Korean J Gastroenterol Vol. 63 No. 6, 361-365
http://dx.doi.org/10.4166/kjg.2014.63.6.361
pISSN 1598-9992 eISSN 2233-6869

ORIGINAL ARTICLE

KJG

빅데이터 시각에서 본 단일기관의 위장관 질환 블로그 분석

최정란, 박효진, 이충현[1]
연세대학교 의과대학 강남세브란스병원 내과학교실, 홍보팀[1]

Analysis of a Blog for Gastrointestinal Disease in the View Point of the Big Data: A Single Institutional Study

Jungran Choi, Hyojin Park and Choong-Hyun Lee[1]
Department of Internal Medicine and Public Relations Team[1], Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, Korea

Background/Aims: With the enormous increase in the amount of data, the concept of big data has emerged and this allows us to gain new insights and appreciate its value. However, analysis related to gastrointestinal diseases in the viewpoint of the big data has not been performed yet in Korea. This study analyzed the data of the blog's visitors as a set of big data to investigate questions they did not mention in the clinical situation.
Methods: We analyzed the blog of a professor whose subspecialty is gastroenterology at Gangnam Severance Hospital. We assessed the changes in the number of visitors, access path of visitors, and the queries from January 2011 to December 2013.
Results: A total of 50,084 visitors gained accessed to the blog. An average of 1,535.3 people visited the blog per month and 49.5 people per day. The number of visitors and the cumulative number of registered posts showed a positive correlation. The most utilized access path of visitors to the website was blog.iseverance.com (42.2%), followed by Google (32.8%) and Daum (6.6%). The most searched term by the visitors in the blog was intestinal metaplasia (16.6%), followed by dizziness (8.3%) and gastric submucosal tumor (7.0%).
Conclusions: Personal blog can function as a communication route for patients with digestive diseases. The most frequently searched word necessitating explanation and education was 'intestinal metaplasia'. Identifying and analyzing even unstructured data as a set of big data is expected to provide meaningful information. **(Korean J Gastroenterol 2014;63:361-365)**

Key Words: Big data; Blogging; Metaplasia; Gastrointestinal diseases

서 론

우리는 어느 때보다 많은 데이터에 근거하여 결정을 내리는 세상에 살고 있다. 사람들은 매일 250경 바이트의 데이터를 생성하고 있으며, 지난 2년 동안에만 현재 전 세계에 존재하는 데이터의 90%가 생성되었다.[1] 이렇게 최근 들어 검토할 데이터 양이 커짐에 따라 큰 규모를 활용하여 과거에는 불가능했던 새로운 통찰이나 가치를 추출해 내는 빅데이터라는 개념이 등장하였다. 현재 이 빅데이터를 이용하여 공공 부문에서 다양한 활용이 시도되고 있으며[2] 국외에서는 빅데이터를 활용하여 구글이 겨울철 미국에서 독감의 확산을 예측할 수 있다는 연구[3]와 워싱턴 D.C.의 Medstar 워싱턴 병원 센터에서 환자의 입원 당시의 정신적 증상과 재입원 가능성의 상관성을 찾은 연구[4] 등 의료 부문에서도 다양하게 활용이 되고 있다. 국내에서도 social media data로 자살률을 예측한다는 연구[5]가 보고되었으나 아직까지 소화기 질환에서 빅데이터 관

Received February 4, 2014. **Revised** April 17, 2014. **Accepted** April 25, 2014.

교신저자: 박효진, 135-720, 서울시 강남구 도곡동 언주로 211, 강남세브란스병원 소화기내과
Correspondence to: Hyojin Park, Department of Internal Medicine, Gangnam Severance Hospital, 211 Eonju-ro, Gangnam-gu, Seoul 135-720, Korea. Tel: +82-2-2019- 3318, Fax: +82-2-3463-3882, E-mail: hjpark21@yuhs.ac

Financial support: None. Conflict of interest: None.

점의 분석은 보고되지 않았다. 따라서 이번 연구는 빅데이터의 시각에서 위장관질환 개인 블로그의 방문자의 데이터를 분석하여 새로운 가치를 발견해 보고자 하였다.

대상 및 방법

연세대학교 강남세브란스병원의 베스트닥터 개인 블로그 중 방문자수와 블로그 게시글의 수가 가장 많았던(Table 1) 소화기내과 교수의 개인 블로그를 대상으로, 2011년 1월부터 2013년 12월까지 접속한 50,084명의 방문자에 대해 방문수의 추이와 블로그 접속 경로, 그리고 방문자들이 블로그 내에서 검색한 단어를 분석하였다. 블로그의 게시글은 총 84개로 프로필, 언론보도, 시와 수필, 소화기질환의 이해, 환자 경험으로 구성 되었다(Table 2). 그 중 소화기질환의 이해가 37개(44.0%)로 가장 많았는데 소화기질환의 이해는 식도, 위, 대장, 기타로 세분화되어 다양한 질환에 대해서 다루고 있었다(Fig. 1). 환자 경험은 외래 혹은 입원 환자 중 환자 입장에서 치료받은 경험을 소개한 글을 정리하였다. 방문자들이 블로그에서 검색한 단어의 분석은 같은 의미의 질환이지만 다른 단어로 검색한 경우, 한 질환에 대해서 세분화된 내용을 검색한 경우 한 종류로 해석하였고 블로그 프로그램의 한계로 2013년 10월 3일부터 2013년 12월 9일까지의 방문자들이 검색한 단어의 누적만이 가능하여 이 기간 검색어의 빈도를 분석하였다. 통계는 빈도 분석과 상관 분석을 이용하였으며 IBM SPSS Statistics 프로그램(version 19.0; IBM Co., Armonk, NY, USA)을 사용하여 $p<0.05$일 때 유의하다고 정의하였다.

Table 1. Blog Statistics (Based on September 2013)

Blog name	Visitor (n)	Post (n)
Blog 1[a]	41,935	79
Blog 2	4,118	3
Blog 3	3,578	4
Blog 4	22,804	42
Blog 5	20,385	12
Blog 6	27,291	23
Blog 7	12,288	54
Blog 8	5,870	6
Blog 9	13,509	14
Blog 10	13,039	28
Blog 11	6,215	3
Blog 12	12,905	46
Blog 13	860	1

[a]This blog was analyzed in our study.

결 과

방문한 총 방문자수는 50,084명으로 월 평균수는 1,535.3명, 일 평균수는 49.5명이었으며 누적된 게시글의 수와 월 방문자수는 양의 상관성을 보이는 것을 확인하였다($p<0.01$) (Fig. 2). 블로그의 방문자 유입 경로로는 blog.iseverance.com이 가장 많았고(42.2%), 그 다음으로는 www.google.co.kr (32.8%), www.daum.net (6.6%) 순이었다(Table 3). 2013년

Table 2. Blog Contents

Category	Post (n)
Profile	1 (1.2)
Press release	17 (20.2)
Poetry and essays	19 (22.6)
Understanding of digestive disease	37 (44.0)
Esophagus	7 (8.3)
Stomach	13 (15.5)
Colon	14 (16.7)
Others	3 (3.6)
Patients experience	10 (12.0)

Values are presented as n (%).

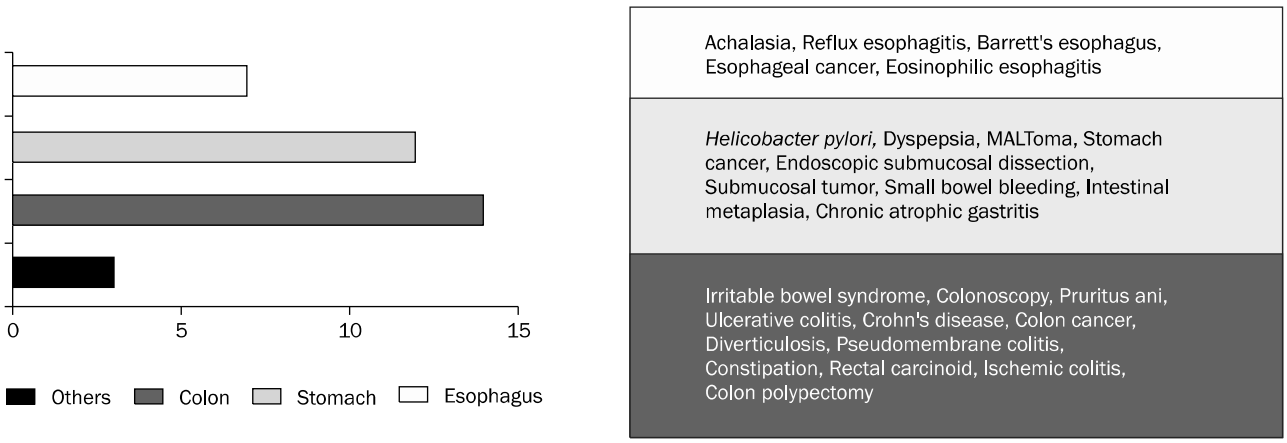

Fig. 1. Understanding of digestive disease. There are various posts about digestive disease in blog.

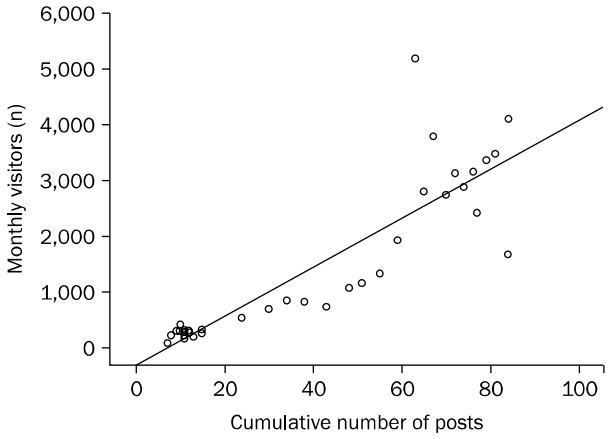

Fig. 2. The correlation between accumulation of posts and the number of monthly visitors. Two variables show a positive correlation, R=0.878 ($p < 0.01$).

Table 3. Ranking on the Access Path of Visitors

	Site	Access path (n)
1.	blog.iseverance.com	10,772 (42.2)
2.	www.google.co.kr	8,356 (32.8)
3.	search.daum.net	1,681 (6.6)
4.	search.naver.com	1,098 (4.3)
5.	www.internetsupervision.com	776 (3.0)
6.	m.search.naver.com	521 (2.0)
7.	gs.iseverance.com	487 (1.9)
8.	www.facebook.com	360 (1.4)
9.	image.postman.co.kr	251 (1.0)
10.	www.iseverance.com	191 (0.7)

Values are presented as n (%).

10월 3일부터 2013년 12월 9일까지 방문자들이 블로그 내에서 검색한 단어는 종류는 502가지이며 단어 검색 건수는 1,339건으로, 이 중 장상피화생이 222건(16.6%)으로 가장 많았고 다음으로는 어지러움 증상 111건(8.3%), 위점막하종양 94건(7.0%) 순이었다(Table 4).

고 찰

지금 세계는 빅데이터로 인해 변화하고 있다. 현재는 정보기술(information technology)이 일상화된 스마트 시대로 국외에서는 경제, 산업, 사회, 행정, 스포츠, 의료, 선거운동까지 다양하게 빅데이터가 활용되고 있다.[6] 국내에서도 공공 부문의 빅데이터 분석과 활용에 대한 수요가 크게 증가하고 있으나 현재 국내 사례 및 가이드가 많지 않으며 소개된 방법론은 대부분 분석 업체의 방법론 또는 기술 소개로, 실제로 활용하기에는 지나치게 기술적인 단점이 있어 국외에 비하여 현저히 빅데이터 활용도가 낮다.[2] 이번 연구에서는 World Wide Web

Table 4. Ranking on Searching Words in the Blog

	Query	Searching word (n)
1.	Intestinal metaplasia	222 (16.6)
2.	Symptom of dizziness	111 (8.3)
3.	Gastric submucosal tumor	94 (7.0)
4.	Diverticulitis/ Colonic diverticulosis	82 (6.1)
5.	Carcinoid tumor	57 (4.3)
6.	Mucosa-associated lymphoid tissue (MALT) lymphoma	30 (2.2)
7.	Crohn's disease	27 (2.0)
8.	Barrett's esophagus	20 (1.5)
9.	Dyspepsia	15 (1.1)
10.	Achalasia	9 (0.7)

Values are presented as n (%).

에 게시되는 토론과 정보 사이트인 블로그를 활용하여 빅데이터 분석을 시도하였는데, 블로그가 다른 정적인 웹 사이트와 구별되는 점인 방문자가 블로그 내에서 자신의 의견을 남기고 주고 받을 수 있다는 특징을 활용하였다.[7-9]

국내 인터넷 순위에서 i세브란스가 건강/의학 사이트의 종합병원 내에서 점유율 5위 이상을 유지하는 것과 이번 연구에서 블로그로 접속하게 되는 유입경로로 blog.iseverance.com이 가장 많은 것을 보았을 때 환자들에게 블로그에 대한 적극적인 홍보가 필요하며, 블로그 게시글의 수와 방문자수가 양의 상관관계를 보인 점으로 미루어 블로그의 활성화가 필요할 것으로 판단된다.[10]

이번 연구에서는 프로그램의 한계로 2013년 10월 3일부터 2013년 12월 9일까지의 38일 동안 블로그 방문자들이 블로그 내에서 검색한 단어의 추적이 가능하였다. 이 기간에 502가지의 종류의 검색어와 1,339건의 검색 건수로 하루 평균 35.2건의 검색이 이루어졌고, 하루 평균 방문자수가 49.5명임을 고려하면 최소 3명 중 2명은 블로그 내에서 검색을 한다는 것을 의미한다. 이 중 장상피화생이 가장 많은 것은 최근 위내시경 검사 건수가 증가하면서 본원에서 장상피화생을 진단받은 환자수가 증가했다는 것이 이유가 될 수 있으나, 상대적으로 다른 질환과 비교하여 인터넷 상에 장상피화생에 대한 정보가 부족한 것도 한 이유로 꼽을 수 있겠다.

이번 연구와 비슷한 연구로 Harsha 등[11]은 시간과 지역에 따른 '정맥류 치료' 검색의 수의 추이를 분석, 계절과 지역의 차이를 나타내어 의료 시술에 대한 인터넷 검색어의 분석이 제한된 자원과 마케팅의 비용에서 효율적인 사용이 될 수 있음을 제시하고 있다. 이는 이번 연구에서도 임상 진료에서는 알 수 없었던 환자들의 실제 의문사항이나 질환의 빈도를 다른 도구 없이도 알 수 있었던 것과 공통되는 점으로, 추후 블로그의 프로그램을 개선하여 더 세부적인 정보를 알아 낼 수 있으리라 판단된다.

빅데이터의 정의는 아직 정립되지 않았으나 volume (규모)이 방대하고 variety (종류)가 다양하며, 여러 종류의 데이터가 융합되고, velocity (속도)는 자료의 수집-처리-분석/예측을 한번에 처리하여 value (가치)를 추출하고 발견하는 것으로 설명된다.[12] 또한 저명한 애널리스트인 Donald Feinberg는 빅데이터는 새로운 것이 아니고 이전부터 존재했던 데이터를 모은 것에 불과하며 다만 보유한 데이터 중 가치있는 부분을 발견하고 분석하는 데에 집중하는 것이 중요하다고 지적한 바 있다.[13] 이번 연구에서도 블로그에 방문자들로 인해 발생된 버려질 수 있는 비정형의 데이터를 분석하여 그들이 궁금해 하는 것이 무엇인지 분석할 수 있었고, 이를 활용해 앞으로 환자와의 의사소통에도 기여할 수 있을 것으로 기대된다.

이번 연구의 제한점으로는 첫째, 블로그 자체 프로그램의 한계로 인하여 블로그 방문자가 어떤 게시글을 열람하였는지 알 수 없었던 점과, 둘째, 블로그가 방문자들의 개인 정보를 등록하지 않기 때문에 방문자들의 기본적 특징을 파악할 수 없었다는 점, 셋째, 개인정보 보호의 문제로 블로그에 접속한 개인 컴퓨터 internet protocol 주소를 추적할 수 없어 처음 방문자와 재방문자의 구분을 할 수 없고 방문자들의 블로그 체류 시간을 분석할 수 없었던 점으로, 현재 빅데이터에서 문제되고 있는 개인정보 보호와 관련한 문제가 이번 연구에서도 나타나고 있었다.

결론으로는 첫째, 개인 블로그가 소화기질환 환자들의 의사소통의 장으로서 기능을 할 수 있겠다. 두 번째로 이번 연구를 통하여 '장상피화생'에 대한 환자들의 많은 관심이 있음을 확인하여 향후 이에 대한 설명과 교육이 필요할 것이라는 정보를 얻을 수 있었다. 마지막으로 이번 연구는 개인 블로그에서도 빅데이터 관점에서의 분석이 가능하다는 것을 시사한다. 하지만 앞으로 더 오랜 기간의 연구와 다른 병원이나 타 블로그의 연계를 통한 더 큰 데이터가 필요하며, 이는 개인정보 보호를 염두에 두고 이를 침해하지 않는 범위 내에서 분석되어야 할 것이다.

요 약

목적: 최근 등장한 빅데이터의 시각에서 위장관질환 개인 블로그의 방문자들을 데이터화하여 경향을 분석하고 임상에서는 알 수 없었던 환자들의 실제 의문 사항들을 알아보고자 하였다.

대상 및 방법: 연세대학교 강남세브란스병원 소화기내과 교수의 개인 블로그에 2011년 1월 1일부터 2013년 12월 9일까지 접속한 방문자의 수, 방문자의 유입 경로, 블로그에서 검색한 단어를 분석하였다. 블로그의 게시글은 총 84개로 프로필 이외에 언론보도, 내 마음의 행로, 소화기질환, 고객경험으로 분류되어 있었다.

결과: 총 블로그의 방문자 수는 50,084명으로 월 평균 1,535명이 방문하였고 하루 평균은 50명이 방문하였으며, 등록된 게시물의 누적수와 방문자수는 양의 상관 관계를 보였다. 방문자들의 유입 경로로는 i세브란스 베스트 닥터(http://blog.iseverance.com)가 가장 많았고(42.2%), 그 다음으로는 구글이었다(32.8%). 블로그 내에서 가장 많이 검색된 검색어는 장상피화생이었으며(16.6%), 다음으로 어지러움 증세(8.3%), 위점막하종양(7.0%) 순이었다.

결론: 블로그의 방문 경로와 검색어를 분석한 결과 개인 블로그는 소화기질환 환자들에게 의사 소통의 장으로서의 기능을 하며, 가장 많이 검색되었던 장상피화생에 대한 설명과 교육이 필요하리라 판단된다. 방대한 양의 비정형적인 데이터라 할지라도 빅데이터의 관점에서 데이터의 경향을 파악하고 분석하는 것이 의료와 환자 간의 소통에 유용한 자료를 제공하리라 기대된다.

색인단어: 빅데이터; 블로깅; 화생; 위장관질환

REFERENCES

1. What is big data? Bringing big data to the enterprise. [Internet]. Armonk (NY): IBM [cited 2014 Jan 10]. Available from: http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html
2. Guideline for analysis of big data. [Internet]. Seoul: National Information Society Agency; 2013 Jan 25 [cited 2014 Jan 10]. Available from: http://www.nia.or.kr/bbs/board_view.asp?BoardID=20111128132107445 8&id=10343&Order=010200&search_target=&keyword=&Flag=010000&nowpage=2&objpage=0
3. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009;457:1012-1014.
4. Microsoft expands presence in healthcare IT industry with acquisition of health intelligence software Azyxxi. [Internet]. Redmond (WA): Microsoft Press; 2006 Jul 26 [cited 2014 Jan 10]. Available from: http://www.microsoft.com/en-us/news/press/2006/jul06/07-26azyxxiacquisitionpr.aspx
5. Won HH, Myung W, Song GY, et al. Predicting national suicide numbers with social media data. PLoS One 2013;8:e61809.
6. An analysis of data for a better future-big data case study in the developed nation. [Internet]. Seoul: National Information Society Agency; 2013 Apr 16 [cited 2014 Jan 10]. Available from: http://www.nia.or.kr/bbs/board_view.asp?BoardID=20111128132107445 8&id=10764&Order=010200&search_target=&keyword=& Flag=010000&nowpage=2&objpage=0
7. Weblogs: a history and perspective. [Internet]. San Francisco (CA): Rebecca's Pocket; 2000 Sep 7 [cited 2014 Jan 10]. Available from: http://www.rebeccablood.net/essays/weblog_history.html

8. Mutum D, Wang Q. Consumer generated advertising in blogs. In: Eastin MS, Daugherty T, Neal M, eds. Handbook of research on digital media and advertising: user generated content consumption. Hershey: Information Science Reference, 2010.
9. Gaudeul A, Peroni C. Reciprocal attention and norm of reciprocity in blogging networks. Jena economic research papers. 2010-020. Jena: Jena University, 2010.
10. Ranking in sites of general hospital. [Internet]. Seoul: MediaChannel [cited 2014 Jan 10]. Available from: http://www.rankey.com/rank/rank_site_cate.php?cat1_id=1&cat2_id=24&cat3_id= 253
11. Harsha AK, Schmitt JE, Stavropoulos SW. Know your market: use of online query tools to quantify trends in patient information-seeking behavior for varicose vein treatment. J Vasc Interv Radiol 2014;25:53-57.
12. Big data analytics. [Internet]. Renton (WA): TDWI; 2011 Sep 14 [cited 2014 Jan 10]. Available from: http://tdwi.org/research/2011/12/sas_best-practices-report-q4-big-data-analytics.aspx?tc=page0
13. In next year the big data bubble will fall. [Internet]. Seoul: DigitalTimes; 2013 Oct 20 [cited 2014 Jan 10]. Available from: http://www.dt.co.kr/contents.html?article_no=2013102102010860718002