**Optimizing Search Engines using Click through Data ---** Thorsten Joachims

보통 사용자가 입력한 키워드에 대해서 검색 엔진은 키워드가 포함된 수천 개의 결과를 보여주게 된다. 그러나 이중에서 실제로 사용자가 필요한 것은 극히 일부분에 지나지 않는다. 그렇기 때문에 사용자가 입력한 쿼리에 대해서 실제로 접속한 페이지를 알 수 있다면 그 데이터를 학습 데이터로 이용할 수 있다. 그러나 일반적으로 이러한 데이터는 사용자의 피드백에 의해 만들어지는데, 이는 매우 적은 양이다. 그렇기 때문에 본 연구에서는 검색엔진에 존재하는 로그 파일을 이용하여 학습 데이터로 사용할 수 있는 정보를 만드는 것을 제안한다. 이 방식은 기존의 피드백 방식을 통해 데이터를 얻는 것과 비교해서 무료이면서, 더 적절하게 수행된다. 본 논문에서는 사용자들이 입력한 쿼리에 대해서 어떤 페이지가 더 많이 클릭이 되었는지에 따라 검색 기능을 학습 시킨다. 이를 위해 non-linear ranking function으로 확장 가능한 SVM 알고리즘을 사용하여 검색 엔진을 학습시켰다.

**Agglomerative clustering of a search engine query log** --- Doug Beeferman and Adam Berger

이 논문에서는 비슷한 쿼리와 비슷한 URLs 의 클러스터를 찾기 위해서 인터넷 검색 엔진으로 사용자 Transaction을 mining하는 기법을 소개한다. 탐색하는 정보는 clickthrough data(각 레코드는 검색 엔진이 제시한 것 중 사용자에 의해 선택된 URL과 함께, 사용자의 쿼리로 이루어져 있다.) 이다. 이 데이터 셋을 Bipartite 그래프로 보며 (쿼리를 노드로 하는 쪽과, URL을 노드로 하는 쪽) 이 그래프에 관련된 쿼리와 URL을 찾기 위한 클러스터링 알고리즘을 적용한다. 이 알고리즘은 쿼리와 URL의 실제 내용은 무시한 채로, 단지 clickthrough data안에서 어떻게 쿼리와 URL이 함께 일어나는 지에만 관심을 갖는다. 논문은 사용자의 웹 검색에 도움을 주기 위해 어떻게 발견해낸 클러스터들을 보여줄지 설명한다. 그리고는 Lycos 검색 엔진에서 발견한 클러스터들의 효율을 측정한다.

논문에서 제시한 알고리즘은 페이지들을 vector space에 끼워 넣지 않더라도 매우 유사한 URL의 클러스터를 찾을 수 있게 한다. 클러스터링 전략은 첫째, URL을 이용하는 것이다. 동일한 정보를 필요로 하는 사용자들은 검색 엔진에 다양한 쿼리를 입력할 것이다. 하지만 같은 url을 선택할 것이므로, 이를 이용해서 쿼리가 비슷함을 찾을 수 있다. 둘째, 같은 쿼리를 입력한 후, 사용자는 다른 url을 방문하게 될 것이라는 사실을 이용한다. 하지만 쿼리가 똑같으므로 이는 URL이 유사하다는 증거가 된다. 논문은 이후, 이 클러스터링 알고리즘을 bipartite graph에 적용하는 문제 등에 대해 다룬다.

**Clustering User Queries of a Search Engine ---** Ji-Rong Wen, Jian-Yun Nie and Hong-Jiang Zhang

검색의 정확도를 높이기 위해서, 어떤 서치 엔진들은 FAQs를 제공한다. 이 논문의 task는 FAQs를 식별하는 것이다. 사용자 로그와 같은 contents에 따라 비슷한 쿼리들을 클러스터링하고 있다. 이렇게 클러스터링 된 결과들은 FAQ 식별을 위한 유용한 정보를 제공한다.

인터넷에서 정보의 폭발적인 증가로, 검색 엔진에 대한 요구가 늘어나고 있다. 기존의 검색 엔진에서는 사용자의 필요와는 관계없는 수 천 개의 문서들을 쿼리의 결과로 내놓기 때문에 검색 엔진에 대한 정확도가 요구되고 있다. 기존의 검색 엔진들이 쿼리에 있는 키워드를 문서의 내용에 맞추는데 집중했다면 새롭게 나오는 검색 엔진들은 사용자의 쿼리를 이해하려고 한다. 그리고는 다른 사용자가 질문했던 비슷한 쿼리들을 제시해서 정확한 답을 내놓으려 한다. 이러한 시스템은 사람들이 비슷한 물음에 관심이 있다는 가정을 하고 있다 (the Frequently Asked Questions/Queries -FAQs). 그러나 사용자에 의해 제시되는 쿼리들은 그 형태와 의도가 매우 다양하다. 편집자가 어떤 쿼리가 FAQs인지 결정하는 것은 여전히 open issue이다. 즉, 어떻게 시스템이 두 물음이 비슷한 것인지 판단하는 것이 중요하기에 기존의 IR에서는 키워드에 따라서 Similarity를 계산했었다. 하지만 이는 키워드의 한계 때문에 제한된다. 그래서 이 논문에서 저자는, 사용자의 로그를 이용해서 쿼리를 클러스터링하는 새로운 방식을 제안한다.

1. 사용자가 다른 쿼리에 대해 같은 문서를 검색했다면 그 쿼리는 비슷하다.
2. 쿼리 set에 대해서 documents set이 선택되었다면 이 문서의 term들은 쿼리의 term들과 어느 정도 관련된다.

그리고 이 실험의 결과, 논문에서 소개된 approach를 이용하면 비슷하다고 판단되었던 쿼리가 실제로 같은 클러스터로 묶임을 알 수 있었다. 논문은 사용자의 로그로 쿼리 클러스터링을 하는 것이 유용함을 보인다.

**Optimizing web search using web click-through data ---** Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Yong Yu, Wei-Ying Ma, WenSi Xi, and WeiGuo Fan

현재 많은 검색 엔진들은 주어진 쿼리에서 웹 페이지 간의 연관성을 찾기 위해 특정 키워드가 포함되어 있는지를 계산한다. 이러한 방법은 검색 단어가 명확하고 특정적일 경우에 대해서만 정확하게 돌아간다. 실제로는 대부분 검색어가 짧고 모호하며 웹 페이지가 noisy information을 가지고 있기 때문에 적용하기가 힘들다. 그래서 이때까지는 이를 조금이라도 해결하기 위해 대표적으로 anchor text를 이용하였다. 이러한 방법은 키워드만을 이용하여 검색하는 것보다는 나은 결과를 보여준다. Anchor text를 이용하는 것 이외의 방법으로는 click-through data를 이용하는 방법이 있는데 이는 아직까지 활발히 연구되지 않는 분야였다.

Click-through data는 검색 엔진에 존재하는 로그 데이터에서 많이 얻을 수 있다. 이러한 click through data는 사용자가 클릭한 페이지에 대한 데이터이기 때문에 실제로 사용자가 어떤 쿼리에 대해서 어떠한 웹 페이지를 선택했는지 알 수 있고, 이를 통해 쿼리에 따라 사용자의 의도를 파악할 수 있다. 본 논문에서는 사용자가 요구하는 정보와 실제 웹 페이지에서의 틈을 줄일 수 있는 방법으로 click through data를 추가적인 metadata로 이용을 제안한다. 이러한 쿼리 로그를 이용한 metadata 생성 방법은 다음과 같은 3가지 특징이 존재한다.

1. Click through data는 사용자의 데이터이다. 원래 웹 페이지의 성능은 사용자가 측정하는 것이지 에디터가 측정하는 것이 아니기 때문에 anchor text를 이용하는 것보다 더 사용자의 관점에서 바라볼 수 있다.
2. 다른 메타 데이터들과 결합하여 함께 사용할 수 있다.
3. Click through data의 누적이 웹 페이지와 쿼리간의 상관 관계로 발전 할 수 있다.

이렇게 click through data를 이용하는 데에는 다음과 같이 해결해야 할 문제들이 많이 남아 있다.

1. Click through data는 noisy하고 불완전하다.
2. 사용자들이 보통 검색 엔진에서 상위의 웹 페이지만 이용하므로 매우 sparse하다는 특징이 있다.
3. 실제 세계에서 나오는 새로운 웹 페이지와 쿼리에 대해서는 활용하기가 어렵다.

본 논문에서는 이러한 이슈들을 해결하고 보강하기 위해 웹 페이지와 쿼리 간의 유사성을 계산하기 위한 알고리즘을 제시한다. 또한, 실험을 통해서 일반적인 검색 시스템에 비해서 157% 정도의 성능을 보여준다고 한다.