ TF-IDF를 발전시킨 모델로, 현재 대부분의 검색엔진이 채택하고 있습니다.

◦ 문서의 길이를 고려하여, 짧은 문서에서 키워드가 자주 나오는 경우에 더 높은 가중치를 줍니다.

4. 검색엔진은 어떻게 진화하고 있는가?

AI와 딥러닝의 결합

이제 검색엔진은 단순한 단어 비교를 넘어 **문맥**을 이해합니다.

- **오타 교정:** 사용자가 '네바시'라고 검색해도 딥러닝을 통해 '세바시'로 이해하고 올바른 결과를 제안합니다.
- **MUM (Multitask Unified Model):** 구글의 최신 기술로, 75개 이상의 언어를 통합 처리하며 텍스트뿐만 아니라 이미지, 영상 정보까지 동시에 이해하여 복잡한 질문에 답합니다.

사용자 최적화와 테스트

- **전문 용어 - A/B 테스트:** 두 가지 버전의 검색 결과(A안, B안)를 사용자에게 무작위로 보여주고, 어떤 쪽의 클릭률이나 재방문율이 높은지 비교하여 시스템을 개선하는 방식입니다.

5. 비즈니스 모델: 검색 광고

검색엔진은 사용자의 의도(쿼리)에 맞는 광고를 노출하여 수익을 얻습니다.

- **전문 용어 - CPC(Cost Per Click):** 광고가 노출될 때는 비용을 받지 않고, 사용자가 실제로 광고를 **클릭했을 때만** 광고주가 비용을 지불하는 방식입니다.
- **광고 랭킹:** 단순히 돈을 많이 낸 광고를 상단에 띄우는 것이 아니라, 클릭률이 높고 사용자에게 유용한 광고를 우선 노출하는 정교한 경매 시스템을 사용합니다.

"검색엔진은 웹상의 거대한 도서관이고, 크롤러는 사서이며, 색인은 도서 목록집" 이다.