빅데이터 분석 기술

최도진

01 빅데이터의 개념과 처리 과정

❖ 빅데이터 처리 과정과 기술

그림 1-7 빅데이터 처리 과정 [09]

05 빅데이터 분석 기술

개요

- 데이터마이닝Data Mining
  - 저장된 방대한 양의 데이터 안에서 자동으로 체계적이고 통계적인 규칙이나 패턴을 찾아내는 데이터베이스 파생 기술.
  - 다른 말로 KDDKnowledge-Discovery in Databases라고도 하는데, 각종 데이터베이스에 내재된 의미 있는 지식의 발견이라는 뜻.
  - 데이터 마이닝 기법은 통계학 분야에서 발전한 탐색적 데이터 분석, 가설 검증, 다변량 분석, 시계열 분석, 일반선형모형 등 방법론과 데이터베이스 분야에서 발전한 OLAPOn-Line Analytic Processing;온라인 분석 처리, 인공지능 분야에서 발전한 SOMSelf-Organizing Map; 자기조직화지도, 신경망, 전문가 시스템 등 기술적 방법론 등에 사용.

05 빅데이터 분석 기술

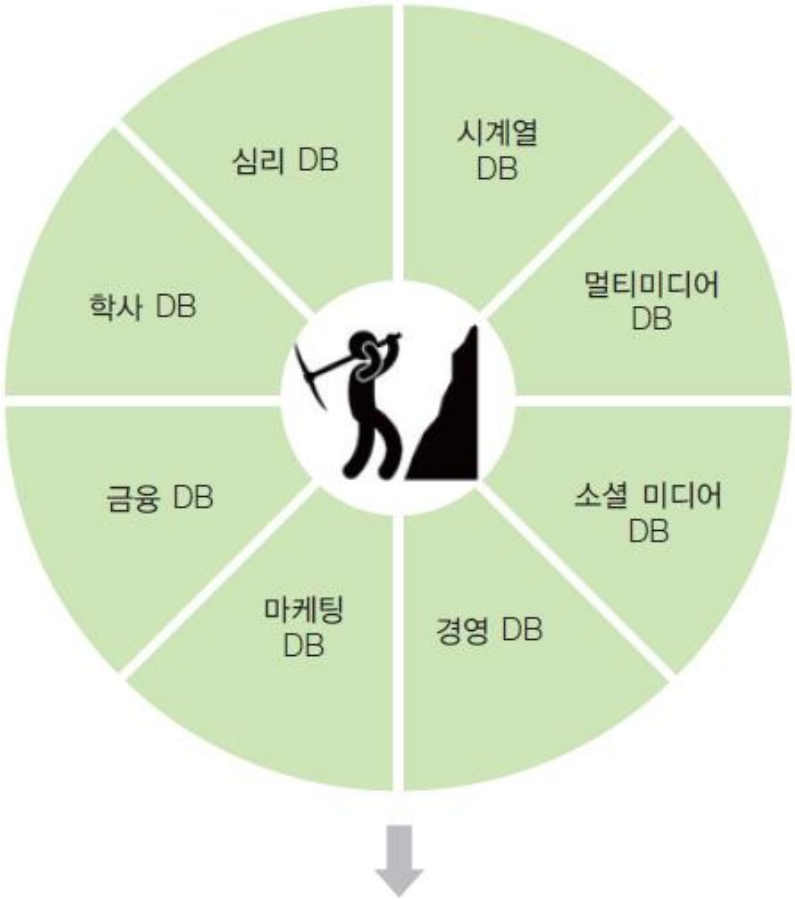

그림 5-1 데이터 마이닝의 대상과 결과

05 빅데이터 분석 기술

- 데이터 마이닝 기술
  - 분류Classification : 일정한 집단에서 특정한 정의를 이용하여 분류 및 구분을 추론함. 예를 들어, 경쟁자에게로 이탈한 고객을 분류해 내는 기술을 들 수 있다.
  - 예측Forecasting : 방대한 양의 데이터 집합의 패턴을 기반으로 미래를 예측. 예를 들어, 수요를 예측하는 기술을 들 수 있다.
  - 시계열Time-Series분석 : 시간의 변화에 따라 일정한 간격으로 연속적인 통계 숫자를 저장한 시계열 데이터에 바탕을 둔 분석 방법. 예를 들어, 매일 주식의 값을 저장하는 시계열 데이터를 분석하는 기술을 들 수 있다.
  - 회귀분석Regression : 하나 이상의 변수 간의 영향이나 관계를 분석 및 추정하는 기술을 들 수 있다.
  - 군집화Clustering : 구체적인 특성을 공유하는 군집을 찾음. 군집화는 미리 정의된 특성의 정보가 없다는 점에서 분류와 다름. 예를 들어, 비슷한 행동 집단을 구분해 내는 기술을 들 수 있다 [01].
  - 연관 규칙Association Rule : 동시에 발생한 사건 간의 관계를 정의. 예를 들어, 장바구니 안에 동시에 들어가는 상품들의 관계를 규명하는 기술을 들 수 있다.
  - 요약Summarization : 데이터의 일반적인 특성이나 특징의 요점을 간략히 정리하는 기술을 들 수 있다.

05 빅데이터 분석 기술

- 연속성Sequencing : 시간에 따라 순차적으로 나타나는 사건의 종속성을 말함. 예를 들어, A 제품을 구입한 고객이 향후 B 제품을 구입할 확률이라든가 작년의 계절적 매출 변동 요인과 올해의 매출 등을 알아내는 기술을 들 수 있다.

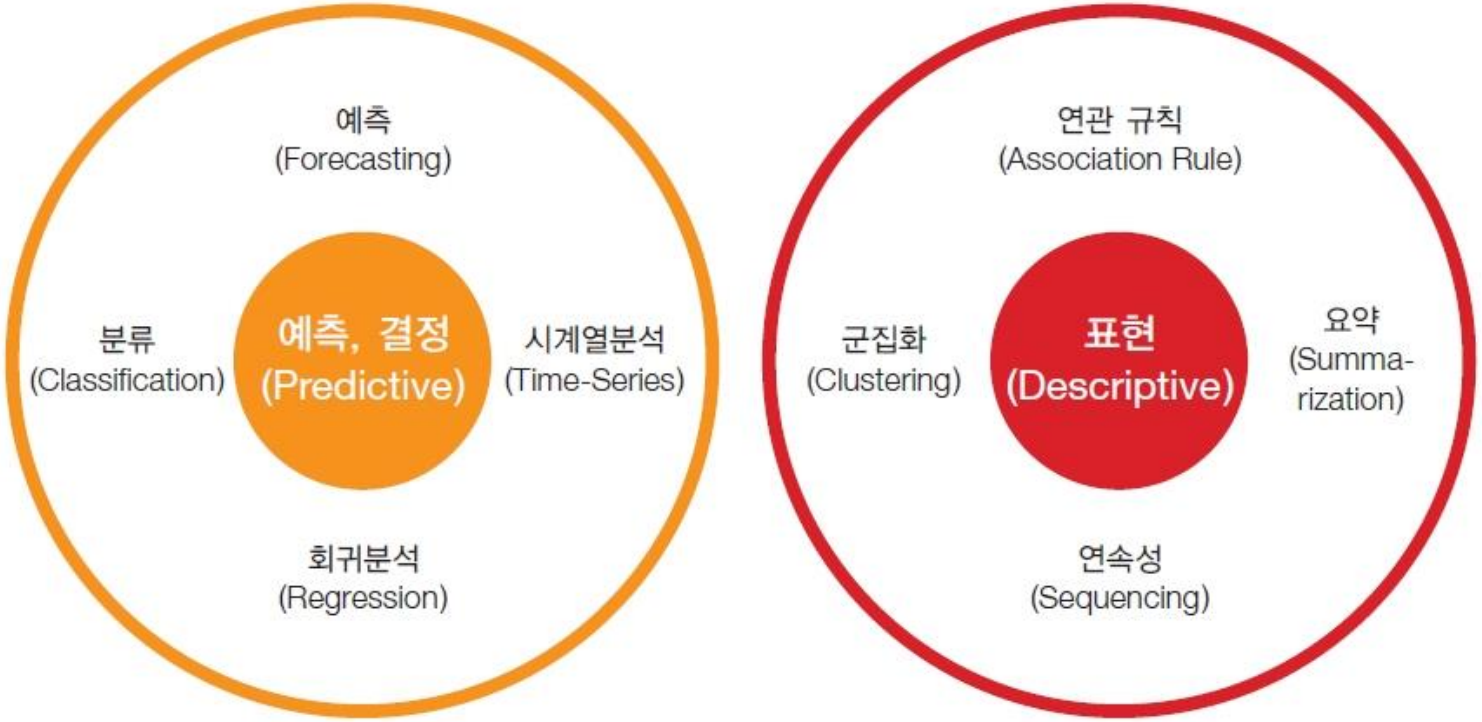

그림 5-2 데이터 마이닝 기술 구분

05 빅데이터 분석 기술

텍스트 마이닝 기술

- 비·반정형 텍스트 데이터로 구성된 빅데이터에서 자연어 처리 기술에 기반하여 의미 있는 정보를 추출하는 기술.
- 데이터 마이닝의 분석 대상은 관계형 데이터베이스, XML 문서와 같은 구조화된 데이터들인 반면, 텍스트 마이닝의 분석 대상은 텍스트 문서, 이메일, HTML 파일 등과 같은 비·반정형의 텍스트 데이터라는 차이점이 있음.

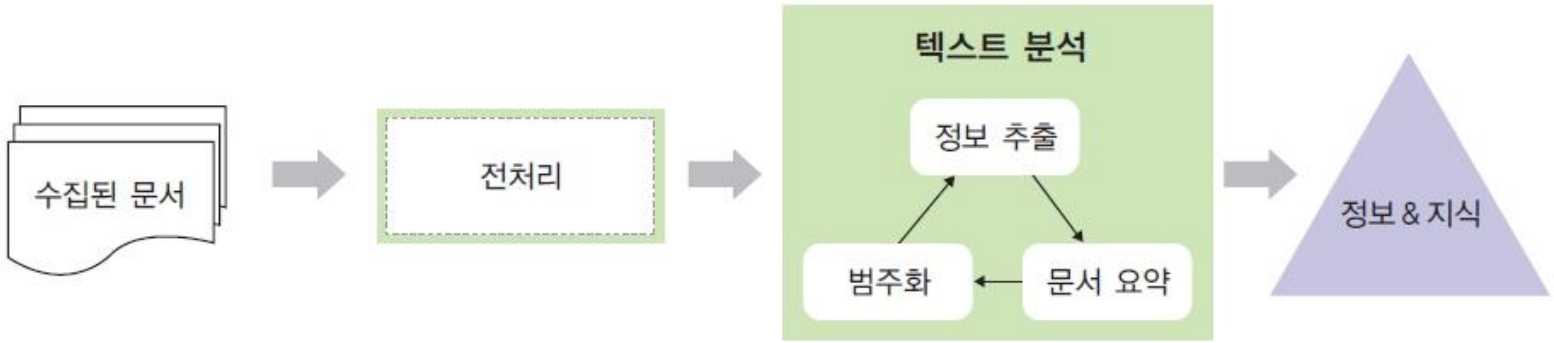

그림 5-3 텍스트 마이닝 과정 [02]

05 빅데이터 분석 기술

- 텍스트 마이닝 절차
  - 정보 수집 : 비·반정형의 텍스트 데이터를 수집하는 단계.
  - 정보 처리 : 대용량의 데이터에서 특정 키워드나 일부 의미 있는 요소를 추출하려고 전처리를 하는 단계.
  - 정보 추출 : 수학적인 모델이나 알고리즘을 이용하여 유용한 정보를 추출해 낸다. 텍스트 마이닝을 위한 정보추출 방법에는 다양한 목적, 조건, 환경 등이 있는데, 이 정보 추출 방법은 텍스트 마이닝에서 가장 중요한 부분 중 하나이다. 특히 정보 추출 방법에는 수많은 수학적 알고리즘과 방법이 있으며, 그 중 간단하면서 가장 강력한 방법인 TF-IDF Term Frequency-Inverse Document Frequency 방식을 많이 사용.
  - 정보 분석 : 최종 키워드나 의미 있는 요소의 우선순위를 도출하는 단계.

05 빅데이터 분석 기술

그림 5-4 텍스트 마이닝의 4단계 과정 [03]

NOTE_ TF-IDF는 정보 검색과 텍스트 마이닝에서 이용하는 키워드의 가중치를 구하는 방법이다. 여러 문서로 된 문서군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한지 나타내는 통계적 수치이다.

05 빅데이터 분석 기술

군집화 기술

- 데이터 마이닝 기술의 한 방법으로, 주어진 빅데이터에서 데이터들의 특성을 고려하여 군집을 정의하고 군집를 대표할 수 있는 대표점을 찾는 것임.

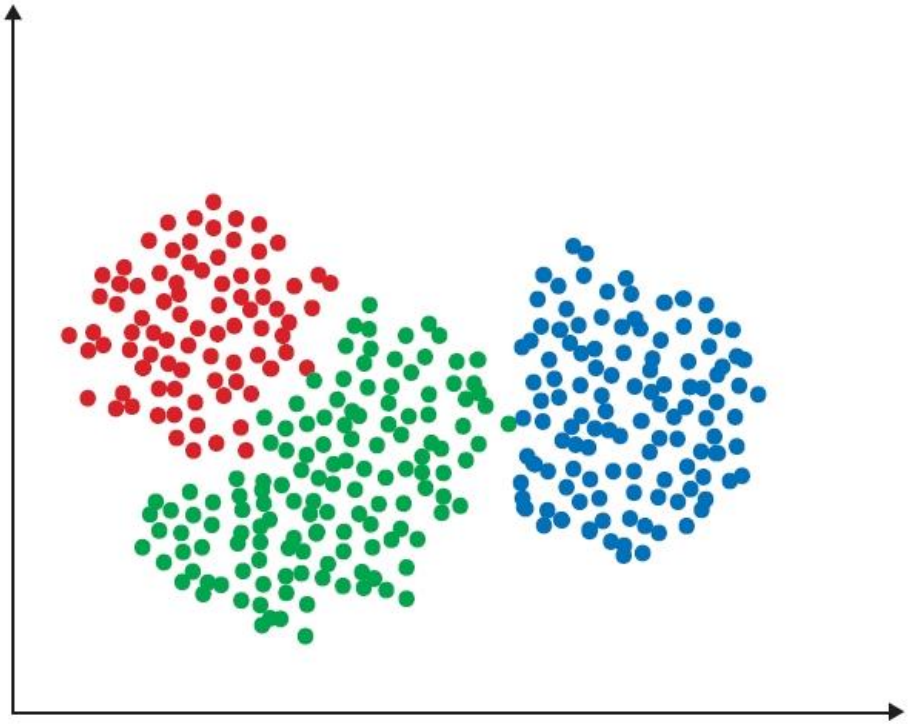

그림 5-13 군집화 기술

05 빅데이터 분석 기술

- 계층적 군집화 기술
  - 각 데이터 점을 하나의 군집으로 설정한 후 이들 간의 거리를 기반으로 하여 분할·합병해 가는 방식.
  - 계층적 군집화 기술에서 데이터 샘플은 분할된 수열로 묶음.

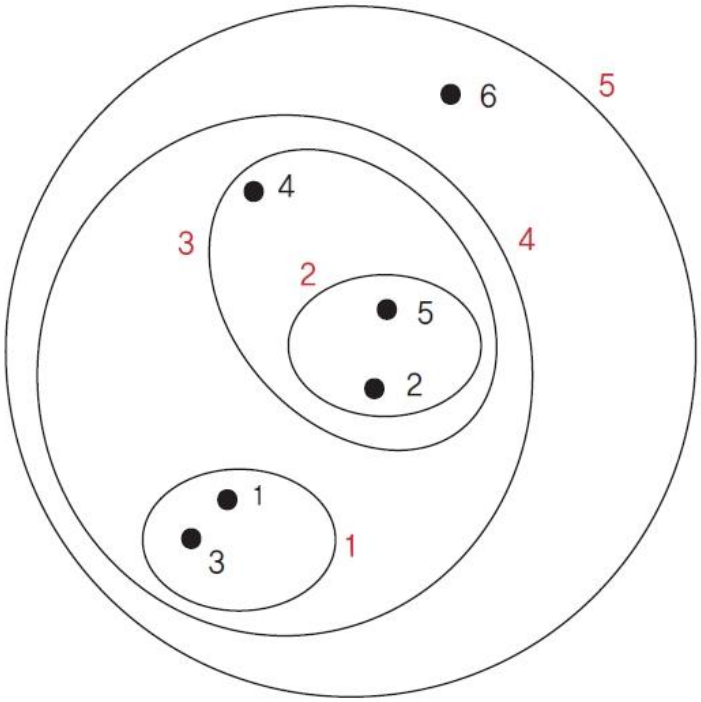

그림 5-14 계층적 군집화 기술 예

05 빅데이터 분석 기술

- 계통도Dendrogram 에서는 각 계층에서 군집들의 유사성을 쉽게 확인할 수 있음
  - 흡수Agglomerative 과정 : 아래에서 위로 처리하여 군집을 흡수, n개의 각 군집과 수열의 형태가 연속적인 흡수 군집화 과정으로 처리됨.
  - 분리Divisive 과정 : 위에서 아래로 분류하는 과정으로, 하나의 군집에 n개의 샘플이 있으며 연속적인 분리 과정으로 수행.

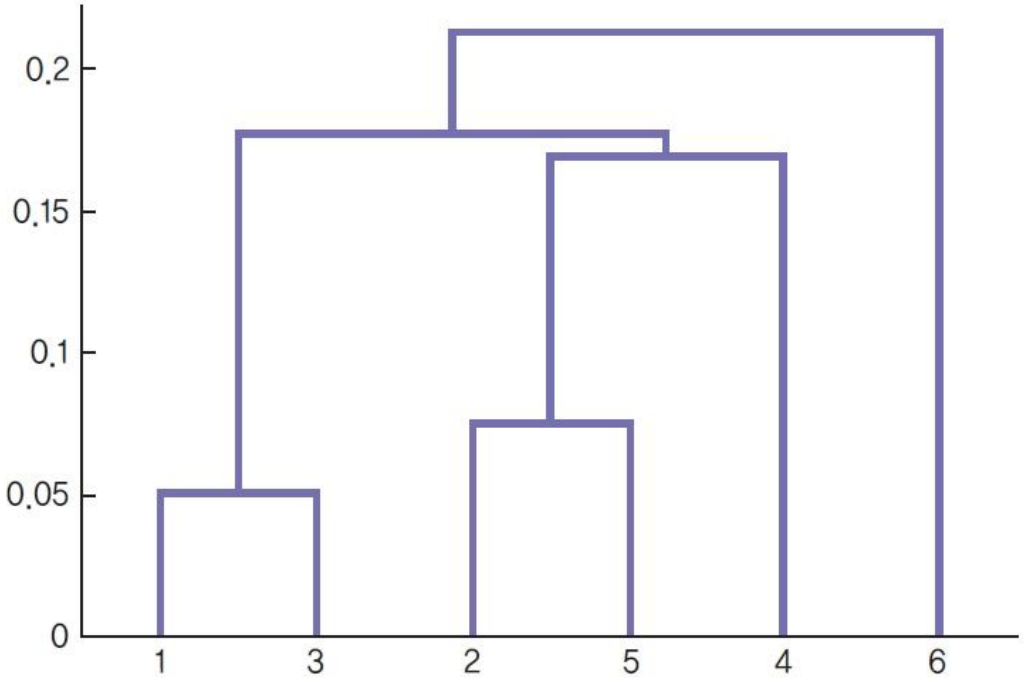

그림 5-15 계층적 군집화 기술을 계통도로 표현한 예 [11]

05 빅데이터 분석 기술

- 분할적 군집화 기술
  - k개의 분할 영역을 결정하는 방법으로 유클리디안 거리Euclidean Distance 계산법에 기반함.
  - K-means 알고리즘
    - 사전에 정한 k개의 군집으로 주어진 데이터를 분류하는 방법.
  - K-means 알고리즘 수행 과정

❶ 군집의 개수인 k를 결정하고 각 군집에 초기값으로 중심 한 개씩을 할당하여 위치를 설정.

❷ 각 데이터를 주어진 중심점을 기준으로 가장 가까운 군집에 할당한다. 중심점과의 거리는 유클리디안 거리 계산 방법에 따름 .

❸할당된 데이터를 중심으로 각 군집은 새로운 중심점을 계산.

❹새로운 중심점이 기존의 중심점과 차이가 있으면 ❷로 되돌아가 반복한다.

새로운 중심점이 기존의 중심점과 차이가 없으면 알고리즘은 끝.

05 빅데이터 분석 기술

- [그림 5-16]은 군집의 중심점을 계속 재형성하는 과정임

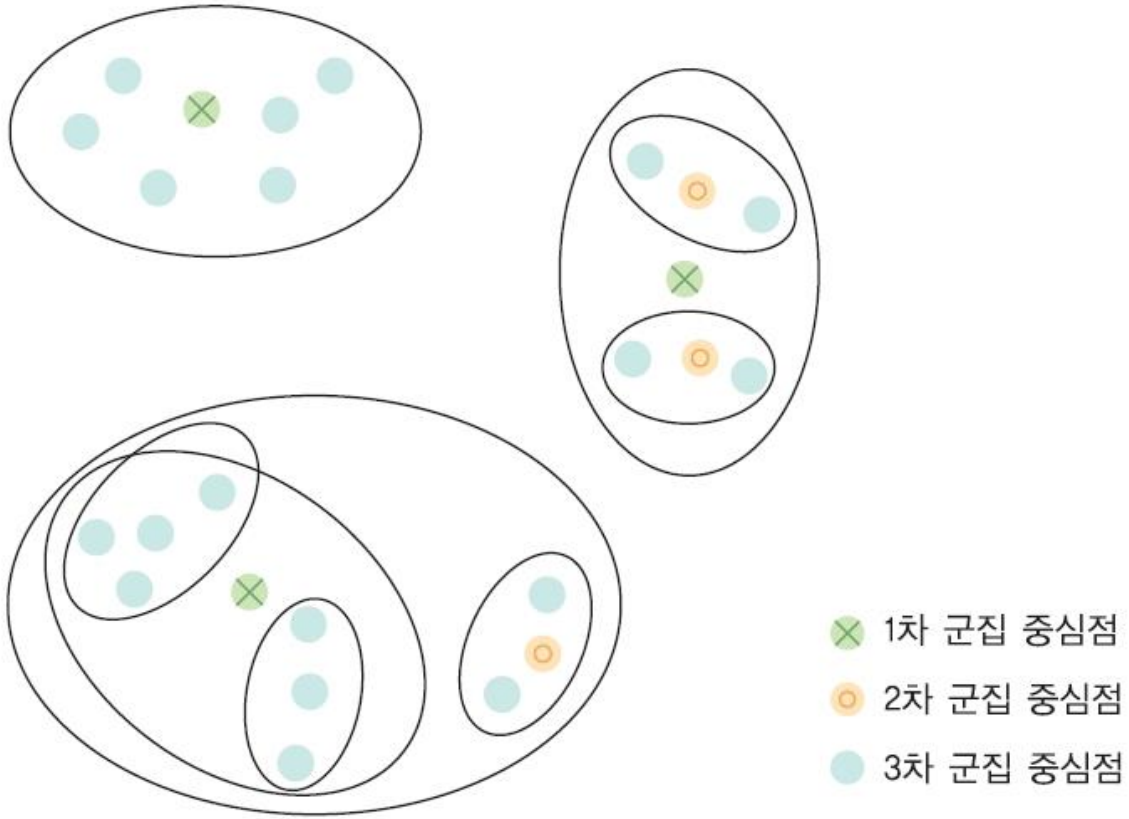

그림 5-16 분할적 군집화 기술의 군집 중심점 생성 과정 [12]

05 빅데이터 분석 기술

❖ 소셜 네트워크 분석 기술

그림 5-17 소셜 네트워크 [14]

- 인접 행렬과 각 항의 관계를 표현한 네트워크

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
A	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
B	0	0	0	0	0	0	1	0	1	1	1	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
D	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
F	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
G	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
I	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0
J	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1
K	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0
L	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
O	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0

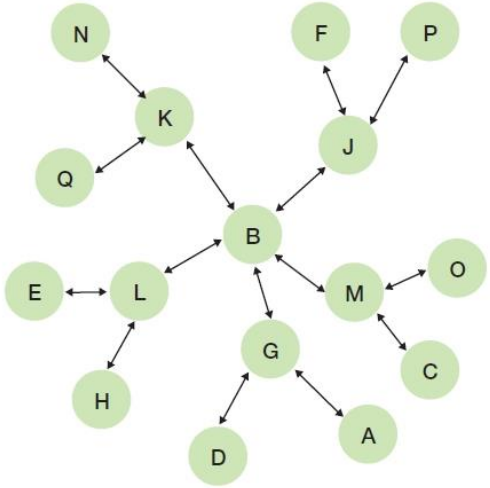

그림 5-18 인접 행렬과 그래프 [14]

05 빅데이터 분석 기술

- 시멘틱Semantic 기술
  - 복잡하고 이질적인 그래프 구조를 트리플Subject-Predicate-Object로 구성된 RDF 그래프로 표현함.
- RDF(S) 의 장점
  - 표준 기반으로 의미의 모호성이 없는 정보 표현과 상호 호환성 확보가 가능.
  - 상호 이질적 정보들을 단일한 표현 체계로 통합 표현·연계가 가능.
  - 그래프 구조뿐 아니라 각 노드의 특성 정보를 통합 표현 가능.
  - 단일 서버에서 대규모 그래프로 표현·저장·관리가 가능(10억 트리플 이상).
  - 그래프 구조에서 강력한 질의와 복잡한 데이터 연산 처리가 가능SPARQL
  - URL에 기반을 두어 분산 그래프 DB 구현과 분산 질의 처리에 유리.
  - 동적으로 변경되는 그래프 정보의 실시간 적용과 질의·연산이 가능.
  - 온톨로지OWL 및 규칙과 결합하여 연역적 논리 추론이 가능.
  - 기계 학습과 연동하여 하이브리드 유형의 분석 체계를 구현. 특히 RDF(S)를 사용하므로 [그림5-19]와 같이 사람의 연결 구조와 정보 연결 구조를 통합하여 표현하고, 필요에 따라 부분적으로 분리하여 단순화하기도 쉬운데, 이는 소셜 네트워크 분석에서 매우 강력한 장점임.

05 빅데이터 분석 기술

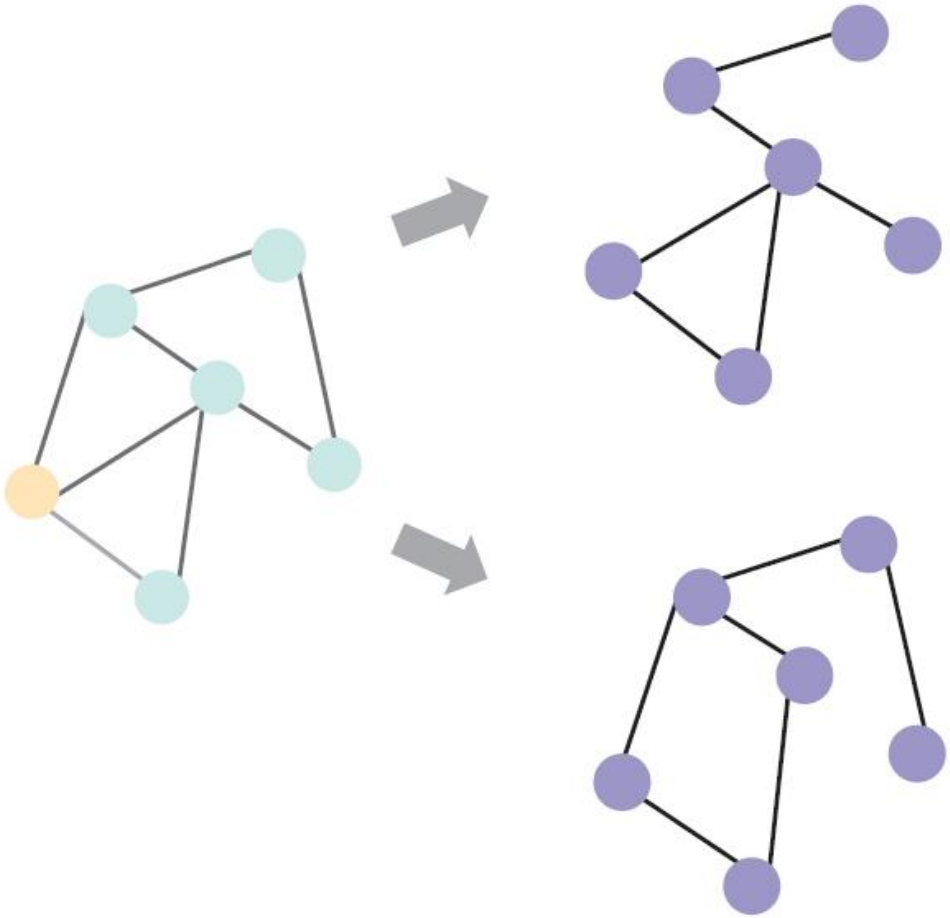

그림 5-19 RDF(S)를 사용한 사람과 정보 네트워크의 통합 표현 및 단순화

05 빅데이터 분석 기술

- 시멘틱 네트워크를 기반으로 한 소셜 네트워크 분석의 특징
  - 거대한 글로벌 네트워크에서 의미적 연관성이 있는 서브 네트워크를 분리하고, 수학적 분석 모델을 적용·해석할 수 있다는 것. RDF에 기반을 둔 시멘틱 네트워크에서 의미 조건을 이용하여 서브 네트워크를 추출하려고 SPARQL을 사용함.

표 5-2 SPARQL을 이용한 시멘틱 소셜 네트워크의 구조 분석 예 [15]

degree centrality	closeness centrality
select ?y count(?x) as ?degree where { {?x $path ?y filter(match($path, star(param[type]))) filter(pathLength($path) <= param[length]) } UNION {?y $path ?x filter(match($path, star(param[type]))) filter(pathLength($path) <= param[length]) } } group by ?y	select ?y ?to pathLength($path) as ?length sum(?length) as ?centrality where { ?y $path ?to filter(match($path, star(param[type]), 's')) } group by ?y

05 빅데이터 분석 기술

- BKNetwork.org의 지식 네트워크 전문가 검색 서비스
  - 소셜 네트워크 분석의 대표적인 응용 시스템.
  - 부산 출신이나 부산 지역에서 활동하는 학계, 산업계, 관계, 문화예술계를 망라한 분야별 전문가와 출향인고향을 떠난 사람 을 검색하는 서비스.

  - 다음 장의 [그림 5-20]은 '삼성'으로 검색한 소셜 네트워크 분석 결과.

05 빅데이터 분석 기술

그림 5-20 BKNetwork 소셜 네트워크 검색 화면 [16]

그림 5-21 BKNetwork 소셜 네트워크 분석 및 사회 관계망 검색 과정 [16]

05 빅데이터 분석 기술

❖ 그래프 마이닝 기술

- 페이스북이나 트위터와 같은 소셜 미디어의 데이터를 표현하는 방법.
- 그래프에서 마이닝 기술을 적용하는 기술, 그래프 마이닝 기술은 일정 빈도수 이상의 특정 패턴을 모두 찾아내는 방법.

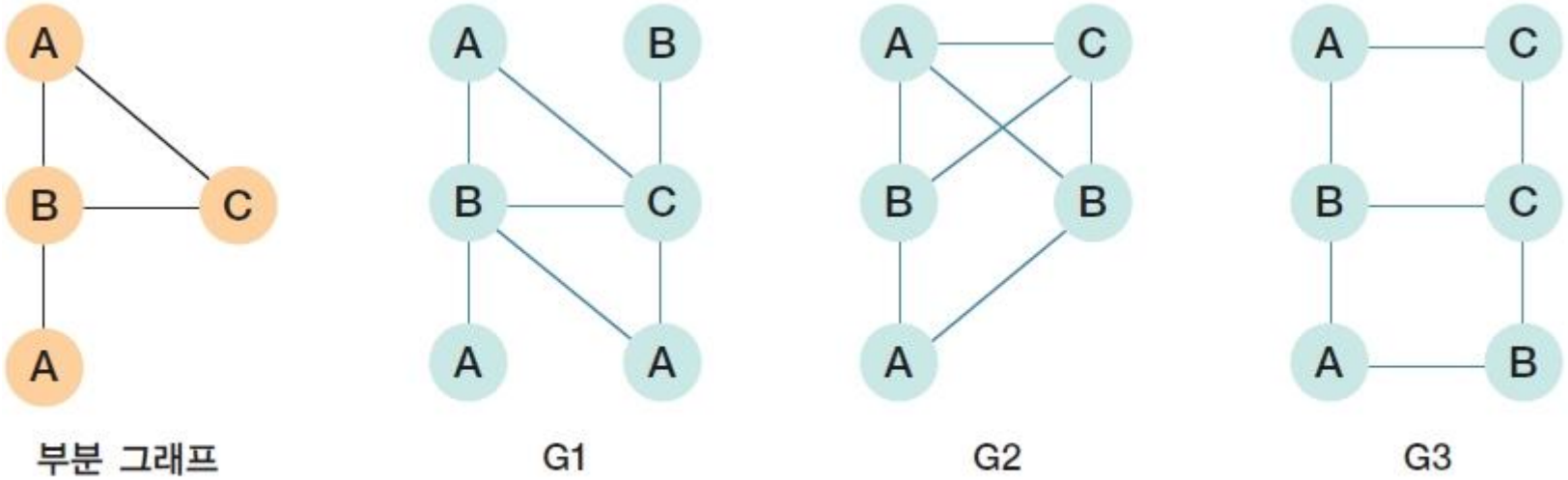

그림 5-22 그래프의 예 [22]

05 빅데이터 분석 기술

- 빈발 부분 그래프 마이닝 기법
  - 전체 그래프에서 자주 발생하는 부분 그래프를 발견함.
  - 그래프 데이터베이스에서 모든 빈발 부분 그래프를 찾는 기법.
  - 성능이 효율적이고 알고리즘의 확장성이 뛰어난 gSpan을 널리 사용함.
  - 도약 탐색Leap Search 방법
    - 분류 성능이 높은 특징을 선택하는 방법으로 측정 기준을 최소 지지도로 정하고, 최소 지지도를 다양하게 변화시키면서 gSpan 과정을 반복하는 방법.
  - 모델 기반탐색 트리Model Based Search Tree
    - 분류 성능이 높은 특징을 선택하는 방법으로, gSpan 과정으로 찾은 빈발 부분 그래프 중에 가장 분류 성능이 높은 빈발 부분 그래프의 포함 여부를 측정한다.
    - 이분할 과정은 그래프 데이터베이스가 완전히 분할될 때까지 반복 수행됨.

05 빅데이터 분석 기술

빅데이터를 처리하는 기존 알고리즘 변형 기술

- 도형 세기 알고리즘의 변형

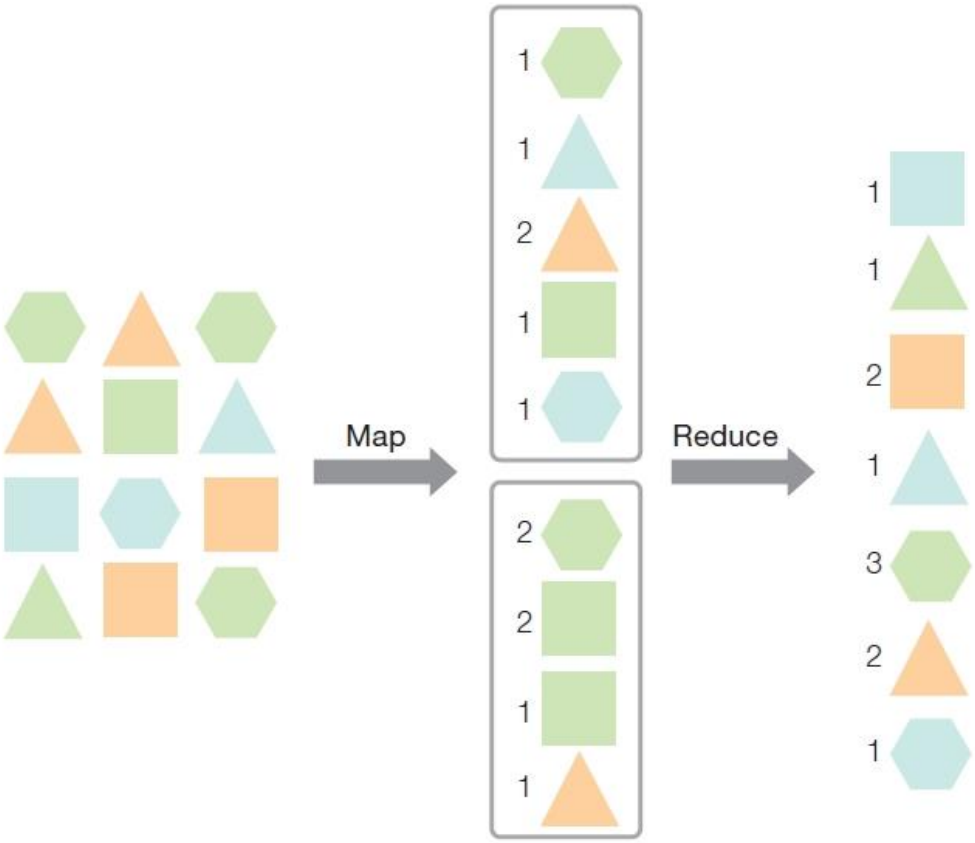

그림 5-24 맵리듀스를 이용한 도형 세기 과정

05 빅데이터 분석 기술

- 컬럼 빈도수 측정 프로그래밍 방안
  - 맵리듀스
    - 맵 함수Mapper 와 리듀스Reducer 함수의 입력과 출력은 모두 <키, 값> 순서쌍으로 정의함.
  - [그림 5-25]
    - 다음의 employees.txt 데이터 파일에서 FIRST별로 빈도수가 얼마인지 측정하는 맵리듀스를 프로그래밍 하고자 함.

employees.txt

# LAST	FIRST	SALARY
Smith	John	$90.000
Brown	David	$70.000
Jahnson	George	$95.000
Yates	John	$80.000
Miller	Bill	$65.000
Moore	Jack	$85.000
Taylor	Fred	$75.000
Smith	David	$80.000
Harris	John	$90.000
...	...	...

05 빅데이터 분석 기술

```
    mapper
def getName (line):
    return line.split('\t')[1]
    reducer
def addCounts (hist, name):
    hist[name] = \
    hist.get(name, default = 0) + 1
    return hist

input = open('employees.txt', 'r')

intermediate = map(getName, input)

result = reduce(addCounts,
                \intermediate, {})
```

```
    mapper
def getName (line):
    return(line.split('\t')[1], 1)
    reducer
def addCounts (hist, (name, c)):
    hist[name] = \
    hist.get(name, default = 0) + c
    return hist

input = open('employees.txt', 'r')

intermediate = map(getName, input)

result = reduce(addCounts,
                \intermediate, {})
```

Key-value iterators

그림 5-25 맵리듀스 예 : First Name을 세는 프로그램

05 빅데이터 분석 기술

- 맵리듀스 에서는 이렇게 하나의 노드 안에서 반복적인 결과를 결합하여 처리하려고 addCounts 함수를 추가로 정의함.

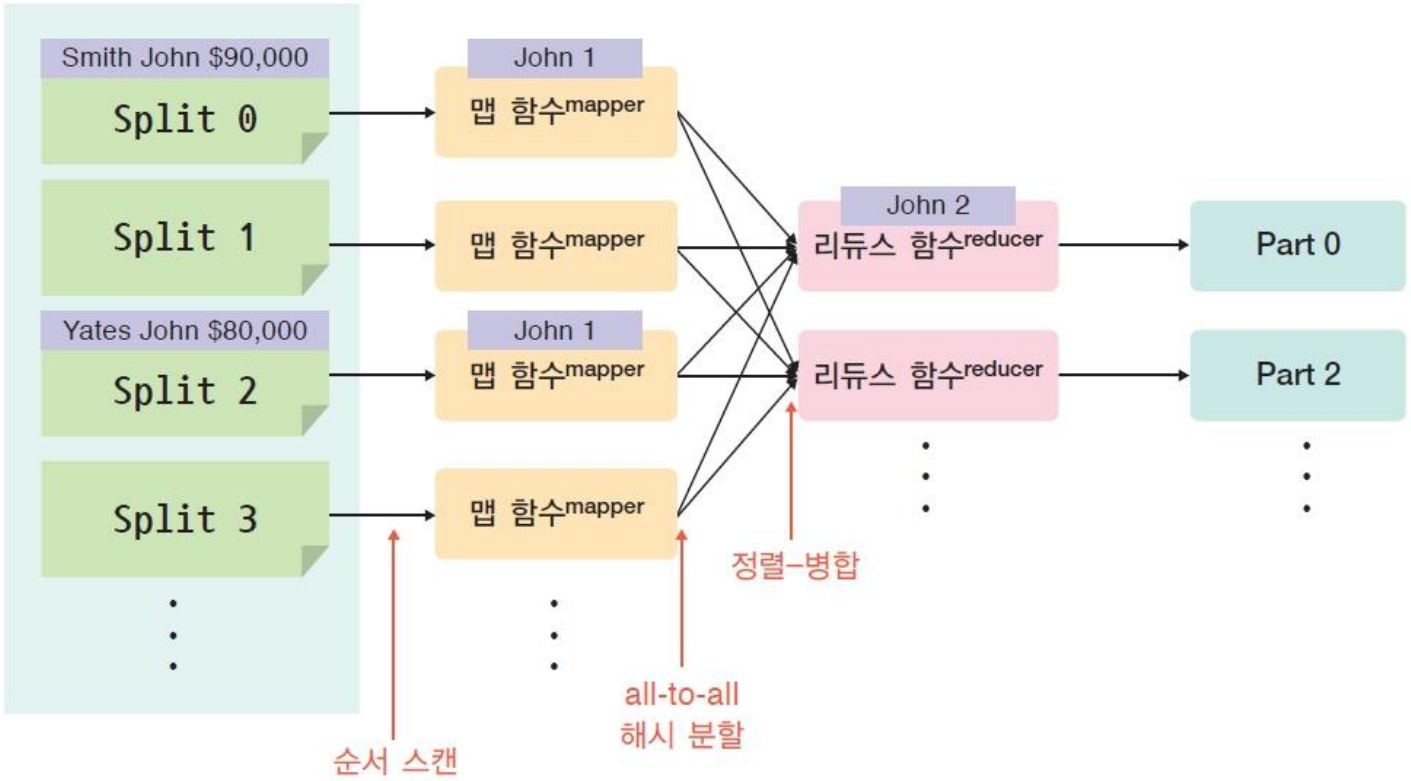

그림 5-26 이름을 세는 프로그램의 흐름

05 빅데이터 분석 기술

- 소셜 네트워크 통계 방안
  - [그림 5-27] 노드가 3개이고 에지에 방향성이 있는 그래프에 맵리듀스를 적용함.

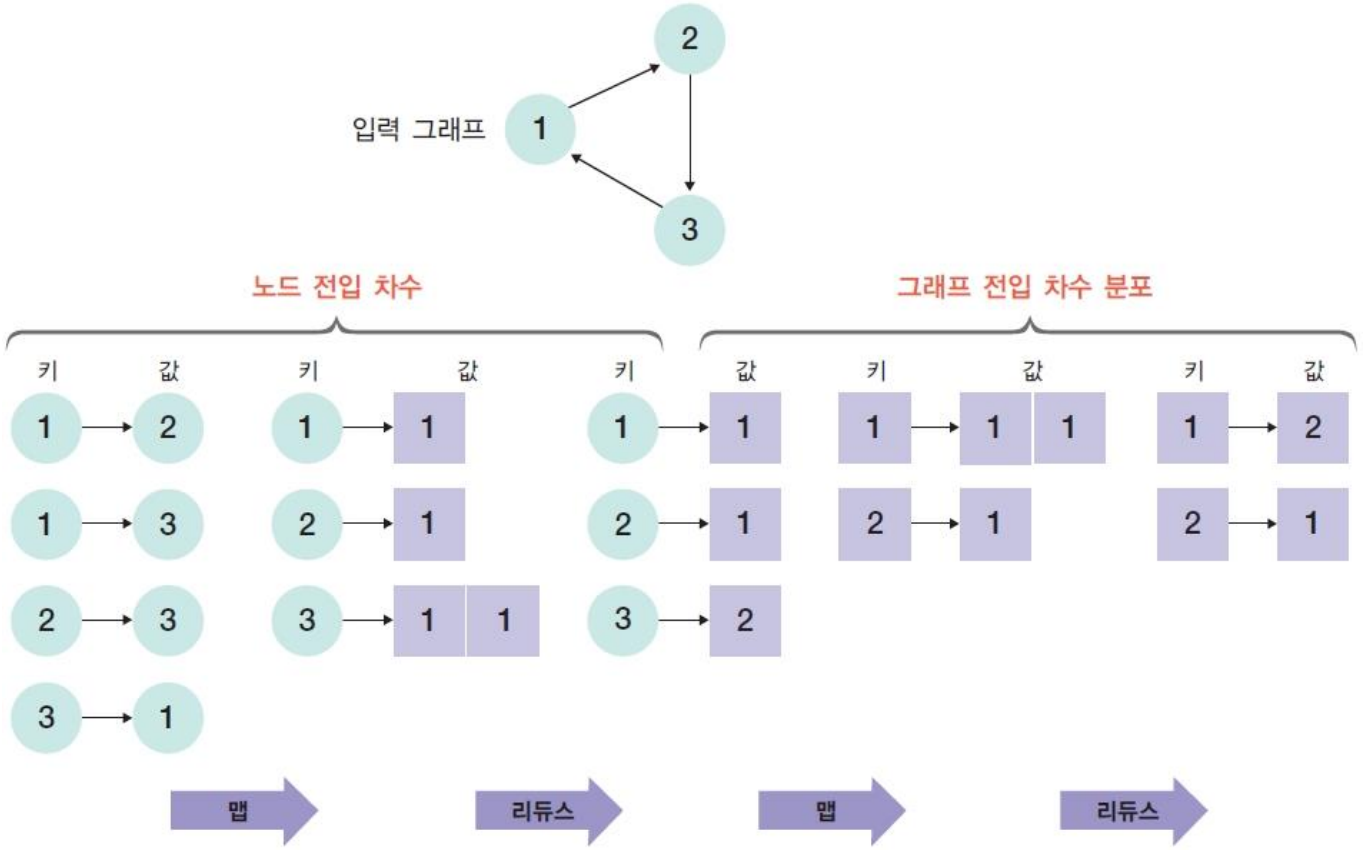

그림 5-27 맵리듀스에서 그래프 통계 계산 방법 [27]

05 빅데이터 분석 기술

❶ 원 데이터를 에지로만 표현.

→ <1, 2>, <1, 3>, <2, 3>, <3, 1>

❷ 각 맵 함수에서는 나누어진 에지의 <키, 값> 순서쌍에서 얻은 값인 목적 노드 ID로 그룹핑함.

→ 컴바이너가 내장된 리듀스 함수로 결합된 결과를 보면, 1은 한 번, 2는 한 번, 3은 두 번 이라는 결과가 나옴.

❸ 출력 정보를 다시 <키,값> 순서쌍으로 표현.

→ <1, 1>, <2, 1>, <3,2>

→ 맵 함수에서는 노드 전입 차수 과정을 그대로 반복함.

❹ <키, 값> 순서쌍에서 얻은 값인 '해당 노드 전입 차수'로 그룹핑 함.

→'해당 노드 전입 차수' 1은 두 번, 2는 한 번 나옴.

05 빅데이터 분석 기술

- 네트워크 분석 기법
  - 군집화 상수Clustering Coefficient
    - 네트워크에서 노드들(컴퓨터들)이 뭉치려는 정도가 얼마나 강한지 측정하는 방법으로, 사람이나 컴퓨터의 관계 응집도를 평가하는 척도로 활용됨.
    - 군집화 상수 함수의 정의

$$\text{클러스터링 상수} = \frac{\text{노드 } v\text{와 이웃 노드 간의 에지의 개수}}{\text{노드 } v\text{와 이웃 노드 간의 가능한 모든 에지의 수}}$$

$$= \frac{\text{노드 } v\text{와 인접한 에지의 수}}{\text{노드 } v\text{의 가능한 모든 쌍의 수} \left(\frac{d_v}{2}\right)}$$

그림 5-28 군집화 상수 함수의 정의

05 빅데이터 분석 기술

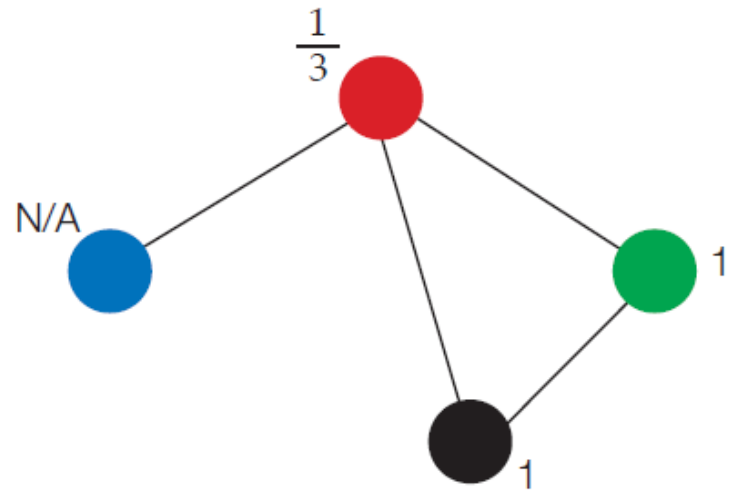

그림 5-29 군집화 상수 함수의 표현

[그림 5-29] cc(v) 식에서 분모는 노드 v의 모든 차수, 즉 가능한 모든 이웃의 크기(d_v)를 2로 Combination하므로 노드 v의 가능한 모든 노드 쌍의 수를 분모로 한다. cc(v) 식에서 분자인 감마(Γ)는 노드 v의 이웃을 말한다. 즉 v에 연결된 노드들의 집합으로 분자의 결과인 <u, w>의 집합은 노드 v에 연결된 모든 이웃의 쌍인 실제로 연결된 에지 개수이다.

05 빅데이터 분석 기술

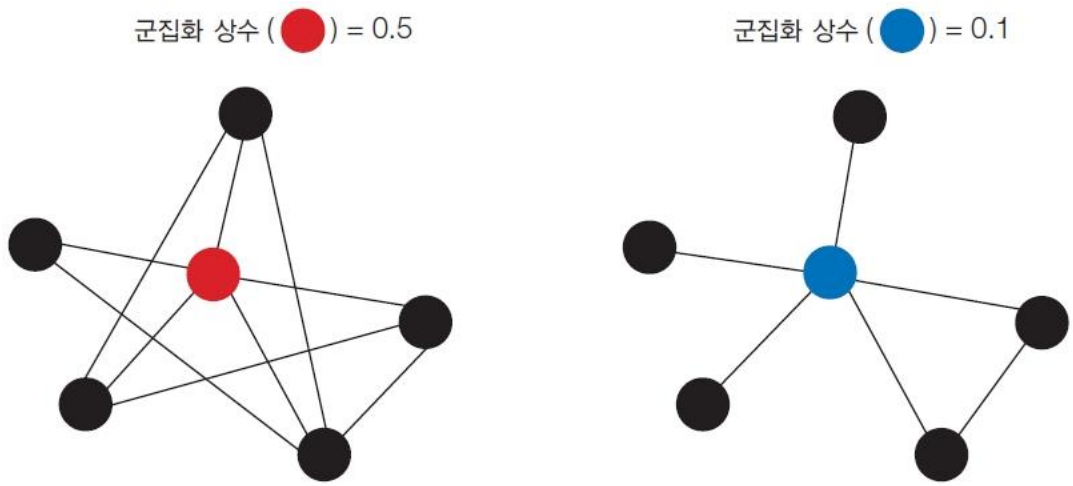

그림 5-30 지역 군집화 상수 예

빨간색 노드의 이웃은 다섯 개이고, cc (v) 식의 분모를 구하려고 5 Combination 2를 계산하면 10임. 이제 cc (v) 식의 분자를 위해 이들 다섯 개의 검은색 노드 이웃끼리 서로 연결된 에지를 세면 다섯개임. 그러므로 빨간색 노드의 군집화 상수는 5/10가 되므로 결과는 0.5이다.

파란색 노드의 이웃은 다섯 개이고, cc (v) 식의 분모를 구하려고 5 Combination 2를 계산하면 10임. 이제cc (v) 식의 분자를 위해 이들 다섯 개의 검은색 노드 이웃끼리 서로 연결된 에지를 세면 한 개임. 그러므로 오른쪽 중심 노드의 군집화 상수는1 /10이 되므로 결과는 0.1이다.

05 빅데이터 분석 기술

- 군집화 상수가 높을수록 그래프 내의 노드 간에 관계성이 많음을 알 수 있다. 이 근거로 군집화 상수는 사회학에서도 많이 활용되며, 군집화 상수가 높을수록 사회적 신뢰도가 높다는 결론을 도출한다.

- [그림 5-31]의 알고리즘은 [그림 5-32]에서 각각의 노드 v 에 인접한 삼각형의 개수를 세어 군집화 상수를 구한다. 이는 순차적인 방법으로 그래프가 매우 커지면 연산의 속도가 떨어져 결과를 보기 어렵다. 그러므로 빅 그래프에서는 맵리듀스로 바꾸어 처리해야 한다.

```
for v ∈ V do
    for u1 w∈ T(v) do
        if(u1 w) ∈ E then
            Triangles [v]++
```

그림 5-31 지역 군집화 상수의 순차적 알고리즘

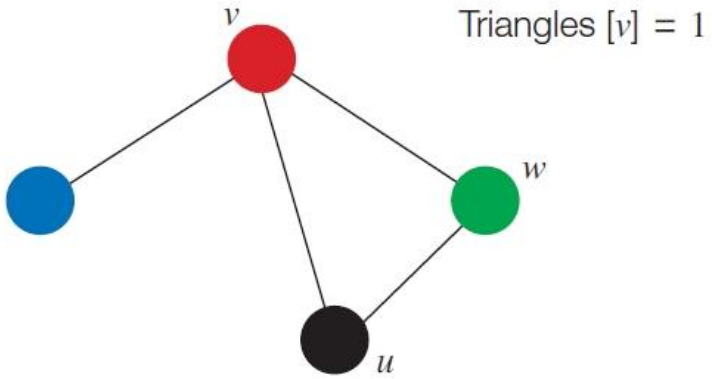

그림 5-32 지역 군집화 상수의 순차적 알고리즘 표현

05 빅데이터 분석 기술

❖ 최신 빅데이터 분석 연구

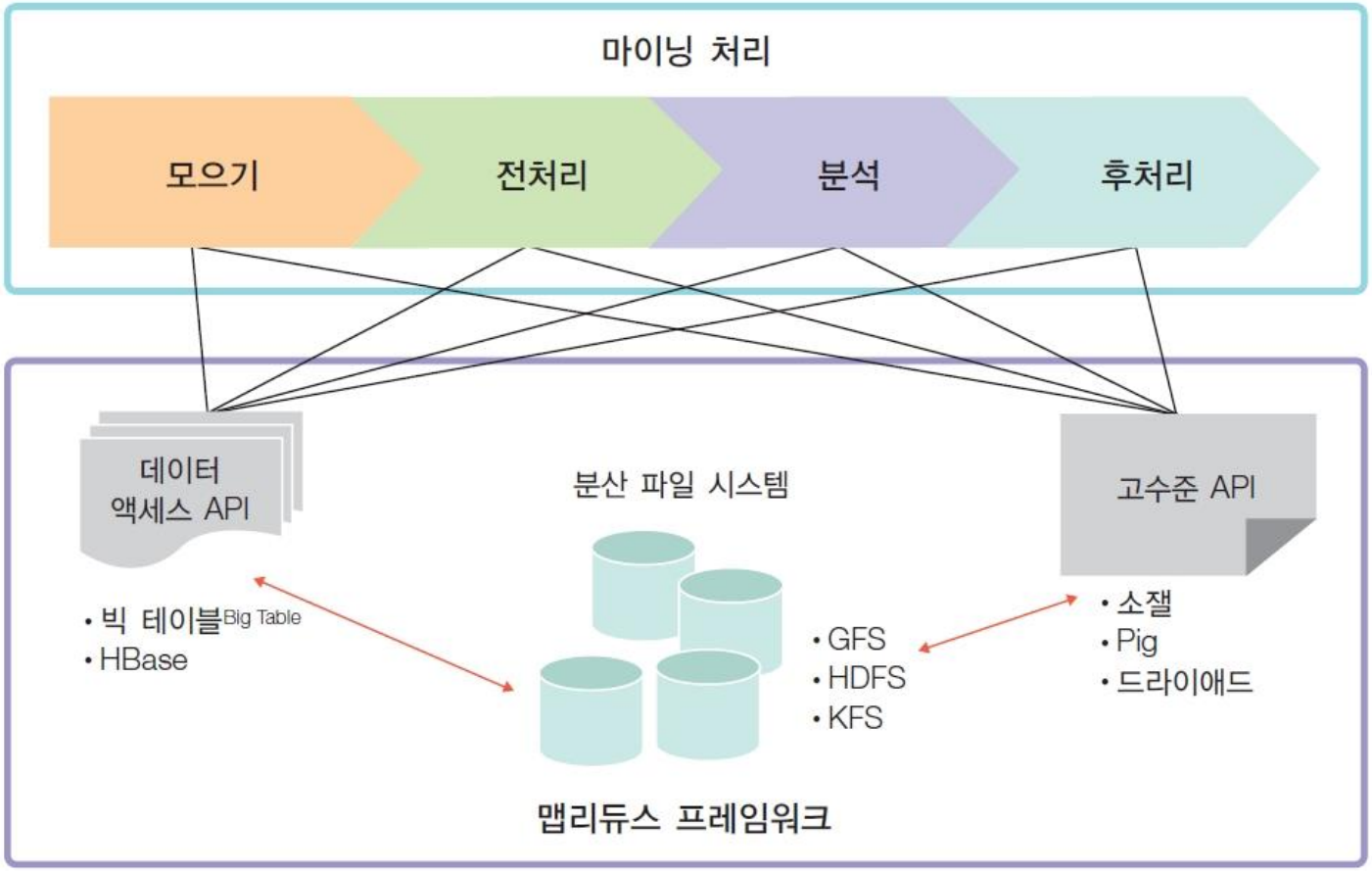

그림 5-36 분산 처리를 이용한 마이닝 처리 단계 [29]

05 빅데이터 분석 기술

- 데이터 마이닝 연관 규칙을 활용한 빅데이터 알고리즘
  - Apriori 알고리즘 [17]은 데이터 마이닝 분야의 대표적인 연관 규칙 알고리즘임.
  - 맵리듀스 알고리즘은 각 트랜잭션의 모든 부분 집합을 만든 후 각 부분 집합의 개수를 세는 방법을 사용하여 이 문제를 해결함.

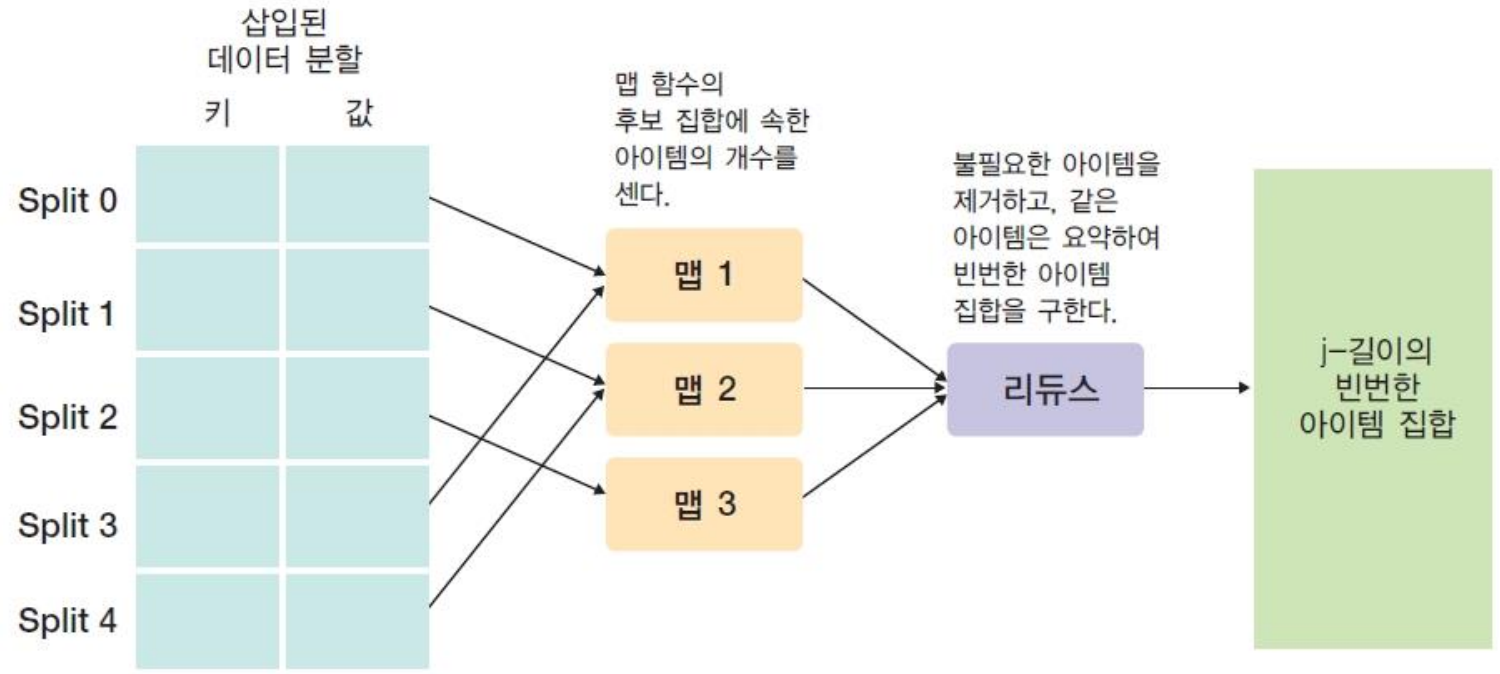

그림 5-37 맵리듀스 기법을 사용한 Apriori 알고리즘 [30]

05 빅데이터 분석 기술

- 빅데이터 주요 분석 기술

표 5-3 빅데이터 주요 분석 기술

제품/기술	최초 개발	최초 공개	주요 기능 및 특징
NLTK	E. Loper	2001년	글의 문장 분할, 문장의 단어 분할 기능 제공
OpenNLP	아파치	2010년	비정형 텍스트에서 의미 있는 용어 추출
Boilerpipe	C. Kohlschutter	2010년	웹 페이지에서 불필요 데이터 제거 및 필요한 정보 추출
WEKA	I. Witten	1993년	데이터 마이닝 및 기계 학습 알고리즘을 포함한 데이터 분석 프로그램
Mahout	아파치	2009년	확장 가능한 기계 학습 알고리즘 개발 프로젝트
scikits_learn	D. Cournapeau	2011년	오픈 소스 기계 학습 라이브러리

05 빅데이터 분석 기술

- NLTK Natural Language ToolKit
  - 특정한 문제를 해결하려고 알고리즘 생성 과정에 필요한 빌딩 블록을 제공한다.
  - PHP, Python 등으로 제공되는 툴킷 들이 있으며, 연구자들이 많이 사용하는 프레임워크로 사용자들이 다양한 방식으로 활용할 수 있다. 또한 일반적인 접미사를 제거하여 중심 단어들을 추출하거나 전체 텍스트에서 동의어를 찾아내려고 기계가 분석 가능한 사전 형태로 데이터를 정형화시키는 기능도 제공한다.
- OpenNLP 오픈NLP
  - 아파치의 OpenNLP는 비정형화된 텍스트에서 사람이나 기관의 이름, 특정 장소, 시간 등을 추출하는 작업을 간편하게 실행할 수 있는 모델을 포함하는 라이브러리.
  - OpenNLP는 Java 기반의 자연어 처리 솔루션.
  - 프로그래밍 언어 기반의 응용 프로그램에 적합한 형태로,자연어 처리 코드를 스스로 생성한다. 비정형화되거나 가공되지 않은 텍스트를 문장과 단어로 분리한 후 그 결과를 다양한 방법으로 클래스화하는 표준 컴포넌트를 많이 포함한다.

05 빅데이터 분석 기술

- Boilerpipe보일러파이프
  - 웹 페이지 내에서 불필요한 부분을 제거하여 실질적으로 필요한 정보만 추출하는 작업을 수행하는 프레임워크.
  - HTML 문서에서 실제로 가장 중요한 콘텐츠를 찾아내는 알고리즘을 적용한 Java 기반 라이브러리로 제공된다. 모든 종류의 웹 콘텐츠에서 거의 완벽한 전처리 도구를 생성할 수 있다.
  - Boilerpipe는 뉴스나 기고 같은 특정 정보가 담긴 웹 페이지의 분석에 많이 사용하다 최근에는 블로그나 SNS 서비스 등다른 형태의 웹 페이지에서 정보를 추출할 때도 유용하게 사용한다.
- WEKA웨카
  - 데이터 분석 프로그램으로, 무료로 배포되고 Java로 작성된 오픈 소스 프로그램이다.
  - 데이터마이닝 알고리즘으로 사용자 개인화가 가능하다는 장점이 있다.
  - WEKA는 플러그인 형식을 취해 개발자가 자신이 개발한 고유 알고리즘을 쉽게 접목시킬 수 있고, 사용이 간편한 명령으로 구성되어 있으며, 이해하기 쉽고 간단한 인터페이스를 제공한다.

05 빅데이터 분석 기술

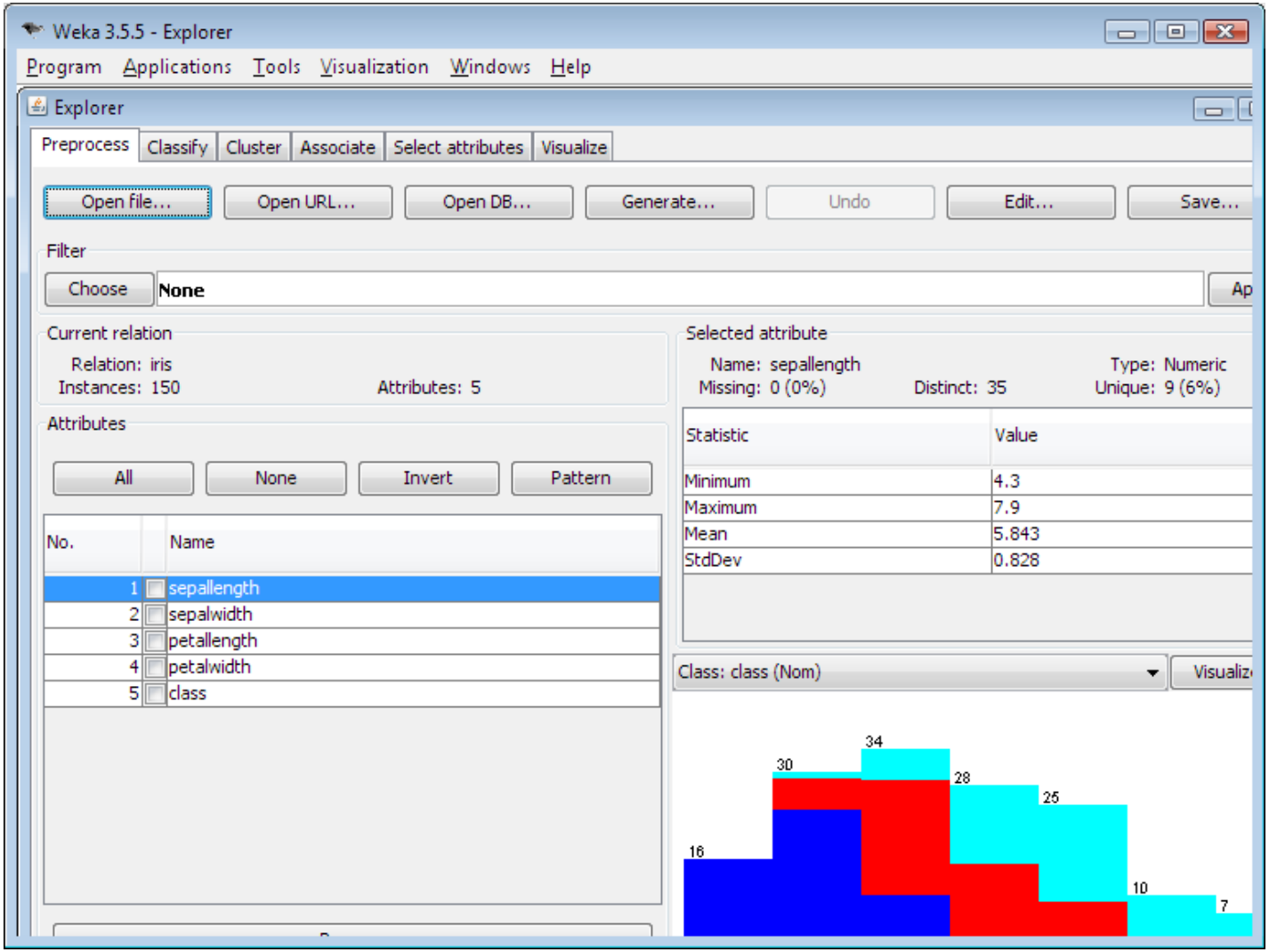

그림 5-43 WEKA를 이용한 분석 결과 예 [33]

05 빅데이터 분석 기술

- Mahout머하웃
  - 확장 가능한 기계 학습 알고리즘을 만드는 것이 주목적으로, 아파치 라이선스가 있으면 무료로 사용할 수 있다.
  - 방대한 양의 데이터 집합에서 기계 학습 알고리즘을 실행할 수 있는 오픈 소스 프레임워크로, 확장성과 처리량을 보장하려고 하둡 기반의 병렬 형식으로 구성한다.
  - 공통 작업이 많은알고리즘에 적합하며, 군집화, 분류, 사용자 행동을 기반으로 품목을 추천해 주는 시스템 등 분석환경에도 적합하다.

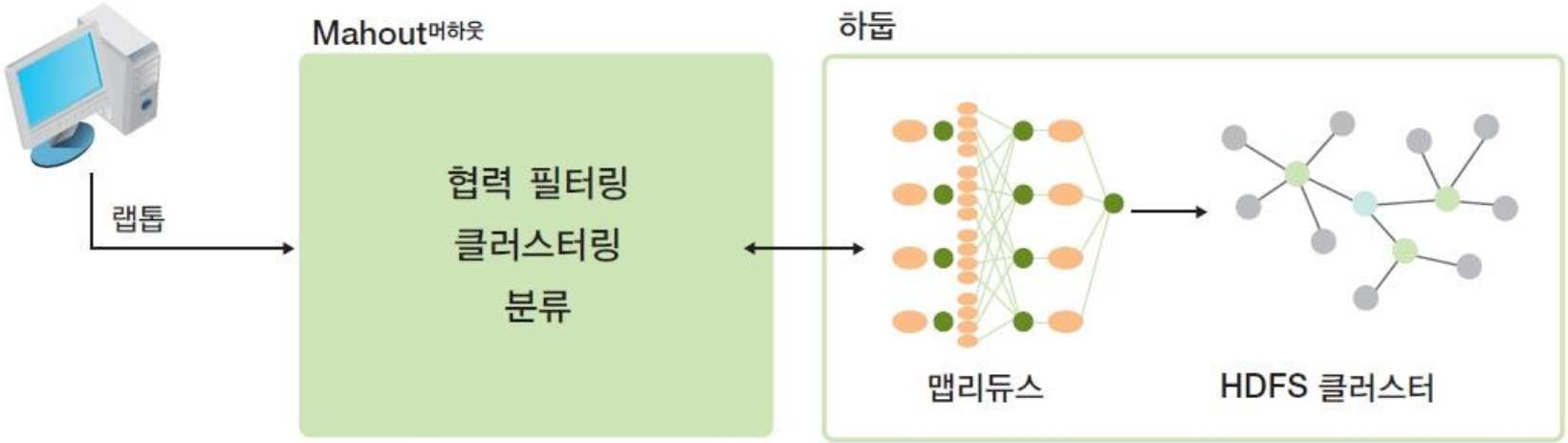

그림 5-44 Mahout과 하둡 연동 구성도 [34]

05 빅데이터 분석 기술

- scikits_learn사이키트런
  - 서포트 벡터 머신,로지스틱 회귀분석, 군집화 등 여러 분석 기법을 제공. 또한 이런 기법을 편리하게 활용할 수 있도록 고수준의 인터페이스를 제공함.
  - scikits_learn은 복잡한 문제를 해결하는 알고리즘이나 문제를 해결하는 과정에 초점을 맞추었기에 과정에 치중하는 사용 환경에 더 적합함.

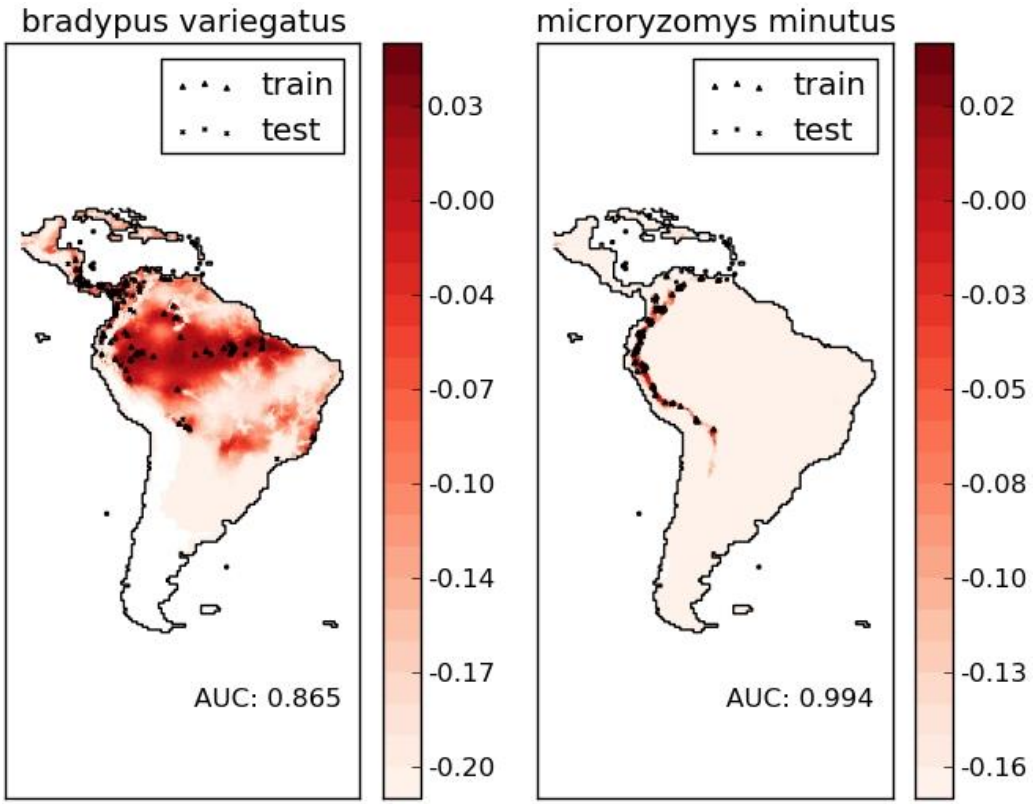

그림 5-45 scikits_learn을 이용한 분석 결과 예 [35]