녹색기술문헌 자동 범주화를 위한 문서 분류기 개발

주원균*, 박민우*, 최기석*
*한국과학기술정보연구원 R&D시스템개발실
e-mail:joo@kisti.re.kr

Document Classification of Green Technology Literature based on Support Vector Machines

Won-Kyun Joo*, Min-Woo Park*, Ki-Seok Choi*
*Dept. of R&D System Development, KISTI

요 약

최근에 이슈화되고 있는 녹색기술문헌의 중요성에 부합하여 녹색기술 문헌을 자동으로 분류해주는 문서 분류시스템 개발하였다. 분류체계로는 14개의 관심 녹색기술 분류 체계를 선택하였고, 다양한 문서 분류 기법 중 SVM(Support Vector Machine)에 기초를 둔 방법을 이용하였다. 문서 벡터를 생성할 때 제목과 본문에 동일한 가중치를 적용하는 방법을 벗어나서 제목의 키워드에 좀 더 높은 가중치를 부여하는 방식을 적용하여 성능평가를 수행하였다.

1. 서론

에너지와 환경, 일자리와 성장동력, 국토개조 및 생활혁명을 포괄하는 신발전 패러다임인 녹색성장이 본격적으로 이슈화되면서 수많은 양의 녹색기술관련 문헌들이 쏟아지고 있다. 이러한 상황에서 녹색기술관련 문헌을 체계적으로 분류하고 서비스를 제공할 수 있는 시스템 개발에 대한 필요성이 증가하고 있다.

본 논문에서는 이러한 시대의 요구에 부합하여 녹색기술문헌을 체계적으로 분류하고 분류된 각 범주에 맞게 문서를 자동으로 분배하는 시스템을 제안한다. 논문은 다음과 같이 구성된다. 2장에서는 관련 연구에 대해서 설명하고, 3장에서는 녹색기술문헌 및 관심 분류 체계에 대해서, 4장에서는 실험 및 분석에 대해서, 5장에서는 결론 및 향후 연구과제에 대하여 논한다.

2. 기존 연구

자동 문서분류 시스템은 문서의 내용을 바탕으로 미리 정의된 분류에 맞게 문서를 학습시킨 후에 새로 입력된 문서를 기존의 분류 성향에 맞게 구분지어서 분별하는 시스템이다. 문서 분류를 위해서 사용되는 알고리즘은 베이지언 네트워크, 최대 엔트로피 모델, 결정 트리, 신경망 등과 같은 다양한 기법들이 연구되어 왔으나 최근에는 지지벡터기계(SVM, Support Vector Machine)를 이용한 문서 분류 시스템의 연구가 활발히 수행되고 있다[1,2,3,4]. SVM은 신경망보다 사용하기에 쉬운 방법임에도 SVM에 친숙하지 못해 만족할 만한 결과를 얻지 못하는 경우가 많은 것으로 알려져 있다. SVM 분류기는 많은 양의 데이터와 높은 차원의 자질집합을 가진 분류작업에 특히 우수한 성능을 보인다.

하지만 스팸메일 필터나 신문기사 분류와 같은 다양한 응용 분야에 적용되고 있음에도 불구하고 최근 들어 강조되고 있는 녹색기술문헌의 범주화에 관한 연구는 아직 시도되고 있지 않고 있다. 이에 본 논문에서는 SVM을 이용하여 최근 부각되고 있는 녹색기술문헌을 효과적으로 분류하는 방법을 제안한다.

3. 녹색기술문헌 및 관심 분류 체계

대통령직속 녹색성장위원회에서는 녹색기술에 관련된 27대 중점기술을 발표하였다[5]. 본 논문에서는 이렇게 발표된 27대 중점 세부 기술 중 성격이 비슷한 분류들을 통합하고 관심 대상 분류만을 추출하여 14개의 분류로 재구성 한다. <표 1>은 14개의 관심 대상 분류 체계를 보여준다.

<표 1> 관심대상 녹색기술 분류

분류코드	분류이름
GT01	기후변화 예측 및 영향 평가
GT02	재생에너지
GT03	원자력/핵융합
GT04	수소연료전지
GT05	친환경 제조 공정/소재 효율성 향상
GT06	화석연료 활용성 향상 및 고효율화
GT07	수송부문 효율성 향상
GT08	녹색국토
GT09	친환경 제조 공정/소재 효율성 향상

GT10	전력 효율성 향상
GT11	대기오염 모니터링 및 제어
GT12	수질환경
GT13	폐기물
GT14	환경보건

4. 실험 및 분석

녹색기술문헌을 위한 분류기의 성능 평가를 수행하기 위해서 KISTI에서 제공하는 글로벌 동향브리핑 컬렉션 중 녹색기술에 특화된 데이터[6]를 사용한다. 본 컬렉션은 <표 1>에서 분류되어 있는 것과 같이 총 14개의 분류 정보를 가지고 있으며, 중복 분류를 허용하는 데이터베이스다. 문서 개수는 3,543건이지만, 중복으로 분류되어 있는 문서도 존재하기 때문에 실제적인 학습 문서의 개수는 3,553건이다.

평가는 SVM을 이용할 때 일반적인 녹색기술문헌 분류에 대한 성능과 데이터 내의 각 항목의 영향력 측정의 2가지 방향으로 진행한다. 첫 번째로 제목과 본문을 구분하지 않고 문서 내에 존재하는 모든 키워드를 대상으로 문서 분류 실험을 수행한다. 두 번째로 제목과 본문을 구분하여 제목에 존재하는 키워드에 대하여 조금 더 가중치를 주어 실험을 수행한다. 일반적으로 문서의 본문에 존재하는 키워드보다는 문서의 제목에 존재하는 키워드의 중요성이 더욱 크기 때문에 제목 키워드를 따로 구분하여 좀 더 가중치를 주는 방식으로 실험을 수행한다.

평가를 위한 SVM은 Chih-Chung Chang 등이 구현한 LIBSVM[7]을 이용한다. LIBSVM에는 선형(linear), 다항식(polynomial), 레이디얼 기초기능(RBF, radial basis function), 시그모이도(sigmoid)의 4가지의 커널이 구현되어 있는데, 경험적으로 RBF가 우수한 성능을 보이고 있어 이 커널을 사용한다. 문서 벡터의 구성 방법은 수식(1)과 같이 일반적으로 많이 사용하는 TF*IDF 방식을 사용한다.

$$Term\ Frequency(TF) = \frac{\text{단어 발생 빈도}}{\text{문서사이즈(문서 내 단어수)}} \quad (1)$$

$$Inverse\ Document\ Frequency(IDF) = \log\frac{\text{문서 개수}}{\text{발생한 문서 개수}}$$

$$Weight = TF \times IDF$$

$$Normalized\ Weight = \frac{Weight}{\sqrt{\Sigma(Weight)^2}}$$

최종적으로 위에서 선택한 커널과 벡터 추출방식을 사용하여 10겹 교차평가를 수행한다. 일반적으로 SVM을 이용한 실험에서는 대부분 n-겹 교차평가(n-fold cross validation)를 통해서 가장 높은 성능을 나타내는 정규화 인자를 선택한다. 많은 연구에서 10겹 교차평가를 선호하며[8], 본 논문에서도 동일한 방식을 선택한다. 실험 결과는 <표 2>에 제시한다.

<표 2> RBF 커널을 이용한 실험 결과

	옵션 값		결과 값
	cost(C)	gamma(γ)	accuracy
제목과 본문의 키워드 가중치를 동일하게 설정	2048	0.00048828125	79.57%
제목의 키워드의 가중치를 본문 키워드의 가중치보다 높게 설정	32768	3.0517578125e-05	81.11%

본 실험을 통해서, 녹색기술문헌을 분류할 때 제목에 가중치를 높게 설정하는 것이 성능 향상에 좀 더 기여할 수 있음을 확인하였다.

5. 결론 및 향후 연구

본 논문에서는 최근에 급부상하고 있는 녹색기술문헌의 중요성에 부합하여 해당 문헌을 자동으로 분류해주는 문서 분류 시스템을 개발하였다. 또한 문서 분류 시스템의 문서 벡터를 생성할 때 제목과 본문에 동일한 가중치를 적용하는 방법을 벗어나서 제목의 키워드에 좀 더 높은 가중치를 부여하는 방식을 적용하여 성능향상을 입증하였다.

향후 연구로는 각 분류의 세부 성능 기준을 파악하여 여기서 미비하게 나온 분류의 문헌을 집중적으로 분석하여 해당 분류의 성능을 올리는 방안이 필요하다. 또한 문서 벡터 생성 시에 해당 키워드가 각 클래스 별로 얼마나 중요한지를 함께 파악하여 분류별 키워드의 가중치를 조절하는 방안이 필요하다.

참고문헌

[1] Vapnik, V., "The Nature of Statistical Learning Theory", Springer-Verlag, 1995.

[2] 오장민, 장병탁, 김영택, "SVM 학습을 이용한 다중 클래스 뉴스그룹 문서 분류", 한국정보과학회 가을 학술발표 논문집(II), 26(2), pp.60-62, 1999.

[3] 정영미, 임혜영, "SVM분류기를 이용한 문서 범주화 연구", 정보관리학회지 제17권 제4호, pp.229-248, 2000.

[4] 운용옥, 이창기, 이근배, "지지 벡터 기계를 이용한 계층적 문서 분류", 2003년도 제15회 한글 및 한국어 정보처리 학술대회 논문집, pp.7-13, 2003.

[5] 27대 중점기술, http://www.greengrowth.go.kr/www/policy/skill/skill.cms

[6] 글로벌 동향브리핑, http://green.kosen21.org/data/global/global_list.jsp

[7] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[8] 최성필, 최윤수, 정창후, 맹성현, "구문 트리 가지치기 및 소멸 인자 조정을 통한 트리 커널 기반 단백질 간 상호작용 추출 성능 향상", 정보과학회논문지, 제37권 제2호, pp.85-94, 2010