# SearchBoth와 DogPile을 combination한 SearchEngine

# SearchEngine Combined with SearchBoth and DogPile

**요 약**

사용자는 많아지고 사용자의 쿼리는 갈수록 짧아지고 연산은 거의 쓰지 않는다. 그리고 검색엔진이 사용자가 의도한 내용을 어떻게 알 수 있는지 판단하는 것이 중요하다. 따라서 본 리포트는 SearchBoth의 사용자가 선택할수 있는 여러가지 옵션과 DogPile의 사용자의 의도에 맞는 검색 결과가 상위에 랭크된 것을 combination한 알고리즘을 제안한다.

## 1. 서 론

SearchBoth는 9개의 검색 카테고리로 분류되며 사용자들이 각자 좋아하는 웹사이트(구글, 네이버, 다음 등)이 있고 이런 모든 웹 정보를 통해 한 검색 창에서 다른 검색 창으로 이동하는 번거로움을 겪지 않아도 한 검색 창 안에서 정보를 검색할 수 있는 좋은 방법론을 취하였다. SearchBoth의 장점으로는 사용자로 하여금 여러가지 옵션들을 선택할 수 있게끔 선택지가 주어진다. 단점은 사용자의 의도와는 다른 결과가 높은 랭킹에 올라 정확도가 떨어진다는 단점이 있다.

DogPile은 InfoSpace에서 개발한 메타 서치 기술을 기반으로한 구글과 야후를 포함한 최고의 검색엔진에서 가장 좋은 결과를 리턴한다. 이로써 사용자들은 더욱 빠르게 원하는 것을 찾을 수 있다. 각 검색엔진에는 자체 검색 방법이 존재하고 각각 다른 결과를 반환한다. 이때 DogPile은 모든 결과값을 가져와 어떤 것이 사용자가 검색한 것과 가장 관련이 있는지 결정하고 중복된 문서를 제거하고 결과를 보여준다. 따라서 다른 어떤 웹 사이트보다 완벽한 결과 목록을 얻을 수 있다고 주장한다. DogPile의 장점으로는 사용자의 의도에 맞는 결과가 상위 랭킹에 올라와 정확도가 높다는 장점이 있지만 단점으로는 사용자가 선택할수 있는 여러가지 옵션이 없이 DogPile 내에서의 알고리즘을 그대로 받아 결과가 표시된다는 점이다.

앞서 설명한 SearchBoth와 DogPile의 장점을 취하고 단점을 없애 SearchBoth와 DogPile의 Combination 방법론을 택한 Combination Search를 제안한다.

## 2. 관련기술

SearchBoth에 필요한 기술은 사용자의 쿼리를 입력받아 각 사이트에 검색을 한 뒤 사용자에게 한 화면에서 보여주는 것으로 특별한 기술이 필요하지 않는다. 다만 여러가지 옵션이 있고 이 옵션 또한 옵션에 해당하는 사이트에 검색을 한 뒤 사용자에게 보여주는 것으로 여러 사이트의 검색을 한 화면에 출력하는 기술이다.

DogPile의 개인정보 이용약관을 보면 우편주소, 이메일 주소, 전화번호등의 개인정보, 네트워크 장비가 엑세스하는 세부사항, 사용 세부정보, IP주소, 쿠키, 웹 비콘 및 기타 추적 기술을 사용하여 정보를 수집하고 자동 데이터 수집 기술을 이용하여 탐색 작업과 패턴에 대한 특정 정보, 트래픽 데이터, 위치 데이터, 로그, 운영체제, 브라우저 유형, 온라인 활동 행동 등을 수집한다. 이를 통해 Personalized Search를 구현하고 키워드 검색 쿼리를 전송하면 서버가 이를 받아 미리 지정한 포털 사이트들에 쿼리를 전송하여 각 포털 사이트의 검색 결과를 모두 받아 사용자가 검색한 것과 가장 관련이 있는지를 결정하고 중복된 문서를 제거하고 결과를 보여주는 메타

서치 기술[1]을 사용하여 보다 정확한 검색을 지원한다.

메타 서치에 관한 기술의 내용은 다음과 같다.

메타 서치는 여러 개의 다른 검색 엔진에 여러 개의 쿼리를 전송함으로써 이 항목의 검색 범위를 확장하고 더 많은 정보를 찾을 수 있다. 그들은 다른 검색 엔진에 의해 만들어진 인덱스를 사용하여 결과를 통합하고 종종 독특한 방법으로 후처리한다. 메타 검색 엔진은 동일한 양의 노력으로 더 많은 결과를 검색할 수 있기 때문에 단일 검색 엔진보다 이점이 있다.  또한 자원을 찾기 위해 다른 엔진의 검색을 개별적으로 입력해야 하는 사용자의 작업을 줄인다 메타 서칭은 사용자가 검색하는 목적이 주제에 대한 개요를 얻거나 빠른 답변을 얻는 것이라면 유용한 접근법이기도 하다. 야후처럼 여러 개의 검색 엔진을 통과해야 하는 대신 메타서치 엔진은 구글을 통해 결과를 비교하는 대신 결과를 신속하게 컴파일하고 결합할 수 있다. 그들은 추가 후처리 없이 질의된 각 엔진의 결과를 나열하거나 결과를 분석하고 그들 자신의 규칙에 따라 순위를 매겨서 할 수 있다.

메타서치의 단점으로는 메타 서치 엔진은 쿼리 형식을 디코딩하거나 쿼리 구문을 완전히 변환할 수 없다. 메타 검색 엔진에서 생성된 링크 수는 제한되어 있으므로 사용자에게 쿼리의 전체 결과를 제공하지 않는다. 대부분의 메타 검색 엔진은 단일 검색 엔진에서 10개 이상의 링크된 파일을 제공하지 않으며 일반적으로 결과를 위해 더 큰 검색 엔진과 상호 작용하지 않는다. 후원 웹 페이지는 우선 순위가 정해졌으며 일반적으로 먼저 표시된다. 메타 서치는 또한 특히 사용자가 대중적이거나 일반적인 정보를 검색하는 경우 질의된 주제에 대한 더 많은 범위가 있다는 환상을 준다. 질의 엔진에서 나온 여러 가지 동일한 결과를 얻는 것이 일반적이다. 또한 사용자가 고급 검색 구문을 사용하여 검색하는 것이 쿼리와 함께 전송되는 것이 더 어렵기 때문에 사용자가 특정 엔진에서 고급 검색 인터페이스를 사용하는 경우만큼 정확하지 않을 수 있다. 이로 인해 간단한 검색을 사용하여 많은 메타 검색 엔진이 생성이 된다.

메타서치의 동작은 다음과 같다. 메타서치 엔진은 사용자로부터 단일 검색 요청을 받아들인다. 이 검색 요청은 다음에 다른 검색 엔진의 데이터베이스로 전달된다. 메타 검색 엔진은 웹 페이지 데이터베이스를 만들지 않지만 여러 소스의 데이터를 통합하기 위한 가상 데이터베이스를 생성한다. 모든 검색 엔진은 고유하며 순위가 매겨진 데이터를 생성하는 알고리즘이 다르기 때문에 복제본도 생성된다. 중복을 제거하기 위해 메타서치 엔진은 이 데이터를 처리하고 자체 알고리즘을 적용한다. 수정된 목록은 사용자를 위한 출력으로 생성된다. 메타 검색 엔진이 다른 검색 엔진에 쿼리를 날릴때, 이러한 검색 엔진은 세 가지 방법으로 반응한다. 이 검색 엔진은 인덱스 데이터베이스에 대한 개인 액세스를 포함하여 메타 검색 엔진에 대한 인터페이스에 대한 완전한 액세스를 협력하고 제공할 것이며, 메타 검색 엔진은 인덱스 데이터베이스에 대한 변경 사항을 통보할 것이다. 검색 엔진은 인터페이스에 대한 액세스를 거부하거나 제공하지 않을 것이다. 이는 조잡한 방식으로 작동할 수 있다. 검색 엔진은 완전히 적대적이고 거부 될 수 있다. 메타 검색 엔진은 데이터베이스 및 심각한 상황에서 전체 액세스를 수행하기 위해 법적 방법을 모색한다고 한다.

랭킹의 구조는 다음과 같다. 많은 검색 엔진에서 높은 순위를 차지하는 웹 페이지는 유용한 정보를 제공하는 데 더 관련이 있을 수 있다. 그러나 모든 검색 엔진은 각 웹 사이트에 대해 다른 순위 점수를 가지며 대부분의 경우 이러한 점수는 동일하지 않다. 검색엔진이 다른 기준과 채점 방식을 우선하기 때문에 웹사이트가 한 검색엔진에서 높은 순위를 차지하고 다른 검색엔진에서도 낮은 순위를 차지할 수 있기 때문이다. 메타 서치 엔진은 신뢰할 수 있는 계정을 생성하기 위해이 데이터의 일관성에 크게 의존하기 때문에 문제가 된다.

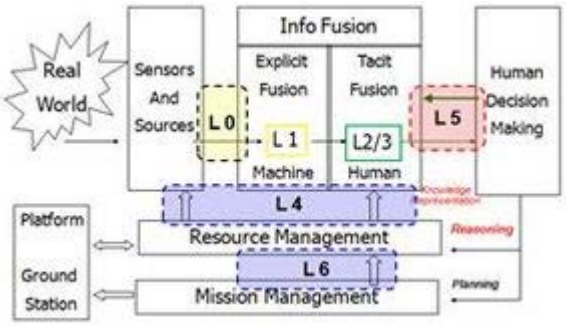


Fig. 1 Data Fusion Model

퓨전은 다음과 같다. 메타 서치 엔진은 퓨전 프로세스를 사용하여 보다 효율적인 결과를 위해 데이터를 필터링한다. 사용된 두 가지 주요 융합 방법은 컬렉션 퓨전 및 데이터 퓨전과 컬렉션 퓨전이다. 콜렉션 퓨전 : 분산 검색 (distributed retrieval)이라고도 알려져 있으며 관련없는 데이터를 색인하는 검색 엔진을 특별히 다룬다. 이러한 소스가 얼마나 중요한지 판별하기 위해 Collection Fusion은 내용을보고 쿼리와 관련하여 관련 정보를 제공할 가능성을 데이터에 표시한다. 생성된 내용

부터 Collection Fusion은 순위에서 최상의 자원을 선택할 수 있다. 선택한 자원은 목록으로 병합된다. 데이터 퓨전: 일반적인 데이터 세트를 색인하는 검색 엔진에서 검색된 정보를 처리한다. 과정은 위와 매우 유사하다. 데이터의 초기 순위 점수는 단일 목록으로 병합된 후 이 문서 각각의 원래 순위가 분석된다. 점수가 높은 데이터는 특정 쿼리와의 관련성이 높으므로 선택된다. 리스트를 생성하려면, CombSum과 같은 알고리즘을 사용하여 점수를 정규화 해야 한다. 이는 검색 엔진이 서로 다른 알고리즘 정책을 채택하여 점수를 비교할 수 없는 결과를 낳기 때문이다.

## 3. 사고실험

SearchBoth에서의 advanced option과 DogPile의 Meta Search를 융합하여 사용자에게 폭넓은 선택과 정확한 결과를 돌려주는 모델을 생성한다. 일례로 구글에서 지원하는 연산자 검색(OR, " ", -, site:, * 등)의 검색 방법이 있는데 이를 간단한 마우스 클릭으로 선택할 수 있게끔 하는 것이다. 원하는 사이트를 사용자가 원하는 대로 선택할 수 있다. 또한 DogPile에서 수집한 개인정보 데이터를 이용하여 Machine Learning을 활용한 개인화 검색의 정확도를 높혀 사용자의 의도를 더욱 정확하게 판단할 수 있고 랭킹 알고리즘의 정확도 또한 높아지게 된다. 이는 단순한 구글링의 결과 보다 더욱 높은 수준의 결과를 리턴할 것으로 예상된다.

개인적으로 추가하고싶은 검색 엔진은 금융과 관련된 검색엔진이다. 주식, 파생상품, 채권, 보험, 부동산등의 일반인들이 자세하게 알기 힘든 금융 상품들을 검색을 통해 사용자들이 알기 쉽게 만드는 것이다. 자본시장통합법 시행 이후 금융시장은 하나로 묶여서 은행과 증권과 보험의 경계가 모호해지고 다양한 금융상품을 팔수 있지만 개개인들은 이러한 상품에 대한 가치나 설명을 객관적으로 찾을 수 없다. 금융상품을 금융상품에 추가하고 빼는 식으로 금융상품은 무한대의 경우의 수를 가질 수 있으며 이는 2008년 서브프라임 모기지 사태의 CDO시장의 확대와 그에 따른 서브프라임 주택담보대출의 확대, 미국 부동산 버블과 같은 결과를 초래할 수 있다. 따라서 소비자는 똑똑하게 상품을 구매해야 하며, 이런 복잡한 금융상품들을 분해 및 비교하는 알고리즘도 필요하다. 또한 주식에 관심있는 사람들에게는 검색한 주식에 대한 Fundamental정보를 객관적으로 보여주고 technical 접근을 통해 해당 주식의 위험성과 기대이익을 확률적으로 제공한다. 또한 사용자가 검색한 주식을 기반으로 프로파일링을 통해 관심있는 테마주의 대장주를 추천해준다.  이로써 일반인들도 금융상품에 대한 이해와 접근을 용이하게 하고 개인의 자산을 지키면서 운용할 수 있는 좋은 서치엔진을 하나 만들고자 한다.

## 4. 결론 및 향후 연구

사용자는 많아지고 사용자의 쿼리는 갈수록 짧아지고 연산은 거의 쓰지 않는다. 그리고 검색엔진이 사용자가 의도한 내용을 어떻게 알 수 있는지 판단하는 것이 중요하다.

위의 내용은 SearchBoth의 기술로 사용자로 하여금 여러가지 선택 옵션을 줄 수 있게 하여 해결하였고 DogPile의 메타서치 검색을 활용하고 더해서 개인정보 데이터를 머신러닝으로 학습시켜 사용자의 의도와 관련된 문서를 찾는 정확도를 높일수 있었다. SearchBoth와 DogPile의 두 사이트의 검색엔진을 Combination 하여 금 새로운 알고리즘이 만들어 지고 이는 사용자에게 더욱 효율적이고 정확한 결과값을 보여주는 알고리즘이 되었다. 이와 관련해 더욱 좋은 알고리즘이 있다면 그 알고리즘의 장점을 취해 제안된 알고리즘에 추가하여 더욱 견고하고 정확한 알고리즘이 될 것이다.

## 5. 참고문헌

**[1] https://en.wikipedia.org/wiki/Metasearch_engine**

.