지상(紙上) 블로그

문헌범주화(Text Categorization) : 자동 주제분류

심 경
shim@irisnet.co.kr
정보학박사
한국도서관협회 평생회원
(주)아이리스닷넷 대표

문헌범주화란 정보검색의 한 분야이다. 사실 문헌정보학 분야에서 학계와 실무사서 간에 관심 폭의 차이가 가장 큰 것이 정보검색분야일 것이다. 우리 분야의 학술지에는 매 호마다 이런 저런 주제로 정보검색을 다룬 논문들이 게재되고 있지만, 도서관 실무자들이 즐겨보는 전문지에는 정보검색에 대해 언급하는 기사들이 거의 없다. 정보검색이론은 다소의 전산이론과 수학에 대한 지식을 필요로 하여 인문사회학 성향을 가진 우리 사서들이 이해하기는 좀 어렵고, 학교에서 배운 것 이상으로 찾아서 공부하기도 내키지 않는 주제이다. 특히 우리가 매일 수행하는 업무와 직접적인 관련이 없는 것으로도 보이니, 이를 파고드는 것은 경제상식에도 어긋난다. 그러나 여기에서 문헌범주화에 대하여 한번 살펴보고자 하는 이유는 정보검색에 대해 본격적으로 다루기보다 도서관 현장에서 그것이 어떻게 "쓸모"가 있는가 하는 점을 얘기하고 싶기 때문이다.

흥미롭다고 해야 할 지, 안타깝다고 해야 할 지… 정보검색 분야에서 수많은 뛰어난 학자들이 지난 수십 년 간 열심히 연구한 결과는 40여 년 전에 크랜필드 실험(Cranfield experiments)에서 결론이 난 정보시스템의 성능인 60~90%의 재현율과 10~25%의 정확률이라는 수치에서 아직도 크게 벗어나지 못함을 보여주고 있다. 그런데 몇 년 전 정보검색기법 중 필자의 관심을 끈 것이 바로 문헌범주화

기법이다. 엄밀히 말하여 기법 자체보다는 시스템 최대 성능이 80% 중반에서 92%에까지 이르는 결과의 정확성을 보여준다는 점이었다. 90%가 넘는 정확성에 도달할 수 있는 기법이라면 어느 업무 분야에 적용되더라도 충분히 수작업을 대치하거나 수작업에 대한 보완수단으로 채택할 가치가 있어 보였다.

지금까지 우리 분야의 전산화는 흔히 도서관자동화 회사에서 개발 · 공급하는 제품을 이해하고 구매하는 데 급급해 왔지만, 이제는 우리가 자동화 회사에 이러저러한 제품을 생산 또는 구현해 달라고 주문하는 개발 주도권을 가질 때가 된 것 같다. 아니 어쩌면 진작부터 그랬어야만 했다. 따라서 이 글에서는 문헌범주화 기법의 원리를 간단히 설명하고 그 기법의 적용분야를 예시함으로써 독자들이 실무에 적용하기 위한 창의성을 발휘하는 데 도움이 되었으면 한다.

문헌범주화 기법의 원리는?

우리는 신문에서 '안정환' 이나 '박지성' 이라는 문구를 보면 그것이 '스포츠' 관련 기사라는 것을 금방 알아차릴 수 있다. 문헌범주화란 그처럼 어떤 글의 내용을 충분히 이해하지 않고도 그 글의 범주(즉, 주제)를 쉽게 알아내는 인간의 능력을 기계에 학습시켜 분류업무를 자동화하는 것이다. 즉, 문헌범주화는 문헌분류를 자동화하는 기법이다.

잠시 이 기법의 개념을 우리가 이미 알고 있는 다른 정보검색기법들과의 차이점을 비교해보자. 문헌범주화는 정보검색의 한 분야로 다루어지기는 하지만, 그 기법 자체가 이용자 질의문을 기반으로 특정 결과세트를 제공하는 검색기법과는 전혀 다르다. 오히려 검색 이전 단계에서 데이터베이스에 있는 문헌 중 유사한 주제성격을 가지는 문헌들끼리 미리 모아놓는다는 관점에서는 문헌 클러스터링(document clustering)[1]과 비슷하다. 즉 두 가지 기법 모두 자동으로 문헌을 분류하는 자동분류(automatic classification)에 속한다. 단지 차이점은 문헌 클러스터링과는 달리 문헌범주화는 미리 정해진 일련의 범주(또는 주제)에 문헌을 분류하거나 문헌에 범주를 부여하는 것이며[2] 학습을 위한 데이터가 미리 분류되어 있다는 점이다. '미리 정해진 일련의 범주나 주제' 란 사람이 문헌을 분류할 항목을 미리 구성해 놓은 분류표를 말한다. 예를

1) 클러스터링이란 정보검색모델의 하나로 벡터(vector)모델이 검색을 위하여 질의문과 개별 문헌 모두를 하나씩 비교하여 검색하는 비효율성을 개선하기 위하여 문헌집단을 대상으로 유사한 주제를 가진 문헌들을 그룹화하여 질의문과 클러스터된 문헌 그룹을 비교하는 정보검색기법이다. 따라서 검색이전에 유사문헌들의 클러스터가 형성되는 것을 전제로 한다. 그런데 개별 클러스터는 "미리" 정해놓은 '경제', '역사', '언어' 등의 주제명을 가지지 않는다.

2) 전자를 범주중심(category-pivoted) 범주화, 후자를 문헌중심(document-pivoted) 범주화라고 하며 이는 개념적이라기 보다는 실용적 구분일 뿐이다.

들면, 신문의 사회, 경제, 정치, 스포츠 등과 같이 미리 정해진 범주를 말하며 문헌(기사)은 이와 같이 미리 정한 하나 또는 그 이상의 개별 범주에 자동으로 할당된다.

이를 가능하도록 해 주는 것이 기계학습(machine learning)이다. 기계학습이란 일반귀납과정(general inductive process)이 미리 분류된 문헌집단에서 각 범주의 특징을 학습하여 텍스트 분류기(classifier)를 자동으로 생성하는 것을 말한다. 여기까지 읽어보면 일단 '기계학습' 이나 '일반귀납과정' 이란 표현에 금방 부담감을 느낀다. 그렇다면 일단 이런 용어들은 접어놓고 우리가 이미 알고 있는 정보검색의 기본부터 다시 한번 더듬어 보자.

정보검색에 사용되는 거의 모든 기법은 문헌에 출현하는 용어의 빈도에 관련된 통계를 기반으로 한다. 이는 1957년 룬(H.P. Luhn)이 한 문헌에 나타난 용어의 출현빈도가 그 문헌의 주제를 표현하기 위한 용어의 중요성을 나타낼 수 있다고 한 것을 말한다. 이를 좀더 쉽게 말하면 어느 문헌에 출현한 용어들을 그 출현빈도 순으로 정렬하였을 때, 출현빈도가 너무 높거나 너무 낮은 용어를 제외한 나머지, 즉 중간빈도로 출현한 용어들이 그 문헌의 주제를 표현하는 용어로 유용하다는 것이다[3]. 문헌범주화 역시 이러한 용어의 출현빈도에 기반하고 있다.

앞서 문헌범주화는 일반귀납과정을 통한 기계학습으로 미리 정해진 일련의 범주나 주제에 문헌을 할당한다고 하였는데, 그림 1의 좌측이 바로 '일반귀납과정' 으로 '기계학습' 을 하여 분류기를 구축하는 과정이다. 이 과정의 결과를 '분류기' 라고 하며(그림 1 우측하단), 문헌을 이 분류기에 통과시키면 그 문헌은 미리 정해놓은 '경제', '역사', '언어' 라는 주제분야 중 하나로 판명되어 주제명이 부여된다. 이것이 문헌범주화시스템(분류기)이 하는 역할이다.

그렇다면 분류기를 구축하는 과정에서 일반귀납과정과 기계학습이란 무엇인가? 학문적인 정의를 배제하고 설명하면, 위에 언급한 역할을 수행하기 위하여 컴퓨터는 미리 분류된 문헌들을 학습한다. 즉, 수작업으로 '경제' 라는 주제에 분류된 수많은 문헌에 출현한 용어들을 추출하여 '경제' 라는 범주에 속한 문헌들이 주로 어떤 용어를 가지는 가를 배운다. 이때 많은 사례들을 가지고 분류기를 형성하므로 귀납과정[4]이라 하고 이러한 과정을 통해 기계가 스스로 특정 규칙이나 패턴을 학습하므로 기계학습이라

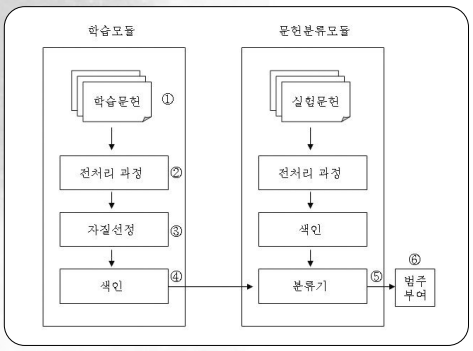

그림 1. 링킹시스템 개념도

3) 물론 용어의 주제중요도를 나타내는 가중치는 이를 출발점으로 다양한 방법으로 표현된다.
4) 고등학교 때 우리는 '공자는 죽는다', '공자는 사람이다', 따라서 '사람은 죽는다' 와 같이 개별적인 사례에서 일반적인 원리를 찾아내는 것을 귀납법이라고 배웠다.

한다. 이와 같은 기계학습을 가능하도록 하는 알고리즘의 종류는 많으나[5] 이는 우리가 고민할 필요없이 전문가에게 맡기면 될 것이다. 단, 우리가 알아야 하는 것은 이들 중 어느 방법이 가장 효율적이며 효과적인가 하는 것이다.

시스템 성능은 어떻게 측정하고 최고 성능을 보이는 분류기는 무엇인가?

문헌범주화는 이미 분류된 문헌들을 활용하여 기계가 학습하도록 함으로써 분류기를 구축하고, 그 분류기로 하여금 새로운 문헌을 분류하도록 한다. 그러면, 그 분류기의 성능은 어떻게 측정할까? 이미 분류된 문헌이란 수작업으로 분류된 문헌을 말하는데, 분류기 입장에서는 이미 답을 가진 문헌들이다. 그렇다면 이와 같이 답을 아는 문헌을 적절히 나누어 일부는 분류기 구축에 사용하고 나머지를 성능실험에 사용할 수 있다. 이때 전자를 학습문헌집합(training set)이라 하며 후자를 실험문헌집합(test set)이라 한다. 그러므로 학습문헌으로 분류기를 구축하고 그 성능은 미리 남겨놓은 실험문헌을 가지고 분류기가 분류를 하도록 한 후 그 결과를 수작업으로 부여한 범주(또는 주제명)와 비교함으로써 그 분류기의 분류성능을 알 수 있다.

그렇다면 당연한 궁금증은 최소 몇 개 정도의 학습문헌을 사용해야 적절한 성능을 가진 분류기가 구축되는가 하는 점인데 이는 실제 상황에 적용하기 위하여 아주 중요한 요소이기 때문이다. 필자가 kNN분류기를 사용하여 실험해 본 결과로는 100개의 학습문헌을 사용하는 것이 최적의 성능을 보였다[6]. 따라서 100개 정도의 수작업 분류문헌을 확보하면 새로운 문헌들은 자동분류가 가능하다는 말이 되며 이는 엄청난 업무효율을 의미한다.

앞서 문헌범주화에는 다양한 알고리즘이 사용된다고 하였다. 이들 중 어느 것을 실무에 적용할 것인가에 대한 평가요소는 일단 그 효과성(즉, 성능 또는 정확성)이 우선이고 그 다음이 시스템 효율성일 것이다. 즉, 전자는 분류정확도를, 후자는 얼마나 많은 문헌을 최소의 시스템 자원으로 처리할 수 있는가를 의미한다. 과거 연구 결과에서 지지벡터기가 가장 높은 성능을 보였으며(84~92%), kNN분류기가 그 뒤를 이었다(82~86%)[7]. 지지벡터기는 성능이 가장 높은

5) 그 종류는 선형분류기(linear classifiers), 확률분류기(probabilistic classifiers), 사례기반 분류기(example-based classifiers), 신경망(neural networks) 분류기, 지지벡터분류기(support vector machines) 등 다양하다. 선형분류기의 대표적인 예는 로치오 분류기(Rocchio's algorithm)가 있으며 확률기반분류기에는 나이브(Naive Bayes) 분류기, 사례기반분류기로는 kNN 분류기가 있다.

6) 심경 (2006). 문헌범주화에서 학습문헌수 최적화에 관한 연구. 정보관리학회지, 23(4), 277-294.

7) 로이터 컬랙션이라고 일컬어지는 표준 실험문헌집단을 대상으로 한 실험들이다. (출처: Sebastiani, F. (2002). Machine learning in automatic text categorization. ACM Computing Surveys, 34(1), 1-47.)

반면 kNN분류기는 개념이 간단하고 학습 효율성이 뛰어나므로 필요에 따라 이들 중 하나를 선택하는 것이 현명할 것이다.

문헌범주화는 어디에 적용되나?

문헌범주화가 적용되는 분야는 스팸메일 필터링, 뉴스기사 필터링, 뉴스기사 등의 주제선정(topic spotting), 웹 문헌 분류, 문학작품의 저자확인, 단어의 중의성 해소 등 범위가 다양하다. 스팸메일 필터링에 어떻게 문헌범주화 기법이 적용되는가 의아해 할 수도 있겠으나 이는 들어오는 이메일 중 스팸메일과 정상메일의 내용을 분석하여 '스팸메일'과 '정상메일'의 두 가지 범주로 나누는 것이다. 따라서 요즘 스팸메일들은 이러한 필터링을 피하기 위하여 제목을 일반메일과 유사하게 만들고 내용은 주로 하이퍼링크를 넣어 스팸메일 필터링을 방해하고 있다.

도서관 환경에서는 이와 같은 기능보다는 문헌을 주제에 따라 자동 분류하는 것에 관심이 있을 것이다. 이 기법이 매우 유용하게 적용될 수 있는 예로는 통합데이터베이스에 저장된 색인 · 초록 레코드를 주제별로 이용자에게 제공할 경우를 들 수 있다. 특히 학술지에 실린 연구논문이나 연구보고서 등의 색인 · 초록 레코드에 범주를 부여하는 것(주제분류)은 주제전문가에게도 많은 시간과 노력을 요하는 고난이도 작업이다. 이러한 경우 100여 개 정도의 문헌을 정확히 분류하여 이를 학습문헌으로 활용하면 신규문헌에 대한 분류업무는 훨씬 단순, 효율적이 될 것이다. 더욱이 문헌범주화는 한 문헌에 단일 범주뿐만 아니라 다수의 범주부여도 가능하므로 그 유용성이 더욱 크다.

그 밖에 근래 웹에서 수집한 문헌을 도서관 홈페이지에 주제별로 분류하여 제공하는 서비스가 증가하는 추세인데 문헌범주화는 이들을 실시간으로 분류할 뿐 아니라, 계층별로 분류하는 것도 가능하기 때문에 도서관에서 매우 쓸모 있다고 할 수 있다.

결언

서지레코드에 오탈자가 하나만 발견되어도 지적 대상이 되는 것이 우리 도서관의 업무이다. 그런데 정확도 80~90%를 믿고 자동분류를 수행하는 것은 상당한 위험을 감수해야 하는 것처럼 보인다. 하지만 수작업 분류 또는 색인과정에서의 색인 일관성 결여(indexing inconsistency)는 널리 알려진 문제이다. 색인 일관성 결여란 서로 다른 색인자가 동일한 문헌을 색인(또는 분류)하였을 때 일관성이 결여되고(inter-indexer inconsistency) 심지어 동일색인자

가 일정한 시간간격을 두고 한 문헌에 부여하는 색인어 조차도 일관성이 없다는(intra-indexer inconsistency) 것으로서 Saracevic[8]의 연구에 의하면 그 일관성 정도는 10~80% 사이라고 한다. 이 결과를 기준으로 보면 전문 색인자에 의한 수작업 색인작업의 성능도 어차피 100%에 미치지 못하므로 문헌범주화 기법은 전문가에 의한 색인작업에 필적하거나 더 나은 정확도를 보인다고 할 수 있다.

요즘의 데이터 수량의 증가와 시스템의 발달로 빠른 시간 내에 많은 정보를 쉽게 검색할 수 있는 반면 수작업에 워낙 많은 시간과 노력을 필요로 하는 데이터의 주제분류 서비스가 점점 사라지거나 이를 낮은 가격의 외주에 의존하는 경우가 있다. 필자가 조사한 한 국가정보기관은 색인 · 초록 레코드에 주제명을 부여하는 업무를 외주에 의존하고 있는데, 그 분류의 정확도는 겨우 16%에 미칠 정도였다[9]. 따라서 문헌범주화는 증가하는 데이터에 반하여 제한된 인력만을 가진 주제분류 업무에 훌륭한 대안이 될 것으로 기대된다. 또한 90%가 넘는 분류 정확도에도 만족할 수 없을 경우, 문헌범주화 결과를 반드시 최종 분류로 볼 것이 아니라 그 결과를 참고로 하여 수작업을 진행한다면, 분류의 품질향상과 업무효율성 제고라는 목표를 동시에 달성할 수 있을 것이다.

8) Saracevic, T. (1991). Individual differences in organizing, searching and retrieving information. Proceedings of the 54th Annual Meeting of the Society for Information Science, 82-86.

9) Kyung Shim, Young-Mee Chung (2006). The effect of the quality of pre-assigned subject categories on the text categorization performance. Journal of the Korean Society for Information Management, 23(2), 266~285.