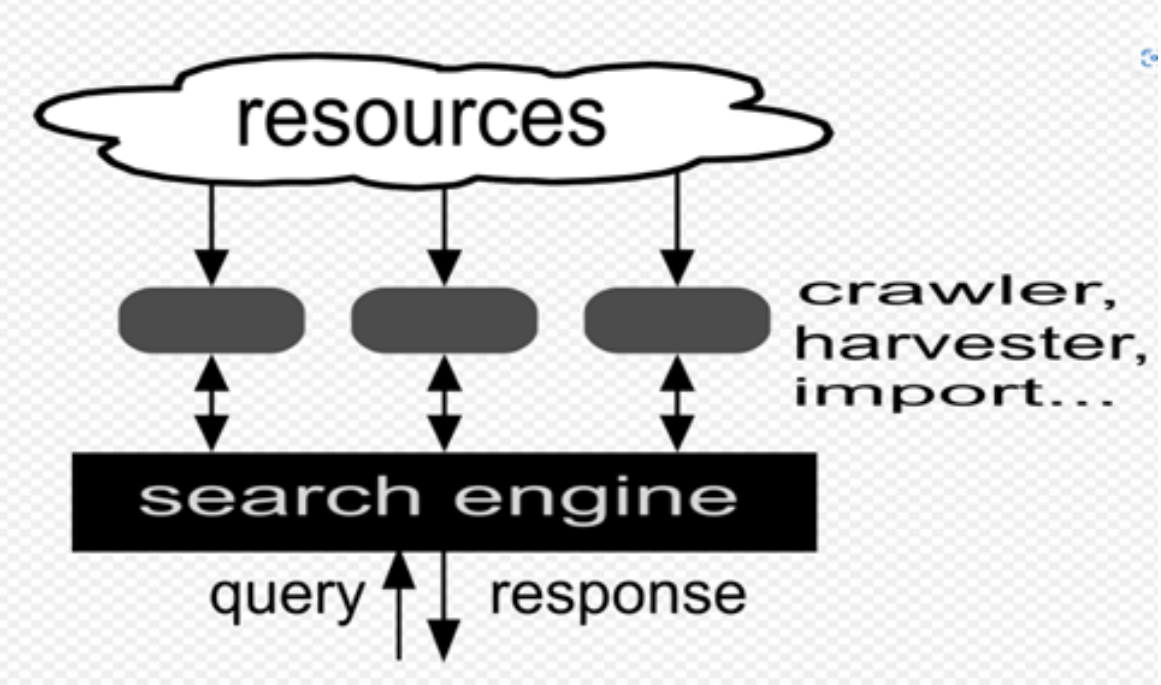
Google Search Engine 과 그 문제점

:"Is Google Getting Worse? A Longitudinal Investigation of SEO Spam in Search Engines"(Bevendorff, Wiegmann, Potthast, Stein, 2024) 논문 분석을 중심으로

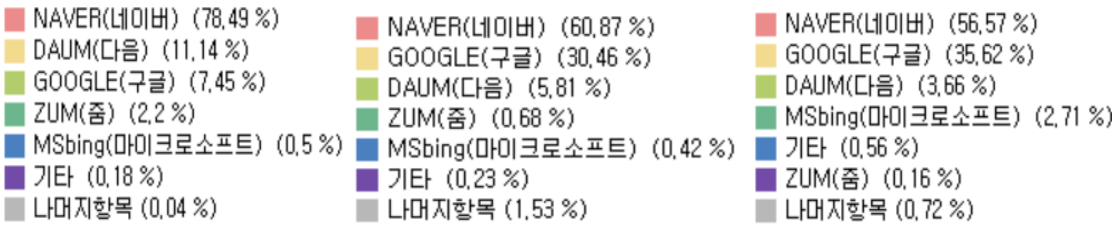
Search Engine 이란

검색 엔진(Search Engine)은 사용자가 원하는 정보를 검색할 수 있는 Web 기반의 Software 로 인터넷에 존재하는 수많은 웹페이지를 탐색하고, 사용자가 입력한 검색어나 질의에 관련된 웹페이지를 찾아주는 역할을 한다. Cloud AI 플랫폼에서는 NLP 기술을 활용하여 검색 엔진을 만들 수 있으며 사용자의 검색 기록과 행동을 분석하여 개인화된 검색 결과를 제공하는 데 활용될 수 있다.



[Figure 1] 검색 엔진의 예시

Search Engine 사용 Trend



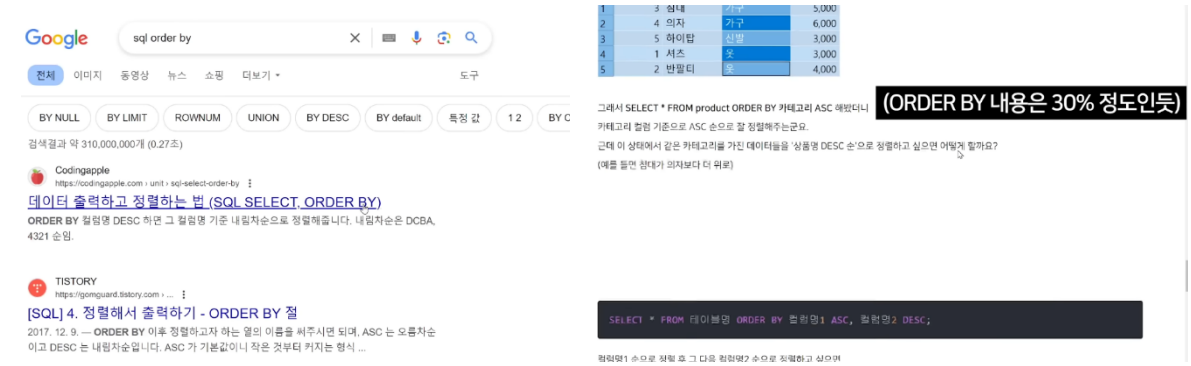
[Figure 2] 연도에 따른 Engine 별 사용 비율, (좌) 2015 년, (중) 2019 년, (우) 2024 년

위 자료에서 보는 것과 같이 국내에서조차 해외 검색 사이트인 Google 의 국내 이용자가 증가하는 변화를 보인다. Google 은 인터넷 상에서 가장 많은 정보를 제공하는 검색 엔진 중 하나로, 사용자들이 다양한 주제와 관한 정보를 찾을 수 있도록 도와준다. 타 검색엔진에 비해 검색 알고리즘이 매우 정교하며 사용자가 원하는 정보를 더욱 정확하게 제공하기 위해 지속적으로 개선되고 있으며 Gmail, Youtube, Google Maps 등 다양한 콘텐츠를 제공하여 국내에서 또한 검색 엔진 사이트 중 Google 의 비율이 높아지고 있다.

Google Search Engine based on Cloud AI

Google Search Engine 은 클라우드 기술을 기반으로 운영된다. Cloud AI 기술은 클라우드 인프라를 통해 검색 엔진의 확장성, 성능, 안전성을 향상시키고, 새로운 기능 및 서비스를 제공하는 데 활용된다. 또한 클라우드 기반 인프라는 필요에 따라 자동으로 확장되거나 축소될 수 있다. 구글 검색 엔진은 급격한 트래픽 증가에 대응하기 위해 클라우드 자원을 신속하게 확장할 수 있다. 이는 검색 서비스의 안정성과 성능을 유지하는 데 큰 장점이 있다. 또한 Cloud AI 기반 인프라를 사용하여 구글은 새로운 기능을 빠르게 개발하고 배포할 수 있다. 이를 통해 검색 엔진은 지속적으로 업그레이드되고 개선되며, 사용자에게 최신 기술과 서비스를 제공할 수 있다.

Search Engine 의 문제점, 그리고 탐구



[Figure 3] Google 에 SQL ORDER BY 문법 검색 결과, 최상단 결과는 내용의 30%만 관련이 있음

이렇게 다방면에서 활용되고 있는 Search Engine 은 최근 많은 불편함을 유발하고 있기도 하다. 예를 들어, Figure 3 처럼 SQL 의 ORDER BY 문에 대해 검색하기 위해 Google 에 "sql order by"라고 검색을 할 경우, 해당 문법에 대해 간결하고 정확하게 정보를 전달하는 결과도 있는 반면 해당 문법에 대한 설명이 "포함되기만 한" 결과가 보여지기도 했고, 심지어 이 경우 그 결과가 최상단에 위치하기도 했다. 이렇듯, 최근 몇 년 동안 많은 이용자들이 검색 엔진이 도출해내는 결과에 대해 불만족을 표시하기도 했다. 이에 반하여, 최근 우리는 다양한 기술 발전을 경험하고, Search Engine Algorithm 부분에서도 이는 예외가 아니었다. 그렇다면 대체 무엇이 이러한 결과를 내는 것일까?

독일의 Leipzig University 소속 Janek Bevendorff 교수는 이러한 Search Engine 의 한계점에 대해 3 명의 다른 연구자들과 함께 탐구하여, ECIR 2024 에 논문 "Is Google Getting Worse? A Longitudinal Investigation of SEO Spam in Search Engines"를 출간하였다. 연구팀은 Google 과 Bing, DuckDuckGo 에서 7392 개의 제품 리뷰 검색 결과를 분석하여, BM25 기반의 검색 엔진인 ChatNoir 로 ClueWeb22 Data Set 에서 같은 검색어에 대한 결과와 비교하였다. 또한, 이러한 검색 결과들에 대해 SEO(Search Engine Optimization; 검색 엔진 최적화)가 이루어지는지를 알아내기 위해 1 년동안 지속적으로 모니터링을 하였다.

Affiliate Marketing 과 Search Engine, 그리고 대응

결론적으로, Google, Bing, DuckDuckGo 등의 상업적 검색 엔진들의 결과인 SERP(Search Engine Result Page)의 상단, 즉 가장 적합한 검색 결과라고 생각되는 것들에 출력되는 결과물은 ChatNoir 과 비교하여 이용자들에게 "이상적"이지 못하다는 평가를 내릴 수 있게 되었다. 최상단, 즉 가장 높은 평가가 나타난 검색 결과들 중 대다수는 제휴 마케팅(Affiliate Marketing)과 관련이 있고, 검색 결과들 중 매우 많은 것들이 SEO Spam 에 해당하는 것으로 밝혀졌다. 비록 Google 의 검색 결과가 Bing 과 DuckDuckGo 보다 좋은 결과를 더 많이 내주었지만, 그래도 ChatNoir 에 비교하면 아직 부족한 수준임이 나타났다.

Affiliate Marketing 이란, Web Business 촉진 기법의 하나이며, 어떤 Website Publisher 가 Partner 의 Website 에 이익을 창출시키는 것에 대해 보상을 받는 방식의 Marketing 기법이다. 이 기법은 특히 여러 기업과 개인이 모여서 Internet Marketing 을 펼칠 때 사용되는데, 이는 즉 보다 더 많이 노출될수록 많은 이익을 취할 수 있다는 특징을 가진다는 것을 시사한다. 다시 말해 Affiliate Marketing 을 사용할 경우, Search Engine 에서 보다 상단에 위치할수록, 즉 Search Engine 에게 더 좋은 평가를 받을수록 더 많은 이익을 취할 수 있기 때문에 SEO Spam(Search Engine 이 좋게 평가하는 항목을 찾아내, 이를 악용하여 자신이 원하는 내용을 SERP 의 상단에 노출될 수 있도록 하는 것)이 많이 발생하고 있고, 앞으로도 더 많이 발생할 것으로 추측되고 있다.

물론 이에 대한 대응이 전혀 없는 것은 아니다. Bevendorff 교수와 연구진들은 1 년동안 모니터링하며 검색 결과들을 지켜본 결과, 대다수의 Review Spam 결과들은 주로 짧은 수명을 갖고, 특히 Google 이 검색을 하는 데에 사용하는 Ranker 를 Update 를 하고 난 뒤에는 즉시 또는 늦어도 2 주 내에 제거됨을 연구팀이 확인하였다. 또한, 연구팀이 연구를 시작한 이후로부터도 Google 의 SEO 는 꾸준히 발전함을 지켜봤지만, 상술했듯이 SEO Spam 의 빈도수는 점차 증가하고 있고, 아직 완벽하고 모든 부분의 Spam 을 막은 것이 아니므로 더 많은 발전을 필요로 한다는 결론이 보다 정확하다. 따라서 현 시점에서의 검색진 이용자들은 자신들이 정확한 결과를 위해 검색하는 방식을 바꿔야 할 필요성이 있다고 사료된다.

이용자들에게 남겨진 방안들

결국 현시점에서 검색 엔진을 활용하려고 하는 일반 이용자들은 좋지 못한 검색 결과에 노출되어 있는 상황이다. 그렇다면 현실적으로 이용자들 입장에서 남은 선택들은 무엇일까? 이하는 이용자들이 활용할 수 있는 "좋은 결과"를 얻기 위한 방법들이다.

1 검색 연산자 사용

1.1 **"검색어"**: 그 검색 어구와 완전하게 일치하는 문구가 포함된 검색, 이를 통해 스팸덱싱(Spamdexing, 검색 엔진 색인(Search Engine Indexing)을 의도적으로 조작하는 행위)의 키워드 반복 기법을 피할 수 있음.

1.2 **site 연산자**: 특정 웹사이트만을 검색 범위로 한정, 이를 통해 스팸덱싱된 웹페이지를 배제할 수 있다.

1.3 **filetype 연산자**: 검색 결과를 특정 파일 형식으로 제한하여 검색함. 웹사이트를 검색하는 것이 아닌 특정 filetype 만을 검색하므로 광고나 불필요한 게시물을 차단할 수 있다.

2 AI 기반 검색 엔진 사용

AI 기반 검색엔진은 불필요한 게시물들을 효과적으로 차단할 수 있다. NLP 모델을 통해 게시물의 반복된 키워드를 추출하여 해당 게시물을 걸러낼 수 있으며 CNN 모델 같은 이미지 인식 모델을 통해 이미지 스팸 및 광고를 효율적으로 차단할 수 있다. 또한 스팸덱싱의 수법 중 하나인 링크 스팸(Link Spam)을 GNN 모델을 통해 각 웹사이트의 구조를 학습하여 스팸 웹사이트를 탐지할 수 있다.

[Figure 4, 5] Copilot(좌), Perplexity AI(우)

참고 문헌

[Figure 1] Wikipedia(위키피디아), https://ko.wikipedia.org/wiki/검색_엔진

[Figure 2] INTERNET TREND, https://internettrend.co.kr/trendForward.tsp

[Figure 3] "구글 검색은 어쩌다가 쓰레기가 되었나", 코딩애플(Youtube), https://www.youtube.com/watch?v=ITc1TJVWZGM

[Figure 4] "Microsoft 365 Copilot Commercially Released", Kurt Mackie 2023, https://redmondmag.com/articles/2023/11/01/microsoft-365-copilot-ga.aspx

[Figure 5] "What is Perplexity AI?", Himanshi Singh 2024, https://www.analyticsvidhya.com/blog/2024/01/perplexity-ai-is-going-to-change-the-way-we-search-beware-google/

"Is Google Getting Worse? A Longitudinal Investigation of SEO Spam in Search Engines", Bevendorff, Wiegmann, Potthast, Stein, 2024, https://downloads.webis.de/publications/papers/bevendorff_2024a.pdf