Journal of the Korean Data & Information Science Society
2014, **25**(6), 1431–1438

http://dx.doi.org/10.7465/jkdi.2014.25.6.1431
한국데이터정보과학회지

부산지역 교통관련 기사를 이용한 비정형 빅데이터의 정형화와 시각적 해석[†]

이경준[1] · 노윤환[2] · 윤상경[3] · 조영석[4]

[1234]부산대학교 통계학과

접수 2014년 10월 3일, 수정 2014년 10월 30일, 게재확정 2014년 11월 7일

요 약

2013년 1월 1일부터 2013년 12월 31일까지의 부산지역지인 국제신문과 부산일보의 기사들 중 제목에 '부산'과 '교통'을 동시에 포함한 2889건의 기사 내용의 관계 또는 관련 있는 데이터에 내재되어 있는 의미 있는 패턴을 찾아내고자한다. 데이터마이닝 (datamining)의 일부인 텍스트마이닝 (textmining)의 기법을 이용하여 사회네트워크분석 (SNA; social network analysis)을 실시하였다. 비정형 데이터의 정형화를 위해 빅데이터의 저장, 처리 및 분석을 위해 자바 기반의 오픈소스 프레임워크인 하둡 생태계 (Hadoop ecosystem)의 HDFS와 맵리듀스 (MapReduce)를 Linux (Ubuntu-12.04LTS) 환경에서 이용하였고, 기존의 R패키지에서 제공되는 사회 네트워크 분석보다 효율적인 시각화를 위해 각 노드 및 선에 비율에 따른 가중치를 주어 색상과 굵기로 해석할 수 있도록 새로운 알고리즘을 구현하였다.

주요용어: 교통, 기사, 빅데이터, 사회 네트워크 분석, 하둡.

1. 서론

빅데이터 (big data)는 인터넷 환경의 발달로 데이터의 트래픽이 폭증했고, SNS 데이터, GPS 위치 데이터 등 그 데이터의 종류도 다양해져 활용 가능성이 늘어났기 때문에 더욱 주목 받고 있다. 특히, 스마트폰 및 태블릿의 확산 등에 따라 유통되는 데이터가 기하급수적으로 증가하면서 빅데이터의 시대가 도래했다. 2010년 Technomy 컨퍼런스에서 구글의 Eric Schmidt는 인류 문명 이래 2003년까지 5EB ($10^1$5byte)가 창출되었으며, 지금은 2일마다 같은 양의 데이터가 신규 창출되고 있다고 하였다.

대량의 데이터는 이미 오래전부터 발생하였다. 하지만 과거에는 이를 축적하는 것만으로도 상당한 시간과 비용이 들었기 때문에 데이터가 일정량을 초과하면 삭제하거나 분석하지 못하고 그대로 저장하여 보관하였다. 하지만, 과학 기술이 발전하여 현재는 저렴한 비용과 시간으로 이들을 저장, 관리할 수 있는 기술이 개발되면서 이들을 분석하고 활용하여 가치있는 정보나 패턴을 찾아내는 것이 매우 중요하게 인식되고 있다. 이러한 빅데이터 분석과 관련하여 Choi 등 (2013), Chae 등 (2013), Kim과 Cho (2013)가 연구하였다.

[†] 본 연구는 미래창조과학부 및 정보통신산업진흥원의 대학IT연구센터육성 지원사업의 연구결과로 수행되었음 (NIPA-2014-H0301-14-1006).
[1] (609-735) 부산광역시 금정구 부산대학로63번길 2, 부산대학교 통계학과, 시간강사.
[2] (609-735) 부산광역시 금정구 부산대학로63번길 2, 부산대학교 통계학과, 석사과정.
[3] (609-735) 부산광역시 금정구 부산대학로63번길 2, 부산대학교 통계학과, 석사과정.
[4] 교신저자: (609-735) 부산광역시 금정구 부산대학로63번길 2, 부산대학교 통계학과, 교수.
E-mail: choys@pusan.ac.kr

본 논문에서는 이러한 빅데이터를 분석하기 위해, 분산 처리 시스템 (distribution processing system)인 맵리듀스 (MapReduce)를 활용하여 비정형 빅데이터를 정형화하고, 이를 분석하고 시각화하기 위해 사회네트워크분석 (social network analysis)를 활용하려 한다. 2013년 1월부터 2013년 12월까지의 부산지역지인 국제신문과 부산일보의 기사 중 부산과 교통이 동시에 언급된 기사들을 추출하여 사용빈도가 높은 단어들을 찾아내고, 그 단어들을 이용하여 각 단어들 간의 특성을 파악하고자 한다. 2절에서는 빅데이터를 처리하는 맵리듀스의 개념과 처리방법 및 사회네트워크분석의 이론에 대해 설명하고, 기존의 사회네트워크분석의 시각적 해석의 추가하여 비정형 데이터를 정형화하여 분석하고 시각화하는 전체적인 작업흐름에 대하여 소개하려 한다. 3절에서는 2절의 기법을 활용하여 국제신문과 부산일보의 지면에 실린 부산지역의 교통이 언급된 기사를 분석하고 시각화하여, 기사에 사용된 단어들 간의 특성을 알아보고 끝으로 4절에서 본 논문을 정리, 요약하겠다.

2. 비정형 빅데이터의 정형화와 사회네트워크분석

2.1. 하둡

최근 빅데이터가 이슈화되면서 빅데이터를 다루는 기술 또한 화두가 되고 있다. 하둡은 대용량 데이터의 분산 저장 및 신속한 처리를 위해 다수의 컴퓨터를 네트워크로 연결하여 하나의 시스템과 같이 사용할 수 있도록 구성한 시스템이다. 하둡 시스템은 주 컴퓨터 (master)들과 종속 컴퓨터들 (slaves)을 하나의 클러스터로 묶어 이루어져 있다 (Ko와 Kim, 2013). 하둡은 오픈소스로 저렴한 가격으로 처리 시스템을 구축할 수 있고, 대용량의 데이터를 분산 처리 시스템을 통해 빠르게 처리할 수 있다는 장점이 있다. 하둡은 크게 분산 데이터 처리를 지원하기 위한 하둡 맵리듀스 (Hadoop MapReduce)와 네트워크를 통해 분산된 데이터를 읽고 쓰기 위한 하둡 분산파일시스템 (Hadoop distributed file system; HDFS), 컬럼 기반의 데이터 베이스로 대규모의 데이터에 빠른 속도로 접근할 수 있게 하는 대용량 데이터 베이스인 에이치베이스 (Hbase)로 구성되어 있다. 하둡은 비즈니스에 효율적으로 적용할 수 있도록 다양한 서브 프로젝트가 제공된다. 이러한 서브 프로젝트가 상용화되면서 Figure 2.1과 같이 하둡 생태계 (Hadoop ecosystem)가 구성되었다. 분산파일시스템과 맵리듀스가 하둡 코어 프로젝트에 해당하며, 나머지 프로젝트는 모두 하둡의 서브 프로젝트이다.

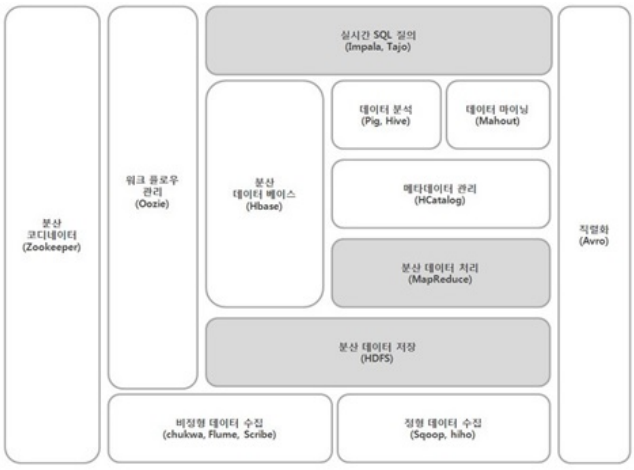


Figure 2.1 Hadoop ecosystem

2.2. 맵리듀스

하둡 맵리듀스는 HDFS 상에서 동작하는 데이터 분석 프레임워크이다. 맵리듀스는 일반 프로그래밍 방법과는 다른 데이터 중심 프로그래밍 모형을 제공한다. 일반적인 분산 환경에서의 프로그래밍은 대개의 프로그래머가 익숙한, 단일 서버에서의 프로그래밍과 달리 분산된 작업의 스케줄링이나 일부 서버의 고장, 서버 간 네트워크 구성 등 많은 문제를 고려해야한다. 맵리듀스에서는 이런 복잡한 문제들이 플랫폼 차원에서 단순화되어 프로그래머는 데이터의 배치 처리를 위한 맵 (mapper)과 리듀스 (reducer) 함수만을 작성하면 되도록 구현되어 있다 (Park 등, 2013).

기본적으로 맵리듀스는 배치 기반의 프로세싱을 수행하며, 대규모의 데이터를 다루기 편리하다. 맵 단계에서는 큰 입력 데이터에 대한 파싱 등의 일을 하고, 리듀스 단계에서는 파싱된 단어들의 수를 합치는 역할을 한다. 맵리듀스를 좀 더 쉽게 이해하기 위해 문장을 구성하고 있는 단어의 빈도수를 계산하는 Figure 2.2를 살펴보면, 입력데이터에 입력된 텍스트는 두 개의 텍스트로 분리된다. 분리된 두 개의 텍스트는 각각의 맵 작업에서 공백을 기준으로 단어들을 분리하고 각 단어에 숫자 1을 부여한다. 맵 단계를 거쳐 생성된 키와 값으로 이루어진 쌍은 다시 리듀스 단계를 거쳐 키를 기준으로 값을 합하고 새로운 키와 값을 가진 쌍을 생성하게 된다.

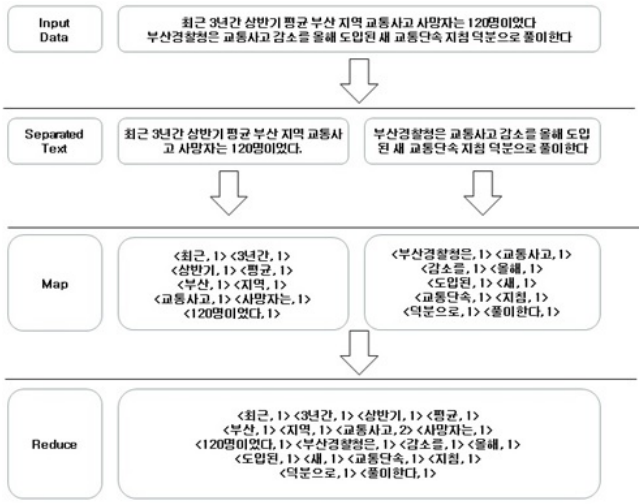


Figure 2.2 MapReduce processing

2.3. 사회네트워크분석

사회네트워크분석은 Barnes (1954)에 의해 처음 사용된 용어로, 사회네트워크 (social network) 현상에서 관계 구조를 시각화하여 중요한 위치에 있는 사람, 기업, 기관 등을 찾아내어 다양한 영역에 적용시킬 수 있는 분석기법이다 (Son, 2002).

사회네트워크란, 다수의 연결된 또는 연결되지 않은 개인 (또는 기관)으로 구성된 사회적 구조를 말하며, 여기서 연결 여부는 친구/친족 관계, 공통 관심, 금융 거래, 친근감, 성 관계, 신뢰도 등 다양하게 정의된다 (Huh, 2010). 사회네트워크분석에서 노드는 꼭지점, 행위자, 구성원 등으로, 연결선은 모서리 또는 이음 등으로 불린다. 사회네트워크분석을 통한 축구경기 분석 (Choi 등, 2011), 사회 연결망 분석

을 이용한 복수전공 유입 및 유출 분석 (Cho, 2012) 등 네트워크에서 가장 중요한 노드나 네트워크 흐름에 가장 중요한 역할을 하는 노드를 찾기 위한 방법으로 사회네트워크 분석이 유용하게 사용되고 있다.

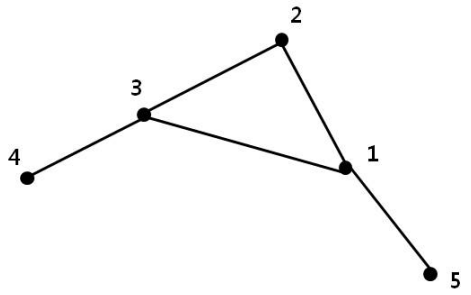
사회네트워크는 행렬 (indicator matrix)로 나타낼 수 있는데 행렬의 행과 열은 각각의 노드를 나타내고, 행과 열이 교차되는 셀에는 노드간의 관계를 나타낸다. 네트워크 이론에서 이와 같은 행렬은 인접행렬(adjacency matrix)이라고 한다. 다음의 예처럼 개의 노드를 가진 인접행렬을 A라 하면, A의 원소 a_{ij}는 노드 i에서 노드 j로 가는 관계를 나타낼 수 있다.

$$A = (a_{ij}) = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \quad a_{ij} \leq 0. \tag{2.1}$$

예를 들어 Figure 2.3의 (a)처럼 1, 2, 3, 4, 5의 번호를 가진 5명의 사람간의 관계를 인접행렬로 나타낼 수 있다. 이 인접행렬을 네트워크 그림으로 나타내면 Figure 2.3의 (b)와 같은 그림이 된다.

	1	2	3	4	5
1	0	1	1	0	1
2	0	0	1	0	0
3	1	0	0	1	0
4	0	0	3	0	0
5	1	0	0	0	0

(a) adjacency matrix



(b) Network

Figure 2.3 Network

그런데 사회네트워크 분석은 노드 간의 관계를 시각적으로 해석하는데 있어서, 노드 간에 어느 정도의 관련이 있는지를 알아보는데 부족함이 많다. 예를 들어 Figure 2.3의 인접행렬에서는 4번을 가진 사람이 3번을 가진 사람에 대한 관계가 다른 사람들 간의 관계보다 높지만 네트워크 그림에서는 이를 알아보기가 어렵다. 따라서 본 논문에서는 이러한 문제점을 해결하고 좀 더 쉽게 시각적으로 해석하고자 새로운 알고리즘을 소개한다.

먼저 기존의 인접행렬 A에 가중치를 부여하여, 새로운 인접행렬 A'을 만든다. 이 때 가중치는 인접행렬 A의 대각원소에는 $1/\sum_{i=j} a_{ij}$을, 대각원소를 제외한 나머지 원소에는 $1/\sum_{i\neq j} a_{ij}$을 준다. 즉, 인접행렬 A의 대각원소들의 합과 대각원소들을 제외한 원소들의 합으로 대각원소들과 나머지원소들을 나누어 준다. 그러면 새로운 인접행렬 A'의 원소들은 각각의 상대비율로 나타나게 된다. 네트워크 그림을 그릴 때 A'의 대각원소는 점의 크기를, 대각원소를 제외한 나머지 원소는 선의 굵기와 색상을 결정한다. 원소들을 상대비율로 나타내어 네트워크 그림에서 노드들 간에 상대적으로 비교하기 쉽다.

$$A^{'} = \left(a_{ij}^{'}\right) = \begin{bmatrix} \frac{a_{11}}{\sum_{i=j} a_{ij}} & \frac{a_{12}}{\sum_{i\neq j} a_{ij}} & \cdots & \frac{a_{1n}}{\sum_{i\neq j} a_{ij}} \\ \frac{a_{21}}{\sum_{i\neq j} a_{ij}} & \frac{a_{22}}{\sum_{i=j} a_{ij}} & \cdots & \frac{a_{2n}}{\sum_{i\neq j} a_{ij}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{a_{n1}}{\sum_{i\neq j} a_{ij}} & \frac{a_{n2}}{\sum_{i\neq j} a_{ij}} & \cdots & \frac{a_{nn}}{\sum_{i=j} a_{ij}} \end{bmatrix} \tag{2.2}$$

선의 색상을 4가지로 나누고, 각각의 4가지 색상에 따라 굵기를 5가지 단계로 나누었다. 선의 비율이 0.25이하이면 노랑색, 0.25초과 0.5이하이면 검정색, 0.5초과 0.75이하이면 빨강색, 0.75초과 1이하이면 파랑색을 사용하였다. 그리고 각 색상안에서 0.05차이마다 5가지의 굵기로 나누었다. 이에 따라 Figure 2.3의 인접행렬을 이용하여 보완된 네트워크 그림을 그려보면 Figure 2.4와 같은 그림을 얻을 수 있다. Figure 2.3의 네트워크 그림에서는 알기 어려웠던 노드 간의 관계를 좀 더 쉽고 자세하게 해석 할 수 있다.

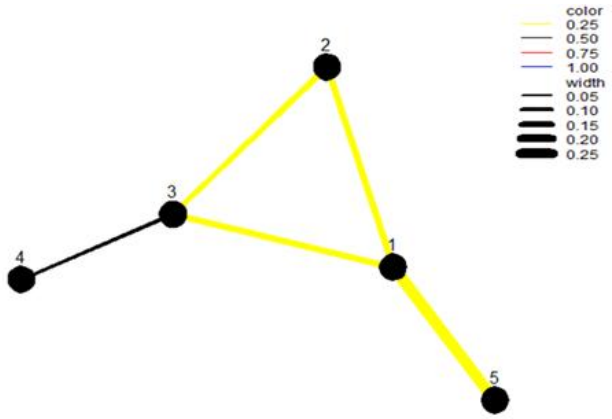


Figure 2.4 Illustration of a modified network analysis

3. 사례 분석

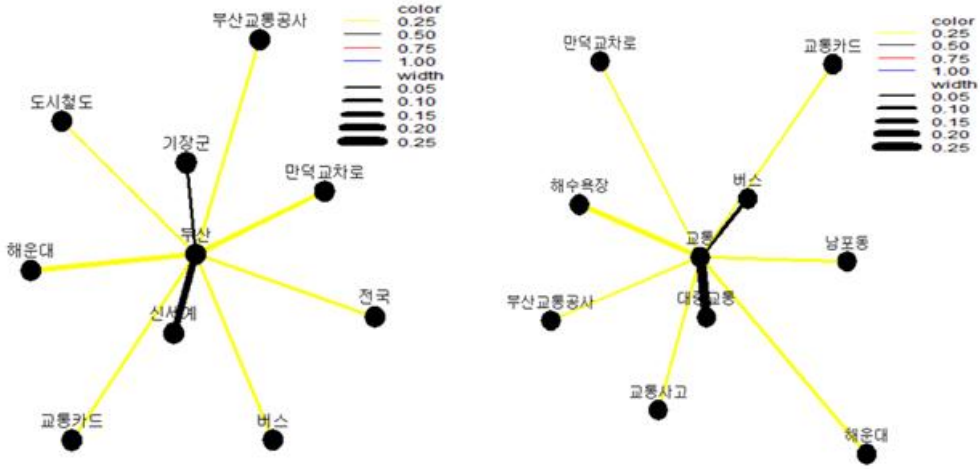
신문 기사는 대표적인 비정형 빅데이터이다. 비정형 데이터를 통계적 방법론을 이용한 분석을 하기 위해서는 정형화 시키는 과정이 필요하다. 본 연구에서는 비정형 빅데이터를 정형화하기 위해 2절에 소개했던 맵리듀스 과정을 활용하였다.

본 연구는 부산 교통과 관련된 기사들의 동향과 이슈를 파악하기 위해 2013년 1월 1일부터 2013년 12월 31일사이의 기간동안 부산지역지인 국제신문과 부산일보의 기사 제목 중 부산과 교통이 동시에 언급된 2889건의 기사들을 사용하였다. 조사목록 리스트를 이용하여 기사에 포함된 조사를 제거하였다. 조사목록 리스트는 국민대학교 강승식 교수의 한글공학-정보검색 연구실 사이트에서 추출한 것이다 (http://nip.kookmin.ac.kr). 그리고 맵리듀스의 맵과 리듀스 과정을 거쳐 8973개의 단어들로 이루어진

인접행렬을 만들었는데, 조사목록에 있는 단어들을 제거한 후 중복을 제거한 단어목록과 신문기사 텍스트를 비교, 순회하여 단어들의 빈도로 이루어진 인접행렬을 만들었다. 특히, 부산과 교통을 각각 포함한 문장에서 단어의 빈도가 높은 10개의 단어들을 이용하여 보완된 네트워크 그림을 보려보았다.

Figure 3.1 (a)는 부산을 포함한 문장을 이용하여 그린 네트워크 그림이다. 부산을 포함한 문장에서 '기장군', '신세계', '도시철도', '해운대', '교통카드', '버스', '부산교통공사', '전국', '만덕교차로'의 빈도가 높게 나타났다. 특히, 그림을 통해 '기장군'과 '신세계'의 빈도가 가장 높음을 알 수 있다. 이것은 실제로 2013년 8월 기장군에 신세계 프리미엄 아울렛이 개장을 했고, 인근지역 주민들의 거센 반발과 교통체증이 극심했으며, 이에 따라 기사화 된 것을 확인 할 수 있었다. 또한, '해운대'는 부산의 대표적인 관광지로서 해마다 많은 기사들이 나오고 있다. '전국'과 '교통카드'는 부산지역 교통카드인 하나로 카드에서 전국 호환 교통카드를 출시하여 많은 이슈가 되었다. '만덕교차로'는 부산지역에서 교통체증이 가장 심한 곳으로 많은 이슈가 되어 많은 기사가 쓰였다.

Figure 3.1 (b)는 교통을 포함한 문장을 이용하여 그린 네트워크 그림이다. 교통을 포함한 문장에는 '대중교통', '버스', '해수욕장', '부산교통공사', '교통사고', '남포동', '만덕교차로', '교통카드', '해운대'의 빈도가 높게 나타났다. '대중교통', '버스', '해수욕장', '해운대'는 부산 관광과 관련된 기사들에서 많이 사용되었고, 특히 여름에 작성된 기사들에서 많이 사용되었다. '만덕교차로'는 Figure 3.1 (a)에서와 만찬가지로 교통체증에 관한 기사에 많이 사용되었다. 사회네트워크분석을 이용하여 2013년 부산지역지인 국제신문과 부산일보의 '부산'과 '교통'에 관련된 기사에 사용된 단어들을 통하여 2013년에 화두가 된 주제에 대해 한눈에 알아볼 수 있었다.



(a) IIllustration of a modified network analysis using the 'Busan'

(b) Illustration of a modified network analysis using the 'Traffic'

Figure 3.1 Illustration of a modified network analysis for examples

4. 결론

빅데이터 분석은 다양한 형태로 축적되어 있는 대용량의 데이터로부터 잠재 되어 있는 가치를 찾아낼 수 있다. 본 연구에서는 맵리듀스를 활용하여 비정형 빅데이터를 정형화하고, 통계적 기법에 적용할 수 있도록 하였다. 더불어, 사회네트워크분석을 활용하여, 정형화된 빅데이터를 시각화하고 해석하였

다. 이에 부산과 교통이 언급된 2013년 1월부터 2013년 12월까지의 국제신문과 부산일보의 기사를 정형화하여 통계적 분석기법에 활용할 수 있도록 하였고, 사회네트워크분석에 적용하여 2013년 부산지역의 화두에 대해 살펴보았다. 그 결과 부산지역 상권과 밀접한 관련이 있는 부산 프리미엄 아울렛의 개장과 이에 대한 인근지역의 극심한 교통체증, 부산의 가장 대표적인 관광지인 해수욕장 특히, 해운대 해수욕장이 큰 화두가 되었다. 그리고 부산지역의 교통체증이 가장 심한 만덕교차로, 전국 호환 교통카드 출시 등의 화제를 알 수 있었다.

이처럼, 대용량의 텍스트를 정독하지 않고도, 빅데이터 분석만으로 많은 정보를 얻을 수 있으며, 나아가 현재에도 폭발적으로 증가하고 있는 빅데이터를 이해하고 분석함으로써, 다양한 분야에서 가치 있는 정보를 얻고, 활용 할 수 있을 것이라고 기대하여 본다.

References

Barnes, J. (1954). Class and committees in a Norwegian island parish. *Human Relations*, **7**, 39–58.

Chae, M., Kang, M. and Kim, Y. (2013). Documents recommendation using large citation data. *Journal of the Korean Data & Information Science Society*, **24**, 999–1011.

Cho, J. (2012). Inflow and outflow analysis of double majors using social network analysis. *Journal of the Korean Data & Information Science Society*, **23**, 693–701.

Choi, H., Park, H. and Park, C. (2013). Support vector machines for big data analysis. *Journal of the Korean Data & Information Science Society*, **24**, 989–998.

Choi, S., Kang, C., Choi, H. and Kang, B. (2011). Social network analysis for a soccer game. *Journal of the Korean Data & Information Science Society*, **22**, 1053–1063.

Huh, M. (2010). *Introduction to social network analysis using R*, Freedom Academy, Seoul.

Kim, Y. and Cho, K. (2013). Big data and statistics. *Journal of the Korean Data & Information Science Society*, **24**, 959–974.

Ko, Y. and Kim, J. (2013). Analysis of big data using Rhipe. *Journal of the Korean Data & Information Science Society*, **24**, 975–987.

Park, J., Lee, Y., Kang, D. and Won, J. (2013). Hadoop and MapReduce. *Journal of the Korean Data & Information Science Society*, **24**, 1013–1027.

Son, D. (2002). *Social network analysis*, Kyungmoon Publishers, Seoul.

Journal of the Korean Data & Information Science Society
2014, **25**(6), 1431–1438

http://dx.doi.org/10.7465/jkdi.2014.25.6.1431
한국데이터정보과학회지

Structuring of unstructured big data and visual interpretation[†]

Kyeongjun Lee[1] · Yunhwan Noh[2] · Sanggyeong Yoon[3] · Youngseuk Cho[4]

[1234]Department of Statistics, Pusan National University

Received 3 October 2014, revised 30 October 2014, accepted 7 November 2014

Abstract

We analyzed the articles from "Kukje Shinmun" and "Busan Ilbo", which are two local newpapers of Busan Metropolitan City. The articles cover from January 1, 2013 to December 31, 2013. Meaningful pattern inherent in 2889 articles of which the title includes "Busan" and "Traffic" and related data was analyzed. Textmining method, which is a part of datamining, was used for the social network analysis (SNA). HDFS and MapReduce (from Hadoop ecosystem), which is open-source framework based on JAVA, were used with Linux environment (Uubntu-12.04LTS) for the construction of unstructured data and the storage, process and the analysis of big data. We implemented new algorithm that shows better visualization compared with the default one from R package, by providing the color and thickness based on the weight from each node and line connecting the nodes.

Keywords: Articles, big data, Hadoop, social network analysis, traffic.

[†] This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2014-H0301-14-1006) supervised by the NIPA(National IT Industry Promotion Agency).

[1] Instructor, Department of Statistics, Pusan National University, Busan 609-735, Korea.
[2] Master's course, Department of Statistics, Pusan National University, Busan 609-735, Korea.
[3] Master's course, Department of Statistics, Pusan National University, Busan 609-735, Korea
[4] Corresponding author: Professor, Department of Statistics, Pusan National University, Busan 609-735, Korea E-mail: choys@pusan.ac.kr