https://doi.org/10.14400/JDC.2017.15.12.333

시계열분석과 인공신경망을 이용한 실시간검색어 변화 예측

정민영
광주여자대학교 실버케어학과

Predicting changes of realtime search words using time series analysis and artificial neural networks

Min-Yeong Chong
Dept. of Silvercare, Kwangju Women's University

요 약 실시간검색어는 지금 바로 이슈가 되는 검색어의 검색 증가율이 단기간에 급상승하는 것을 중심으로 하기 때문에 일정기간 지속적으로 관심도를 유지하고 있는 이슈를 나타내지 못하고 이들이 가까운 미래에 어떤 변화를 보이는지에 대한 것도 알 수 없는 한계를 가지고 있다. 본 논문에서는 이러한 한계를 극복할 수 있도록 일정기간 동안 상위 10위 안에 속한 적이 있는 실시간검색어에 대해 일자별, 시간별 지속성을 평가하여 꾸준히 관심을 받는 검색어를 추출한다. 그런 다음, 이들 중 상위에 속하는 검색어의 관심도가 어떻게 변화하는지를 알 수 있게 하는 시계열 분석과 신경망을 이용하는 방법을 제시하고 이를 통해 도출한 실제 예를 통해 가까운 미래의 변화량을 예측한 결과를 보인다. 일자별로는 시계열 분석을, 시간별로는 인공신경망의 학습을 통해 예측하는 것이 좋은 결과를 보인다는 것을 알 수 있다.

주제어 : 실시간검색어, 빅데이터 분석, 시계열 분석, 인공신경망, 웹 마이닝, 텍스트 마이닝

Abstract Since realtime search words are centered on the fact that the search growth rate of an issue is rapidly increasing in a short period of time, it is not possible to express an issue that maintains interest for a certain period of time. In order to overcome these limitations, this paper evaluates the daily and hourly persistence of the realtime words that belong to the top 10 for a certain period of time and extracts the search word that are constantly interested. Then, we present the method of using the time series analysis and the neural network to know how the interest of the upper search word changes, and show the result of forecasting the near future change through the actual example derived through the method. It can be seen that forecasting through time series analysis by date and artificial neural networks learning by time shows good results.

Key Words : Realtime search word, Bigdata analysis, Time series analysis, Artificial neural networks, Web mining, Text mining

1. 서론

최근 들어 사물인터넷을 중심으로 하는 연결이 확대되고, 모바일컴퓨팅과 클라우드컴퓨팅의 활용범위가 넓

* 본 논문은 2017 학년도 광주여자대학교 교내연구비 지원에 의하여 연구되었음(KWUI17-021).

Received 2 November 2017, Revised 30 November 2017
Accepted 20 December 2017, Published 28 December 2017
Corresponding Author: Min-Yeong Chong
(Kwangju Women's University)
Email: mychong@kwu.ac.kr

ISSN: 1738-1916

어짐에 따라 지속적으로 발생하고 있는 빅데이터의 분석과 활용에 대한 관심이 높아지고 있다[1,2]. 또한 '4차 산업혁명'을 촉발시킨 핵심에 해당되는 지능정보기술을 진화시키는 인공지능 소프트웨어 개발과 불가분의 관계를 갖고 있는 빅데이터 관련 기술은 다양한 분야에서 그 활용도를 높이고 있다[3,4].

특히 고수준의 정보검색 서비스를 중심으로 사람들을 모이게 하고 이들이 궁금해 하는 검색어를 수집하여 보관하고 이를 집계하고 분석하여 제공하는 실시간검색어는 포털 사이트에 따라 실시간급상승검색어(Naver), 실시간이슈검색어(Daum), 인기급상승검색어(Google) 등으로 불리며[5,6], 검색증가율이 높은 순으로 상위 10개를 제시하여 사용자의 관심도를 반영한 인터넷검색의 출발을 지원해주는 중요한 서비스로 포털 사이트의 간판으로서 역할하고 있다[7,8]. 사용자 관심도가 높은 검색어를 통해 현안에 대한 신속한 접근이 가능하게 함으로써 검색서비스 만족도를 높여서 포털 사이트의 선호도를 높일 뿐 아니라 수집된 검색어의 빅데이터 분석을 통한 다양한 분야의 서비스에 활용할 수 있는 데이터의 원천으로서 역할을 하고 있다[9,10].

하지만, 주로 사용자의 관심도가 일시적으로 급상승하는 검색어를 제공하는 서비스를 중심으로 하기 때문에 일정기간 동안 지속적으로 관심을 받는 검색어에 관한 것은 알 수 없을 뿐 아니라 지속성이 높은 검색어가 가까운 미래에 어떻게 변화할지에 대한 흐름을 파악하는 정보를 제공하기는 힘들다[11,12]. 이러한 서비스와 유사하게 현재까지의 일정기간 동안 변화를 알려주는 것으로 구글 트랜드가 있지만 이 서비스는 특정 검색어의 최대 검색량에 대한 상대적 지표만을 알려주므로 다른 검색어들과의 비교할 수 있는 질적인 차이를 파악하기 힘들다. 또한 과거 특정 시점에서 현재 시점까지의 변화만 알려주고 다가오는 시점에 대한 정보는 알려주지 못하는 한계가 존재한다[13,14].

따라서 본 논문에서는 이러한 한계를 극복할 수 있도록 일정기간 실시간검색어 상위 10에 속한 적이 있는 것에 대한 검색 지속성을 일자별, 시간별로 평가하여 일자별 지속성이 큰 상위 10개와 시간별 지속성이 큰 상위 10개를 구하고, 이를 바탕으로 일정기간 동안 변화량을 분석하여 가까운 미래의 변화를 예측하는 방법과 이 방법을 적용한 실제적인 사례를 제시하고자 한다.

이를 위한 검색 지속성을 평가할 수 있는 것으로는 각 검색어별로 일자(date) 단위 변화량을 집계한 것을 기초로 검색어별 출현일수를 구하여 추출하는 이른바, '출현일자 상위 10 검색어'와 각 검색어와 일자별로 시간(hour) 단위 변화량을 집계한 것을 기초로 검색어별 출현일수별 출현시간을 구하여 추출하는 '출현시간 상위 10 검색어'가 필요하다. '출현일자 상위 10 검색어'는 검색 지속성에 대한 일자별 변화를 파악하여 그 추세를 일자별 시계열 분석을 통해 현재 이후의 2일 동안 변화 추이를 예측하는데 활용한다. '출현시간 상위 10 검색어'는 검색 지속성에 대한 시간별 변화를 파악하여 이를 인공신경망을 통해 학습시킨 다음 현재 이후의 12시간 동안 변화 추이를 예측하는데 활용한다.

본 논문에서 이루어지는 데이터 수집과 저장, 집계 및 정렬, 그리고 이를 기반으로 하는 분석 및 예측은 R언어를 통해 수행한다[15].

2. 실시간검색어 수집 및 그룹별 집계

2.1 실시간검색어 수집 대상 및 방법

실시간검색어는 네이버(Naver)의 실시간급상승검색어와 다음(Daum)의 실시간이슈검색어를 대상으로 수집한다. 두 포털 사이트 모두 검색 요청된 검색어의 증가비율이 가장 높은 것부터 내림차순으로 상위 10개씩 보여주는 서비스로, 이전 시점에 비해 상대적으로 증가 비율이 급격하게 상승한 것을 기준으로 한다. 실시간검색어는 시간 제약을 갖는 서비스 특성상 1분이라는 시간 간격으로 검색 순위가 집계되어 게시되므로 한 시간에도 수십 차례 새롭게 변경될 수 있는 특성을 가지고 있다. 따라서 실시간검색어는 일정기간 단순 검색횟수 만으로 집계하는 검색어와는 다른 성격을 가지며, 오히려 검색횟수 증가율로 비교하므로 일시적으로 사용자의 관심도가 증가하는 것 중심으로 정해지는 한계가 있다. 또한 그 자체만으로 검색어의 검색 지속성을 파악하여 사용자의 지속적인 관심을 받고 있는 검색어를 파악하거나 그 검색어의 관심이 앞으로 어떻게 변화할 것인지를 예측하기는 힘들다.

이러한 실시간검색어의 일시성의 한계를 극복할 수 있는 방안은 두 포털 사이트의 홈페이지 원시코드에서

수집되는 실시간검색어를 검색어 순위에 따라 일정간격으로 실시간검색어에 대한 점수를 부여하여 저장하는 방식으로 데이터를 수집하여, 특정 시점을 기준으로 일정기간 동안의 두 포털 사이트의 자료를 병합한 다음, 이를 대상으로 각 검색어별로 일자(date) 단위 변화량과 시간(hour) 단위 변화량을 추출할 수 있게 함으로써 검색 지속성을 평가할 수 있는 기초를 마련하는 것이다.

본 논문에서 핵심이 되는 시간 단위 변화량은 근본적으로 검색 지속성을 평가하는 가장 기초적인 역할을 하며 이를 근간으로 시간별, 오전오후별, 일자별, 주간별, 월별, 년별로 변화량을 분석할 수 있다는 것을 의미한다.

실시간검색어 자료는 2016. 11.11일 0시 0분에서 2016. 11.22일 23시 59분까지 1분 간격으로 수집한 것이다.

2.2 실시간검색어 그룹별 집계

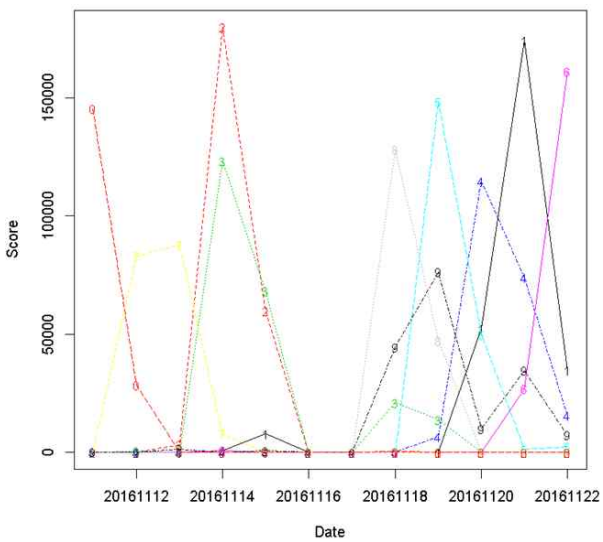
현재 포털 사이트에서 검색횟수 순간증가율을 근거로 제공하는 실시간검색어의 일시성을 개선하기 위해, 데이터 수집기간 동안 상위 10 검색어에 대해 순위별로 점수를 부여하여 일단 저장해둔다. 그리고 이것을 읽어서 검색어를 그룹으로 하여 점수를 집계하고 그 점수합계가 큰 것부터 순서대로 정렬해서 상위 10개를 추출해낸 것이 '관심도 상위 10 검색어'이며 '손연재', '슈퍼문', '추미애', '김연아', '천호식품', '불야성', '김제동', '박근혜계엄령', '장시호', '길가에버려지다' 순으로 나타났다.

[Fig. 1]은 '관심도 상위 10 검색어'에 대한 일자별 점수 변화를 나타낸 것으로 해당 기간 동안 관심도의 흐름을 파악할 수 있게 해준다. 최상위에 속하는 '손연재'의 경우, 3일간의 집중적인 관심도에 의해 전체 기간의 최대 관심도를 갖는 검색어로 결정된 것으로 볼 때, '관심도 상위 10 검색어'는 1~3일 정도의 비교적 짧은 기간 동안 관심도가 높은 검색어를 선정하는 용도로 활용하는 것이 바람직하고, 지속성이 높은 검색어를 추출하는 것에는 분명한 한계가 존재한다는 것을 확실히 보여주고 있다.

본 논문에서는 이를 개선하여 실시간검색어의 지속성을 추출하기 위한 방법으로 제시한 것이 일자 단위 변화량을 기준으로 집계하는 방법과 시간 단위 변화량을 기준으로 집계하는 방법이다.

먼저 일자 단위 변화량을 기준으로 집계하기 위해서는 검색어와 일자별로 group_by() 함수에 의해 그룹핑하고 summarise() 함수에 의해 검색어와 일자 그룹별 건수, 점수 합계, 점수 평균을 구한 다음, 그 결과에 대해 다시 summarise() 함수를 적용하여 검색어 그룹별 건수, 일자별 건수의 합계, 점수 합계, 점수 평균을 구한다. 여기서 검색어 그룹별 건수는 검색어별 일자에 대한 건수이므로 검색어별 출현일수가 된다.

다음으로 시간 단위 변화량을 기준으로 집계하기 위해서는 일단 검색어와 일자와 시간별로 group_by() 함수에 의해 그룹핑하고 summarise() 함수에 의해 검색어와 일자와 시간 그룹별 건수, 점수 합계, 점수 평균을 구한 다음, 그 결과에 대해 다시 summarise() 함수를 적용하여 검색어와 일자 그룹별 건수, 일자별 건수의 합계, 점수 합계, 점수 평균을 구한다. 여기서 검색어와 일자 그룹별 건수는 검색어와 일자와 시간에 대한 건수이므로 검색어와 일자별 출현시수가 된다. 그리고 그 결과에 대해 한 번 더 summarise() 함수를 적용하여 검색어 그룹별 건수, 일자별 출현시수의 합계, 일자별 건수의 합계, 점수 합계, 점수 평균을 구한다. 여기서도 검색어 그룹별 건수는 검색어별 일자에 대한 건수이므로 검색어별 출현일수가 된다. 따라서 최종적으로 검색어별, 출현일자별, 출현시수별, 출현시간 건수별, 점수합계별, 점수평균별 분류가 가능하다.



[Fig. 1] Change in score of interest top 10 search words

3. 실시간검색어 지속성 평가 및 선정

3.1 일자별 변화량 기반 지속 검색어 선정

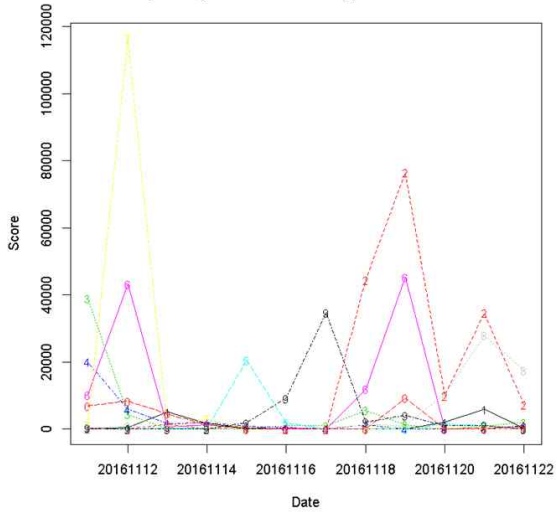
'출현일자 상위 10 검색어'는 검색어가 얼마나 사용자의 관심을 지속적으로 받았는가를 알려주는 출현일자가 많은 검색어를 뜻하며, 구체적으로는 각 검색어별로 일자(date) 단위 변화량을 기초로 검색어별 일수에 해당하는 출현일수가 가장 많은 상위 10개의 검색어에 해당된다. 일단 각 검색어별로 일자 단위 변화량을 구하고 이를 기준으로 검색어별 출현일수를 집계한 다음 이를 arrange() 함수에 의해 출현일수 내림차순으로 정렬하고 head()함수에 의해 상위 10개를 추출하는 과정을 통해 구할 수 있으며 실제로 '박근혜', '장시호', '정유라', '이정현', '문재인', '팬텀싱어', '박진주', '이재명', '역도요정 김복주', '이번주아내가바람을핍니다' 순으로 나타났다.

[Fig. 2]는 '출현일자 기준 상위 10 검색어'에 대한 점수 변화를 나타낸 것으로. 최상위인 '박근혜'의 경우, 전체 기간(12일간)중 10일 동안의 지속적인 관심도에 의해 최대 관심도를 갖는 것으로 결정되었다. 이는 최상위지만 누락일(2일간)이 존재할 수 있다는 것과 시간 단위의 보다 세부적인 변화의 흐름은 파악할 수 없다는 한계를 보여주면서도 거의 매일 꾸준히 주목받고 있는 검색어로서는 의미가 있다는 것을 보여준다. 그러나 [Fig. 1]의 '관심도 상위 10 검색어'에는 속하지 않는 것으로 보아 검색 순간증가율은 그다지 높지 않은 것으로 나타났다.

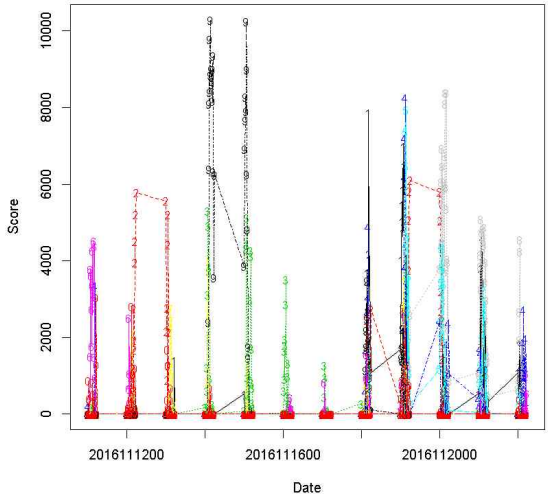
3.2 시간별 변화량 기반 지속 검색어 선정

'출현시간 상위 10 검색어'는 검색어가 얼마나 사용자의 관심을 지속적으로 받았는가를 알려주는 출현시수가 많은 검색어를 뜻하며, 구체적으로는 각 검색어와 일자별로 시간(hour) 단위 변화량을 기초로 검색어와 일수별 시수에 해당하는 출현시수가 가장 많은 상위 10개의 검색어에 해당된다. 일단 각 검색어와 일자별로 시간 단위의 변화량을 구하고 이를 기준으로 검색어와 일수별 시수를 집계한 다음 이를 arrange() 함수에 의해 출현시수와 출현일수의 내림차순으로 정렬하고 head()함수에 의해 상위 10개를 추출하는 과정을 통해 구할 수 있으며 실제로 '장시호', '그것이알고싶다', '수능', '김기춘', '박태환', '김연아', '미세먼지', '정유라', '추미애', '이번주아내가바람을핍니다' 순으로 나타났다.

[Fig. 3]은 '출현시간 상위 10 검색어'에 대한 점수 변화를 나타낸 것으로, 최상위인 '장시호'의 경우, 전체 기간(288시간)중 78시간 동안의 지속적인 관심도에 의해 최대 관심도를 갖는 것으로 결정되었다. 이는 최상위지만 누락시간(210시간)이 존재할 수 있다는 한계를 보여주면서도, 취침 등 주요 활동시간 이외의 시간을 고려했을 때 출현시간 만큼 시간별로 꾸준히 주목받고 있는 검색어로서는 의미가 있다는 것을 보여준다. 또한 [Fig. 1]의 '관심도 상위 10 검색어'에서는 9위로 어느 정도 검색 순간증가율이 높고, [Fig. 2]의 '출현일자 상위 10 검색어'에서는 2위로 일자별 검색 지속성이 상당히 높았다.



[Fig. 2] Change in score of top 10 search words by appearance days



[Fig. 3] Change in score of top 10 search words by appearance hours

4. 실시간검색어 변화 분석 및 예측

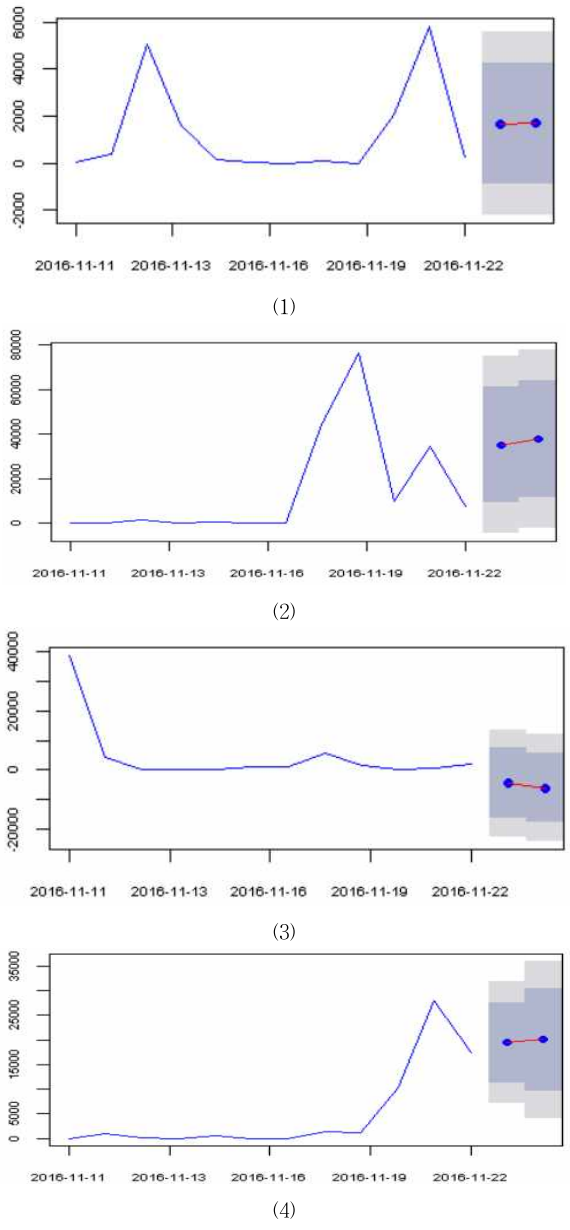
4.1 실시간검색어 시계열 분석에 의한 예측

시계열 분석은 시간에 따른 변화량을 계열화한 시계열 데이터를 시간의 흐름에 따라 변화하는 함수로 표현하고 이를 통해 흐름을 분석하는 것으로, 과거에서 현재까지의 변화량을 분석함으로써 가까운 미래에 대한 예측하는데 활용되고 있다[16]. 대표적인 시계열 분석 방법에는 회귀법, 이동평균법, 지수평활법, 요소분할법 등이 있다. 이 중에서 지수평활법은 단기간에 발생하는 불규칙 데이터를 평활화 하되 최근의 시계열 자료에 더 가중치를 두어 예측하는 방법이므로 가까운 미래를 예측하는데 많이 사용한다[17]. 본 논문에서는 지수평활법에 의한 시계열 분석을 통해 가까운 미래를 예측하고자 한다.

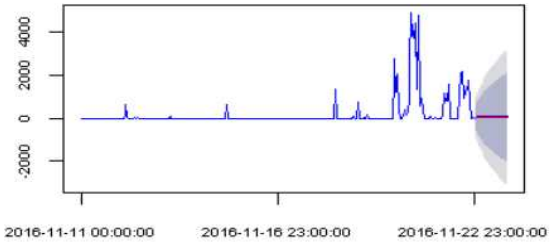
R언어의 forecast 패키지에 있는 ets 함수는 지수평활법을 수행하므로 이를 통해 시계열 데이터를 만들고 forecast() 함수를 이용하여 2일간의 미래를 예측한다. R의 기본적인 시계열 데이터는 년과 월을 기준으로 하기 때문에 일자와 시간까지 그 기준을 확대시켜야 하므로 이를 가능하게 하는 xts 패키지의 xts 함수를 사용한다.

[Fig. 4]는 일자 단위 변화량 기반의 '출현일자 상위 10 검색어'에 대한 일자별 시계열 분석을 하기 위해서 먼저 2016년 11월 11일~2016년 11월 22일 까지 12일간 일자에 따른 각 검색어의 변화량을 xts 함수에 의해 확장된 시계열 데이터로 만들고 이를 지수평활법을 수행하는 ets 함수에 적용하여 12일 이후 2일간 예측 결과까지 포함하는 시계열 분석 결과를 나타낸 것이다. 시계열 분석에 의한 일자별 예측 실험 결과, '(1)박근혜'는 미세한 상승, '(2)장시호'는 높은 상승, '(3)정유라'는 하향, '(4)이재명'은 약간 상승의 추세를 보인 것으로 예측되었다.

[Fig. 5]는 시간 단위 변화량 기반의 '출현시간 상위 10 검색어'에 대한 시간별 시계열 분석을 하기 위해서 먼저 2016년 11월 11일 0시~2016년 11월 22일 23시 까지 288시간 동안 시간에 따른 각 검색어의 변화량을 xts 함수에 의해 확장된 시계열 데이터를 만들고 이를 지수평활법을 수행하는 ets 함수에 적용하여 48시간 예측 결과까지 포함하는 시계열 분석 결과이며, 실시간검색어 '이재명'에 대한 것을 나타낸 것이다. 이러한 48시간 예측결과는 다른 모든 '출현시간 상위 10 검색어'에 대한 결과와 같이 아무런 변화가 없는 것으로 나타났다.



[Fig. 4] Time series analysis by day



[Fig. 5] Time series analysis by hour

4.2 실시간검색어 인공신경망에 의한 예측

본 논문에서는 '출현시간 상위 10 검색어' 중 특정 검색어에 대한 일자와 시간별 시간 단위 점수 변화량를 집계하고 이를 기초로 [Fig. 5]와 같은 시계열분석의 결과를 개선하여 의미 있는 예측결과를 도출하기 위하여, 학습을 통해 생성되는 인공신경망 모델을 적용하여 현재 이후의 12시간 동안 변화 추이를 예측한다.

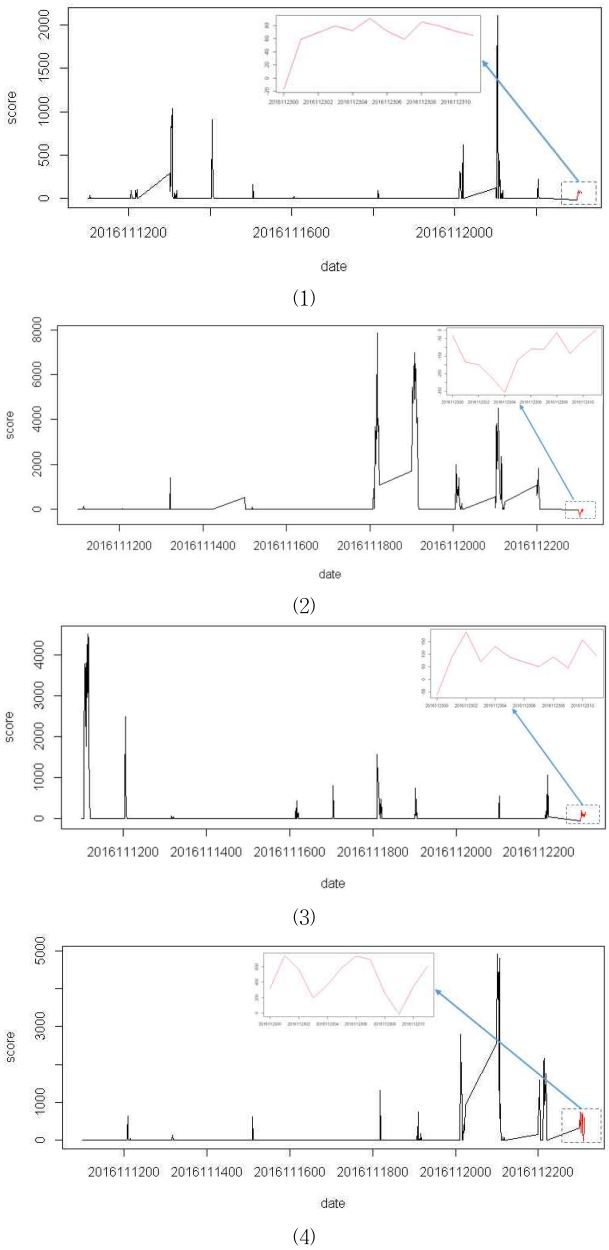
입력층과 출력층, 그리고 1 단계의 은닉층을 갖는 다층 인공신경망의 지도학습 방식을 사용하여 최적의 인공신경망 모델을 선정하는 최적 모델 선정 과정과 선정된 모델을 사용하여 가까운 미래를 예측하는 미래 예측 과정을 수행한다.

최적 모델 선정 과정은 인공신경망구조 정의, 학습데이터 준비, 테스트, 최적모형 선정 등의 단계로 이루어진다. 인공신경망 구조 정의 단계에서는 입력층, 은닉층, 출력층의 각 노드의 수를 결정하게 되는데 일반적으로 입력층의 노드 수가 k라면, k/2, k, 2k, 2k+1개의 은닉층 노드수를 갖는다[18]. 학습데이터 준비 단계에서는 최적모형 선정을 위한 학습 입력 데이터, 학습 목표 데이터, 예측 입력 데이터, 예측 비교 실제 데이터 등을 준비한다. 테스트 단계에서는 정의된 인공신경망구조를 토대로 학습 입력 데이터와 학습 목표 데이터를 적용하여 학습을 반복적으로 수행함으로써 학습된 인공신경망 모델을 생성한다. 그리고 예측 입력 데이터를 새롭게 생성된 인공신경망 모델에 적용하여 예측 결과를 추출하고 이것이 예측 비교 실제 데이터와 일치하는지를 평가하여 오차를 구한다. 최적모형 선정 단계에서는 여러 가지 형태로 정의된 인공신경망구조에 따라 시험한 결과로 나온 학습된 인공신경망 모델 중에서 오차가 작은 것을 골라 최적의 모형으로 선정한다.

본 논문에서는 <Table 1>과 같이 입력층의 노드수에 따라 다양한 은닉층, 출력층의 노드수를 정의하여 최적의 인공신경망 모델을 선정하고자 테스트하였고 그 결과 평균오차가 비교적 작고 12시간을 예측할 수 있는 모델 6을 최적의 모델로 선정하였다. 은닉층이 1개 층인 경우에 사용할 수 있는 R언어의 nnet 패키지에 있는 nnet()함수[19]를 사용하여 인공신경망 학습을 수행하고 최적의 인공신경망 모델에 의해 12시간까지의 미래를 예측한다.

<Table 1> Artificial neural network model by node number of input, hidden, output nodes

Model No	input node(k)	hidden node	output node(k/2)	iteration	errors
1	16	32(2k)	8	500	248,269,314,294,228
2	16	33(2k+1)	8	500	359,265,354,248,293
3	18	18(k)	9	500	424,397,441,462,438
4	20	10(k/2)	10	500	340,422,383,353,303
5	20	20(k)	10	500	313,337,287,331,317
6	24	12(k/2)	12	500	201,240,202,233,231



(1) (2) (3) (4)

[Fig. 6] Forecasting with artificial neural networks by hour data

미래 예측 과정은 예측을 위한 학습데이터 준비, 테스트, 예측 단계로 이루어진다. 학습데이터 준비 단계에서는 예측 결과를 추출하기 위한 학습 입력 데이터, 학습 목표 데이터, 예측 입력 데이터를 준비한다. 테스트 단계에서는 최적 모델로 선정된 신경망 모델 6에 적용된 인공신경망구조의 정의를 토대로 학습 입력 데이터와 학습 목표 데이터를 적용하여 학습을 반복적으로 수행함으로써 학습된 인공신경망 모델을 다시 새롭게 생성한다. 예측 단계에서는 예측 입력 데이터를 다시 새롭게 생성된 인공신경망 모델에 적용하여 12시간의 예측 결과를 추출한다.

[Fig. 6]은 '출현일자 상위 10 검색어' 중에서 최상위에 해당되는 '(1)박근혜', '출현일자 상위 10 검색어' 중에서 2위이고 '출현시간 상위 10 검색어' 중에서 최상위에 속하며 '관심도 상위 10 검색어'에서는 9위를 차지하는 '(2)장시호', '출현일자 상위 10 검색어' 중에서 3위이고 '출현시간 상위 10 검색어' 중에서 6위에 속하는 '(3)정유라', '출현일자 상위 10 검색어' 중에서 8위에 속하는 '(4)이재명' 등에 대해 미래 예측 과정을 통해 추출한 12시간 예측 결과를 나타낸 것이다.

5. 결론

본 논문에서는 실시간검색어가 단기간 급상승 검색 증가율을 기준으로 상위 10개를 선정하는 미시적 관점 때문에 일정기간 꾸준히 관심도를 유지하는 검색 지속성과 검색어가 가까운 미래에 어떤 변화를 보이는지에 대한 검색 방향성을 알 수 없는 한계를 극복할 수 있도록 하기 위하여 일정기간 실시간검색어의 검색 지속성을 일자별, 시간별로 변화량 중심으로 평가하여 일자 단위 변화량 기반의 '출현일자 상위 10 검색어'에 대한 일자별 시계열 분석과 시간 단위 변화량 기반의 '출현시간 상위 10 검색어'에 대한 인공신경망 학습에 의한 변화 예측을 하는 방법을 제시하였다.

시계열 분석은 최근의 자료에 중점을 두고 가까운 미래를 예측하는 지수평활법을 적용하였고, 양과 음의 양쪽 방향 변화가능성을 나타내는 실례와 함께 2일간의 변화가능성을 예측한 결과를 보여주었다. 그리고 인공신경망은 입력층, 은닉층, 출력층을 갖는 다층신경망으로 구성하되 각 노드의 수를 변화시켜 정의한 인공신경망구조를 적용한 결과로 비교적 적은 오차를 갖는 것을 최적의 모델로 선택하고, 여기에 예측을 위한 데이터를 사용하여 반복 학습을 시켜서 새로운 인공신경망 모델을 만들고, 미리 준비한 예측 데이터를 적용하여 12시간의 예측치를 추출하여 그 결과를 보였다. 이는 실시간검색어에 내재된 미시적 관점을 확대하여 보다 거시적 관점에서 검색 지속성을 평가할 수 있는 기본틀을 제공하여 향후 누적되어가는 데이터의 크기만큼 기간별 범위를 넓혀서 유의미한 분석결과를 낳을 수 있는 가능성을 제공했다는 측면에서 의의가 있다.

그러나 시간 단위 변화량 기반의 검색어에 대한 시계열 분석과 일자 단위의 변화량 기반의 인공신경망 학습에 의한 변화 예측은 비교적 연속성이 부족한 실시간검색어의 특성과 수집된 실험 데이터 크기의 한계 때문에 유의미한 결과를 갖지 못했다. 이를 위해서 데이터 수집을 보다 안정적으로 장기간 할 수 있는 틀을 갖추고 인공신경망의 은닉층을 확대하여 딥러닝[20]을 할 수 있는 심층신경망을 구성하는 방법과 SNS 분석[21]에 대한 추가적인 연구가 필요하다.

ACKNOWLEDGMENTS

This work was supported by Research Funds of Kwangju Women's University in 2017(KWUI17-021).

REFERENCES

[1] Min Chen, Shiwen Mao, and Yunhao Liu, "Big Data: A Survey", Mobile Netw Appl, Vol. 19, pp. 171-209, 2014.

[2] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan, "The rise of big data on cloud computing:Review and open research issues", Information Systems, Vol. 47, pp. 98-115, 2015.

[3] Su-Hyeon Namn, "Knowledge Creation Structure of Big Data Research Domain", Journal of Digital Convergence, Vol. 13, No. 9, pp. 129-136, 2015.

[4] Shinkon Kim, Sukjun Lee, and JeonggonA Kim, "Study on the Development of Phased Big Data Distribution Model Based on Big Data Distribution Ecology", Journal of Digital Convergence, Vol. 14, No. 5, pp. 95-106, 2016.

[5] Naver Search Help, "Realtime hot searches", https://help.naver.com/support/service/main.nhn?serviceNo=606&categoryNo=1989, 2015.

[6] Daum Search Help, "Realtime hot issues" http://cs.daum.net/faq/15/14957.html#28971, 2016.

[7] Min-Yeong Chong, "Selecting a key issue through association analysis of realtime search words", Journal of Digital Convergence, Vol. 13, No. 12, pp. 161-169, 2015.

[8] Min-Yeong Chong, "Extracting week key issues and analyzing differences from realtime search keywords of portal sites", Journal of Digital Convergence, Vol. 14, No. 12, pp. 237-243, 2016.

[9] Kyoung-HoChoi,Jeong-Hye Park, "The Analysis of Public Awareness about Literary Therapy by Utilizing Big Data Analysis - The aspects of convergence literature and statistics", Journal of Digital Convergence, Vol. 13, No. 4, pp. 395-404, 2015.

[10] Matthew A. Russell, "Mining the Social Web:Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub and More", p.411, O'Reilly Media, Inc., 2013.

[11] Xiao Fang, and Olivia R. Liu Sheng, "Designing a better web portal for digital government: a web-mining based approach", Proceedings of the 2005 national conference on Digital government research. Digital Government Society of North America, pp. 277-278, 2005.

[12] KISO Validation Committee, "The fourth validation report about realtime hot searches of Naver", 2015.

[13] Simon Dennis, Peter Bruza and Robert McArthur, "Web Searching: A Process-Oriented Experimental Study of Three Interactive Search Paradigms", Journal of the American Society for Information Science and Technology, Vol. 53, No. 2, pp. 120-133, 2002.

[14] Seong-Hoon Lee and Dong-Woo Lee, "Current Status of Big Data Utilization", Journal of Digital Convergence, Vol. 11, No. 2, pp. 229-233, 2013.

[15] Jon Starkweather, "Introduction to basic Text Mining in R", p.10, University of North Texas, 2014.

[16] George E. P. Box,Gwilym M. Jenkins,Gregory C. Reinsel, and Greta M. Ljung, Time Series Analysis: Forecasting and Control, John Wiley & Sons, 2016.

[17] Alysha M De Livera, Rob J Hyndman, and Ralph D Snyder, "Forecasting time series with complex seasonal patterns using exponential smoothing", Journal of the American Statistical Association, Vol. 106, pp. 1513-1527, 2011.

[18] Guoqiang Zhang, B. Eddy Patuwo, and Michael Y. Hu, "Forecasting with artificial neural networks: The state of the art", International Journal of Forecasting, Vol. 14, pp. 35-62, 1998.

[19] Frauke Günther and Stefan Fritsch, "neuralnet: Training of Neural Networks", The R Journal Vol. 2, No. 11, pp. 30-38, 2010.

[20] Yoon-Su Jeong, "Subnet Generation Scheme based on Deep Learning for Healthcare Information Gathering", Journal of Digital Convergence, Vol. 15, No. 3, pp. 221-228, 2017.

[21] Eun-Jung Choi, Sea-Won Choi, Se-Yeon Lee, and Myhung-Joo Kim, "Analysis of the effect of the mention in SNS on the result of election", Journal of Digital Convergence, Vol. 15, No. 2, pp. 191-197, 2017.

정 민 영(Chong, Min Yeong)

・1991년 2월 : 숭실대학교 전자계산학과(공학사)
・1993년 2월 : 숭실대학교 전자계산학과(공학석사)
・2004년 8월 : 전남대학교 컴퓨터정보통신공학과(공학박사)
・1996년 3월 ~ 현재 : 광주여자대학교 실버케어학과 교수
・관심분야 : 빅데이터분석, 소프트웨어공학, 컴퓨터응용
・E-Mail : mychong@kwu.ac.kr