Regular paper http://dx.doi.org/10.13067/JKIECS.2014.9.11.1249

데이터 변형성 기반 유사성 연결을 위한 시각화 알고리즘

김분희*

Visualization Algorithm for Similarity Connection based on Data Transmutability

Boon-Hee Kim*

요 약

사람에 의해 만들어진 수많은 데이터를 기반으로 하는 빅 데이터는 유용한 정보를 얻기 위해 사용된다. 컴퓨터 프로그램의 특징에 인간 메모리의 변형성을 추가 한 기계 학습 기법을 적용 할 경우 보다 유용한 정보를 얻을 수 있다. 그리고 빅 데이터는 이러한 결론을 사용하여 예측된다. 인간은 원래의 데이터와 유사한 데이터를 기억하는 경향이 있다. 그래서 빅 데이터 처리 기술은 인간의 이러한 특성을 반영해야 한다. 본 연구에서는 정보의 선택성을 제공하는 알고리즘을 제안한다. 이 알고리즘은 위 요인들을 반영한 기술이다. 이 알고리즘은 데이터의 변형 특성에 기초하여 유사한 데이터를 결정하는 데 높은 선택성을 가진 데이터를 선택한다.

ABSTRACT

Big data based on numerous data made by the people are used in order to obtain useful information. We can obtain more useful information if it can apply machine learning techniques added deformation of human memory on the characteristics of the computer program. And big data is predicted by using these conclusions. Humans are used to remember similar data as an original data, so big data processing technology should reflect these human characteristics. In this study, this algorithm to provide the selectivity of information is proposed. This algorithm is the technology to reflect the above factors. This algorithm is selected the data with high selectivity to determine similar data based on the deformation characteristics of the data.

키워드

Big Data, Database, Similarity, Transmutability, Visualization
빅데이터, 데이터베이스, 유사성, 변형성, 가시화

Ⅰ. 서 론

빅 데이터는 데이터의 수집은 일반적인 데이터베이스 시스템을 이용하여 처리하기 어렵다. 빅 데이터는 대용량 데이터이며 빠르게 변화하는 특성이 있으며, 데이터가 일정한 형태를 유지하는 것이 아니라 다양성의 특성을 갖는다. 빅 데이터에 대한 데이터 처리 기술은 기존 기술에 비해서 매우 어렵다.

빅 데이터를 처리하는데 있어서 기존의 데이터베이스 기술은 대부분의 관계형 데이터베이스 관리 시스템을 이용하여 작업하는 특성으로 인하여 빅 데이터의 특성 상 적합하지 않다[1-6].

* 교신저자(corresponding author) : 동명대학교 자율전공학부(bhkim@tu.ac.kr)
접수일자 : 2014. 09. 10 심사(수정)일자 : 2014. 10. 20 게재확정일자 : 2014. 11. 10

그림 1에 나타난 바와 같이 빅 데이터는 3V(Volume, Variety, Velocity)의 특성을 갖는다. 즉 빅 데이터는 대용량, 다양한 데이터 소스, 빠른 업데이트의 특성을 갖는다. 사람들에 의해 만들어진 수많은 데이터를 기반으로 유용한 정보를 얻기 위해 빅데이터 처리 시스템이 사용된다. 컴퓨터 프로그램의 특징에 인간 메모리의 변형성이라는 특성을 추가 한 기계 학습 기법을 적용 할 경우 보다 유용한 정보를 얻을 수 있다.

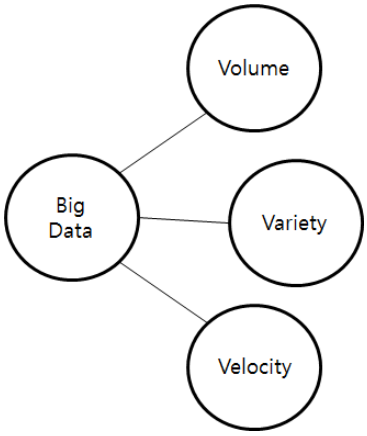

그림 1. 빅데이터의 특성
Fig. 1 Characteristics of big data

본 연구에서는 정보의 선택성을 제공하는 알고리즘을 제안한다. 이 알고리즘은 위 요인들을 반영한 기술로 데이터의 변형 특성에 기초하여 유사한 데이터를 결정하는 데 높은 선택성을 가진 데이터를 선택하는 기법을 적용하고 있다.

II. 관련 연구

정보 시각화는 1999년 Stuart K. Card에 의해 사용한 이후로 다양하게 사용되고 있는데, 컴퓨터를 기반으로 다양한 매체와의 상호작용에 있어서 그래픽적인 요소를 데이터 표현에 사용한 것으로 정의할 수 있다. 정보 시각화의 프로세스는 그림 2에서와 같은 절차를 거친다. 정보 시각화와 사용자의 인지 절차는 원래의 데이터를 조직화 하고, 이를 시각화 하는 과정을 거치게 되어 결국 사용자가 그 데이터를 인지하게 된다. 그런 다음 사용자는 시각화된 데이터와의 인터렉션을 통하여 하나의 의미 있는 정보가 생성되는 것이다.

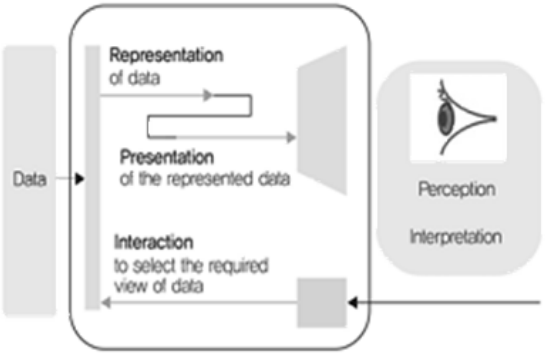

그림 2. 정보 시각화 처리[6]
Fig. 2 Process of information visualization[6]

정보 시각화 절차의 첫 번째 단계인 정보 조직화 단계는 사용자의 정보 인지에 관여하는 단계로써 다양한 종류의 데이터를 그대로 받아들이는 것이 아니라 정보의 종류에 따라 데이터를 분류하고 이를 다시 조직화하여 다양하고 무질서한 데이터에 질서를 부여하는 작업을 의미한다. 무질서한 데이터를 본 사용자는 그대로 인지하는 것이 아니라 나름의 기준으로 분류하고 조직화하여 받아들이는 것이다. 이러한 과정을 거쳐야 분석이 가능한 데이터로써의 의미가 부여되는 것이다.

정보 조직화 단계를 거쳐 정보 시각화 단계에 이르면 시각화 대상이 되는 데이터가 숫자와 같은 정량적인 데이터라면 통계적인 정보를 제공하기 위한 수단의 시각화가 진행된다. 그러나 시각화의 대상이 정성적인 데이터라면 시각화의 방법은 눈에 보이지 않는 다양한 사상 등의 표현을 위한 것으로써 눈에 보이지 않는 것을 시각화함으로써 그 정보는 원 데이터에 비해 이해하기 좋은 형태로 표현될 것이다. 그러므로 정량적인 데이터에 비해 정성적인 데이터의 표현 범위가 더욱 자유롭게 진행되는 측면이 있다. 이러한 시각화의 방법은 점자 출력과 같은 1차원적 방법, 위치 정보와 같은 속성을 이용한 그래프 기반의 2차원적 방법, 3차원의 공간적인 표현이 가능한 3차원적 방법 등이 있다. 이러한 방법 가운데 어떤 표현 방법을 이용할 것인지는 데이터가 가지는 속성에 따라 다르다.

이렇게 정보 시각화 단계를 거친 후 인터렉션 단계를 거치게 되는데, 사용자와 정보간의 상호작용 방법은 미디어의 발달과 함께 많은 연구가 진행되고 있다. 아이콘 기반의 시각적인 요소를 이용한 GUI는 컴퓨터 내의 기능을 이용하는데 있어서 정보를 시각화 하여 사용자에게 직관적이며 효율적인 화면을 제공하였듯이, 최근 터치 기반의 인터페이스가 개발되면서 상호작용 방법은 제스처, 음성, 시선 등의 방법을 통하여 입력되는 형태로 인터랙션의 방법에 많은 변화가 진행되고 있다.

본 연구에서는 이러한 다양한 정보 시각화의 연구 방향 가운데 정성적인 데이터의 정보 시각화 단계에 대한 연구를 빅데이터의 데이터 변형성 측면과 관련하여 진행한다.

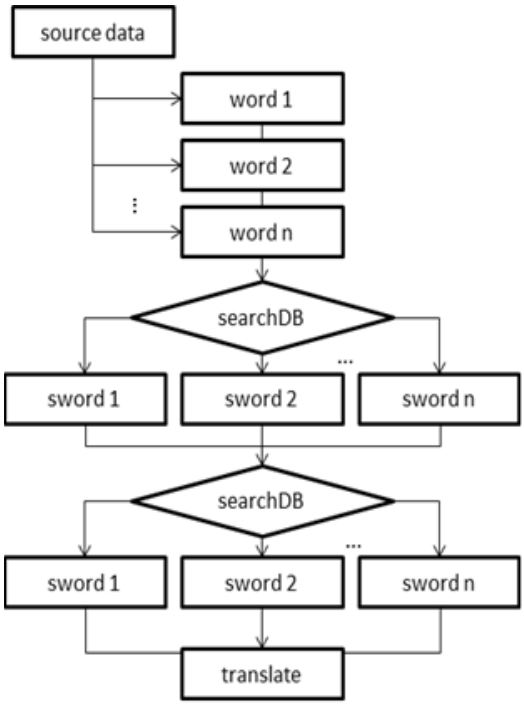

그림 3. 유사성 연결 알고리즘[7]
Fig. 3 Algorithm for similarity connection[7]

III. 제안 알고리즘

인간의 사고는 시간의 요소가 적용되는 과정에서 원래의 데이터는 변형의 과정을 밟게 된다. 이러한 변형의 과정이 있기 때문에 인간의 창의적인 사고가 가능하게 되었다는 측면에서 봤을 때 변형 데이터의 유용성은 빅데이터 시대에서 더욱 그 가치를 발하고 있다. 참고문헌 [6]에서와 마찬가지로 그림 3은 이러한 데이터 변형성에 과정에서 선택된 최종 데이터에서 유용한 데이터를 선택하는 과정을 나타내고 있다. 원본 데이터(source)를 기준으로 추출한 의미어(word)를 기반으로 유사어 데이터베이스를 통한 검색(searchDB)를 통하여 추출된 유사어(sword)를 기준으로 설문 기반의 선택 단어를 유의미한 통계의 결과를 기반으로 기준(mid)을 결정하고 해당 기준 이상의 유사어를 선택(sel)하는 과정을 거친다. 이러한 유사어 선택의 과정에서 원본 데이터와 선택된 유사어와의 밀도를 기준으로 한 선택을 통하여 변형된 데이터를 추출(translate) 할 수 있다.

$$average(high30) = \left(\frac{\sum high30(0kv)}{\sum_{i=A}^{M} 0KV_i} \right) * 100 \quad (1)$$

식1은 상위 30% 단어의 평균 선택율을 나타낸다. 상위 30%의 평균 선택율이 각 단어별 선택율과 비교 대상이 되어 핵심 단어를 결정하는데 이용될 수 있다. 이를 변형하여 각 단어별로 변수를 교체하면 전체 각 단어별 선택율을 계산할 수 있다.

$$\sum_{k=1}^{n} \frac{0kv_i * \frac{alikev_i}{B}}{A} \quad (2)$$

식2는 혼돈율을 나타낸다. 원문자를 선택한 총 개수(A), 유사문자를 선택한 총 개수(B), 개별 원문자 선택 개수(okVi), 개별 유사문자 선택 개수(alikeVi)를 기반으로 혼돈율이 개산된다. 값은 0에서 1의 범위로 혼돈율이 높을수록 원문자와 유사문자와의 혼돈정도가 높은 것으로 나온다.

그림 4는 시각화 알고리즘이다. 시각화 알고리즘은 각각의 원단어 선택 개수의 합(sumofWord)을 기준으

로 원 단어 별로 존재하는 유사단어에 따른 별도의 레이어(layer)를 둔다. 여기서 레이어는 포토샵과 같은 그래픽 프로그램에 존재하는 레이어의 역할과 동일하다. 각 레이어마다 투명도(transparent)를 조정하여 아래쪽 레이어와 위쪽 레이어 간의 농도(density)가 반영되도록 조정하는 작업이 필요하다. 농도에 따라 그리기 메서드(draw)가 실행되어 전체 원단어와 유사 단어 간의 선택 개수에 따른 차이를 그래프로 보여준다.

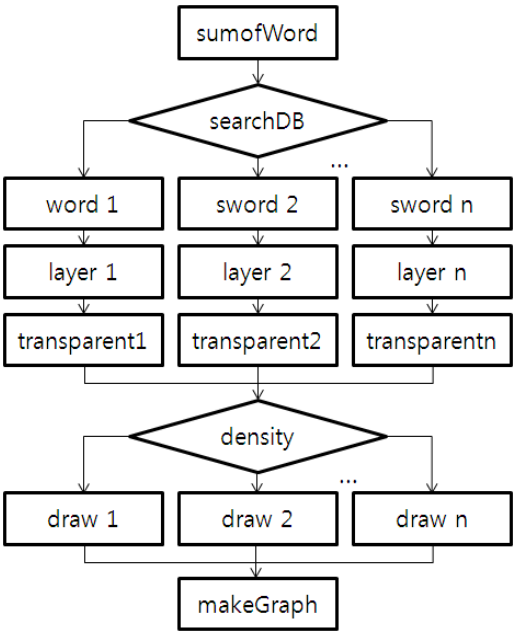

그림 4. 시각화 알고리즘
Fig. 4 Algorithm for visualization

IV. 실험 결과 및 결론

실험의 과정은 시각화 알고리즘에 대하여 Windows 7 32비트 운영체제 상에서 자바 SE 패키지 jdk1.6.0을 기반으로 작성되었다. 그림 4와 같은 과정을 통하여 원단어와 유사단어 간의 선택 개수와 농도에 따라 그래프로 표현된 예는 그림 5와 같다. 이러한 과정을 통하여 단어의 선택 정도를 색의 농도라는 시각적 요소에 의해서 표현됨으로써 즉시적으로 선택의 정도를 파악할 수 있다.

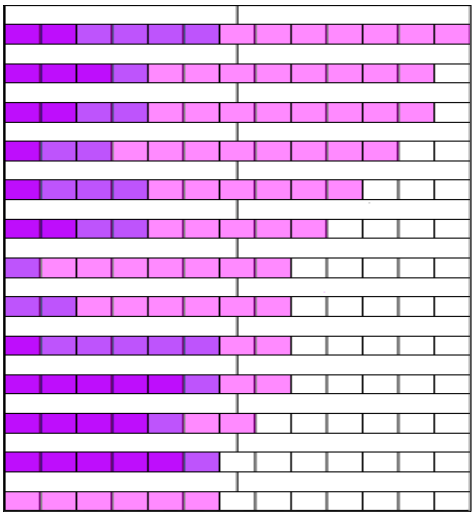

그림 5. 시각화 알고리즘의 결과
Fig. 5 Result of visualization algorithm

그림 6은 원 단어와 이에 대응하는 2개의 유사단어에 대해서 시각화 알고리즘을 수행한 결과이다. 단어별 항목에 대해 원 단어를 첫 번째 바 그래프, 유사단어를 두 번째 바 그래프, 다음 유사 단어를 세 번째 바 그래프로 표현하였다. 대부분의 단어들에 대해 원 단어의 선택 정도가 높았다. L의 경우 유사 단어와의 혼돈의 정도가 높았으며 M의 경우 원 단어에 대한 선택이 압도적인 경우로 유사 단어와의 차이가 컸다.

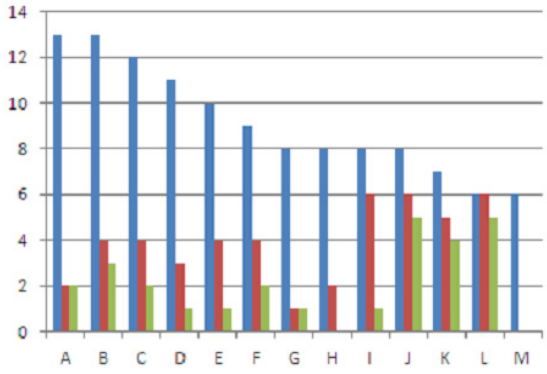

그림 6. 실험 결과
Fig. 6 Experiment results

향후 연구로는 유비쿼터스 환경에서의 많은 연구[8-9]를 바탕으로 사용자가 선택한 유사 단어에 대하

여 원 단어 추천 시스템을 구축하고자 한다.

References

[1] J.-S. Kim and T.-M. Song, “A study on Initial Characterization of Big Data Technology Acceptance-Moderating Role of Technology User & Technology Utilizer,” *J. of the Korea Contents Association*, vol. 14, no. 9, Sept. 2014, pp. 538-555.

[2] J.-H. Song and J.-S. Kim, “ Analysis of the best practices big data services,“ *J. of the Korea Contents Association*, vol. 12, no. 1, Mar. 2014, pp. 32-37.

[3] J.-S. Kim, “Big data analysis Technologies and practical examples,” *J. of the Korea Contents Association*, vol. 12, no. 1, Mar. 2014, pp. 14-20.

[4] B.-Y. Lee, J.-T. Lim, and J. Yoo, "Utilization of Social Media Analysis using Big Data," *J. of the Korea Contents Association*, vol. 13, no. 2, Mar. 2013, pp. 211-219.

[5] L. Boyd, W. Boyd, and G. Vanderheiden, “The Graphical User Interface: Crisis, Danger, and Opportunity,” *J. Visual Impairment Blindness*, vol. 84, no. 10, 1990, pp. 496-502.

[6] H. Shin, J. Lim, and J. Park, "Information Visualization and Information Presentation for Visually Impaired People," *Electronic and Telecommunications Trends*, vol. 28, no. 1, 2013, pp. 81-91.

[7] B. Kim, "Selection Algorithm for Similarity Connection based on Data Transmutability," *Proc. of the Korea Institute of Electronic Communication Sciences,* vol. 7, no. 1, June 2013, pp. 233-234.

[8] H. Kang, Y. Lee, and W. Han, "Energy-Efficient Hierarchical Cluster-Based Routing for Ubiquitous Sensor Networks,“ *J. of the Korea Institute of Electronic Communication Sciences,* vol. 4, no. 3, 2009, pp. 243-246.

[9] H. Lee, H. Lee, and H. Shin, "A Study On Ubiquitous Sensor Network Technologies,“ *J. of the Korea Institute of Electronic Communication Sciences,* vol. 4, no. 1, 2009, pp. 68-74.

저자 소개

김분희(Boon-Hee Kim)

2005년 2월 중앙대학교 컴퓨터공학과(공학박사)

1999년～(주)CEDAR.com 연구원

2005년～현재 동명대학교 자율전공학부 조교수

※ 관심분야 : 분산시스템, P2P 검색 기법, HCI 응용