The Journal of the Convergence on Culture Technology (JCCT)
Vol. 4, No. 4, pp.337-341, November 30, 2018. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2018.4.4.337
JCCT 2018-11-43

실시간 검색어 분석을 이용한 인터넷 정보 관심도 분석

An Analysis on Internet Information using Real Time Search Words

노기섭*

Giseop Noh*

요 약 온라인 미디어의 지속적인 발전과 최근 모바일 컴퓨팅 사용 환경이 급격하게 개선됨에 따라 인터넷 정보의 유통이 공급자 중심 단방향에서 소비자 중심의 양방향으로 빠르게 변화하였다. 이에 따라 인터넷 정보의 관심도를 측정하는 것이 공급자와 소비자에게 중요한 문제로 대두되었다. 본 논문에서는 국내 인터넷 정보제공 업체에서 제공하는 실시간 검색어를 자동화된 소프트웨어를 구현하여 1개월간의 데이터를 수집하고, 실시간 검색어의 지속시간을 분석하여 인터넷 정보 관심도를 분석하였다.

주요어 : 실시간 검색, 키워드, 정보 지속성, 관심도

Abstract As the online media continues to evolve and the mobile computing environment has improved dramatically, the distribution of Internet information has rapidly changed from one-sided to consumer-oriented. Therefore, measuring the interest of Internet information has become an important issue for suppliers and consumers. In this paper, we analyze the Internet information interest by analyzing the duration of real - time query by collecting data for one month by implementing real - time search word provided by domestic Internet information provider.

Key words : Realtime search, Keywords, Information Persistence, Interestingness

Ⅰ. 서 론

온라인 미디어 기술의 비약적인 발전으로 정보생성 및 공유의 속도가 빠르게 증가하고 있다. 과거의 정보 통신 방법은 신문, 방송, 잡지, 케이블 등 정보 생성자가 일방적으로 제공하고 정보 소비자는 컨텐츠 선택권이 없는 단방향 서비스가 주를 이루었다. 그러나 현대는 모바일 환경을 기반으로 양방향 서비스가 더 큰 영향력을 미치고 있다. 양방향 서비스란 정보 소비자가 원하는 정보를 선택하거나 요청할 수 있고 정보 제공자는 정보 소비자의 요구에 맞추어 정보를 제공하는 형태의 정보제공 서비스를 의미한다. 이러한 양방향 서비스는 정보 소비자가 관심을 갖는 이슈를 파악하는 것이 중요하다.

정보 제공자는 실질적인 정보 소비자의 관심도를 파악하여 요구되는 정보를 신속하게 노출 또는 제공하여 다양한 제공자가 존재하는 정보제공 환경에서 정보 소비자에 대한 영향력을 장악할 수 있다. 이러한 소비자

*정회원, 청주대학교 소프트웨어융합학부 (제1저자)
접수일: 2018년 8월 11일, 수정완료일: 2018년 9월 13일
게재확정일: 2018년 9월 23일

Received: August 11, 2018 / Revised: September 13, 2018
Accepted: September 23, 2018
*Corresponding Author: kafa46@gmail.com
Dept. of Software Convergence, Cheongju Univ, Korea

관심도를 실시간으로 관찰할 수 있는 방법은 각종 포털 싸이트에서 제공하는 실시간 검색어를 확인하는 방법이다. Naver, Daum, Nate 등 우리나라의 대표적인 정보 제공 포털 싸이트에서는 공통적으로 실시간 검색어를 제공하고 있으며, 일부에서는 스마트폰을 이용한 온라인 플랫폼을 정치분야로 확장하려는 노력이 있다[1].

실시간 검색어 분석을 통한 정보 관심도는 다양한 형태로 분석이 가능하다. 예를 들면, 실시간 검색어가 가지는 정서적 특성, 시간대별 특성, 문화적 특성 등에 대한 분석이 가능하다. 실시간 키워드를 시간대별로 분석하고 온라인 뉴스와의 연관성을 키워드 영향력과 뉴스 생성 및 지속성과의 상관관계를 분석한 연구[2], 실시간 검색어의 순위변경 예측에 관한 연구[3], 검색 포털 싸이트에서 검색어를 분석하는 연구[4] 등이 있다. 반면에 실시간 검색어는 온라인 정보 흐름을 왜곡시키는 현상을 발생시키기도 한다. 만약, 실시간 검색어가 운영자에 의해 자의적으로 편집(의도적 배제 또는 추가)될 경우 온라인 정보 제공자(콘텐츠 제공자)의 영업이익을 저해하거나 온라인 여론을 왜곡할 가능성이 있다. 실제로 한국인터넷자율정책기구(www.kiso.or.kr)에서는 네이버 실시간급상승검색어 공정성을 검증하기도 하였다[5].

실시간 검색어의 영향력과 부작용을 고려할 때 실시간 검색어 분석을 통한 인터넷 정보의 관심도를 정확히 분석할 필요가 있다. 그러나 그간의 연구와 검증 방법에 있어서는 깊이 있는 분석이 부족하였다. 본 논문에서는 한국의 대표적인 정보제공 사업자인 Naver에서 발생하는 실시간 검색어를 분석하고, 이를 바탕으로 관심정보의 지속성을 분석하고, 정보 관심도를 관찰 방안을 제시한다.

본 논문의 구조는 Ⅱ장에서 실시간 검색어 수집 방법 및 수집된 데이터셋에 대하여 설명하고, Ⅲ장에서는 실시간 검색어의 지속성과 관심도 분석결과를 제시한다. Ⅳ장에서 결론으로 본 논문을 마무리 한다.

Ⅱ. 실시간 검색 데이터 수집

1. 타겟 시스템 및 자동화 수집도구 구현

네이버(www.naver.com)의 경우 실시간 급상승 검색어를 1위부터 20위까지 30초 단위로 제공하고 있다. 네이버는 메인 페이지에 검색어 순위를 노출하고 있으며 수년간 수집된 실시간 검색어 데이터베이스[1]를 통하여 제공하고 있다(그림 1 참조).

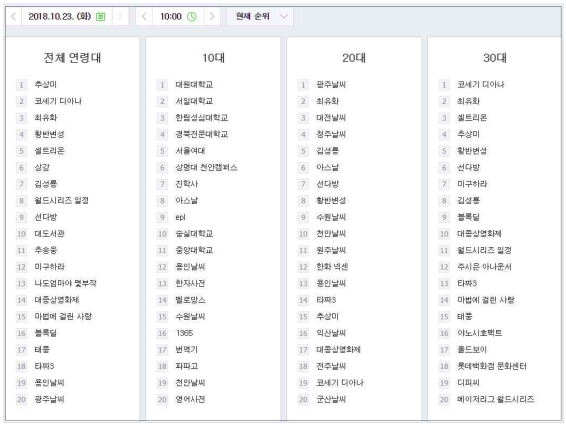

그림 1. 네이버 데이터랩 화면
Figure 1. Screenshot of Naver Datalab

본 논문에서 필요한 실시간 검색어 데이터 수집을 위해 자동화된 수집기(crawler) 설계를 수행하였다. 이를 위해 필요한 인터넷 데이터 수집기는 Selenium, Chromedriver, Beautifulsoup, Python언어를 이용하여 소프트웨어로 구현하였다. 구현된 소프트웨어는 데이터 수집 시작시간, 종료시간, 수집 간격을 선택하여 실행할 수 있도록 하였다. 추가적으로 수집된 키워드의 출현, 소멸 시기를 측정하기 위한 분석 소프트웨어를 Python을 활용하여 구현하였다.

2. 수집된 실시간 검색 데이터

본 논문에서는 실시간 검색어 수집기를 통해 2017년 4월 1일 00시 00분부터 2017년 5월 1일 00시 00분까지 30초 간격으로 총 43,201회에 걸쳐 생성된 실시간 검색어를 수집하였다. 이를 통해 1개월간 수집된 전체 키워드 개수는 5,004개 이다.

Ⅲ. 검색어 지속성 및 관심도 분석

네이버의 실시간 검색어는 [5]에서 제시된 바 있으며 검색 순위를 결정하기 위한 점수 산출 수식은 수식 (1)과 같다.

1) datalab.naver.com/keyword/realtimeList.naver

$$Score = \left[\frac{\text{관측회수} - \text{기대회수}}{\text{표준편차}}\right] + \text{순위차보정} \quad (1)$$
$$+ \text{관측회수보정}$$

$$\text{기대횟수} = \max\left[\frac{\text{과거1주일평균검색횟수}}{\text{어제검색횟수}}\right] \cdot \begin{bmatrix}\text{시간대특징보정}\\ \text{전체검색량보정}\\ \text{실급검노출보정}\end{bmatrix}$$

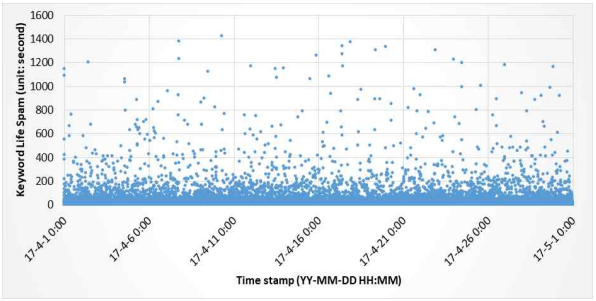

그림 2. 시간에 따른 검색어 지속시간(30일)
Figure 2. Keywords persistence on time series (30 days)

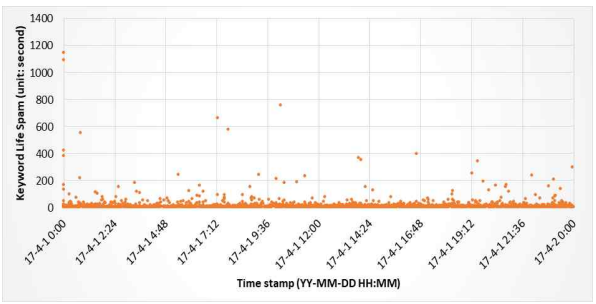

그림 3. 시간에 따른 검색어 지속시간(1일)
Figure 3. Keywords life time on time series (1 day)

수식 (1)은 검색량 증가, 반복패턴 제거, 특정기간, 주간 단위 반복, 과거 검색 경향 등에 따른 가중치를 통해 검색어의 실시간 순위를 산출할 수는 있지만, 실제 어떤 종류의 검색어가 얼마나 오랜 기간 동안 검색어 순위(20위) 내에 존재할 수 있는가에 대한 실증적 분석은 불가능하다. 매 30초 단위로 순위 내에 새로 진입하는 검색어, 지속적으로 존재하는 검색어, 삭제되는 검색어에 대한 추가적 분석이 필요하다. 본 논문에서는 첫 번째 분석으로 새로 진입한 검색어가 얼마 정도 순위 내에 지속할 수 있는지를 분석하였다.

특정 키워드가 순위 내에 새로 진입하여 존재하다가 순위 밖으로 사라지는 시간까지를 라이프 타임(life time, LT)라 하자. LT의 형태를 그림 2와 그림 3에 표시하였다. 그림 2, 3에서 확인할 수 있는 바와 같이 새로 진입한 검색어의 시간 흐름에 따른 특별한 연관관계는 존재하지 않는다. 그러나 LT가 큰 것부터 내림차순으로 정렬한 후 관찰하면 특정 패턴을 파악할 수 있다.

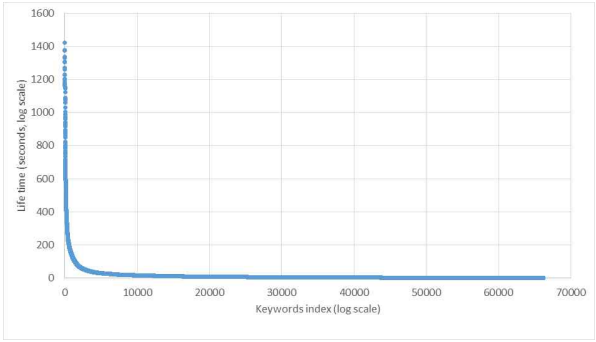

그림 4. LT 내림차순 정렬에 의한 분포
Figure 4. LT value distribution sorted by descending order

그림 4는 LT가 큰 키워드를 순서대로 정렬한 그림으로 long tail의 형태를 보여주고 있다. 극히 일부의 키워드들이 높은 LT를 보이고 있지만 대부분의 키워드들은 100분 미만의 지속시간을 보이고 있다. 그림 4에서 지속시간이 1200분(20시간) 이상인 키워드는 21개이며, 600분(10시간) 이상 검색어 순위에 머무르는 키워드의 경우는 101개로 전체 키워드의 2.02%에 불과하다.

이러한 현상은 발생 패턴을 수학적으로 정형화함으로써 발생 경향을 보다 정확하게 파악할 수 있다. 이를 위해 본 논문에서는 long tail 형태를 보이는 LT 분포가 Power Law (PL)를 따르는지를 분석하였다. 자동화된 분석 도구는 [6]에서 제공하는 도구를 사용하였다.

PL 분포는 heavy tail을 가지는 대표적인 분포로써 분포의 오른쪽 꼬리 부분이 차지하는 확률이 상당히 많은 분포이다. PL 분포는 기존의 확률분포 또는 평균/표준편차 기반으로 표현하기 어려운 현상들을 표현할 수 있다. 예를 들면, 2000년 미국의 인구통계 조사에 따르면 도시(town, city, village)당 평균 인구는 8,226명 이었으나, 대부분의 사람들은 도시에 살고 있어(LA, 뉴욕 등 대도시에 집중) 일반적 평균을 이용한 분포로 현상을 설명할 수 없는 경우를 들 수 있다. PL 현상은 자연적 또는 인공적 결과물을 관찰할 때 빈번히 발생하며, 승수값(power)을 이용하여 잘 설명할 수 있다. 대표적 분야는 천체물리, 언어, 신경망 등이 있다 [7-10]. 일반적으로 수식 (2)와 같이 표현될 수 있다.

$$p(x) \propto x^{-\alpha} \quad (2)$$

$p(x)$는 확률, x는 관측값, α는 상수
파라미터(exponent 또는 scaling parameter로 칭함)

수식 (2)에서 x의 최대값 범위는 한정되지 않기

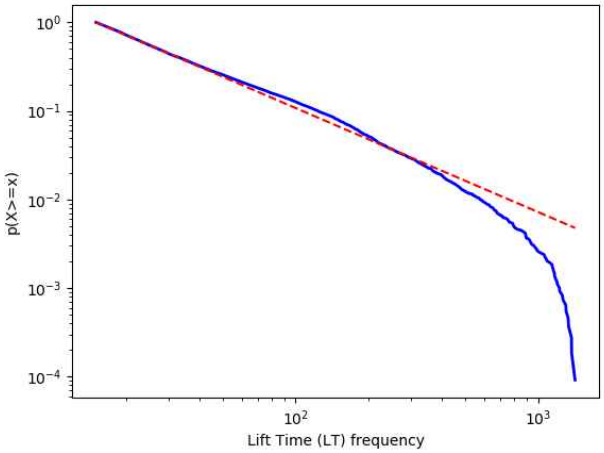

그림 5. LT CCDF 분포 $p(X \geq x)$
Figure 5. LT CCDF distribution $p(X \geq x)$

표 1. Power Law 분포 분석결과
Table 1. Analysis Results on Power Law Distribution

Parameter	Description	Value
α	Scaling parameter	2.174
σ	Standrad deviation	0.111
$x_{\min}$	x 구간 중 최소 값	15.000

때문에 $\alpha < 3$ 범위에서 표준편차가 정의되지 않을 수 있으며, $\alpha < 2$ 범위에서는 평균이 정의되지 않을 수 있기 때문에 평균, 표준편차를 기반으로 도출되는 정규분포 등에서 관측값 분석 한계점에 대한 분석이 가능하다. 이러한 특성 때문에 PL 분포는 Scale-free system으로 해석하기도 한다.

2장에서 수집한 데이터를 분석한 결과는 표 1에 정리하였다. 분석 결과 Scaling parameter 2.174로 PL 분포를 따르는 것을 확인하였으며, LT 존재 분포 complementary cumulative distribution function (CCDF)는 그림 5에 표현하였다. 그림 5에서 파란색 실선은 실측값, 붉은색 점선은 이론적 예측값을 표현한 것으로 $x_{\min} = 15$ (즉, LT값이 15분 이상인 모든 자료)에 대하여 PL 분포와 실제로 유사하게 나타남을 확인 하였다.

Ⅳ. 실험결과 분석

1. 인터넷 정보 관심도 측정 지표

인터넷 정보 중 사용자들의 관심사를 실시간으로 측정할 수 있는 방법은 실시간 검색어를 관찰하는 것이다. 본 논문에서는 특정 포털 정보제공 서비스(네이버)를 대상으로 실시간 검색어의 LT를 관찰하였다. LT는 시간대별로 다양한 특성을 예측할 수 있으나, 인터넷 이용자의 관심도는 특정 검색어의 지속시간을 관찰함으로써 수치화가 가능하다. 본 논문에서는 검색어(LT)의 지속시간 분포 성향과 특징을 기반으로 인터넷 정보 관심도를 분석한다.

2. 인터넷 정보 관심도 분석

실시간 검색어를 분석한 결과 PL 분포의 LT값이 15분 이상에서 유의미한 분포를 의미하는 것을 Ⅲ장에서 확인하였다. 따라서 인터넷 정보에 있어 유의미한 관심도를 예측할 수 있는 검색어는 약 15분 이상 실시간 검색어 순위에 지속되어야 유의미한 의미를 가지는 것을 확인하였다.

또한 대부분의 실시간 검색어는 30분 이내의 LT값을 나타내었다. 본 논문 데이터의 경우 전체 검색어의 99.58% 검색어가 30분 미만의 지속시간을 가지고 있다. 즉 일반적으로 인터넷에 생성되는 정보에 대한 관심도를 LT 측면으로 분석할 경우 99% 이상의 관심 정보(검색어)는 단기간의 관심을 보인 이후 쉽게 소멸되는 것을 확인할 수 있었다. 다만 상위 2%의 특정 검색어는 10시간 이상 관심도가 지속되는 것을 확인할 수 있었는데, 이러한 특징은 인터넷 정보 관심도가 PL 특성을 지닌다는 것을 입증하고 있다.

Ⅴ. 결 론

본 논문에서는 인터넷 정보의 관심도를 분석하기 위하여 국내 인터넷 종합 정보제공 업체의 실시간 검색어를 분석하였다. 분석을 위해 자동화된 자료수집 도구를 설계/구현하였으며, 데이터 분석 및 가공 시스템을 추가로 구현하였다. 1개월 검색어별 지속시간(LT)를 조사하여 LT가 보이는 특성을 파악하였다. 검색어는 PL분포를 따르는 것을 확인하였으며, 인터넷 검색어의 관심도를 LT관점에서 분석하였다. 향후, 추가적인 측정 지

표를 구상하여 보다 정교한 분석을 통해 인터넷 정보 관심도 측정의 정확도를 향상시켜 나갈 예정이다.

References

[1] Yong Jun and S. Lee, "Implementation of a Political Online Platform Using Mobile Phones," *Journal of the Convergence on Culture Technology,* vol. 4, pp. 205-209, May 31, 2018.
DOI: 10.17703/JCCT.2018.4.2.205

[2] Junyep Oh, Seungkyu Lee, and Jooyoup Lee, "A Study on the user attributes for aquisition of information by analyzing the durability of real-time issues," *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology,* vol. 7, pp. 299-314, 04 2017.

[3] K. Dohyeong, K. Byeong Ho, and L. Sungyoung, "'Hot Search Keyword' Rank-Change Prediction," *Journal of KIISE,* vol. 44, pp. 782-790, 8 2017.
DOI: 10.5626/JOK.2017.44.8.782

[4] S. C. Han, Y. Liang, H. Chung, H. Kim, and B. H. Kang, "Chinese trending search terms popularity rank prediction," *Inf. Technol. and Management,* vol. 17, pp. 133-139, 2016.
DOI: 10.1007/s10799-015-0238-0

[5] KISO, "Validation Report on 'Naver' Realtime Arising Keywords (III)," Korea Internet Self-Governance Organization, 2014.

[6] J. Alstott, E. Bullmore, and D. Plenz, "powerlaw: a Python package for analysis of heavy-tailed distributions," PloS one, vol. 9, p. e85777, 2014.
DOI: 10.1371/journal.pone.0085777

[7] M. Michel, H. Kirk, and P. C. Myers, "Mass distributions of stars and cores in young groups and clusters," The Astrophysical Journal, vol. 735, p. 51, 2011.
DOI: 10.1088/0004-637X/735/1/51

[8] G. K. Zipf, The psycho-biology of language: An introduction to dynamic philology: Routledge, 2013.

[9] J. M. Beggs and D. Plenz, "Neuronal avalanches in neocortical circuits," Journal of neuroscience, vol. 23, pp. 11167-11177, 2003.
DOI: 10.1523/JNEUROSCI.23-35-11167.2003

[10] O. Shriki, J. Alstott, F. Carver, T. Holroyd, R. N. Henson, M. L. Smith, et al., "Neuronal avalanches in the resting MEG of the human brain," Journal of Neuroscience, vol. 33, pp. 7079-7090, 2013.
DOI: 10.1523/JNEUROSCI.4286-12.2013

※ 이 논문은 2018학년도에 청주대학교 산업과학연구소가 지원한 학술연구조성비(특별연구과제)에 의해 연구되었음.