http://dx.doi.org/10.3745/KIPSTD.2011.18D.6.493

용어 활용주기 모델링을 이용한 기술용어 트렌드 분석

황 미 녕[†] · 조 민 희[†] · 황 명 권[†] · 정 도 헌[††]

요 약

기술용어 트렌드는 특정 연구 분야의 세부적인 주제가 시간의 흐름에 따라 변화하는 양상을 표현한다. 그런데 학술 문헌이나 특허의 경우에는 그 데이터가 방대하여 인적 자원을 활용하여 트렌드를 분석하는 것이 용이하지 않다. 본 논문은 용어의 활용주기를 모델링하고, 이를 통해 학술 논문에 나타나는 기술용어 트렌드를 탐지하고 분석할 수 있는 방법을 제안한다. 제안된 기법은 다음과 같은 과정으로 구성된다. 먼저 논문 데이터에서 추출된 기술용어를 대상으로 일정 주기별 용어지배값을 측정한다. 용어지배값 획득되면 이를 기반으로 용어 활용주기를 모델링한다. 이 모델링 과정에서 활용주기의 시계열 패턴이 유사한 기술용어들은 동일 트렌드 범주로 분류한다. 본 논문의 기술용어 트렌드 분석 실험을 위해 한국과학기술정보연구원이 운영 중인 국가과학기술정보센터(NDSL) 학술 논문 데이터를 활용하였다.

키워드 : 기술용어, 용어 활용주기, 트렌드 탐지, 트렌드 분석

Trend Analysis of Technical Terms Using Term Life Cycle Modeling

Mi-Nyeong Hwang[†] · Min-Hee Cho[†] · Myung-Gwon Hwang[†] · Do-Heon Jeong[††]

ABSTRACT

The trends of technical terms express the changes of particular subjects in a specific research field over time. However, the amount of academic literature and patent data is too large to be analyzed by human resources. In this paper, we propose a method that can detect and analyze the trends of terms by modeling the life cycle of the terms. The proposed method is composed of the following steps. First, the technical terms are extracted from academic literature data, and the TDVs(Term Dominance Values) of terms are computed on a periodic basis. Based on the TDVs, the life cycles of terms are modeled, and technical terms with similar temporal patterns of the life cycles are classified into the same trends class. The experiments shown in this paper is performed by exploiting the NDSL academic literature data maintained by KISTI.

Keywords : Technical Term, Term Life Cycle, Trend Detection, Trend Analysis

1. 서 론

시맨틱 웹 기술은 처음 정의된 1999년부터 가트너 그룹이 기대 절정기에 진입했다고 진단한 2010년까지 지속적인 관심 속에 발전을 거듭했다[1]. 현재는 온톨로지(ontology)를 이용하여 정보나 서비스에 대한 메타데이터를 표현하거나, 지능형 에이전트가 과학기술 분야의 유망 신기술을 자동으로 탐지할 수도 있으며[1], 이종 플랫폼의 정보 원천에서 데이터 의미를 추출하고 추론하여 사용자에게 제공하는 단계[2]에까지 이르게 되었다[2,3].

시맨틱 웹 기술의 성장에 있어 중요한 축을 담당하고 있는 것이 언어 처리 기술이다. 분석의 대상이 되는 주요 기술 개체를 효과적으로 식별하여 추출하고, 이를 해석하는 기술이 지능형 정보 분석의 핵심 요소 기술이 되고 있다. 이 가운데 최근 주목받는 분야가 트렌드 분석(trend analysis)과 토픽 탐지 추적 기술(TDT: topic detection and tracking)이다.

트렌드(trend)는 방향성을 가진 경향, 동향, 추세 등을 의미한다. 이러한 트렌드는 문서 등에 일정한 양상을 띠면서 출현하는 사물이나 개념을 통해 발견된다. 웹 문서나 특정 말뭉치를 이용하여 트렌드를 탐지하는 다양한 연구가 있었다[4-9]. 이러한 연구에서 트렌드 분석 대상은 문헌에 명사구로 나타나는 일반 개념이나 토픽이다. 과거 연구들의 한계는 트렌드 분석을 위해 단어의 출현 빈도 중심으로 중요

† 정 회 원 : 한국과학기술정보연구원 연구원
†† 정 회 원 : 한국과학기술정보연구원 선임연구원(교신저자)
논문접수 : 2011년 8월 1일
수 정 일 : 1차 2011년 9월 20일
심사완료 : 2011년 9월 26일

1) FUSE(Foresight and Understanding from Scientific Exposition)
2) CUBIST(Combining and Uniting Business Intelligence with Semantic Technologies)

도를 측정하고, 이러한 중요도가 시간에 따라 변화하는 동향을 보여주는 정도에 머물러 있다는 것이다. 또한 지금까지의 주요 연구들은 새롭게 시장에 진입하는 신출 트렌드(emerging trends)의 탐지에 집중하였다[9].

일반적으로 특정 분야의 과학기술 문서를 분석하면 시간 변화에 따라 세부 주제의 트렌드도 변화함을 알 수 있다[8]. 과학 분야의 논문의 경우 발행 연도를 기준으로 해당 분야의 트렌드의 예측이 가능하고, 토픽 탐지 및 추적 연구를 통해 특정 사건의 발생 시점부터 시간 경과에 따른 사건의 전개를 추적할 수도 있다. 따라서, 특정 분야의 문서 집합에 나타나는 기술 개체를 추출 분석함으로써 유의미한 과거 기술 트렌드를 탐지할 수 있고, 미래의 트렌드를 예측할 수 있다.

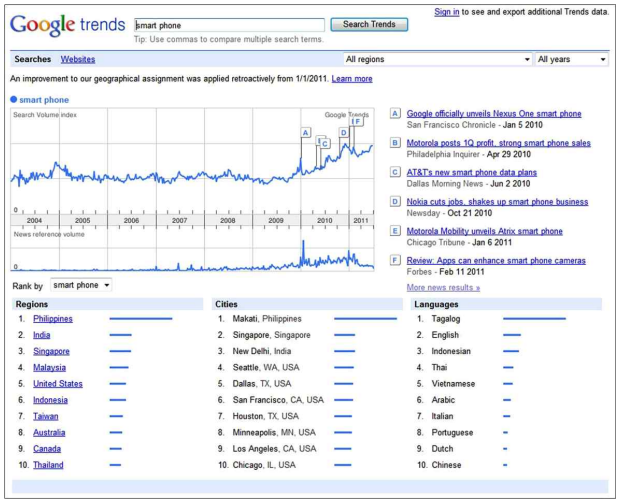
본 논문에서는 연구 트렌드 탐지를 위해 학술 논문에서 사용된 기술용어의 트렌드를 분석하는 방법을 제안한다. 이를 위해 논문 데이터에서 기술용어를 추출하고, 이 기술용어에 대해 일정 주기별로 용어지배값을 측정한 후, 같은 시계열 패턴을 보이는 기술 용어의 군집끼리는 동일한 트렌드를 갖는 것으로 간주한다. 일반적인 분석에서는 트렌드를 신출 트렌드와 사양 트렌드로 이분하지만, 본 연구에서는 성장형, 지속성장형, 성장후둔화형, 쇠퇴형, 소멸형, 유지형, 재생형 트렌드로 세분하여 그 트렌드를 분석하였다.

본 논문의 구성은 다음과 같다. 2장에서는 대용량 문서 집합에서 트렌드를 탐색하는 관련 연구들을 고찰한 후, 기존 연구의 문제점에 대해 기술한다. 3장에서는 기술용어 트렌드 탐지를 위한 용어 활용주기 모델링 과정에 대해 기술한다. 4장에서는 이 모델링 방법론을 실제 논문 데이터에 적용한 결과에 대해 해석한 후 5장에서 결론을 맺는다.

2. 관련 연구

트렌드 분석 기술 중에서 관심이 집중되는 분야는 신출 트렌드 탐지(emerging trend detection) 연구이다. 이는 단어 출현 빈도수를 기반으로 여러 트렌드 중에서 신출 트렌드를 탐지하는 것이 주목적이다[9]. 페이턴트마이너(PatentMiner)는 형태 정의 언어(shape definition language)를 도입하여 사용자가 정의한 형태의 트렌드를 특허 데이터로부터 찾아낸다[22]. 그러나 이 방식은 사용자에게 높은 수준의 전문지식을 요구한다는 단점을 가진다. HDDI는 특허 문서로부터 일반적인 명사구를 트렌드로 간주하고 이들의 출현 빈도를 자질로 하는 신경망 학습을 적용하여 각 트렌드가 신출 트렌드인지 아닌지를 판단하였다[17]. 탐지된 트렌드들 중에서 특징적인 트렌드를 찾아서 사용자에게 제공하기 위해 트렌드 순위 결정 함수를 정의한 연구도 있었다[21].

상용 트렌드 분석 서비스도 존재한다. (그림 1)에서 보듯이 구글 트렌드(Google trends)는 사용자가 입력한 검색 질의어의 입력 횟수 변화를 시각화하여 보여주는 서비스를 제공한다[4]. 구글 토픽(Google topic)은 사용자들이 가장 많이 입력한 검색어뿐만 아니라 최근의 국내외 뉴스, 블로그 등에 많이 등장한 주제를 분석하여 제시하는 서비스도 제공한다[5]. (그림 2)의 블로그펄스(BlogPulse)는 웹 블로그로부터 주요 명사구를 추출하고, 각 명사구마다 지난 2주간의 평균 추출 빈도와 당일 추출 빈도의 비율을 출현두각도(burstiness)로 정의함으로써 트렌드를 탐지하였다[6]. 트렌즈맵(Trendsmap)은 전세계 트위터의 지역별 동향을 실시간 제공하고 있다[7].



(그림 1) 구글 트렌드(Google Trends) 서비스 화면

(그림 2) 블로그펄스(BlogPulse) 서비스 화면

학계에서도 연구자들이 관심을 가지는 분야를 파악하기 위해서 용어의 사용 정도를 분석하는 연구가 진행되어 왔다. 출현 빈도를 기반으로 용어의 사용 정도를 등고선 지도 형태로 표현하고 높낮이와 인접도에 따라 용어의 중요도를 시각화한 연구도 있다[11]. 하지만 이러한 방식은 시간 정보를 고려하지 않아 트렌드를 파악하는 데에는 한계를 가진다. 등고선 지도를 활용하여 국내외 생명 연구의 트렌드를 분석하여 미래 유망 연구 테마를 도출한 연구도 있다[12]. 이 연구는 학술 논문의 제목과 초록을 대상으로 임계값 이

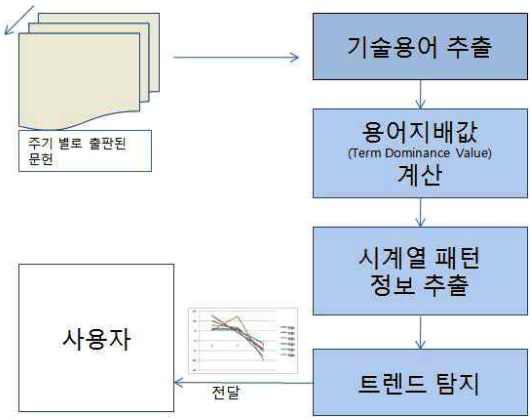
상의 고빈도 키워드를 추출하고 연도별 중요도 등고선 비교를 통해 변화되는 부분을 관심 연구 분야로 선정하였다. 또 다른 연구에서는 시계열 정보가 있는 문서 집합에서의 용어의 TF-IDF(Term Frequency-Inverse Document Frequency) 값을 시간대별로 구하여 용어의 트렌드를 파악하기도 하였으나, 신조어, 사어, 인기어의 3개 영역으로만 분류하여 다양한 트렌드 정보를 획득하기가 쉽지 않다[13,14]. 일정 기간 단위로 용어의 생명주기를 측정하기 위해 출현 빈도와 기간 가중치를 적용한 용어지배값 연구도 있었다[15,16]. 이 연구는 구간별 용어지배값 변화에 따라 생명주기 상의 생성, 성장, 쇠퇴, 소멸 등의 의미를 부여하였으나 이전 연구들처럼 신조어와 사어를 구분하는 트렌드 해석에 머물렀다.

본 연구는 이 용어지배값을 이용하여 기술용어의 활용주기를 모델링하고, 동일한 시계열 패턴을 보이는 기술용어 군집을 추출하여 트렌드 분석을 하는 방법을 제안한 뒤, 향후 서비스 방안을 소개하고자 한다.

3. 용어 활용주기 모델링

이 장에서는 기술용어의 트렌드 탐지를 위해 용어의 활용주기를 모델링하는 과정을 기술한다. 이 과정의 전체적인 프레임워크는 (그림 3)과 같다. 우선 시간 정보가 있는 대용량 문서 집합에서 기술용어를 추출하여 선정하고 용어지배값을 계산한다. 이후 시간 경과에 따라 용어지배값이 변화하는 시계열 패턴을 추출하고, 동일한 패턴을 가진 용어들을 군집화한다. 최종 트렌드 분석은 군집단위로 이뤄진다. 이 과정을 단계별로 정리하면 다음과 같다.

1. 기술용어 추출 - 대용량 문서 데이터 집합에서 기술용어 추출 및 용어 선정
2. 용어지배값 계산 - 시간 구간별로 각 기술 용어의 용어지배값 계산
3. 군집화 - 용어지배값의 시계열 패턴 기반 군집화
4. 트렌드 분석 - 군집별 트렌드 의미 부여



(그림 3) 용어 활용주기 모델링 과정

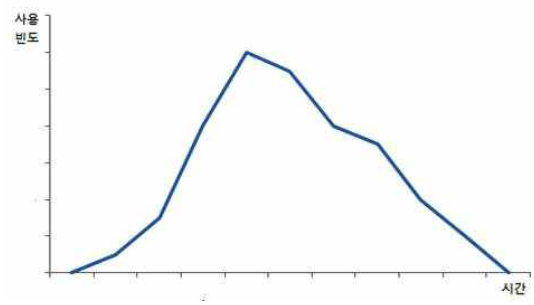
활용주기 모델링 과정의 1단계는 대용량 문서 데이터 집합에서 기술용어를 추출한다. 이를 위해 텍스트 마이닝 시스템인 SINDI(Scientific INtelligent DIscovery)를 활용하여 원문 내의 비정형 기술용어를 추출하였다[18]. 추출된 기술용어에 대한 시간대별 지배 정도를 측정하기 위해서는 용어지배값을 이용하였다. 이 용어지배값의 개념은 이전 연구[15,16]를 통해 잘 정립되어 있으므로, 본 논문에서는 수식을 이용하여 실제 계산과 활용을 위한 기술적 설명만을 할 것이다.

3.1 용어지배값

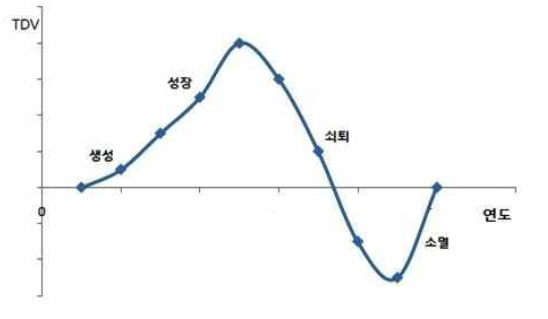
용어지배값(TDV: Term Dominance Value)은 특정 용어가 특정한 기간 중에 생성, 성장, 쇠퇴하는 정도를 정량적으로 측정하기 위해 해당 기간 내에 관찰되는 패턴을 수치화한 값이다. 이 값을 통해 특정 용어의 기간별 지배 정도를 파악할 수 있다. $TDV_{(t)\text{시작}-\text{종료}}$는 용어 t에 대해 시작부터 종료기간까지 관찰하여 얻은 용어지배값이다. 이를 계산하는 데에 사용되는 NTF_i는 해당 용어가 특정 시점 i에서 발생하는 정도를 나타내는 정규화된 용어 빈도값이며, $ANTF$는 용어 t가 시작시점부터 종료시점 내의 매 시점마다 가지는 정규화된 용어 빈도값의 평균값이다. 또한 PW_i는 특정시점 i에 대한 기간 가중치로 최근 시점이 더 큰 가중치를 가지도록 설정할 수 있다. PF_i는 용어 t의 출현 시점 빈도이다.

$$TDV_{(t)\text{시작}-\text{종료}} = \frac{\sum_{i=\text{시작}}^{\text{종료}} \left((NTF_i - ANTF) \times PW_i^2 \right)}{PF_i} \quad (1)$$

일관성 있는 관찰을 위해서 시작시점과 종료시점 사이의 기간은 고정한다. 관찰 기간이 짧을수록 용어의 생명주기 단계는 용어의 출현 빈도 변화에 민감하게 반응하고, 급격하게 움직인다. 관찰 기간이 긴 경우에는 용어의 생명주기 단계가 완만한 흐름을 보인다[16]. 일반적인 용어의 시간대별 사용 빈도가 (그림 4)와 같이 나타난다면 (그림 5)는 용어의 생명주기에 따른 용어지배값의 전체 추이를 보여준다. 용어가 생성되는 시점에서는 0, 성장 단계는 값의 증가, 쇠퇴 과정에서는 값의 감소, 그리고 소멸 되는 시점에서는 다시 0이 된다.



(그림 4) 용어의 시간대 별 사용빈도(TF)



(그림 5) 용어의 생명주기에 따른 용어지배값(TDV) 추이

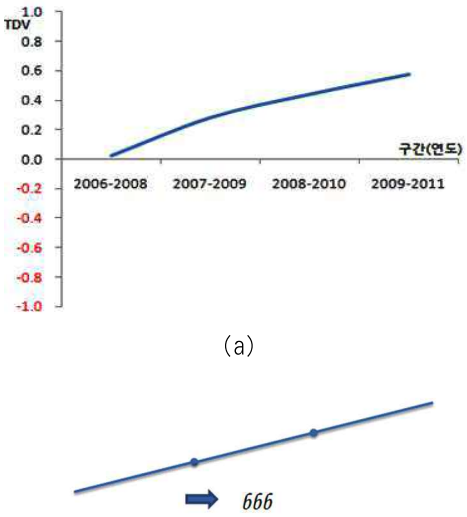
용어지배값의 추이를 관찰함에 따라 특정 용어가 현재 생명주기 상의 어떤 단계에 있는지 알 수 있다. 이전 연구 [15,16]에서는 'Term Life cycle'을 용어의 '생명주기'로 정의하였으나, 본 논문에서는 용어가 자생하는데 중점을 두지 않고, 용어가 특정 분야에서 활용되는 정도에 초점을 맞추기에 이를 용어의 '활용주기'로 재 정의하기로 한다. 다음 절에서는 기술용어의 트렌드 변화를 탐지하기 위하여 용어의 활용주기 그래프를 시계열 패턴으로 변환하는 모델링 단계에 대해 설명한다.

3.2 기술용어 활용주기 모델링

본 논문의 기법은 어떤 기술용어가 활용주기 모델에서 어떠한 위치에 있는지를 시간의 흐름에 따라 관찰함으로써 해당 용어의 트렌드를 분석할 수 있다는 가정을 전제로 한다. 이러한 분석을 위해서는 우선 기술 용어의 지배값을 어떠한 시간 간격으로 측정할 것이며, 측정되는 시점에 고려해야할 시간 범위를 설정해야 한다. 이 시간 간격을 *관찰주기*, 시간의 범위를 *관찰구간*으로 정의한다. 본 논문에서는 1년 간격으로 용어 지배값을 계산하였으며, 매번 용어지배값을 계산할 때 고려하는 관찰구간은 3년으로 설정하였다.

본 논문의 기법은 용어의 트렌드를 분석하기 위한 시계열 패턴을 추출한다. 이 시계열 패턴은 용어의 활용주기 상 위치를 표현하는 용어 지배값을 통해 얻었다. 시간축을 따라 용어의 지배값을 측정된 순서에 따라 배치하면, 이 시계열 패턴이 변화하는 양상을 인접한 두 값을 연결하는 간선(edge)의 위상으로 파악할 수 있다. 간선이 상승, 하강, 또는 유지되는 양상이나 용어지배값의 부호 등을 이용하여 다양한 간선 위상을 정의할 수 있고, 이러한 위상에 특정한 기호를 부여할 수 있다. 본 논문의 트렌드 분석 기법은 이러한 간선의 위상을 표현하는 기호의 연속열(sequence)를 분석패턴으로 이용하였다.

그래프 간선의 위상에 대응하는 기호는 (그림 6)과 같이 정의한다. 간선이 상승하는 조건에 따라 {*a, b, c, d, e*}로 정의한다. 이전 구간의 누적 빈도($ACCTF$)가 0이고 현재 구간의 용어지배값이 0보다 큰 경우는 용어가 생성되는 구간으로 *a*로 정의하였다. 간선이 상승하는 것은 동일하지만, 이미 생성되어 있던 용어가 더욱 활발히 사용되며 출현빈도가 증가하는 경우(즉, 현재 구간의 용어지배값이 이전 구간의

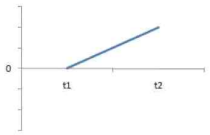
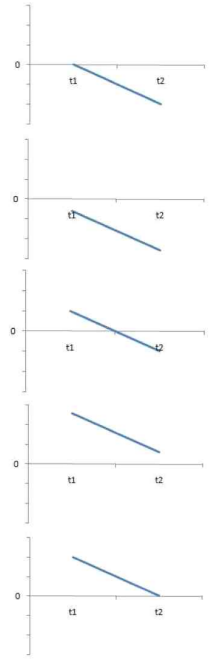
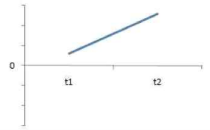
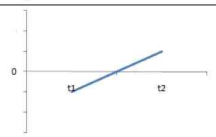
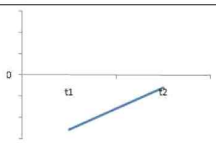
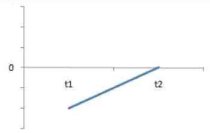
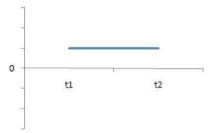
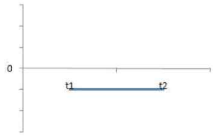
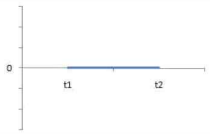


(그림 6) 'clustering algorithm'의 용어지배값을 이용하는 활용주기 그래프(a)와 패턴 비교를 위한 기호화(b)

용어지배값보다 크고, 현재 구간의 용어지배값과 이전 구간의 용어지배값의 차이가 ε보다 큰 경우)를 *b*로 정의한다. *c*는 용어가 재생되는 구간으로 용어가 쇠퇴, 소멸하지 않고 다시 사용이 증가하는 이전 구간의 용어지배값은 0보다 작고, 현재 구간의 용어지배값은 0보다 큰 경우이다. *d, e*는 용어가 소멸해가는 구간으로 현재 구간의 용어지배값과 이전 구간의 용어지배값의 차의 절대값이 ε보다 큰 경우인데, 현재구간의 용어지배값의 위치에 따라 *d*와 *e*로 구분하여 정의한다. *i*는 용어가 쇠퇴하는 구간으로 현재 구간의 용어지배값이 이전 구간의 용어지배값보다 작고, 두 값이 가진 차의 절대치가 ε보다 큰 경우로 <표 1>에서 보듯이 5가지 위상을 포함한다. 마지막으로 출현빈도가 유지되는 구간은 $|TDV_{t2} - TDV_{t1}| \leq \epsilon$인 구간을 의미하는데 간선이 양, 음, 혹은 0의 위치 가운데 어디에 있는가에 따라 *o, p, q*로 세분하였다. 본 논문에서는 임계치 ε를 전체 그래프가 나타나는 구간 평균값의 10%로 설정하였다.

예를 들어, <표 2>에서 보듯이 'clustering algorithm'이라는 기술용어는 2006년에서 2011년의 전체 기간에서의 누적 빈도가 641회이고, 관찰구간을 3년으로 하여 용어지배값을 측정하면 $TDV_{(2006-2008)}$=-0.0209, $TDV_{(2007-2009)}$=0.2819, $TDV_{(2008-2010)}$=0.4408 , $TDV_{(2009-2011)}$=0.5764로 나타난다. 이렇게 일정한 구간별로 계산된 용어지배값을 (그림 6)의 (a)와 같이 해당 용어의 활용주기 그래프로 시각화할 수 있다. 이 그래프의 구간별 간선의 위상을 <표 1>에 정의된 기호로 치환하면 그래프가 패턴 '*bbb*'로 변환된다. 이러한 방식으로 각 기술용어의 활용주기 그래프를 패턴으로 변환할 수 있으며, 이 시계열 패턴의 군집화 과정을 거쳐 기술용어 군집들의 트렌드 분석이 가능해진다. 다음 절에서는 실제 사용되고 있는 학술 논문 데이터를 대상으로 용어 활용주기 모델링 과정을 거쳐 동일 트렌드 범주의 기술용어를 분류하기로 한다.

〈표 1〉 용어의 활용주기 그래프의 간선의 위상 별 기호화

기호	TDV 변화 형태	비고	기호	TDV 변화 형태	비고
a		$ACCTF_{t1}=0$ and $TDV_{t2}=0$	i		$(TDV_{t2}-TDV_{t1})<0$ and $\lvert TDV_{t2}-TDV_{t1}\rvert>\varepsilon$
b		$(TDV_{t2}-TDV_{t1})>0$ and $\lvert TDV_{t2}-TDV_{t1}\rvert>\varepsilon$			
c		$TDV_{t1}<0$ and $TDV_{t2}>0$			
d		$\lvert TDV_{t2}-TDV_{t1}\rvert>\varepsilon$ and $TDV_{t1}<0$ and $TDV_{t2}<0$			
e		$TDV_{t1}<0$ and $ACCTF_{t2}=0$			
o		$\lvert TDV_{t2}-TDV_{t1}\rvert\le\varepsilon$ and $TDV_{t1}>0$ and $TDV_{t2}>0$	p		$\lvert TDV_{t2}-TDV_{t1}\rvert\le\varepsilon$ and $TDV_{t1}<0$ and $TDV_{t2}<0$
q		$TDV_{t1}=0$ and $TDV_{t2}=0$			

〈표 2〉 NDSL 논문 데이터에서의 'clustering algorithm' 기술용어에 대한 연도별 TF, NTF, TDV 값

연도	*TF*	*NTF*	*TDV*			
2006	38	19	-0.0209			
2007	70	33		0.2819		
2008	47	22			0.4408	
2009	199	51				0.5764
2010	195	57				
2011	92	92				

4. 기술용어 트렌드 분석

과학기술 논문에 나타나는 기술용어의 트렌드를 탐지하기 위해서 국가과학기술정보센터(NDSL: National Discovery for Science Leaders, http://www.ndsl.kr)가 보유하고 있는 논문 데이터를 사용하였다. NDSL은 학술 문헌들의 서지정보와 원문을 제공하는 정보 검색 서비스로 검색 편의성이 높고, 논문, 특허, 동향 보고서 등의 다양한 문헌 정보를 제공하지만, 문헌에 대한 텍스트 마이닝이나 연구 트렌드 정보를 제공하는 기능은 미비한 실정이다.

이번 실험은 NDSL 해외 논문 가운데 IT 분야로 대상을 한정하여 기술용어를 추출하였다. NDSL의 문서들은 국과과학기술표준분류체계를 기본으로 하여 분류되어 있고, 이들 중에서 대분류 코드가 J(정보), I(전기,전자), K(통신)인 문서들만을 대상으로 하였다. 이 기술용어 추출은 텍스트 마이닝 시스템인 SINDI-CORE를 통해서 각 논문 초록에서 기술용어를 추출하고, 이를 정규화하는 과정을 통해 이뤄졌다[18].

4.1 실험 데이터

<표 3>[3]에서 보듯이 10만 여건의 문헌에서 추출한 기술용어의 총 개체 수는 3,728,923건 이며, 이 가운데 기술용어의 개체 수는 1,672,045이었다. 또한 전체 데이터에서 95.6%

3) NDSL 논문 데이터의 입수 일자는 2011년 3월 말 기준임

에 해당하는 기술용어가 6년 동안 5회 이하로 나타나는 저빈도 용어였다. 저빈도 기술용어에 대해 트렌드를 탐지하는 것은 적절하지 않으므로 누적 빈도수를 기준으로 상위에 있는 기술용어를 우선적으로 선별하여 분석을 진행하였다. 누적 빈도수 기준 상위 500개의 기술용어에 대해 관찰주기가 1년일 때의 용어지배값을 계산하고, 이전 절에서 언급한 활용주기 모델링 과정을 통해 자동 군집화 실험을 실시하였다. 용어지배값의 추이를 일관성 있게 관찰하기 위해서 관찰구간은 고정하며, 총 구간이 6년이기 때문에 본 논문에서는 3년을 사용하였다.

〈표 3〉 NDSL에서 추출한 용어의 연도별 분포 현황

연도	문헌(논문)의 수	추출된 용어의 수
2006	14,738	506,793
2007	15,734	548,422
2008	15,399	529,967
2009	26,601	1,010,438
2010	22,645	874,086
2011	6,483	259,217
합계	101,600	3,728,923

4.2 기술용어 트렌드 분석

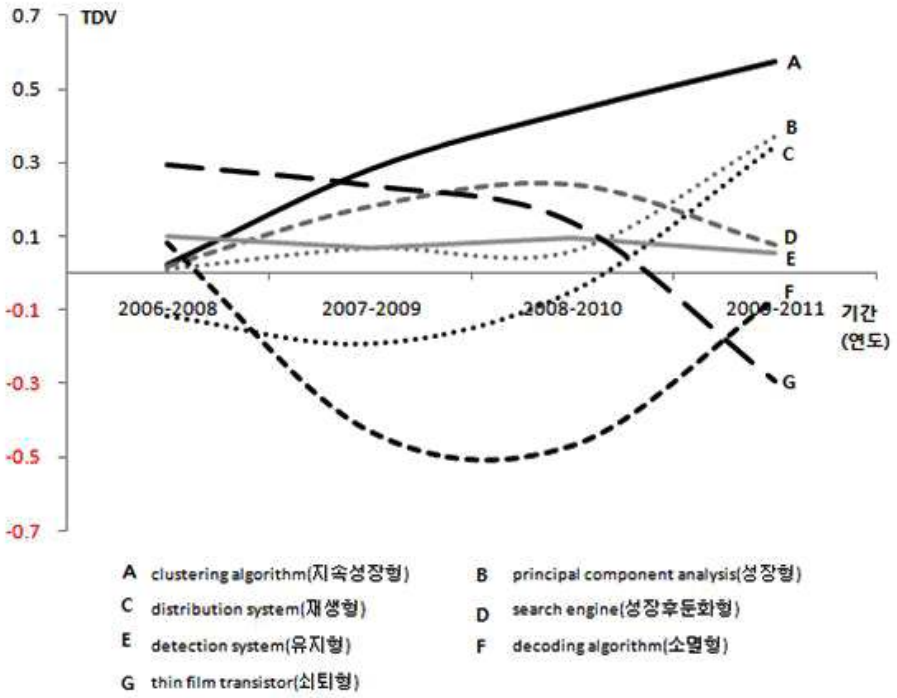
NDSL의 학술 논문 초록에서 추출된 기술 용어에 대해 관찰기간이 3년인 용어지배값을 측정하여 시계열 패턴에 따라 군집화한 결과 전체 500개의 용어를 79개의 군집으로 분류할 수 있었다. 각 군집의 패턴을 해석하여 해당 군집의 트렌드 의미를 <표 4>와 같이 성장형, 지속성장형, 성장후둔화형, 쇠퇴형, 소멸형, 유지형, 재생형으로 부여하였다.

성장형, 지속성장형, 성장후둔화형은 모두 기술용어의 사용이 증가하고 있는 모델들이지만, 성장의 지속성을 기준으로 다시 세분류한 것이다. 성장형 모델은 기술용어의 사용이 점진적으로 증가하고 있는 군으로 'recognition system', 'web service' 등이 포함되어 있다. 지속성장형 모델은 성장률이 감소하지 않고 지속적인 성장을 보이는 군집으로 'clustering algorithm' 등의 기술용어가 대표적이다. 성장후둔화형 모델은 'thin film' 등이 속한 군으로 기술용어가 신규로 등장하여 사용 추이가 증가하다가 정체 단계로 접어든 모델이다.

기술용어의 사용이 감소하고 있는 모델에 대해서는 그 추이에 따라 쇠퇴형 모델과 소멸형 모델로 정의한다. NDSL 논문 데이터의 6년 기간 동안 지속적인 사용 추이를 보이는 유지형 모델은 'search algorithm', 'detection system' 등의 용어가 포함된다. 재생형 모델은 기술용어의 사용 추이가 쇠퇴했지만 소멸하지 않고 다시 사용이 증가하는 군이다. 재생형 모델에는 'fuel cell'이 포함되어 있다. 가트너 그룹에서 발표한 Hype Cycle for Emerging Energy Technologies (2010)에 의하면 이 용어는 '거품이 꺼진(sliding into the trough)' 기술용어로 평가되는데, 이는 해당 기술에 대한 환상이 제거되고, 시장의 과도한 관심에서 벗어나 재조명받는 시기에 속한다[19]. NDSL 논문 데이터에서 추출한 2006~2011년의 트렌드 탐지에서도 기술용어의 활용이 쇠퇴하였다가 다시 재생되고 있는 기술용어로 분석되었다.

〈표 4〉 기술용어 트렌드와 대표 용어

트렌드	패턴	설명	대표 용어
성장형	bio, bob, cbo, cio, cob	성장->유지	principal component analysis, recognition system, heuristic algorithm, linear model, web service, decision tree, support vector machine, wavelet transform, image segmentation
지속 성장형	*bbb, bib, oob, obb*	성장->성장 ->성장	clustering algorithm, swarm optimization, expert system, fuzzy logic, prediction model, feature extraction
성장후둔 화형	*bbi, boo, coo, dco*	신규 등장-> 성장->정체	thin film, x ray diffraction, decision maker, performance evaluation
쇠퇴형	*bii, boi, cbi, cii, iii, coi*	성장->쇠퇴	mathematical model, network model, chaotic system, dynamical system, information retrieval, analytical solution, periodic solution, neural network model, bayesian network, neural network, artificial neural network, solar cell, information system, pattern recognition, computer science, thin film transistor
소멸형	*idd, iid, ipd*	쇠퇴->소멸	finite element method, communication system, monte carlo simulation, finite element analysis, decoding algorithm, wireless communication
유지형	*ooo*	유지	search algorithm, detection system
재생형	*idc, ipc, pdc, ppc, iic*	쇠퇴->소멸되지 않고->재생	sensor network, wireless network, distribution system, scheduling algorithm, speech recognition, access network, fuel cell, synchronous motor, real time, embedded system, wireless sensor network, mobile device, distribution network



(그림 7) NDSL 기술용어의 활용주기 모델링 결과로 탐지된 트렌드

이 6가지 모델들에 대한 도식화는 (그림 7)에서 확인할 수 있다. 그림에서 볼 수 있는 바와 같이 본 논문의 기법은 기존의 기법에 비해 매우 다양한 트렌드를 분석해 낼 수 있다. 이러한 분석 능력은 신출 트렌드나 소멸 트렌드만 감지하는 일반적인 기법에 비해 더욱 다양한 용어들을 트렌드 분석의 대상으로 삼을 수 있으며, 이에 따라 더욱 구체적이고 복합적인 연구 전략 수립 등을 가능하게 한다.

4.3 기술용어 트렌드 분석의 정성적인 평가

학술 문헌에서 추출한 기술용어를 대상으로 용어 활용주기 모델링 과정을 통해 시계열 패턴이 유사한 기술용어들을 동일한 트렌드로 분류하였다. 이전 절에서 제시한 성장형, 지속성장형, 성장후둔화형, 쇠퇴형, 소멸형, 유지형, 재생형 모델들은 해당 학술 분야의 연구가들에게 기술용어의 과거에서 현재로 이어지는 트렌드를 제시해준다. 전문 연구가들에게는 해당 기술용어의 성장, 쇠퇴의 추이를 보면서 관련 기술의 성장, 쇠퇴 원인을 파악하게 해주는 시점을 제시해 줄 수 있다. 관련 분야의 연구자들은 신규로 등장하는 기술용어들을 보면서 새로운 연구 분야에 대한 동향을 파악하거나, 쇠퇴하는 기술용어를 통해 관련 기술이 쇠퇴하게 되는 원인을 분석하게 하는 계기를 제공해주며, 소멸되지 않고 새로이 재생되는 기술용어들을 통해 재사용되는 관련 기술을 연구 분야로 선정할 수 있다. 예를 들어, 재생 모델에 속한 'distribution network', 'distribution system', 'mobile device' 등의 기술용어가 다시 부각되는 것을 보면서 이 기술들이 관련된 새로운 연구/산업 분야인 클라우드 컴퓨팅(cloud computing)에 대해 유추해 낼 수 있고, 이를 통해 새로운 아이템을 발굴하여 연구 및 사업을 진행하는데 도움을 줄 수 있다.

5. 결 론

본 논문은 학술 문헌을 통해 연구의 트렌드 탐지를 수행할 수 있는 방법을 제안하였다. 이를 위해 학술논문 데이터에서 기술용어를 추출하고, 각 용어에 대해 일정 주기별 용어지배값측정을 수행하였다. 시간 변화에 따라 변동하는 용어지배값을 이용하여 해당 용어의 활용주기 그래프를 분석이 용이한 시계열 패턴으로 치환하여 군집화하는 방법을 제안하였다. 이러한 방법을 통해 각 용어 군집의 활용주기를 파악할 수 있으며, 해당 군집에 포함된 용어들의 트렌드 의미를 부여할 수 있었다. 일반적인 트렌드 분류가 신출 트렌드와 사양 트렌드를 구분하는 것에 머무르는 데에 반해, 본 논문에서 제안하는 기법은 성장형, 지속성장형, 성장후둔화형, 쇠퇴형, 소멸형, 유지형, 재생형 트렌드로 세분화하여 트렌드 분석을 할 수 있다. 이러한 기술용어 트렌드 분석 기능은 R&D 전략 수립 과정 및 의사 결정을 지원할 수 있으며, 새로운 형태의 테크놀로지 인텔리전스 서비스인 InSciTe(Intelligence Science & Technology)를 통해 서비스할 계획이다[20].

본 논문에서는 특정한 연구 분야에 한정된 기술 용어를 대상으로 실험을 실시하였다. 이러한 연구를 바탕으로 향후 적용 학술 분야를 확대할 예정이며, 학술 연구 분야별로 트

렌드 탐지 기법을 적용하는 데에 있어 고려해야할 차별성 등이 존재하는지에 대한 연구를 추가로 수행할 예정이다. 이를 통해 다양한 학술 연구 분야의 트렌드 탐지에 적합한 기술용어의 활용주기 모델링 방법을 연구할 예정이다.

참 고 문 헌

[1] "Hype Cycle for Web and User Interaction Technologies, 2010", http://www.gartner.com/DisplayDocument?id=1407814& ref='g_fromdoc'

[2] http://www.iarpa.gov/solicitations_fuse.html

[3] http://www.v3.co.uk/v3/news/2268590/sheffield-researchers-tap

[4] Google Trends: http://www.google.com/trends.

[5] Google 토픽: http://www.google.co.kr/topicsearch.

[6] N. S. Glance, M. Hurst and T. Tomokiyo, "BlogPulse: Automated Trend Discovery for Weblogs", WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004.

[7] Trendsmap: http://trendsmap.com.

[8] Q. Mei and C. Zhai, "Discovering Evolutionary Theme Patterns from Text - An Exploration of Temporal Text Mining", Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, August 21-24, 2005, Chicago, Illinois, USA.

[9] A. Kontostathis, L. M. Galitsky, W. M. Pottenger, S. Roy and D. J. Phelps, "A Survey of Emerging Trend Detection in Textual Data Mining", In Survey of Text Mining: Clustering, Classification, and Retrieval, 2003.

[10] S. Morinaga and k. Yamanishi, "Tracking Dynamics of Topic Trends Using a Finite Mixture Model", Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, August 22-25, 2004, Seattle, WA, USA.

[11] A. Skupin, "The world of geography: Visualizing a knowledge domain with cartographic means", Proceedings of the National Academy of Sciences of the United States of America, Vol.101, No.Suppl 1. pp.5274-5278, 2004.

[12] 생명공학정책연구센터, "바이오 연구 트렌드 분석 및 미래유망 연구 테마 도출", 2011.

[13] H. Abe and S. Tsumoto. "Trend Detection from Large Text Data", SMC, pp.310-315, 2010.

[14] H. Abe and S. Tsumoto. "Analysis of Research Keys as Temporal Patterns of Technical Term Usages in Bibliographical Data", AMT, pp.150-157, 2010.

[15] 정한민, 구희관, 이병희, 성원경. "효율적인 자원 운영을 위한 전문용어 생명주기 관리 연구", 한국컴퓨터종합학술대회, Vol.32, No.1(B), pp.457-459, 2005.

[16] 정한민, 성원경. "과학기술 용어에 대한 용어 생명주기 고찰", 한국콘텐츠학회, Vol.4, No.2, pp.84-89, 2006.

[17] L. E. holzman, T. A. Fisher, L. M. Galitsky, A. Kontostathis, W. M. Pottenger, "A Software Infrastructure for Research in Textual Data Mining", pp.112, ITCTAI'03, 2003.

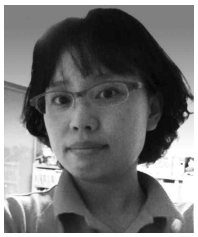
[18] 최윤수, 정창후, 조현양. "과학기술 핵심개체 인식기술 통합에 관한 연구", 정보관리학회지, 제28권 제1호, pp.89-104, 2011.

[19] "Hype Cycle for Emerging Energy Technologies, 2010 ", http://www.gartner.com/DisplayDocument?doc_cd=205231&ref=g_noreg

[20] 이미경, 정한민, 김평, 성원경, "연구개발 전략 수집 지원을 위한 테크놀로지 인텔리전스 서비스", 정보과학회지, 제 17권 제 5호, 2011.

[21] 오홍선, 최윤정, 신욱현, 정윤재, 맹성현, "자동 트렌드 탐지를 위한 속성의 정의 및 트렌드 순위 결정 방법", 정보과학회논문지:소프트웨어 및 응용, 제 36권 제 3호, 2009.

[22] B. Lent, R. Agrawal and R. Srikant, "Discovering Trends in Text Databases", KDD-97, 1997.



황 미 녕

e-mail : mnhwang@kisti.re.kr
2000년 부산대학교 전자계산학과(학사)
2002년 부산대학교 전자계산학과(이학석사)
2002년~현 재 한국과학기술정보연구원 연구원
관심분야 : 데이터 마이닝, 텍스트 분석, 계산 이론

조 민 희

e-mail : mini@kisti.re.kr
2003년 연세대학교 전산학과(학사)
2005년 연세대학교 전산학과(석사)
2005년 한국과학기술정보연구원 연구원
관심분야 : 자연어처리, 정보검색, 텍스트마이닝

황 명 권

e-mail : mgh@kisti.re.kr, mg.hwang@gmail.com
2004년 조선대학교 컴퓨터공학부(학사)
2006년 조선대학교 전자계산학과(이학석사)
2011년 조선대학교 컴퓨터공학과(공학박사)
2011년~현 재 한국과학기술정보연구원 연구원
관심분야 : 데이터마이닝, 시맨틱웹, 의미기반처리

정 도 헌

e-mail : heon@kisti.re.kr
2011년 연세대학교 문헌정보학과(박사수료)
2003년~현 재 한국과학기술정보연구원 선임연구원
관심분야 : 텍스트 마이닝, 시맨틱 웹, 정보검색