**통계데이터 연관 분석을 통한 향후 동향 분석**

2010104086 진호조

**개 요**

정보를 저장하는 방식으로 서적이 아닌 전부를 데이터로 저장하는 시대가 되면서 다양하게 저장되어 있는 데이터들을 단순히 저장만 해 두는 것이 아닌 그것을 활용할 방법이 필요하게 되었다. 그런 다양한 데이터들을 분류하고 분석하는 것을 빅데이터 처리라고 한다.

이러한 빅데이터는 다양한 분야에서 이용된다. 주식, 은행과 같은 금융정보나 SNS와 같은 개인 커뮤니티 그리고 일반적인 시장움직임의 동향과 같은 경제적인 지표에도 활용되고 있다. 물론 이러한 빅데이터에서도 사람이 자신이 필요한 데이터와 필요하지 않은 데이터들도 존재한다. 그렇기에 이것을 분류, 추출, 정리를 해야 하는데 그것이 데이터마이닝이다. 이러한 데이터들을 데이터마이닝을 처리 그리고 원하는 데이터를 통해서 추가적인 작업을 하는 것이 빅데이터를 활용하는 핵심적인 부분이다.

통계 데이터 샘플을 구해 그것을 통해서 데이터를 정리 및 분류하여 정리된 데이터들을 비교하여 데이터의 상관관계 및 앞으로의 동향을 예측하는 2차 데이터를 얻는 프로그램을 개발해 보고자 한다.

1. **서론  
   1.1. 연구배경**

지금 우리 생활에 있어서 모든 것들이 정보로 이루어져 있다고 볼 수 있다. 우리가 간단하게 입력하는 휴대폰의 문자나 각종 통계수치 자료들 그리고 우리들의 신체 정보 그 자체도 전부다 디지털화 된 정보로서 저장되어 있는 정보화 시대에서 이런 막대한 데이터를 정리할 필요가 있다.

실제로 우리 생활에서뿐만이 아니라 각종 기업이나 공공 기관에서는 더욱 데이터의 저장 및 정리를 필요로 하고 있으며 독자적으로 이런 데이터를 분류하고 정리하는 시스템들을 만들어 두었다.

기본적으로 하나의 관련된 데이터에 대해서 분류를 하는 경우도 많지만 정보는 다른 정보를 통해서 추가적인 새로운 정보를 발견하기도 하고 축적된 정보를 통해서 앞으로 어떤 식의 정보가 생성되거나 특정 상황에 대해서 어떤 식으로 처리되어야 하는 예측 및 대처도 가능하게 할 수 있다.

* 1. **연구목표**

최근에는 통계청과 같은 통계 정보나 다양한 데이터들이 있는 빅데이터 사이트들이 존재하며 이런 빅데이터 사이트에서 해당되는 자료의 일부를 입수 그리고 그 것을 제목이나 태그로 분류하여서 그것과 관련된 자료와 연관을 지을 기본적인 분류를 시작하며 내부의 데이터들을 각각

해당 기준에 따라서 분류 심화적으로는 도표화를 통해서 데이터의 동향을 파악하여 앞으로의 데이터의 변화 동향을 파악하도록 구현하고 추가적인 부분으로 다른 쪽에서 분류된 데이터와 비교하여 서로의 상관관계가 존재한다면 그것까지 도출해낼 수 있도록 처리를 하는 것이 궁극적인 목표이다.

우선 그러기 위해서는 가장 필요한 것은 역시 데이터이다. 그리고 보통 통계와 같은 데이터들은 엑셀파일 형태로 많이 구현되어 있기 때문에 이러한 엑셀파일을 처리할 수 있는 방법을 모색해야 하며 수많은 데이터들을 어떤 형태로 저장 수치화 및 도표화를 시킬 것인지에 대한 방법도 생각하여야 한다.

기본적으로 도표나 그래프 형태를 처리하기 위해서는 Window API 형태를 이용하는 것이 가장 무난하나 데이터 처리를 하는 언어에 따라서 UI의 형태는 변화할 가능성이 높다.

1. **관련 연구**

**2.1. 데이터 마이닝(Apriori 알고리즘)**

서로 다른 데이터를 두개를 연관 시키기 위해서 연관 규칙 알고리즘을 구현하는 것이 필요하다.

* 연관 규칙 탐사에서의 주요 정의

연관성 탐사에 사용되는 주요 용어에는 거래(transaction )，항목 집합(item set) ，빈발

항목 집합(large item set), 연관성 규칙(association rule), 지지도(support)，신뢰도

(confidence), 기여도(lift) 등이 있는데 이들은 ⑵의 정의를 그대로 따르기로 한다.

* 연관 규칙 탐사의 절차

연관성 탐사는 다음과 같이 크게 두 단계로 이루어진다.

• 1 단계 : 높은 지지도를 가지는 항목 집합 즉 최소 지지도를 만족하는 빈발 항목집합들을

찾는다.

• 2 단계 : 빈발 항목집합들의 내부로부터 최소 신뢰도를 만족하는 연관 규칙을 도출한다.

여기서는 주어진 빈발 항목 집합을 모든 가능한 두개의 부분집합(X，Y)로 구분

하여 R : X→ Y가 신뢰도를 만족할 경우 연관 규칙으로 출력한다.

연관성 탐사의 어려움은 주로 1 단계에 존재하며 따라서 대부분의 연관성 탐사 알고리즘은 1 단계를 효율적으로 수행하기 위한 방안에 집중하며 본 논문에서도 이들 빈발 항목

집합을 찾는 알고리즘들을 비교，분석한다.

* 빈발 항목 집합 발견을 위한 기초 알고리즘(Apriori 알고리즘)

최근 연구되고 있는 대부분의 빈발 항목 집합 발견 알고리즘의 기본은 ‘Apriori’에 근간을

두고 있는데， 이 방법은 초기 빈발 항목 집합을 발견한 후 그것을 바탕으로 다음 단계의 빈발항목집합을 찾는 과정을 반복하므로 연역적 접근 방식이라 할 수 있다. Apriori에서는 다음과 같은 용어를 정의한다.

• Lk (k = 1, 2, 3, •••，m) : k 단계에서 발견된 모든 빈발 항목 집합들의 집합

• Ck (k= 1, 2, 3, •••，m) : Lk 를 위한 후보 빈발 항목집합들의 집합

• D : 거래들이 들어있는 데이터베이스

위의 정의를 바탕으로 Apriori 알고리즘은 다음과 같이 연역적으로 진행한다〔3,4].

1. C1을 생성한다(각 항목 하나 하나를 원소로 하는 후보 빈발 항목집합들을 구성).
2. C1에 속한 각 후보 빈발항목집합 즉 각 원소 하나하나 단위로，DB를 스캔하여 지

지도를 계산한다.

1. L1 을 생성한다(C1중에서 최소 지지도를 만족하는 빈발 항목집합들만 선택).
2. k = 2로부터 시작하여 다음을 반복한다.
3. Lk 로부터 Ck+1을 생성한다. (Ck+i= Lk \* Lk : 현재 k-item set 들로 구성된 Lk

로부터 각 k-item set 간의 관계를 살 펴서 (k+ l)\_ item set 들을 생성함)

(2) Ck+1의 각 원소((k + l)- item se t)에 대하여 DB를 스캔하여 지지도를 계산한다.

(3)Ck+1로부터 최소 지지도를 만족하는 것만 선택하여 Lk+1을 구성한다.

(4) k = k + 1

(5) 가 NULL 이면 반복을 중단한다.

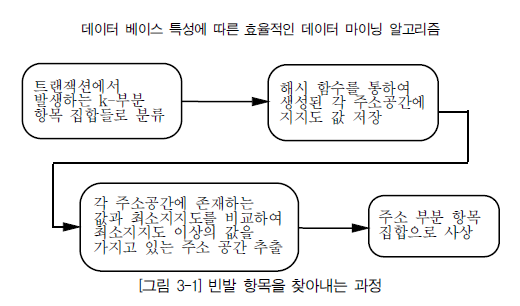
위의 내용을 알고리즘으로 표현하게 되면 아래와 같다.



**2.2. 데이터 마이닝 알고리즘 “EDHPA”**

해슁 기법을 이용하여 빈발 항목 집합을 생성하는 DHP 알고리즘을 리모델링 하여 한

번의 데이터베이스 스캔으로 각 트랜잭션에서 발생 가능한 모든 k-부분집합들의 지지도를 계산함으로써 빈발 항목집합들을 생성하는 EDHPA(Efficient DHP Algorithm) 알고리즘을 제안한다. EDHPA 알고리즘은 비교적 길이가 짧은 트랜잭션 데이터베이스를 대상으로 전체 트랜잭션을 이루는 모든 부분 항목집합들을 버켓이라는 기억 공간 속에 저장하는 알고리즘이다. EDHPA 알고리즘은 트랜잭션들의 길이가 짧거나 데이터베이스를 이루는 항목들의 수가 비교적 적은 트랜잭션 데이터베이스에서 해슁 기법을 사용하여 기존의 알고리즘보다 더 짧은 시간에 빈발항목집합을 발견할 수 있는 효과적인 연관 규칙 탐사 알고리즘이다. EDHPA 알고리즘에서의 빈발항목집합을 찾아내는 과정은 [그림 3-1]과 같다.

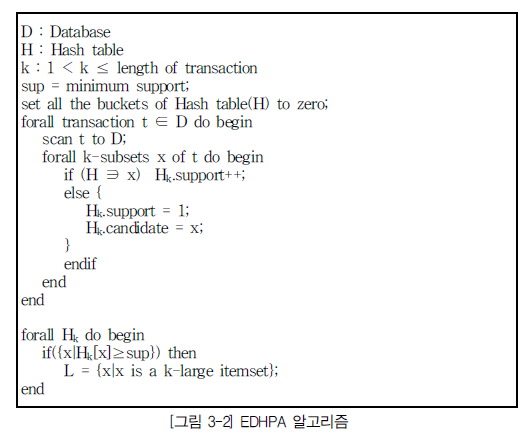


기존의 알고리즘에서는 1-항목 부분집합들에 대한 지지도를 먼저 확인하고, 1-빈발 항목 집합들이 될 수 있는지의 여부를 확인한다. 다시, 1-빈발 항목 집합들은 기존의 후보 항목 집합들을 생성하기 위한 알고리즘에 의해서 2-후보 항목집합들이 생성되고 이는 2-빈발 항목 집합을 찾기 위해 사용된다. 이러한 순차적인 과정이 기존의 대부분의 알고리즘들에서는 사용되지만 EDHPA 알고리즘에서는 이러한 반복 과정을 거치지 않고 데이터베이스를 한 번 스캔한 후에 모든 부분 항목집합들의 지지도가 계산되기 때문에 매 단계마다 확인을 할 필요는 없다.

연관 규칙 탐사에서 1-부분집합은 기존의 알고리즘들에서 보듯이 다음단계의 후보를 생성하기 위해서는 필요하지만, 항목들의 연관성을 찾는 탐사에서 1-빈발항목집합은 의미가 없다. 왜냐하면, 하나의 항목에 대한 연관성은 존재하지 않기 때문이다.

분류된 각 항목집합들은 해쉬 함수를 통하여 생성된 각 주소 공간, 즉 버켓에 부분 항목집합들이 아닌 지지도 값이 저장된다. 데이터베이스를 이루는 각 트랜잭션에서 발생 가능한 모든 부분 항목집합들은 해쉬 함수로 인해서 그들만의 유

일한 주소를 가지게 되고 그 주소에는 같은 부분 항목집합들이 존재할 때마다 1씩 증가하게 된다. 결국 모든 부분 항목집합들에 대해서 수행이 완료되면 모든 트랜잭션에 대한 부분 항목집합들의 주소가 생성되고 이들 주소에는 결국 부분 항목집합들의 지지도가 저장되게 된다. 두 번째는 각 베켓에 존재하는 값과 최소 지지도를 비교하여 최소 지지도 이상의 값을 가지고 있는 버켓들을 추출하고 주소에 해당하는 부분 항목집합으로 사상 시켜주면 결국 우리가 원하는 빈발 항목집합들을 찾게 된다. EDHPA 알고리즘은 [그림 3-2]과 같다.



**2.3. 기존 연구와의 차이점 및 변경 방법**

지금 이 연구 자료의 경우 기본적으로 DBMS를 기반으로한 데이터 베이스 기반으로 알고리즘 및 코드가 구성되어 있다. 하지만 내가 베이스 기반으로 할 데이터들은 빅데이터 사이트에서 받아오는 실제적인 데이터들이기 때문에 DBMS를 이용하지 않을 경우도 발생 할 수 있다. 그렇기 때문에 DBMS를 이용하지 않은 통상적인 텍스트 파일과 엑셀 파일에 대해서 동향을 분석 및 처리할 수 있도록 처리를 해주어야 한다.

즉 기본적인 데이터를 처리 하는 방법을 2가지로 나누어 볼 수가 있게 된다. 기존에 있는 엑셀 데이터를 DBMS로 저장 그 이후의 처리, 두번째로 기존에 있는 엑셀 데이터를 처리하지 않고 그대로 처리하는 부분으로 2가지의 방법이 존재한다.

물론 관계 데이터의 알고리즘도 DBMS의 기반이 아닌 기본적으로 받아올 수 있는 엑셀 파일을 기반으로 처리해야 할 가능성이 상당히 높아진다. 그렇기에 1차적으로 분류되서 나오는 자료의 모음에 대해서 알고리즘의 로직을 적용 그것을 내가 만든 데이터 정리 형태에 맞춰서 조정할 필요가 있다.

그렇기에 기본적인 C나 java가 아닌 R을 이용하여서 구현을 할 가능성도 상당히 높을 것이다. 그렇기 때문에 데이터 처리 및 데이터 연관관계 알고리즘을 R의 형태로 재구성할 필요가 있다.2. 1. 연관 규칙 탐사에서의 주요 정의연관성 탐사에 사용되는 주요 용어에는 거래(transaction )，항목집합(itemset) ，빈발항목집합(large itemset), 연관성 규칙(association rule), 지지도(support)，신뢰도

(confidence), 기여도(lift) 등이 있는데 이들은 ⑵의 정의를 그대로 따르기로 한다.

2.2■ 연관 규칙 탐사의 절차

연관성 탐사는 다음과 같이 크게 두 단계로 이루어진다.

• 1단계 : 높은 지지도를 가지는 항목집합 즉 최소 지지도를 만족하는 빈발 항목집합들을

찾는다.

• 2단계 : 빈발 항목집합들의 내부로부터 최소 신뢰도를 만족하는 연관 규칙을 도출한다.

여기서는 주어진 빈발 항목집합을 모든 가능한 두개의 부분집합(X，Y)로 구분

하여 R : X ᅳ Y가 신뢰도를 만족할 경우 연관 규칙으로 출력한다.

연관성 탐사의 어려움은 주로 1 단계에 존재하며，따라서 대부분의 연관성 탐사 알고리

즘은 1 단계를 효율적으로 수행하기 위한 방안에 집중하며 본 논문에서도 이들 빈발 항목

집합을 찾는 알고리즘들을 비교，분석한다.

2 .2 . 1 . 빈발 항목집합 발견을 위한 기초 알고리즘(Apriori 알고리즘)

최근 연구되고 있는 대부분의 빈발 항목집합 발견 알고리즘의 기본은 ‘Apriori’에 근간을

두고 있는데，이 방법은 초기 빈발 항목집합을 발견한 후，그 것을 바탕으로 다음 단계의

빈발항목집합을 찾는 과정을 반복하므로 연역적 접근방식이라 할 수 있다. Apriori에서는

다음과 같은 용어를 정의한다.

• L k (k = l , 2, 3, •••.，m) : k 단계에서 발견된 모든 빈발항목 집합들의 집합(k\_th

large itemset 다 set of k-itemset))

• Ck(k= 1, 2, 3, .•••，m) : 기를 위한 투보 빈발 항목집합들의 집합(k\_th candidate

large item se t여 set of k-itemset))

• D : 거래들이 들어있는 데이터베이스

연관 규칙 탐사 알고리즘의 비교 분석

71

위의 정의를 바탕으로 Apriori 알고리즘은 다음과 같이 연역적으로 진행한다〔3,4].

① C i을 생성한다(각 항목 하나 하나를 원소로 하는 후보 빈발 항목집합들을 구성).

② C i에 속한 각 후보 빈발 항목집합 즉 각 원소 하나하나 단위로，DB를 스캔하여 지

지도를 계산한다.

③ 쇼1을 생성한다(C i중에서 최소 지지도를 만족하는 빈발 항목집합들만 선택).

④ k = 2로부터 시작하여 다음을 반복한다.

O Z&로부터 Ck+i을 생성한다. (Ck+i = Z& \* Z& : 현재 k-itemset 들로 구성된 Z&

로부터 각 k-itemset 간의 관계를 살 펴서 (k+ l)\_ item se t 들을 생성함)

@ Ck+i 의 각 원소((k + l)- item se t)에 대하여 DB를 스캔하여 지지도를 계산한다.

@ Ck+i로부터 최소 지지도를 만족하는 것만 선택하여 Lk+i을 구성한다.

© k = k + 1

0 Lk7> NULL 이면 반복을 중단한다.

위의 내용을 알고리즘으로 표시하면〔그림 1〕과 같다. 이 알고리즘의 핵심 특징은 아래

와 같다.2.2■ 연관 규칙 탐사의 절차

연관성 탐사는 다음과 같이 크게 두 단계로 이루어진다.

• 1단계 : 높은 지지도를 가지는 항목집합 즉 최소 지지도를 만족하는 빈발 항목집합들을

찾는다.

• 2단계 : 빈발 항목집합들의 내부로부터 최소 신뢰도를 만족하는 연관 규칙을 도출한다.

여기서는 주어진 빈발 항목집합을 모든 가능한 두개의 부분집합(X，Y)로 구분

하여 R : X ᅳ Y가 신뢰도를 만족할 경우 연관 규칙으로 출력한다.

연관성 탐사의 어려움은 주로 1 단계에 존재하며，따라서 대부분의 연관성 탐사 알고리

즘은 1 단계를 효율적으로 수행하기 위한 방안에 집중하며 본 논문에서도 이들 빈발 항목

집합을 찾는 알고리즘들을 비교，분석한다.

2 .2 . 1 . 빈발 항목집합 발견을 위한 기초 알고리즘(Apriori 알고리즘)

최근 연구되고 있는 대부분의 빈발 항목집합 발견 알고리즘의 기본은 ‘Apriori’에 근간을

두고 있는데，이 방법은 초기 빈발 항목집합을 발견한 후，그 것을 바탕으로 다음 단계의

빈발항목집합을 찾는 과정을 반복하므로 연역적 접근방식이라 할 수 있다. Apriori에서는

다음과 같은 용어를 정의한다.

• L k (k = l , 2, 3, •••.，m) : k 단계에서 발견된 모든 빈발항목 집합들의 집합(k\_th

large itemset 다 set of k-itemset))

• Ck(k= 1, 2, 3, .•••，m) : 기를 위한 투보 빈발 항목집합들의 집합(k\_th candidate

large item se t여 set of k-itemset))

• D : 거래들이 들어있는 데이터베이스

연관 규칙 탐사 알고리즘의 비교 분석

71

위의 정의를 바탕으로 Apriori 알고리즘은 다음과 같이 연역적으로 진행한다〔3,4].

① C i을 생성한다(각 항목 하나 하나를 원소로 하는 후보 빈발 항목집합들을 구성).

② C i에 속한 각 후보 빈발 항목집합 즉 각 원소 하나하나 단위로，DB를 스캔하여 지

지도를 계산한다.

③ 쇼1을 생성한다(C i중에서 최소 지지도를 만족하는 빈발 항목집합들만 선택).

④ k = 2로부터 시작하여 다음을 반복한다.

O Z&로부터 Ck+i을 생성한다. (Ck+i = Z& \* Z& : 현재 k-itemset 들로 구성된 Z&

로부터 각 k-itemset 간의 관계를 살 펴서 (k+ l)\_ item se t 들을 생성함)

@ Ck+i 의 각 원소((k + l)- item se t)에 대하여 DB를 스캔하여 지지도를 계산한다.

@ Ck+i로부터 최소 지지도를 만족하는 것만 선택하여 Lk+i을 구성한다.

© k = k + 1

0 Lk7> NULL 이면 반복을 중단한다.

위의 내용을 알고리즘으로 표시하면〔그림 1〕과 같다. 이 알고리즘의 핵심 특징은 아래

와 같다.

2 . 연관 규칙 탐사 알고리즘

2. 1. 연관 규칙 탐사에서의 주요 정의

연관성 탐사에 사용되는 주요 용어에는 거래(transaction )，항목집합(itemset) ，빈발

항목집합(large itemset), 연관성 규칙(association rule), 지지도(support)，신뢰도

(confidence), 기여도(lift) 등이 있는데 이들은 ⑵의 정의를 그대로 따르기로 한다.

2.2■ 연관 규칙 탐사의 절차

연관성 탐사는 다음과 같이 크게 두 단계로 이루어진다.

• 1단계 : 높은 지지도를 가지는 항목집합 즉 최소 지지도를 만족하는 빈발 항목집합들을

찾는다.

• 2단계 : 빈발 항목집합들의 내부로부터 최소 신뢰도를 만족하는 연관 규칙을 도출한다.

여기서는 주어진 빈발 항목집합을 모든 가능한 두개의 부분집합(X，Y)로 구분

하여 R : X ᅳ Y가 신뢰도를 만족할 경우 연관 규칙으로 출력한다.

연관성 탐사의 어려움은 주로 1 단계에 존재하며，따라서 대부분의 연관성 탐사 알고리

즘은 1 단계를 효율적으로 수행하기 위한 방안에 집중하며 본 논문에서도 이들 빈발 항목

집합을 찾는 알고리즘들을 비교，분석한다.

2 .2 . 1 . 빈발 항목집합 발견을 위한 기초 알고리즘(Apriori 알고리즘)

최근 연구되고 있는 대부분의 빈발 항목집합 발견 알고리즘의 기본은 ‘Apriori’에 근간을

두고 있는데，이 방법은 초기 빈발 항목집합을 발견한 후，그 것을 바탕으로 다음 단계의

빈발항목집합을 찾는 과정을 반복하므로 연역적 접근방식이라 할 수 있다. Apriori에서는

다음과 같은 용어를 정의한다.

2 . 연관 규칙 탐사 알고리즘

2. 1. 연관 규칙 탐사에서의 주요 정의

연관성 탐사에 사용되는 주요 용어에는 거래(transaction )，항목집합(itemset) ，빈발

항목집합(large itemset), 연관성 규칙(association rule), 지지도(support)，신뢰도

(confidence), 기여도(lift) 등이 있는데 이들은 ⑵의 정의를 그대로 따르기로 한다.

2.2■ 연관 규칙 탐사의 절차

연관성 탐사는 다음과 같이 크게 두 단계로 이루어진다.

• 1단계 : 높은 지지도를 가지는 항목집합 즉 최소 지지도를 만족하는 빈발 항목집합들을

찾는다.

• 2단계 : 빈발 항목집합들의 내부로부터 최소 신뢰도를 만족하는 연관 규칙을 도출한다.

여기서는 주어진 빈발 항목집합을 모든 가능한 두개의 부분집합(X，Y)로 구분

하여 R : X ᅳ Y가 신뢰도를 만족할 경우 연관 규칙으로 출력한다.

연관성 탐사의 어려움은 주로 1 단계에 존재하며，따라서 대부분의 연관성 탐사 알고리

즘은 1 단계를 효율적으로 수행하기 위한 방안에 집중하며 본 논문에서도 이들 빈발 항목

집합을 찾는 알고리즘들을 비교，분석한다.

2 .2 . 1 . 빈발 항목집합 발견을 위한 기초 알고리즘(Apriori 알고리즘)

최근 연구되고 있는 대부분의 빈발 항목집합 발견 알고리즘의 기본은 ‘Apriori’에 근간을

두고 있는데，이 방법은 초기 빈발 항목집합을 발견한 후，그 것을 바탕으로 다음 단계의

빈발항목집합을 찾는 과정을 반복하므로 연역적 접근방식이라 할 수 있다. Apriori에서는

다음과 같은 용어를 정의한다.

1. **프로젝트 내용**

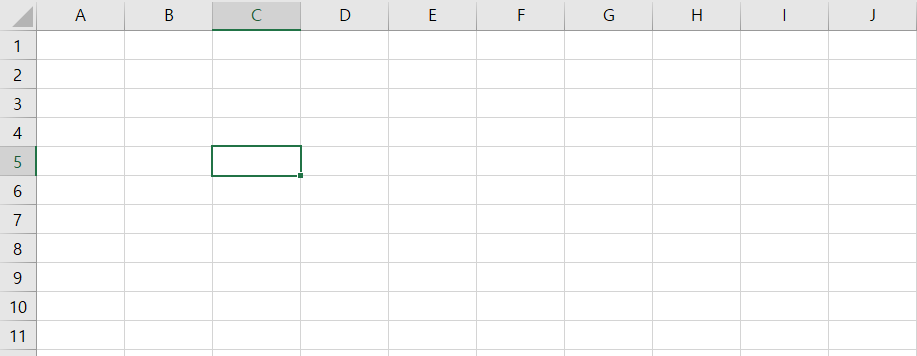
**3.1. 시나리오**

**3.1.1. 데이터 분석**

C++ or R

엑셀파일

데이터 산출

[](http://terms.naver.com/entry.nhn?docId=1073536)

기본적으로 받아올 자료들이 대부분 excel형태의 파일로 도표 형태의 상태가 많을 것이다. 그렇기에 DBMS를 이용하지 않게 된다. DBMS를 이용하지 않고 도표 형태의 데이터를 처리하기 위해서 R을 이용하여서 통계 프로그래밍을 통해서 데이터를 처리할 것이다 이렇게 처리된 데이터를 R이나 C++를 통해서 데이터를 정리 및 출력 처리를 할 것이다.

**3.1.2. 데이터마이닝**

[](http://terms.naver.com/entry.nhn?docId=1073536)

결과데이터

(새로운 데이터)

연관데이터 알고리즘

[](http://terms.naver.com/entry.nhn?docId=1073536)

다음으로 데이터 마이닝 과정이다. 데이터 마이닝의 경우 위의 통계자료로 얻은 데이터들을 토대로 최소 2개 혹은 그 이상의 데이터를 이용 서로 연관되는 부분에 대해서 찾고 그 연관된 데이터들끼리을 서로 비교하여 표현할 수 있는 또 다른 형태의 데이터를 구해야 하며 그 데이터를 출력 및 다른 데이터와 비교하기 위해 데이터를 excel이나 텍스트 형태로 저장을 해두어야 한다.

* 1. **요구사항**
     1. **입력 데이터**
* 입력데이터는 Excel도표로 이루어져 있는 파일을 사용하게 된다.
* 입력데이터의 경우 기준이 되는 범례가 설정되어 있어야 하며 결과의 수치값을 표현할 수 있어야 한다.
* 데이터가 표시하는 것은 하나가 아닌 여러가지의 수치를 표현할 수도 있다.
  + 1. **어플리케이션**
* 어플리케이션은 최초에 들어오는 통계데이터들을 일정한 기준을 가지고 분류하여 처리할 수 있어야한다.
* 분류된 데이터들을 도표나 그래프 형태로 표현할 수 있는 기능을 가지고 있어야한다.
* 이렇게 분류된 1차 정보 데이터들을 모아서 연관관계를 파악할 수 있는 알고리즘을 통해서 2차 정보를 도출 및 표현할 수 있어야한다.
* 이렇게 산출된 2차적으로 생성된 정보들도 다른 정보들과 연관시켜 새로운 정보로 만들 수 있으므로 생성된 정보를 하나의 데이터로 저장해 둘 수 있어야한다.

1. **결론(기대효과)**

앞으로 4차산업혁명 시대가 다가오게 되면서 일반적인 통계 데이터뿐만이 아니라 다양한 데이터들이 점점 늘어나게 될 것이다. 이러한 데이터들 중에서도 지금도 사용하고 있는 기본적인 통계데이터를 이용하여 데이터의 분석 및 향후 데이터의 동향과 데이터의 연관관계를 파악하는 것은 앞으로 통계 데이터뿐만 아니라 데이터의 형태를 확장시켜서 사용자의 어플리케이션의 이용패턴을 파악하는 곳이나 더 나아가서는 사물인터넷의 데이터나 인공지능과 같은 다양한 데이터들이 교차하여서 데이터마이닝을 통한 2차정보를 더 많이 얻어야 되는 환경에서 이용될 수 있다. 즉 지금 연구하고 있는 부분은 이러한 데이터간의 상호연관 처리의 기본적인 틀이 되는 모듈을 구현하는 점에 있다.

또한 이러한 데이터마이닝을 통한 2차 정보는 출력만 하는 것이 아닌 또 다른 정보를 얻기 위한 하나의 정보로서 저장되어서 데이터마이닝의 하나의 샘플이 될 수 있게 된다. 즉 2차 정보가 아닌 들어오는 정보와 생성된 새로운 정보를 통해서 연관관계를 찾으면 또 다른 새로운 정보를 얻을 수 있게 된다.

1. **참고문헌**

연관규칙 알고리즘 : http://www.dbpia.co.kr/Article/NODE00994608

EDHPA알고리즘 : http://www.dbpia.co.kr/Article/NODE00992037