과학기술 문헌 기반 주제 분석

정한민° 강인수 성원경
한국과학기술정보연구원 정보서비스연구팀
{jhm, dbiask, wksung}@kisti.re.kr

Topic Analysis based on Science and Technology Articles

Hanmin Jung° In-Su Kang Won-kyung Sung
ISRL Lab. KISTI

인터넷의 발달과 함께 접근 가능한 말뭉치가 전 분야에 걸쳐 급격히 늘어나고 있다. 특히, 과학기술 분야는 매우 빠른 발전 속도를 보이는 동시에 세부 분야 간 융·복합 현상도 빈번히 일어나는 특징을 가지고 있다. 과학기술정보 말뭉치로부터 상기 특성을 분석해 내는 작업은 연구 주제 추이를 분석하고 주제 간 연관 관계를 파악하는 것과 밀접한 관련이 있다. 또한, 주제의 효용성을 높일 수 있는 방안 마련이 필요하지만 아직까지 주제가 실제 응용 시스템에서 어떻게 활용될 수 있는지를 연계하여 연구하는 일이 드물다. 본 연구의 목표는 대용량 말뭉치로부터의 주제 추출 및 분석을 통해 불용어 사전 구축, 주제 추이 시각화, 연관 주제 제시 등의 응용에 주제가 어떻게 활용될 수 있는 지와 연구개발 전주기 지원 시스템인 OntoFrame에서 주제가 어떠한 역할을 할 수 있는지를 사례를 통해 살펴보는 것이다.

실험 대상 말뭉치가 영어로 구성된 대용량 Citeseer 말뭉치이므로 과학기술 분야의 영어 시소러스를 처리 범위 (Coverage)를 보장하는 수준으로 구축하기 어려운 점을 고려하여 말뭉치로부터 추출된 전문용어 사전을 이용한다. 주제 추출을 위한 자질로는 문서로부터 추출된 색인어 목록, 말뭉치 (제목+초록)로부터 추출된 전문용어 목록, 색인어 별 어휘 빈도 (Term Frequency), 색인어와 매칭된 전문용어 별 어휘 빈도가 있다. 개념어 별 시소러스 깊이와 개념어 별 개념 패싯 (Conceptual Facet)은 전문용어 사전이 시소러스와 같은 계층 구조를 가지고 있지 않아 이용하지 않는다. 주제간 관계 설정은 한 문서 내에서 선정된 상위 N개 주제를 상호 연결하는 방식으로 이루어진다. 공기 (Coocurrence) 정보의 하위 집합이나, 문서를 대표할 수 있는 주제에 대해서만 주제간 관계를 설정하기 때문에 주제 대표성과 유의미한 주제 추이를 관찰하는데 도움이 된다.

본 연구에서는 주제 분석을 3가지 관점에서 보고자 한다. 불용어 사전 구축, 주제 추이 시각화, 연관 주제 제시가 여기에 해당된다. 불용어 사전 구축을 위한 불용어 선정은 주제가 가지는 관계 수를 이용한다. DF (Document Frequency)가 극단적인 값 (최소값과 최대값)을 가질수록 불용어일 확률이 높아진다는 유사한 실험 결과가 있지만, DF와 주제가 가지는 관계 수가 일치하는 것은 아니며, 상호보완적으로 적용할 필요가 있다. 주제 추이 시각화는 특정 주제와 관계를 가지는 주제들이 시간 추이에 따라 변화하는 모습을 시각화시키고 관찰하는 것으로, Tag Cloud 등을 이용할 수 있다. 먼저 연도별로 말뭉치를 분리하고 각각의 말뭉치에서 관찰하고자 하는 주제와 직접적 관계를 가지는 주제들과 그 빈도를 추출한다. 시간 추이에 따라 새롭게 나타나는 주제를 파악함으로써 관찰 대상 주제의 적용 영역 변화, 적용 방법 변화, 주제간 융·복합화 (예. 'Protein Structure'-'Metabolic Pathway'-'Information Extraction') 등을 확인할 수 있다. 연관 주제 제시는 특정 주제와 직접 또는 간접적으로 관련된 주제를 이용하여 주제간 관련도를 수치화시키는 방법이다. 연관 주제는 통합 검색 시 질의어로 입력된 주제와 연관된 주제를 보여주거나 질의어 확장에 이용할 수 있는데, 연관 주제에 가중치를 두면 가시화 범위나

확장 범위를 조절할 수 있다.

2006년도 OntoFrame은 국내 학술대회 논문을 서비스 대상으로 하였으며, 논문에서 추출된 주제들을 활용하여 논문을 분류하였다. 논문의 주제는 추론 엔진을 통해 해당 저자에게 전파 (Propagation)되고, 저자에게 모인 주제는 재계산되어 저작 당시의 기관에 전파되는 방식으로 서비스 전반에 걸쳐 적용되었다. 특히, 주제를 시소러스 개념어로 통제함으로써 불용어와 분야 부적합어들을 배제할 수 있어 정교한 서비스 구현이 가능하였다. 2007년도 OntoFrame은 논문, 인력, 주제, 출처 등 개체 별 서비스를 강화하는 방향으로 구현되었다. 특히, 주제 페이지를 별도로 구성하여, 해당 주제에 대한 중첩 주제, 연관 주제, 논문, 인력, 기관, 위치, 연구자 네트워크, 학술대회, 도서 정보 등을 제시한다. 중첩 주제는 해당 주제와 조어적 관계에 있는 주제로 상위 주제, 하위 주제, 형제 주제 등이 있으며, 연도별 연관 주제 추이는 해당 주제 관련 방법론 진화, 응용 분야 확대 등의 환경적 변화를 인지하는데 도움을 줄 수 있다. 주제 추출 및 분석은 불용어 사전 구축, 주제 추이 시각화, 연관 주제 제시 등 다양한 응용에 활용될 수 있는 방법이다. 특히, 수작업으로 처리하기 어려운 대용량 말뭉치를 대상으로 하는 경우 자동적인 말뭉치 처리 기법을 사용할 수밖에 없다.

향후 연구는 다음 두 가지 작업을 포함할 것이다. 첫째는 불용어 처리를 통해 주제로서 적합한 용어를 선별하는 것이며, 둘째는 다양한 응용 분야에의 적용을 통해 본 연구에서 제시한 주제 분석 방안의 효용성을 보이는 것이다.