13장: 빅데이터 컴퓨팅

13장: 빅데이터 컴퓨팅

- 빅데이터 개요
- 빅데이터 컴퓨팅 방법
- 빅데이터 마이닝
- 빅데이터 시각화

이장의 목적

- 빅데이터란 무엇인가?
- 빅데이터 컴퓨팅을 위한 방법은?
- 빅데이터 컴퓨팅 절차는?
- 빅데이터 마이닝 방법은?
- 빅데이터를 시각화하는 방법은?

13.1 빅데이터 개요

빅데이터 개요

- IoT 장치 및 서비스, 스마트폰 등의 확산
  - 데이터의 양이 폭발적으로 증가
  - 데이터 형태도 다양화
- 데이터의 양
  - 메가→기가→테라→페타→제타
- 데이터의 형태
  - 정형 → 비정형 데이터

- 이러한 양과 형태의 데이터를 빅데이터라 함
- 빅데이터 분석을 통한 효과적인 의사 결정을 수행

빅데이터 정의

- 빅데이터
  - 데이터베이스 관리 시스템에서 수집, 저장, 관리, 분석할 수 있는 대량의 정형 또는 비정형 자료의 모음
- 빅데이터 컴퓨팅
  - 대용량의 데이터를 모아 저장하고 분석하는 컴퓨터 작업
  - 빅데이터에서 가치를 추출하고, 결과를 분석하는 작업
- 빅데이터 컴퓨팅 활용 예
  - 카드 회사에서는 소셜 빅데이터와 카드 결제 정보를 분석하여 소비 유형과 특성 등 다양한 데이터를 추출하고 연계하여 금융 분야의 타겟 마케팅에 활용

빅데이터 컴퓨팅 과정

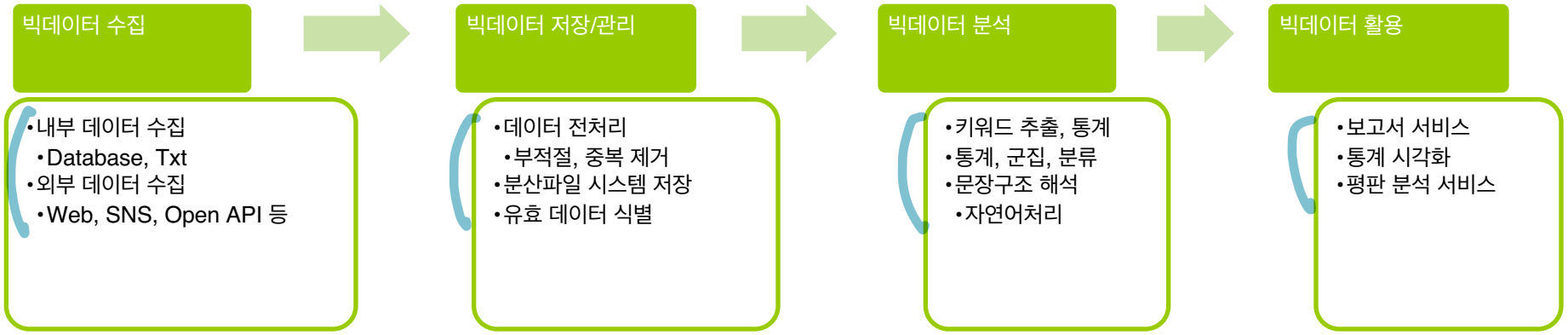


그림 12.1 빅데이터 컴퓨팅 과정

13.2 빅데이터 수집

12.2.1 빅데이터 수집 개요

- 필요한 데이터를 시스템 내부외부에서 주기적으로 수집하는 활동
- 서비스결정→ 수집할 원천데이타 탐색 --> 수집방법(주기,데이터위치, 수집방법) → 저장

- 원천 데이터 탐색에서
  - 수집난이도, 비용, 안정성, 수집주기, 데이타위치, 수집방법,

12.2.2 빅데이터 종류

- 데이터를 검색하여 수집하고
- 변환 과정을 통한 잘 정리된 데이터를 확보하는 과정
- 원본 위치에 따른 분류
  - 내부 데이터 수집
    - 내부 파일 시스템이나 데이터베이스 관리 시스템, 센서 등의 데이터
  - 외부 데이터 수집
    - 외부에서 인터넷으로 수집
- 데이터의 유형에 따른 분류

데이터 유형	데이터 종류	수집기술
정형 데이터	RDB. 스프레드 시트	ETL, FTP, Open API
반정형 데이터	HTML. XML. JSON. 웹문서. 웹로그. 센서 데이터	Crawling, RSS, Open API, FTP
비정형 데이터	소셜 데이터. 문서(워드, 한글). 이미지. 오디오. 비디오. IoT	Crawling, RSS, Open API, Streaming, FTP

- 정형데이터
  - RDB의 relation
- 반정형데이터, XML
  - 메타데이타를 포함하여 메타데이타가 자료를 설명
  - <학생>
    - <이름> 홍길동 </이름> <나이> 50</나이> <주소>대구</주소>
  - </학생>
- 비정형데이터
  - 언어 분석이 필요한 자료
  - 책, 이미지, 동영상
  - html은 mark 가 있는 (메타데이타) 자료이지만 필요한 자료는 내부에 텍스트 형태로 존재하여 자연어처리가 필요함

- 내부데이타
- 외부데이타: 자료가 조직의 외부에
  - 수집대상자료를 분석하여 수집시스템을 구축
    - 대부분 비정형, 환경통제불가, 자료수집 불가 때를 대비한
    - 서비스 관리 정책 필요

- 빅데이터 수집방법
  - 크롤링 기술
  - Open API
  - FTP
  - RSS
  - Log Aggregator
  - RDB Aggregator

13.3 빅데이터 저장/처리

12.3.1 빅데이터 저장/처리

- 빅데이터를 효율적으로 저장하고 관리하는 과정

- NoSQL : MongoDB, Cassandra, HBase 등 NoSQL DB
- 분산 파일 시스템 : HDFS

12.3.1 빅데이터 저장/처리

- 빅데이터 전처리
  - 필터링: 불필요 자료 제거
  - 데이터 유형 변환
  - 정제 작업 : 손실값 보정, 잡음제거, 자료에라 중복제거
- 빅데이터 후처리
  - 분석하기 쉽게 일관성 있는 형식으로 변환
  - 평활화 → 집계 → 일반화→ 정규화 → 속성 생성
- 데이터 통합
  - 여러 곳(장치)에서 수집된 데이터를 관련성이 있는 데이터끼리 결합

12.3.2 NoSQL

- NoSQL : MongoDB, Cassandra, HBase 등 NoSQL DB
- Relation 사용하지 않음 == table 구조의 자료 및 연산 사용하지 않음.
- SQL DB의 주요특징(일관성 및 유효성을 보장하지 않음.

12.3.3 분산화일시스템

- 분산 파일 시스템
  - 이론적으로 무제한 용량의 자료 저장 시스테
  - 구글 파일 시스템, 하둡 분산파일 시스템, 클라우드 스토어
- 하둡
  - 분산 파일 시스템 / 분산 병렬 처리를 수행하는 맵리듀스
  - 맵리듀스는 구글 분산 파일 시스템에서 빅데이터를 분산 병렬 처리를 간단하게 할 수 있는 소프트웨어
  - 하둡 분산 파일 시스템은 분산 환경에서 다양한 형태, 초대용량의 데이터를 안전하게 저장할 뿐 아니라, 저장되어 있는 데이터를 빠르게 처리

하둡과 맵리듀스 활용

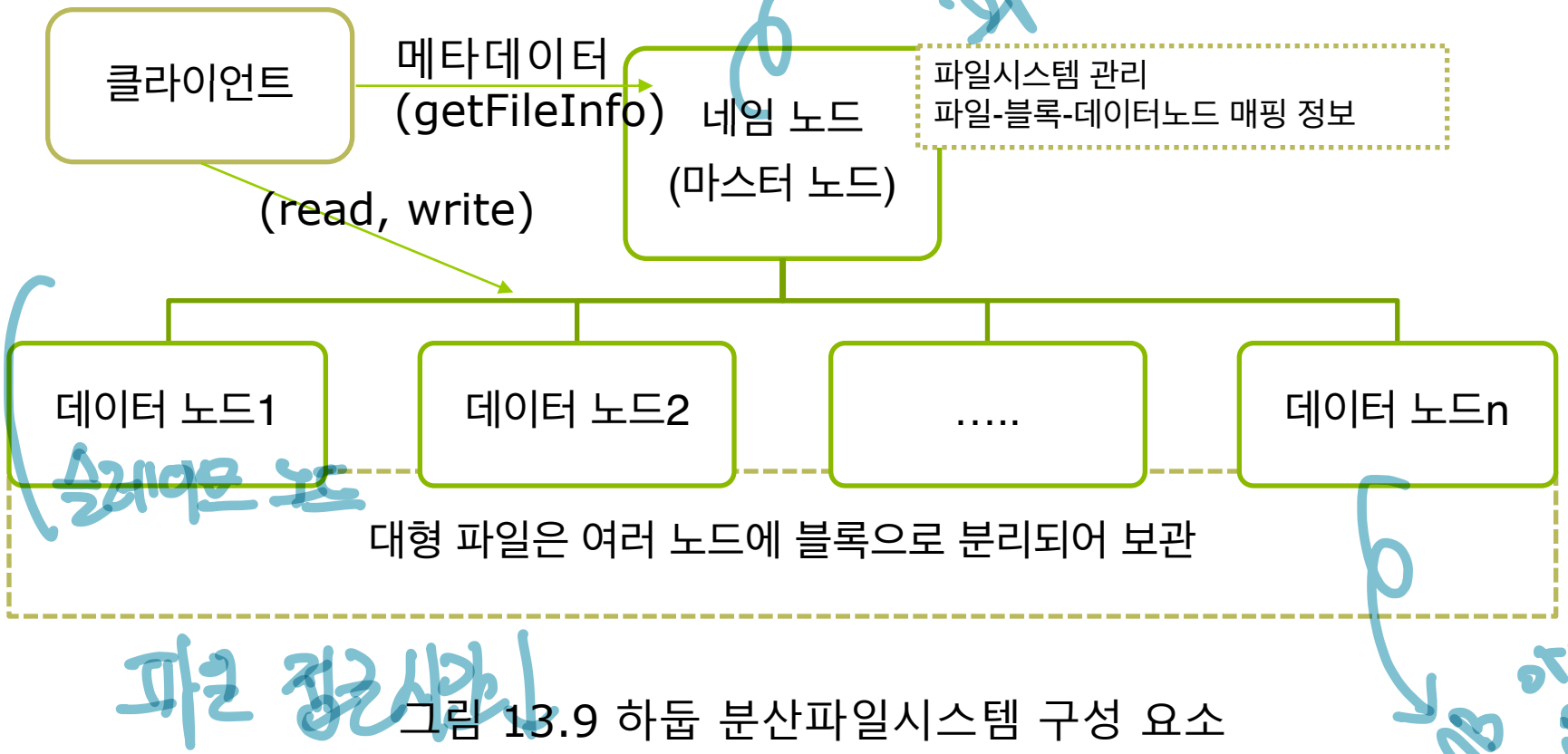


그림 13.9 하둡 분산파일시스템 구성 요소

하둡과 맵리듀스 활용

- 하둡 분산파일시스템
  - 마스터 노드가 동작하는 서버와 슬레이브 노드가 동작하는 서버
  - 마스터 노드는 네임 노드로 구성되며 분산 파일 시스템에서 사용하는 모든 슬레이브 노드를 관리
  - 네임 노드는 파일 시스템의 네임 스페이스를 관리하는 서버
    - 파일 시스템의 트리, 모든 파일과 디렉터리 구조, 액세스 권한 등의 메타 데이터를 관리하고 블록에 대한 배치 정보를 관리
  - 데이터 노드는 특정 파일을 분할하여 블록으로 저장하며
  - 데이터 노드 사이에 데이터 복제를 통해 데이터를 신뢰성을 유지
  - 보조 네임 노드는 네임 노드에서 관리하는 파일시스템의 이미지 정보를 백업하고 네임스페이스 이미지와 주기적으로 병합하는 기능

하둡과 맵리듀스 활용

- 맵리듀스
  - 병렬로 분산 프로그래밍 모델
  - 범용 프로그래밍 언어로 맵 함수와 맵 리듀스 함수를 사용하여 병렬 처리
- 하둡 에코 시스템
  - 하둡은 빅데이터 저장과 처리를 위한 기본 기능만 제공
  - 이를 보완하기 위해 하둡 에코 시스템을 둠
  - 하둡의 기능을 보완하는 부속 오픈 소스 소프트웨어
  - 빅데이터를 좀 더 보다 효율적으로 분석

12.4 빅데이터 분석

12.4.1 빅데이터 분석(마이닝) 개요

- 의사결정 즉시성이 떨어짐
- 장기적, 전략적-때로는 일회성 거래처리나 행동분석 지원
- 처리의 복잡도 높음
- 실시간 데이터 분석에 다소 부적합

- 데이터 저장소 ➔ 새로운 상관관계, 패턴, 추세 등을 발견하는 과정

- 예) 빅데이터 통계, 데이터 마이닝, 텍스트마이닝, 예측 분석, 최적화, 평판 분석 소셜네트워크분석, 소셜 빅데이터 분석 등

12.4.2 빅데이터 분석(마이닝)

- 문자 마이닝, 오피니언 마이닝, 웹 마이닝 소셜 네트워크 분석, 군집분석 등을 활용
- 문자 마이닝
  - 문자 분석, 지식 발견, 문서 마이닝
  - 자연어 처리 방법을 주로 활용
  - 문서 분류, 문서 군집, 메타데이터 추출, 정보 추출 등으로 구분
- 오피니언 마이닝
  - 소셜미디어에서 정형/비정형 문자의 긍정, 부정, 중립의 선호도를 판별
  - 특정 주제에 대한 사람들의 의견을 모아 문장을 분석
  - 브랜드 모니터링, 버즈 모니터링, 온라인 인류학, 시장 영향력 분석, 대화 모니터링, 온라인 소비자 이해 등

웹 마이닝

- 웹 로그 정보나 검색어 분석
- 웹 구조 마이닝
  - 웹 사이트의 노드와 연결 구조를 분석하는 기법
  - 하이퍼링크에서 패턴을 찾아냄
- 웹 사용 마이닝
  - 웹서버 로그 파일 분석을 통해 웹 사이트 개선이나 고객 특성을 반영한 맞춤형 서비스를 제공
- 웹 콘텐츠 마이닝
  - 콘텐츠에서 사용자가 원하는 정보를 빠르게 찾는 기법
  - 예 : 웹 페이지에서 특정 상품의 설명이나 독자의 상품평과 같은 필요한 정보를 추출하는 작업
- 웹 크롤러- 웹페이지를 찾아다니며 자료를 수집
  - 스파이더, 웜, 로봇, 봇
  - 범용, 포커스, 토픽 크롤러

12.4.3 분류, 군집, 연관 관계

- 분류
  - 일정한 빅데이터에서 특별히 정한 기준에 따라 데이터를 분류
  - ex) 다른 그룹으로 이동한 구성원 분류, 매주 토요일에 정해진 물건을 구입하는 사람
- 군집
  - 어떤 정해진 특성을 공유하는 데이터 그룹을 찾는 것
  - 미리 정한 기준 없이 유사한 특성을 공유하는 데이터 끼리 그룹화
  - ex) 일찍 등교하는 학생과 매일 지각하는 학생으로 그룹화
  - 계층적 군집 :
    - 데이터 사이의 유클리드 거리를 계산하고 유사도로 군집화
  - K-mean 군집 :
    - 유클리드 거리를 구하고 그 평균에 있는 데이터와 각 데이터 사이의 거리를 구하여 군집

분류, 군집, 연관 관계

- 연관 관계 분석
  - 데이터 항목 사이의 종속 관계를 찾아냄
  - ex)
    - 등산화를 팔면 배낭도 같이 팔리는가 → A 제품을 구입한 고객에게 어떤 제품을 함께 팔 수 있을지를 추축가능
    - 카드 구매 유형 간에는 어떤 관계가 있는지 등
  - ‘항목 A는 항목 B이다’ 혹은 ‘항목 A 그리고 항목 B는 곧 항목 C’라는 형식의 연관규칙을 만들 수 있음
    - 그러므로 등산화를 구입하는 소비자는 반드시 배낭을 구입할 확률이 98%라면, 매장에서는 등산화 옆에 배낭을 전시
  - ‘국가’ 라는 단어가 나타나면, ‘국경’ , ‘국민’ , ‘애국심’ 이라는 단어가 나올 확률도 높음

분류, 군집, 연관 관계

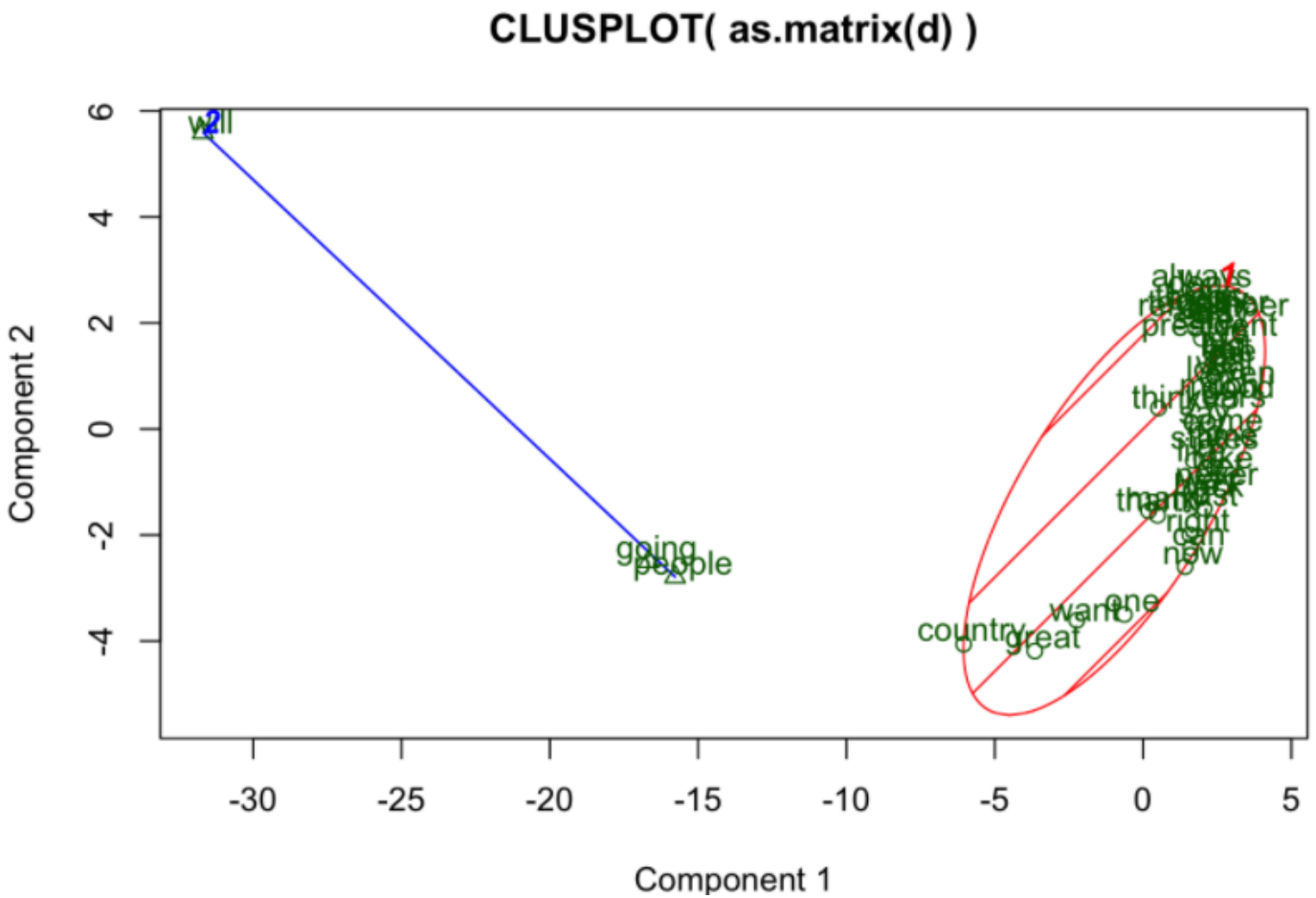


그림 12.9 Clusplot를 이용한 군집의 시각화 예

12.4.4 스트림 데이터 마이닝

- 실시간 스트림 데이터
  - 지속적인 데이터 질의가 가능해야 하며 순차적인 접근도 가능해야 함
  - 통화 정보 분석, 침입탐지 시스템, CCTV와 같은 비디오 기반 감시 환경에서 이상 행동분석, 주가 분석 등
- 특징
  - 대량성: 데이터 크기가 커서 분석이 어렵고 해석이 어려움
  - 연속성: 데이터 수집이 완료도지 않고, 계속 실시간으로 축적
  - 가변성: 데이터는 시간의 흐름에 따라 패턴이 변함
- 바운드 데이터
  - 저장소에 저장되고 더 이상 증가하거나 변경이 없는 상태로 유지
- 언바운드 데이터
  - 데이터 수가 정해져 있지 않고 계속하여 추가되는 데이터

스트림 데이터 마이닝

- 언바운드 데이터를 처리하는 방식
  - 일괄 처리 방식과 스트리밍 처리 방식
  - 일괄 처리 방식
    - 스트리밍 데이터를 일정 시간 단위로 모은 후, 일괄로 처리

- 스트리밍 처리 방식
  - 필터링 방식
  - 내부 조인 방식
  - 윈도우 방식

스트림 데이터 마이닝

- 필터링 방식
  - 스트리밍 데이터 중 특정 데이터만 필터링하여 저장
  - ex) 웹 로깅 데이터를 수집하고 특정 인터넷 프로토콜이나 국가로 들어오는 데이터만 필터링
- 내부 조인 방식
  - 두 개의 데이터 스트림에서 들어오는 값을 비교하여 매칭
  - ex) [모바일 뉴스 앱에서 보고 있는 콘텐츠에 대한 데이터] JOIN [지도 앱을 통한 사용자의 현재 위치] → 사용자가 어떤 위치에서 어떤 뉴스를 보고 있는지를 분석 가능
- 윈도우 방식
  - 고정 윈도우 방식
  - 슬라이딩 윈도우 방식
  - 세션 윈도우 방식

12.4.4 스트림 데이터 마이닝

- 고정 윈도우 방식
  - 일정 시간 단위로 시간 윈도우를 나누어 분석
- 슬라이딩 윈도우 방식
  - 현재 시간으로 부터 ±m 시간 전후의 데이터를 매 n 시간 마다 추출
- 세션 윈도우 방식
  - 세션 시작에서 부터 반응이 없어지는 시간까지 한 세션으로 묶어서 처리
  - 세션 타임아웃이 20분일 경우
    - 데이터가 1:00, 1:03, 1:17, 1:40분에 스트리밍 될 경우
    - 1:00 이후에 20분 동안 (1:20까지) 데이터가 들어오지 않았기 때문에, 1:00,1:03,1:17는 한 개의 세션이 되고, 1:40은 새로운 세션이 시작

12.5 빅데이터 시각화

- 자료를 분석하여 한 번에 볼 수 있도록 도표나 차트 등으로 보이는 작업
- 시각화는 데이터에서 정보를 파악하는 시간이 절감되므로 직감적이고 즉각적인 상황판단이 가능
- 시각화 도구에는 구글 차트, 파이썬, R, MATLAB 등

- 문자 데이터 시각화
- 지리 데이터 시각화
- 지리 데이터 시각화

문자 데이터 시각화

- 비정형인 문자 데이터에서 의미가 있는 정보를 연관 짓고 찾아내어 시각화
- 관련성 있는 문자, 문서들의 클러스터링, 추출, 문서 요약 등에 활용
- R의 플로팅과 워드 클라우드로 시각화

- 빈도수 플로팅

```
>library(ggplot2)
> p <- ggplot(subset(wf, freq>50), aes(x = reorder(word, -freq), y = freq))
+ geom_bar(stat = "identity") + theme(axis.text.x=element_text(angle=45,
hjust=1))
```

- 워드 클라우드

```
> dark2 <- brewer.pal(6, "Dark2")
> wordcloud(names(freq), freq, max.words=100, rot.per=0.2,
colors=dark2)
```

문자 데이터 시각화

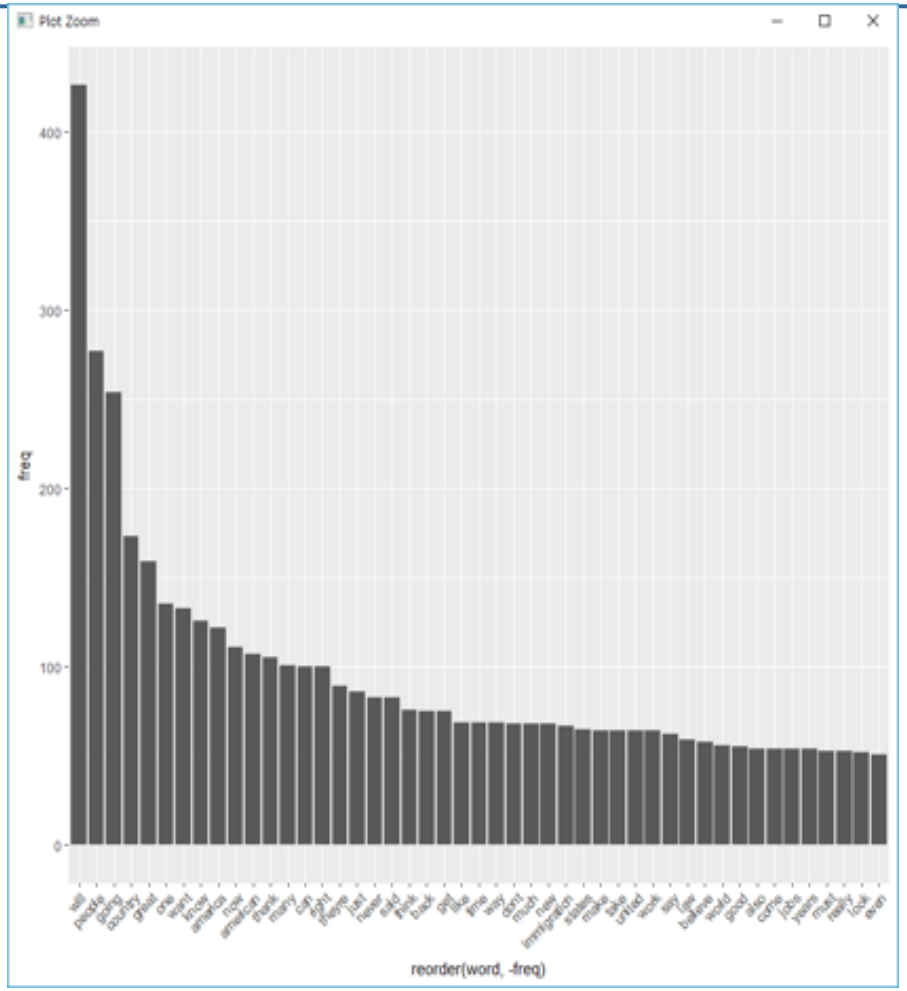


그림 13.11 단어 빈도 수 차트

문자 데이터 시각화

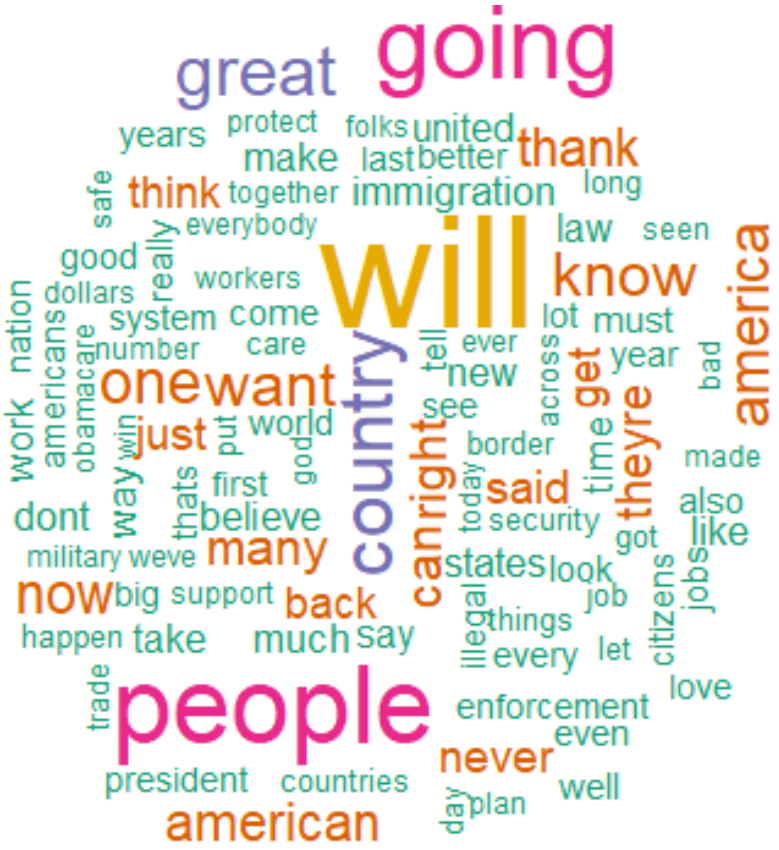


그림 13.12 워드 클라우드의 예 : 단어 빈도수

지리 데이터 시각화

- 지역 데이터를 시각화
- 지도에 이동 동경로와 시간, 위도, 경도, 한글지명 등을 포함하고 있는 데이터 프레임을 구축
- 데이터를 지도에 표시

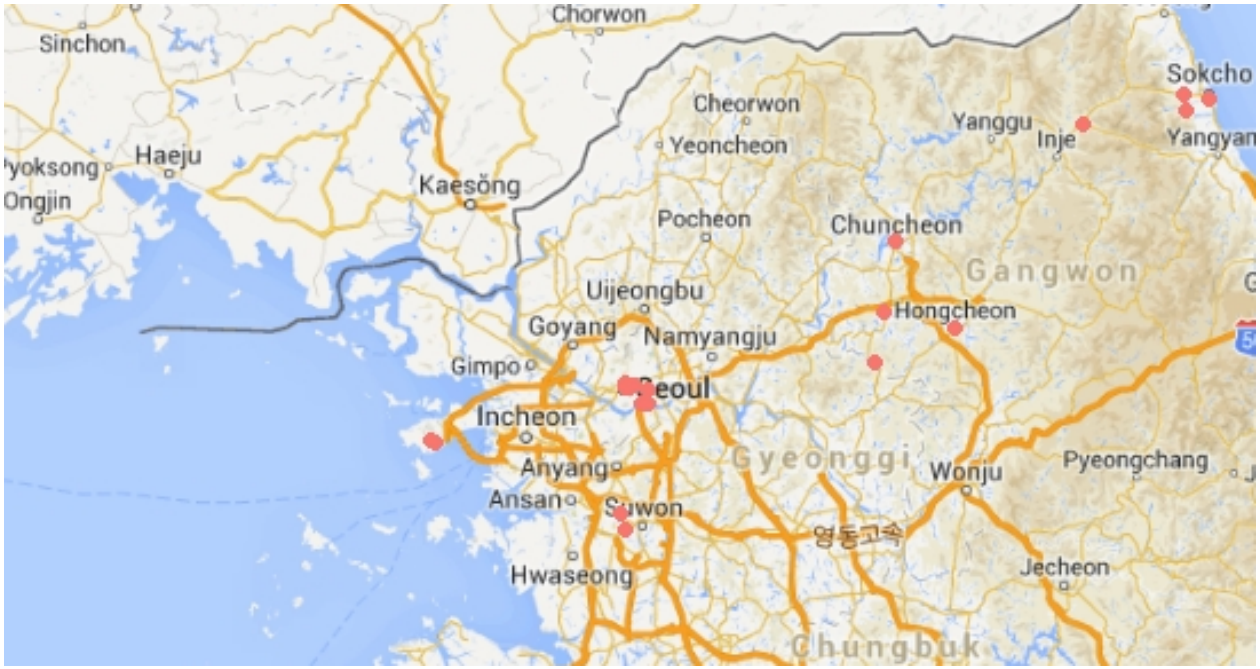
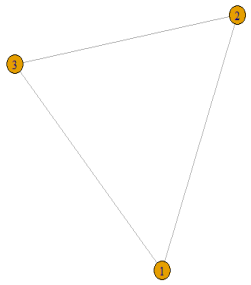


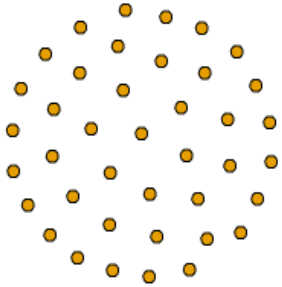
그림 12.12 데이터를 지도에 시각화

네트워크 시각화

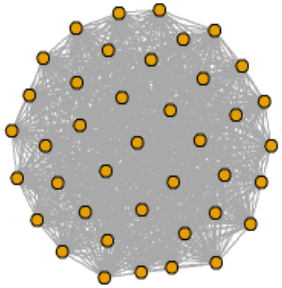
- 수송망, 교통망, 전화망, 인간 관계, 사물인터넷 망 등 연결 구조 표현
- 사회와 자연 현상을 네트워크 형태로 모델링하고 특성을 분석
- 간선에 가중치를 주어 노드 사이의 관계를 나타낼 수 있음
- 네트워크 분석
  - 최소 경로
  - 친구 관계, 협업 관계, 인터넷 연결망 구조, 분자 그래프 등을 분석
- 소셜네트워크 분석
  - 결과를 통해 광고나 여러 목적으로 활용
- 네트워크 시각화 패키지
  - R 패키지에는 igraph, statnet, RSiena

네트워크 시각화

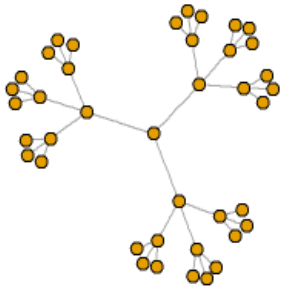




```
eg <- make_empty_graph(40)
plot(eg, vertex.size=10,
vertex.label=NA)
```



```
fg <- make_full_graph(40)
plot(fg, vertex.size=10,
vertex.label=NA)
```

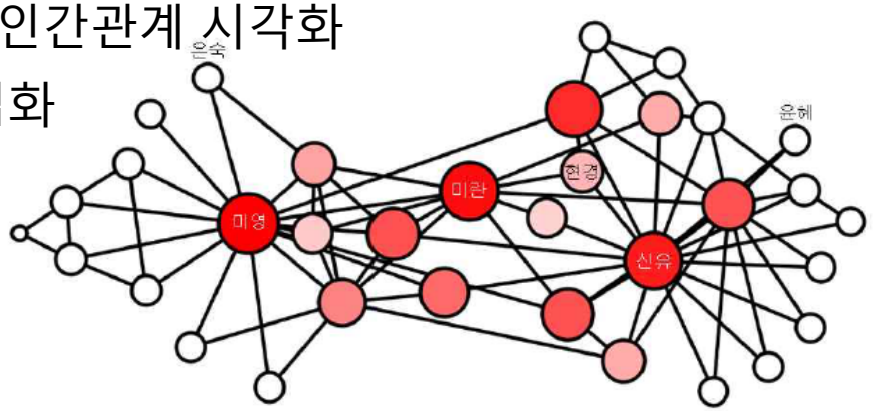


```
tr <- make_tree(40, children = 3, mode =
"undirected")
plot(tr, vertex.size=10, vertex.label=NA)
```

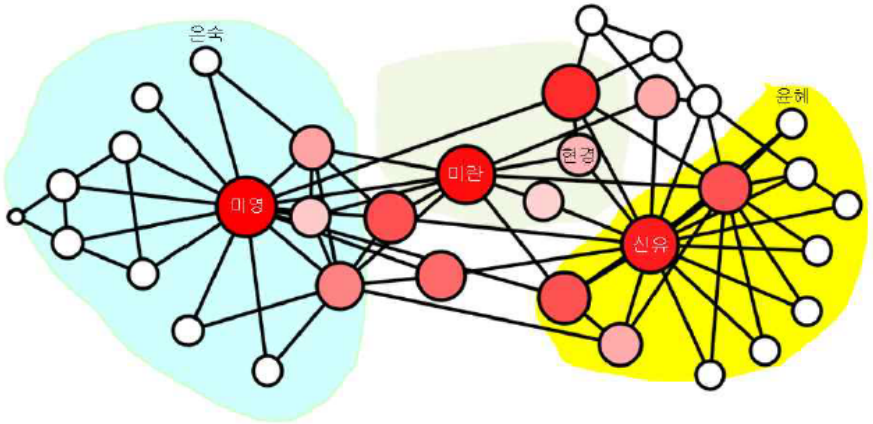
그림 12.14 다양한 모양의 네트워크 시각화

네트워크 시각화

- 소셜네트워크 인간관계 시각화
- 사람관계 군집화



(a) 단순 관계망



(b) 인간관계의 군집화

그림 12.14 소셜 네트워크의 시각화