THEME 03

의미 기반 문서 검색 기술 동향

함 경 준 ETRI 방송통신미디어연구소 연구원 | e-mail : hahm@etri.re.kr

기업 내 업무 효율을 높이고 기업 간 협업 효율화를 위해서는 방대한 제품 설계 데이터로부터 사용자가 필요한 문서를 제공해 주는 검색 시스템 구축이 필요하다. 이 글에서는 용어의 의미 처리를 바탕으로 검색 성능을 높일 수 있는 의미 기반 문서 검색 기술에 대해 소개하고자 한다.

스마트폰의 보급은 우리를 검색 엔진과 더욱 밀접하게 하였다. 이 글을 읽고 있는 독자들도 하루에 한 번은 검색 버튼을 클릭할 것이고, 필자가 속해 있는 ETRI 방송통신미디어연구소가 무슨 기관인지 궁금한 독자는 스마트폰에 검색을 시도하려고 할지도 모른다. 검색은 일상생활뿐만 아니라 업무를 원활히 수행하기 위해 반드시 필요한 기능이기도 하다. 제품의 초기 설계부터 생산되기까지의 제품개발 과정에서 작업자들은 주어진 업무를 수행하기 위해 상당한 양의 설계 문서를 찾아 검토하게 된다. 따라서 이전에 생성되어 관리되고 있는 방대한 설계 문서 중에 밀접히 관련된 문서를 찾을 수 있다면, 작업자의 생산성 향상에 도움이 될 것이다. 검색의 과정을 단순화 하여 핵심 구성 요소와 이들 간의 관계를 도식화 해보면 그림 1처럼 나타낼 수 있다.

단순히 질의어와 문서를 매칭 시키기에는 다양한 형태로 작성된 용어 때문에 검색 결과가 만족스럽지 않다. 용어의 의미를 분석하여 하나의 형태로 바꿔 주어야 비로소 원하는 검색 결과를 얻을 수 있다.

즉, 사용자가 찾고자 하는 정보를 구체적이고 명확하게 질의어로 표현하고 이렇게 표현된 질의어가 실제로 사용자가 찾고자 하는 문서에 다행히 포함되어 있다면 사용자가 원하는 검색 결과를 얻을 수 있을 것이다. 하지만 오늘날의 분산 협업 설계 환경은 서로 다른 조직이나 팀에 속한 다양한 구성원이 문서를 작성하게 되므로 같은 의미의 단어일지라도 저마다 다른 형태로 작성될 가능성이 높다. 따라서 만족할만한 검색 결과를 얻기가 힘들어진다. 사용자에게 만족할만한 검색 성능을 제공하려면 기존의 키워드 기반 검색의 한계를 어떻게 넘어야 할 것인가?

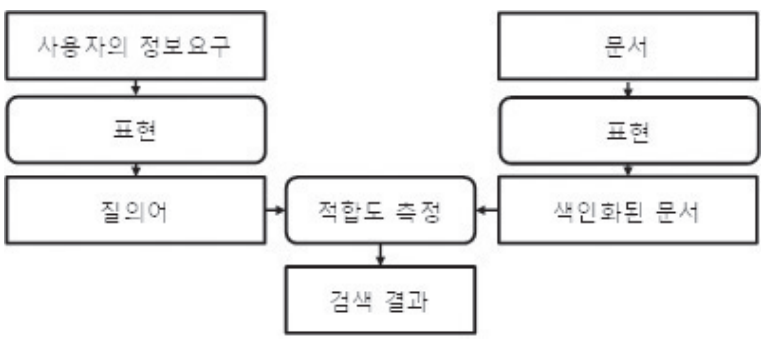

그림 1 검색의 기본 절차

의미 기반 검색에 대한 소개 : 시멘틱 프레임워크

키워드 기반 검색시스템은 구현의 용이성과 빠른 검색이 가능하다는 점 때문에 널리 보급되어 사용되고 있다. 그러나 제조 도메인에서 설계 문서는 다양한 형태와 모호한 의미의 용어로 구성되어 있으며, 사용자의 질의

어 또한 마찬가지여서 단순한 키워드 매칭으로는 원하는 문서가 검색되지 않을 것이다. 결국 사용자가 원하는 문서를 찾아내기 위해서는 사용자의 질의어가 찾고자 하는 문서에 포함되어 있거나 사용자의 질의어를 찾고자 하는 문서에 표현되어 있는 형태로 작성해야 한다. 즉, 문서 내의 용어와 질의어 간의 의미적 모호성을 해소하고 동일한 형태로 표현한다면 키워드 기반의 검색의 한계를 뛰어 넘을 수 있게 된다. 이러한 접근 방법을 의미 기반 검색 방법이라고 한다. 문서나 질의어를 구성하고 있는 용어에 대하여 의미적인 모호성을 해소하고 동일한 형태로 표현을 하여 키워드 매칭을 시켜 사용자에게 만족할 만한 검색 성능을 제공하게 되는 것이다. 의미 기반 검색 방법의 핵심사항은 용어가 갖고 있는 의미를 컴퓨터가 어떻게 정확하게 파악해 내느냐의 문제로 귀결될 수 있다. 보다 정확하게는 문서 내의 문장으로부터 적절한 형태소 처리와 품사 태깅(tagging)을 수행하여 의미를 파악할 용어를 추출하고, 각 용어에 대한 의미적 모호성을 해소하기 위해 문서 집합으로부터 추출한 통계정보를 이용한다거나 온토롤지와 같은 외부 리소스를 이용하여 정확한 의미를 파악하게 된다. 예를 들어 '배를 타고 가다 배가 고파서 배를 먹었다' 라는 문장이 있다면 문장 내에 있는 각각의 배에 대한 의미적 모호성을 해결해야 한다. 적절한 형태소 분석과 품사 태깅을 통해 예제문서 내의 주요 명사와 동사가 추출이 되면, 이를 토대로 의미 분석을 수행하게 된다. 통계정보를 이용하는 경우 방대한 문서에 대해 사람이 직접 개입하여 문맥상에서 단어의 의미를 파악하고 이를 바탕으로 일종의 학습을 시켜 확률이나 패턴을 구하게 된다. 즉 교통수단의 배인 경우 '타다' 라는 동사와 동시에 빈번히 출현하므로 위의 예제 문장에서 첫 번째 '배' 라는 단어는 확률상 혹은 패턴 매칭에 의해 교통/운반 수단으로 의미가 파악된다. 통계정보를 이용한 의미 모호성 해소 방법은 일반적인 문서나 웹 문서를 대상으로 많은 연구가 이루어지고 있으나, 설계 문서와 같은 제조 도메인에서는 전문 용어 및 용어의 축약 표현 등의 이슈로 인해 의미 분석의 정확도가 떨어지는 것으로 보고되고 있다.

반면 도메인 온톨로지(ontology)를 구축하여 용어에 대한 의미적 모호성을 해결하려는 시도가 제조 분야에 보다 적합한 접근 방법으로 여겨지고 있다. 온토롤지 구축비용의 발생이라는 점을 제외한다면, 문서 내에 사용되는 용어에 대한 체계적인 정의를 통해 정확한 의미 분석을 수행할 수 있게 된다. 각각의 용어에 대한 계층 구조와 해당 용어와 관계를 맺고 있는 용어가 정의되어 있는 온톨로지를 바탕으로 문장 내에 포함되어 있는 용어 간의 의미적 거리를 정량화할 수 있게 되고 이러한 수치를 토대로 의미 모호성이 해소된다.

이미 구축되어 사용되고 있는 기술 용어 사전을 이용한 접근도 가능하다. 기술 용어 사전에는 각각의 용어에 대한 설명이 나와 있는데, 설명 내에 포함되어 있는 단어가 해당 용어의 의미 분석에 중요한 증거가 될 수 있다. 즉, 한 용어가 문서 내에 축약된 형태로 작성되어 있어서 의미적인 모호성을 갖고 있을 때 이 용어의 문서 내 주변 용어를 분석해 보니 특정 용어의 설명에 포함되어 있는 용어들이 다수 존재한다면, 이 용어의 의미 모호성이 해소된다.

키워드 검색의 한계로부터 이를 해결하기 위한 의미 기반 검색에 대해 알아보았고 특히 용어에 대한 의미 모호성 해결 방안에 대해 더욱 자세히 알아보았다. 의미 기반 처리를 통해 문서의 용어와 사용자의 질의어를 통일된 형태로 변환하여 매칭 시킨다면 적어도 기존의 키워드 기반 방법보다 정확한 검색 결과를 제공할 것으로 예상한다. 하지만 몇 가지 추가적인 부분을 고도화 한다면 보다 월등한 검색 성능을 제공할 수 있게 된다.

검색의 고도화 : 질의어 확장, 전처리, 문서랭킹, 문서분할, 군집화

검색의 성능을 높이기 위한 방안은 여러 가지가 존재한다. 물론, 앞에서 소개된 검색의 기본 구성 요소는 큰 틀에서 변화가 없지만, 각각의 구성 요소 내에서 다양한 접근 방법을 도입하여 사용자에게 정확한 검색 결과를 제공하

기 위한 연구가 활발히 이루어지고 있다. 검색성능을 객관적으로 측정하기 위해 재현율과 정확률 척도가 흔하게 사용되며, 이 두 가지 척도를 혼합하여 사용하는 척도도 존재한다. 재현율은 전체 적합 문서 중에 검색 결과에 몇 개의 적합 문서가 있는지를 나타내는 척도이고, 정확률은 검색 결과의 문서 중에 적합 문서가 몇 개인지 나타내는 척도이다. 그렇다면, 이 두 가지 척도를 높이기 위해 사용되는 대표적인 검색 고도화 방법을 알아보기로 하자.

▶ 질의어 확장

통계에 의하면 사용자의 질의어는 평균적으로 2~3개의 키워드로 이루어진다고 한다. 이 통계는 영어권 국가의 사용자를 대상으로 측정된 수치이기는 하나, 우리나라의 경우도 비슷한 수의 키워드로 질의어를 작성한다고 볼 수 있다. 이렇게 짧은 질의어로는 사용자가 찾고자 하는 사항이 무엇인지 정확하게 표현하는 데 한계가 있다. 즉, 사용자의 정보 요구가 모호하게 표현되기 때문에 아무리 좋은 검색엔진을 갖다 놓아도 별다른 효과를 보지 못한다. 질의어 확장은 이렇게 모호한 사용자의 정보 요구를 보다 구체적이고 명확하게 표현하기 위해 의미적으로 관련 있는 키워드를 추가하여 질의어를 확장시켜 검색을 수행하는 방법이다. 이렇게 질의어를 확장시켜 검색을 수행하게 되면 일반적으로 재현율은 높아지지만, 정확률이 현격하게 저하되는 경향이 있다. 따라서 두 가지 성능 척도를 향상시키기 위해서는 의미적으로 관련 있는 키워드를 적절하게 선별하여 추가하는 것이 중요하다. 질의어 확장 방법은 수동과 자동으로 구분이 되며, 자동 확장 방법은 통계기반과 지식 리소스 기반으로 구분된다. 통계 기반은 해당 질의어와 빈번하게 동시에 출현하는 용어를 문서 집합으로부터 찾아내어 질의어 확장을 수행하는 방법이다. 지식 리소스 기반 질의어 확장은 워드넷과 같은 공용 시소러스, 용어사전, 온톨로지를 이용하여 질의어와 의미적으로 연관이 있는 용어를 찾아내어 질의어 확장을 수행한다. 온톨로지를 이용한 질의어 확장은 질의어 내 키워드들이 온톨로지 상에서 매칭되는 컨셉트를 식별한 후에, 해당 컨셉의 부모

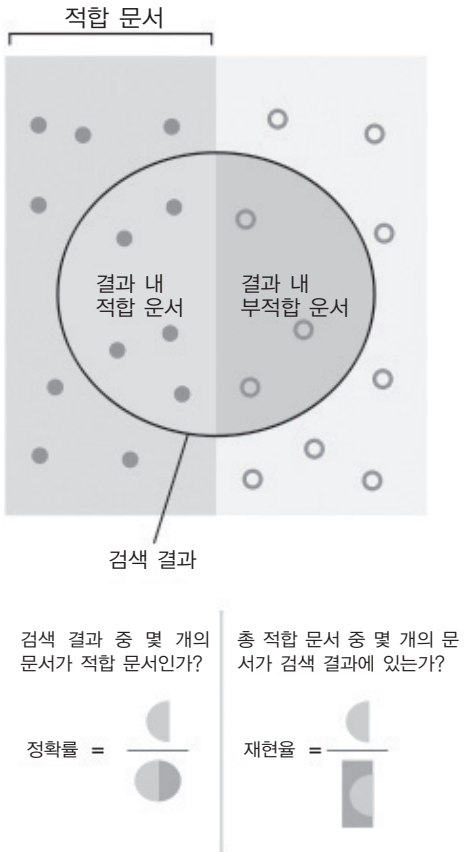

그림 2 정확률과 재현율

컨셉트, 자식 컨셉트, 형제 컨셉트, 인스턴스 등을 이용하여 질의어를 확장하는 방식으로, 설계 문서와 같이 특정 도메인을 대상으로 하는 검색의 경우 온톨로지 기반 질의어 확장 방식이 적절할 것으로 예상된다.

▶ 전처리

질의어에 대하여 사용자의 정보 요구를 구체적이고 명확하게 표현하는 것이 중요한 반면, 검색 대상이 되는 문서에 대한 정확한 전처리 과정도 검색의 성능을 높이기 위해 중요하다. 특히 한글로 작성된 문서의 전처리 과정에서 형태소 분석 및 품사 태깅은 어려운 문제로 인식이 되고 있는데 형태소 분석 및 품사 태깅이 적절하게 수행되어야 올바른 색인어가 추출된다. 복합명사가 빈번히 등장하는 설계 문서의 경우 띄어쓰기 오류로 인하여 의도하

지 않은 색인 결과를 접하게 되는데, 복합 명사에 대한 정확한 식별 또한 문서 전처리에서 중요한 이슈 중 하나이다. 전처리 과정을 구성하는 형태소 분석, 품사 태깅 등의 접근 방법들은 규칙 기반이나 통계 기반 혹은 이 둘을 혼합적으로 사용하는 하이브리드 기반으로 분류될 수 있다. 규칙 기반의 경우 특정 도메인의 문서에 맞춤형으로 구축하여 비교적 빠르게 전처리를 수행할 수 있다는 장점이 있으나 확장성이 낮고 유지보수의 어려움이 존재한다는 단점이 있다. 통계 기반의 경우 확장성과 범용성이 높다는 장점이 있으나 학습을 위해 방대한 문서를 대상으로 사용자가 수작업으로 전처리를 수행해야 하는 비용이 발생한다는 단점이 존재한다.

▶ 문서랭킹

대부분의 검색 시스템은 사용자에게 검색결과를 제공하는 과정에서 질의어에 대하여 해당 질의어를 포함하고 있는 문서를 찾아내고, 찾아낸 문서를 대상으로 사용자의 정보 요구에 대한 적합도를 정량화하여 순위를 매겨 사용자에게 검색 결과를 제공한다. 사용자에게 수십 개의 문서가 포함된 검색 결과가 제공된다 하더라도 대부분의 사용자는 상위 10개 혹은 20개의 검색결과에서 원하는 문서가 있는지 파악하기 때문에 문서랭킹은 검색 성능에 상당히 많은 영향을 미친다. 앞서 언급한 검색 성능의 척도 중에서 이러한 문서랭킹의 결과를 반영할 수 있는 척도는 없어서 이를 위해 상위 n개의 검색 결과에서의 정확률을 따지는 수정된 척도가 사용되기도 한다. 질의어에 대한 문서 적합도를 계산하는 방식은 크게 통계기반 모델, 벡터기반 모델, 그래프기반 모델의 세 가지로 나뉠 수 있다. 각각의 방식은 문서를 어떤 형식으로 표현하여 문서의 적합도를 계산하는지에 따라 나뉘게 되는데, 통계기반의 경우 확률모델로 벡터기반의 경우 벡터모델로 그래프기반의 경우 그래프 형태로 문서를 나타내어 적합도를 계산하게 된다. 문서랭킹 방식을 나누는 또 다른 기준이 있는데 문서 내 용어 간의 의존관계 고려 유무에 따라 용어 독립 모델과 용어 비독립 모델로 나뉠 수 있다. 용어 독립 모델의 경우 한 용어의 출현은 다른 용어의 출현과 관계없이 독립적인 사건이라는 가정을 전제로 한다. 문서의 적합도를 계산하는 과정에서도 질의어 내 키워드와 매칭되는 각 용어에 대한 가중치의 합을 기반으로 한다. 용어 간의 의존관계를 고려한 용어 비독립 모델의 경우 문서의 적합도를 계산하는 과정에서 질의어 내 키워드와 매칭되는 각 용어와 이 용어와 통계적으로 혹은 의미적으로 연관된 용어까지 같이 고려한다. 용어 간의 의존관계가 실제로 존재하기 때문에 용어 독립 모델의 문서 랭킹은 적절한 방법이 아닐 것처럼 예상되지만 의외로 용어 비독립 모델보다 좋거나 비슷한 검색 성능을 제공하는 것으로 알려져 많이 쓰이고 있는 상황이다. 용어 비독립 모델의 경우 2000년대에 들어 연구가 많이 이루어지면서 상당한 발전을 이룬 상태이다. 최근의 연구에 의하면, 방대한 양의 문서 학습을 통해 용어 간 의존관계를 확률모델로 구축하면 용어 독립 모델보다 더 좋은 성능을 제공하는 것으로 알려져 있다. 또한, 온톨로지를 이용하여 용어 간의 의존 관계를 의미적으로 고려하여 문서의 적합도를 평가하는 연구도 진행되고 있다.

▶ 문서분할/군집화

문서분할이나 군집화는 검색 성능 효과를 높이기보다는 사용자의 편의를 높이거나 시스템이 효과적으로 작동될 수 있도록 도와주는 접근법이다. 검색 대상의 문서가 길이가 긴 문서, 즉 많은 수의 단어로 이루어진 문서인 경우 문서분할이 필요하다. 왜냐하면 사용자의 정보 요구가 일치한다 하더라도 사용자가 찾고자 하는 부분은 문서의 일부분일 것이고, 해당 부분을 긴 문서에서 찾아내야 하는 것도 사용자에게는 상당한 부담이 된다. 문서분할은 문서를 특정 기준의 단위로 분할하여 다수 개의 하위 문서를 생성하고, 사용자에게는 하위 문서에 대한 검색 결과를 제공하는 방법이다. 분할 기준은 고정적 또는 가변적으로 설정할 수 있다. 고정적인 분할 기준은 단어 수로 분할을 수행한다. 가변적인 분할 기준은 문서의 문맥을 고려하여 분할을 수행하게 되는데 문단, 장(chapter) 등의

문서 구조를 이용하거나, 도메인 온톨로지를 이용하여 의미적인 분석을 통해 문서를 분할하는 방법도 가능하다.

검색대상이 되는 문서의 양이 방대할 경우 적절한 군집 알고리즘을 사용하여 문서를 군집화하게 되면, 검색 공간이 군집 단위로 줄어서 검색 속도를 향상시킬 수 있다. 즉, 어떤 사용자가 특정 군집의 문서를 주로 열람한다면 검색 시 해당 군집의 문서 내에서 검색을 수행하는 것이 검색의 정확도와 속도를 높일 수 있을 것이다.

검색의 성능을 높이기 위한 방법은 여러 가지가 존재하며 환경과 상황에 따라 적절히 조합하여 쓰인다. 이와 더불어 사용자 맞춤형 검색 결과가 제공된다면 검색 만족도는 배가 될 것이다.

개인화 검색

앞서 소개한 의미 기반 검색 방법과 이를 바탕으로 여러 고도화 기술을 적용한다면 충분히 체감할 정도의 만족스러운 검색 결과를 제공할 것으로 기대한다. 하지만 한 단계 더 진보된 검색 시스템을 위해서는 사용자의 검색 취향과 의도를 반영하여 맞춤형 검색 결과를 제공하는 개인화 검색 방법이 필요하다. 이미 상당수의 업체들이 개인화 검색이나 추천 시스템을 이용하여 서비스를 제공하고 있는데, 주로 음악, 비디오, 웹과 같은 일반 도메인을 위한 서비스가 대부분이다. PLM/PDM과 같은 기간시스템에 사용자의 관심 사항을 기록하는 사용자 프로파일을 생성 및 관리하게 되면, 사용자 맞춤형 검색 결과를 제공하게 되어 작업자의 업무 지원을 더욱 원활하게 할 것이다. 예를 들어, 사용자가 자주 열람하는 문서나 도면 등으로부터 핵심 키워드를 식별하여 프로파일을 구축하고, 이를 토대로 검색 결과 제공 시 해당 키워드를 포함하고 있는 문서나 도면을 상위에 위치시키면 검색의 정확도가 월등히 높아질 것이다.

현주소, 앞으로의 방향

설계 문서의 경우 예전에는 비교적 구조화된 문서가 주를 이루었다면 현재는 다양한 포맷의 비정형문서가 주를 이루며 방대해졌다. 따라서 콘텐츠 기반 검색 시스템의 중요도가 높아졌으며 상용 PLM 벤더 업체들도 PLM에서 관리되는 문서에 대한 검색의 중요성을 인식하여 유명 검색 엔진 회사를 인수하거나 최신 기술의 검색 기술을 적용하려는 시도를 하고 있다. 하지만 앞서 언급하였던 것처럼 일반적인 문서를 대상으로 하는 검색 기법을 특정 도메인의 문서 검색을 위해 적용하기에는 상당한 한계점이 존재한다. 또한 의미 기반 검색 접근을 도입 하더라도 검색 성능의 결정적인 역할을 하는 도메인 온톨로지의 구축에 있어서 업종마다 회사마다 다른 지식 구조와 체계가 존재하기 때문에 범용적으로 접근한다면 제한이 따를 것으로 예상된다. 즉 기업 간 협업 효율화를 위해서는 해당 기업별로 업종별로 특화된 도메인 온톨로지 구축과 접근 방법이 필요하다. 따라서 향후 문서 검색 기술은 업종별 특징과 성향을 반영하여 맞춤형 검색 플랫폼 개발로 진화될 것으로 예상되며, 궁극적으로는 엔지니어에게 작업에 필요한 관련 문서를 자동으로 선별하여 추천해 주는 추천시스템 개발이 이루어질 것으로 보인다.