http://dx.doi.org/10.14400/JDC.2015.13.12.161

실시간 검색어 연관 분석을 통한 핵심 이슈 선정

정민영
광주여자대학교 실버케어학과

Selecting a key issue through association analysis of realtime search words

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

요 약 포털 사이트의 실시간 검색어는 현재 관심이 급상승하고 있는 이슈를 보여주기 위해 주로 검색횟수가 많은 순서에 따라 몇 초 간격으로 제공되고 있다. 그렇지만 너무 짧은 시간 내에 순위가 바뀌는 실시간 검색어의 특성 때문에 하루의 핵심 이슈를 비켜가는 문제가 발생한다. 본 논문에서 이러한 문제를 보완하기 위해 검색어들 사이의 연관 분석을 통하여 검색어들이 관련된 핵심 이슈를 도출하는 방법을 제안하고자 한다. 이를 위해 먼저 실시간 검색어를 순위와 상대적 관심도를 기반으로 점수화하여 집단별 기술통계를 통해 최상위 10개의 검색어를 도출한다. 그 다음으로 지지도와 신뢰도를 기반으로 연관 규칙을 추출하고 이를 가시화하는 그래프 결과를 바탕으로 핵심 이슈를 선정한다. 실험 결과는 단일 최상위 실시간 검색어보다 연관분석을 통해 높은 점수로 선정된 핵심 이슈가 더 큰 의미를 갖는다는 것을 보여준다.

주제어 : 실시간 검색어, 연관 분석, 텍스트 마이닝, 웹 마이닝, 빅데이터

Abstract Realtime search words of typical portal sites appear every few seconds in descending order by search frequency in order to show issues increasing rapidly in interest. However, the characteristics of realtime search words reordering within too short a time cause problems that they go over the key issues of the day. This paper proposes a method for deriving a key issue through association analysis of realtime search words. The proposed method first makes scores of realtime search words depending on the ranking and the relative interest, and derives the top 10 search words through descriptive statistics for groups. Then, it extracts association rules depending on 'support' and 'confidence', and chooses the key issue based on the results as a graph visualizing them. The results of experiments show that the key issue through association rules is more meaningful than the first realtime search word.

Key Words : realtime search words, association rules, text mining, web mining, big data

1. 서론

최근 들어 사물인터넷의 발전과 더불어 빅데이터 분석과 활용이 우리 생활 깊숙한 곳까지 영향을 미치고 있는 가운데, 비정형 데이터 마이닝의 중요성이 대두되고 있다[1,2,3,4,5]. 특히 텍스트 마이닝에 대한 관심은 웹 마

* 본 논문은 2015 학년도 광주여자대학교 교내연구비 지원에 의하여 연구되었음
Received 12 October 2015, Revised 15 November 2015
Accepted 20 December 2015
Corresponding Author: Min-Yeong Chong
(Kwangju Women's University)
Email: mychong@kwu.ac.kr

ISSN: 1738-1916

이닝과 더불어 웹 서비스를 통해 일반적인 궁금증을 해결하고자 하는 욕구만큼 꾸준히 증가하고 있다[6,7,8,9]. 이에 검색어 중심의 정보검색 서비스를 제공하는 포털 사이트들은 보다 양질의 정보 서비스가 보다 많은 사용자를 모이게 한다는 측면에서 그 무엇보다 중요하게 다루고 있으며 경쟁도 치열할 수밖에 없다. 사용자가 입력하는 검색어 속에 들어있는 궁금증과 문제, 그리고 현상과 의미를 오피니언 마이닝이나 소셜네트워크 분석을 통하여 파악하여 사용자의 요구에 맞는 서비스를 새로 개발하거나 기존의 서비스를 개선함으로써 사용자의 관심을 유도하여 보다 적극적인 사용자가 많아지도록 노력하고 있다[10,11,12]. 실제로 '네이버'의 경우, '실시간 급상승 검색어'와 '핫토픽 검색어', '다음'의 경우 '실시간 이슈 검색어'와 소셜메트릭스, 구글의 경우, 구글트렌드의 인기 급상승 검색어와 인기 급상승 이야기 같이 빅데이터 분석을 통해 사용자의 관심사를 파악하고 주된 관심사에 대한 정보를 제공하는 서비스를 실시간으로 지속적으로 제공하고 있다.

그렇지만 단순한 단어로만 이슈를 파악해야 하는 한계와 짧은 시간에 급상승된 검색어의 한계가 존재하여 순간순간 변화하는 단어 중심 이슈의 일시성을 극복하지 못하고 있으며 이를 보완한 것처럼 여겨지는 네이버의 핫토픽 키워드조차도 여전히 단어 중심에서 벗어나지 못하고 있고 구글트랜드의 인기 급상승 이야기의 경우 관련 검색어들을 묶은 형태의 이슈를 보여주지만 관련 검색어 사이의 관계가 누락되어 있고 이마저도 대한민국에서 서비스 되고 있지 않다. 오피니언이나 사용자 자신만의 주요 관심사를 잘 파악할 수 있도록 한 '다음'의 소셜메트릭스 서비스도 SNS 상에서 토론을 전제해야 하는 한계가 있다.

따라서 본 논문에서는 너무 짧은 시간 내에 변화하는 실시간 검색어들로 인해 그날의 핵심 이슈를 향하는 집중도가 떨어지는 문제를 보완하기 위해 먼저 실시간 검색어를 순위와 상대적 관심도를 기반으로 점수화하여 집단별 기술통계 분석을 통해 실시간 검색어별 점수를 집계하고 합산된 점수의 내림차순으로 정렬하여 시간별 상위 실시간 검색어를 하루 동안 수집한다. 그리고 이를 토대로 하루 동안 점수를 가장 높게 받은 이른바 '오늘의 이슈 검색어'와 하루 동안 수집된 시간별 실시간 검색어 사이의 연관 분석을 통하여 실시간 검색어들이 수렴하고 있는 이른바 '오늘의 핵심 이슈'를 도출하는 방법을 제안한다. 연관 분석은 한 번에 시간별 상위 실시간 검색어를 하루 동안 선택한 것을 분석 대상으로 삼아, 지지도와 신뢰도가 일정한 수준 이상인 연관 규칙을 구하고 상승도 내림차순으로 정렬하여 상위의 규칙을 추출하는 것이다. 그 결과를 그래프로 출력하여 핵심 이슈를 선정한다. 분석 대상은 '네이버'의 실시간 급상승 검색어로 하며 분석 도구는 R 언어를 사용한다.

2. 실시간 검색어 분석

2.1 실시간 검색어의 종류와 특성

포털 사이트의 실시간 검색어는 단위 시간 동안 입력되는 검색어의 입력 횟수를 분석하여 그 증가 비율이 가장 큰 검색어부터 차례로 나타내는 서비스이다. 따라서 단순히 일정기간 동안 검색어의 입력 횟수가 많은 것을 보여주는 '종합 검색어'와 달리 현재 급증하는 비율을 보이는 사용자의 관심사에 대한 흐름을 살펴서 이슈와 트렌드를 파악할 수 있는 기초를 제공한다는 측면에서 포기할 수 없는 중요한 서비스로 인식하고 있다. 사용자들이 입력한 검색어들을 집계하여 게시함으로써 보다 많은 사용자들이 보다 많은 콘텐츠와 엮어지게 하고, 공감대 확산을 통해 여론을 형성하는데 도움을 주며, 모이는 사람이 많은 만큼의 비즈니스로 연결시킬 수 있으며 새로운 문화공간으로서 역할을 할 수 있다. 현재 포털 사이트의 대표적인 실시간 검색어로는 '네이버'의 실시간 급상승 검색어, '다음'의 실시간 이슈 검색어, '네이트'의 실시간 검색어, 그리고 구글트렌드의 인기 급상승 검색어 등이 있다.

네이버의 실시간 급상승 검색어는 총 10개의 검색어를 1위에서 10위까지 현재 순위 바로 오른쪽 옆에 함께 표시하며 검색어 바로 오른쪽에는 상승 화살표와 함께 상대적 관심도에 해당되는 것을 표시한다. 이것은 해당 검색어 자체에 대한 상대적 관심도로서 검색어의 실시간 검색 횟수 및 순위를 과거의 것과 비교한 것을 보여주는 하나의 지표이다. 만약 검색어의 실시간 검색 횟수 및 순위와 비교할 과거 데이터가 없는 경우에는 숫자 대신 'NEW'라고 표시한다[13].

'네이버'의 실시간 급상승 검색어 서비스와 거의 유사

한 서비스로는 '다음'의 실시간 이슈 검색어와 '네이트'의 실시간 검색어가 있는데 포털 사용자의 관심 분야 및 경향, 검색 점유율, 선정 기준, 갱신 주기의 차이에서 오는 특성으로 인해 어느 정도 다른 결과와 관점을 보여주지만 모두가 짧은 시간에 주목받는 검색어를 중심으로 치우쳐 있는 점은 비슷하다. 따라서 일반적으로 많은 사용자가 관심을 갖는 이슈에 집중되는 특징을 가지고 있으나 특정 분야에 관련된 이슈나 특정 주제에 대한 이슈에는 취약하고 트렌드를 파악하거나 검색어 사이의 연관관계는 보여주지 않는다.

이런 측면에서는 구글트렌드의 인기 급상승 검색어와 인기 급상승 이야기 서비스는 현재 관심을 보이는 몇 가지 주제를 비교하면서 관심도에 대한 트렌드를 파악할 수 있게 하는 측면에 나름대로의 의미가 있으므로 주목해볼 필요가 있다. 인기 급상승 검색어는 기본적인 실시간 검색어 서비스에 추가하여 중심의 관심도 트렌드를 제공하는데 비해 인기 급상승 이야기는 몇 개의 관련 검색어들이 모인 그룹 형태로 관심도에 대한 트렌드를 제공한다.

<Table 1>은 유사한 특성을 보이는 국내 대표 포털 사이트를 대표하는 네이버와 구글트렌드에서 제공하는 실시간 검색어에 대한 특징을 비교한 것이다. 네이버로 대표되는 국내 포털 사이트의 실시간 검색어의 갱신주기는 대부분 15초 정도이므로 검색어들 사이의 연관성을 찾기에는 너무 짧은 주기를 가지고 있다.

〈Table 1〉 Comparison of realtime search words between portals

Portal name / comparison item	Naver	Google Trends
Name	realtime hot searches	Trending searches
Posted ranking	Top 10	Being different by country
Refresh cycle	15 seconds	one hour
Region	Republic of Korea	Around the world
Related Services	Hourly hot topic keyword	Interest over time, Regional interest

반면에 구글트렌드의 인기 급상승 검색어는 1시간의 갱신 주기를 가지고 있으며[14] 각 검색어에 대한 시간별 흐름을 게시하고 구글트렌드의 인기 급상승 이야기 서비스를 통해 검색어 그룹별 트렌트를 제공하고 있지만 여기서도 검색어 사이의 관계를 찾아서 함께 보여주지는 않는다.

2.2 실시간 검색어 선정 기준

포털의 실시간 검색어를 결정하는데 가장 중요한 역할을 하는 것은 선정 기준이다. 사용자가 검색바에 입력한 검색어를 분석하여 가장 주목받는 검색어를 선정하여 다시 사용자에게 되돌려 주어 공감대를 형성함으로써 보다 많은 사용자들이 모이게 하는 것이므로 검색어 선정 기준은 보다 객관적이고 합리적이며 명확해야 한다. 그러므로 포털에서는 보다 공신력 있는 기관의 검증을 비롯하여 공감대를 넓힐 수 있는 기준이 되도록 하는 노력을 끊임없이 해야 한다.

네이버의 경우, 실시간 급상승 검색어의 순위를 선정할 때 몇 가지 기준에 입각하여 자동으로 선정하는 방식을 사용하는 것을 원칙으로 한다. 먼저 일정기간을 기준으로 사용자가 입력한 검색어가 과거 시점이나 다른 검색어에 비해 상대적으로 급격하게 상승한 비율을 기준으로 순위를 선정한다. 따라서 일상적으로 많이 입력되어 일정기간 동안 검색횟수가 크지만 기준 시간 당 검색 횟수 비율에 큰 변화가 없는 검색어는 상위 순위에 오르지 못한다. 그 다음으로 동일인이 일정기간을 기준으로 같은 검색어를 두 번 이상 입력할 경우, 한 번 입력한 것과 동일하게 계산하고 차트에 이미 노출되고 있는 검색어를 클릭한 경우이거나 특정 시간대에 일상적으로 많이 입력되는 검색어는 검색 횟수에 포함시키지 않는다. 네이버 실시간 급상승 검색어 서비스는 임의로 검색어를 추가하거나 제외할 수 없으며, 검색어를 임의 조정하지 않지만 불법/범죄/반사회성, 서비스품질 저해, 성인/음란성, 명예훼손, 개인정보, 기타 법령, 행정/사법기관의 요청 등의 사유에 해당될 경우 검색어 노출을 제외할 수 있다. 네이버 실시간 급상승 검색어 순위 점수는 (관측횟수-기대횟수)를 표준편차로 나누고 순위차 보정값과 관측횟수 보정값을 더하여 계산한다. 여기서 관측횟수는 순수한 검색횟수를 말하며, 기대 횟수는 과거 일일 평균 검색횟수와 어제 검색횟수 중 최대값을 취한 값에 시간대특징보정값, 전체검색량보정값, 실급검노출보정값을 곱하여 구한 값을 말한다. 이러한 점수는 15초 마다 계산되어 갱신되는 것을 기본으로 하고 있다[15].

2.3 실시간 검색어의 한계

실시간 급상승 검색어를 선정할 때 사용자가 검색바에서 입력한 검색어만을 가지고 검색횟수를 평가하지 못

하고 여러 가지 현실적 상황에 따라 앞서 제시한 제외어 같은 기준을 사용할 수밖에 없는 근본적인 한계가 분명히 존재하지만 이 부분은 연구 범위에 해당되지 않는다.

다만 실시간 급상승 검색어는 '실시간' 이라는 말 자체가 시간의 제약을 받는 것이라는 것을 의미하므로 순식간에 증가비율이 높아질 경우 매우 짧은 시간에도 검색어 순위가 수차례 새롭게 선정될 수 있고 사용자의 검색 형태에 따라 검색어의 순위 변동 폭이 커질 수 있으며, 설사 순위권 밖으로 내려가 표시되지 않아 없어진 것처럼 보이는 검색어도, 10위권 안으로 재 진입하기만 하면 언제라도 다시 표시될 수 있으므로 짧은 기간 동안 특정 이슈를 집중 부각시키려는 집단의 개입을 크게 차단시키지 못한다거나, 많은 사용자가 어느 정도의 기간 동안 관심을 가지고 있는 이슈가 10위권 안에 들지 못함으로써 주목을 받지 못하고 사라질 수도 있다는 한계는 개선 가능성을 열어두고 주목해서 볼 필요가 있다.

포털 '다음'의 소셜메트릭스는 주로 트위터 상에 실시간으로 올라오는 글들을 모아서 단어 단위로 쪼갠 뒤 빅데이터 분석을 통해 검색어를 추출하고 연관어를 추가로 분석하여 오피니언까지 파악할 수 있게 한다. 이는 검색어가 사용자에게 실제로 어떤 의미를 갖는지 보다 깊은 정보를 제공할 수 있으므로 실시간 급상승 검색어의 한계를 극복하는 하나의 방법이 될 수 있다. 그렇지만, 이 서비스는 소셜네트워크 서비스를 통해 어느 정도 토론이 이루어진 후에야 가능하고 처음부터 토론에 적합하지 않은 것은 아예 누락될 수밖에 없다는 한계가 여전히 존재한다.

현재 일부 다른 주요 국가에서만 서비스되고 있는 구글트렌드의 인기 급상승 이야기 서비스의 경우 검색어 그룹에 대한 주제별, 시간별, 지역별 관심도에 기반을 둔 트렌드를 제공하고 있지만 검색어 그룹 속의 검색어들 사이의 관계를 보여주지 못하는 한계가 있으므로 이를 개선하거나 보완하고자 하는 추가적인 노력이 필요하다.

3. '오늘의 핵심 이슈' 선정

3.1 '오늘의 핵심 이슈' 선정 과정

본 논문에서는 네이버의 실시간 급상승 검색어를 기반으로 하여 [Fig. 1]과 같은 과정을 통해서 오늘의 핵심 이슈를 선정한다. 먼저 실시간 급상승 검색어 수집 단계에서는 실시간 급상승 검색어를 파싱하여 중요도 점수를 계산하고 이를 검색어와 함께 저장하는 것을 1분 간격으로 24시간 반복한다. 다음으로, 이렇게 저장된 실시간 급상승 검색어와 중요도 점수를 읽어 들이고 시간별 이슈 검색어 추출 단계로 들어간다. 이 단계에서는 1시간 단위로 실시간 급상승 검색어를 기준으로 집계하고 그 결과로 나온 점수합계의 내림차순으로 정렬한 다음, Top 10을 뽑아 24회(0~23시) 누적시킨다. 그런 다음 일일(24시간)동안 누적된 Top 10들을 대상으로 하여 오늘의 이슈 검색어(Top 20) 추출 단계와 시간별 이슈 검색어의 연관 분석 단계를 수행하고, 오늘의 이슈 검색어 및 연관 규칙 가시화 단계를 거쳐 오늘의 핵심 이슈를 선정한다.

[Fig. 1] The selection process for today's key issue

'오늘의 이슈 검색어'는 하루 동안 누적된 Top 10들을 대상으로 검색어 기준으로 집계한 일일 누적 점수합계를 구하고 이것의 내림차순으로 정렬하여 추출한 최상위 검색어로서 하루 동안 가장 주목받은 검색어를 말한다.

'오늘의 핵심 이슈'는 하루 동안 누적된 Top 10들을 시간별 검색어로 분리한 다음, 이를 대상으로 연관 분석을 하여 지지도와 신뢰도가 일정 수준 이상인 연관 규칙

을 구하고 상승도 내림차순으로 정렬하여 상위 규칙을 추출하고 이를 그래프로 나타내어 검색어 사이의 연관성이 높은 것을 선정한 것으로서 하루 동안 실제적으로 가장 큰 이슈가 되는 관련 검색어들의 집단을 말한다.

따라서 독립적으로 급상승된 결과를 토대로 선정한 네이버의 실시간 급상승 검색어보다 실질적으로 관심도가 높은 이슈를 선정할 수 있다.

3.2 실시간 급상승 검색어 수집과 점수 기록

실시간 급상승 검색어는 1분 단위로 네이버 홈페이지 원시코드를 읽어서 네이버의 실시간 급상승 검색어 순위 선정 기준을 토대로 게시된 10개의 실시간 급상승 검색어를 비롯하여 그와 관련된 순위, 상승상태, 상대적 관심도를 파싱(parsing)하여 추출한다. 파싱은 먼저 동적으로 얻는 홈페이지 원시코드에 대해 R언어의 문자열과 정규식 처리 함수를 이용하여 불필요한 태그를 뛰어 넘어 실시간 이슈 검색어 제목이 있는 위치를 찾은 다음, 이 후에 일정한 패턴으로 나타나는 검색어와 상승상태, 그리고 상대적 관심도를 추출하고 차례대로 순위를 획득하는 것을 10회 반복하는 것을 말한다.

여기서 순위는 1~10, 상승상태는 '상승' 또는 'NEW', 상대적 관심도는 상승상태가 '상승'인 경우에 숫자로 표시되고, 상승상태가 'NEW'인 경우에 값이 표시되지 않는다.

```
#Looking for the maximum value of the relative interest
max <- -1
for(i in 1:10)
{
  if(issueDataFrame$concern[i] > max)
  {
    max <- issueDataFrame$concern[i]
  } # end if
} # end for
# If the status 'NEW' happens more than once,
# increase the maximum value of the relative interest as it
k <- max
for(i in 1:10)
{
  if(identical(issueDataFrame$up[11-i], "NEW"))
  {
    k <-k+1;
    issueDataFrame$concern[11-i] = k
  }
}
#Importance score = Relative interest(v) + Ranking score(g)
for(i in 1:10)
{
  v<-9*issueDataFrame$concern[i]/k  # Maximum 9 points
  g<-(10-i)*10+1 #Ranking 1 (91points), ranking 10 (1point)
  issueDataFrame$score[i] = v + g
}
```

[Fig. 2] The algorithm for scoring the importance of realtime search words

그리고 이들을 기반으로 [Fig. 2]와 같은 알고리즘에 의해 실시간 급상승 검색어의 중요도 점수를 계산한다. 먼저 상대적 관심도의 최대값을 찾은 다음, 'NEW'가 존재하는 만큼 최대값을 증가시킨다. 그런 다음, 상대적 관심도는 최대 9점으로 환산하고, 순위점수는 1점에서 91점 까지 10점 단위로 부여하여 이들을 합산한 값을 최종 점수로 삼는다.

3.3 '오늘의 이슈 검색어' 추출

'오늘의 이슈 검색어'를 추출하기 위해서는 먼저 시간별 Top 10 검색어와 중요도 점수를 24시간 동안 누적한 일일 누적 Top 10을 대상으로 집단별 기술통계 분석을 통해 각 검색어별 케이스수, 누적점수평균, 누적점수합계를 집계한다. 그런 다음, 누적점수합계를 키 항목으로 하여 내림차순으로 정렬하면 된다. 이 정렬 결과 중에서 최상위에 있는 검색어가 하루 동안의 누적점수합계가 가장 높은 '오늘의 이슈 검색어'에 해당되며 여기서 오늘의 이슈 검색어 Top 20과 Top 10도 추출할 수 있다.

집계 결과, 검색어의 케이스수가 많은 것은 적어도 하루 동안 지속적인 많은 관심을 받았다는 것을 의미하고, 평균이 높다는 것은 하루 중 어느 순간 그 만큼 강한 인상을 남겼다는 것을 의미한다. 그리고 합계가 높다는 것은 그 만큼 강한 인상을 적어도 하루 동안은 지속적으로 남겼다는 것을 의미하므로 '오늘의 이슈 검색어'로 선택하는 근거로서 역할을 한다.

3.4 시간별 이슈 검색어의 연관 분석

'오늘의 핵심 이슈'를 추출하기 위해서는 먼저 일일(24시간)동안 누적된 Top 10들을 시간별(0~23시) 검색어로 분리하고, 이를 대상으로 검색어 사이의 연관 분석을 한다. 즉, 일정 수준 이상인 지지도(0.2)와 신뢰도(1.0)를 갖는 연관 규칙을 생성하여 상승도 내림차순으로 정렬하고 상위 규칙 Top 30을 추출하여 이를 그래프로 나타내면서 검색어 사이의 연관성이 드러나게 만든다.

여기서 연관 규칙은 조건과 조건에 따른 결과를 만드는 규칙으로서 조건에 해당되는 검색어들에 따라 결과에 해당되는 검색어가 나오게 함으로써 조건과 결과 사이의 연관성을 표현하는 것이다. 지지도는 검색어들이 동시에 Top 10에 선택될 확률을 말하며 전체적인 선택 경향을 파악할 수 있는 기준이다. 신뢰도는 조건 검색어가 선택

되었다는 가정 하에 결과 검색어가 선택될 확률로서 연관 규칙의 강도를 파악할 수 있는 기준이다. 향상도는 결과 검색어가 선택될 확률을 기준으로 조건 검색어에 따른 결과 검색어가 선택될 확률의 비율에 해당되며 1보다 크면 양의 상관관계이고 1보다 작으면 음의 상관관계를 갖는 기준이다[16,17,18].

그리고 결과 검색어 중심 연관 규칙 그래프를 통해서 결과 검색어 중심으로 묶어진 검색어 그룹을 일차적인 핵심 이슈 대상으로 선정하고, 지지도와 향상도 중심 연관 규칙 그래프를 통해서 향상도가 높아 양의 상관관계가 강하게 나타나는 것을 이차적인 핵심 이슈 대상으로 선정한다. 마지막으로 이차 핵심 이슈 대상의 각 검색어의 중요도 점수의 합을 구하여 가장 큰 검색어 그룹을 '오늘의 핵심 이슈'로 선정한다.

4. '오늘의 핵심 이슈' 실증 사례

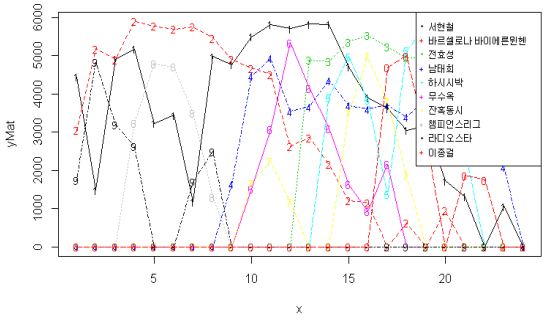
4.1 '오늘의 이슈 검색어' 사례

[Fig. 3]은 네이버에서 2015년 5월 7일 0시 0분 부터 23시 59분 까지의 실시간 급상승 검색어 상위 10개씩을 갱신주기 15초보다 긴 1분 단위로 수집하여 분석한 결과로 추출된 오늘의 이슈 검색어 상위 20개의 워드 클라우드를 나타낸 것이다. 오늘의 최상위 이슈 검색어는 [Fig. 3]의 중앙에 크게 자리잡은 '서현철'이다.

[Fig. 3] A word cloud of today's top 20 issue searches

네이버 실시간 급상승 검색어의 경우 어느 순간에 관심이 급상승하고 있는 검색어를 독립적으로 제시하기 때문에 하루에도 수많은 변화가 있어서 그 날의 이슈를 특정할 수 없지만 [Fig. 2]에서 제시한 알고리즘을 기반으로 계산된 점수를 기준으로 선정된 '오늘의 이슈 검색어'는 하루 동안에 관심을 주목받은 이슈로서의 의미를 가진다.

[Fig. 4]는 오늘의 이슈 검색어 상위 10개만을 뽑아서 시간별 점수의 변화를 나타낸 것이다. 오늘의 최상위 이슈 검색어인 '서현철'에 대한 중요도 점수의 시간별 변화를 살펴보면 가장 넓은 시간대에 비교적 높은 점수를 유지한 것으로 나타났다. 그 다음으로 '바르셀로나 바이에르뮌헨'은 비교적 넓은 시간대와 높은 점수를 유지하고 있으면서 특이하게 그 안에 '챔피언스리그'를 완전히 내포하고 있는 것으로 나타났다.



[Fig. 4] Hourly changes in interest of today's top 10 issue searches

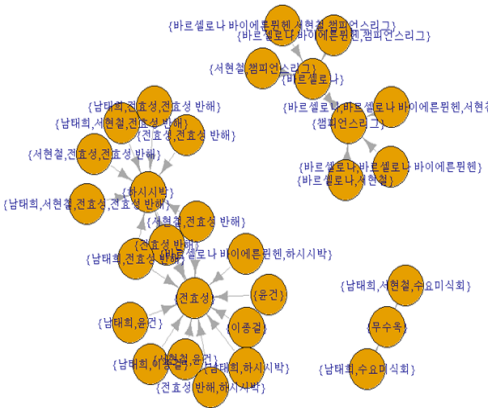
4.2 '오늘의 이슈 검색어' 연관 분석 사례

[Fig. 5]는 연관 규칙 상위 30개의 조건과 결과를 매트릭스로 나타낸 것이다. 왼쪽에 보이는 첫 번째 열이 조건이고, 오른쪽에 보이는 두 번째 열이 그 결과이다. 즉 각 행은 조건이 일어나면 결과가 발생하는 규칙을 나타낸다.

또한 [Fig. 6]은 [Fig. 5]의 조건에 따른 결과가 보다 가시적으로 돋보이도록 연관규칙 상위 30을 그래프 형태로 나타낸 것이다. 그 중에서도 특히 규칙의 결과로 나타나는 대표적인 검색어는 바르셀로나, 챔피언스리그, 하시시박, 무수옥, 전효성 등이 있다. 이들이 이슈의 중심에 있는 검색어들인데, 이 중 바르셀로나와 챔피언스리그는 필요충분조건이어서 밀접한 관계가 있는 이슈로 볼 수 있다.

```
> rulemat
      [,1]                                                  [,2]
 [1,] "{챔피언스리그}"                                      "{바르셀로나}"
 [2,] "{바르셀로나}"                                        "{챔피언스리그}"
 [3,] "{바르셀로나 바이에른뮌헨,챔피언스리그}"              "{바르셀로나}"
 [4,] "{바르셀로나,바르셀로나 바이에른뮌헨}"                "{챔피언스리그}"
 [5,] "{서현철,챔피언스리그}"                               "{바르셀로나}"
 [6,] "{바르셀로나,서현철}"                                 "{챔피언스리그}"
 [7,] "{바르셀로나 바이에른뮌헨,서현철,챔피언스리그}"       "{바르셀로나}"
 [8,] "{바르셀로나,바르셀로나 바이에른뮌헨,서현철}"         "{챔피언스리그}"
 [9,] "{전효성 반해}"                                       "{하시시박}"
[10,] "{전효성,전효성 반해}"                               "{하시시박}"
[11,] "{남태희,전효성 반해}"                               "{하시시박}"
[12,] "{서현철,전효성 반해}"                               "{하시시박}"
[13,] "{남태희,수요미식회}"                                 "{무수옥}"
[14,] "{남태희,전효성,전효성 반해}"                         "{하시시박}"
[15,] "{서현철,전효성,전효성 반해}"                         "{하시시박}"
[16,] "{남태희,서현철,전효성 반해}"                         "{하시시박}"
[17,] "{남태희,서현철,수요미식회}"                          "{무수옥}"
[18,] "{남태희,서현철,전효성,전효성 반해}"                  "{하시시박}"
[19,] "{이종걸}"                                           "{전효성}"
[20,] "{윤건}"                                             "{전효성}"
[21,] "{전효성 반해}"                                      "{전효성}"
[22,] "{하시시박}"                                         "{전효성}"
[23,] "{남태희,이종걸}"                                    "{전효성}"
[24,] "{남태희,윤건}"                                      "{전효성}"
[25,] "{서현철,윤건}"                                      "{전효성}"
[26,] "{전효성 반해,하시시박}"                             "{전효성}"
[27,] "{남태희,전효성 반해}"                               "{전효성}"
[28,] "{서현철,전효성 반해}"                               "{전효성}"
[29,] "{남태희,하시시박}"                                  "{전효성}"
[30,] "{바르셀로나 바이에른뮌헨,하시시박}"                 "{전효성}"
```

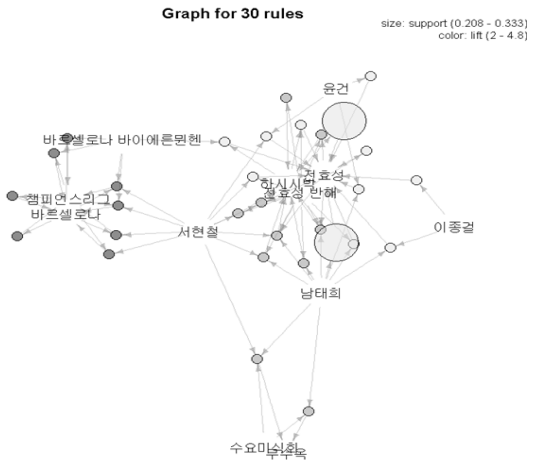
[Fig. 5] A matrix of top 30 association rules



[Fig. 6] A result-oriented graph for top 30 association rules

4.3 '오늘의 핵심 이슈' 선정 사례

[Fig. 7]은 지지도(support)와 향상도(lift)를 중심으로 연관 규칙 상위 30을 나타낸 그래프이다. 원의 크기는 지지도 크기이고, 원 색깔의 농도는 향상도의 크기를 나타낸다. [Fig. 6]이 조건에 대한 결과로 나타나는 검색어 중심의 그래프라면 [Fig. 7]은 조건과 결과로 나타나는 검색어들의 연관성을 보여주는 그래프이다. 따라서 [Fig. 7]을 통해 검색어들이 지지도와 향상도가 반영되어 어울려 있는 덩어리로 나타나므로 이들을 통해 핵심 이슈를 선정할 수 있다.



[Fig. 7] A 'support and lift' oriented graph for top 30 association rules

[Fig. 6]에서 필요충분조건으로 나타났던 '바르셀로나'와 '챔피언스리그'가 [Fig. 7]에서 '바르셀로나 바이에른뮌헨'과 더불어 향상도가 강한 형태로 뭉쳐있으므로 가장 주목을 받는 핵심 이슈로 볼 수 있다. [Fig. 6]에서 비교적 다수의 필요조건으로서 역할을 한 '하시시박'과 '전효성'은 '전효성 반해'와 뭉쳐 있고, '무수옥'이 수요미식회'와 뭉쳐있으나 비교적 향상도가 약한 상태에 있는 이슈로 볼 수 있다. 그리고 [Fig. 6]에서 가장 간단한 연결형태로 나타난 '무수옥'은 '수요미식회'와 뭉쳐서 이슈를 형성한 것으로 볼 수 있다. 그런데 오늘의 최상위 실시간 급상승 검색어로 추출된 바 있는 '서현철'이 [Fig. 6]에서 검색어 그룹의 중심으로 나타나지 않고 [Fig. 7]에서도 오히려 다른 검색어들과 뭉쳐있지 않은 것으로 나타나므로 강하게 연관된 검색어가 없는 이슈로 볼 수 있다.

실험결과, '바르셀로나 바이에른뮌헨'(67437.701), '챔피언스리그'(17551.515), '바르셀로나'(14209.429)의 점수의 합(99198.645)이 오늘의 최상위 이슈 검색어인 '서현철'(84867.494)보다도 더 높은 것으로 나타났다. 따라서, '바르셀로나 바이에른뮌헨', '챔피언스리그', '바르셀로나' 검색어 그룹을 '오늘의 핵심 이슈'로 선정한다.

5. 결론

본 논문에서는 실시간 검색어가 검색횟수의 크기 순서에 따라 몇 십초 간격으로 제공됨으로써 인기 검색어

만을 알 수 있는 한계를 극복하기 위해 1분 간격으로 검색어를 수집하고 1시간 단위로 상위 10개의 이슈 검색어를 추출하여 24시간 동안 누적한 다음, 이를 사용하여 집단별 기술 통계 분석을 통해 '오늘의 이슈 검색어'를 선정함과 동시에 이슈 검색어들 사이의 연관 분석을 실시하여 연관성 있는 검색어 집단을 추출하여 가장 주목받는 이슈인 '오늘의 핵심 이슈'를 선정하고 이에 대한 실제적인 예를 제시하였다.

본 논문에서 제안하는 방법은 단순한 개별 단어로만 이슈를 파악해야 하는 한계를 넘어서 연관된 특정 단어들을 한 덩어리의 이슈로 파악하게 함으로써 보다 실제적인 핵심 이슈로 활용할 수 있게 했다는 측면에서 의미가 있다. 특히 개별 단어의 일시성을 보완하여 연관 단어들의 덩어리로서의 얼마간의 지속성을 확보할 수 있고 이를 기반으로 일자별 핵심 이슈를 예측하는 모델을 형성하는데 도움을 줄 수 있으며, SNS를 통하여 보다 심도 있는 토론과 논의를 제공하는 이슈로도 사용될 수 있다. 또한 웹 마이닝을 통해 획득한 텍스트 데이터에 대한 메타 분석의 한 방법이라는 측면에서 차별성이 있으며 이를 바탕으로 일주일, 한달, 일년 단위로 확장시켜 나가는 씨앗 역할로서 관련 연구에 유용하게 활용될 수 있다.

하지만 주요 포털의 실시간 검색어는 포털별 검색어 선정기준, 점유율이나 사용자의 경향의 차이로 인해 다르게 나타날 가능성이 존재하고 네이버에 국한된 짧은 기간의 데이터를 기반으로 한 분석 결과라는 한계가 존재한다. 이를 극복하기 위한 노력의 일환으로 보다 긴 기간에 걸쳐 수집된 여러 포털의 실시간 검색어를 함께 비교분석하는 추가적인 연구가 필요하다.

ACKNOWLEDGMENTS

This work was supported by Research Funds of Kwangju Women's University in 2015.

REFERENCES

[1] Guandong Xu, Lin Li, and Yanchun Zhang, Web Mining and Social Networking: Techniques and Applications. Springer, 2011

[2] Han, Jiawei, and Chi Wang. "Mining latent entity structures from massive unstructured and interconnected data." Proceedings of the 2014 ACM SIGMOD international conference on Management of data. ACM, 2014

[3] Scott Spangler and Jeffrey Kreulen, "Mining the Talk: Unlocking the Business Value in Unstructured Information", IBM, 2007

[4] Ronen Feldman and James Sanger, "The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data", Cambridge University, 2007

[5] Kyoo-Sung Noh, "A Exploratory Study on Big-data based Election Campaign Strategy Model in South Korea ", Journal of Digital Convergence, v.11, no.12, 113-120, 2013

[6] Miner G, Elder J, Hill T, Nisbet R, Delen D, and Fast A, "Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications", Elsevier Academic Press, 2012

[7] Bing Liu, "Web Data Mining: Exploring Hyperlinks: Contents and Usage Data", Springer, 2011

[8] Guandong Xu, Lin Li, and Yanchun Zhang, "Web Mining and Social Networking: Techniques and Applications". Springer, 2011

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

[10] Bing Liu, "Sentiment Analysis and Subjectivity", Handbook of Natural Language Processing, 2010

[11] Reis Pinheiro and Carlos Andre, "Social Network Analysis in Telecommunications". John Wiley & Sons, 2011

[12] Golbandi, Nadav Golbandi, et al. "Expediting search trend detection via prediction of query counts." Proceedings of the sixth ACM international conference on Web search and data mining. ACM, 2013

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

[14] Google Trends Help, Trends Searches, "https://support.google.com/trends/?hl=en#topic=62

48107", 2015
[15] KISO Validation Committee, "The third validation report about realtime hot searches of Naver", 2014
[16] Lee, Changyong, Bomi Song, and Yongtae Park. "Design of convergent product concepts based on functionality: An association rule mining and decision tree approach." Expert Systems with Applications Vol. 39, No. 10, pp.9534-9542, 2012
[17] Hahsler, Michael, and Sudheer Chelluboina. "Visualizing Association Rules: Introduction to the R-extension Package arulesViz.", R project module, pp.223-238, 2011
[18] KeunWon Kim, DongWoo Kim, Kyoo-Sung Noh, and Joo-Yeoun Lee, "An Exploratory Study on Improvement Method of the Subway Congestion Based Big Data Convergence", Journal of Digital Convergence, v.13, no.2, 35-42, 2015

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