통합검색시스템의 이해 : 전자도서관 통합검색솔루션 분석

김덕현_이패스(주)

1. 개요

Federated Search System은 다양하고 산재한 웹자원에 대한 통합 일괄검색을 진행하여 이용자에게 실시간으로 정보를 제공하는 검색엔진이다. 흔히 메타검색으로도 불리우는 Federated Search System은 다방면에서 활용이 가능하고, 도서관에서도 특화하여 다양한 방식으로 활용할 수 있다.

도서관에서 활용할 수 있는 통합검색시스템의 2가지 방식인 Federated Search System과 DB Search type의 비교 분석을 통해 우리 도서관에서 적합하게 활용할 수 있는 통합검색에 대해 살펴보고자 한다.

2. 검색엔진의 검색 알고리즘 구현 방식 분석

Our main goal is to improve the quality of web search engines. In 1994, some people believed that a complete search index would make it possible to find anything easily.

(우리의 주요 목표는 검색 엔진의 품질을 향상시키는 것이다. 1994년에 사람들은 검색 인덱스를 완성하고 나면 무엇이든 쉽게 찾을 수 있을 것이라고 생각했다.)

However, the Web of 1997 is quite different. Anyone who has used a search engine recently, can readily testify that the completeness of the index is not the only factor in the quality of search results. "Junk results" often wash out any results that a user is interested in.

(하지만 1997년의 웹은 꽤 다르다. 최근에 검색 엔진을 사용해 본 사람이라면 누구나 인덱스를 완성하는 것만으로는 좋은 품질의 검색 결과를 얻을 수 없다는 것을 안다. '쓰레기 정보'가 종종 사용자들이 진정 관심있어 하는 정보를 가려버린다.)

.....

People are still only willing to look at the first few tens of results.

(사람들은 여전히 검색 결과 중 처음 몇 십 개 정도만 살펴볼 뿐이다.)

.....

In particular, link structure and link text provide a lot of information for making relevance judgments and quality filtering. Google makes use of both link structure and anchor text1).

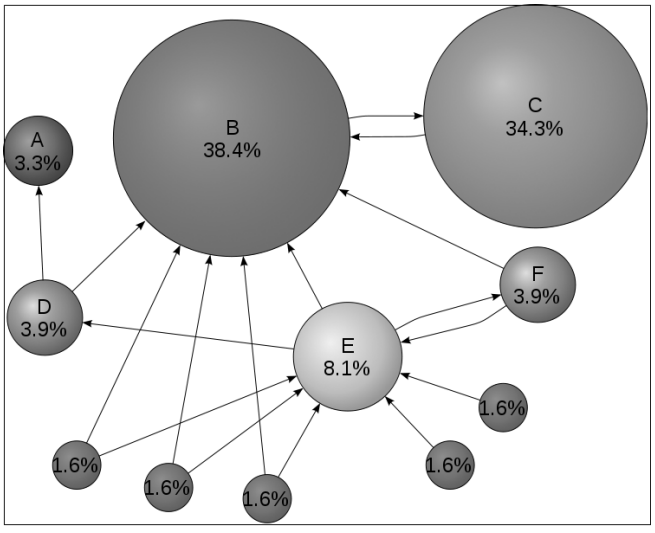
(특히, 웹 페이지 사이의 연결 관계가 상당히 유용한 정보를 제공해줄 수 있다. 구글은 이러한 링크 구조와 링크 달린 텍스트를 이용한다.)

1) Brin, Sergey and Page, Larry 1998. 『The Anatomy of a Large-Scale Hypertextual Web Search Engine』

1998년 세르게이 브린과 래리 페이지가 작성한 논문은 이렇게 시작하고 있다. "google" 이라는 기업 브랜드 가치로 474억 6,300만 달러(약 54조)의 거대 기업을 탄생시킨 시발점이 된 논문이다. 참고로 삼성그룹의 브랜드 가치는 381억 9,700만 달러(43조 5,255억원)이다[2]. 구글 이전에도 검색엔진은 존재했다. 유명한 야후부터 알타비스타 등의 검색엔진이 이미 존재했다. 야후는 여전히 디렉토리 검색으로 그 특성을 보여주고 있지만, 2012년 현재 세계 검색 엔진 시장에서 구글의 시장점유율로 놓고 보자면 그 이외에 검색엔진의 시장점유율은 매우 낮은 수준이다. 세계 인터넷 검색엔진 시장 점유율은 구글이 80% 이상의 압도적인 우위를 점하고 있으며, 야후나 빙, 중국 검색 엔진인 바이두 정도가 나머지를 차지하고 있을 뿐이다. netmarketshare의 웹 트래픽 조사 통계에 따르면 Google 82.1%, Yahoo 7.0%, Bing 4.6%, Baidu 4.3%의 점유율을 가진다. 한국의 경우를 보자면, 지금도 Web browser를 여는 순간 등장하는 네이버는 한국에서 만큼은 점유율 80% 이상이다. 구글의 엄청난 온라인 시장 점유율이 한국에서만큼은 아주 미미한 수준이다. 약 3% 정도의 시장점유율이니 말이다. 구글의 한국시장 초기에는 네이버의 데이터베이스가 막힌 상태였기 때문에 구글 등 다른 검색엔진이 수집할 수 있는 정보에는 한계가 있었다. 혹자는 네이버를 비롯한 국내 포털들이 "포털 바깥에는 쓸만한 정보가 없다"는 가정하에 모든 정보를 자신의 데이터베이스에 담으려고 했고, 검색 결과에서 자신의 포털 안에 들어있는 정보를 가장 먼저 보여주려고 하여, 좋은 정보가 생생하게 살아 숨 쉬어야 할 인터넷 생태계가 파괴되었다고 말하고 있다. 물론 여기서 구글과 네이버 중 어느 것이 우수한지 확인하려는 것은 아니다.

과연 우리 도서관들이 통합검색시스템을 바라보았을 때 어떻게 이해하면 좋을까? 구글과 네이버의 검색알고리즘을 잠시 확인하면서 이 질문에 대한 답을 풀어가고자 한다.

"PageRank extends this idea by not counting links from all pages equally, and by



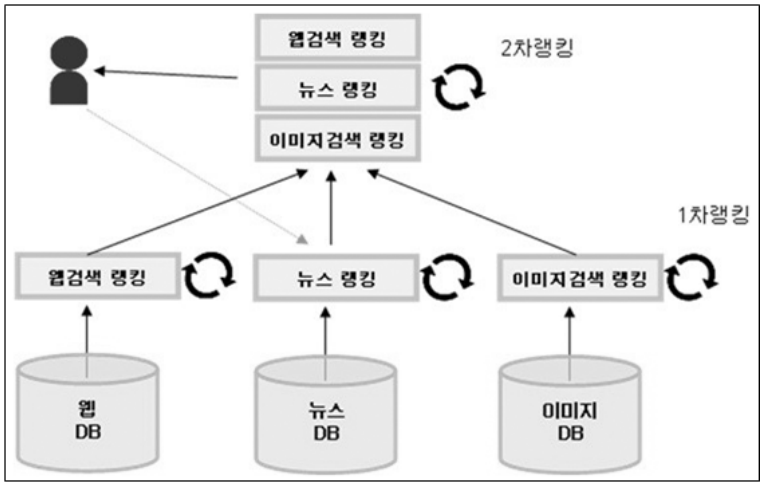
〈그림 1〉 구글의 No.1 검색 알고리즘인 Page Rank 기법

2) 세계적인 브랜드평가 컨설팅업체인 영국의 Brand Finance brand Finance Global 500 2012 참고

normalizing by the number of links on a page(...다른 페이지에서 오는 링크를 같은 비중으로 세는 대신에, 그 페이지에 걸린 링크 숫자를 '정규화(normalize)' 하는 방식을 사용한다.)"

〈그림 1〉과 같은 이 기법은 간단한 수학적인 Vector값을 구하는 것에서 차츰 복잡해진다. 간단히 요약하면, 각각의 웹 페이지는 포워드 링크(forward link; outedges)와 백 링크(backlink; inedges)를 갖는데, 이를 활용하여 상대적인 중요도를 계산하여 검색결과의 성능효과를 반영하는 것이다. 물론, 구글의 검색기법은 수 천 가지가 넘으며, 지금도 매일 새로운 검색기법이 적용되고 있다. 그럼에도 불구하고, 이 논문에서 소개된 PageRank 알고리즘은 14년이 지난 지금에도 구글 검색 엔진의 핵심을 이루고 있다. PageRank는 구글이 야후보다 월등히 좋은 검색 결과를 낼 수 있었던 비결이었고, 결국 야후를 넘어 검색 엔진의 대명사로 불리워진 출발점이었다.

네이버의 검색 알고리즘 중 흥미있는 부분은 어떤 것이 있을까? 네이버는 Collection Rank라는 검색 알고리즘을 사용하고 있다. 블로그 지식iN, 카페, 뉴스 등 통합검색의 각 컬렉션 중 어떤 것을 먼저 상위에 노출시킬지 결정하는 알고리즘이다. 이는 〈그림 2〉에 나타난 바와 같은 Multi-Ranking System이다.



〈그림 2〉 네이버의 검색알고리즘인 Multi-Ranking System

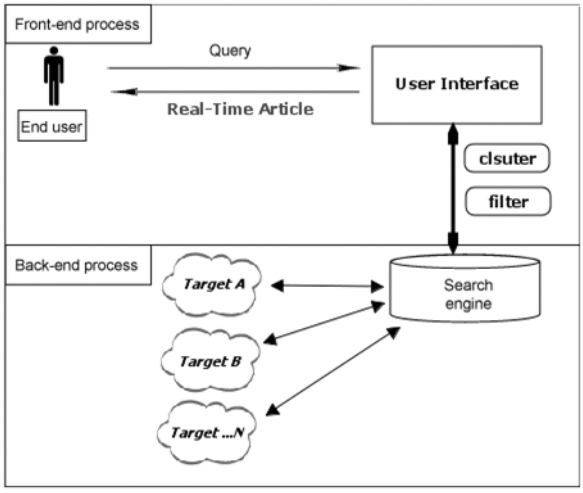
잠깐 쉬면서 네이버 검색창에 좋아하는 가수의 명칭과 FTA와 같은 키워드를 각각 검색해 보자. 각 콜렉션들이 어떤 순서로 출력되는가를 살펴보면 Multi-Ranking System의 원리를 간단히 이해 할 수 있다. 나는 원하지 않았지만, 검색질의어에 따라 뉴스 또는 웹페이지가 먼저 나오기도 한다.

3. 도서관의 검색 엔진 활용 및 검색시스템의 비교 · 분석

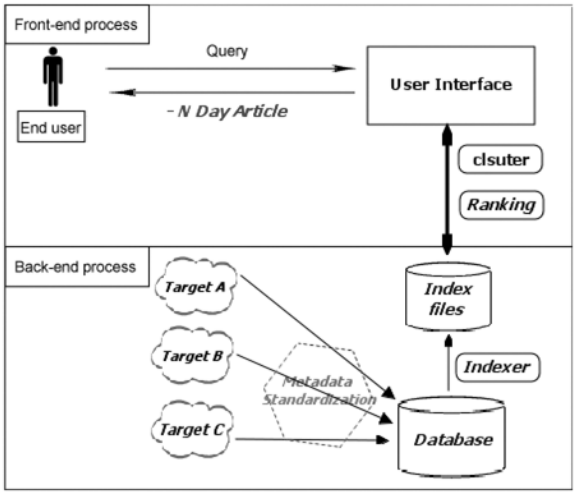
우리 도서관에 이러한 검색엔진의 개념을 도입하는 것을 생각해 보자. 최근 통합검색이라는 명칭으로 도서관에서 사용되는 검색엔진 솔루션들이 많이 나타나고 있다. 메타데이터를 사전에 DB에 저장하여 서비스하는 방식인 통합검색 솔루션들의 이용자 인터페이스(User Interface)를 보면 마치 구글의 검색창을 연상시킨다. 내가 무엇이든 입력하면 원하는 정확한 정보를 바로 보여줄 것 같다는 느낌이다. 과연 그럴까? 도서관에서 사용되는 통합검색시스템의 모호함을 검증해 볼 필요가 있다.

우선 구글과 같이 하나의 검색창에서 간단히 검색만 하면 내가 원하는 정보를 첫 페이지에 보여주는가? 우리 도서관에서 검색엔진이라는 검색솔루션은 크게 두 가지가 있다. 자관의 DB에 직접 접근하여 검색하는 Search Engine과 다양한 형식을 가진 온라인 플랫폼들에 대해 일괄 검색을 하여 이용자에게 결과를 보여주는 Federated Search Engine이 있다. 전자는 자관의 소장정보 검색을 떠올리면 되고, 후자는 흔히 메타검색이라고 부르는 검색엔진을 떠올리면 된다. 최근의 통합검색엔진들은 이를 통합하여 하나의 검색창에서 검색하면 이용자가 원하는 정보를 첫 페이지에 올려줄 것처럼 강조하고 있다. Federated Search(메타검색)와 메타데이터를 사전에 DB에 저장하여 검색을 하는 시스템은 그 search process가 전혀 다르다. 〈그림 3〉과 같이 이용자의 검색 질의어 입력과 동시에 결과를 가져오는 Real Time Search 방식과 〈그림 4〉와 같이 사전에 메타데이터를 DB에 모아서 규격화하여 색인을 거쳐 탐색을 진행하는 방식에 차이가 있다.

〈그림 4〉와 같은 DB 색인(일반적으로 Discovery type search라고 표현한다) 방식의 검색시스템은 하나의 검색창에서 모든 것을 보여줄 것 같은 의미를 전달하지만 상당한 모호함이 있다. 도서관에서의 정보검색은 일상의 흥미위주 탐색이 아니라, 문헌의 내재적인 가치를 지닌 학술정보를 찾아 활용하는 것이 그 목적이다. 여기서 출발하는 통합검색 솔루션의 문제점은 바로 정교한 학술정보에 대한 검색성능효



〈그림 3〉 Federated Search System(Real Time Search)



〈그림 4〉 Discovery Type Search System

율을 어떻게 부여해 줄 것인가 하는 것이다. 우리가 다루는 학술논문은 학문적인 논의 및 심사를 거쳐 학술지에 게재되는 것으로 구글의 pagerank를 적용할 수 없는 평면적인 문서이다. 즉 하이퍼텍스트 구조를 가지지 않는다는 것이다. 이는 곧바로 검색성능효율의 문제와 연결된다.

세르게이 브린과 래리 페이지의 논문을 떠올려 보면 "People are still only willing to look at the first few tens of results."(사람들은 여전히 검색 결과 중 처음 몇 십 개 정도만 살펴볼 뿐이다.) 학술적인 논의 및 심사를 거쳐 각 저널공급사의 온라인 플랫폼에 수록되는 학술논문이 단순한 IT적인 색인기법을 통하면 검색효율이 뒤죽박죽이 될 수도 있다. 학술지를 공급하는 유수의 출판사들은 학술논문 브라우징 검색에 많은 심혈을 기울인다. 학술논문 혹은 저자의 영향도나 최근 해당 분야의 주제(Issue) 등을 등한시 할 수 없다. 학술적인 검토가 수반된 수많은 학술지의 논문검색을 통합검색시스템에서 브라우징 할 때 실시간 탐색(Real Time Search) 방식이 아닌, 사전에 메타데이터를 DB에 모아놓고 색인을 거쳐 브라우징하는 방식의 통합검색시스템은 반드시 검색효율에 문제가 발생할 수밖에 없는 것이다.

각 학술지 고유의 학술적 검색성능 탐색을 어떤 방식으로 조정할 것인가? 사전에 규격화해 놓은 메타데이터에 어떠한 검색성능을 나타낼 것인가? "to look at the first few tens of results" 몇 가지 선택이 있는데 LIKE search, 자연어처리시스템[3] 등 IT기술을 활용할 수 있다. 물론, 이 경우에도 수십 가지 온라인 플랫폼에서 제공하는 여러 종의 학술지의 검색결과값 상호 간에 조화라는 측면을 배제해야한다.

여기에는 형태소 분석(Morphological analysis) 기법이 사용될 것이고, 형태소 분석은 그나마 한국어, 일본어보다 분석이 쉬운 영어에서조차 품사의 중의성(lexical ambiguity)이라는 문제가 있다. 명사형과 동사형이 동일한 단어를 떠올려 보면 쉽게 이해할 수 있다. 일본어의 경우에는 단어를 띄어 쓰는 공백이 없어 단어에 대한 명확한 정의를 내리는 것이 어렵고, 한국어는 교착어로 문장 속에서 활용할 때 단어

3) 윤성희 2004. 『Multiway Retrieval using Syntactic Component Extraction and Keyword Replacement』. 공학기술연구, 상명대학교 공학기술연구소
– 자연어 처리 시스템은 영어, 일본어, 중국어, 한국어와 같은 자연어를 입력 받아 이를 분석하는 시스템을 의미한다. 이러한 자연어 처리 시스템은 크게 형태소분석, 구문 형태소 추출, 구문 분석 결과로부터 검색어 추출로 나뉘며, 이 단계에서는 각각 애매성이 발생하여 이를 처리하기 위한 과정을 자연어 처리라고 한다.

(어간)에 조사나 어미(어미)를 붙여 상황에 맞게 사용하기 때문에 형태소의 품사에 대한 중의성이 매우 많다. 띄어쓰기의 오류가 있을 경우 그 중의성은 더 높아지기 때문에 검색엔진이 처리하기가 매우 어렵다. 즉 이러한 기계적인 처리로 전문 학술 논문을 검색하는 것도 문제가 있을 뿐만 아니라, 이중에서도 구문 분석 혹은 어휘 분석(Lexical Analysis)처리는 제대로 수행되지 않고 있다.

또 다른 중요한 문제는 학술 메타데이터를 모아놓고 규격화한다면 사전에 整地작업을 해야 한다. 즉, 업데이트의 문제가 발생하는 것이다. 한 가지 쉬운 예를 들면, 구글에서 웹크로링시에도 대상 사이트가 크롤링하는 범위를 정할 수가 있다. 구글에 "우리 홈페이지의 정보를 가져가지 마세요!" 라고 e-mail를 보낼 필요도 없다. 간단한 탐색로봇 설정으로 막을 수 있다.

이것을 통합검색시스템에 대입해 보면 학술 메타데이터를 사전에 모아서 규격화하여 제공하는 통합검색시스템들은 바로 그 메타데이터를 누군가에게 받아야 한다. 당장 머릿속에 떠오르는 수십개의 출판사들에서 메타데이터를 자신들의 온라인플랫폼에 업로드하면서 무관한 검색엔진에 실시간으로 보내줄 것인가? 만약에 모든 출판사가 준다하더라도 이용자가 시간차를 못 느낄 만큼 즉시 메타데이터를 규격화하여 브라우징이 가능할 것인지 의문이다.

데이트하기 좋은 곳은 한 달 늦게 검색이 되어도 상관이 없다. 원천 기술 연구에 매진하고 있는 우리 기관의 연구자가 도서관에서 제공하는 통합검색시스템이 이상 없이 검색결과를 보여주는 것으로 믿고 있다가, 한 동안 경쟁하는 외국의 어떤 연구자가 발표한 최신 논문을 파악하지 못한다면 문제는 다르다.

정확히 알려진 바는 없지만, 혹자에 의하면 구글의 웹크롤링 서버가 90만대에 이르고, 색인 서버만 4만 여대에 이른다고 한다. 참고로 구글은 그동안 공식적으로 자신들이 가지고 있는 세계 각 지역의 40여개 데이터센터의 서버 수를 공개하지 않았다. Jonathan Koomey교수가 2011년 발표한 보고서[4]에 따르면 구글이 운영 중인 데이터센터 서버는 2010년 기준으로 90만대 이상인 것으로 나왔다. 이는 세계에서 가장 방대한 서버팜(server farm)을 구축한 구글 데이터센터의 전기소모량을 계산해 서버대수를 도출한 것이다.

우리가 다루는 학술정보의 양이 구글의 정보에 비하면 턱없이 적은 양의 수준이겠지만, 마찬가지로 도서관의 통합검색솔루션 공급 기업들도 구글이 아니다. 즉, 엄청난 양의 데이터를 매순간 매초 저장하고 초 단위로 데이터를 규격화하여 색인을 거쳐 이용자에게 브라우징 할 수 있느냐는 다시 한 번 생각해 볼 문제이다. "미국전기전자학회(IEEE)는 매달 평균 5만 페이지분량을 업데이트 한다"

4) Koomey, Jonathan 2011. 『GROWTH IN DATA CENTER ELECTRICITY USE 2005 TO 2010』 A report by Analytics Press, completed at the request of The New York Times Consulting Professor, Stanford University

4. 결론

우리 도서관에 Meta Search(Real Time Search)나 Discovery type Search가 왜 필요한가? 고가의 전자저널 및 Web DB의 이용률을 높여 주어 투입비용대비 효과를 극대화하기 때문이다. 유료 온라인 플랫폼 이용에 익숙하지 않은 이용자들에게 검색의 편의성을 제공하기 위함이며, 우리 기관과 관련된 무수히 많은 세계 각 국의 연구소, 정부기관, 아카이브에 대해 일괄검색을 하기 위함이다. 검색시스템은 종착점이 아니다. 한국의 포털사이트들이 약간 다른 방향으로 가고 있지만, 검색엔진은 어디까지나 토털(Total)이 아닌 포털(Portal)이어야 한다.

즉, 도서관의 통합검색시스템은 학술정보를 찾아가기 위해 최적화한 검색도구이라야 한다. 학술정보의 근원지인 도서관에서 이용자들에게 자신 있게 서비스하고 있는 검색솔루션이라면 전문적인 학술정보의 검색성능을 왜곡해서는 안되며, 데이터의 부재가 발생하면 안된다. 우리 도서관에서 접근하고 있는 유료의 모든 정보원의 학술정보를 제공할 수 있어야 하며, 유료 정보원 이외에도 우리 기관의 특성에 맞게 전문적인 검색 카테고리 구성도 가능해야 한다. 검색엔진의 정확성과 유연성이 필요한 것이다. 이러한 명제에 실시간 탐색 방식인 메타검색시스템과 메타데이터를 사전에 모아서 규격화하는 DB 색인 방식(Discovery type System) 중 어느 것이 부합하는지 실제 검증해 볼 필요가 있다. Federated Search System 역시 검색의 신속성 측면이나 타겟 사이트의 작은 변화에도 검색의 지속성을 유지할 수 있는 자동화된 운영시스템을 갖추고 있는지 검토해 봐야한다.

도서관의 통합검색시스템은 이미 전자정보원의 공급만큼이나 필수적인 요소로 기능하고 있다. 우리 도서관에서 다루는 정보는 전문적인 학술정보임을 다시 한 번 상기하며, 구글과 네이버와는 다른 통합검색을 제공해야 한다. 다양한 국내외 학술정보에 대한 접근을 비용대비 효과 측면에서도 자관에서 운영하는 통합검색시스템이 검색결과에 잘 반영하고 있는가를 생각해 볼 필요가 있다. 우리 도서관의 통합검색시스템은 정확한 결과를 실시간으로 제공해야 할 의무가 있다.