



비즈니스 로직을 반영한 검색 랭킹 모델링

비즈니스 목적에 맞는 검색결과를 제공하는 방안

2008.09.02

모든 IT 시스템은 비즈니스의 목적과 요구사항에 맞게 구축되어야 한다. 검색시스템도 예외가 아니다. 기업의 검색 담당자는 “검색 결과 좀 좋게 할 수 없나?”라는 상사의 요구사항에서부터 “검색 결과를 부서에 따라 다르게 보여주세요”, “이번에 이벤트 상품을 검색결과 상위에 나타나게 해주세요” 등과 같은 실무자들의 요구사항들을 접하게 되는데, 이와 같이 기업의 다양한 비즈니스 요구사항을 논리화 시킨 것이 비즈니스 로직이다.

목차

랭킹 모델링 개념

랭킹 모델 트렌드

랭킹 모델링 방법론

필요한 환경들

결론

기업이 처한 비즈니스 환경이 급변함에 따라 요구사항, 즉 비즈니스 로직은 빈번히 발생하며 변경되고 다양해지는데, 검색 담당자는 그때 마다 이를 반영하기 위해 많은 어려움을 겪는다. 이러한 어려움과 함께 검색시스템 구축 경험과 지식, 전문 인력의 부족으로 인해, 사실상 대부분의 기업에서는 비즈니스 로직에 대한 충분한 고려 없이 검색시스템을 구축함으로써 효과적인 검색서비스를 제공하지 못하고 있다. 본 백서는 검색시스템에서 가장 중요한 요소 중의 하나인 랭킹 모델을 구현할 때, 비즈니스 로직을 빠르고 쉽게 반영하기 위한 방안을 제시한다.

랭킹 모델링 개념

기업 검색에 있어서 검색에 관한 이슈를 보면 검색시스템의 안정성과 검색 속도, 검색 정확도, 여러 정보원으로부터 데이터를 수집하고 다양한 시스템과 연동하기 위한 시스템 유연성, 그리고 이것들을 쉽고 편리하게 관리하기 위한 관리도구, 사용자의 여러 요구사항을 처리하기 위한 다양한 검색기능, 그리고 중요하지만 문제가 터지기 전까지는 수면 아래 묻혀 있는 보안 등이 있다. 이중에 검색 담당자 입장에서 보면 시스템 안정성과 검색 정확도가 가장 중요한 요소라 할 수 있다. 현재 검색솔루션 업체의 솔루션 완성도 측면에서 시스템적인 안정성은 어느 정도 확보 되었다고 보기 때문에 가장 중요한 이슈 중에 하나는 검색 정확도를 높이기 위한 랭킹 모델이다.

랭킹 모델은 검색 대상 콘텐츠를 검색질의(이하 검색어)에 나타난 사용자 의도에 맞게 순위화 시켜 주는 방법이다. 2004년 코넬(Cornell) 대학교 Joachims 교수의 구글(Google)을 이용한 사용자의 검색 패턴에 대한 연구를 보면 약 79%의 사용자가 상위 3개까지, 약 88%의 사용자가 상위 5개까지, 약 99%는 상위 10개 이전까지의 검

색결과만을 본다고 한다¹. 연구에 따르면 상위 10개 이후의 검색결과는 거의 소용이 없고, 상위 3개 혹은 5개 이내에 사용자가 원하는 검색결과가 있어야 만족스러운 검색서비스라 할 것이다. 그러므로, 검색결과의 순위를 정하는 랭킹 모델이 검색 정확도와 검색서비스 만족도에 미치는 영향은 절대적이라 할 수 있다.

랭킹 모델 트렌드

랭킹 모델의 트렌드는 연관성(relevance)만을 고려한 모델에서 점차 연관성뿐만 아니라 콘텐츠 자체의 질(quality)까지 고려하는 모델로 진화하고 있다. 연관성은 검색어와 콘텐츠간에 얼마나 밀접한 관련이 있는가를 나타내고 질은 콘텐츠 자체의 품질이 얼마나 좋은가를 말한다.

연관성을 고려한 랭킹 모델로 현재 가장 많이 쓰이는 것은 TF*IDF 모델이다. TF는 단어 빈도(Term Frequency), IDF는 문서 빈도의 역(Inverse Document Frequency)을 말한다. TF*IDF 모델에서 검색어와 문서의 연관성은 TF가 높고, IDF가 높을수록(IDF가 낮을수록) 커진다. 즉, 사용자가 입력한 단어가 여러 개 포함된 콘텐츠일수록 연관성이 높으며 여러 콘텐츠에 두루 쓰이는 공통적인 단어는 연관성이 적어 덜 중요하다라는 것이다. 그러나, TF*IDF, 불리언(boolean) 모델 등 연관성을 기반으로 한 랭킹 모델의 근본적인 문제점은 검색어와 연관성이 높은 콘텐츠라도 품질이 좋지 않으면 정보로서의 가치가 떨어지므로 결과적으로 검색서비스의 만족도가 떨어진다는 것이다.

따라서, 최근에는 검색어와의 연관성도 높으면서도 품질도 우수한 콘텐츠를 검색결과 상위로 올려주는 랭킹 모델이 다양하게 연구되고 있다. 최근의 모델 중에서 가장 인지도가 높은 것은 구글의 페이지랭크(PageRank)다. 페이지랭크는 웹문서 콘텐츠의 품질을 측정하기 위해 해당 웹문서에 얼마나 많은 링크가 걸려있는지를 측정한다(in-link의 개수 측정). 이는 많이 인용되는 논문은 품질이 높을 것이라는 기본 생각에서 고안되었다. 이 외에도 펄질을 통해 중복된 콘텐츠와 최신 콘텐츠가 질이 좋다는 첫눈(www.1noon.co.kr)의 스노우랭크(SnowRank), 사용자의 관심도(attention)를 반영한 나루 검색(www.naroo.co.kr) 등이 있다. 이러한 모델은 TF*IDF 모델을 이용해 검색어와 연관성 있는 콘텐츠를 추출하고 연결된 링크의 개수, 콘텐츠의 중복도와 최신성, 댓글수, 조회수, 스크랩 수 등의 콘텐츠의 질을 측정하는 기준을 사용해서 연관

¹ Joachims, Thorsten(etal.)(2004), "Eye-Tracking Analysis of User Behavior in WWW Search", Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval

성이 높을 뿐만 아니라 질도 높은 콘텐츠를 순위화해서 제공하고자 한다.

웹검색 뿐만 아니라 기업 검색의 경우도 위와 같은 트렌드가 반영되고 있다. 인터넷 쇼핑물의 상품검색에서는 검색어와 상품명 혹은 카테고리명과의 연관성을 고려하는 것뿐만 아니라 상품의 품질을 나타내는 판매지수, 상품평수, 상품평점 등과 함께, 고객은 좋은 서비스와 저렴한 가격의 상품을 선호한다는 의도를 반영하여 가격이 저렴하며 배송일이 짧고, 판매자의 만족도가 높은 상품을 상위로 올려주는 랭킹 모델을 적용하고 있다. KMS, EDMS 등 인트라넷 정보검색의 경우도 검색어와 게시물 제목, 내용의 연관성뿐만 아니라 조회수, 댓글수 등의 사용자의 관심도(attention)를 랭킹 모델에 반영함으로써 사용자에게 양질의 검색결과를 제공하고자 한다.

랭킹 모델링 방법론

랭킹 모델링의 방법을 자세히 설명하기 위해 와인 포털을 예로 들어 설명한다.

와인 판매 회사가 인터넷 판매를 위해 와인상품과 블로그와 와인 커뮤니티 등의 콘텐츠를 제공하는 와인 포털을 만들어서 서비스하고 있다고 하자. 사장님이 와인 포털을 이용하다가 검색결과가 마음에 들지 않아서 "고객에 관심을 가질 만하면서도 회사 이익에 도움이 될 와인이 검색결과에 잘 나오도록 하게"라고 검색 담당자에 지시를 했다고 하자. 여러분이 검색 담당자라면 이 뜬구름 잡는 듯한 사장님 지시를 어떻게 처리할 것인가?

먼저 비즈니스 로직을 반영하는 프로세스를 구체화해야 한다. 현재 비즈니스 로직을 검색시스템에 반영하기 위한 프로세스 모델링 방법론이 나와 있지 않은 관계로 가장 일반적인 모델링 방법론인 식스시그마(Six Sigma)의 DMAIC 방법론을 검색 시스템 프로세스 모델링에 응용해 보자. 이 방법론은 문제를 해결하기 위해 지속적으로 정의(Define), 측정(Measure), 분석(Analysis), 개선(Improve), 조작(Control)을 반복하는 방법으로 이를 응용하면 그림 1과 같이 검색시스템에 비즈니스 로직을 반영하는 프로세스를 모델링 할 수 있다.

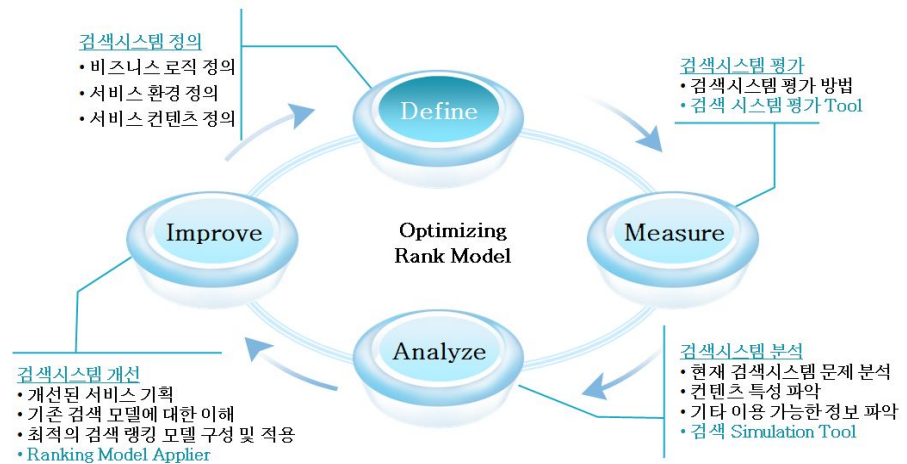


그림 1 비즈니스 로직을 검색 시스템에 반영하기 위한 프로세스 모델링

정의 (Define)

검색시스템에 관한 정의는 서비스와 콘텐츠, 비즈니스 로직 정의로 이루어진다.

- **서비스 정의**
 - ⊆ 일반인을 대상으로 한 와인 판매 포털
 - ⊆ 와인상품과 와인에 관련된 블로그, 커뮤니티 정보 제공
- **서비스 콘텐츠 정의**
 - ⊆ 와인상품 : 와인명, 국가, 지역, 연도, 타입(레드/화이트), 가격, 판매량, 세일여부 등
 - ⊆ 블로그 : 제목, 본문, 태그, 날짜 등
 - ⊆ 커뮤니티 : 제목, 본문, 작성자, 작성일, 조회수 등
- **비즈니스 로직 정의**
 - ⊆ 고객이 관심을 가질 만한 와인 상품을 상위에 배치
 - ⊆ 회사에 이익에 도움될 만한 와인 상품을 상위에 배치

측정 (Measure)

검색시스템의 검색 품질이 고객에게 만족스러운지 아닌지를 어떻게 파악할 것인가? 검색 품질은 검색 랭킹 모델에 의해 결정 되므로 검색 랭킹 모델을 평가하기 위한 방법으로도 사용할 수 있다. 검색의 품질을 측정하는 고전적인 방법은 테스트 세트를 통해 검색의 정확률(precision)과 재현율(recall)을 측정하는 것이다. 이 방법은 테스트 할 검색어 세트를 만들고 검색 대상 콘텐츠를 일일이 살펴서 테스트 검색어마다 정답 콘텐츠를 찾아서 정답 세트를 만든 후에 검색시스템이 테스트 검색어에 대해 얼

마나 정확하게 정답을 찾는지 측정하는 것이다. 이 방법을 적용하기엔 시간과 비용이 너무 많이 드는 문제점이 있다. 따라서, 사용자가 많이 검색하는 정해진 수의 인기검색어에 대해 상위 5개, 10개처럼 정해진 등수까지 검색결과 정확률을 측정하는 방법을 사용한다. 또한, 사용자가 검색한 후, 특정 검색결과를 클릭하거나, 마음에 드는 검색결과가 없을 때, 다음 검색페이지(next page)를 클릭하는 등의 클릭 정보를 활용하는 방법이 사용되고 있다.

분석 (Analysis)

분석은 기업의 비즈니스 로직을 랭킹모델에 반영할 때, 필요한 속성(feature)을 정하는 단계로써 현재의 문제점 분석을 통해 개선사항을 도출하고 기업내의 요구 사항을 반영할 수 있는 방법을 찾는 과정이다. 측정 단계에서 실시한 검색 시스템 평가 결과가 아래와 같을 때를 예로 들어 분석 과정을 설명해 보자.

- 인기 검색어에 대한 정확률 측정 결과
 - ⊆ 결과 1 : 상위 5개의 인기검색어에 대해 검색결과가 아예 없음
 - ⊆ 결과 2 : 상위 40개의 인기검색어에 대해서는 원하는 검색 결과가 나오지 않음
- 사용자 클릭 정보
 - ⊆ 결과 3 : 20%의 사용자가 첫 페이지에서 검색결과를 클릭하지 않음
 - ⊆ 결과 4 : 상위 8~10위 사이의 검색결과에 대한 클릭이 많음

우선 결과 1을 분석해 보면 검색 결과 자체가 없다는 것은 대부분 형태소 분석이 잘못되거나 유사어 확장 검색이 잘못 되는 경우에 발생하므로 검색 시뮬레이션 도구를 이용해 검색어가 어떻게 분석되고 확장되는지를 검사해서 문제를 해결해야 한다. 결과2, 결과 3는 검색 랭킹 모델이 잘못되어 원하는 검색 결과를 찾을 수 없을 때 나타나는 것으로 검색 정확도를 높이기 위한 방법을 찾아야 한다. 이 때 도움을 줄 수 있는 것이 결과 4와 같은 사용자의 로그 분석 정보나 콘텐츠 분석 정보이다. 우선 결과 4의 상위의 8위~10위의 와인 상품을 살펴 보니 현재 커뮤니티에서 이슈가 되어 고객들이 관심이 높아진 상품들로 대부분 커뮤니티에서 관련 콘텐츠의 조회수가 상위를 차지하는 상품이었다고 한다면 여러분은 이 정보를 검색시스템에 반영하여 검색의 정확률을 올릴 수 있을 것이다. 또, 구매이력정보를 이용한 데이터 마이닝 결과나 직원을 대상으로 한 설문도 도움이 될 수 있다. 예를 들어 회사 내에 “회사에 이익에 도움이 되는 상품은 어떤 것일까?”라는 질문에 대해 임직원 들이 대부분이 “판매량이 많으면서 가격이 높아 마진이 좋은 상품” 이라고 응답했다면 “회사에 도움이 되는 상품”을 찾기 위한 검색결과에 “판매량과 가격”을 반영하는 것이 좋은 방법

일 것이다.

이와 같이 분석 단계에서는 비즈니스 로직을 랭킹 모델에 반영하기 위한 콘텐츠나 서비스의 특성을 찾아 내게 된다. 예를 들면, 위의 결과로부터 “고객이 관심을 가질 만 한 것”이라는 비즈니스 로직을 랭킹 모델에 반영하기 위해서는 커뮤니티의 ‘조회수’, 또 다른 비즈니스 로직 “회사에 이익에 도움이 될만한 것”을 반영하기 위해서는 와인 상품의 ‘판매량’과 ‘가격’ 속성을 사용할 것을 결정하게 된다.

개선 (Improvement)

개선 단계에서는 분석단계에서 결정된 속성과 이를 이용한 랭킹 모델을 결정하고 검색시스템에 반영하는 단계이다. 랭킹 모델은 그림 2에서 보는 것과 같이 속성과 각 속성이 반영될 가중치(weight)를 가지는 수식으로 나타낸다. 검색시스템에서는 이런 수식을 반영하는 방법을 제공하고 있으며, 일반적인 검색시스템에서는 검색 랭킹 모델의 변화가 거의 없으므로 프로그래밍을 통해 직접 반영되기도 하고, 일부 기업용 검색 솔루션에서는 검색 랭킹 모델의 변화를 쉽게 반영하기 위해 가중치를 외부 설정 파일을 통해 설정할 수 있도록 한다. 이렇게 랭킹 모델이 결정되면 데이터 마이닝이나 검색 담당자의 경험을 토대로 가중치 초기값을 설정하여 검색시스템에 반영하며 반복적인 테스트와 가중치 조절을 통해 최적의 가중치를 도출하게 된다.

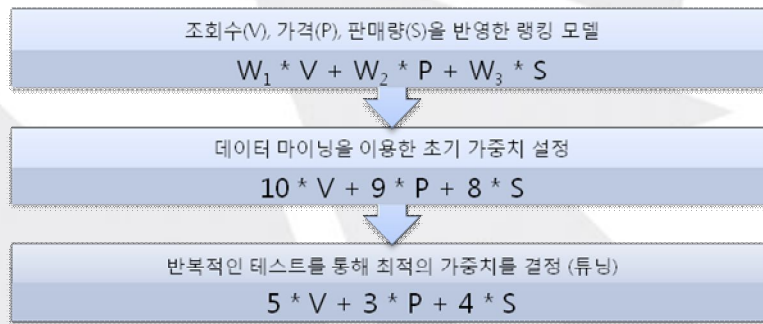


그림 2 비즈니스 로직을 반영한 랭킹 모델의 구성

필요한 환경들

지금까지 기업의 요구사항, 즉 기업의 비즈니스 로직이 검색 랭킹 모델링을 통해 어떻게 검색 시스템에 반영되는가에 대해 설명했다. 그렇다면 이런 과정은 검색시스템 도입 시 한번만 발생하는 것일까? 많은 검색 담당자는 검색결과 개선이나 새로운 서비스에 대한 요구에 빈번히 직면하게 되는데, 이것을 쉽고 빠르게 처리하기 위해서는

검색 랭킹 모델링을 쉽고 편리하게 할 수 있는 아래와 같은 도구들이 필요하게 된다.

- **검색 시뮬레이션 도구**

사용자의 검색어 입력에서 검색결과가 생성되기까지의 언어 분석, 유사어 처리, 랭킹 모델에 의한 랭킹 점수의 산정까지의 과정을 모니터링 해주는 도구이다. 프로그래밍 시 사용하는 디버그 툴과 같이 검색결과에의 문제점을 찾기 위한 필수 도구이다.

- **검색 프로파일 관리 도구**

다양한 요구사항에 대응하려면 랭킹 모델이 검색시스템 내에 다수 존재하게 된다. 개인화를 위해서는 각 개인마다 랭킹 모델이 달라질 수 있으며 KMS, EDMS 등 인트라넷에서는 부서별, 직무별로 맞춤형 검색결과를 보여 주기 위해서는 그룹별로 다른 랭킹 모델을 가져야 한다. 이와 같이 서비스별, 부서별, 직무별, 개인별로 다양한 랭킹 모델을 지원해야 하는데, 컬렉션이나 검색서버 단위로 처리하면 맞춤형 검색을 적용하기 힘들다. 따라서, 서비스별, 그룹별, 개인별로 검색 프로파일을 만들고 관리하며 이를 검색시스템에 반영할 수 있는 관리도구가 필요하다.

결론

검색 랭킹 모델링은 검색의 정확도를 높이기 위한 방법으로 현재 성공을 거둔 많은 인터넷 기업들은 각자 고유의 검색 랭킹 모델을 가지고 있다. 여러분도 제2의 구글, 제2의 아마존을 꿈꾸고 있다면 여러분 기업만의 장점을 살린 검색 랭킹 모델을 개발하여 이를 검색시스템에 반영해야 한다. 기업 검색에서 랭킹 모델에 정답은 없으며 각 기업의 비즈니스 환경을 분석하여 도출한 비즈니스 로직을 반영해서 검색 랭킹 모델을 만들어야 한다. 이는 경쟁이 심화되고 비즈니스 환경이 급변하는 상황에서 비즈니스 요구사항에 맞는 서비스를 적시에 출시하는 것이 중요하므로 신속하고 편리하게 이루어 져야만 하며 이것은 검색 담당자의 분석 능력과 이를 지원하는 검색 솔루션 업체의 기술력에 의해 좌우 된다.

다이퀘스트는

2000년 설립되어 검색전문가들로 구성된 자연어처리 솔루션 및 검색엔진 전문업체로 범람하는 정보들을 정교하고 강력한 검색기술로 정제함으로써 이용자가 원하는 정보를 손쉽게 획득할 수 있도록 새로운 검색환경을 제공합니다. 장영실상 및 대통령표창, GS 인증과 다수의 고객 사이트를 통해 그 기술력을 인정받고 있는 전문 검색엔진 업체입니다.



서울특별시 구로구 구로동 222-14번지
에이스 하이엔드타워 2차 604호
대표전화: 02-3470-4300
Fax: 02-3470-4301
문의: sales@diquest.co.kr