버블 힙: 2 개의 시계열데이터 세트를 사용하는 다차원 시각화 기법

Bubble Heap: Visualization technique using two different time series data set

이만재, Manjai Lee*, 온병원, Byung-Won On**

요약 시계열 데이터를 시각화하는 가장 기본적인 방식은 x 축을 시간 축으로 사용하고 y 축을 표현하고자 하는 데이터의 값을 사용하는 선형 그래프이다. 선형 그래프의 경우 하나의 그래프에서 여러 개의 변량을 표시하고자 하면 데이터의 선이 중복되어 그래프의 의미를 파악하기 어렵다. 이러한 문제는 컴퓨터의 인터랙션 기능을 사용해 필요에 따라 데이터를 달리 보여 줌으로 해결할 수 있다. 다차원 시계열 데이터의 경우 시간을 매개로 하여 다른 데이터 세트의 관계를 보여 주는 것이 가능하며 GapMinder 는 이러한 기능을 최대한 활용한 사례이다. GapMinder 의 경우 기본적으로 3 개의 데이터 세트를 사용하며 2 개 데이터 세트를 종합적으로 분석하는 목적에 적합하다. 2 개 차원의 관련성을 보고자 하는 것이 아니라면 분석의 대상이 되는 차원을 하나만 사용하는 버블 힙이라는 새로운 시계열 시각화 기법을 제시한다.

Abstract The basic technique of time series data visualization is linear graph using x axis and the remaining data for y axis. This technique is not suitable for large number of multi-variate data analysis. Interaction technique can be used to resolve the problem by showing subset of data. For multi-dimensional case, using separate data sets, GapMinder shows its strength to understand the relation between two different data and one additional data for the circle size. It is very powerful technique for comparing time-series data, such as national activity of the world. We propose a new technique called Bubble Heap, which is more suited in case we want to investigate data with less dimension.

핵심어: *Time series, Visualization, Force directed layout, Bubble heap*

1. 서론

빅 데이터에 관한 관심이 증가하면서 빅 데이터 분석 결과의 시각화에 대한 관심이 증가하고 있다. 데이터 분석에 사용되는 엑셀과 같은 소프트웨어는 스프레드 시트 계산 결과를 시각화하여 보여주는 기능을 갖고 있으나 막대그래프, 원 그래프, Scatter Plot 과 같은 종이 매체에 적합한 시각화 기법을 그대로 사용하고 있어 컴퓨터를 이용하는 장점을 살리지 못하고 있다. 빅 데이터 분석 기능과 시각화에 자주 사용되는 R 의 경우 추가적인 패키지를 통해 네트워크 그래프, 지도를 사용하는 그래프 등 다양한 시각화 방법을 시도할 수 있다.

데이터 시각화에서 자주 등장하는 데이터로 시계열(Time series) 데이터를 들 수 있다. 시계열 데이터는 하나의 변수로 시간을 사용하는 모든 데이터를 말하며, 시간이라는 변수를 매개로 한다. 증권을 예로 들면, 주식의 거래가를 시간 단위나 일 단위로 표시하고자 할 경우 선 그래프를 사용한다. 하나의 주식은 하나의 선으로 표시하며 여러 개의 다른 주식을 표현할 경우 여러 개의 선을 사용한다.

본 논문은 2013 년 차세대융합기술연구원의 연구비 (No. 2012-P3-22) 지원에 의하여 연구되었음.

*주저자: 서울대학교 차세대융합기술연구원 특임연구위원 e-mail: manjai@snu.ac.kr

**교신저자: 서울대학교 차세대융합기술연구원 연구교수 e-mail: bwon@snu.ac.kr

월드뱅크나 OECD 와 같은 국제기구는 정책 개발의 필요성 때문에 회원국의 데이터를 지속적으로 유지하고 있다. 개별 국가의 경우에도 지역별 또는 부처별 비교를 통한 정책개발을 위해 시계열 통계를 유지하고 있다. 대부분 시계열 형식으로 표시된 이러한 통계 데이터는 적절한 분석 도구가 주어진다면 정책 수립에 큰 도움이 될 수 있다.

본 연구에서는 기존의 시계열 데이터 분석 기법이 어떻게 사용되고 있는지를 살펴보고 국제기구나 국가와 같이 회원국이나 산하기구의 특징을 새로운 시각으로 분석할 수 있는 버블 힙(Bubble Heap)이라는 새로운 시각화 기법을 제시한다.

2. 시계열 시각화 기법

시계열 데이터를 다루기 전에 본 논문에서 사용되는 용어를 정의한다. 시간의 변화에 따른 인구의 증가를 표시하는 데이터를 예로 들자. 우리나라의 인구 변화 데이터를 하나의 데이터 세트라고 할 수 있다. 만약 우리나라 외에 미국, 중국, 일본 등 다른 국가의 인구 데이터를 같이 표시해 분석할 경우 이를 다변량(Multivariate) 분석으로 정의한다[1].

인구 데이터 세트 외에도 시간의 변화에 따른 다른 데이터가 있을 수 있다. 한국, 미국, 일본, 중국의 수출 데이터가 있다면 이는 또 하나의 데이터 세트라고 본다. 이렇게 두 가지 다른 데이터 세트를 시간이라는 매개변수를 사용하여 인구와 수출 간의 관계를 분석하고자 할 경우 데이터의 차원이 다른 것으로 보아 다차원(Multi-dimensional) 분석으로 정의한다. 다차원 분석은 컴퓨터의 도움 없이는 어려운 분석으로 최근에 활성화된 분야이다.

2.1 다변량 시각화 기법

현재까지 가장 널리 사용되는 시계열 그래프 방식은 선 그래프(Line graph)로 볼 수 있다. x 축을 시간축으로 하고 데이터 세트의 값을 y 축에 표시하는 방식으로 데이터가 표시되는 점을 선으로 연결하여 사용하는 방식으로 이해하기 쉽고 그리기 쉽기에 모든 분야에서 사용된다. 선 그래프의 단점은 여러 개의 선이 중복될 경우 확실히 구별되는 소수의 현상에 대해서만 관심을 끌기 어렵다는 문제를 갖는다. 또한 하나의 변량에 대한 분석만이 가능하기에 다른 데이터 세트와의 관계를 보는 것은 불가능하다.

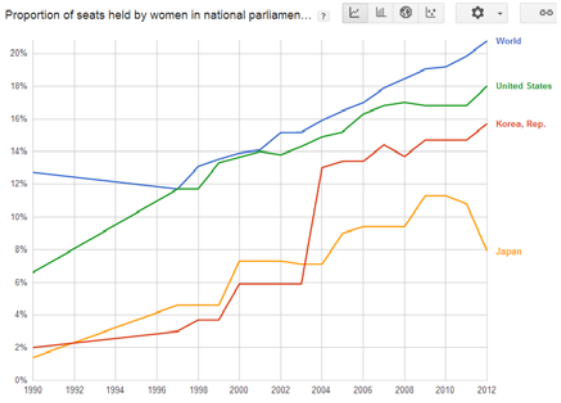


그림 1. 선 그래프 예

그림 1은 1990년부터 2012년까지 여성의원 비율을 한국, 미국, 일본의 경우와 전세계 평균값을 보여주고 있다. 변량의 수가 적을 경우 전달하고자 하는 메시지를 정확히 보여주고 있다. 데이터와 그래프는 구글이 제공하는 공공 데이터 익스플로러를 이용하였다.

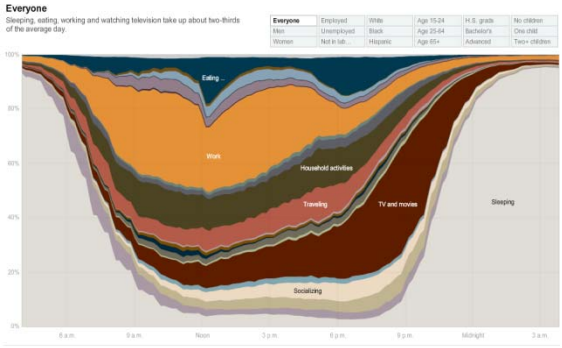


그림 2. 일반인의 누적선 그래프의 예

선 그래프에서 전체 데이터를 보고자 할 경우에는 누적 그래프(Cumulative graph)가 사용되다. 누적 그래프는 개별 변량의 변화와 함께 개별 변량과 함께 누적된 데이터를 볼 수 있도록 한다. 선 그래프와 누적 그래프는 일반 사무환경에서 자주 사용되기에 엑셀의 기본 기능으로 포함되어 있다. 그림 2는 평균적인 사람이 하루를 어떻게 보내는가를 시각화한 것이다[2]. 가장 아래의 수면시간이 밤시간에 가장 큰 비중을 차지하는 것과 식사시간이 점심시간과 저녁시간에 비교적 많은 비중을 차지하는 것을 볼 수 있다.

그림 2는 앞의 그림 1에는 없는 인터랙션 기능을 포함하고 있다. 연령별, 성별, 직장 유무에 따라 실제 하루 생활이 어떻게 다른 가를 보여 주는 것이 가능하다. 그림 3은 직업이 없는 실업자를 선택한 경우 실업자의 하루를 보여주는 그래프로 업무라는 영역이 없어지고 대신 집안일이 차지하는 비율이 증가한 것을 볼 수 있다.

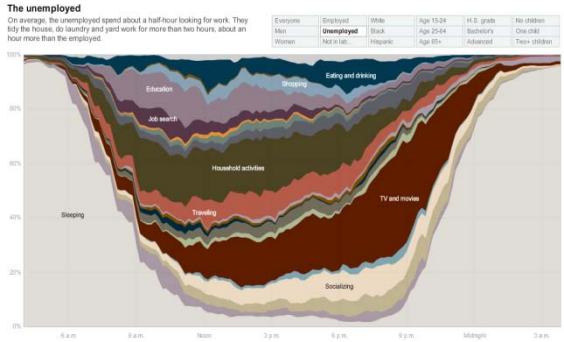


그림 3. 실업자의 하루 일과 누적 선 그래프

선 그래프나 누적 선 그래프 모두 제한된 수의 데이터 이상을 표현하는 데 문제가 있으며 잘 설계된 데이터 세트가 아닌 경우 대략 5개 내외의 데이터 세트에서 의미 전달이 가능하다. 이러한 문제를 해결하기 위해서는 관심을 갖는 데이터 세트만을 따로 그리는 복수 그래프(Multiple graph) 방식[3]을 사용할 수도 있다.

2.2 다차원 시각화 기법

앞의 사례는 시계열에서 x 축을 시간의 변화에 할당하고 y 축에 데이터 값을 매핑 하는 방식으로 그래프를 그리도록 하였다. 시간이라는 변수는 데이터 세트에서 자주 등장하기에 시간을 매체로 하고 x 축과 y 축에 다른 데이터 세트 값을 적용한다면 새로운 시각화가 가능하다. 이러한 방법을 처음 시도한 것은 한스 로스링의 갭마인더(GapMinder)[4]를 들 수 있다. 인터넷이라는 매체를 최대한 활용하여 통계에 재미라는 요소를 추가한 대표적인 사례이다. 갭마인더는 전세계의 국가의 백여 개 이상의 변수를 다루는 통계 세트 중 3개의 상이한 데이터 세트를 임의로 선택하여 볼 수 있도록 한다.

갭마인더에서는 인구, 수출입, 환경, 교육 등 전 세계 주요 국가의 최근 200년간의 데이터를 내장하고 있어 하나의 어플리케이션으로 볼 수 있다. x 축과 y 축에 각각 다른 데이터 세트를 할당하면 해당 데이터가 유효한 기간을 슬라이더 형태의 타임라인으로 표시된다. 재생(Play)버튼을 누르면 시간의 흐름에 따라 국가를 대표하는 원이 2차원 평면에서 이동하게 된다. 원은 인구나 GDP 등 국가를 대표하는 값과 비례하도록 할 경우 시각적으로 해당 국가의 중요성을 인지할 수 있다.

전쟁이나 기근과 같은 특이한 경우에는 인구의 감소, 평균 수명의 변화 등 특정국가의 이상한 움직임을 애니메이션 행태로 볼 수 있어 데이터로부터 스토리를 얻을 수 있는 우수한 사례이다.

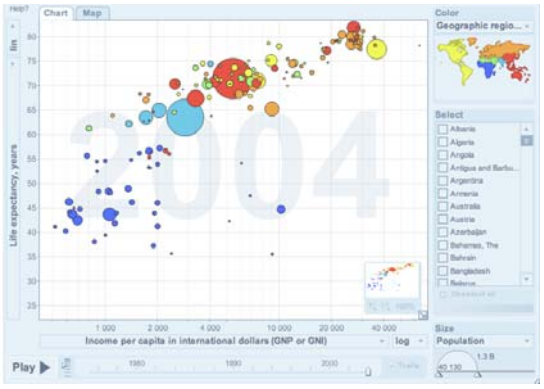


그림 4. 갭마인더 그래프의 예

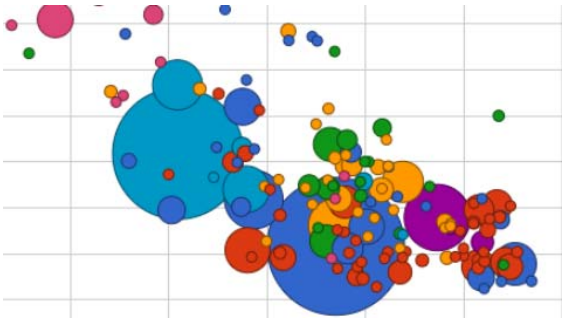
그림 4는 x 축에 1인당 GNP 를, y 축에 기대수명을 적용하고 원의 크기는 인구에 비례하도록 한 결과이다. 시간은 1975년부터 2005년까지를 대상으로 하고 있고 2004년의 전 세계 국가를 해당좌표에 표시하고 있다.

2.3 시각화 기법 비교

앞서 설명한 시계열 시각화 기법은 각각의 장점과 단점을 갖는다. 가장 오래 전부터 사용된 선 그래프는 데이터가 준비되어 있다면 가장 그리기 쉽다는 장점을 갖고 있다. 엑셀과 같은 오피스 도구에 기본 기능이 탑재되어 있다는 것도 큰 장점으로 꼽힌다. 그러나 데이터 세트의 값이 골고루 분산되어 있지 않다면 선의 겹침이 발생하여 다섯 개 이상의 데이터 세트에 사용하기에는 무리가 따른다. 그러나 컴퓨터 기능을 사용하여 버튼을 추가한다면 데이터를 바꾸어 보여줌으로 어느 정도 해결이 가능하다. 누적 선 그래프의 경우도 선 그래프와 장단점을 공유한다.

갭마인더는 구현에 많은 사전 데이터 준비작업이 필요하였기에 쉬운 기법은 아니나 기본 플랫폼이 준비되었다면 데이터를 추가하면 쉽게 변화를 볼 수 있다. 갭마인더는 많은 장점을 갖고 있지만 이를 다른 목적으로 사용하고자 할 경우 약점을 갖고 있다. 그림 5에서 볼 수 있는 바와 같이 원으로 표시되는 국가가 중첩되게 배치될 경우 일부 국가가 보이지 않는다. 면적이 큰 원을 뒤로 보내거나 국가 선택 메뉴를 중요한 국가를 하이라이트 표시함으로 문제를 일부 해결할 수 있으나 중첩이라는 근본적인 문제를 해결하지는 못한다.

두 번째 문제는 갭마인더를 사용하기 위해서는 두 개의 서로 다른 데이터 세트를 선택해야 그래프를 볼 수 있다는 것이다. 하나의 변수에만 관심을 갖는 경우 추가적인 데이터 세트로 무엇을 선택해야 하는가 고민을 해야 한다. 여기에서 중첩문제에서 자유로운 새로운 기법이 필요함을 알 수 있다.



.

그림 5. 갭마인더에서 여러 원이 중첩된 예

3. 버블 힙

새롭게 제시한 시각화 기법은 시계열 데이터를 기본으로 하고 두 개의 데이터 세트를 기본으로 선택한 대신 하나의 데이터 세트의 변화를 집중적으로 보기 위한 기법이다. 갭마인더에서 국가를 표시하는 원을 버블로 명명하고 버블이 함께 뭉쳐 있는 그래프 모양을 따라 버블 힙이라 명명하였다.

버블 힙에서는 갭마인더와 같이 제3의 데이터 값을 원의 크기로 표시하는 방식은 동일하다. 핵심이 되는 데이터 차원을 둘에서 하나로 줄인 만큼 버블 힙에서는 x 축의 중심역할을 하며 y 축은 보조역할만을 담당하다. x 값은 주어진 데이터 값에 가급적 가까운 값을 사용하도록 하며 y 값은 버블의 중복이 일어나지 않도록 임의로 정하는 것을 원칙으로 한다. 여기서 x 축의 값을 그대로 사용하지 않고 가급적 가까운 값을 사용하도록 한 것은 데이터의 값이 정수를 사용하여 같은 값이 발생할 경우 원하는 그래프가 만들어 지지 않는 것을 방지하기 위한 것이다. 버블 힙의

배치 알고리즘은 이러한 원칙을 지키는 한도 내에서 여러 가지 방식을 사용할 수 있다. 가장 단순한 방법은 y 값에 임의의 랜덤 값을 적용한 다음 이미 배치된 버블에 충돌하지 않는 한도 내에서 가급적 x 축에 가깝게 이동하도록 하는 것이 가능하다.

보다 발전된 알고리즘은 목표 위치의 좌표를 (x_i 0)로 정하고 랜덤 값으로 정해진 초기 위치로부터 목표로 이동하도록 하는 방식이다. 이동 과정에서 버블 간 충돌이 일어나며 충돌 결과를 반영하여 다음 좌표를 정하는 과정을 반복한다. 이 때 마찰력을 적용하여 이동 속도를 점차로 줄여나가며 이동거리가 정해진 값 이하일 경우 계산을 멈추는 방식을 사용할 수 있다.

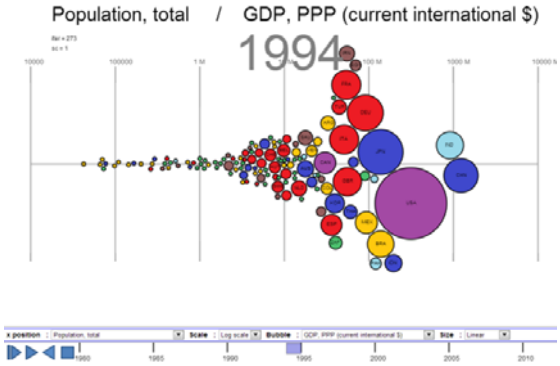


그림 6. 국가의 인구와 GDP 데이터를 사용한 버블 힙

그림 6은 버블 힙 기법을 월드뱅크 데이터에 적용한 프로토타입이다. 프로토타입은 두 개의 데이터 세트를 선택할 수 있는 드롭 다운 메뉴를 포함하고 있어 첫 번째 데이터 세트는 x 좌표를 정하는 데 사용하며 두 번째 데이터는 버블의 크기를 정하는 용도로 사용한다. 데이터의 성격에 따라 스케일을 선형(Linear) 값 또는 로그(Log) 값을 사용할 수 있다. 그림 7에서 국가는 GDP 값으로 면적이 결정되며 국가의 인구로 x 축 값을 결정한다. 그림에서 가장 큰 버블은 미국을 표시하며 가장 오른 쪽에 배치된 두 개의 버블은 각각 중국과 인도를 표시한다. x 스케일은 국가 간 차이가 크기에 로그 스케일을 적용하고 있다.

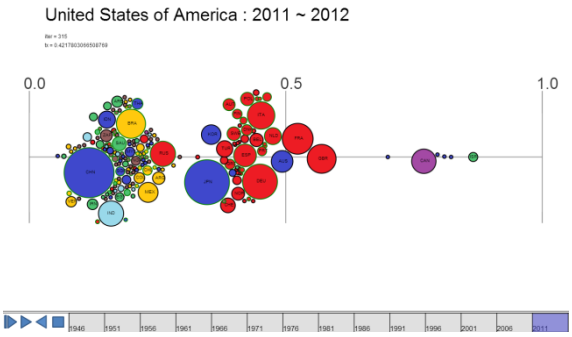


그림 7. 미국과 다른 UN 회원국간의 관계 버블 힙

버블 힙의 용도가 여러 용도에 활용할 수 있음을 보여주기 위해 UN 회원국간의 투표 유사도의 시각화에 사용하였다. 그림 7은 미국을 기준으로 한 다른 국가와의 관계를 표시한다. 190여 개의 회원국과 미국과의 관계를 동시에 볼 수 있기에 갭마인더와는 다른 통찰력을 제공할 수 있다.

버블 힙에서 버블 배치 알고리즘은 기본 원칙을 지키는 한도 내에서 여러 가지 새로운 방법을 제시할 수 있으며 이는 추후 연구과제로 남는다.

4. 결론

시계열 데이터를 시각화하는 기법의 발전을 기술하였다. 선형 그래프로부터 시작된 기법은 누적 그래프와 같이 오랜 기간 사용되어 왔다. 컴퓨터의 인터랙션 기능이 추가되면서 같은 선형 그래프라도 다양한 변화를 가져올 수 있게 되었다. 그러나 가장 큰 변화는 시간 축을 매개로 하는 두 개 이상의 데이터 세트를 2차원 평면에 보여주는 갭마인더와 같은 다차원 시각화 기법의 등장이다.

버블 힙은 2개 데이터 세트를 정해야만 가능한 2차원 배치방법을 대신해서 하나의 데이터 세트만을 정하고 나머지 좌표는 버블 간에 중첩이 발생하지 않도록 하는 원칙을 제시한 새로운 시각화 기법이다. 월드뱅크의 데이터 세트와 UN 회원국의 투표 유사도에 적용하여 버블 힙 아이디어의 사용 가능성함을 보였으며 버블 배치 알고리즘에 대해서는 추가적인 연구가 필요하다.

참고문헌

[1] Wolgang Mullër and Heidrun Schumann, “Visualization methods for time-dependent data-an overview,” Simulation Conference, 2003. Proceedings of the 2003 Winter. Vol. 1. IEEE, 2003

[2] For the Unemployed, the Day Stacks Up Differently, http://www.nytimes.com/2009/08/02/business/02metrics.html

[3] Jeffrey Heeret et al., “A tour through the visualization zoo,” Communication of the ACM, Vol. 53, No. 6, pp59-67, 2010

[4] Hans Rosling, Rönnlund A. Rosling, and Ola Rosling, “New software brings statistics beyond the eye,” Statistics, Knowledge and Policy: Key Indicators to Inform Decision Making. Paris, France: OECD Publishing (2005): 522-530