막대그래프 왜곡 탐지 및 교정 시스템

A System for Detecting and Correcting Distorted Bar Graphs

이규민	박미현	변유진	이지은	박규동*
Gyumin Lee	Mihyun Park	Yoojin Byeon	Jieun Lee	Kyudong Park
광운대학교 정보융합학부	광운대학교 정보융합학부	광운대학교 정보융합학부	광운대학교 정보융합학부	광운대학교 정보융합학부
Kwangwoon University	Kwangwoon University	Kwangwoon University	Kwangwoon University	Kwangwoon University
luckykr1234@gmail.com	3093977@naver.com	yjasd1234@naver.com	dlwldmscjstk@naver.com	kdpark@kw.ac.kr

요약문

현대 인터넷에서는 수많은 데이터들이 그래프 형태로 제공된다. 하지만 이러한 정보의 시각화는 정보 설계자의 실수 또는 의도로 인해 진의가 왜곡될 수 있는 여지가 존재하고, 결국 수용자로 하여금 잘못된 정보를 받아들일 수 있는 가능성을 내재하게 된다. 본 연구에서는 이러한 그래프 정보의 왜곡을 줄이고자 웹페이지상에서 이를 탐지하고 교정하여 올바른 그래프를 제공할 수 있는 시스템을 제안한다. 접속한 웹페이지의 모든 이미지가 서버로 전송되며 그래프인 경우에 대해 표시된 텍스트 값과 그래프 이미지의 크기를 인식하여 왜곡 여부를 판단하고, 왜곡이 감지할 경우 그래프를 재구성하여 웹페이지에 표출한다. 본 연구는 시각적 데이터와 텍스트 데이터의 차이가 존재하는 왜곡된 정보를 가진 막대그래프를 교정하는데 성공하였지만 추정 가능한 대상이 특정 조건을 갖춘 막대그래프라는 한계를 갖는다. 향후 다양한 그래프 유형에 대한 정보 왜곡을 판단할 수 있도록 모델을 개선할 예정이며, 해당 확장 프로그램으로 웹사이트 상의 정보 왜곡을 파악하는 것이 정보 수용자에게 어떤 영향을 미치는지 사용성 평가를 통해 연구해 보고자 한다.

주제어

그래프 왜곡, 머신러닝, 정보디자인, 데이터시각화

1. 서론

1.1 현대 사회의 그래프 왜곡

현대 사회에서는 데이터의 분석 결과를 쉽게 이해할 수 있도록 그래프라는 수단을 통해 명확하고 효과적으로 전달된다. 이러한 데이터의 시각화를 진행할 때 이해하기 쉽게 전달하는 것이 중요하지만, 언론사와 같이 대중의 인식에 많은 영향을 끼치는 기관은 공정하고 중립적인 입장에서 정보를 제공해야 한다. 하지만, 미디어 매체나 정부 기관에서 왜곡된 시각화를 제공하는 경우를 어렵지 않게 찾아볼 수 있다[1][2]. 심지어 Science 저널에서도 전체 그래프 중 30%가 오류를 포함하는 것이 드러나기도 하였다[3]. 그래프는 문자 대비 직관적인 정보 전달이 가능한데, 정보 수용자들은 꼼꼼하게 이를 살피지 않고 피상적 수준에서 이해하기 쉽다. 결국, 정보 설계자의 의도 또는 실수가 들어가 있다는 것을 알지 못하는 정보 수용자들은 왜곡된 정보를 수용하게 될 가능성이 높다[4].

1.2 연구 목표

그래프 왜곡의 종류들은 실제 시각화와 수치를 의도적으로 다르게 표시하거나 축을 비롯한 공간적 요소를 의도적으로 왜곡하는 등 다양하게 존재한다[5]. 이러한 그래프 왜곡으로 인한 잘못된 정보 전달을 줄이고자 본 연구에서는 막대그래프 상에 수치와 시각화의 불일치가 존재하는 그래프의 왜곡을 감지하여 이를 본래의 데이터에 걸맞게 재구성할 수 있는 프로그램을 제안하고자 한다.

기존에는 이러한 왜곡을 감지하기 위해서 정보 수용자가 직접 그래프 상의 수치와 시각화 요소를 일일이 확인해야 하는 번거로움이 존재한다. 또한, 매우 교묘하게 왜곡된 그래프의 경우 감지가 어렵고 정보 수용자로 하여금 피로를 가중시킨다.

이런 한계점을 극복하기 위해 제안하고자 하는 시스템은 크롬 확장 프로그램 플랫폼을 이용하여 웹사이트 상의 그래프를 자동으로 수집하고 해당 그래프에 존재하는 수치와 시각화의 데이터를 정량적으로 분석 및 비교한다. 불일치가 존재하여 오류가 있다고 판단되는 경우 텍스트 데이터를 바탕으로 정확한 그래프를 제작하여 웹사이트 상에 원본 그래프와 비교하여 표출하는 기능을 갖고 있다.

2. 제안 시스템

2.1 프로그램 데이터 흐름

웹페이지 상의 그래프 수집, 그래프 분류, 그래프 상의 데이터 추정, 정확한 그래프 생성의 기능으로 구성된 확장 프로그램을 개발하였으며 해당 프로그램의 데이터 흐름은 그림 1 과 같다.

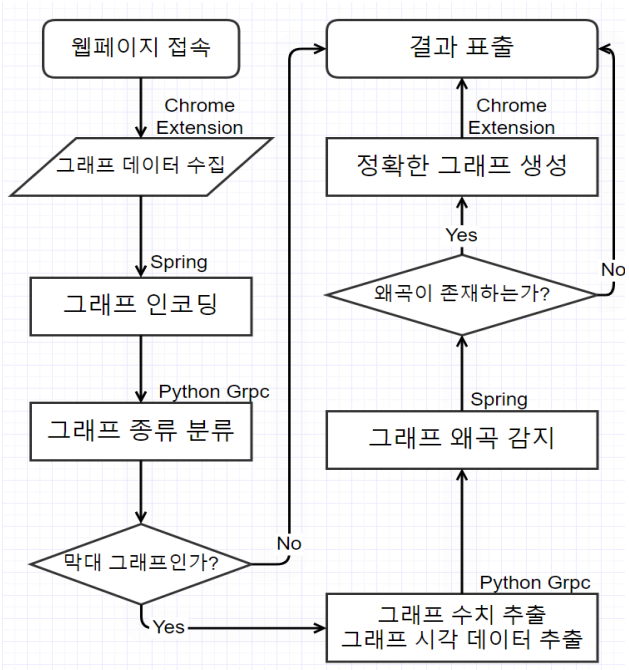

그림 1 크롬 확장 프로그램의 데이터 흐름

2.2 그래프 종류 분류

그래프의 종류를 분류하는 작업은 [6]에서 제시한 학습된 그래프 분류 모델을 이식하여 구현하였다. 해당 모델은 VGG-19 기반의 ImageNet 데이터 세트로 사전 훈련되어 있는 모델[7]을 바탕으로, 해당 연구에서 추가로 수집한 차트 데이터들을 학습시켜 조정된 모델이다. 해당 모델은 그래프를 총 13 가지의 종류로 분류하며 평균 검증 정확도는 82.11%를 나타내는 것을 확인할 수 있었다[6]. 이를 그래프 추정 전 단계에 배치하여 입력 받은 이미지가 막대그래프인지 아닌지를 판단하여 막대그래프인 경우 데이터 추정을 진행한다.

2.3 그래프 데이터 추정

그래프 내부의 수치 텍스트와 시각화된 데이터를 수치화하는 작업은 [6]에서 제시한 막대그래프 데이터 추정 방법을 중심으로, 한글 환경에서 그래프 왜곡 여부 판별에 필요한 데이터를 수집할 수 있도록 추가적인 수정을 가하는 등 본 연구에 맞게 커스터마이징하여 이식하였다. 제시된 추정 방법은 그래프 내에 존재하는 텍스트의 위치 정보와 축의 위치정보를 추정해내고 이를 바탕으로 x 축과 y 축의 텍스트 정보, 범례 정보를 구분하고 범례 정보의 색상을 탐지한다. 수집된 데이터를 토대로 데이터 막대들의 위치를 파악해 막대의 픽셀 길이와 y 축 텍스트의 y 좌표 값을 대조하여 해당 막대가 어떠한 수치를 표현하는지 파악한다.

2.3.1 텍스트 인식

본 연구에서는 그래프상의 텍스트 정보를 분석하기 위해 Naver 의 Clova OCR API 를 사용하였다. 이는 그래프상에 존재하는 모든 텍스트의 좌표와 크기, 그리고 내용을 인식하여 수집한다. 텍스트의 좌표와 크기는 해당 텍스트가 어디에 위치하는지 파악하여 어떤 역할을 지니는지 추정하는 데 사용한다.

2.3.2 x,y 축 추정

그래프에 존재하는 축의 위치를 추정하기 위해 그래프 사진을 회색조로 변환하여 휘도가 200 이상인 모든 픽셀을 1(흰색)으로 바꾸고 그 외의 모든 픽셀은 0(검정)으로 변환하여 사진을 이진화한다. 변환된 사진에서 연속된 검정 픽셀이 제일 길게 존재하는 행 중에서 가장 아래에 있는 행을 x 축이라고 판단하고 제일 길게 존재하는 열중에서 가장 좌측에 있는 열을 y 축이라고 판단한다.

2.3.3 막대 위 숫자 인식

수집한 텍스트들의 정보와 x, y 축의 위치 정보를 바탕으로 x 축보다 오른쪽에, y 축보다 위쪽에 있는 텍스트들을 검사한다. 이들 중 숫자로만 이루어져 있는 텍스트들을 수집하여 이를 데이터 막대에 매칭되는 숫자라고 판단한다.

2.3.4 범례 추정

먼저 수집한 숫자들을 제외하고 남은 텍스트 중 가로나 세로로 정렬되어 있는 텍스트 집합을 검사하여 이들 중 가장 많이 정렬되어 있는 집합을 범례라고 판단한다. 이 중 거리가 10 미만인 텍스트들은 연결되는 단어라고 판단하여 한 단어로 집계하고, 오차 막대를 잘못 판단하여 생성되는 “I”문자는 집계 시 제외한다.

2.3.5 범례 – 막대 색상 추정

범례로 추정되는 박스들 중 흰색을 제외한 가장 큰 크기를 가진 동일 색상 박스를 해당 범례에 매칭되는 막대의 색상이라고 판단한다.

2.3.6 막대 데이터 수치 추정

앞서 추정한 범례-막대 색상 값을 바탕으로 해당 색상 값을 지닌 막대들의 세로 길이를 파악한다. 그리고 그래프 상의 내부 픽셀 길이 값과 실제 데이터 수치의 값을 연결하기 위해 y 축 텍스트들의 간격과 이들의 y 좌표를 대입하여 수치 : 픽셀 길이의 비율을 계산한다. 이렇게 계산된 비율을 바탕으로 각 막대들의 길이 값을 대입하여 해당 막대가 나타내는 수치를 추정한다.

2.4 그래프 왜곡 판별

앞선 단계에서 추정한 데이터들을 기반으로 그래프의 막대 데이터 값과 각 막대 위에 존재하는 수치들의 값을 비교하여 왜곡이 존재하는지 판별한다. 여러 그래프를 테스트를 진행하였을 때 정상적으로 추정된 막대 데이터 값은 최대 10%의 오차가 존재한다고 판단하여 막대 위의 텍스트 수치와 해당 막대가 나타내는 추정 값의 차이가 텍스트를 기준으로 10% 이상의 차이가 존재할 경우 그래프가 왜곡되었다고 판단한다.

2.5 크롬 확장 프로그램

본 연구 시스템은 웹브라우저인 크롬의 확장 프로그램으로 배포되어 페이지 내 오류가 존재하는 막대그래프 이미지가 감지되었을 때 해당 이미지의 오류를 즉시 확인할 수 있도록 개발되었다. 사용자는 크롬에서 운영하는 확장 프로그램 사이트에서 해당 시스템을 크롬 브라우저에 설치할 수 있고, 사용자는 페이지 내의 원본 막대그래프 이미지와 교정된 그래프를 비교하며 오류 정보를 얻을 수 있다.

2.5.1 시스템 사용법 및 조작 제공

해당 프로그램은 사용자가 시스템을 손쉽게 조작할 수 있도록 크롬 확장 프로그램 플랫폼 자체에서 제공하는 팝업 기능을 활용하여 시스템 사용법을 사용자에게 제공한다. 사용자는 팝업창을 이용해 시스템의 소개 및 목적을 확인할 수 있고, 시스템의 오류 교정 방식을 알 수 있다. 또한 팝업창은 그림 2 와 같이 교정 그래프와 오류 내역 텍스트 표시 각각의 on · off 버튼을 제공하여 사용자가 원하는 교정 형태로 페이지가 구성될 수 있도록 돕는다.

2.5.2 그래프 재구성

웹사이트 상에 존재하는 그래프 이미지를 전송하여 해당 사진에 왜곡이 존재한다고 판단되는 결과를 받으면 같이 전달받은 그래프 데이터를 바탕으로 그래프를 재구성하여 페이지에 표시한다. 그래프는 chart.js 라이브러리를 사용하여 서버에서 분석한 수치 텍스트 데이터를 바탕으로 재구성한다. 재구성되는 그래프의 크기는 원본 이미지의 크기를 따르며 y 축 범위와 범례 별 색상, 텍스트 데이터를 그대로 인식하여 원본 그래프와 흡사하게 제작한다.

2.5.3 왜곡 내역 표시

왜곡이 존재할 경우 해당 왜곡에 대한 구체적인 정보 역시 표시된다. 특정 막대가 왜곡됐다는 결과를 받았을 때, 해당 막대에 대한 데이터와 어느 정도 왜곡되었는지에 대한 정보를 표시한다. 이에 사용자는 오류가 존재하는 막대그래프 이미지의 어떤 막대에서 오류가 존재하는지 파악할 수 있고, 텍스트로 정의된 데이터 수치와 막대 데이터 수치 간의 차이를 확인할 수 있다. 또한 오류 수치와 대응되는 해당 막대가 무엇인지를 한눈에 파악하기 쉽도록 오류 내역 텍스트에는 해당 막대의 색상이 표시된다.

그림 2 크롬 확장프로그램 팝업 창

3. 결과

크롬 확장 프로그램을 통해 보여주는 그래프 왜곡 여부 판별 및 그래프 재구성 결과는 다음 그림 3, 4 와 같다. 그림 3 은 원본 그래프를 나타내며, 그림 4 는 교정된 이후에 새롭게 생성된 그래프를 나타낸다. 재구성된 그래프 하단에 어떤 항목의 어떤 막대가 얼마나 왜곡되었는지에 대한 구체적인 왜곡 정보를 텍스트로 같이 제공한다.

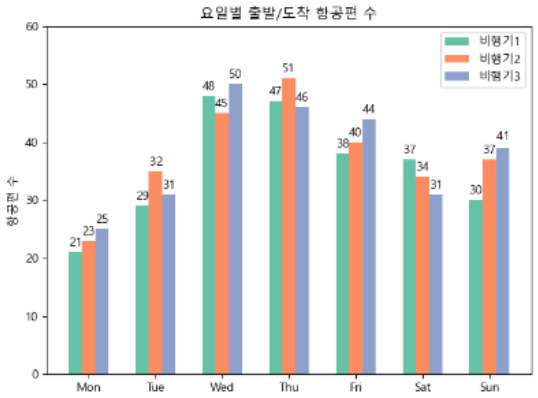

그림 3 원본그래프 예시

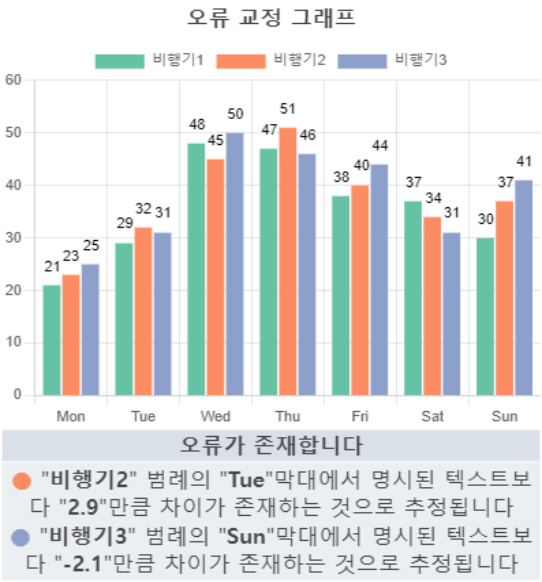

그림 4 시스템을 통해 교정된 그래프 및 설명

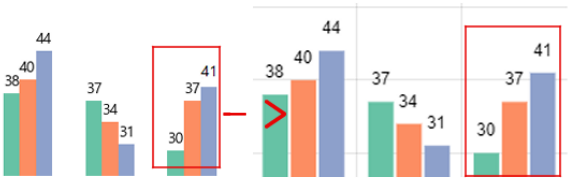

그림 5 감지된 오류의 수정 사례 (Sun 항목)

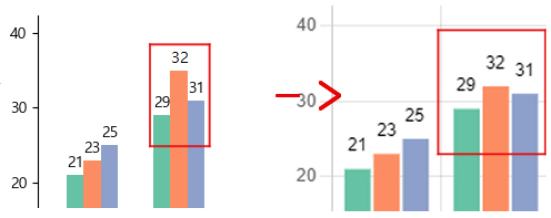

그림 6 감지된 오류의 수정 사례 (Tue 항목)

그림 5 는 x 축의 “Sun”항목에서 “비행기 3”이 텍스트에 비해 막대를 짧게 나타내어 해당 수치를 과소평가할 수 있게 하는 정보 왜곡 사례를 교정하였다. 그림 6 의 “Tue”항목에서는 “비행기 2”가 수치에 비해 막대를 길게 나타내어 해당 수치를 과대평가할 수 있게 하는 정보 왜곡을 교정한 것을 확인할 수 있다.

4. 토의 및 결론

본 연구는 [6]의 연구를 바탕으로 이미지 상의 그래프를 분류하고 데이터를 추정하여 텍스트와 막대의 데이터 간에 왜곡 여부를 판별하는 서버를 구축하였다. 그리고 이를 크롬 확장 프로그램을 통해 정보 수용자가 접하게 되는 웹 상의 그래프의 정보 왜곡 여부를 판별하여 제공하고, 왜곡되었을 경우에 교정된 그래프를 같이 제공함으로써 웹 상에서 노출되는 정보의 왜곡을 줄일 수 있는 프로그램을 구현하였다.

해당 프로그램은 별다른 조작 없이 웹사이트 상의 그래프 사진을 자동으로 수집하고 왜곡을 판단한 결과를 사용자에게 제공한다. 이는 기존에 정보 수용자가 스스로 그래프의 왜곡 여부를 판단해야만 했던 과정을 자동화하여 편리함을 제공하게 된다. 왜곡이 존재할 경우 비교적 정확한 대체 그래프 역시 제공함으로써 원본 그래프와 비교하여 어떠한 부분이 왜곡되었는지 직접적으로 확인할 수 있고, 정보 설계자의 의도 역시 쉽게 파악할 수 있어 정보 전달 측면에서 긍정적인 영향을 미칠 수 있을 것이라 생각된다.

또, 크롬 계열의 브라우저에서 모두 사용할 수 있는 크롬 확장 프로그램이라는 플랫폼을 기반이므로, 보다 더 많은 사람들이 간편하게 해당 프로그램을 이용할 수 있을 것으로 기대된다.

4.1 한계

해당 프로그램은 [6]의 연구에서 드러나는 한계점을 그대로 답습하고 있다. 해당 연구에선 ICDAR 2019[8]의 데이터 세트를 이용해 막대그래프의 데이터 추정을 시도하였을 때 198,010 건의 전체 데이터에서 100 건을 무작위 추출했을 때는 27.19%의 정확도를, 전체 데이터에서 대상을 범례가 RGB 로 표현된 수직 막대 그래프로 제한하여 100 건을 무작위 추출했을 때는 45.45%의 정확도를 가졌음을 확인할 수 있었다. 이처럼 정확도가 낮은 이유는 데이터 추정에 필요한 모든 값의 탐지에 성공해도 단 하나의 정보의 추정에 실패하면 결과적으로 막대의 데이터 자체를 추정할 수 없기 때문이다.범례가 존재하지 않거나 색상 의외의 방식으로 범례가표기되어 있어 범례와 막대의 관계를 추정해내지 못하는 경우, 축에 해당하는 선이 존재하지 않거나 텍스트들

의 위치가 상정한 범위를 벗어날 때 바로 추정할 수 없는 문제가 발생한다.

4.2 향후 연구

향후 연구에서는 해당 프로그램의 막대그래프 분석에서 발생하는 각종 제약 사항을 완화할 수 있도록 연구를 진행하고자 한다. 차트 데이터 추정 과정을 개선하여 정보 왜곡 탐지의 정확도를 높이고, 단순히 텍스트와 막대 데이터 간의 왜곡만이 아닌 축 왜곡 등 다른 막대그래프에서 나타날 수 있는 다른 정보 왜곡 또한 감지하여 이를 교정할 예정이다. 또한, 크롬 확장 프로그램을 사용하여 정보 왜곡을 판단하게 되는 것이 정보 수용자에게 어떤 영향을 미칠 수 있을지 사용성 평가 방식을 활용하여 연구하고자 한다.

사사의 글

이 연구는 “A 사업”의 지원을 받아서 수행되었다.

참고 문헌

1. Engledowl, C., & Weiland, T. Data (Mis) representation and COVID-19: leveraging misleading data visualizations for developing statistical literacy across grades 6-16. *Journal of Statistics and Data Science Education*, *29*(2), 160-164. 2021.
2. Kwon, O. N., Han, C., Lee, C., Lee, K., Kim, K., Jo, G., & Yoon, G. Graphs in the COVID-19 news: A mathematics audit of newspapers in Korea. *Educational Studies in Mathematics*, *108*(1), 183-200. 2021.
3. Cleveland, W. S. Graphs in scientific publications. *The American Statistician*, *38*(4), 261-269. 1984.
4. Driessen, J. E., Vos, D. A., Smeets, I., & Albers, C. J. Misleading graphs in context: Less misleading than expected. *PloS one*, *17*(6), e0265823. 2022.
5. Cairo, A. Graphics lies, misleading visuals. In *New challenges for data design* (pp. 103-116). Springer, London. 2015.
6. Rane, C., Subramanya, S. M., Endluri, D. S., Wu, J., & Giles, C. L. ChartReader: Automatic Parsing of Bar-Plots. In 2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI) (pp. 318-325). IEEE. 2021.
7. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211-252. 2015.
8. Davila, K., Kota, B. U., Setlur, S., Govindaraju, V., Tensmeyer, C., Shekhar, S., & Chaudhry, R. ICDAR 2019 competition on harvesting raw tables from infographics (chart-infographics). In *2019 International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1594-1599). IEEE. 2019.