情報管理研究 vol. 31, no. 3, 2000, pp. 1-20

정보관리분야의 데이터 마이닝 기법 적용에 대한 연구

A Study of Data Mining Application in Information Management Field

최 희 윤*

Heeyoon Choi

= 초 록 =

정보화가 진행됨에 따라 급속하게 축적되는 대량의 데이터로부터 가치있는 지식을 추출하려는 시도가 다양하게 진행되고 있으며, 그 일환으로 데이터 마이닝 기법이 관심의 대상이 되고 있다. 따라서 날로 증가하는 디지털 문헌에 대한 효율적인 처리와 서비스, 시스템 구축이 필요한 정보관리분야에 있어서도 이러한 기법의 적용 필요성은 증대하고 있다. 본고에서는 다양한 기법의 개발과 함께 활발하게 적용이 시도되고 있는 데이터 마이닝에 대한 이론적 배경 및 실제 적용 사례를 개관해 보고, 최근 들어 그 필요성이 대두되고 있는 정보관리분야에의 데이터 마이닝 기법의 적용 영역 및 활용 가능성을 관련 연구의 분석을 통해 살펴 보고자 한다.

= 키워드 =

데이터 마이닝, 정보관리, 텍스트 마이닝, 지식탐사, 데이터웨어하우스

- ABSTRACT -

A variety of trials selecting necessary and valuable information from rapidly increasing volume of data are made, and as one of them, data mining methods is an interest. This methodology is increasingly applied to information management field which consists of efficient processing and systemizing increasing digital documents for user service. This article analyzes theoletical background and empirical case studies of data mining, and predicts the possibility of its application to information management area.

- KEYWORDS -

Data Mining, Information Management, Text Mining, Knowledge Discovery

* 포스코경영연구소 지식자산실장
(Director, Knowledge Asset Center, POSRI)

1. 서론

정보화 사회의 도래와 함께 정보시스템의 구축이 활발히 진행됨에 따라 축적된 많은 양의 데이터를 가치있는 지식으로 변환하려는 노력이 확산되고 있다. 최근에 도입되는 데이터 마이닝(data mining)과 같은 개념은 "대량의 데이터에서 업무상 의미있는 패턴을 발견하는 것"으로 그 목적이 추상적으로 표현되고 있지만 이러한 사회적 필요성을 반영하고 있다고 볼 수 있다. 즉 최근 20년간 데이터와 정보는 폭발적인 증가를 보여왔으며 이러한 데이터와 정보의 증가는 데이터 저장소의 역할과 그 영역의 증대를 가져왔다. 그러나 이러한 데이터 저장소의 역할과 영역확대는 필요로 하는 데이터를 찾기 어렵게 한다는 문제점을 야기시켰을 뿐만 아니라 데이터 처리를 위한 시간과 비용을 크게 증대시켰다(Dilly 1995).

이와 같은 데이터 크기의 비약적인 증대는 사용하고자 하는 데이터의 추출과 데이터 저장공간 확보 등의 문제점을 유발하였고 이러한 문제를 해결하기 위하여 다양한 분야에서 데이터 마이닝을 적용한 연구가 수행되고 있다. 특히 대규모의 디지털 문헌의 효율적인 처리와 시스템 구축을 통해 이용자에게 양질의 정보를 신속하고 정확하게 제공하는 것을 최대의 목표로 하는 도서관이나 정보센터와 같은 여러 유형의 정보시스템에서 이러한 연구의 필요성은 더욱 크다고 하겠다.

본 고에서는 다양한 기법의 개발과 함께 활발하게 적용이 시도되고 있는 데이터 마이닝에 대한 이론적 배경 및 실제 적용 사례를 개관해 보고, 최근 들어 그 필요성이 대두되고 있는 정보관리 분야에의 데이터 마이닝 기법의 적용 영역 및 활용 가능성을 관련 연구의 분석을 통해 살펴 보고자 한다.

2. 이론적 배경

2.1 개념

"Mine"이란 의미는 채광하다, 즉 거대한 더미에서 가치있는 무언가를 캐낸다는 것이다. 즉 데이터 마이닝이라는 것은 방대한 양의 데이터 속에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 과정이라고 말할 수 있다. 이러한 데이터 마이닝의 개념에 대해서는 지식발견이라는 막연하고 추상적인 목적에 부합될 수 있도록 많은 사람에 의해서 다양하지만 기본적으로는 유사한 정의들이 부여되고 있다.

Fayyad(1996)는 데이터 마이닝은 엄청난 양의 대규모 데이터로부터 쓸모 있는 유용한 지식을 발견하고 이들 정보를 활용하기 위한 기술로서, 데이터베이스로부터 실제 데이터에서 알려지지 않은 일정한 패턴을 찾아내는 일련의 과정이라고 정의하고 있으며, Zekulin과 Parsaye(1996)는 데이터 마이닝은 대량의 데이터 사이에 묻혀 있는 패턴을 발견하고, 규칙을 추론함으로써 의사결정을 지원하고, 그 효과를

예측하기 위한 기법이라고 소개하고 있다. 또한 John(1996)은 데이터 마이닝이란 데이터에서 이로운 패턴을 발견해 내는 과정으로, Ferruza(1996)는 방대한 데이터베이스로부터 숨겨진 지식, 예상치 않았던 패턴 및 새로운 규칙 등을 발견하는 지식탐사(KDD : Knowledge Discovery in Database) 과정의 일부분으로 데이터 마이닝을 정의하고 있다(Friedman, 1997). 결국 데이터 마이닝은 가트너그룹과 SAS Institute에서 각각 지적한 바와 같이 "하나의 분석기법이 아니라 여러 기법과 방법들의 적절한 조합으로 이루어진 일련의 과정"으로, 또한 "대용량의 데이터에 숨겨져 있는 데이터간의 관계, 패턴을 탐색하고 이를 모형화하여 업무에 적용할 수 있는 의미있는 정보로 변환함으로써 기업의 의사결정에 적용하는 일련의 과정"으로 정의할 수 있겠다.

2.2 출현 배경

데이터 마이닝은 축적된 정보기술의 발달과 비즈니스적 요구에 의해 시장에 등장하게 되었다고 볼 수 있다. 1980년대 이후 급속한 성장을 이룬 정보기술의 발달에 근거하여 기업들은 수십 기가 이상에 이르는 방대한 양의 데이터를 저장하고 관리하기 위한 데이터베이스의 구축에 많은 투자와 노력을 기울여 왔다. 그리고 이러한 방대한 양의 데이터베이스를 실제 업무에 있어 활용도를 높이기 위한 방편으로 정제되고 일관성있게 통합된 형태로 쌓아두고자 하는 데이터웨어하우스의 구축이 시도되고 있다. 즉 기업이 얻고자 하는 정보의 근간은 바로 각 기업이 보유하고 있는 고객, 상품, 경쟁사 등과 관련한 데이터 등과 매일 발생하는 거래 혹은 이용 데이터이다. 정보기술 및 시스템의 발전으로 이러한 데이터들은 보다 손쉽게 접근할 수 있고 효과적으로 활용될 수 있도록 관리할 수 있게 되었다.

또한 과다해지는 기업간 경쟁의 상황에서 더 다양화되고 개성화되고 있는 고객들의 요구에 대한 적절하고 빠른 대응이 기업의 경쟁력 척도가 되고, 지속적인 경쟁우위를 확보하기 위해서는 효과적이고 합리적으로 결정하는 신속한 경영전략의 결정이 중요한 요소가 되고 있다. 따라서 기업들은 최적의 경영전략이나 의사결정을 뒷받침해 줄 수 있는 의미있는 고급정보를 필요로 하게 되었다. 이러한 환경에서는 자연스럽게 데이터를 쌓아놓기만 하는 단계를 벗어나 이 거대한 창고에서 보다 가치 있는 정보를 효과적으로 찾아내는 데 새로운 관심이 모아지고 있다. 이를 위하여 시스템과 운영기법을 모두 개발해야 할 필요가 대두되는데 여기에 해당하는 것이 바로 데이터웨어하우스와 데이터 마이닝이다. 정확하고 올바른 결과를 얻기 위해서는 적용대상이 되는 데이터에 오류가 없는 정제되고 표준화된 체계적인 구조의 데이터가 준비되어야 한다. 따라서 데이터웨어하우스는 데이터 마이닝을 위한 효율적인 출발점이 된다고 볼 수

있으며 상호 보완관계를 가지고 발전하고 있다.

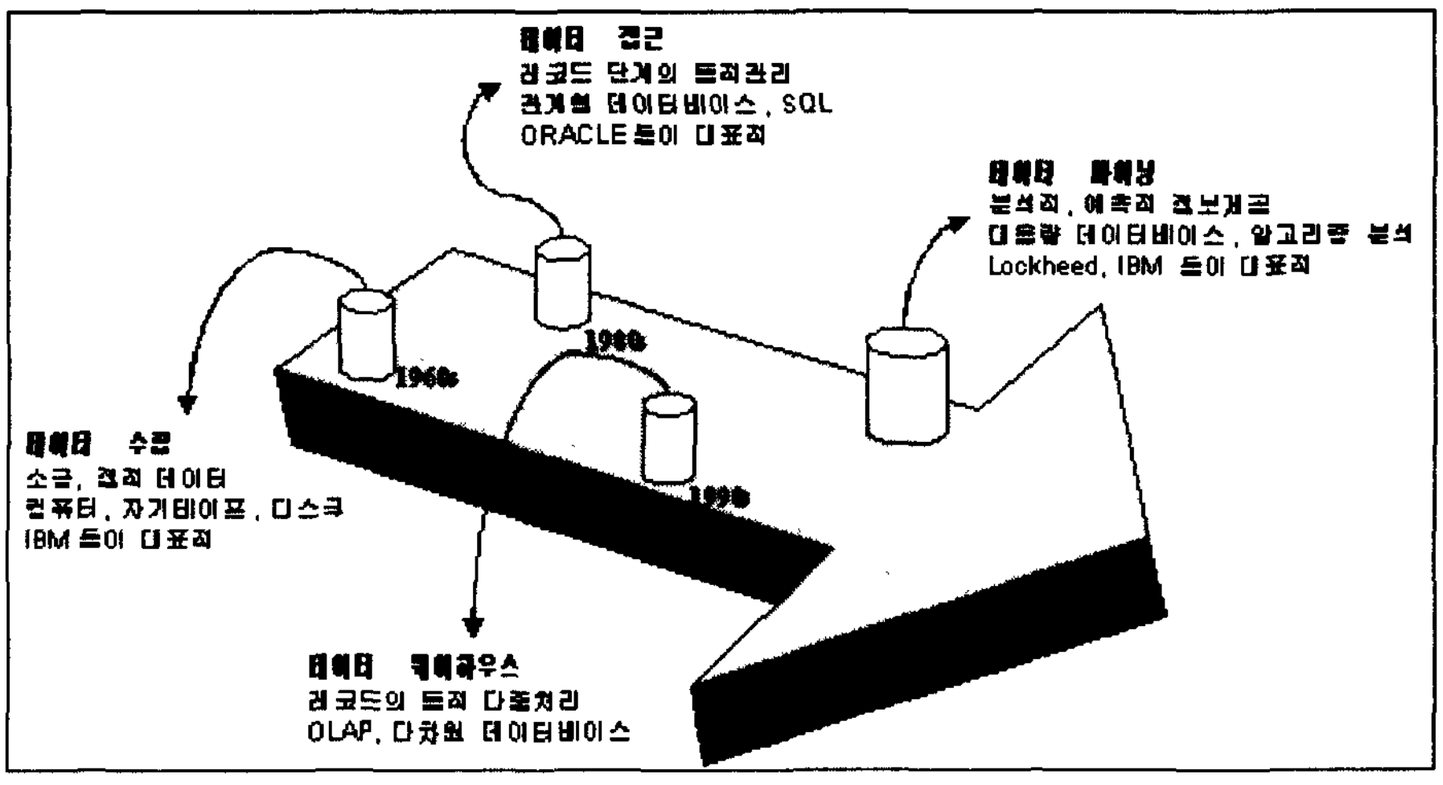
2.3 발전 단계

컴퓨터가 개발된 이래로 데이터의 관리는 몇가지의 발전단계를 거쳐 왔다. 제1단계는 1960년대로 데이터의 수집에 중점을 둔 시대이다. 주로 테이프나 디스크 등이 사용되었다. 제2단계는 데이터 액세스 단계로서 1980년대 이후의 관계형 데이터베이스로 대표되는 시기이다. 제3단계는 데이터 웨어하우스 단계로서 과거의 데이터를 동적으로 여러 수준을 통해 처리해 주는 단계이다. 제4단계는 대량의 데이터로부터 분석과 예측을 수행하는 데이터 마이닝의 단계로 볼 수 있다. 〈그림 1〉은 데이터 마이닝으로의 발전단계를 정리한 것이다.

2.4 수행과정

데이터 마이닝은 한두가지의 특정 기법이 아니라 일련의 기법들의 조합으로 이루어진 과정으로 어떤 특정 문제에만 적용되도록 준비된 방안이 아니라 여러 분야에서 적용할 수 있는 정보추출의 개념적인 방법론이다. 따라서 데이터집합으로부터 시작하여 지식을 추출하기까지의 과정을 구분하면 5가지 단계로 나눌 수 있다. 먼저 선정(selection) 즉 샘플링(sampling)은 어떤 기준에 따라 데이터를 선택하고 분할하느냐에 관한 것으로 예를 들면 자동차를 소유한 사람에 대한 데이터를 선정하는 것이다. 두번째 단계인 사전처리(preprocessing)은 데이터 정리 단계로서 불필요한 데이터를 제거하여 속도가 느려지지 않도록 미리 가공하는 것이다. 세번째 단계인 변형(transformation) 혹은 조정단계에서

〈그림 1〉 데이터 마이닝으로의 발전단계



데이터는 단순히 이동되기 보다는 중첩과 같은 변형이 가능하다. 네번째 단계는 패턴 추출 혹은 모델링 단계로서 이 단계를 데이터 마이닝이라고 하기도 한다. 다섯번째 단계는 해석 및 평가로서 시스템에서 발견한 패턴이 인간의 의사결정을 지원할 수 있도록 지식으로 해석하는 부분이다.

2.5 관련 기법

1) 데이터 마이닝과 지식탐사

지식탐사(knowledge discovery)란 대용량의 단편적인 데이터베이스에서 유용한 정보를 추출하고 이용하는 일련의 과정을 일컫는 용어이다. 최근 이 분야는 데이터베이스, 연역 데이터베이스, 인공지능, 기계학습, 통계학 등 기존 연구분야를 기반으로 독자적인 연구분야로 구축되고 있다. 지식탐사는 주어진 데이터베이스로부터 대상이 되는 데이터를 선별하는 과정, 선별된 대상 데이터를 적절한 형태로 변환하는 과정, 변환된 대상 데이터에 대해 마이닝 알고리즘을 적용하여 수행하는 과정, 수행결과를 이해하기 쉽게 표현하는 과정으로 이루어지고 있다. 지식탐사의 응용분야로는 정보관리, 질의처리, 의사결정, 프로세스 제어 등 다양하며, 최근 발전하는 온라인 정보서비스와 웹 정보서비스에도 보다 나은 정보의 제공과 사용자의 행동양식 등을 파악하는 등에 중요한 역할을 수행할 것으로 예상되고 있다(김성민 외 1998).

2) 데이터 마이닝과 데이터웨어하우스

기본적으로 오류가 있는 데이터에는 아무리 뛰어난 기법을 적용한다고 해도 그 결과에 대해서는 신뢰할 수 없을 것이다. 마찬가지로 데이터 마이닝을 통하여 정확하고 올바른 결과를 얻기 위해서는 적용될 데이터가 정제되고 표준화되어, 일관성 있는 체계적인 구조로 준비되어야 한다. 그러나 일반적으로 기업이 가지고 있는 데이터들은 다양한 운영시스템으로부터 모아지므로 각기 다른 형식과 값을 가질 수 있고 따라서 서로 일치하지 않거나 부정확한 값을 가지고 있는 경우가 많다. 그러므로 이러한 문제점에 대한 해결이 선행되어야 비로소 효과적인 데이터 마이닝을 수행할 수 있다. 이러한 부분들은 데이터웨어하우스가 구축되어 있다면 보다 쉽게 해결될 수 있다. 따라서 데이터웨어하우스는 효율적인 데이터 마이닝을 위한 출발점이 된다고 할 수 있겠다.

3) 데이터 마이닝과 OLAP

데이터 마이닝과 OLAP(On-Line Analytical Processing) 기법의 차이는 다음과 같다. OLAP은 대용량의 데이터베이스에 이용자가 직접 접근하여 다차원적 질의와 상호작용을 통해 정보를 탐색해 나가는 과정이다. 즉 OLAP이 이용자가 보고자 하는 관점에서의 데이터를 요약하는 기법이라면 데이터 마이닝은 이용자의 본질적인 의문에 해답이 될 수 있는 지식을 발

견하는 기법이다. 따라서 이들은 상호 보완적인 관계에 있다. OLAP을 이용하여 얻을 수 있는 정보를 기반으로 데이터 마이닝에 접근한다면 보다 효과적으로 고급정보를 획득할 수 있다. 또한 데이터 마이닝의 결과는 데이터베이스에 적용되어 OLAP을 통해 그 효과가 극대화될 수 있다. 따라서 데이터 마이닝의 목표 중 하나는 예측과 설명이 된다. 예측이란 관심 주제의 아직 알려지지 않은 미래값을 추정하기 위해 데이터베이스 내에 있는 기존 변수를 이용하는 것이며, 설명이란 데이터와 이에 따른 결과로 발생하는 패턴의 발견을 목적으로 한다(이희석, 장재경 1997).

2.6 구현 기법

데이터 마이닝의 모형화 단계는 다양한 기법이 적용될 수 있으나 특정문제에 적용하는 기법이 정해져 있지는 않다. 또한 하나의 기법으로 모든 문제를 해결할 수 있는 것은 아니며 얻고자 하는 결과나 데이터의 상태에 따라 적용할 수 있는 기법은 다를 수 있다. 데이터 마이닝의 구현기법에는 일반적으로 통계적 기법을 포함하여 연관성 측정(association), 연속성 탐사, 분류규칙, 클러스터링(clustering), 의사결정트리(decision trees), 신경망(neural network) 등의 기법이 있다. 또한 실제 데이터 마이닝을 수행할 ㅁ에는 대량의 데이터를 빠르게 모델링하여 표준 형태의 데이터로 변환하는 데이터 정리(data reduction)기술과, 데이터를 추출하는 예측방법(prediction method)도 필요하다(Weiss & Indurkhya 1997).

1) 연관성 측정

연관성 측정은 어떤 특정문제에 대해 아직은 일어나지 않은 해답을 얻고자 하는 예측의 문제나 고객을 특정목적에 따라 분할(segmentation)하는 문제가 아니라, 상품 혹은 서비스의 거래기록 데이터로부터 항목간의 연관성 정도를 측정하여 연관성이 많은 항목을 그룹화하는 클러스터링의 일종으로서 동시에 이용될 가능성이 많은 항목들을 찾아내는 과정에 쓰이며, 가장 많이 활용되는 데이터 마이닝의 지식탐사 기법이다. 연관성 측정에서 얻어지는 결과물인 연관규칙(association rule)은 A → B와 같은 형식으로 표현되며 이는 "상품 A가 판매된 경우에는 상품 B도 판매된다"와 같이 해석된다. 유용한 연관규칙을 유도하기 위해서는 먼저 어떠한 항목들이 어느정도의 연관성을 가지고 있는지 측정해야 할 것이다. 즉 관심있는 규칙이 얼마나 가능성이 있는가를 측정하는 연관성측정은 다른 기법에 비해서 기본개념이 간단하며 신뢰도(confidence), 지지도(support), 역상관도(lift)와 같은 측정지수가 사용되게 된다. 신뢰도는 A에 대한 거래중에서 B의 거래는 어느 정도인가를 측정하며, 지지도는 전체거래중 A와 B를 모두 포함하는 거래는 어느 정도인가를 측정하며, 역상관도(lift)는 임의로 B가 구매되는 경

우에 A와의 관계가 고려되어 구매되는 비율을 의미한다. 일반적으로 신뢰도의 값이 크면 좋지만 신뢰도가 높은 연관규칙이 최상은 아니다. 즉 신뢰도가 크면서 역상관도도 큰 값이 관찰되어야 유용한 정보라고 할 수 있을 것이다.

2) 순차적 패턴발견

동시에 구매될 가능성이 큰 상품군을 찾아내는 연관성 측정에서 시간이라는 요소를 도입하여 순차적으로 구매가능성이 큰 상품군을 찾아내는 것이다. 예를 들어 "새 컴퓨터를 구입한 사람들 중 25%는 그 다음달에 레이저 프린터를 구입한다"와 같은 연관규칙을 찾아내는 것을 순차적 패턴발견이라고 한다. 여기에서의 연관규칙은 일반적으로 "상품 A가 판매된 일정시간 후에는 상품 B가 판매된다"와 같이 해석된다. 즉 순차적 패턴발견은 구매의 순서가 고려되어 상품간의 연관성이 측정되고 이의 정도에 따라 유용한 연관규칙을 찾는 기법이다. 따라서 위의 연관성측정에서의 데이터 형태에서 각각 고객으로부터 발생한 구매시점에 대한 정보가 포함되어야 한다.

3) 클러스터링 · 분할

이 기법은 전체를 어떤 특정 기준에 따라 유사한 그룹끼리 나누는 것이다. 하나의 군집은 유사성에 따라 모인 일련의 개체이다. 유사성에 따른 클러스터링 기법은 계량적인 방법으로 유사성을 측정하는 기법이라고 볼 수 있다. 이는 스스로 규칙을 발견하고 학습해 가는 학습형으로 시스템에서는 대상 집합에 관련된 군집을 발견하고 나서 각 클러스터링에 대한 설명을 찾게 된다.

4) 분류(classification)

데이터 마이닝 도구는 데이터베이스에서 어떤 목표가 되는 문제를 파악하여 이를 예측하기 위하여 분류를 사용한다. 한번 분류가 정의되고 나면 시스템에서는 각 클래스에 대한 학습내용에 근거하여 분류를 주도할 규칙을 유추하게 된다. 생성된 분류 규칙은 조건문과 결과문으로 구성되게 된다.

5) 예측(prediction)

어떤 사건에 대한 가능성을 예측한다기 보다는 어떤 변수에 대한 미래를 예측하게 되는데 예를 들면 총수입 또는 지불청구 등의 예측이다. 예측모델이 수립되면 what-if 분석에 이를 활용할 수 있다. 예를 들어 광고, 상품위치, 가격 등과 같은 변수를 다양하게 변동시켜서 판매에 대한 영향을 예측할 수 있다.

2.7 사용 알고리즘

데이터 마이닝에 사용되는 대표적인 알고리즘은 대략 5가지로 구분할 수 있는데 이들은 다음과 같다.

1) 연관규칙

연관규칙(association rule) 알고리즘은 업무의 순서 및 관련성 파악이나

유사성 분석에 적합하다. 일반적으로 이 알고리즘을 이용한 도구는 지지도 수준을 계산하여, 모든 레코드에서 복수의 사건이 일어나는 확률을 구하게 된다. 다음에 신뢰도 수준을 계산하게 되는데 어떤 사건이 일어나고 다른 사건이 다시 일어나는 경우에 대한 경우의 확률을 구하게 된다.

연관규칙은 항목의 집합으로 이루어진 트랜잭션들의 데이터베이스에서 같은 트랜잭션에 나타나는 항목들간의 연관성을 표현하는 규칙으로서, A → B의 형식으로 표현된다. 이때 A와 B는 항목집합이며, 이 규칙이 뜻하는 것은 A 항목집합이 나타나는 트랜잭션에는 B 항목집합도 나타나는 경향이 있다는 의미이다. 연관규칙은 그 유용성이 크고 데이터 마이닝의 기본 연산중 하나인 중요한 지식패턴이라 할 수 있다(이동하 외 1998).

연관규칙을 이용한 전형적인 응용사례는 IBM의 쇼핑바구니 분석이다. 이에 따라 특정 브랜드의 토스터를 살 때 주방용 장갑과 커버를 함께 사는 고객이 20%라는 연관규칙을 표현할 수 있다. 연관규칙을 사용한 다른 예로는 의료보험 회사에 환자가 제출한 청구서를 분석하는 것이다. 각 청구서에는 환자가 받은 의료서비스 내역이 기재된다. 연관함수를 이용하여 환자가 치료받은 일련의 의료서비스를 분석하여 의료 서비스간의 관계를 발견하는 것이다.

2) K-최소이웃법

K-최소이웃법(K-nearest neighbor)은 군집분석과 분류에 가장 적합한 알고리즘이다. 우선 레코드를 다차원속의 한 개의 점으로 나타낸 후, 유사성이 높은 목적 속성을 갖고 있는 데이터 점을 하나의 클러스터가 되도록 각 차원의 가중치를 조정한다.

3) 의사결정 트리

의사결정 트리(decision tree)는 한정된 수의 클래스로 데이터를 분류하는 알고리즘으로 클러스터링과 분류기법에 가장 적합하다. 클러스터링으로는 모집단을 유사한 특성에 따라 세그먼트로 나누는 것이며 이 알고리즘에서는 예측을 보다 정교하게 하기 위하여 각 분할된 클러스터에 통계적인 유의성을 부여한다.

4) 신경망

신경망(neural network) 알고리즘은 예측, 클러스터링, 분류에 사용된다. 신경망은 복잡하고 불완전한 데이터로부터 의미를 유추하는데 탁월하며 패턴을 추출하거나 매우 복잡한 동향을 탐지하는데 이용될 수 있다. 훈련된 신경망은 전문가 수준에 오를 수도 있어 문제에 대해 예측하고 what-if 질문에 대한 응답에 이용될 수도 있다. 신경망은 뇌의 신경단위와 유사한 일련의 노드를 사용한다. 정해진 순서로 처리하는 전통적 알고리즘과는 달리 이런 노드에서는 데이터에 일단 노출이 되면 데이터 내의 패턴을 식별할 수 있는 네트웍으로 상호 연결된다.

신경망 알고리즘은 실제 비즈니스 문제에 폭넓게 활용되고 있으며 이미 많은 산업분야에서 성공적으로 적용되고 있다. 그러나 신경망은 출력결과를 제대로 설명할 수 없으며 장기간의 학습을 위한 데이터가 늘어날수록 결과가 나빠지는 문제점이 지적되고 있다.

5) 유전자 알고리즘

유전자 알고리즘은 진화이론에서와 같이 진화적인 방법으로 계산을 변형시키는 방법이다. 어떤 문제에 대한 잠정적인 해결방안을 상호 경쟁적으로 두어 생존원리에 따라 최선의 해결방안이 살아남게 된다. 유전자 알고리즘은 신경망과 자주 결합하여 사용하는데 신경망에서 몇가지의 모델을 발생시키면 유전자 알고리즘에서 최선의 모델을 선택하게 되는 것이다. 예를 들면 양호한 대출신청자를 식별하는 일과 같이 많은 변수를 갖고 있는 어려운 최적화 문제를 해결하는데 적합하다.

2.8 응용분야

데이터 마이닝에 대한 관심이 급속히 발전하게 된 것은 정보기술의 급속한 발전과 이러한 새로운 기술을 자신의 업무에 활용하기를 원하는 사용자 요구 때문이다 (Adrianns & Zantinge 1996). 데이터 마이닝은 산업계 전반에 걸쳐 다양한 분야에 응용될 수 있으며, 대표적인 응용분야는 다음과 같다.

1) 유통 · 마케팅 분야

데이터 마이닝이 현재 가장 많이 활용되고 있는 것은 유통 · 마케팅 분야이다. 기업이 가지고 있는 고객에 대한 데이터베이스를 중심으로 데이터 마이닝 기법을 활용하고 있다. 첫 번째 응용은 고객의 인구통계적 정보나 구매패턴 정보를 기반으로 고객을 세분화(segmentation)하고 그 특성을 요약하여, 그 결과를 바탕으로 마케팅에 활용함으로써 적은 비용으로 최대의 효과를 얻기 위한 분야를 들 수 있다. 두 번째는 고객의 성향을 파악함으로써 경쟁업체로의 전환 가능성이 있는 고객과 이탈 고객의 특성을 분석하여 고객 유지율을 향상시키고 고객과의 지속적인 관계를 유지해 나가는데 이용하고 있다. 세 번째로 현재 고객 자료를 토대로 고객을 순위별로 분류하고 이에 따라 새 고객의 유치에 활용하거나 우수 고객에 대한 서비스를 개선할 수 있다. 이외에도 고객응답을 분석하여 경제적이고 효율적인 광고전략을 수립하는 등 데이터베이스 마케팅에 활용되고 있다. 또한 소매업에서도 데이터 마이닝을 활용하고 있다. 가장 큰 활용범위로는 시장바구니 분석(market basket analysis)를 들 수 있다. 즉 생산품들의 각 구매시점을 이용하여 생산품간의 관련성을 찾아내고 이익을 극대화하기 위한 전략을 세우는데 활용이 된다. 미국의 경우 편의점 체인에서 아기 기저귀와 맥주 판매와의 아주 강한 관련성을 찾아내어 이를 수익성 증대에 효율적으로 활용한 사례가 있다. 또 다른 활용 예로는 유

통업자에게 재고결정을 할 수 있도록 해주는 시간패턴 분석이 있다. 이를 통해 어떤 고객이 캠코더를 구매하였다면 언제 별도의 배터리와 녹화용 테이프를 구매하게 될지를 예측할 수 있다. 또한 예측 모델로서 디자이너에 따른 구매 또는 충동구매와 같은 일시적인 유행이나 구매행동을 읽을 수 있는 순차적인 패턴을 탐지한 판매전략을 구사하여 특정 소비패턴을 보이는 고객의 프로파일을 개발할 수 있다. 또한 브랜드명과 POS 시스템을 이용함으로써 유통업체는 쇼핑행위에 관한 상세한 정보를 얻을 수 있다. 이는 여러 가지의 고객분석을 더 이해하기 쉽게 해준다. 먼저 고객유사성 분석으로 어떤 상품을 고객들이 함께 구입하는지를 분석하는 부문이 있으며 유통업자에게 재고결정을 할 수 있도록 해주는 시간패턴 분석이 있다. 이를 통해 어떤 고객이 캠코더를 구매하였다면 언제 별도의 배터리와 녹화용 테이프를 구매하게 될지를 예측할 수 있다. 또한 예측 모델로서 디자이너에 따른 구매 또는 충동구매와 같은 특정 소비패턴을 보이는 고객의 프로파일을 개발할 수 있다.

2) 금융 분야

금융분야에서도 잠재된 신용카드의 도용문제를 해결하기 위하여 신용카드 이용패턴을 탐색하는데 사용되고, 이자율이나 유통시장에서의 환율 변동의 예측에 데이터 마이닝 기법이 쓰여지고 있다. 또한 주식과 채권의 포트폴리오 관리에도 응용되고 있으며 신용위험 관리(risk management)와 기업대출에서의 도산예측(bankruptcy prediction)도 데이터 마이닝 기법을 활용하고 있다. 은행은 사기 탐지, 고객분할, 라이프사이클 예측 관리 등의 분야 등에서 여러 가지로 데이터 마이닝의 이익을 취할 수 있다. 먼저 은행의 크레디트 카드 업무에서 사기와 관련된 문제는 매우 큰 비용이다. 나중에 사기임이 밝혀진 지난 거래를 분석함으로써 은행은 어떤 패턴을 찾을 수 있다. 예를 들어 하나의 경보신호로서 단기간 가전제품 상가에서의 연속적인 구매행위를 찾아낼 수 있다. 은행에서는 이러한 지식을 활용하여 의심이 가는 거래를 승인하지 않도록 조처할 수 있다. 또한 은행은 고객을 분할함으로써 특정 그룹별로 다양한 서비스를 제공할 수 있게 된다. 예를 들면 은행에서는 여행을 자주 가는 고객에게 특정한 카드 상품을, 적시에 대금을 납부하는 고객에게는 다른 상품을 판매할 수 있다. 특정 홍보로부터 가장 많은 이익을 줄 수 있는 지점을 식별하는데도 고객분할을 이용할 수 있다. 또한 데이터 마이닝은 은행이 각 고객의 라이프사이클을 예측하고 각각에 적절하게 서비스할 수 있도록 한다. 은행별로 현재 이익을 낼 수 있는 고객의 프로파일을 개발하고 이전 몇해동안의 이들의 공통된 특성을 파악하는데 데이터 마이닝 기법을 이용하고 있다.

3) 보험, 증권 분야

보험회사는 매년 대량의 데이터를 수집하고 있으며 이에 기반하면 효과적인 계획을 수립할 수 있다. 보험회사는 클레임이 높은 분야에서 청구자간 발생빈도가 높은 관계를 분석해 봄으로써 사기행위를 감소시킬 수 있으며, 지불된 클레임에 대한 요인을 분석함으로써 책임에 대한 위험을 감소시킬 수 있다. 예를 들어 한 증권회사는 지난 2년간의 심각한 클레임에 대한 분석을 시도하여 기혼자의 클레임 총액은 미혼자의 2배에 달함을 발견하였다. 회사에서는 이러한 결과를 토대로 총괄적인 기혼자용 컨설팅 정책을 재평가하였다.

4) 통신 분야

통신부문은 매우 높은 경쟁에 직면해 있기 때문에 기존의 고객을 유지하고 새로운 고객을 유인하기 위한 공격적인 홍보, 마케팅 및 가격 프로그램을 제공하고 있다. 통신회사는 상세한 통화기록을 보유하고 있는데 이를 마케팅을 위하여 유사한 이용 패턴으로 고객을 분할함으로써 가격과 관련된 특징적인 고객군을 구별할 수 있을 것이다. 또한 한번 연결에 장시간 이용하는 고객의 특징을 식별함으로써 이에 대한 대책을 강구할 수 있다.

5) 기타 활용 분야

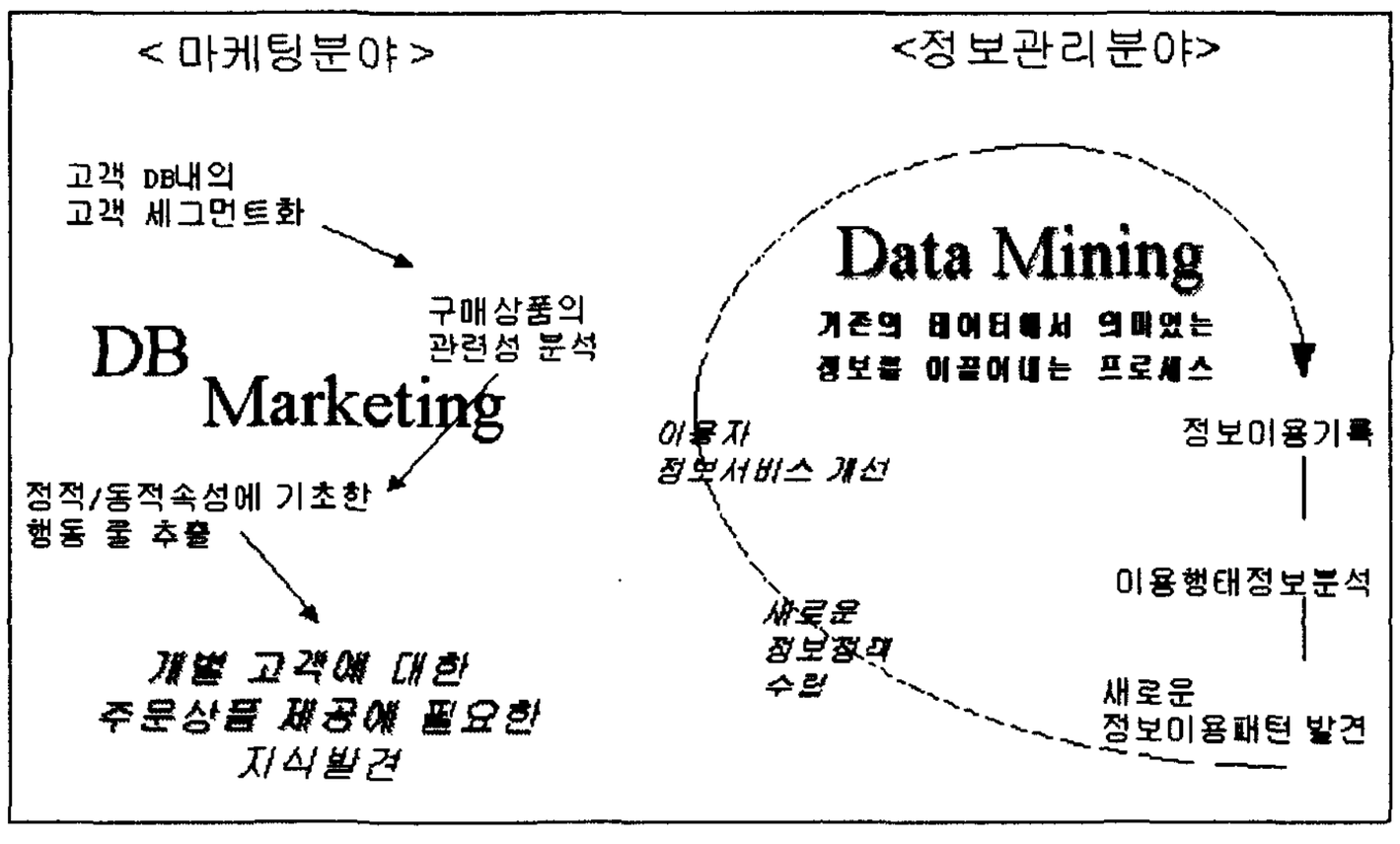
제조업 분야에서도 작업의 효율성과 품질의 우수성을 동시에 요구하는 현대의 제조업 성격에 맞게 데이터 마이닝 기법이 여러 분야에서 응용되고 있다. 공정제어 문제, 공정과정의 최적화, 에너지 소비문제, 품질관리 및 자동검사 등의 영역에 데이터 마이닝이 적용되고 있으며 보건 의학분야의 보험사기문제, 경제적 의료제공원 탐색 등에 활용되고 있음을 볼 수 있다 (SAS 1998).

자동차 제조업자는 자동차 주문을 받기전에 어떤 외양이 선호도가 높을 것이며, 함께 주문할 것으로 여겨지는 사양은 어떤 것이 있을지 예측할 필요가 있다. 또한 항공사에서도 자주 탑승하는 고객군의 특징을 분석할 수 있다. 예를 들면 마일리지는 충분하지 않지만 단거리 여행을 자주 하는 고객을 분할하도록 하여 여행횟수로도 혜택을 수혜할 수 있도록 할 수 있다.

6) 응용사례

응용 사례로는 금융업계의 우수고객 분류 및 특성 파악에 관한 문제해결 과정을 들 수 있다. 또한 소매업계의 대표적인 마케팅 전략인 Cross-Selling을 위하여 이용할 수도 있다. 즉 소매점에서 고객이 함께 구입할 가능성이 높은 상품들을 서로 가까운 곳에 노출되도록 하기 위하여 데이터 마이닝을 이용하는데 연관성 분석이나 연속패턴 분석등의 기법을 사용할 수 있다. 신용카드 회사에서는 악성 연체 회원의 카드 사용패턴을 발견하여 장기 악성연체로 가는 확률을 줄일 수 있고 은행 및 보험회사에서는 서로 연관성 있는 금융상품의 구매패턴을 발견하여 고객의 흥미를 유발하는 상품을 캠페인할

〈그림 2〉 마케팅과 정보관리 분야의 데이터 마이닝 적용 비교



수 있는 마케팅 전략을 수립할 수 있다(조성원, 1998). 그러나 이러한 분석은 충분히 대량의 데이터를 확보한 이후에나 가능한 일이며 건실한 데이터베이스가 없이는 불가능하다.

3. 정보관리분야의 데이터 마이닝 적용

3.1 적용영역

정보관리 분야에서도 데이터 마이닝을 적용해야 할 필요성이 높아지고 있다. 최근 이용자들이 구할 수 있는 디지털 문헌의 양은 급증하고 있고 그 내용과 이용자들의 요구도 매우 다양해지고 있다. 따라서 최근에 요구되는 정보시스템의 기능은 매우 방대한 데이터를 처리해야 하며 미리 정해지지 않은 다양한 정보요구에 대한 신속하고 유연한 대처가 요구된다. 또한 이용자의 정보이용 행태를 분석함으로써 정보나 서비스의 우선순위를 설정할 수 있으며, 이를 통해 효율적이고 전략적인 정보정책을 수립할 수 도 있다. 〈그림 2〉는 마케팅의 경우와 비교하여 정보센터의 경우 동일한 이용자 분석을 모형화한 예이다.

여기에는 도서관이나 정보센터에서 여러 유형의 정보시스템을 사용하는 이용자들의 정보이용 행태이나 패턴을 분석하여 정보정책의 수립이나 정보서비스 개선에 사용하는 이용자 연구분야와, 지능형 정보검색에서 질의처리 분야, 그리고 색인어의 추출을 위하여 디지털 원문 텍스트를 분석하는 텍스트 마이닝 등의 영역이 있다.

특히 최근에는 순차적 데이터 분석에 대한 관심이 높아지고 있고 텍스트 역시 순차적 데이터로 간주할 수 있기 때문에 데이터 마이닝의 연관규칙을

응용한 사건규칙(episode rule) 생성기법을 통하여 원문으로부터 단순한 키워드가 아닌 핵심 단어구를 추출하는 방법 또한 연구되고 있다(Helena Ahonen et al., 1998). 또한 데이터 마이닝과 웹 기술의 영향으로 웹 페이지 접근 기록으로부터 수집한 웹 로그 레코드에 대한 웹 로그 마이너의 설계를 통한 데이터 마이닝 기법과 데이터웨어하우징 기법의 적용이 시도되고 있다(Osmar et al., 1998).

3.2 이용자의 정보이용 행태분석

최근에 일반 이용자가 손쉽게 접근하는 정보시스템으로 가장 많이 활용되고 있는 것이 웹 서비스이다. 이들은 방대한 데이터를 담고 있으며 이용자들은 네트웍을 통하여 다양한 웹 서버에 접근하고 있다. 웹 서버에 대한 모든 접근은 로그 파일로 기록되며 대부분의 서버는 표준 포맷[1]에 따라 로그파일을 생성하게 된다 (Luotonen, 1995). 이들 각각의 로그기록은 접속한 이용자의 IP주소, 접근한 시각, 접근방법, 접근한 대상 문헌, 전송에 사용된 프로토콜, 에러기록, 전송데이터 용량 등과 같은 정보를 포함하고 계속 누적되어 방대한 데이터 집합을 구성하게 된다. 따라서 이러한 웹 로그 파일은 데이터를 관리하거나 이용자의 성향을 분석할 수 있는 기초적인 데이터로 사용할 수 있다. 이들 데이터를 분석하는 도구가 나와 있는 상태이지만 기초적인 통계 수준에 머무르고 있어 웹 서버 이용 패턴에 대한 통찰력 있는 분석을 제공하지는 못한다. 따라서 웹 로그 기록과 같은 방대한 데이터를 데이터마이닝을 통해 분석하여 온라인 정보서비스를 이용하는 이용자의 정보이용 행태를 탐사하는 연구가 시도되고 있다(Osmar et. al. 1998).

웹 로그 기록과 같이 순차적으로 누적되는 데이터와 같은 경우에는 데이터 마이닝의 주요 지식패턴인 연관규칙과 순차패턴 기법을 적용하는 추출하는 방법이 사용된다. 연관규칙은 이미 설명한 바와 같이 항목집합 A와 B에 대하여 A → B의 형태로 나타내며 항목집합 A가 나타날 때 항목집합 B도 함께 나타나는 경향이 있음을 의미하는 규칙이다. 여기에는 신뢰도(confidence)와 지지도(support)가 주요한 인자로 작용하게 된다. 신뢰도는 규칙이 실제로 맞아 떨어지는 정도를 나타내는 인자이고 지지도는 연관규칙을 반영하는 트랜잭션이 전체 데이터베이스에서 차지하는 비율을 나타내는 것이다(Agrawal et al. 1995). 한편 순차패턴은 시간적 순서에 따라 항목집합의 등장순서가 어떻게 나타나는지를 반영한다. 이는 항목집합 A, B에 대하여 {A, B}의 형태로 나타내며, 항목집합 A가 나타난 후에 항목집합 B가 나타나는 경향이 있음을 의미한다 (Agrawal, 1993). 즉 이러한 패턴은 특정한 페이지를 읽은 사람이 다음에 읽을 가능성이 높은 페이지를 예측할 수 있는 단서를 제공한다. 순차패턴에서

1)CERN과 NCSA에 의해 HTTP 프로토콜의 일부로 명시된 "Common Log Format"을 사용하고 있다.

는 지지도가 중요한 인자가 된다. 남도원 등(1997)은 로그 파일을 변환하지 않고 배경지식을 사용하는 모델을 정의하였다. 이 모델에 의하면 웹 로그 기록으로부터 다음과 같은 형태의 연관규칙을 발견할 수 있다.

1) 웹 페이지(URL) 사이의 연관규칙

웹 페이지간의 연관규칙에서는 사용자의 성향을 찾아낼 수 있다. 예를 들어 "테니스에 관한 사이트에 접근하는 사용자의 45%는 골프에 관한 사이트도 접근하려는 경향이 있다"라는 연관규칙이 발견되었다면 테니스에 관한 사이트에 골프 사이트에 대한 광고를 게재하는 방법을 통하여 보다 큰 효과를 얻을 수 있다.

2) 접근시간에 대한 연관규칙

이것은 이용자의 서비스 이용시간의 분석을 통하여 응용할 수 있다. 예를 들어 "오전 10시경에 접속한 사용자는 오후 5시경에 다시 접속하는 경향이 있다"라는 연관규칙이 발견되었다면 웹 사이트 기사의 갱신은 적어도 오후 5시 이전에 갱신되는 것이 좋다는 것으로 해석할 수 있을 것이다.

3) 웹 페이지간의 순차패턴

이용자가 접속하는 웹 페이지가 어떤 순서를 가진다면 이는 이들 사이트간의 계층적 관계를 재조정하는데 이용할 수 있다. 이용자가 스키, 스키복, 스노우보드의 순서로 정보를 검색하는 경향이 발견되었다면 정보의 네비게이션을 이러한 순서로 재조정하는 과정이 수행될 것이다.

4) 이용자간의 연관규칙

같은 사이트에 접속하는 이용자간에 연관규칙이 성립한다면 이들이 서로 공통된 관심사를 가지고 있다고 해석할 수 있다. 따라서 A, B, C가 비슷한 취향을 가진 연관규칙을 발견한다면 A와 B가 접속했으나 C는 아직 접속하지 못한 사이트가 있다면 이를 C에게 알려주는 형태로 이용자 관리를 발전시킬 수 있다.

5) 이용자간의 순차패턴

어떤 사이트에 접속한 이용자간에 접속 순서가 존재한다면 이 연관규칙은 정보의 흐름이 어떻게 이어지는지를 파악하기 위한 자료가 된다. 또한 이러한 개선의 방향은 바로 접속해 오는 이용자들을 대상으로 이루어져야 하는만큼 발견한 규칙을 이용자마다 재구성하는 동적 웹 서버 체계를 구성하여 각 이용자가 필요로 하는 정보를 제공할 수 있는 데이터 마이닝 모듈과 결합하는 응용방안도 생각해 볼 수 있다. 이와 같이 이용자의 웹 사이트 접속 패턴을 데이터 마이닝을 통하여 분석함으로써 이용자를 위한 정보서비스 및 시스템의 개선과 효율적인 운영에 관련된 지식을 발견할 수 있다.

3.3 지능형 질의처리

1) 개념 및 관련 연구

지능형 질의 처리란 정보시스템이

이용자의 질의에 대하여 질문에 대한 단편적인 응답만을 제공하는 것이 아니라 보다 많은 정보 또는 이용자의 의도를 추측하여 이를 반영한 정보를 제공하는 기법이며 시스템이 보유하고 있는 지식을 활용하는 방법이 포함된다. 여기에는 확장질의, 비교질의, 제안질의 등의 구체적인 방법론이 연구되고 있으며 관련되는 분야로는 협조적 질의 응답(Cuppens & Demolombe, 1988) 등을 들 수 있다.

이러한 분야에 데이터 마이닝을 비롯한 지식탐사 기법을 응용하려는 연구가 시도되었으며 이들의 목적은 데이터베이스에 접근하는 이용자의 질의와 관련된 행동양식을 분석함으로써 보다 협조적인 모습을 보이는 시스템을 구축하려는 것이다(Han, et al. 1996). 예를 들어 "가장 빠른 서울행 비행기편은?"이라는 질문에 단순히 'CK103'이라는 응답 대신에 "CK103 입니다만 출발시간이 5분밖에 남지 않았습니다. 서두르십시오"와 같은 협조적이고 이용자의 의도를 파악한 응답을 제시하려는 시도이다.

이러한 시도는 질의에서 유용한 정보를 확보하여 데이터베이스의 표현양식인 튜플 형태로 표현한 후 이로부터 연관규칙을 추출하는 방식으로 이루어진다. 따라서 데이터베이스와 이용자간에 이루어지는 질의와 응답을 추상화하여 연관규칙을 추출하는 것이 지능형 질의처리의 주된 연구영역에 속한다. 먼저 질의패턴과 응답패턴에 대한 모델을 정의하고 이용자가 작성하는 SQL 문으로부터 질의패턴을 추출하게 된다. 이용자의 질의에는 이용시간, 이용자, 데이터베이스 속성집합, 조건 등이 포함되어 있으며 이로부터 질의패턴이 추출된다. 응답패턴은 데이터베이스가 반환하는 응답으로부터 추출되며 응답이 제출되는 시간, 질의를 요청한 이용자, 응답에 사용되는 속성, 이에 반환되는 속성값 등으로 구성된다. 보통 하나의 질의에 대하여 여러개의 응답패턴이 생성된다.

2) 지능형 질의처리 기법

일반적으로 지능적인 질의 처리에 사용되는 기법에는 이용자 모델링, 질의내 가정 처리, 내연 응답, 일반화 등의 기법이 다음과 같이 사용되고 있다.

(1) 이용자 모델링

이용자의 행동양식을 관리하여 시스템을 지능적이고 효율적으로 동작하도록 하는 시도에는 연속적인 질의응답에 대해 상황 전체를 모델화 하는 기법(discourse modeling)을 들 수 있다. 이러한 기법은 인공지능적인 기법이며 일반적으로는 이용자의 선호도(preference), 요구(need), 의도(intent) 등의 정보를 활용하여 이용자 모델을 구축한다.

(2) 질의내 가정 처리

이용자의 질의에는 대상 데이터베이스에 대한 가설이 전제된다. "우리 회사에서 빨간 자전거를 가지고 있는 구성원은?"이라는 질문에 단순히 "없

다"라는 응답보다는 "우리회사에는 자전거를 가지고 있는 사람이 전혀 없다"라는 응답이 보다 많은 정보를 전달하고 있다고 할 수 있다. 이러한 질의 처리를 위하여 연결관계를 이용한 기법이 제시되고 있다(Janas, 1981). 또한 이용자의 질의에 들어있는 가설에는 질문을 잘못 생각하는 오해가 있다. 예를 들어 "어떤 교수가 물리학개론 과목을 수강하는가?"와 같은 형태로 데이터베이스 스키마에 대한 잘못된 가정이 있을 수 있고, "고래의 아가미는 어느 부분에 있는가?"와 같이 데이터베이스 내의 튜플에 대한 잘못된 가정이 있을 수 있다. 이러한 오해에 대한 올바른 응답으로 단순히 "없다"라는 대답보다는 "교수는 강의를 하고, 학생은 수강한다"와 "고래는 아가미가 없고 포유류이기 때문에 폐호흡을 한다"라는 응답을 시도하게 된다.

(3) 내연 응답

내연 응답(intensional answer)이란 질의의 결과를 내연 데이터베이스를 이용하여 제공한다는 의미이다. 연역적인 데이터베이스(deductive database)에서는 데이터베이스를 외연(extension)과 내연(intension)으로 구분하는데, 외연은 데이터베이스에 그대로 저장된 형태의 데이터이고, 내연 데이터베이스는 관계형에서의 뷰 정의(view definition)와 무결성 제한(integrity constraint)과 같이 데이터베이스에 저장되어 있는 데이터를 한정짓고 관리하는 데에 필요한 지식이다. 따라서 내연 응답이란 데이터베이스의 스키마 정보만을 이용하여 질의를 수행하거나 질의의 결과를 스키마 정보를 이용하여 가공한 후 이용자에게 제공하는 것으로 정의할 수 있다. 예를 들어 "이번 보너스 지급대상은 누구인가?"라는 질문에 대하여 외연 데이터베이스에서는 모든 명단의 나열만이 가능하지만, 내연 데이터베이스를 이용한 응답에서는 "A공장과 B사무소 모두입니다"라는 유연한 대답이 가능하다.

(4) 일반화

지능형 질의 처리에서는 질의에 대한 일반화 기법이 포함된다. 이의 목적은 데이터베이스에 전달된 이용자의 질의를 관심분야를 고려하여 확장하는 경우와, 데이터베이스 내에 질의의 조건을 만족하는 데이터가 없을 경우 질의를 재구성하여 수행하는 방법이 있다. 질의의 재구성 방법에는 질의에 포함된 조건 중 일부를 삭제하거나 완화시키는 기법 등이 있다. 연역 데이터베이스에서는 내연 데이터베이스 내의 규칙을 이용하여 일반화된 질의를 생성할 수 있다.

3.4 텍스트 분석을 통한 색인어 추출

최근 디지털 문헌은 더욱 급속히 증가하고 있으며 이용자의 요구와 정보내용도 다양해 지고 있어 이러한 원문데이터에 대해서도 데이터 마이닝 기법의 적용이 필요다. 이러한 시도로서 원문 텍스트에 대한 데이터 마이닝을

통하여 색인어를 추출하려는 연구들(Helena Ahonen et al. 1998; Heikki Mannila, Hannu Toivonen 1996; Heikki Mannila 1996)이 수행되고 있다. 이를 통해 최근 원문 텍스트가 증가함에 따라 이용자들이 방대한 레코드에서 정보 내용의 특정한 단면을 찾기를 원하는 경향을 발견할 수 있다.

1) 출현 배경

정보검색에 있어 키워드와 단어구는 질의 과정의 출발점으로 사용되고 있다. 일반적인 정보검색 과정은 이용자가 정보시스템에 질의를 제공함으로써 정보요구를 표현하고, 시스템을 질의를 문헌과 비교함으로써 검색을 수행한다. 그런데 매번 단순히 텍스트를 스캐닝하는 작업은 불가능한 일이며, 따라서 문헌을 대표할 수 있는 키워드 집합이 선택되어 문헌에 첨부되어야 한다. 그러나 문헌의 양이 기하급수적으로 늘어남에 따라 이러한 목적을 위하여 만들어진 단일 키워드가 너무 많아지게 된다. 이러한 현상은 검색의 정확성을 떨어뜨리고 너무 많은 검색결과의 홍수를 초래하게 된다. 정보시스템에서의 많은 경험은 단일 키워드보다는 관련된 주제를 보다 정확히 표현할 수 있는 여러 단어로 구성된 색인 어구의 도입이 더 특정한 의미를 전달할 수 있음을 발견하였다. 따라서 단순한 키워드 선정에 그치지 않고 원문 텍스트로부터 의미있는 핵심 단어구를 발견할 수 있는 방법의 필요성이 대두되는데 여기에 데이터 마이닝을 적절하게 활용할 수 있고, 이러한 일련의 작업을 텍스트 마이닝이라고 말하기도 한다.

2) 텍스트 마이닝 과정

정보검색 분야에서는 색인과 키워드의 선택 방법론이 많이 연구되었지만, 텍스트 마이닝을 통한 적용에서는 두 개 이상의 단어로 이루어진 핵심 단어구(key phrase)의 선정을 그 목적으로 한다. 이러한 복합키 용어(compound key term)은 텍스트 검색의 품질을 개선할 수 있는 핵심적인 역할을 수행한다. 또한 이들 핵심 단어구의 정확한 형태와 선택을 유도하기 위해 각 성분에 대한 가중치를 결정하기 위해서는 많은 실험이 필요하다.

텍스트는 순차적인 데이터로 간주할 수 있으며 이는 다른 시간적 기록을 수행하는 관찰시스템에 의하여 얻어진 데이터와 많은 측면에서 비슷하다. 일반적인 지식탐사 프로세스에 따르면 시작점은 원문 데이터가 되며 최종 산출물은 데이터의 빈도 현상을 설명해 주는 정보이다. 이를 위하여 연관규칙은 사건과 사건규칙이라는 용어로 표현할 수 있다. 이러한 과정을 단순화하면 사전처리과정, 텍스트 마이닝 과정, 사후처리 과정으로 요약할 수 있는데 특히 사전처리과정은 시스템의 효율성에 결정적인 영향을 주고 있으며, 전체 노력의 80%를 차지하고 있다. 텍스트 데이터의 사전처리과정에는 특별한 측면이 존재하고 있다. 텍스트는 단어와 특수문자와 구조적

정보로 이루어져 있으며, 사전처리과정은 이들을 처리하는데 있어 결과의 이용의도에 많은 부분을 의존하고 있다. 대체로 특수문자와 구조적 정보(HTML, SGML 등의 태그)를 부호로 치환함으로써 표준화한다. 이중 일부는 완전히 제거되기도 하고 특수한 경우 가공을 통하여 변환되기도 한다. 또한 사전처리과정에서는 자연언어 분석을 수행하기도 한다. 이때 미리 마련된 어형분석 프로그램을 통하여 문헌 텍스트를 일반화 하는 방법으로 자연어의 문형을 표준화하게 된다. 만약 발견하고자 하는 규칙에 대한 명확한 개념이 정의되지 않은 단계라면 이후 과정에서 다소 시간이 소요되더라도 사전처리과정에서 과다한 부하를 부여하지 않고 사후처리과정에서 수행하도록 데이터 정리 작업을 알맞게 수행하는 것이 중요하다. 그러나 방대한 문헌을 다룰 때는 효율성이 참작되어야 하며 검색공간의 크기를 제한하고 줄이기 위하여 사전처리과정에서 제거가 수행되어야 한다.

데이터 마이닝 과정에서 핵심적인 것은 연관규칙의 변형인 사건(episode)과 사건규칙(episode rule)의 적용이다. 텍스트는 유사한 데이터의 순차집합으로 볼 수 있고 이에 따라 속성과 색인으로 구성된 사건을 작성하게 된다. 사건 규칙의 작성에는 지지도(support)와 신뢰도(confidence)가 기본적으로 적용되며 이들은 일정 크기 이상의 확률을 초과하는 경우에만 선택되게 된다.

데이터 마이닝 결과의 사후처리과정에는 보다 많은 요소를 도입하고 있다. 여기에는 최대 윈도우 크기에 대한 사건규칙의 절대 길이를 반영한 길이(length)가 있으며 사건규칙의 양변을 연산하여 나눈 견고성(toughness) 지수가 있을 수 있다. 또한 상호 신뢰도를 계산하기 위하여 각 변의 지지도와 전체 지지도의 비율을 합하여 나누어주는 방법을 이용한다. 또한 문헌을 구분하게 하는 특정한 핵심어구를 발견하기 위해서는 문헌집합 전체에서의 단어구 분포도를 고려해야 한다. 이를 위해 문헌내 각각의 사건규칙에 대하여 문헌집합에서 이 규칙이 얼마나 보편적인지를 설명해 줄 수 있는 역문헌빈도(inverse document frequency, IDF)를 계산하게 된다. 이들 모두를 종합하여 최종적인 사건규칙의 가중치가 산출되게 되고 사후처리과정 결과로 이들 가중치의 순서대로 사건집합이 기술되는 것이다. 이러한 과정이 모두 완료되면 큰 문헌집합에 대해서 변별력을 갖춘 색인어구를 발굴할 수 있다.

4. 결론

정보관리분야에도 데이터 마이닝을 적용해야 할 필요성은 점점 높아지고 있으며, 현재 인터넷 로그데이터 등 디지털문헌을 대상으로 한 다양한 연구가 시도되고 있다. 정보센터나 인터넷서비스 등 상업적인 정보서비스기관에서도 변화하는 정보환경과 이용자

요구에 신속하게 대응할 수 있는 분석 도구의 하나로 데이터 마이닝을 활용함으로서 전략적인 정보정책의 수립이나 이용자서비스 개선 등을 도모할 수 있다.

데이터마이닝은 현재와 같이 대량의 디지털 데이터를 다루어야 하는 요구에 의하여 개발되었으며, 이에 따라 숨겨져 있는 새로운 지식을 발견하는 핵심적인 기법으로 활용될 수 있다. 지속적인 부가가치의 창출을 위한 효율적이고 경제적인 정보시스템의 구축과 정보서비스를 추구하는 정보관리 분야에 있어 데이터 마이닝의 활용 잠재력은 매우 크다고 보며, 새로운 적용 영역의 개발 및 다양한 연구의 수행이 필요하다고 하겠다.

참고문헌

- 강성희, 박승수. 1998. "데이터 마이닝 환경에서 다중 에이전트 기반 지식 검증 및 통합 방법," 98 한국전문가시스템학회 추계학술대회 : 75-80
- 김성민 외. 1998. "관계형 데이터베이스에서의 지식 탐사를 위한 개념 계층의 제어기법," 98 한국전문가시스템학회 추계학술대회 : 83-90
- 김신곤. 1997. "데이터 마이닝과 지식 발견," 전문가시스템학회 1997 춘계학술대회 : 239-248.
- 나민영. 1997. "2000년대의 DB 응용 기술 데이터 마이닝," *데이터베이스 월드* 1997.9 http://www.dpc.or.kr/dbworld/document/9709.spec.html.
- 남도원 외. 1998. "웹 로그에서의 사용자 접근 패턴 분석," http://iislab.postech.ac.kr/dongha/Publish/9802hciCirene.ps
- 양경식, 김현수. 1998. "판매 및 재무 데이터에 대한 데이터 마이닝 기법 적용연구," 98 한국전문가시스템학회 추계학술대회 : 105-112
- 이동하외.1998. "관계형 데이터베이스에서의 연관규칙 탐사를 위한 객체 일반화 트리의 구성," 98 한국전문가시스템학회 추계학술대회 : 91-96
- 이동하, 이전영. 1997. "지능 질의 처리를 위한 지식탐사 시스템의 설계," http://iislab.postech.ac.kr/dongha/Publish/9702sigdbCdongha.ps
- 이희석, 장재경. 1997. "데이터 마이닝과 OLAP," *컴퓨터월드* : 132-141.
- 조성원. 1998. "컨설팅 방법론에 근거한 데이터 마이닝 접근 방법 및 사례 연구," 98 한국전문가시스템학회 추계학술대회 : 130-138
- 조재희, 박성진. 1996. 데이터웨어하우스와 OLAP, 서울 : 대청
- Adrianns, Pieter & Zantinge, Dolf. 1996. *Data Mining*, New York : Addison-Wesley.
- Agrawal, R., Imielinski, T., Swami, A. 1993. "Mining association rules between sets of items in large database," *Proceedings of the ACM SIGMOD Conference on Management of Data* : 207-216.
- Agrawal, R., et. al. 1995. "Mining Sequential Patterns," *Proceedings of the 11th International Conference Data Engineering* : 3-14.
- Ahonen et al. 1998. "Applying Data Mining Techniques fo Descriptive Phrase Extraction in Digital Document Collections," *'98 Advances in Digital Library Conference.*
- Berry, Michel J. A., Linoff, Gordon. 1997. *Data Mining Technique*, John Wiley & Sons.
- Bigus, Joseph P. 1996. *Data Mining with*

Neural Networks, McGraw Hill.
- Cuppens, F., Demolombe, R. 1998. "Cooperative answering : a methodology to provide Intelligent Access to Databases," *Proceedings of 2nd International Conference Expert Database Systems* : 621-643.
- Davies, Roy. 1989. "The Creation of New Knowledge by Information Retrieval and Classification," *Journal of Documentation,* 45(4) : 273-301
- Digital Equipment Corporation(DEC). 1997. *Data Mining*, http://www.digital.com/datamine.
- Dilly, Ruth. 1995. *Data Mining An Introduction Student Notes*, The Queen' s University of Belfast, Version 2.0, http://www.qub.ac.uk.
- Edelstein, Herb. 1996. "Mining Data Warehouses," *Information Week*, (561) : 48-54.
- Etzioni, Oren, 1996. "The World-Wide Web : Quagmire or Gold Mine?" *Communications of the ACM*, 39(11) : 65-68.
- Fayyad, Usama M., et al. 1996. *Advances in Knowledge Discovery and Data Mining*, AAAI Press.
- Fayyad, Usama M. 1996. "Data Mining and Knowledge Discovery : Making Sense Out of Data," *IEEE Expert,* 11(5) : 20-25
- Fayyad, Usama M., Uthurusamy, Ramasamy. 1996. "Data Mining and Knowledge Discovery in Database," *Communications of the ACM,* 39(11) : 24-26
- Friedman, Jerome H. et al. 1997. "Data Mining and Statistics : What' s the Connection?,"
- Gray, Jim. 1996. "Evolution of Data Management," *IEEE Computer,* Oct 1996, 29(10) : 38-46
- Han, J., Huang, Y., Cercone, N., Fu, Y. 1996. "Intelligent Query Answering by Knowledge Discovery Techniques," *IEEE Trans. Knowledge and Data Engineering,* 8(3) : 373-390
- Information Discovery, Incorporated (IDI) http://www.datamining.com
- Janas, J. M. 1981. "On the Feasibility of Informative answers," *Advances in Database Theory,* Vol.1 : 397-414, New York : Plenum Press
- KD Nuggets. 1996. "Data Mining and Knowledge Discovery References," http://www.kdnuggets.com/references.html
- Lesk, Michael. 1996. "The Seven Ages of Information Retrieval," http://community.bellcore.com/lesk/ages/ages.html
- Luotonen, A. 1995. 'The common log file format," http://www.w3.org
- Mannila, Heikki. 1996. Data Mining: Machine Learning, Statistics, and Databases, *Proceedings of the 8th International Conference on Scientific and Statistical Database Management* : 1-6.
- Mannila Heikki, Toivonen Hannu. 1996. "Discovering Generalized Episodes Using Minimal Occurences," In Evangelos Simoudis, Jiawei han, and Usama Fayyad, editors, *Proceedings of the 2nd international Conference on Knowledge Discovery and Data mining* : 146-151.
- Park, S., Kang, S. 1998. "Agent-based Knowledge Management in Data Mining," *Proceedings of CAINE-98*
- Pilot Software Incorporated. 1996. "The Scope of Data Mining," http://www.pilotsw.com/dmpaper/dmscop.htm
- SAS. 1998. "데이터 마이닝 솔루션," http://www.sas.com/offices/asiapacific/korea/solution/mining/wp/mining-w_wp3.html
- Weiss, Sholom M., Indurkhya, Nitin 1997. *Predictive Data Mining a Practical Guide*, Morgan Kaufmann Publishers, Inc.