인기 검색어의 순위 변화 예측

'Hot Search Keyword' Rank-Change Prediction

저자
(Authors)

김도형, 강병호, 이승룡
Dohyeong Kim, Byeong Ho Kang, Sungyoung Lee

출처
(Source)

정보과학회논문지 44(8), 2017.8, 782-790 (9 pages)
Journal of KIISE 44(8), 2017.8, 782-790 (9 pages)

발행처
(Publisher)

한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL

http://www.dbpia.co.kr/Article/NODE07226476

APA Style

김도형, 강병호, 이승룡 (2017). 인기 검색어의 순위 변화 예측. 정보과학회논문지, 44(8), 782-790.

ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 44, No. 8, pp. 782-790, 2017. 8
https://doi.org/10.5626/JOK.2017.44.8.782

인기 검색어의 순위 변화 예측
('Hot Search Keyword' Rank-Change Prediction)

김 도 형 † 강 병 호 †† 이 승 룡 †††
(Dohyeong Kim) (Byeong Ho Kang) (Sungyoung Lee)

요 약 인기 검색어 리스트는 현재 가장 인기 있는 검색어의 순위를 보여주는 서비스로서 네이버와 같은 포털사이트가 제공한다. 이 리스트에서의 순위 변화는 특정 검색어에 대한 사람들의 관심의 변화를 반영한다. 본 논문은 인기 검색어의 순위 변화를 예측하기 위해 시계열 모델링 프레임워크를 제안한다. 제안한 프레임워크는 과거 순위와 기계학습 모델이 적용되었고, 여기서 해결해야 할 두 가지 문제점이 있다. 첫째, 과거 순위 데이터를 분석한 결과, 70% 이상의 검색어가 리스트에서 소멸 후 재출현하는 현상을 보였다. 소멸 후의 순위는 손실 값으로 볼 수 있으며, 이를 해결하기 위해서 다양한 처리 방법을 적용하였다. 둘째, 과거 순위 데이터는 시계열 데이터이므로 최적 윈도우 크기를 계산하는 것이 중요하다. 본 논문에서는 최적 윈도우 크기는 동일한 검색어들이 서로 다른 두 시점에서 내용상 의미가 달라지는 최단 소멸기간으로 볼 수 있음을 밝혔다. 성능 평가를 위해서 4가지의 기계학습 기법과 2년 동안 수집한 네이버, 다음, 네이트의 인기 검색어 리스트 데이터를 사용하였다.

키워드: 인기 검색어, 시계열 예측, 검색어 예측, 기계학습

Abstract The service, 'Hot Search Keywords', provides a list of the most hot search terms of different web services such as Naver or Daum. The service, bases the changes in rank of a specific search keyword on changes in its users' interest. This paper introduces a temporal modelling framework for predicting the rank change of hot search keywords using past rank data and machine learning. Past rank data shows that more than 70% of hot search keywords tend to disappear and reappear later. The authors processed missing rank value, using deletion, dummy variables, mean substitution, and expectation maximization. It is however crucial to calculate the optimal window size of the past rank data. We proposed an optimal window size selection approach based on the minimum amount of time a topic within the same or a differing context disappeared. The experiments were conducted with four different machine-learning techniques using the Naver, Daum, and Nate 'Hot Search Keywords' datasets, which were collected for 2 years.

Keywords: hot search keyword, temporal prediction, search keyword prediction, machine learning

· This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning(2011-0030079). This work was supported by the Industrial Core Technology Development Program (10049079, Develop of mining core technology exploiting personal big data) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea).

† 학생회원 : 경희대학교 컴퓨터공학과
dhkim@oslab.khu.ac.kr
†† 비 회 원 : University of Tasmania, School of Engineering and ICT 교수
byeong.kang@utas.edu.au
††† 종신회원 : 경희대학교 컴퓨터공학과 교수(Kyung Hee Univ.)
sylee@oslab.khu.ac.kr
(Corresponding author임)

논문접수 : 2016년 12월 27일
(Received 27 December 2016)
논문수정 : 2017년 4월 27일
(Revised 27 April 2017)
심사완료 : 2017년 5월 20일
(Accepted 20 May 2017)

정보과학회논문지 제44권 제8호(2017. 8)

1. 서 론

인터넷 사용자는 검색엔진, 소셜미디어 및 뉴스 집계 사이트와 같은 다양한 유형의 웹 기반 서비스를 사용하여 정보를 공유하고 검색할 수 있다. 이러한 웹 기반 서비스는 개인 정보의 공유 및 수집을 증가시킴으로써 거대한 정보 공유 패러다임으로의 전환을 가져왔다. 이 현상은 "소셜 데이터 혁명"이라고 불리며 전례 없는 양의 소셜 데이터가 축적되고 있다. 이처럼 수많은 사용자가 만들어 내는 소셜 데이터는 21세기의 미 개척된 금광과도 같다고 할 수 있다. 대표적으로 웹 기반 업체가 유저들 사이에서 검색되는 가장 인기 있는 검색어를 표시하는 실시간 인기 검색어 서비스를 제공하고 있다. 예를 들면, 포털 사이트인 네이버는 '실시간 급상승', 다음은 '실시간 이슈'라는 실시간 인기 검색어(실검) 리스트를 제공한다. 이 리스트에는 현재 화제가 되는 상위 10개의 인기 검색어가 표시되며 포털사이트 내에 인터페이스의 일부로 보이므로 모든 사용자가 현재에 화제가 되는 항목을 쉽게 식별할 수 있다. 실시간 검색어는 사람들의 관점에서 현재 사회의 이슈를 반영한다. 이러한 인기 검색어 중 85% 이상은 뉴스 속보 헤드라인과 관련이 있으며, 각 인기 검색어와 관련된 뉴스는 더 구체적인 정보를 제공할 뿐만 아니라 댓글을 통해 사용자 의견을 반영하여 더 구체적인 정보(댓글)를 제공한다[1].

실검 리스트는 인기가 높은 순서대로 상위 10개의 인기 검색어를 보여준다. 이 검색어 순위 변화는 해당 검색어의 현재 인기 정도를 파악하는 척도가 된다. 예를 들어, 2014년 4월 16일, 진도 해상에서 세월호가 침몰하게 된다.[1)] 이 기간 동안 '세월호'라는 검색어가 실검리스트에 나타났다. 그림 1은 검색어 '세월호'가 실검 리스트에 처음 등장한 시점부터 24시간 동안의 시간대별 순위 변화를 보여준다. 검색어가 사람들의 관심 변화에 따라서 한 시간 단위로 순위가 변하는 것을 볼 수 있다. 시간대별 순위 변화는 '상승, 하강 및 무 변동'이라는 세 가지 범주로 분류할 수 있고 이 순위 변화는 사람들의 관심이 높아지거나 낮아지거나 변하지 않음을 반영한다. 따라서 실시간 검색어의 순위 변화를 매시간 예측하는 것은 그 검색어가 가까운 미래에 사람들에게 얼마나 관심을 받을지, 그리고 사회에 어떠한 영향을 끼칠지 파악할 수 있다. 하지만, 실검 리스트에는 한정된 정보(검색어, 순위 및 업데이트 날짜, 시간)만을 제공하는 한계가 있다. 본 논문은 이 한정된 정보만으로 실시간 검색어의

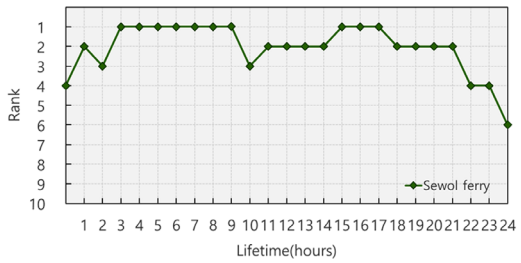


그림 1 24시간 동안 검색어 '세월호'의 실검 리스트 순위 변화

Fig. 1 The rank change of the search keyword 'Sewol Ferry' for 24 hours

향후 순위 변화 예측을 위한 연구를 하였다.

본 논문의 연구 목적은 "인기 검색어의 순위 변화(상승, 하강, 무 변동)를 예측할 수 있는가?"라는 의문을 해결하는 것이다.

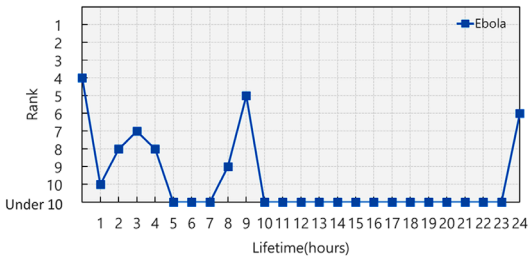
이 연구 목적을 달성하기 위해 본 논문은 과거 검색어 순위 데이터와 기계학습 기법을 사용하여 시계열 모델링 프레임워크를 제안하였다. 이 프레임워크에서 FRC (Future Rank Change) 즉, 시간 t에서, 실시간 검색어 T_x의 향후 순위 변화 예측은 다음과 같이 표현할 수 있다.

$$FRC(T_x) = f[r_{t-n},\ldots,r_{t-1},r_t], (n>0) \quad (1)$$

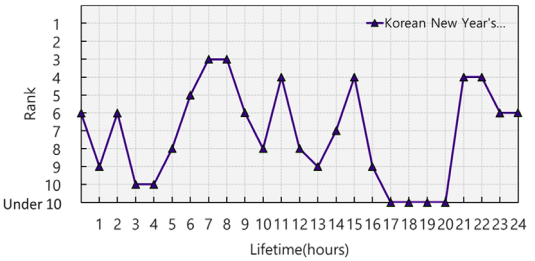
위 수식에서 f는 기계학습 기법이고 n 시간 동안의 실시간 검색어 T_x의 과거 순위는 $[r_{t-n},\ldots,r_{t-1},r_t]$ 이다. 실시간 검색어에 대한 예측된 검색어의 순위 변화 FRC는 세 가지 클래스(상승, 하강, 무 변동)로 분류될 수 있다. 예를 들어 검색어 '세월호'의 순위 변화가 20시에서부터 21시까지 하강한다고 예측을 가정할 때, 0시에서 20시까지의 과거 순위 데이터를 사용할 수 있다.

순위 변화 예측을 위해서 과거 순위 데이터를 사용하려면 조사해야 할 문제가 있다. 실검 리스트의 검색어는 사라지고 다시 나타나는 경향이 있으므로 검색어가 리스트에서 사라졌을 때 정확한 순위를 아는 것은 불가능하다. 그림 2는 '에볼라'와 '설날 인사말'이라는 두 가지 다른 검색어의 실종 문제 예시를 잘 보여준다. 이 예시는 실검 리스트에서 검색어가 처음 발견된 시점에서 24시간 동안 검색어 순위 패턴을 보여준다. (a)는 2014년 8월 21일 정오에 발생한 검색어 '에볼라'의 24시간 순위 변동을 보여준다. '에볼라'는 유행성 전염병으로 새로운 환자가 발생하거나 위급할 때마다 검색어는 실검 리스트에 출현하고 순위가 올라간다. 발병 소식이 없는 5~7시, 10~23시에는 소멸하였다가 환자 혹은 위급 사태가 발생하면 재출현하였다. '설날 인사말'은 계절적으로 유행하는 검색어이다. 이 시점에 네이버 사용자들은 설날에 대해서 검색을 하고 있다. 2015년 설날은 2월 19일

1) 진도해상서 350여명 탄 여객선 조난신고…침수 중(1보), 〈연합뉴스〉, 2014/04/16 10:36, http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=102&oid=001&aid=0006864701(접속시간: 2016/12/09 00:00)



(a) The rank change of the search Keyword 'Ebola' for 24 hours (From 12am, 21 August 2014)



(b) The rank change of the search Keyword 'New Year's Greetings' for 24 hours (From 10am, 17 February 2015)

그림 2 실검 리스트에서 검색어 소멸

Fig. 2 Search keyword disappearance from the hot search keyword list

이며 (b)는 설날 연휴 전날인 17일에 검색어 '설날 인사말'에 대해서 오전 10시부터 한 시간마다 변화를 보여준다. 그림에서 보이듯, 17일 아침 10시에 실검 리스트에서 처음 출현했고 새벽 3시부터 6시 사이에 소멸하였고 설 연휴 시작 아침 7시에 재출현하였다.

과거 순위 데이터에서 검색어의 70%는 사라지고 이후에 다시 나타나는 경향을 보인다. 그러므로 예측 모델에서 '소멸과 재출현' 현상을 반영하는 것이 중요하며 이는 손실 값(Missing value)과 윈도우 크기를 다루는 것과 연관되어 있다. 우선, 검색어가 사라진 동안 누락 순위 값을 처리해야 한다.

본 논문의 3.1절에서는 누락 순위 값을 처리하기 위해 사용할 수 있는 네 가지의 다른 손실 값 처리 방법에 관해서 기술하였고 5.2절에서는 그 네 가지 방법 중 검색어 순위 예측에서 가장 적합한 방법을 실험을 통해 도출하였다. 또한, 시계열 특성을 가진 과거 순위 데이터를 예측 프레임워크에 사용하기 위해서는 최적의 윈도우 크기를 선택해야 한다. 기존의 방법과는 다르게 임의의 윈도우 크기를 선택하기보다는, 본 논문에서는 검색어의 순위 변화를 예측하기 위해 적절한 윈도우 크기를 결정하는 방법을 3.2절에서 제안하였다.

2. 관련연구

소셜 미디어를 사용하여 정치적 사건을 예측하기 위한 여러 연구가 있다. 가장 대표적인 연구 중 하나는 Chung과 Mustafaraj[2]의 연구로서 선거 예측을 위해 트위터와 뉴스 분석을 적용하는 것이었다. 저자는 후보자의 이름이 포함된 모든 관련 트위터와 뉴스를 수집하였고, 정서 분석을 수행하였다. 또한, 다음 총선은 Franch[3]에 의해 예측되었다. 저자는 다양한 소셜 미디어의 소셜 데이터를 사용하였고 이러한 미디어는 각 해당 미디어별로 분류하였다. 예측 성능은 ARIMA(auto regressive integrated moving average)를 사용하여 평가되었으며, 실제 투표 점유율에서 0.48%와 0.83%를 달성하였다.

소셜 데이터는 경제 분야에서도 많은 주목을 받고 있다. 주식 가격은 현실 세계의 사건과 사람들의 관점을 기반으로 동적으로 변한다. 따라서 소셜 미디어는 주가 추세를 예측하는 위한 가장 효과적인 자원이 될 수 있다. Sprenger[4]는 트위터를 주식 관련 여론을 추출하기 위한 집단지성으로 활용하는 포럼으로 간주하였다. Bollen et al.[5]는 정서 분석을 하고 정서가 주식 시장에서 의사결정에 실제로 어떻게 영향을 미칠 수 있는지를 조사하였다. 제안된 접근법은 일일 예측에서 86.7%의 정확도를 얻었다. Si et al.[6]은 2012년부터 1년간의 캐시 태그($)가 포함된 트윗 메시지들을 기반으로 시맨틱 주식 네트워크를 구축하였으며, 사회적 감정 상태를 분류하여 주식시장을 예측하는 데 사용하였다.

사람들은 소셜미디어 사이트로부터의 데이터가 사용자들의 관심사를 잘 표현하는 것으로 보임에 따라서 소셜미디어 사이트들은 인기 검색어로 불리는 가장 많이 화제 되고 검색된 검색어를 게시하기 시작하였다[7]. 이러한 인기 검색어 서비스는 많은 관심을 받고 있다. 예를 들어, 트위터가 제공하는 실시간 이벤트 검색서비스인 '트위터 인기 검색어'는 가장 빈번하게 언급되거나 게시된 단문, 단어, 헤시 태그를 보여 준다[8]. 그러나 검색어에 대한 용어와 순위는 자세하게 설명하지 않았다.

따라서 많은 연구자는 검색어의 분명한 의미를 밝히기 위해 다양한 요약 및 추출 방법을 적용하였다. Han and Chung[9]는 방법을 분명하게 하기 위해서 단순한 용어 빈도(Term Frequency) 접근법을 대표 키워드를 추출하는 데에 적용하였다. 또한, 검색어의 의미를 밝혀내기 위한 가장 성공적인 접근 방법은 단순한 용어 빈도임을 증명하였고 이는 대학원 학생 20명에 의해 평가되었다. Han et al.[10]은 중국 검색엔진인 바이두에서 검색어의 트렌드를 분석하는 방법을 제안하였다. Jaidka

et. al[11]은 점진적 클러스터링 방식을 사용하여 실시간으로 트윗 메시지들을 군집화하고 인기단어를 트위터 인기 검색어 서비스에 나타나기 전에 예측할 수 있는 성과를 내었다. Vakali et al.[12] 또한 같은 목적으로 실시간 분류 방법을 사용하여 인기 검색어를 서비스에서 제공하기 전에 예측할 수 있는 프레임워크를 제안하였다.

일부 연구자들은 검색어를 분류하여 조사하였다. Lee et al.[13]은 라벨링과 기계학습 기법을 적용하여 검색어를 18개의 일반적인 카테고리로 분류하였다. Kim et al.[14]은 검색엔진, 소셜미디어, 뉴스 사이트를 각각의 온라인 커뮤니티로 설정한 후 각각에 제공되는 인기 검색어를 분석하였다. 이 연구를 통하여 각 커뮤니티의 사용자들이 어떤 식으로 관심도가 변화하고 상호 작용하는지에 대해서 예측할 수 있는 프레임워크를 제안하였다.

소셜 미디어를 이용한 다양한 유형의 예측 연구가 수행되었다. Nikolov와 Shah[15]는 트위터에서 화제어를 조기에 감지하기 위한 새로운 알고리즘을 제안하였다. 성능 평가에서 95% 정확도를 달성하였으나, 검색어 순위와 순위 변화에 대한 예측은 다루지 않았다.

3. 실시간 인기 검색어 변화 모델

본 논문은 과거 순위 패턴과 기계학습 기법을 사용하여 검색어의 순위 변화를 예측하기 위한 시계열 모델링 프레임워크를 제안한다.

제안된 모델은 다음 수식으로 설명할 수 있다.

$$FRC(T_x) = ML(PRP(T_x)) \quad (2)$$

특정 검색어 T_x의 다음 순위 변화 FRC(Future Rank Change)를 예측하기 위해서 검색어 T_x의 과거 순위 패턴(PRP: Past Rank Pattern) 데이터를 사용하였다. 제안된 모델을 학습시키기 위해서 여러 가지 기계학습 기법(ML)이 적용되었고, 그 결과는 5.2에서 볼 수 있다.

식 (3)은 시간 t에서 특정 검색어 T_x의 과거 순위 패턴의 예를 설명한다. 이 수식은 특정 기간 n에 있는 검색어 T_x의 모든 과거 순위 패턴을 나타낸다. FRC는 다음 시간에 검색어 순위의 추세를 나타낸다. 현재 순위와 한 시간 이후의 순위를 비교하여 그 순위 변화의 예측 형태는 up, down, unchanged 세 가지 클래스 중 하나가 된다.

예를 들어, 한 시간 이후 순위인 r_{t+1}의 값이 현재 순위 r_t값보다 높으면 FRC는 'down'으로 표현된다.

$$PRP(T_x) = [r_{t-n}, ..., r_{t-1}, r_t], (n > 0) \quad (3)$$

$$FRC(T_x) = \begin{cases} up, & \text{if } r_t - r_{t+1} > 0 \\ down, & \text{if } r_t - r_{t+1} < 0 \\ unchanged, & \text{if } r_t - r_{t+1} = 0 \end{cases} \quad (4)$$

제안된 모델에 과거 순위 데이터를 적용하려 할 때, 누락된 순위 처리와 윈도우 크기 선택이라는 두 가지 문제가 존재한다.

첫째, 앞서 말했듯 많은 검색어가 실검 리스트에서 사라지고 다시 나타나는 경향이 있다. 본 연구는 검색어가 사라진 동안에 누락된 순위 값을 처리하는 방법을 실험을 통해 제안한다. 둘째, 검색어의 과거 순위 패턴은 시계열 데이터이므로, 예측을 위해서는 적절한 윈도우 크기를 선택하는 것이 중요하다. 본 연구는 실험에 사용된 검색어 순위 도메인에 맞는 최적 윈도우 크기를 계산하는 방법을 제안한다. 이 두 가지 문제에 대해서 자세한 내용은 3.1 누락 순위 처리와 3.2 윈도우 크기 선택 절에서 확인할 수 있다.

3.1 누락 순위 처리

실검 리스트는 그 순간의 상위 10개의 검색어를 보여주기 때문에, 순위 1에서 10까지의 검색어가 나타난다. 그러므로 검색어가 갑자기 실검 리스트에서 사라지면 검색어의 순위가 11위인지 50위인지 정확한 순위를 알 수 없다. 본 논문에서 검색어들을 관찰하고 조사한 결과 이러한 소멸과 재출현 패턴이 검색어의 모든 유형에서 나타난다는 것을 발견하였다. 뉴스 속보나 상시 뉴스(예: TV 프로그램 또는 스포츠 경기) 등 다양한 유형의 인기 검색어가 무작위로 사라지고 다시 출현한다. 따라서 본 연구는 실제로 얼마나 많은 검색어가 사라지고 다시 나타나는지를 분석하였다. 표 1은 검색어가 사라진 후에 재출현하거나 아예 출현하지 않는 검색어의 비율을 보여주며 재출현하는 검색어의 비율은 약 72.64%이다.

제안된 예측 모델에서 누락된 순위 데이터를 처리하는 것이 중요하다. 제안된 모델은 Allison[16]이 검토한 방법 중 4가지 손실 값 처리 방법(Pairwise deletion, dummy variable, Mean substitution과 Expectation Maximization(EM))이 적용됐다. 이러한 접근법의 예측 결과는 '5절 평가 결과'에서 논의된다.

표 1 재출현 또는 미출현 하는 검색어 비율
Table 1 The percentage of hot search keywords that reappeared or "non-reappeared"

	Reappearance	Non-reappearance
Percentage	72.64%	27.36%

3.2 윈도우 크기 선택

제안된 시계열 모델은 과거 검색어 순위 데이터를 사용하고 기계학습 기법으로 학습하므로 훈련 및 테스트에 대해서 동일한 윈도우 크기의 시퀀스를 사용해야 한다. 그러나 우선 해결해야 할 문제는 적합한 학습 기법을 사용하여 시행착오 없이 예측을 위한 최적의 윈도우

표 2 인기 검색어 '세월호'에 대한 두 가지 다른 이벤트
Table 2 Two different events related to hot search keyword, 'Sewol Ferry'

Collected Date	Extracted Contents
2014/04/16	세월호, 침몰, 청해진, 여객선, 진도, 참사, 실종, 승객, 단원고, 학생, 구조, 해경, 청해진해운, 전원, 침수
2014/11/07	세월호, 특별법, 세월호 특별법, 희생자, 국회, 여야, 유가족, 가족대책위, 통과, 유병언법, 세월호3법, 청해진해운, 진도, 기자회견, 피해자

크기를 선택하는 것이다.

본 연구는 네이버의 실제 실검 리스트의 순위 데이터를 분석하였다. 이 데이터를 분석해본 결과 때때로 검색어들이 실검 리스트에서 소멸하였고 이 소멸하는 길이가 특정 시간을 초과했을 때 다시 나타난 검색어가 소멸하기 이전의 검색어와 단어는 동일하지만 서로 다른 사건과 연관되어 있을 수 있음을 발견하였다. 예를 들어, 표 2는 두 가지 이벤트에 대한 동일한 검색어인 '세월호'를 분석한 것이다.

표는 수집된 날짜와 검색어에 대한 콘텐츠의 대표적인 키워드를 나타낸다. 2014년 세월호와 관련된 두 가지 이벤트가 있다. 첫째, 2014년 4월 16일 진도 해상에서 청해진해운이 운영하는 여객선 '세월호'가 침몰해 승객이 죽거나 실종된 사고가 발생하였다. 둘째, 2014년 11월 7일 여야가 10.31 세월호 3법 합의안에 동의하고 유가족 측도 수용함으로써 국회 본 회의에서 세월호 특별법이 일괄 통과되었다. 표는 서로 다른 날짜에 동일한 검색어 '세월호'로 수집했을 때 각 날짜에 포함되는 키워드가 서로 다르다는 것을 보여준다. 따라서 약 7개월 전후로 '세월호'라는 검색어가 두 개의 서로 다른 이벤트로 분리된다.

본 연구는 두 시점(검색어가 사라진 전후)에서 사건의 내용 간 유사성을 비교하여 서로 다른 사건의 내용을 갖는 검색어가 사라지는 최소 길이를 식별하려는 방법을 제안하였다. 앞서 언급했듯이, 검색어는 단어 또는 단문으로 구성되지만, 설명은 제공하지 않는다. 자세한 정보를 추출하지 않고 검색어의 정확한 의미를 파악하는 것은 거의 불가능하다. 따라서 본 연구는 각 인기 검색어에 대한 대표적인 키워드들을 추출하고 그 검색어가 사라진 시점과 다시 나타난 시점, 즉 이 두 시점의 대표 키워드들의 유사성을 비교하는 것을 최적의 윈도우 크기 계산 방법으로 제안한다. 제안된 방법은 다음 네 가지 절차를 따른다. 1) 1시간 전에 게시된 인기 검색어와 관련된 뉴스를 수집, 2) 불용어를 제거하여 관련 뉴스 기사를 전처리, 3) TF(단어 빈도, Term Frequency)를 사용하여 15가지 대표 단어를 추출하고, 4) 서로 다른 두 시점에서 특정 인기 검색어에 대한 사건 내용의 코사인 유사도를 계산한다.

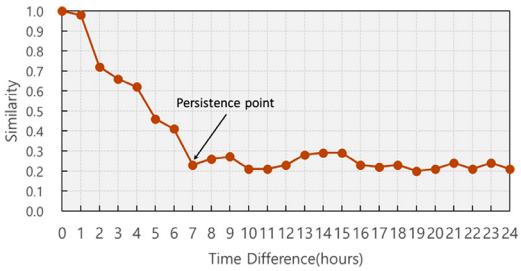


그림 3 검색어 소멸 시간 기반 콘텐츠 유사도 평균값
Fig. 3 The average of content similarity based on disappearance time of search keyword

그림 3은 실험에 사용된 전체 검색어가 실검 리스트에서 연속적으로 소멸한 기간을 기반으로 두 시점의 대표 단어들의 유사도 결과의 평균을 그래프로 보여준다. x축은 인기 검색어의 연속 소멸 기간을 나타내며, y축은 그 소멸 전후 두 시점에 대한 대표 단어의 평균 유사도를 보여준다(1은 완전히 동일, 0은 완전히 다름). 그림 3에서 볼 수 있듯이, 검색어가 7시간 이상 연속적으로 사라질 경우 두 시점의 사건의 내용 유사도는 0.2로 매우 낮았다. 검색어 소멸 시간의 길이가 7시간 이상이 될 경우 유사도는 더 낮아지지 않고 0.2 정도를 유지한다. 즉, 특정 인기 검색어 'A'가 7시간 이상 실검 리스트에서 나타나지 않고 그 후 다시 등장할 경우, 맨 처음 출현한 검색어 'A'와 재출현한 검색어 'A'는 서로 사건의 내용이 다르다고 할 수 있다. 따라서 최적의 윈도우 크기는 동일한 검색어들이 두 시점에서 사건의 내용이 달라지는 최단 소멸 기간으로 볼 수 있다. 그러므로 그림 3을 토대로 네이버의 인기 검색어에 대한 최적의 윈도우 크기는 7이어야 한다. 다른 검색엔진인 다음, 네이트의 윈도우 크기에 대한 평가는 5.1절(윈도우 크기 선택 조사)에서 수행되었다.

4. 실험 설정

이장에서는 인기 검색어의 순위 변화 예측을 평가하기 위해 수집된 데이터와 기계학습 기법을 설명한다. 알고리즘 1은 예측 알고리즘을 보여주며, 알고리즘을 평가하는 데 필요한 데이터와 여러 기계학습 기법을 이용하였다.

4.1 실험 데이터

제안된 모델의 성능 평가를 위해서 인기 검색어와 이와 관련된 뉴스 기사, 그리고 검색어에 대한 순위 패턴을 수집하였다. 또한, 네이버뿐만 아니라 다음, 네이트 이 세 가지 검색엔진의 API를 사용하여 2년(2012년 6

월 30일~2014년 6월 30일) 동안의 인기 검색어와 이와 관련된 뉴스 및 순위를 크롤링(crawling)하였다.

총 57,359개의 서로 다른 인기 검색어가 수집됐고 평가를 위한 2년간의 데이터를 훈련시켰다. 인기 검색어와 이와 관련된 기사를 포함한 자세한 데이터 수집 방법은 알고리즘 1에서 확인할 수 있다. 3절의 공식 2를 달성하기 위해서, 훈련 데이터는 피처(feature)로써 과거 순위 패턴과 클래스로써 미리 정의된 미래순위를 포함한다. 피처의 수는 최적 윈도우 크기에 기반을 두어 변화한다. 훈련 데이터를 사용하여 예측 모델을 구축하기 위해서 4가지의 기계학습 기법(Naive Bayes[2], Neural Networks[3], Support Vector Machines[4], Decision Trees[5])을 적용하였다. 각 기계학습 기법은 기계학습 도구인 Weka[17]에 기본 설정되어 있는 구조와 파라미터를 사용하였다.

알고리즘 1 인기검색어의 순위 변화 예측

1: 네이버에서 수집 시간 (h)에 대한 검색어(T)와 해당 순위(r)를 수집한다.
2: 뉴스 검색 API에 T를 입력하여, 한 시간 동안($h-1$~h) 게시된 T와 관련된 뉴스 ra를 얻는다.
3: 검색어 리스트에 T가 예전에 나타난 적이 있는지 확인한다.
  3.1: 예전에 나타난 적이 있다면, TF를 이용하여 수집된 검색어의 의미를 대표하는 단어를 ra에서 추출한다.
  3.2: 반면에, 나타난 적이 없다면, T는 새로 검색된 키워드이며 5단계로 건너뛴다.
4: 특정 시간($h-n+1$~h) (n=window size)에 T의 모든 과거 순위(PR)를 얻는다.
  PR을 기계학습으로 훈련된 모델에 입력 데이터로 사용한다.
5: 만일, T가 최초 검색되어 과거 순위가 없다면 현재 순위만을 PR로 이용한다.
6: 한 시간 이후에 T의 FRC가 up, down, unchanged 중 어떻게 변할지 예측한다.

5. 평가 결과

4.1절에서 설명한 2년간의 훈련 데이터에 대해서 10-fold cross validation을 수행하였다. 평가 결과를 바탕으로, 제안된 시계열 모델의 예측 결과와 서로 다른 누락 순위 처리 방법, 그리고 윈도우 크기를 비교하고 정리하였다.

5.1 윈도우 크기 선택 조사

본 논문에서는 인기 검색어 순위 변화의 예측을 위한 최적의 윈도우 크기를 선택하는 방법을 제안하였다. 최적의 윈도우 크기는 3.2절에서 다루었듯이 동일한 검색어가 서로 다른 두 시점에서 사건의 내용이 달라지는 최단 소멸 기간으로 계산할 수 있다. 이를 통해 네이버의 인기 검색어 데이터의 최적의 윈도우 크기는 7로 계산되었고, 다음과 네이트의 실검 리스트 순위 데이터에도 제안된 방법을 적용하였다.

그 결과, 표 3에서 보는 것과 같이 다음과 네이트의 최적 윈도우 크기는 각각 6과 8임을 발견하였다. 본 논문은 제안된 접근 방법이 서로 다른 데이터에 대해서 최적의 윈도우 크기를 선택하는지를 조사하기 위해 이러한 윈도우 크기로 예측 성능을 평가한다.

표 3 네이버, 다음, 네이트 검색 엔진에 대한 최적 윈도우 크기
Table 3 Optimal window sizes for search engines of 'Naver', 'Daum', and 'Nate'

	Naver	Daum	Nate
Optimal window size	7	6	8

5.2 예측 평가

본 논문의 실험 목적은 제안된 시계열 모델을 테스트하기 위해서 설계되었다. 실험은 4.1절에서 언급하였듯이 네 가지 기계학습 기법을 적용하여 모델의 성능 평가를 수행하였다. 실험 결과는 서로 다른 윈도우 크기와 다른 누락 순위 처리 기법이 적용되었다. 표 4는 네이버의 실검 리스트 순위 변화를 서로 다른 윈도우 크기(5,7,9)와 네 가지 다른 손실 값 처리기법인 Pairwise deletion, Dummy variable, Mean, EM을 적용하여 예측했고 분석된 결과를 보여준다.

3.2절에서도 언급했듯이 본 논문에서는 네이버 실검 리스트 데이터에 대한 최적 윈도우 크기는 7로 계산했으며, 실험 결과는 윈도우 크기가 5,7,9 중에서 7일 때 예측 성능이 가장 높았다는 것을 보여준다. 윈도우 크기 7과 9의 예측 성능은 거의 차이가 없으므로 9보다 7이 더 나을지 결정하기는 어렵다. 그러나 윈도우 크기 7과 9의 차이가 없을 때 7을 사용하면 데이터 크기와 속도를 포함한 성능 면에서 더 효과적일 수 있다고 추측할 수 있다.

누락된 순위 데이터 처리의 경우, EM에서 계산된 lowest+1 (즉 11위)로 누락 순위를 대체하는 것이 가장 높은 예측 정확도를 달성하였다. 이 이유로는 다른 세

2) http://weka.sourceforge.net/doc.dev/weka/classifiers/bayes/NaiveBayes.html
3) http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/MultilayerPerceptron.html
4) http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/SMO.html
5) http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html

표 4 손실 값 처리 방법과 윈도우 크기에 따른 실시간 검색어 변화 예측 정확도

Table 4 Rank-change prediction accuracies according to different missing rank handling approaches and window size

Window Size	Missing Value	NB	NN	SVM	DT
5	Dummy value	79.74%	88.22%	79.92%	88.76%
5	EM	80.13%	88.95%	80.84%	89.88%
5	Mean	75.14%	86.57%	77.32%	87.52%
5	Pairwise deletion	75.96%	85.47%	77.55%	85.78%
7	Dummy value	83.94%	93.59%	85.41%	93.09%
7	EM	83.04%	93.72%	86.07%	94.06%
7	Mean	80.29%	91.11%	83.26%	92.95%
7	Pairwise deletion	82.98%	92.83%	83.97%	90.13%
9	Dummy value	83.92%	92.56%	85.37%	93.05%
9	EM	83.03%	92.57%	85.66%	93.94%
9	Mean	80.41%	91.46%	83.36%	92.21%
9	Pairwise deletion	82.94%	90.96%	83.94%	90.14%

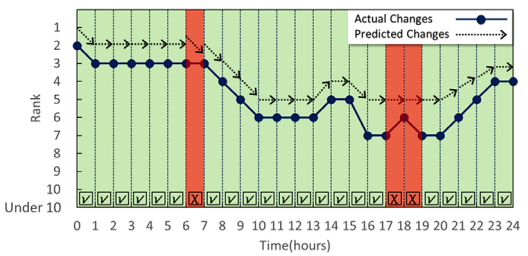


그림 4 검색어 '메르스'의 실검리스트 순위 변화 예측

Fig. 4 Future Rank Prediction of 'MERS' from the hot search keyword list

가지 접근법인 Pairwise deletion, Dummy variable, Mean 기법이 실검 리스트 순위의 본질을 고려하지 않았기 때문이다. 그리고 모델을 학습하는 네 가지 기계학습 기법 중에서는 C4.5 결정 트리 알고리즘이 다른 기법보다 우수한 성능을 보였다. 마지막으로, 제안된 모델의 용이성이 네이버 실검 리스트에만 국한되지 않는지 다른 두 검색엔진인 다음과 네이트에서도 성능을 분석하였다. 앞서 5.1절에서 언급했듯이, 다음과 네이트의 실검 리스트의 최적의 윈도우 크기는 각각 6과 8이었고, 다음은 윈도우 크기 6과 EM을 사용하여 92.54%로 그 도메인 내에서는 최고 성능을 달성했으며, 네이트는 윈도우 크기 8과 EM을 사용하여 80.13%의 정확도를 보여주었다.

그림 4는 2015년 5일 20일 대한민국 메르스 유입이 확인되었을 당시 검색어 '메르스'에 대한 24시간 실검 리스트 순위와 그 예측을 데모한 것이다. 그림에서 볼 수 있듯이, 모든 실시간 예측은 제안된 손실 값 처리 방법 (EM)과 윈도우 크기 처리 방법을 사용하고 과거 실검 리스트 순위 데이터만을 이용하여 높은 예측률을 달성하였다.

5.3 추가적인 피처

본 논문에서는 실검 리스트에서 제공하는 한정된 데이터만으로 학습하여 모델을 만들었다. 하지만 과거 순위 데이터 외에 다른 피처를 추가하여 훈련 데이터 (training data)에 적용하였고, 결과를 비교해 보았다. 추가적인 피처는 토픽 피처로서 검색어가 무슨 토픽 (연예, 정치, 스포츠 등)인지를 확인하는 것이다.

각 검색어의 토픽을 찾기 위해서 본 논문에서는 네이버 뉴스 분류 서비스를 사용하여 실검 리스트 검색어들을 분류하였다. 왜냐하면, 인기 검색어는 실시간 이벤트, 즉, 본 적 없는 새로운 단어일 가능성이 높으므로 기존 검색어 분류 온톨로지는 적용할 수 없기 때문이다. 인기 검색어의 토픽이 일반적인 단어 분류 온톨로지를 이용해 추출된다면, 의미론적으로 관련된 토픽으로 분류될 것이다. 예를 들어, 검색어가 '갤럭시'라면, 일반적인 의미는 '우주'이므로, 의미론적으로 관련된 토픽은 항상 '순수과학'일 것이다. 하지만 관련된 뉴스를 사용하여 분류하게 되면 검색어가 실검 리스트에 올라온 시간에 따라서 우주가 내포된 '순수과학'이 될 수도 있고, 삼성 갤럭시 폰을 뜻하는 '기술과학'이 될 수도 있다.

토픽 피처를 추출하기 위해서 다음과 같은 절차로 실험을 진행하였다. 우선, 네이버 뉴스 분류 서비스에서 실검 리스트에 올라온 인기 검색어를 검색한다. 검색할 때, 뉴스가 작성된 시간은 인기 검색어가 처음 등장한 때로 설정한다, 그 후, 그날에 해당 검색어로 게시된 관련 기사들이 나오면, 네이버 뉴스 카테고리 중 가장 많은 양의 기사가 포함된 토픽을 선택하여 분류한다. 즉, 특정 시간대에 '갤럭시'라는 검색어로 뉴스를 검색하여 가장 많은 토픽인 '기술과학'을 선택하여 분류하는 것이다. 표 5는 2년 동안에 네이버의 인기 검색어를 분류한 결과이다. 표에서 찾아볼 수 있듯이, 실검 리스트의 검색어들은 약 80%가 연예, 스포츠, 정치 토픽과 관련된 것으로 보인다. 이 토픽 피처를 훈련 데이터 세트에 추가한 후, 제안된 손실 값 처리 EM과 윈도우 크기 7로

표 5 실검리스트의 검색어 주제 분포

Table 5 Topic distribution of hot search keywords

Topic	Entertainment	Sports	Politics	Fashion	World	Obituaries	Health	Technology
%	42%	28%	10%	6%	5%	4%	3%	2%

C4.5 결정 트리 방법을 사용하여 학습시켰고, 거기서 나온 예측 정확도는 94.85%로 과거 순위 패턴만 사용했을 경우(94.06%)와 비교하여 다소 높았다.

6. 토 의

본 논문에서는 인기 검색어 순위 예측에 대한 문제를 다뤘다. 네이버 인기 검색어 서비스는 단지 인기 검색어와 각 인기 검색어의 순위만을 제공한다. 따라서 사람들은 데이터만을 사용하는 예측 모델이 실제적인 예측 결과를 제시할 수 있는지에 대해 의문을 가질 수 있다. 본 논문에서는 윈도우 크기와 손실 값 처리를 고려하여 과거 데이터를 사용하는 간단한 순위 예측 방법을 제안한다. 놀랍게도, 제안한 방법은 C4.5 결정 트리에서 94.06%로 매우 우수한 성능을 달성하였다. 한편으로, 이는 변화하는 추세가 순위 예측에 대해서 매우 중요한 요소라는 것을 암시한다. 반면에, 이는 순위 예측의 성능을 향상하는 것이 가능할 수 있다는 것을 의미한다. 예를 들어 피처(토픽 피처)를 추가했지만, 순위를 완벽(100%) 하게 예측하는 것은 어려울 것으로 보이며, 그 이유는 알고리즘적인 요인뿐만 아니라 인기 검색어 순위, 즉, 사람들의 관심도가 불규칙하게 변화하는 자연적인 혹은 사회적인 현상 때문이다.

7. 결 론

본 논문에서는 시간별 검색어의 순위 변화를 예측하기 위한 시계열 모델링 프레임워크를 제안하였다. 개발된 프레임워크는 인기 검색어의 과거 순위 패턴에 기반을 두어 인기 검색어 순위 예측을 모델링 하는 것에 중점을 뒀다. 또한, 과거 순위 패턴만을 사용하여 예측하기 위해서 누락 순위 처리와 최적의 윈도우 크기를 계산할 수 있는 새로운 접근법을 제안하였다. 이러한 방법들을 사용하여, 복잡한 특징을 사용하기보다는 검색어에서 제공되는 과거 순위 패턴을 기계학습 기법을 적용하여 높은 예측률(94.06%)을 달성하였고, 우수한 검색어 순위 예측 서비스를 제공하는 것이 가능하다는 것을 증명하였다.

References

[1] H. Kwak, et al., "What is Twitter, a social network or a news media?," *Proc. of the 19th international conference on World wide web. ACM*, pp. 591-600, Apr. 2010.

[2] J. E. Chung and E. Mustafaraj, "Can collective sentiment expressed on twitter predict political elections?," *AAAI*, Vol. 11, Apr. 2011.

[3] F. Franch, "(Wisdom of the Crowds) 2: 2010 UK election prediction with social media," *Journal of Information Technology & Politics*, Vol. 10, No. 1, pp. 57-71, 2013.

[4] T. O. Sprenger, "TweetTrader. net: Leveraging Crowd Wisdom in a Stock Microblogging Forum," *ICWSM*, pp. 663-664, May, 2011.

[5] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of computational science*, Vol. 2, No. 1, pp. 1-8, 2011.

[6] J. Si, et al., "Exploiting Social Relations and Sentiment for Stock Prediction," *EMNLP*, Vol. 14, pp. 1139-1145, Oct. 2014.

[7] M. Naaman, H. Becker, and L. Gravan, "Hip and trendy: Characterizing emerging trends on Twitter," *Journal of the Association for Information Science and Technology*, Vol. 62, No. 5, pp. 902-918, 2011.

[8] H. Becker, M. Naaman, and L. Gravano, "Beyond Trending Topics: Real-World Event Identification on Twitter," *ICWSM*, Vol. 11, pp. 438-441, 2011.

[9] S. C. Han. and H. Chung, "Social issue gives you an opportunity: discovering the personalised relevance of social issues," *Pacific Rim Knowledge Acquisition Workshop, Springer Berlin Heidelberg*, pp. 272-284, Sep. 2012.

[10] S. C. Han, et al., "Chinese trending search terms popularity rank prediction," *Information Technology and Management*, Vol. 17, No. 2, pp. 133-139, 2016.

[11] K. Jaidka, et al., "SocialStories: Segmenting Stories within Trending Twitter Topics," *Proc. of the 3rd IKDD Conference on Data Science, 2016*, ACM, p. 1, Mar. 2016.

[12] A. Vakali, N. Kitmeridis, and M. Panourgia, "A distributed framework for early trending topics detection on big social networks data threads," *INNS Conference on Big Data. Springer International Publishing*, pp. 186-194, Oct. 2016.

[13] K. Lee, et al., "Twitter trending topic classification," *2011 11th IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 251-258, Dec. 2011.

[14] D. Kim, et al., "Predicting the Scale of Trending Topic Diffusion Among Online Communities," *Pacific Rim Knowledge Acquisition Workshop,* Springer International Publishing, pp. 153-165, Aug. 2016.

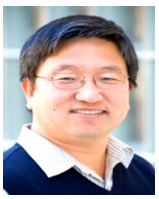
[15] S. Nikolov, and D. Shah, "A nonparametric method for early detection of trending topics," *Proc. of the Interdisciplinary Workshop on Information and Decision in Social Networks (WIDS 2012)*, MIT, Nov. 2012.

[16] P. D. Allison, *Missing data*, Thousand Oaks, CA: Sage, 2000.

[17] I. H. Witten, et al., *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, 2016.

김 도 형

2009년 광운대학교 컴퓨터공학과(학사)
2011년 광운대학교 컴퓨터공학과(석사)
2012년~현재 경희대학교 컴퓨터공학과 박사과정. 관심분야는 유비쿼터스 컴퓨팅, 전문가시스템, 지식관리, 인공지능



강 병 호

1988년 부산대학교 수학과(학사). 1990년 University of Tasmania, Computer Science(석사). 1995년 University of New South Wales, Computer Science(박사) 1996년~1999년 호서대학교 조교수. 2008년~현재 University of Tasmania, School of Information and Communication Technology, professor 관심분야는 인공지능, 전문가시스템, 지식관리, 웹기반서비스, 지능형 시스템, 지식기반시스템

이 승 룡

정보과학회논문지
제 44 권 제 4 호 참조