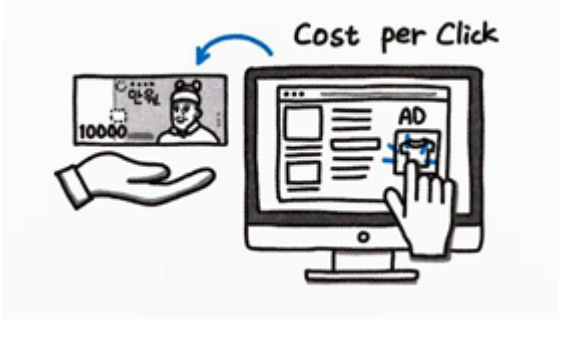
제 4장. 검색엔진

검색엔진이 돈을 벌어들이다?

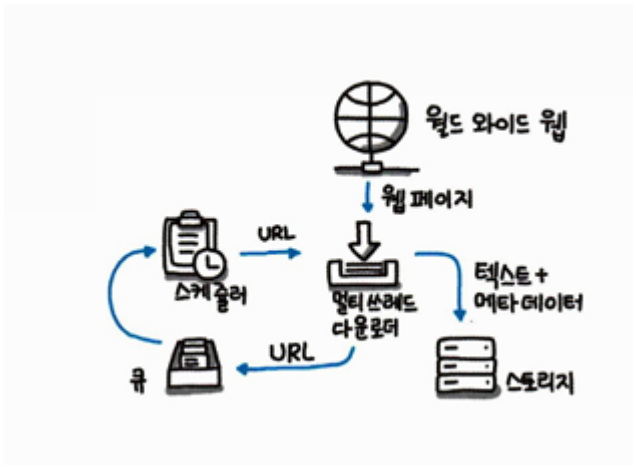


- 검색엔진이 시장의 흐름을 주도하면서 검색광고를 도입
- 쿼리에 적합한 광고를 매번 다르게 보여주는 타깃 마케팅 및 CPC(Cost Per Click) 방식을 도입

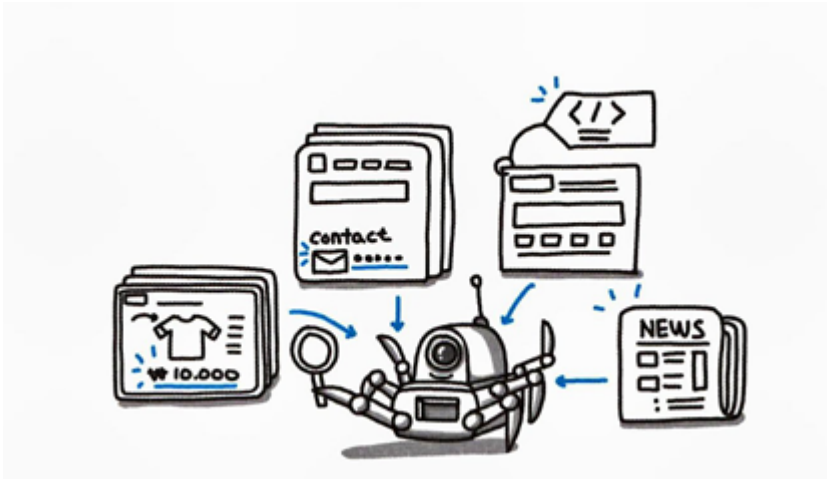
검색엔진의 수많은 문서

- 검색엔진이 인터넷에 있는 문서를 수집하여 검색에 적합하게 저장하는 것 → 색인(index)
- 구글은 약 300조 이상의 문서를 고가의 컴퓨터 몇대에 저장하는 방식이 아닌 사양이 낮은 컴퓨터 수천 대에 나눠 저장하는 방식을 선택
  - 구글 파일 시스템 (GFS) 를 만들어 냄
- 이렇게 많은 문서들 중 우리가 쿼리하는 문서를 어떻게 빨리 찾아내고 있는것일까?

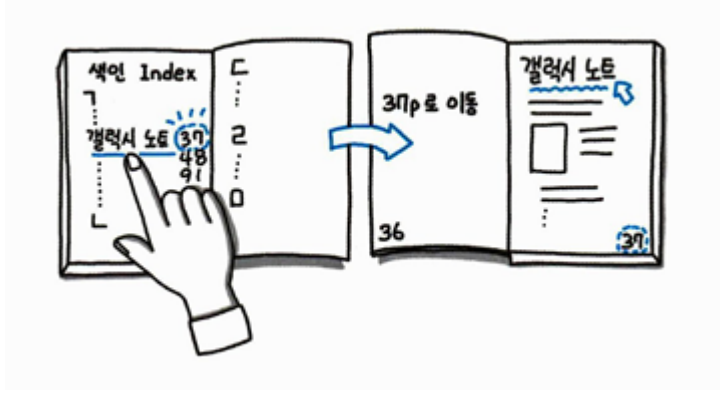
문서 수집



- 문서를 수집하기 위해선 웹사이트 구석구석을 돌아다녀야 함
  - 웹 문서를 갈고리처럼 긁어온다그래서 웹 사이트 수집 봇의 경우 웹 크롤러 (Web Crawler) 라고 부름
- 크롤러가 웹 문서를 방문할 때 마다 '색인(특수한 데이터베이스)' 에 정보를 추가
- 결국 크롤러의 목적 : 색인을 만드는 일.



검색엔진은 어떻게 검색할까?



- 먼저 책의 색인 처럼 항목을 정리 하는 과정이 필요 → 색인 구축

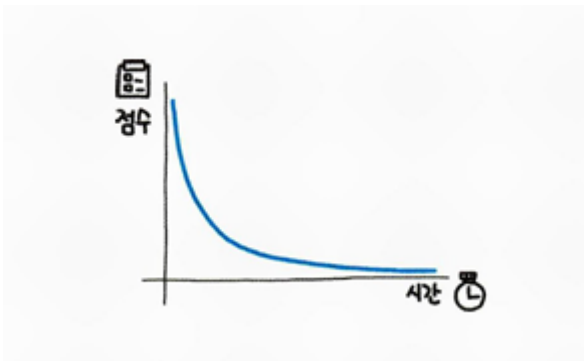
- 색인 구축 이후 검색엔진이 결과를 가져오는 방법
- 위 예에서 “파란색 나무 단추” 라고 검색하면?
  - 파란색 나무 단추의 공통 페이지는 11 페이지라 11페이지를 보여주면 됨
- 다른 예로 파란색을 검색하면 10,11,72,101 페이지 중 어느 페이지를 보여줘야 하는가?
  - 랭킹 개념 등장
    - 어떤 문서를 가장 노출할지 결정하는 알고리즘

랭킹, 수집 조 가치의 줄 세우기 기술

- 여러 종합적인 면을 모두 고려해야 함.
- ex. 구글의 경우 약 200여 가지의 랭킹 조건을 이용한다고 알려짐

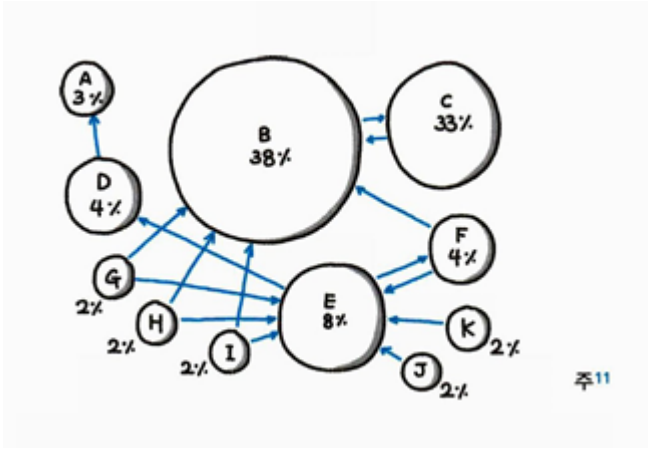
• 쿼리가문서 제목에 포함되어 있는가?
— 지금은 예전만큼중요도가 높진 않지만 제목에 쿼리가 포함되어 있다면 중요한 문서 라고 판단합니다.
• 긴문서인가?
— 짧은 글보다는 긴 글이 품질 점수가 높습니다.
• 문서 로딩이 빠른가?
— 빠른문서는더 좋은 경험을줍니다.
• 사이트에 접속할 수 없는 상황이 자주 발생하는가?
— 빠른 로딩과 함께 사이트의 안정성은 매우 중요합니다. 자주 다운되는 사이트에 있는 문서라면 아무리 내용이 좋아도 문서를 보기가 어렵 겠지요.
• 모바일에서 잘보이는가?
— 이제 모바일 인터넷 트래픽이 PC를 앞질렀습니다. 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 바탕이 기도 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 性} 야겠죠. 이 부분도 판단하여 점수에 포함합니다.
• 본문내용의 수준이 지적인 내용인가, 욕설로 가득한내용인가?
— 글의 품위도평가기준으로삼습니다.
• 저작권이 정식으로 표기되어 있는가?
— 저작권을 제대로 표시한 문서가 좋은 문서일 가능성이 높습니다.
• SNS에 링크가걸려 있는가?
느 좋은 문서라면 트위터 같은 SNS에 링크가 퍼진 경우가 많겠죠.

최신문서일수록...



- 최신문서일수록 점수가 높음
  - 오래되면 오래될수록 정보의 의미가 많이 약해진다고 보기 때문

품질 좋은 문서, 페이지 랭크



- 세그게이 브린과 래리 페이지라는 사람이 문서의 품질을 평가하는 획기적인 알고리즘을 고안
  - 유명한 사이트가 많이 가리킬수록 문서의 점수가 올라가는 알고리즘
  - 이를 페이지 랭크 알고리즘 이라 함
- 자신을 향하는 화살표의 수, 즉 링크된 횟수가 많을수록 좋음

쿼리에 딱 맞는 문서 찾는 법

- 유사도 점수
  - 사용자가 입력한 쿼리가 문서 어디쯤에 위치하느냐가 중요
  - 또한 순서대로 매칭되어 있는지도 중요
    - 이를 근접도 라고 함
    - 단어와 단어 사이의 간격이 좁을수록 더 유사한 문서라고 판단하고 높은 점수를 줌

```
• 얼마전 삼성전자에서 갤럭시 노트의 신제품을 출시했습니다. 새로운
기능에 대한 고객들의 기대가큽니다.
• 신제 품이 나온지 벌써 1년여 가 지 났지 만노트에 글을 쓰는 느낌 을 주
는 갤럭시는 아직 등장하지 않고 있습니다


--> 1번이 더 높은 점수를 받음
```

TF-IDF 그리고 마법같은 BM25

- 근접도 보다 더 훌륭한 유사도 알고리즘이 존재
  - TF-IDF 알고리즘
- 여기서 TF는 ‘Term Frequency’로 ‘단어의 출현 빈도’를 의미하고 IDF는 ‘inverse Document Frequency’로 ‘문서 출현 빈도의 역수’를 의미합니다.

- IDF는 다른 문서에 많이 출현할수록 작은 값이 된다 는 얘기죠. 원래 IDF 점수는 로그로. 계산하지만 여기서는 쉽게 설명하기 위해 단순히 출현 빈도만 이용해 '1/문서출현빈도' 형태의 간단한 분수로 나타내겠습니다. 이 렇게 하면 해당 문서에 많이 출현할수록, 다른 문서에는 적게 출현할수록 TF-IDF 점수가 커집니다.

문서	내용
A	갤럭시 노트 신제품 출시
B	갤럭시 노트 신제품 출시 새로운 노트 만나보세요
C	갤럭시 노트 과연 기존 노트 시리즈와 차별화된 노트 될까
D	갤럭시 전용 케이스 얇아진 두께에 따라 더욱 도드라져 보이는 디자인
E	삼성전자 역대 최고 실적 기록 반도체의 힘 파운드리 상반기 최대 매출

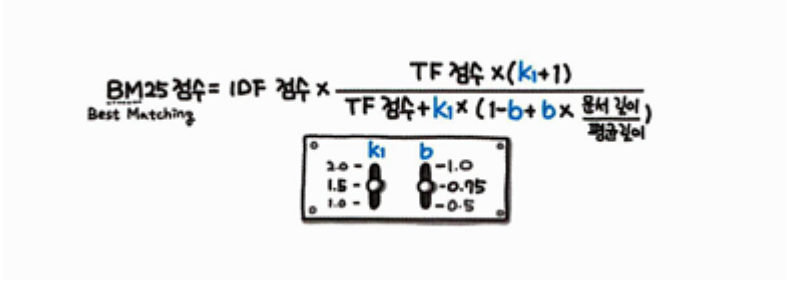
문서	내용	'갤럭시' TF-IDF 점수
A	갤럭시 노트 신제품 출시	0.25
B	갤럭시 노트 신제품 출시 새로운 노트 만나보세요	0.25
C	갤럭시 노트 과연 기존 노트 시리즈와 차별화된 노트 될까	0.25
D	갤럭시 전용 케이스 얇아진 두께에 따라 더욱 도드라져 보이는 디자인	0.25
E	삼성전자 역대 최고 실적 기록 반도체의 힘 파운드리 상반기 최대 매출	0

문서	내용	'갤럭시' TF-IDF 점수	'노트' TF-IDF 점수	'신제품' TF-IDF 점수	'갤럭시 노트 신제품' TF-IDF 점수
A	갤럭시 노트 신제품 출시	0.25	0.33	0.5	1.08 (3등)
B	갤럭시 노트 신제품 출시 새로운 노트 만나보세요	0.25	0.66	0.5	**1.41 (1등)**
C	갤럭시 노트 과연 기존 노트 시리즈와 차별화된 노트 될까	0.25	1	0	1.25 (2등)
D	갤럭시 전용 케이스 얇아진 두께에 따라 더욱 도드라져 보이는 디자인	0.25	0	0	0.25
E	삼성전자 역대 최고 실적 기록 반도체의 힘 파운드리 상반기 최대 매출	0	0	0	0

- “갤럭시 노트 신제품” 이라는 키워드로 질의 했을 때 가장 높은 점수를 받는 과정

실무에 쓰이는 좀 더 정교한 점수 계산 방식인 BM25

- Best Matching 25 의 약자
- 구글, 네이버, 다음 등 거의 모든 검색엔진이 채택한 유사도 계산 방식



- BM25 의 수식
- TF-IDF 는 문서 길이는 평가하지 않기 때문에 긴 문서가 무조건 유리함
  - 하지만 BM25 는 현재 문서 길이와 전체의 평균 길이를 비교하면서 가중치를 조절하는 장치가 있어서 같은 조건에서는 오히려 짧은 문서가 유리함
- 이외에도 BM25 수식에는 정체를 알 수 없는 k 와 b 라는 값이 존재
  - 이 두 값은 랭킹을 모델링하는 개발자가 임의로 정하는 값. 매개변수임
  - 스위치 역할을 함
  - 기본값으로는 k=1.5 , b=0.75 를 사용
- 위에서 구한 TF-IDF 점수를 구한건 매우 단순하게 계산했지만 실제로 IDF 점수를 정교하게 구하려면 몇가지 로그 계산을 거쳐야하는데 이 과정은 미리 계산했다고 가정하고 시작

단어	문서 출현 빈도	IDF 점수
갤럭시	5개 문서 중 4개 문서	0.2343
노트	5개 문서 중 3개 문서	0.2343
신제품	5개 문서 중 2개 문서	0.3364

- ‘갤럭시’ 와 ‘노트’의 출현 빈도가 다르지만 왜 점수는 같을까?
  - 출현빈도가 두 단어 모두 빈번하기 때문
  - 횟수는 많지 않지만 전체 비율로 보면 각각 60%, 80% 에 해당함
- 과반수가 넘게 출현하는 단어는 BM25계산식에 따를 때 IDF 점수가 마이너스로 나올 수 있어서 보정 작업이 필요함
  - 각각 최솟값으로 보정해야 되고 위 예에서는 그 값이 0.2343
- 이제 남은 수식인 문서길이/평균길이 인데 이는 무슨 역할을 할까?
  - BM25 수식에서 문서의 길이로 점수의 가중치를 조절.
- “갤럭시 노트 신제품 출시” 라는 문장에 들어간 띄어쓰기로 구분한 단어 수는 4
  - 이렇게 구한 현재 문서 길이값을 전체 문서의 평균 길이의 값으로 나누면 BM25 점수에 가중치 부여 가능
  - 4+8+9+10+11/5 = 8.4 (전체문서의 평균길이)
  - 4/8.4 = 0.48 이 나옴

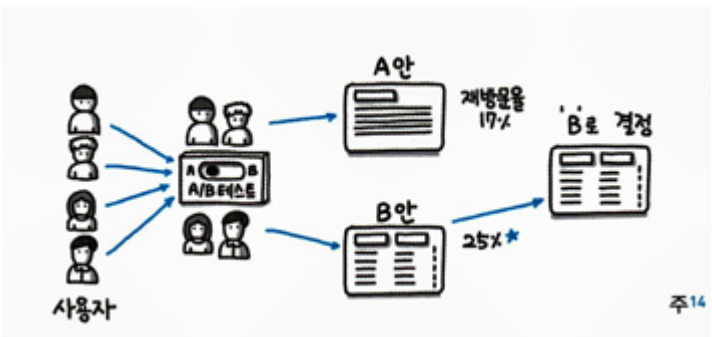
$$\text{'갤럭시' BM25 점수} = 0.2343 \times \frac{1 \times (1.5 + 1)}{1 + 1.5 \times (1 - 0.75 + 0.75 \times \frac{4}{84})} = 0.3066$$

문서	내용	'갤럭시 노트 신제품' TF-IDF 점수	'갤럭시 노트 신제품' BM25 점수
A	갤럭시 노트 신제품 출시	1.08(3등)	**1.05 (1등)**
B	갤럭시 노트 신제품 출시 새로운 노트 만나보세요	**1.41 (1등)**	0.92(2등)
C	갤럭시 노트 과연 기존 노트 시리즈와 차별화된 노트 될까	1.25(2등)	0.61(3등)
D	갤럭시 전용 케이스 얇아진 두께에 따라 더욱 도드라져 보이는 디자인	0.25	0.21
E	삼성전자 역대 최고 실적 기록 반도체의 힘 파운드리 상반기 최대 매출	0	0

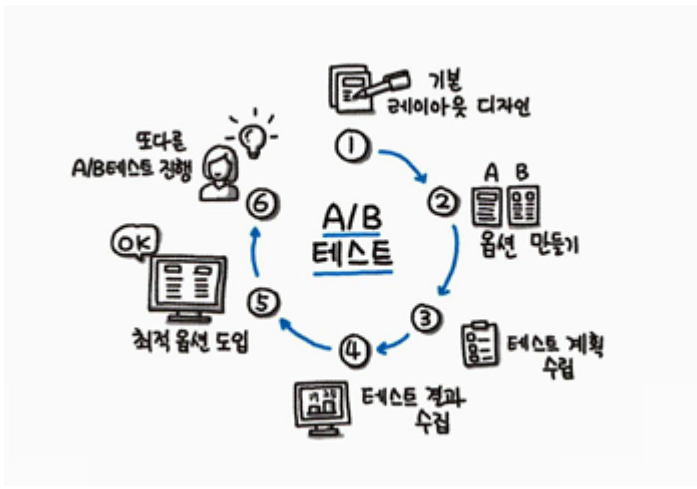
- 우리가 기대했던 순서와 동일하게 나옴!

A/B 테스트

- 시초는 과학분야에서 사용했던 무작위 대조 시험임



- 실제로 페이스북에서 A/B 테스트를 적극 활용한다고 함
  - 내가 로그인해서 보는 페이스북 화면과 친구가 로그인해서 보는 페이스북 화면이 다름
- 말 그대로 사용자에게 2가지(혹은 여러가지) 다른 결과를 보여주고 반응을 살핌

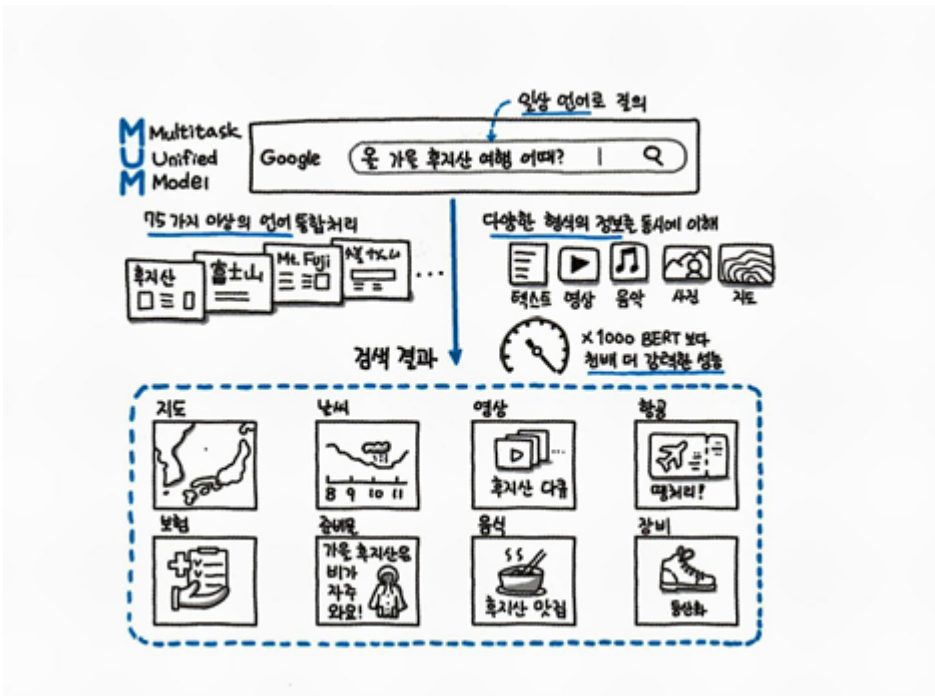


검색엔진 최적화

- 창과 방패의 싸움임. 관리자는 랭킹을 쉽게 올릴 수 없게 끊임없이 방어로직을 넣음

더 똑똑해지는 구글 검색의 진화

- 이제는 딥러닝을 이용해 문장의 의미를 정확하게 이해하고 이에 맞는 정답을 찾아냄
- 오타 교정의 경우도 딥러닝이 활용됨
- 사실... 과거에는 편집거리 (edit distance) 라고 해서 정상적인 철자와 얼마만큼 틀렸는지 계산하여 오타를 교정했었음...
    - 실제 편집거리 알고리즘이라고하여 존재하고, 구글링 하여 직접 적용해봤었음
- 이제는 수많은 데이터로 학습한 딥러닝으로 오타를 훨씬 더 정교하게 교정 해낸다!!



- 2021년 구글이 발표한 MUM 기술 (멀티태스킹 유니파이드 모델)