자동 트렌드 탐지를 위한 속성의 정의 및 트렌드 순위 결정 방법

(Trend Properties and a Ranking Method for Automatic Trend Analysis)

오 흥 선 † 최 윤 정 † 신 욱 현 † 정 윤 재 † 맹 성 현 ††
(Heung-Seon Oh) (Yoonjung Choi) (Wookhyun Shin) (Yoonjae Jeong) (Sung-Hyon Myaeng)

요 약 특허, 뉴스, 블로그와 같이 시간 정보가 있는 문서들로부터의 자동적인 트렌드 분석(trend analysis)은 토픽탐지 및 추적 기술(TDT: Topic Detection and Tracking)과 더불어 중요한 연구 분야로 대두되고 있다. 과거 연구들은 대부분 트렌드과 관련된 단어의 출현 빈도 정보를 이용하여 주어진 개념의 중요도를 측정하고 이 개념의 시간에 따른 트렌드 라인을 보여주는 것에 초점을 맞췄다. 신출 트렌드(emerging trend)를 탐지하기 위해서는 주어진 개념의 출현 빈도수 변화와 같은 간단한 방법이나 학습 데이타와 비교하여 차이를 탐지하여 제시하는 방법이 사용되었다. 그러나 여러 트렌드 중에서 특징적인 트렌드를 찾아서 사용자에게 제공하기 위해서는 트렌드 순위 결정 함수가 필요하다. 본 논문은 트렌드의 다양한 측면을 정량화하기 위하여 출현 빈도로 구성된 트렌드 곡선으로부터 네 가지 속성 (변동성, 지속성, 안정성, 누적량) 을 정의하고 이를 활용한 트렌드 순위 결정 방법을 제안한다. 일련의 실험을 통하여 각 속성의 유용성을 검증하고 속성들의 조합이 순위 결정에 어떤 영향을 미치는지 분석하였다. 실험결과로부터 네 가지 속성을 모두 조합할 경우 특징적인 트렌드 탐지에 더욱 기여하는 것을 알 수 있다.

키워드 : 트렌드 탐지, 트렌드 속성, 트렌드 순위 결정

Abstract With advances in topic detection and tracking(TDT), automatic trend analysis from a collection of time-stamped documents, like patents, news papers, and blog pages, is a challenging research problem. Past research in this area has mainly focused on showing a trend line over time of a given concept by measuring the strength of trend-associated term frequency information. For detection of emerging trends, either a simple criterion such as frequency change was used, or an overall comparison was made against a training data. We note that in order to show most salient trends detected among many possibilities, it is critical to devise a ranking function. To this end, we define four properties(change, persistency, stability and volume) of trend lines drawn from frequency information, to quantify various aspects of trends, and propose a method by which trend lines can be ranked. The properties are examined individually and in combination in a series of experiments for their validity using the ranking algorithm. The results show that a judicious combination of the four properties is a better indicator for salient trends than any single criterion used in the past for ranking or detecting emerging trends.

Key words : Trend Detection, Properties of Trend Lines, Trend Ranking

· 연구는 Microsoft Research Asia 지원과 2008년 2단계 BK 21 사업의 부분 지원으로 수행되었습니다.

† 비 회 원 : 한국과학기술원 정보통신공학과
ohs@kaist.ac.kr
choiyj35@kaist.ac.kr
wshin@kaist.ac.kr
hybris@kaist.ac.kr
†† 종신회원 : 한국과학기술원 정보통신공학과 교수
myaeng@kaist.ac.kr

논문접수 : 2008년 10월 23일
심사완료 : 2009년 2월 6일

정보과학회논문지: 소프트웨어 및 응용 제36권 제3호(2009.3)

1. 서 론

트렌드(trend)는 일반적인 방향성이나 경향성을 가지고 있는 현상을 지칭한다[1]. 따라서 일정한 경향을 가지고 문서에 출현하는 사물(entity)이나 개념(concept)들을 통해 트렌드를 발견할 수 있다. 예를 들어 휴대폰의 사용자인터페이스 관련 용어가 웹 문서에 출현하는 정도의 변화나 웹 포탈 서비스의 인기 검색어 목록으로부터 트렌드를 관찰할 수 있다.

트렌드는 여러 면에서 유용하게 사용될 수 있다. 첫째, 우선 사람들은 사물이나 개념에 대해서 전체 정보나 사전 지식 없이 어떤 결정을 내리려고 할 때 트렌드를 알고 싶어 한다. 휴대폰 영역을 예로 들면, 청소년 구매자들이 터치 스크린이 탑재된 휴대폰을 사려고 할 때, 많은 경우 어떤 모델이 일반적인 인기를 얻고 있는지 파악하고 가장 대중적인 모델을 구입하려고 한다. 둘째로, 자신들의 지적 호기심을 충족을 위하여 트렌드를 알고 싶어 한다. 다시 말하면 사람들은 그들의 관심 분야에 어떤 변화가 있는지 앎으로써 지적 호기심을 충족시키려고 한다.

웹에 있는 정보와 사용자들이 증가할수록, 사람들은 웹을 이용하여 자신들이 관심 있어 하는 다양한 사물의 트렌드를 찾으려고 한다. 하지만 웹 문서를 검색하여 읽고 그 결과를 분석하여 트렌드가 무엇인지 파악하는 것은 많은 시간을 요구한다. 이런 면에서 시간 정보가 부여된 특허, 뉴스, 블로그 문서들로부터의 자동적으로 트렌드를 찾아 내는 기술은 토픽을 자동 탐지하고 추적하는 TDT(Topic Detection & Tracking) 기술과 더불어 최근에 부각되는 연구 분야이다.

웹 문서나 주어진 말뭉치로부터의 트렌드 탐지는 여러 관련 연구들에서 찾아 볼 수 있다[2-8]. 이들 연구에서 트렌드의 대상은 텍스트에 있는 명사구로 표현된 일반적인 개념이나 토픽(topic)이다. 과거 연구들은 대부분 트렌드와 관련된 단어의 출현 빈도 정보를 이용하여 해당 개념의 중요도를 측정하고 그 개념의 시간에 따른 트렌드 곡선을 보여주는 것에 초점을 맞췄다. 신출 트렌드(emerging trend)를 탐지하기 위해서는 주어진 개념의 출현 빈도수 변화와 같은 간단한 방법 혹은 학습 데이타와 비교하는 방법이 사용되었다. 그러나 많은 트렌드 중에서 가장 특징적인 트렌드를 찾기 위해서는 순위 결정 함수를 고안하는 것이 중요하다. 시스템 관점에서 볼 때 트렌드 간의 우선순위 결정을 하지 않고는 사용자에게 특징적인 트렌드를 우선적으로 보여 줄 수가 없다.

일정 기간 동안, 주어진 개념의 출현 빈도수가 높을 경우 그 개념은 다른 개념들보다 대중적인 인기도(popularity)를 가지고 있다고 할 수 있다. 그러나 트렌드를 발견하는데 있어 출현 빈도수의 변화 형태를 고려하지 않고 누적 출현 빈도만으로 그 순위를 결정하는 것은 문제가 있을 수 있다. 예를 들면, 일정 기간 동안에 동일한 누적 출현 빈도를 가지는 두 개 트렌드 곡선이 있을 경우 상대적인 순위를 결정하기 위해서는 두 곡선의 전반적인 기울기도 고려하는 것이 트렌드의 중요도를 판단하는데 중요한 역할을 할 것이다.

본 논문에서는 트렌드의 다양한 측면을 정량화하기 위한 4 가지 트렌드 속성 및 이를 이용한 새로운 트렌드 순위 결정 방법을 제안한다. 제안하는 트렌드 순위 결정 방법은 기존의 출현 빈도 정보를 활용하여 4 가지 속성값을 결정하고 이를 조합함으로써 하나의 특징적인 값을 계산하고 비교하는 것으로 이루어져 있다.

트렌드의 형태에는 여러 가지가 있을 수 있지만 본 논문에서는 신출 트렌드(emerging trend)와 사양 트렌드(submerging)로 그 형태를 분류하였다. 제안하는 트렌드 속성 및 순위 결정 방법은 트렌드의 형태에 종속적이지 않으므로 동일한 방법으로 신출 혹은 사양 트렌드 순위 결정 방법에 적용할 수 있다. 본 논문에서는 관련 연구에서와 같이 신출 트렌드 순위 결정에 초점을 두었다.

본 논문은 다음과 같이 구성 되었다. 2장에서는 기존의 트렌드 분석 관련 연구를 조사 하였다. 3장에서는 기존 출현 빈도 수를 이용한 트렌드 순위 결정 방법에 대한 문제 제기를 설명하고 4장과 5장에서 4 가지 트렌드 속성의 정의 및 정량화와 이를 활용한 트렌드 순위 결정 방법을 설명한다. 6장에서는 실험 및 결과에 대하여 설명하고 7장에서는 결론을 맺는다.

2. 관련 연구

트렌드 분석은 여러 연구에서 찾을 수 있다. 예를 들어 주식 시장 분석에서, 트렌드는 시간에 따른 주식 가격의 변화를 나타낸다. [5,9,10]에서는 각기 다른 방법을 이용하여 주식 가격이 상승할지 혹은 하강할지를 예측하는 시스템을 제안하였다.

토픽탐지 및 추적(topic detection and tracking, TDT) 분야는 탐지된 토픽을 트렌드로 확장하려고 노력하였다. [3,7,8]에서는 신경망과 유한 혼합 모델(finite mixture model) 등을 이용하여 토픽을 탐지 및 추적하고 탐지된 토픽의 트렌드는 시간에 따른 출현 빈도수나 확률의 변화로써 보여주었다. 하지만 TDT에서는 탐지된 토픽 중에서 어떤 것이 특징적인 트렌드를 갖는 토픽인지에 대해서는 고려하지 않고 있다.

신출 트렌드 탐지(emerging trend detection)[6] 연구는 단어 출현 빈도수를 기반으로 여러 트렌드 중에서 신출 트렌드를 탐지하려고 시도하였다. 예를 들어

PatentMiner[11]는 Shape Definition Language를 소개하며 이를 활용하여 특허 데이타로부터 사용자가 정의한 형태의 트렌드를 찾는다. 이는 사용자에게 전문적인 지식을 요구하므로 대중적인 트렌드를 탐지하는데 적합하지 않다. HDDI[17]는 특허 문서로부터 일반적인 명사구를 트렌드로 간주하고 각 트렌드가 신출 트렌드인지 아닌지를 판단하였다. 이를 위해서 과거 기간에 각 명사구가 출현한 빈도수를 자질(feature)로 활용하여 신경망을 학습시켰다.

트렌드 분석을 위한 여러 가지 상용 서비스가 존재한다. GoogleTM Trends[4]는 사용자가 입력한 검색 질의의 입력 횟수 변화를 보여주는 서비스를 제공한다. BlogPulse[2]는 웹블로그로부터 주요 명사구(key phrase)를 추출하고 각 명사구마다 지난 2주간 문서에서 추출된 평균 횟수와 당일에 추출된 횟수 간의 비율을 burstness로 정의함으로써 트렌드를 탐지하였다. Autonomy[16]는 상업용 토픽 분석 도구로서 말뭉치를 입력받아 토픽 군집합을 생성하고 그 결과를 시간에 따라 보여줌으로써 트렌드를 탐지한다.

기존의 연구는 트렌드의 대상이 되는 개념, 객체, 토픽 등의 탐지에 초점을 두고 탐지된 대상의 시간에 따른 출현 빈도 수의 그래프를 트렌드로 보여주거나 단순하게 출현 빈도 합을 활용하여 트렌드의 순위를 결정하였다. 탐지할 수 있는 트렌드의 대상이 많을수록 그 중 가장 특징적인 트렌드를 결정하는 문제가 대두된다. 출현 빈도의 합을 이용한 방법은 여러 가지 면에서 특징적인 트렌드를 결정하는데 부족하다.

3. 문제 제기

본 연구에서 트렌드는 한 객체(object)에 대하여 주어진 기간 동안의 인기도 분포(popularity distribution)로 정의한다. 트렌드의 대상 객체는 사용자가 흥미를 가지는 토픽 중에서 명시적인 대상을 나타낸다. 예를 들어, 휴대폰의 모델명, 노트북의 모델명, 혹은 자동차 차종이 트렌드의 객체가 될 수 있다. 대상 객체의 인기도란 주어진 시간 동안 웹 문서로부터 대상 객체에 대한 긍정적인 표현의 수로 나타낸다. 하지만 이를 위해서는 모든 문서에서 대상 객체에 대한 감성 정보 분석(sentiment analysis)이 선행되어야 한다. 계산의 단순성 및 명료성을 위해서, 본 논문에서는 객체의 인기도를 특정 기간에 그 객체를 포함하고 있는 문서 빈도(document frequency)로 정의하였다.

특정 기간에 웹 문서나 말뭉치에서 높은 출현 빈도를 가지는 객체가 낮은 출현 빈도를 가지는 객체들보다 대중적인 인기를 얻고 있다고 볼 수 있다. 하지만 출현 빈도만을 이용하는 방법은 트렌드의 동적인 특징을 제대로 반영하지 못한다. 두 개의 서로 다른 객체(예: 두 휴대폰 모델)가 있다고 하자. 특정 기간 동안 각 객체의 누적 출현 빈도가 서로 같은 경우, 각 분포의 모양에 따라서 다른 해석을 할 수 있다. 표 1과 같은 경우, 모델1의 출현 빈도가 단조 증가하는 반면, 모델2의 출현 빈도는 변화량이 크지 않다.

표 1 동일 기간에 누적 출현 빈도가 같은 두 휴대폰

시각	5/1~	5/8~	5/15~	5/22~	5/29~	합
모델1	100	300	500	700	900	2500
모델2	500	500	400	500	600	2500

표 2는 누적 출현 빈도만을 이용하여 트렌드의 강도를 파악할 경우의 또 다른 한계를 보여준다. 표 2의 경우, 모델1이 모델2에 비해 큰 누적 출현 빈도를 가짐에도 불구하고 출현 빈도가 지속적으로 증가하는 모델2에 비해 모델1의 트렌드 강도가 더 크다고 판단하기가 명료하지 않다.

표 2 동일 기간에 누적 출현 빈도가 다른 두 휴대폰

시각	5/1~	5/8~	5/15~	5/22~	5/29~	합
모델1	180	200	220	240	260	1100
모델2	100	150	200	250	300	1000

이와 같이 누적 출현 빈도만을 이용하여 트렌드의 강도를 측정하고 이를 활용하여 순위를 결정한다면 트렌드의 동적인 특징을 반영한 세분화된 순위 결정을 할 수 없다. 다음 장에서는 이를 극복하기 위해서 4 가지 트렌드 속성을 제안한다.

4. 트렌드 속성

트렌드의 다양한 측면을 정량화하기 위해 정의한 트렌드 속성은 다음과 같다: 변동성(change), 지속성(persistency), 안정성(stability), 누적량(volume). 변동성은 트렌드 곡선의 시작점과 끝점의 차이와 같은 빈도 변화의 정도를 나타낸다. 지속성은 일정한 방향성을 가지고 증가/감소 하는 정도를 나타낸다. 안정성은 증가/감소의 폭이 변하지 않고 균일하게 증가하는 정도를 가리킨다. 마지막으로, 누적량은 트렌드 곡선의 형태와 무관하게 누적 빈도 값의 정도를 가리킨다. 본 논문에서 T_{start}와 T_{end}로 주어진 시간 T는 n 개의 시간 구간인 $<t_1,t_2,...,t_n>$으로 나누어지며, 각 구간은 일주일 단위이다. 네 가지 속성에 대한 정의 및 설명은 아래와 같다.

4.1 변동성

변동성은 트렌드 곡선의 기울기를 나타낸다. 이를 위하여 선형 회귀 분석된 트렌드 직선을 사용한다. 계산된

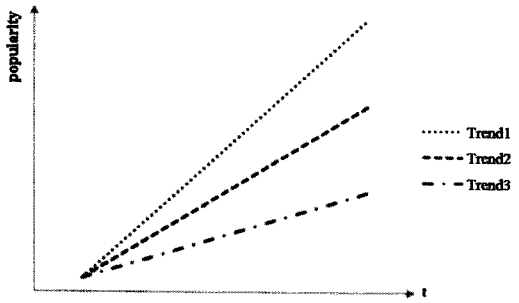


그림 1 서로 다른 변동성을 가지는 트렌드 직선

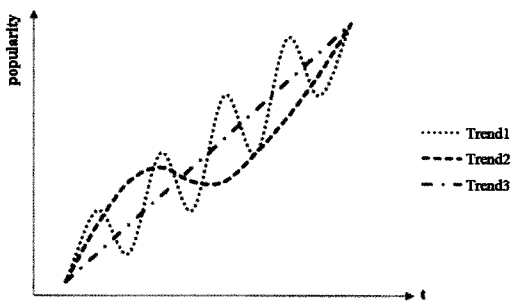


그림 2 서로 다른 지속성을 가지는 트렌드 직선

기울기 값이 양수인지 음수인지에 따라서 트렌드는 신출 트렌트 혹은 사양 트렌드(submerging trend)로 나뉜다. 두 경우 모두 기울기의 절댓값이 커질수록 그 트렌드 강도가 커진다.

그림 1에서 각 직선은 선형 회귀 분석된 트렌드를 나타낸다. 변동성은 사람들이 직선의 기울기가 클수록 트렌드 강도가 크다고 생각하는 것을 정량화한 속성이다. 그러므로 가장 가파른 기울기를 가지는 트렌드1의 트렌드 강도가 가장 크다.

시간을 정의역으로 하고 인기도를 공역으로 하는 트렌드 f에 대하여, 처음과 마지막의 인기도의 차이를 다음과 같이 정의한다.

$$diff(f) = LR_f(T_{end}) - LR_f(T_{start}) \quad (1)$$

위 식에서 f는 T_{start} 와 T_{end} 사이의 트렌드를 나타내며, LR_f는 f의 선형 회귀 분석을 가리킨다. $diff(f)$값이 양수이면 신출 트렌드, 음수이면 사양 트렌드를 나타낸다. 본 논문에서는 신출 트렌드에만 초점을 맞춘다.

변동성은 시간축과 선형 회귀 분석된 트렌드 사이의 기울기를 사인(sine) 함수를 이용함으로써 계산한다.

$$Change(f) = \frac{diff(f)}{\sqrt{diff(f)^2 + (T_{end} - T_{start})^2}} \quad (2)$$

4.2 변동성

지속성은 주어진 기간 동안 트렌드 곡선의 방향 변화 횟수를 나타낸다. 이는 방향 변화가 적은 트렌드가 방향 변화가 많은 트렌드보다 트렌드 강도가 크다라는 가정에 기반한다. 지속성은 런 길이 부호화(run-length encoding)[12]를 이용하여 측정하였다.

그림 2의 트렌드1, 2, 3은 서로 다른 방향 변화 횟수(8, 2, 0)를 가진다. 지속성 외에 다른 속성들이 같다고 한다면 방향 변화 횟수가 가장 적은 트렌드3의 트렌드 강도가 가장 크다.

런 길이 부호화는 간단한 형태의 압축 기법이다. 이는 주어진 데이타를 반복되는 값과 그 길이로 표현한다. 예를 들어, 위 그림에서 각 트렌드에 대하여 런 길이 부호화를 수행한 결과는 표 3에 나타나있다. U, D, S는 각각 증가, 감소, 수평을 의미하며, Seq, Comp는 각각 그래프의 연속된 방향, 압축된 방향을 나타낸다. 런 길이 부호화를 이용하여 곡선의 방향 변화와 지속성은 아래와 같이 정의 할 수 있다.

$$Persistency(f) = 1 - \frac{RUN(f)}{n} \quad (3)$$

RUN은 계산된 런의 수를 나타낸다.

표 3 런 길이 부호화를 통한 방향 변화 계산 예시

	Seq	Comp	# of runs
트렌드1	U D U D U D U D U	U1D1U1D1 U 1D1U1D1U1	9
트렌드2	U U U D S U U U U	U3D1S1U4	4
트렌드3	U U U U U U U U U	U9	1

4.3 변동성

안정성은 트렌드 곡선과 선형 회귀 분석된 직선의 사이의 거리차이를 나타내는 척도이다. 이는 트렌드 곡선이 선형 회귀 분석된 직선과 차이가 적을수록 트렌드 강도가 크다라는 가정에 기반한다. 즉, 일정하게 증가/감소하는 트렌드 곡선이 불규칙하게 증가/감소하는 트렌드 곡선보다 사람들에게 주목 받을 것이라는 것이다.

그림 3의 3가지 트렌드 곡선은 동일한 선형 회귀 분석 결과를 가진다. 이 중 트렌드3은 선형 회귀 분석 결

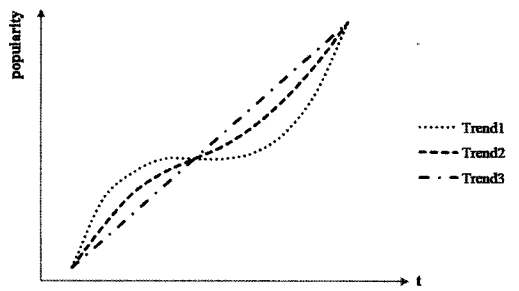


그림 3 서로 다른 안정성을 가지는 트렌드 직선

과에 가장 일치하기 때문에 트렌드 강도가 가장 크다. 트렌드2와 트렌드3은 동일한 변동성과 지속성을 가지지만, 트렌드3이 트렌드2보다 큰 안정성을 가지므로 트렌드 강도가 크다. 안정성을 정량적으로 측정하기 위하여, 트렌드 곡선과 선형 회귀 분석된 직선의 차이의 합을 다음과 같이 계산한다.

$$SumOfDiff(f)=\sum_{i=1}^{n}\left(f(t_i)-LR(t_i)\right)^2 \quad (4)$$

안정성은 *SumOfDiff*를 이용하여 다음과 같이 계산된다.

$$Stability(f)=(k)^{SumOfDiff(f)} \quad (5)$$

위 식에서 k는 0과 1사이의 상수이다. 본 연구에서는, k를 0.9로 설정한다. *SumOfDiff*가 커지면 안정성은 0에 가까워지며, *SumOfDiff*가 작아지면 안정성은 1에 수렴한다.

4.4 변동성

누적량은 주어진 기간의 누적 빈도를 의미한다. 누적량이 커질수록 트렌드 강도가 크다는 것을 의미한다. 누적량과 같이 출현 빈도를 이용하여 트렌드 강도를 측정하는 방법은 기존의 트렌드 분석방법에서 이미 널리 사용하고 있다. 그림 4에서 트렌드2의 인기도 합이 트렌드 1 보다 크므로 트렌드 강도 역시 크다.

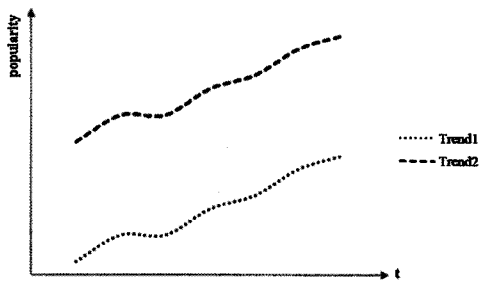


그림 4 서로 다른 누적량을 가지는 트렌드 직선

누적량은 아래 수식과 같이 평균 빈도수에 기반하여 계산된다.

$$AvgPopularity(f)=\frac{\sum_{i=1}^{n}f(t_i)}{n} \quad (6)$$

$$Volume(f)=\frac{AvgPopularity(f)-min}{max-min} \quad (7)$$

위 식에서 *max*와 *min*은 각각 기간 내에 주어진 모든 트렌드 중 인기도의 최댓값과 최솟값을 나타낸다. 분모는 누적량 값을 0에서 1사이가 되도록 정규화한다.

5. 트렌드 순위 결정

트렌드 강도(trend strength)는 소개된 4가지 속성의 선형적인 합을 통하여 계산할 수 있다. 하지만 모든 속성이 트렌드 순위 결정에 동일한 영향을 준다고 볼 수 없으므로 트렌드 강도는 각 속성과 그 속성의 기여도인 가중치에 대한 선형적인 합으로 나타낼 수 있다.

$$S(f)=w_c\times Change(f)+w_p\times Persistency(f)+w_s\times Stability(f)+w_v\times Volume(f). \quad (8)$$

위 식에서 w_c,w_p,w_s,w_v는 각각 각 속성의 가중치를 나타낸다($0\le w_c,w_p,w_s,w_v\le 1$).

트렌드 강도의 가중치를 찾는 문제는 위의 선형식을 목적함수로 하고, 목적함수의 결과와 학습 트렌드 간의 상관관계를 최대화하는 가중치를 찾는 최적화 문제로 변환할 수 있다. 최적화 문제의 해결책으로 여러 방법이 있으나 본 논문에서는 유전 알고리즘을 이용하였다[13]. 4개의 가중치들을 하나의 염색체로 부호화하고, 스피어만 상관 계수(Spearman's rank correlation coefficient) [14]를 적합도 함수(fitness function)로 사용하였다.

6. 실험 및 결과

본 논문에서 제안하는 속성들의 유용성과 순위 결정 방법의 우수성을 증명하기 위하여 이전 연구에서 널리 사용한 누적량만을 고려한 트렌드 순위 결정 결과와 다른 속성들을 함께 또는 개별적으로 사용한 트렌드 순위 결정 결과를 비교하였다. 평가는 스피어만 상관 계수를 사용하였다.

6.1 데이타

학습 데이타는 총 350개의 케이스(case)를 구축하고 사용하였다. 각 케이스는 5개의 서브 트렌드로 구성되어 있고 각 트렌드는 서로 다른 속성값을 가지고 있다. 따라서 350개의 케이스가 4가지의 속성으로 가능한 모든 조합을 포함할 수 있다고 가정한다.

350개의 케이스 중 200개의 케이스는 실제 데이타부터 구축하였다. 본 논문에서 휴대폰 영역의 데이타를 선택하였는데 이는 휴대폰 영역에 빠르게 변화하는 다양한 트렌드가 존재하기 때문이다. 트렌드의 대상 객체로 모바일디아[1]에서 690개의 휴대폰 모델명을 수집하고, 각 휴대폰 모델명을 [2]구글 블로그 검색 서비스의 검색 질의로 사용하여 검색 가능한 모든 블로그 공간에서 2002년부터 2007년까지 한 주 단위로 휴대폰의 인기도 분포를 생성하였다. 각 인기도 분포를 25주 단위의 여러 개의 서브 트렌드로 나누었다. 그리고 모든 서브 트렌드 수집하여 하나의 트렌드 저장소(trend pool)를 구축하였다. 이 중, 무작위로 5개의 서브 트렌드를 선출하여 하나의 케이스를 만든다. 이런 방법을 통하여 학습 데이타

1) http://www.mobiledia.com/
2) http://blogsearch.google.com/

로 쓰일 200개의 케이스를 제작하였다.

200개의 학습 데이타가 본 논문에서 제안하는 4 가지 속성을 모두 반영하고 있다고 단정지을 수 없다. 그렇기에 네 가지 속성을 명시적으로 반영하고 있는 150개의 케이스를 가우시안 분포를 이용하여 인공적으로 구축하였다. 각 케이스는 25개의 인기도를 가지고 있는 5 개의 인공적인 트렌드로 구성된다.

생성된 학습 데이타는 정의된 4가지의 속성을 알고 있는 3명의 저자가 각 케이스 내에 있는 트렌드에 대해서 1-5위까지 순위 결정하였다. 강도가 가장 큰 트렌드에 1를 부여하고 가장 작은 트렌드에 5를 부여하였다. 이 순위를 점수화하여 각 트렌드에 따른 누적점수를 계산한 후 최종 순위를 결정하였다. 표 4는 하나의 학습 케이스에 대한 트렌드 순위 결정의 예를 보여준다.

표 4 학습 트렌드 순위 결정 예제

	저자1	저자2	저자3	누적점수	최종순위
트렌드1	1	2	1	4	1
트렌드2	2	1	3	6	2
트렌드3	3	3	2	8	3
트렌드4	4	4	4	12	4
트렌드5	5	5	5	15	5

평가 데이타로는 트렌드 저장소로부터 학습 데이타로 사용된 것을 제외한 나머지 트렌드 중에서 무작위로 5개의 트렌드를 선택하여 50개의 케이스를 생성했다. 50개의 케이스에 대하여, 정의된 4가지 속성을 모르는 4명의 사람이 각 케이스에 대해서 1-5위까지 순위 결정하고 학습 데이타 순위 결정과 동일한 방법을 적용하여 최종 순위를 결정하였다. 이렇게 최종 순위가 결정된 평가 데이타를 금본위 데이타(gold standard)로 이용하였다.

6.2 실험 결과 및 분석

본 실험의 목적은 앞서 정의한 4가지의 속성들이 트렌드 강도를 측정에 있어 유용성을 입증하는 것이다. 4가지의 속성 중 누적량은 누적 출현 빈도수와 같은 의미로 이미 다른 연구에서 일반적으로 자주 쓰고 있는 것이지만, 다른 3가지의 속성은 본 논문에서 처음 정의되는 것이기 때문에 각 속성의 유용성 입증이 필요하다. [15]는 웹 문서에서 키워드를 찾기 위해 사용된 속성들의 유용성을 검증하기 위해서 기본 속성(base attribute)에 여러 속성들을 추가하고 제거함으로써 각 속성의 유용성을 증명하였다. 본 논문의 실험에서도 정의한 속성들의 유용성을 증명하기 위하여 같은 방법의 실험을 수행하였다.

첫 번째 실험에서 기본 속성으로는 이미 이전 연구에서 자주 쓰이고 있는 누적량을 선택하였고, 이에 다른 3가지 속성들을 각각 추가하여, 3가지 속성의 유용성을 보였다(표 5). 각 속성이 추가되어 결합할 때마다 가중치를 결정하기 위하여 유전 알고리즘이 적용되었다. 표에서도 볼 수 있듯이, 각 속성을 추가함으로써, 기본 속성만 사용된 경우보다 제안한 순위 결정 방법의 결과와 금본위 데이타 사이의 상관 관계가 증가함을 알 수 있다. 3가지의 속성 중, 변동성을 추가하였을 경우에 가장 높은 성능 향상을 보였다. 안정성과 지속성도 성능 향상에 영향을 주었지만 변동성보다는 향상 폭이 적었다.

표 5 누적량에 다른 속성을 추가함에 따른 효과

	Attribute	*Correlation*
Case1	Volume (baseline)	0.4320
Case2	Volume + Change	0.6780
Case3	Volume + Persistency	0.4800
Case4	Volume + Stability	0.5020

표 6은 4가지 속성을 사용한 결과에서 하나의 속성을 제거한 결과를 비교하여 보여준다. 이 실험을 통하여, 모든 속성을 적용하였을 때(0.7580)는 누적량만을 사용했을 때 (0.4320)에 비해 높은 성능 향상을 보였다. 또한, 표에서 알 수 있듯이, 각 속성을 제거함으로써, 속성마다 편차는 다르지만, 모든 경우가 전체 속성을 사용하였을 경우보다 낮은 상관 관계를 나타내었다. 특히, 변동성이 제거되었을 때 최하의 성능을 보였고, 안정성 제거가 두 번째로 낮은 성능을 보임으로써 변동성과 안정성이 트렌드 강도를 결정하는데 중요한 속성이란 것을 알 수 있다. 지속성의 제거는 다른 속성들에 비해 아주 작은 성능 하락을 보였다. 하지만 지속성의 추가와 제거가 성능에 영향을 미치고 있으므로 지속성 역시 트렌드 순위 결정에 필요한 속성임을 알 수 있다. 흥미로운 사실은 누적량의 제거가 기대했던 것만큼 성능 하락에 큰 영향을 주지 않는다는 것이다.

표 6 각 속성을 제거함에 따른 효과

	Attribute	*Correlation*
Case1	All Attributes	0.7580
Case2	- Change	0.4920
Case3	- Persistency	0.7560
Case4	- Stability	0.6960
Case5	- Volume	0.7200

표 5와 표 6을 통하여 각 속성이 전체 성능에 미치는 영향을 알 수 있었다. 표 7에서는 누적량을 기본 속성으로 하여 순차적으로 다른 속성을 적용함으로써 그 성능 향상의 폭을 측정하였다. 다른 속성을 추가할수록 성능 향상의 폭은 줄어들지만 성능 향상을 보여주었다.

표 7 누적량에 순차적으로 속성을 추가한 결과

	Attribute	*Correlation*
Case1	Volume	0.4320
Case2	Volume + Change	0.6780
Case3	Volume + Change + Stability	0.7560
Case4	Volume + Change + Stability + Persistency	0.7580

표 7에서 볼 수 있듯이 다른 속성들과 다르게 변동성을 추가하면 성능이 월등히 향상됨을 알 수 있다. 이를 고려하여, 앞에서 한 실험과 유사하게 변동성을 기본 속성으로 하여 순차적으로 다른 속성들을 추가해 결과를 비교해 보았다(표 8). 놀라운 결과 중 하나는 변동성만 고려했을 경우(0.5880)가 기본 시스템(baseline)으로 생각했던 누적량만 고려했을 경우(0.4320) 보다 높은 성능을 갖는다는 것이다. 또한, 남은 3개의 속성을 추가했을 때, 가장 높은 성능 향상을 보인 것이 누적량이 아니라 안정성을 추가한 경우였다.

일련의 실험을 통하여 사람들이 트렌드 순위 결정에 있어서 가장 많이 고려하는 것이, 우리가 기존에 알고 있었던 누적량, 즉 누적 출현 빈도가 아니라 변동성임을 알 수 있었다. 그리고 각 속성 또는 속성들의 조합마다 트렌드 순위 결정에 미치는 영향이 다르다는 것을 알 수 있었다.

표 8 변동성에 순차적으로 속성을 추가한 결과

	Attribute	*Correlation*
Case1	Change	0.5880
Case2	Change + Stability	0.7200
Case3	Change + Volume	0.6780
Case4	Change + Persistency	0.6100
Case5	Change + Stability + Volume	0.7560
Case6	Change + Stability + Volume + Persistency	0.7580

7. 결 론

이전 연구들은 트렌드의 대상 객체(개념, 객체, 토픽 등)의 탐지에 초점을 두고 탐지된 대상 객체의 트렌드를 출현 빈도 수를 그래프로 부여주거나 출현 빈도 수의 합을 이용하여 순위를 결정하였다. 트렌드를 탐지할 수 있는 대상의 수가 많아질수록 많은 트렌드 중 특징적인 트렌드를 선별하는 문제가 중요해진다. 본 논문은 과거 출현 빈도 합을 이용한 트렌드 순위 결정 방법의 한계를 소개하며 이를 극복하기 위한 4 가지 속성(변동성, 지속성, 안정성, 누적량)을 정의하고 4 가지 속성과 유전 알고리즘을 이용하는 새로운 트렌드 순위 결정 방법을 제안하였다. 각 속성들과 제안된 순위 결정 방법을 검증하기 위하여 실제 웹 문서로부터 50개의 평가 케이스와 그에 상응하는 금본위 트렌드를 구축하고 이용하여 여러 가지 실험을 하였다. 실험 결과로부터 트렌드 순위 결정에 있어서 정의된 4 가지 속성의 유용성을 검증하고 이를 조합하여 사용한 새로운 순위 결정 방법의 우수성을 입증하였다.

참 고 문 헌

[1] Firminger, L., Trend Analysis: a collection of sub methodologies, Swinburne University of Technology, 2003.

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

[3] Mei, Q. and C.X. Zhai., Discovering evolutionary theme patterns from text: an exploration of temporal text mining, *In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005.

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

[5] Lavrenko, V., et al., Language models for financial news recommendation, *In Proceedings of the ninth international conference on Information and knowledge management*, 2000.

[6] Kontostathis, A., et al., A Survey of Emerging Trend Detection in Textual Data Mining, In Survey of Text Mining: Clustering, Classification, and Retrieval, 2003.

[7] Rajaraman, K. and A.H. Tan, Topic Detection, Tracking, and Trend Analysis Using Self-Organizing Neural Networks, *In Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2001.

[8] Morinaga, S. and K. Yamanishi, Tracking dynamics of topic trends using a finite mixture model. *In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004.

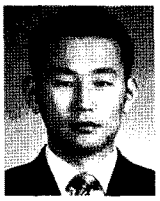
[9] Fung, G.P.C., J.X. Yu, and W. Lam, News Sensitive Stock Trend Prediction. *In Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 2002.

[10] Lee, J., S. Cho, and J. Baek, Trend detection using auto-associative neural networks: Intraday KOSPI 200 futures, *In Computational Intelligence for Financial Engineering*, 2003.

[11] Lent, B., R. Agrawal, and R. Srikant, Discovering trends in text databases, *In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining* (KDD), 1997.

[12] Run-Length Encoding: http://en.wikipedia.org/wiki/Run-length_encoding.

[13] Wright, A.H., Genetic algorithms for real parameter optimization. Foundations of Genetic Algorithms, 1991.
[14] Budanitsky, A. and G. Hirst, Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 2006.
[15] Yih, W., J. Goodman, and V.R. Carvalho, Finding advertising keywords on web pages. *In Proceedings of the 15th international conference on World Wide Web*, 2006.
[16] Verity, http://www.verity.com
[17] Holzman, L.E., Fisher, Fisher, T.A., Galisky, L. M., Kontostathis, A., and Pottenger, W. M., A Software Infrastructure for Research in Textual Data Mining. *The International Journal of Artificial Intelligence Tools*, volume 14, 2004.



오 홍 선
2006년 한국항공대학교 컴퓨터공학과 학사. 2009년 한국정보통신대학교 전산학 석사. 2009년~현재 한국과학기술원 정보통신공학과 박사과정. 관심분야는 Relation Extraction, Named Entity Recognition, Ads Placement, Trend Analysis

최 윤 정
2007년 한국정보통신대학교 전산학 학사 2007년~현재 한국과학기술원 정보통신공학과 석사과정. 관심분야는 Trend Analysis, Topic Detection, Mood Detection, Ads Placement, User Interest

신 욱 현
2007년 한국정보통신대학교 전산학 학사 2007년~현재 한국과학기술원 정보통신공학과 석사과정. 관심분야는 Blog Search, Social Intelligence, Ads Placement, Trend Analysis

정 윤 재
1998년 포항공과대학교 학사. 2007년 한국정보통신대학교 전산학 석사. 2007년~현재 한국과학기술원 정보통신공학과 박사과정. 관심분야는 Knowledge Discovery, Social Intelligence, Complex System

맹 성 현
1983년 미국 캘리포니아 주립대학 학사 1985년 미국 Southern Methodist University(SMU) 석사. 1987년 미국 Southern Methodist University(SMU) 박사. 1987년~1988년 미국 Temple University 교수. 1988년~1994년 미국 Syracuse University 교수(tenured). 1994년~2003년 충남대학교 컴퓨터과학과 교수. 2003년~2009년 한국정보통신대학교 교수. 2009년~현재 한국과학기술원 교수. 관심분야는 정보검색, 텍스트 마이닝, HCI, 상황인지 컴퓨팅 등