공공데이터를 기반으로 하는 데이터 가시화의 활용방법

박선희*, 이희만**, 이정배***, 배종환****
*아이에이치테크, **서원대학교 멀티미디어학과,
***부산외국어대학교 컴퓨터공학과, ****공주대학교 군사정보학과
e-mail:sadal@hanmail.net

Reinterpreting and utilizing data visualization based on public data

Seon-Hui Bak*, Hee-Man Lee**, Jeong-Bae Lee***, Jong-Hwan Bae****
*IHTECH, **Dep. of Multimedia SeoWon University,
***Dept. of Computer Engineering., Busan University of Foreign Studies
****Dept. Military Science and Information, Kongju University

요 약

IoT, 소셜미디어, 스마트 폰, 웨어러블 기기의 등장함에 따라 발생하는 데이터가 폭발적으로 증가해 바야흐로 “빅 데이터” 시대가 다가왔다. 이에 정부와 기업에서는 빅 데이터를 효율적으로 사용하기 위한 정책을 추진하고 전문 인력 양성에 힘쓰고 있다. 그 중 빅 데이터를 이용한 시각화는 빠른 의사결정을 도와주고, 자료로부터 데이터를 얻는 시간을 단축하고 즉각적인 상황판단이 가능해지는 등 다양한 장점을 가지고 있다. 그러나 무수히 많은 데이터 중 공공데이터를 활용한 시각화에 관한 연구는 현재까지 잘 이루어지지 않고 있다. 따라서 본 논문에서는 공공데이터를 기반으로 데이터 가시화의 활용방법에 대해 제안한다.

1. 서론

1.1 빅데이터와 공공데이터

최근 IoT(Internet of Things), 소셜미디어, 스마트 폰, 웨어러블 기기의 확산에 따라 발생하는 데이터가 폭발적으로 증가하고 있다. 이에 따라 정부와 기업에서는 효율적인 빅 데이터 활용을 위해 전문 인력 양성과 기술개발에 힘쓰고 있다. 그러나 개인정보, 사생활침해, 저작권 등의 문제로 빅 데이터를 활용하기 위해서는 아직 해결해야할 과제들이 많다[1].

현재 빅 데이터의 시대가 도래해옴에 따라 빅 데이터의 정의 또한 다양해지고 있다. 빅 데이터란 디지털 환경에서 생성되는 데이터로, 그 규모가 매우 방대하다. 또 생성주기가 짧고, 형태가 다양하다. 빅 데이터는 크게 3V로 나눌 수 있다. 첫째, 볼륨(Volume)이다. 볼륨은 데이터의 양으로, 사용하지 않고 관리하지 않았던 데이터를 분석함으로써 새로운 지식이나 비즈니스가 되는 것이다. 둘째, 데이터 입출력 속도(Velocity)이다. 빅 데이터의 입출력 속도란, 데이터가 생성, 저장 되며 데이터의 가시화가 되는 과정이 어느 속도에 이뤄졌는가에 대한 정의이다. 셋째, 다양성(Variety)이다. 앞서 말했듯이 데이터에는 수많은 데이터가 있는데, 대표적으로 크게 정형 데이터와 비정형 데이터로 구분된다. 정형 데이터란 고정된 필드에 저장되는 것을 의미하며, 일정한 형식을 갖추고 있기 때문에 데이터베이스와 스프레드시트 등을 예로 들 수 있다. 비정형 데이터는 고정된 필드에 저장되어 있지 않은 데이터를 의미한다. SNS에 올라오는 영상과 페이스북, 이미지 파일, 음원파일 등을 예로 들 수 있다. 수많은 형태의 데이터들 중에서도 비정형 데이터는 빅데이터에서 2/3를 차지하고 있다고 해도 과언이 아니다. 즉, 비정형 데이터의 정확한 정보 추출을 위해서는 기술력 확보가 중요하다.

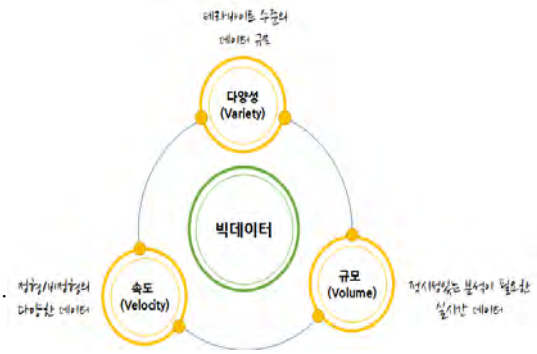


그림 1. 빅 데이터의 3V

수많은 데이터들 중 공공데이터(Public Data)는 정부 또는 공공기관의 업무과정에서 발생한 데이터를 의미한다. 따라서 공공데이터는 상대적으로 가치가 높지만 이를 분석의 용도로 활용하지 못하고 있다는 한계가 있다.

1.2 데이터 가시화

빅 데이터의 활용 방법은 대표적으로 데이터 가시화가 있다. 데이터 가시화란, 데이터 분석 결과를 쉽게 이해 할 수 있도록 시각적으로 표현한 것으로, 도표(graph)라는 수

단을 통해 데이터를 명확하고 효과적으로 전달하는 것을 의미한다. 데이터 시각화를 통해 자료로부터 데이터의 정보를 얻는 시간을 단축하고, 즉각적인 상황 판단이 가능하여 빠른 의사결정이 가능하다.

대표적인 데이터 가시화 방법으로는 인포그래픽이라는 기술이 있다. 인포그래픽은 그림 2와 같이 중요한 정보를 한 장의 그래픽으로 표현해 사람들이 해당 정보를 쉽고 빠르게 이해할 수 있도록 만드는 그래픽 메시지이다[2].

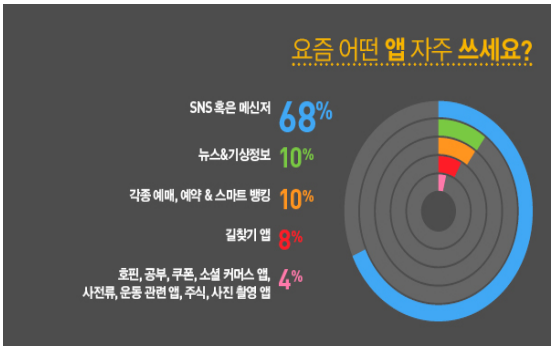


그림 2. 인포 그래픽스 이용한 데이터 시각화

본 논문에서는 공공데이터의 효율적인 분석과 빠른 의사결정을 위한 데이터 시각화 방법에 대해 제안한다.

Spotfire 솔루션을 기반으로 정부 3.0 보건의료분야에서 제공하는 공공데이터인 암 발생률을 활용하여 가시화를 구현하였다.

3. 본론

빅 데이터를 통한 분석은 현재 발생하는 다양한 문제들의 원인 분석이 가능하며 신속한 의사결정이 가능하다. 이를 위해 빅 데이터의 시각화의 중요성은 점점 더 고도화되고 있다. 빅 데이터의 시각화는 데이터를 기반으로 객관적 표현에 더 초점을 맞추는 경우가 많다. 그렇기 때문에 정보형 메시지를 전달하여 데이터 시각화 작업을 하는 경우가 많고, 데이터를 기초로 해석된 의미의 설득형 메시지를 전달하기 위한 경우에는 인포그래픽을 사용하는 경우가 많다.

본 논문에서는 데이터 시각화를 위해 Spotfire의 솔루션을 사용했다. Spotfire의 주요 이점은 정보 분석 및 의사결정을 시속하게 마무리 한다. 어떤 비즈니스 프로세스 및 데이터에 대해서도 공통적인 다이나믹한 분석 환경을 가능하게 한다. 또 Spotfire에서는 다양한 차트들을 제공한다는 점에서 강점을 갖는다[6].

기본적인 차트로는 막대그래프, 선 그래프가 있다. 막대그래프는 데이터의 여러 범주에 대한 값을 비교할 수 있으며 선 그래프는 대부분 시간별 추세를 표시하는데 사용한다. 콤비네이션 차트는 단일 시각화에서 막대와 선 모두를 표시하는 옵션을 가지고 있다. 오버레이 효과로 인해 막대 상단에 선이 그려지므로, 데이터의 여러 칼럼이나 범주에 대한 값을 쉽게 비교할 수 있다.

파이 그래프는 섹터로 분할된 원형 그래프이다. 이 그래프는 상대적 기준에서 데이터의 여러 범주에 대한 값을 비교하는 데 사용한다. 각 파이 섹터는 특정 범주를 나타내고, 해당 크기는 전체 값에서 범주가 차지하는 부분을 백분율로 표시하여 나타낸다.

Spotfire에서는 다소 생소한 차트들도 제공하는데 그 예로는 산점도 좌표계와 트리맵, 평행 좌표 그래프가 있다.

산점도 좌표계의 산점도는 표식이 2차원 좌표계로 표시되며, 이는 두 차원에서 데이터가 분포되는 방식의 개요를 얻는데 유용하다. 트리맵은 계층적으로 구조화 할 수 있는 대용량 데이터를 표시하는 데 사용된다. 트리맵은 다르게 크기가 지정되고 색상이 지정된 사각형을 사용하여 표시한다. 평행 좌표 그래프는 단일 시각화 내에서 유형이나 크기가 완전히 다른 데이터 값을 비교하는 데 사용되며 패턴 검사에도 유용하다. 그림 3에서 설명하는 평행좌표계 그래프의 내용은 높은 SATA평균 점수를 기록할수록, SAT 응시자 비율이 낮은 것을 한눈에 볼 수 있다. 이렇듯 평행좌표 그래프는 여러 변인들의 상관관계를 한 화면에 볼 수 있다는 점에 비해 많은 데이터가 있으면 복잡해져 보기 어렵다는 문제점이 있다[4].

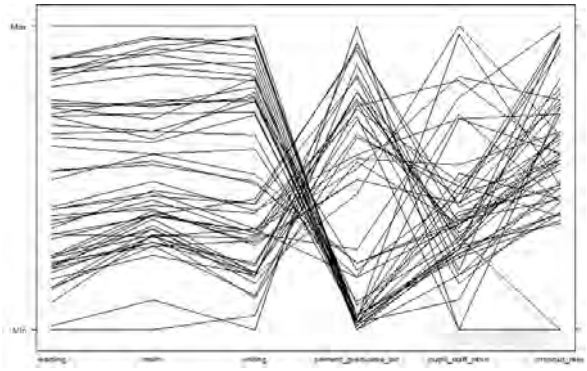
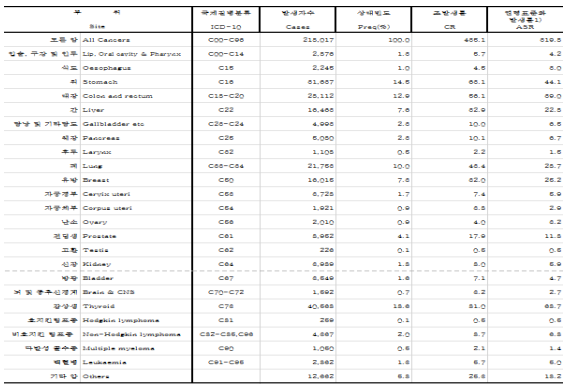


그림 3. 평행좌표계 그래프

본 논문에서 시각화할 공공데이터는 그림 4와 같다. 본 자료는 정부 3.0 홈페이지에서 Excel 파일을 직접 다운로드 받을 수 있다.



부위 Site	국제질병분류 ICD-10	발생자수 Cases	상대빈도 Prev(%)	조발생률 CR	인구표준화 발생률1 ASR
모든 암 All Cancers	C00-C96	215,017	100.0	485.1	319.9
입술, 구강 및 인두 Lip, Oral cavity & Pharynx	C00-C14	2,576	1.5	5.7	4.2
식도 Oesophagus	C15	2,246	1.0	4.5	3.0
위 Stomach	C16	61,667	14.5	65.1	44.1
대장 Colon and rectum	C18-C20	25,112	12.9	56.1	39.0
간 Liver	C22	16,466	7.6	32.9	22.5
담낭 및 기타담도 Gallbladder etc	C23-C24	4,996	2.6	10.0	6.5
췌장 Pancreas	C25	5,050	2.6	10.1	6.7
후두 Larynx	C32	1,105	0.6	2.2	1.6
폐 Lung	C33-C34	21,766	10.0	45.4	25.7
유방 Breast	C50	16,015	7.6	32.0	26.2
자궁경부 Cervix uteri	C53	3,725	1.7	7.4	6.9
자궁체부 Corpus uteri	C54	1,421	0.9	6.8	2.9
난소 Ovary	C56	2,010	0.9	4.0	3.2
전립선 Prostate	C61	8,962	4.1	17.9	11.5
고환 Testis	C62	226	0.1	0.5	0.5
신장 Kidney	C64	6,959	1.5	8.0	5.9
방광 Bladder	C67	3,549	1.6	7.1	4.7
뇌 및 중추신경계 Brain & CNS	C70-C72	1,392	0.7	6.2	2.7
갑상선 Thyroid	C73	40,668	13.6	81.0	68.7
호지킨림프종 Hodgkin lymphoma	C81	259	0.1	0.5	0.5
비호지킨 림프종 Non-Hodgkin lymphoma	C82-C85,C96	4,667	2.0	8.7	6.8
다발성 골수종 Multiple myeloma	C90	1,060	0.5	2.1	1.4
백혈병 Leukaemia	C91-C95	2,982	1.5	6.7	5.0
기타 암 Others		12,662	6.5	26.6	19.2

그림 4. 공공데이터 보건의료분야 정부3.0 암 발생률 자료

Spotfire의 가시화에 앞서 엑셀파일을 Spotfire에 import 한다. [add data tables]-[add] - [fire]를 클릭하여 자신이

import하고 싶은 데이터를 찾아 클릭한다.

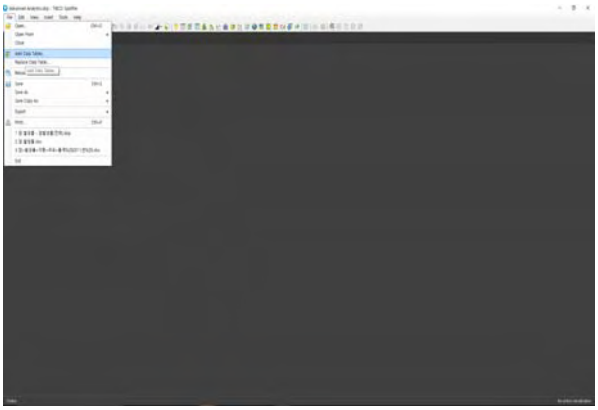


그림 5. Spotfire에서 import 하기

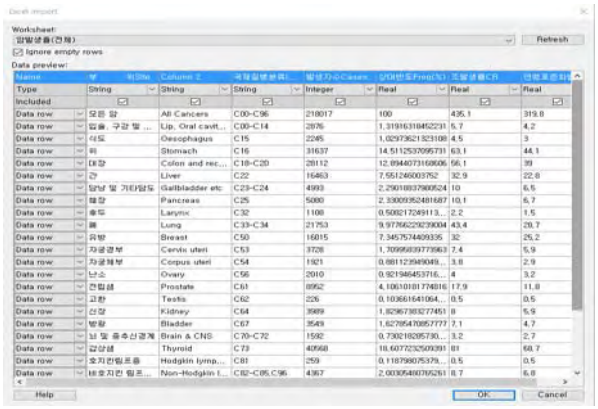


그림 6. 엑셀 데이터의 테이블을 선택할 수 있는 화면

Spotfire는 자신이 import한 자료에서 필요하지 않은 자료를 가시화하지 않도록 테이블을 선택할 수 있다. 본 논문에서는 불필요한 전체 암 발생률과 국제질병분류 Column은 ignore를 체크하여 비활성화 하였다.

테이블을 활성화 시키면 그림 7과 같은 테이블에 맞는 시각화를 권한다. 사용자가 직접 시각화 할 필요 없이 원하는 차트를 선택하면 자동적으로 시각화를 실행한다.

그림 7. Spotfire Recommended visualizations 화면

그림 8은 본 논문에서 사용한 데이터를 트리맵 방식을 통해 시각화한 것이다. 그림 4와 같은 엑셀 데이터는 어떤 암이 제일 많이 발생하고, 위험한지를 한눈에 가늠하기 어려운 반면, 그림 8과 같이 데이터가 시각화 된다면 한 눈에 알아 볼 수 있고, 병원에서는 의사결정과 행동에 있어서 큰 도움이 될 것이다.

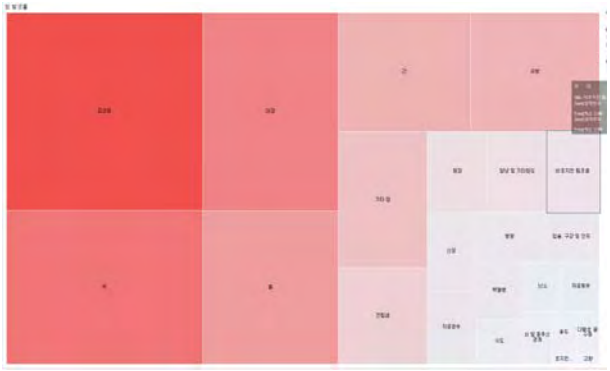


그림 8. 엑셀 데이터를 tree map 형식으로 시각화

그림 9는 남자 암 발생률 및 여자 암 발생률의 차이점을 통해 성별에 따른 암 발생률이 어떻게 다른지도 한눈에 파악 할 수 있는 연구 결과이다.

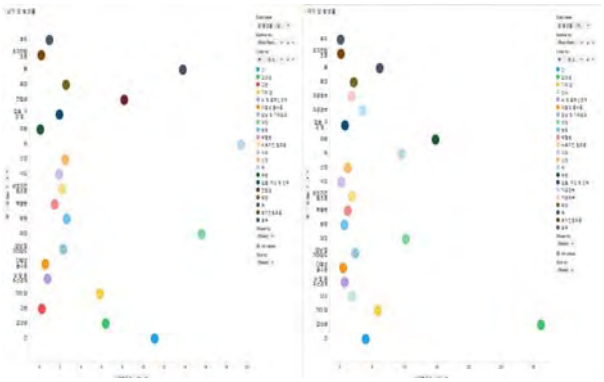


그림 9. 남녀 암 발생률 비교 산점도 그래프

4. 결론

현재 까지 본 논문에서는 정부 3.0 보건 의료 분야에서 제공하는 '암 발생률'에 관한 데이터를 활용하여 가시화하는 방법에 대해서 제안하였다. 국내의 인터넷이 활성화 되었던 2000년대부터 방대해지는 데이터들을 어떻게 처리해야 할지는 오늘날에도 하나의 과제로 남아있다. 또 최근 전 세계적으로 스마트폰의 사용이 활성화됨에 따라 소셜 네트워크 서비스가 많아지면서 정형화데이터뿐만 아니라 반 정형, 비정형 데이터도 막대해지고 있는 추세이다. 그렇기 때문에 데이터의 시각화의 중요성은 더욱 고도화 되고 있다. 이에 대하여 빅 데이터의 가치를 높일 수 있도록 가시화의 방식과 차트들의 특징을 종합적으로 분석하고, 이를 기반으로 한 데이터를 통해 사용자들이 직관적이면서 한 눈에 파악할 수 있게 구성을 하였다. 이를 통해 사용자들은 가시화된 데이터를 보고 정확한 의사 결정이 가능하게 된다. 향후 연구로는 본 논문을 토대로 3D 형태의 시각화가 있다.

참고문헌

[1] Michael Schoroeck, Rebecca Shockley, janet smart, Dolores Romero-Morales, Peter Tufano "Analytics The Real World Use Of Big Data KRV3" p.1-10, 2012

[2] Lee ji seon, “빅데이터 분석 시각화 분석 : 1장 시각화 정의 2장 프로세스” p.5, 2014
[3]Jong Hee LEE, Taeg Yun OH, Jae Bong LEE1, Young Il SEO, Jung Hwa CHOI, Jung Yun KIM, DongWoo LEE “Seasonal variations in species composition by the stow nets and the stow net on boat fisheries in the Han River Estuary, Korea” Journal of the Korean Society of Fisheries Technology Vol.48 No.4 pp.452-468
[4] Joung Woo Ryu, Jin-Hee Song “Visualization for Big Data” The Korea Contents Association Review 12(1), 2014.3, 21-26 (6 pages)
[5] Ben Fry, “Visualizing data” O’REILLY, 2016. 05. 31
[6] Kim SeongKi, “Quick Decision Making Using Visual Dynamic Mining Tool : Sptfire” Korean Institute of Intelligent Systems 18(1), 2008.4 89-91 (3page)