한국컴퓨터정보학회 하계학술대회 논문집 제18권 제2호 (2010. 7)

사용자 검색 패턴 기반의 공공보건 시스템

박정은°, 정진영**, 박구락***
°*대전보건대학 바이오정보과
**대전보건대학 바이오정보과 교수
***공주대학교 컴퓨터공학부 교수
e-mail: love0110je@hanmail.net, jyjung@hit.ac.kr, ecgrpark@kongju.ac.kr

Public Health System Using Search Engine Query Trends

Jung-Eun Park°, Jin-Young Jung**, Koo-Rack Park***
°Dept of Bio Information, Daejeon Health Sciences College
**Dept of Bio Information, Daejeon Health Sciences College
***Dept. of Computer Science and Engineering, Kongju National University

● 요 약 ●

웹을 통해서 수천 또는 심지어는 수백만 명의 정보 수집이 가능해짐에 따라 이러한 사용자들로부터 생성된 데이터를 결합하는 알고리즘을 사용하여 새로운 비즈니스를 창출하는 집단지성이 크게 대두되고 있다. 최근 건강정보에 있어서도 웹을 통하여 사용자들이 정보의 획득이 일반화되면서 웹을 이용하는 사용자의 패턴을 이용하여 식중독이나 독감 같은 공공보건 관련 예후를 예측하는데 사용될 수 있다. 본 논문에서는 인터넷 사용자들의 검색 동향을 통해 독감의 유행을 예측하기 위해 국내외의 인플루엔자 표본감시 데이터 및 검색 동향을 비교하였다. 이러한 사용자들이 독감 관련 검색어의 증가는 실제 독감의 유행과 높은 상관관계(p=0.5, p=0.76)를 보였으며, 이는 인터넷 검색 동향만으로도 초기 단계에서 감시하고자 하는 질병의 발생 양상과 유행 양상의 전개를 예측하는데 중요한 역할을 수행할 수 있음을 의미하는 것으로 인터넷 검색 동향을 통해 공공보건을 예측하는 시스템을 제시한다.

키워드: 공공보건(Pulblic health), 집단지성(Collective intelligence), 예측(Prediction)

I. 서론

독감(Influenza)은 인플루엔자 바이러스에 의한 급성 호흡기 질환으로 전 세계에서 발생하며 전염성이 강할 뿐만 아니라 노인이나 소아 및 다른 질환을 앓고 있는 사람이 걸리면 사망률이 증가하고 합병증의 발생이 증가하는 성향을 보인다. 새로운 종류의 독감 바이러스가 짧은 시간에 넓은 지역에 유행하게 되면 젊은 사람도 사망할 수 있다. 이러한 독감은 일반 감기와는 원인균과 병의 경과가 다르기 때문에 감기와는 일반적으로 구별하고 있다[1].

정부에서는 인플루엔자 표본 감시 시스템을 통해 각 지역의 의료기관을 내원한 환자에 대한 정보를 수집하고 이를 주 1회 발표하고 있고 있다. 이러한 표본 감시는 데이터를 수집하고 보고한 후 다시 집계하는 과정에서 실제 데이터와 약 1~2주 정도의 차이가 발생하게 된다.

질병에 대한 이상 징후에 대해서 좀 더 빠른 감지를 위한 방법으로 전화를 통한 설문조사 및 해당 질병에 대한 약품의 판매 정보를 이용하는 방법들이 사용되기도 한다[3][4]. 요즘에는 인터넷을 통한 정보의 습득이 건강 및 의료 분야로 확대됨에 따라 인터넷 검색을 이용하는 방법이 사용되기도 한다[5].

미국에서는 매년 90만 명 이상의 성인들이 의료 및 건강 문제에 대한 내용을 인터넷을 통해 검색하고 있는 실정이다[6]. 이것은 웹 검색 질의가 의료 및 건강에 대한 경향을 파악하는 소스로서의 가치를 입증하는 것으로, 인터넷 검색 경향을 통해서 질병의 초기 단계에서 질병 발생의 양상을 감지하여 실제로 질병이 대량 발생하여 최종 진단 또는 확진이 내려지기 이전에 필요한 대처를 수행할 수 있다.

스웨덴에서는 스웨덴 메디컬 웹 사이트를 방문하는 방문자의 질의 형태를 분석하여 이상 징후 감시에 대한 연구를 수행하기도 하였다[7]. 캐나다에서는 검색 키워드 광고에서의 검색어와 광고를 통한 독감 이상 징후 연구를 수행하였다[13]. 그러나 이러한 방법들은 특정 지역의 특정 웹 사이트를 사용하는 사용자들에게 국한되기 때문에 검색엔진의 질의 패턴을 분석함으로써 이러한 한계를 극복하려는 움직임을 보이고 있다. 구글에서는 자신들의 축적된 검색 로그를 기반으로 검색 경향과 독감 이상 징후와의 연관성을 보려는 연구가 활발히 진행되고 있다 [8][9].

본 연구에서는 지역이나 언어에 따른 독감 관련 검색어를 추출하고 독감 관련 검색어와 해당 검색어에 대한 질의 동향을 통해 국내외의 독감 관련 검색과 실제 독감과의 연관성을 분석하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구로 변수들간의 함수관계를 규명하기 위한 회귀분석 방법과 구글 통계 분석 서비스에 대해 알아본다. 3장에서는 본 논문에서 제안하는 독감 감지 시스템을 설계한다. 4장에서는 실험 및 결과에 대하여 알아본다. 끝으로 5장에서는 결론 및 향후 연구 방향을 제시한다.

II. 관련 연구

본 장에서는 회귀분석과 상관분석 및 구글 검색 통계에 대해서 기술한다.

1. 회귀분석

회귀분석은 어떤 현상에 영향을 주고 있는 변수들 사이의 함수관계를 규명하기 위하여 관찰된 변수들에 대해 독립변수와 종속변수 사이의 인과관계에 따른 수학적 모델인 선형적 관계식을 구하고 어떤 독립변수가 주어졌을 때 이에 따른 종속변수를 예측하게 된다. 또한 이 수학적 모델이 얼마나 잘 설명하고 있는지를 판별하기 위한 적합도를 측정하는 분석 방법으로 회귀 모형이 적합한지 확인하기 위해 결정계수 R2를 사용한다. 이는 회귀모형의 독립변수가 종속변수의 변동의 몇%를 설명하고 있는지를 나타내는 지표이다. 선형회귀는 회귀분석의 하나로 종속변수 Y 및 독립변수 Xi(i=1,...,p)와 임의의 항 e와의 관계를 모델링 하며 모형은 다음과 같이 나타낸다.

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta pXp + \varepsilon$$

여기서 $\beta 0$는 절편이고, βi는 각 독립변수의 계수이며 p는 선형회귀로 추정되는 모수의 개수이다. 선형회귀는 비선형 회귀와 대비되며 종속변수가 독립변수에 대해 선형함수의 관계에 있을 것이라고 추측하게 된다.

2. 구글 검색 통계

구글 검색 통계는 구글에서 발생하는 웹 검색의 일부에 대해서 분석한 내용으로 일정기간 구글에서 실행된 총 검색 수 대비 사용자가 입력한 용어의 검색 수를 계산하여 이루어진다. 이러한 구글 검색 통계는 트래픽이 많이 발생하는 검색어에 대해서만 이루어지며 특정 용어에 대한 일정 기간 동안의 관심도를 검색량 그래프를 통해 보여주게 된다. 검색 통계에서 사용되는 데이터는 표준화 과정을 거쳐 0~100의 값으로 환산되어 표시되며, 검색량 그래프는 각 지점의 값을 최고값인 100으로 나누는 방식으로 이루어진다. 예를 들어 11월에 스웨덴에서 스키에 대한 관심도가 크게 증가했다고 가정했을 때 검색 통계에서는 이때의 최고값을 100으로 지정하고 12월에 관심도가 큰 폭으로 떨어지고 두 번째 값이 11월 최고값의 절반 수준일 경우 이 값을 50으로 지정해서 보여주게 된다[10].

III. 독감 감지 시스템의 설계

본 장에서는 독감 예측 모형 구축을 위한 독감 발병 데이터와 독감과 관련된 검색어를 선별하였다.

1. 시스템 구조

시스템은 크게 데이터 수집단계와 수집된 데이터를 기반으로 선형관계를 분석하는 단계로 나눠진다. 데이터 수집 단계에서는 독감 감시 시스템을 운영 중인 나라들의 독감 정보 데이터와 각 나라들의 독감 관련 검색어 정보를 수집하게 된다.

2. 국내 독감 데이터 수집

국내의 경우 독감 정보는 일일 표본감시결과와 주간 표본감시결과를 제공하고 있으며, 주간 표본감시결과의 경우에는 일요일~토요일까지의 한주 단위로 작성되며, 2009년 2주인 경우 2009년 1월4일 ~ 1월10일까지(일요일~토요일)의 인플루엔자 의사환자(ILI, Influenza-like illness)분율을 통해 보고된다.

인플루엔자 의사환자는 38℃ 이상의 갑작스러운 발열과 더불어 기침 또는 인후통의 증상을 보이는 것을 말하며 인플루엔자 의사환자분율은 총 진료환자에 대한 인플루엔자 의사환자에 대한 천분율로 다음과 같이 구해진다.

인플루엔자 의사환자분율 = 의사환자건수/총 진료환자건수 x 1000

인플루엔자 의사환자분율의 기본이 되는 진료환자와 인플루엔자 의사환자의 수는 전국의 680여개 보건의료기관을 통해 보고되고 있으며 주간 표본 감시에 사용되는 주간 데이터는 전국 100여개 표본감시 기관에서 수집한 데이터를 기반으로 작성되어진다. 질병관리본부에서는 지난 2005년 8월부터 주간 인플루엔자 의사환자분율 정보를 “인플루엔자 표본감시 소식지”를 통해 제공하고 있다[11].

본 연구에서는 질병관리본부에서 제공하는 데이터를 기반으로 일주일간의 인플루엔자 의사환자분율 데이터를 작성하였다. 의사환자분율 데이터의 구조는 그림 2와 같이 해당 주의 시작 날짜, 전국, 서울, 부산, 대구, 인천, 광주, 대전, 울산, 경기, 강원 충북, 충남, 전북, 전남, 경북, 경남, 제주의 16개 지역과 전국평균 의사환자분율을 콤마(,)로 구분한 CSV 형태로 작성하였다.

일반적으로 보건당국에서는 인플루엔자 유행 판단 기준을 1,000명당 인플루엔자 의사환자 2.6명을 기준으로 하고 있다[11]. 국내 독감 발생 데이터의 경우 매년 12월을 전후로 해서 가장 많은 인플루엔자 의사환자가 발생하며 2월을 기준으로 급격히 의사환자가 줄어들었다가 다시 4월에 12월의 절반 정도의 의사환자가 발생하는 경향을 보이고 있다. 이러한 뚜렷한 경향은 2007년의 경우 2월이 되어도 줄어들지 않은 것을 제외하고는 2005년부터 2009년까지 매년 반복되는 양상을 보이고 있다.

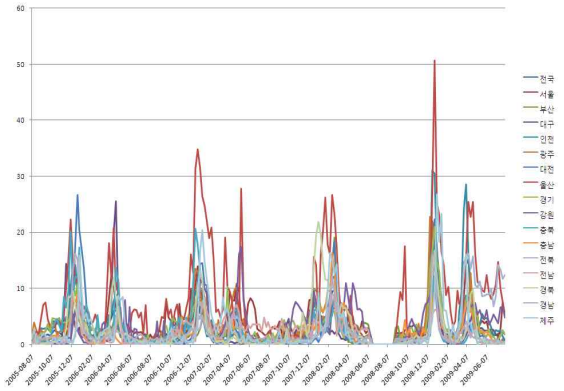

그림 1. 국내 지역별 인플루엔자 의사환자 분율(2005년 8월~2009년 6월)
Fig. 1. ILI in Korea(2005.8~2009.6)

그림 1은 2005년부터 2009년 6월까지 인플루엔자 의사환자분율을 지역별로 표시한 것으로 울산 지역이 타 지역보다 대체적으로 많은 의사환자를 보이고 있다. 그림 2에서는 국내 평균 의사환자분율 데이터와 서울, 부산의 데이터를 비교한 것으로 전국 평균과 유사한 경향을 보이는 것을 확인 할 수 있다.

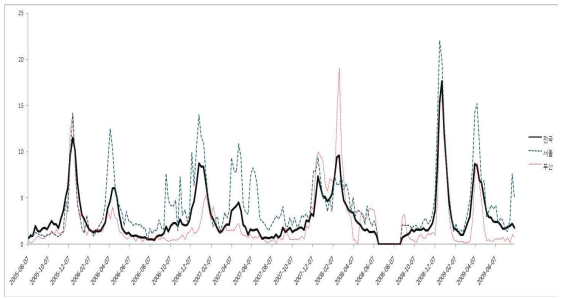

그림 2. 서울, 부산 지역과 전국 평균 데이터와의 인플루엔자 의사환자분율 비교(2005년 8월~2009년 6월)
Fig. 2. Regional Compare
(Seoul, Busan ILI and Average ILI, 2005.8~2009.6)

3. 프랑스 독감 데이터 수집

프랑스에서는 Sentinelles network을 통해서 인플루엔자 의사환자, 위장염, 수두, 천식에 대해서 주간 정보를 제공하고 있다[12]. 표 1은 2개국의 인플루엔자 의사환자분율 데이터를 보여주고 있으며, 국내의 경우 2005년 8월을 시작으로 현재까지의 데이터를 제공하고 있다.

표 1. 국가별 의사환자분율 데이터
Table 1. ILI of France and Korea

국 가 명	기 간
대한민국	2005년 8월 ~ 현재
프랑스	1984년 ~ 현재

4. 독감 관련 검색어 및 검색 동향 데이터

본 연구에서는 독감과 관련이 있는 다양한 단어에 대해서 단일 검색어 또는 단어들의 조합을 생성하였다. “감기” 단어 하나만 사용하는 경우와 “감기” 또는 “독감”의 경우 “감기 + 독감”, “감기”라는 검색에서 “조류”라는 단어를 제외하는 경우 “감기 -조류”, 2개 이상의 단어로 된 정확한 구문을 검색하기 위해서 따옴표를 사용하여 “독감 증세”로 표현하여 구글 검색통계를 사용하여 통계정보를 추출하였다.

“독감”이라는 단어의 경우 그동안 “조류 독감”이나 “돼지 독감”등의 유행으로 급작스런 검색어의 증가를 보이거나 또는 사람에게서 발생하는 독감과 관련이 없기 때문에 “-”를 이용하여 잘못된 검색어를 배제하였다. 그림 3은 “감기”라는 단일 검색어에 대해서 국내에서 발생된 검색 질의에 대하여 2004년 6월부터 2009년 6월까지의 검색 통계 정보를 보여주고 있다. 해당 단어에 대한 검색이 가장 많은 기간인 2004년 10월의 100을 기준으로 각 기간별 검색 통계 정보가 보여 진다.

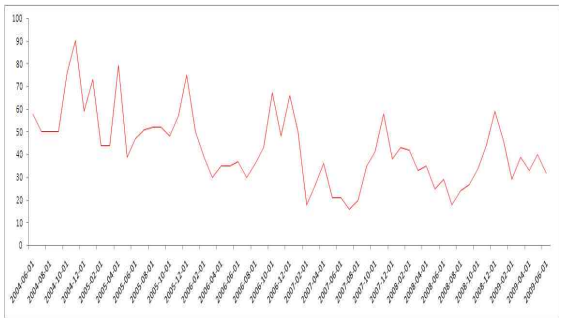

그림 3. “감기” 단일 검색어에 대한 구글 검색 통계 결과
Fig. 3. Google Trend Result of “감기”

IV. 실험 및 결과

본 장에서는 선형회귀를 통해 독감 관련 검색어와 인플루엔자 의사환자간의 상관관계를 분석하였다.

1. 독감 발생과 관련 검색어

인플루엔자 의사환자와 관련 검색어 데이터는 일주일 동안의 데이터로 인플루엔자 의사환자의 경우 천분율 데이터, 검색어 데이터는 100을 기준으로 표준화한 데이터를 각각 사용하였다. 그림 5는 국내 인플루엔자 의사환자와 검색어를 비교한 것으로 붉은색 점선의 경우 검색어 동향을 나타내며, 검은색 실선은 의사환자를 나타낸다. 그림 6은 프랑스에서의 “grippe(독감)” 검색어와 인플루엔자 의사환자를 비교한 것으로 붉은색 실선이 의사환자, 검은색 실선이 검색어를 각각 나타낸다.

그림에서도 볼 수 있듯이 각 검색어와 실제 인플루엔자 의사환자 간에는 비교적 유의한 상관성을 보이고 있으며, 실제 통계 패키지인 R을 통해서 분석한 결과 국내의 경우 0.5, 프랑스의 경우 0.76의 강한 상관관계를 보였다. 이는 통계적으로 해당 검색어가 실제 인플루엔자 의사환자를 예측하는데 있어서 유용한 지표가 됨을 의미한다.

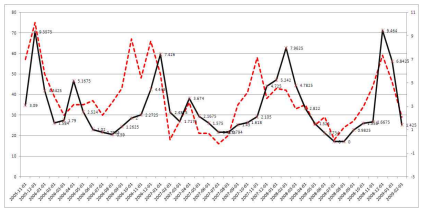

그림 4. 국내 인플루엔자 의사환자분율과 독감 관련 검색어 추이
Fig. 4. Trend of ILI and Flu relate search term

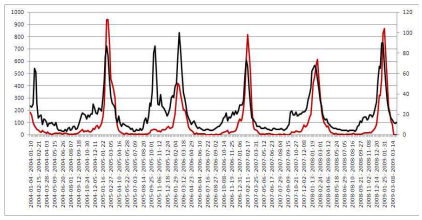

그림 5. 프랑스 인플루엔자 의사환자분율과 독감 관련 검색어 추이
Fig. 5. Relation of ILI and Search Term

2. 독감 예측 모형

국내와 프랑스 각각에 대해서 독감을 예측하기 위한 회귀공식을 유도하였으며, 해당 선형회귀를 그래프로 보면 그림 6과 같다. 프랑스가 국내보다 보다 명확한 선형회귀를 보이고 있다.

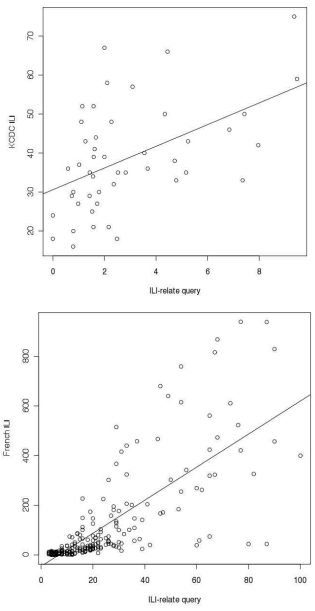

그림 6. 국내와 프랑스의 의사환자와 검색어와의 선형회귀 그래프
Fig. 6. Regression Graph of ILI and Search Term

V. 결론

본 연구에서는 국내 및 프랑스에 대해서 인플루엔자 의사환자 데이터와 독감 관련 검색어와의 상관관계를 살펴보았다. 다양한 검색어에 대한 조건을 통해서 검색어를 추출하고 이를 실제 의사환자 데이터와 비교를 통해 국내의 경우 0.5, 프랑스의 경우 0.76의 강한 상관관계를 보였다. 이는 프랑스보다 국내 인터넷 사용자들이 구글 검색을 많이 활용하지 않기 때문인 것으로 파악되며, 이러한 검색어와의 상관관계를 통해 독감을 예측하는 지표로 활용함으로써 1주일 정도 지연된 공식 독감 데이터에 비해 검색어를 기반으로 좀 더 신속한 예상 수치를 도출함으로써 독감의 출현을 조기에 감지하여 감염자 줄이기 위한 정책 수립에 도움을 줄 수 있을 것으로 기대된다. 향후 구글 이외의 검색엔진을 이용하여 보다 상관관계를 높여 정확도를 개선할 계획이다.

참고문헌

[1] 네이버 건강 검색, http://search.naver.com/
[2] 질병관리본부, http://www.cdc.go.kr
[3] Espino, J., Hogan, W. & Wagner, M. Telephone triage: A timely data source for surveillance of influenza-like diseases. AMIA: Annual Symposium Proceedings 2003. 215-219.
[4] Magruder, S. Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease. Johns Hopkins University APL Technical Digest 24, 2003. 49-353.
[5] Fox, S. Online Health Search 2006. Pew Internet & American Life Project. 2006.
[6] Johnson, H. et al. Analysis of Web access logs for surveillance of influenza. MEDINFO. 2004. 1202-1206.
[7] Hulth, A., Rydevik, G. & Linde, A. Web Queries as a Source for Syndromic Surveillance. PLoS ONE 4(2): e4378. doi:10.1371/journal.pone.0004378, 2009.
[8] Polgreen, P. M., Chen, Y., Pennock, D. M. & Forrest, N. D. Using internet searches for influenza surveillance. Clinical Infectious Diseases 47, 2008. 1443-1448.
[9] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski & Larry Brilliant, Detecting influenza epidemics using search engine query data, Nature Vol 457, 2009, 1012-1014.
[10] 구글 검색 통계, http://google.com/insights/search/
[11] Public Health Weelky Report, KCDC, Korean Influenza Surveillance Report, 2008.
[12] Sentinelles, http://sentiweb.fr
[13] G Eysenbach, Infodemiology: Tracking Flu-Related Searches on the Web for Syndromic Surveillance, AMIA Annual Symposium Proceedings 2006, 244-248