데이터마이닝 분야의 연구, 개발 및 활용

백성욱, 책임 연구원

목 차

요약문

이 보고서는 데이터마이닝의 필요성과 중요성에 대해 강조하며 국내 데이터마이닝의 연구 및 개발의 촉진을 위하여 그리고 국내 데이터마이닝의 다양한 활용을 위해 쓰여졌다. 데이터마이닝은 통계학, 데이터베이스, 패턴인식, 기계학습, 데이터 비주얼라이제이션, 그리고 고성능 계산 기법 등과 같은 여러 학문들이 관여하는 분야이므로 그들 각각의 연구와 개발뿐만 아니라 그들 학문들의 효과적인 협력이 매우 중요하다. 그리고 데이터마이닝의 새로운 활용 분야는 점점 커지고 있다. 예를 들면 인터넷의 빠른 보급으로 전자 상거래와 같은 이비즈니스(E-business)의 발달로 인하여 데이터의 온라인 분석과 같은 새로운 요구가 있으며 바이오 인포메틱스 같은 분야에서 방대한 유전자 정보의 분석을 필요로 한다. 또한 지질학, 해양학, 기상학과 같은 수많은 관측된 데이터들을 이용하는 학문에서도 최근에 데이터마이닝이 매우 중요한 기술로 등장했다. IT 선진국인 미국은 과거의 데이터마이닝과 관련된 여러 학문들을 바탕으로 활발한 데이터마이닝 연구를 하고 있으며 여러 데이터마이닝 소프트웨어들을 이미 개발하여 상품화시켜 미국 시장은 물론 한국을 비롯한 아시아 국가들의 시장을 대부분 점유하고 있다. 그러나 여러 다양한 분야의 문제들을 해결하기 위한 유일한 데이터마이닝 솔루션은 없기 때문에 국내 산업 환경에 익숙한 국내 데이터마이닝 전문가와 과학자들에 의해 국내 특성에 맞는 데이터마이닝 연구와 개발이 이루어 질 수 있도록 정부와 기업들의 후원이 필요하다.

1. 개요

노벨상 수상자인 펜지어스 박사는 (Dr. Penzias) 1999년 1월에 컴퓨터 월드 (Computer World) 잡지사와의 인터뷰에서 "데이터 마이닝은 앞으로 기업들에게 매우 중요해질 것이다. 그리고 그 기업들이 관리하고 있는 고객들의 모든 정보는 너무나도 귀중해서 빠짐없이 보관되어질 것이며, 만약에 지금 어떤 기업이 데이터 마이닝과 그것을 위한 준비 작업을 하고 있지 않다면 그 기업은 반드시 도산될 것이다". 라고 했다 [Groth, 2000]. 현대는 더 많은 고객들을 확보하려는 기업들은 과대한 경쟁 하에 더욱 더 다양화되고 개성화 되고 있는 고객들의 요구에 대해 적절하고 빠른 대응을 해야한다. 그러기 위해서는 신속하고 정확한 마케팅 전략과 여러 가지 상황에 대한 적절한 의사결정을 위한 의미 있는 고급 정보 혹은 지식 (Knowledge) 들을 필요로 할 수밖에 없다.

데이터마이닝의 정의를 간단히 정리하면 다음과 같다. 데이터웨어하우스(Data Warehouse)나 데이터마트(DataMart)안에 저장되어 있는 대량의 데이터는 분석 없이 그 자체만으로는 의미가 없다. 그 이유는 그 데이터의 용량이 너무 방대하고 복잡해서 보통 사람들이 그 데이터를 이해하기란 매우 힘들기 때문이다. 데이터 분석가들도 그들의 필요한 부분만을 선택하여 심층 분석을 한다. 그러므로 그 이외의 데이터들은 저장이 된 후에 한번도 사용을 안 하게 되는 경우가 흔히 발생하게 되고, 그 사용하지 않은 데이터들로부터의 의미 있는 정보나 지식 등을 놓치는 경우가 발생 하게된다. 데이터마이닝은 모든 데이터로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 과정을 말한다. 여기서 유용한 정보란 사람이 이해할 수 있고 활용할 수 있는 지식의 형태를 말한다. 이 형태의 예들은 유용한 패턴, 새로운 법칙, 앞으로 예상 할 수 있는 추세 등이며 이를 실제 경영의 의사결정이나 마케팅 전략 등에 활용할 수 있다. 이러한 지식을 유출해내는 방법은 어떤 특정 기법과 그 기술 자체만을 의미하는 것이 아니고, 비즈니스 문제나 과학용 데이터들을 분석하는 문제 등에서 주어진 상황을 이해하고 그 특정한 문제를 해결하기 위하여 여러 학문 분야의 방법들을 적용하는 포괄적인 과정을 의미한다. 다시 말해 데이터마이닝은 통계학, 데이터베이스, 패턴인식 (pattern recognition), 기계학습 (machine Learning), 데이터 비주얼라이제이션 (data visualization), 그리고 고성능 계산기법 (high-performance computing) 등과 같은 여러 학문들이 관여하는 분야이다.

인터넷과 분산 데이터베이스의 발달들로 인해 온라인 데이터들의 급증은 물론 방대한 과학용 데이터들의 분석의 필요성으로 말미암아 데이터마이닝의 필요성은 더욱 더 절실해졌다. 이미 데이터마이닝은 비즈니스 분야에서 CRM (Customer Relationship Management; 고객 관계 관리 시스템)의 핵심적인 기술로 사용되고 있으며 또한 바이오테크놀로지 (예를 들면 바이오인포메틱 (bio-informatics))의 발달로 인해 그 분야의 방대한 데이터를 분석하는 기술로 부각되었다. 인터내셔널 데이터의 산업 분석가 헨리 모리스는 데이터마이닝과 관련된 어플리케이션의 전 세계의 매출량은 1999년에 3억 4천 3백만 달러였고 앞으로 매년 32 퍼센트씩 증가하여 2004년에는 14억 달러에 육박할 것이라고 예측했다 [Cahlink, 2000].

2. 데이터마이닝의 태스크 (Data Mining Tasks)

데이터 마이닝은 여러 가지 목적을 위해 수행하기 위해 다양한 방법들을 포함한다. 데이터마이닝의 이해를 돕기 위해 그 방법들을 이 장에서 간단히 소개한다 [Goebel, 1999; Fayyad, 1998].

데이터 프로세싱 (Data Processing): 유용한 패턴과 같은 지식들을 추출하기 전에 데이터를 선택하거나 여과 (filter), 통합(aggregate), 표집 (sample), Noise 제거 (clean), 변형 (transform) 등과 같은 과정을 통해 업무량을 줄이거나 분석의 효과를 높인다.

예측 (Prediction): 주어진 데이터나 예측 모형을 통해 아직 알려지지 않은 데이터를 예측한다.

회귀 분석 (Regression): 주어진 데이터 세트에서 어떤 한 데이터의 속성이 다른 속성들과의 의존관계를 분석하는 것이다.

분류 (Classification): 이미 정의한 분류 클레스들(class)로 특정한 데이터들의 클래스를 결정한다.

군집 분석 (Clustering): 데이터들을 비슷한 속성을 가진 것들끼리 모아 묶는다.

요약 (Summarization): 한 그룹으로 모여있는 데이터들을 간단 명료한 설명을 찾아낸다.

편차 발견 (Deviation Detection): 데이터 중에서 다른 데이터와 특별하게 구별되는 비정상적인 데이터를 발견한다.

의존성 모델링 (Dependency Modeling): 데이터의 변수들간의 중요한 의존성을 설명하는 모델을 발견한다.

연관성 측정 (Association): 데이터 세트로부터 데이터간의 연관성 정도를 측정하여 연관성이 많은 데이터들을 그룹화 하는 clustering의 일종이다.

모델 비주얼라이제이션 (Model Visualization): 방대한 데이터로부터 지식이나 모델이 유출되었어도 그들이 여전히 이해하기 어려울 때가 있다. 그때 지식이나 모델들을 사람이 직관적으로 이해하기 쉽게 이차원이나 삼차원공간에서 그래픽으로 그 모델들을 나타내 준다.

탐색적 데이터 분석 (Exploratory Data Analysis):사전의 설정된 가정이나 모델에 전적으로 의존하지 않고 데이터의 그래픽 표현의 도움으로 인간의 직관에 의해 흥미 있는 패턴들을 찾는다.

3. 데이터마이닝의 방법론의 소개

오늘날 SAS의 E-Miner, SPSS의 Clementine 그리고 IBM의 Intelligent Miner 등과 같은 선두주자격인 데이터마이닝의 툴들은 여러 다양하고 복잡한 문제들을 더욱더 잘 해결하기위해 여러 다양한 데이터마이닝 알고리즘들을 사용하고 있다. 그 알고리즘들은 데이터마이닝에 적용될 때 각자 장점과 단점들을 갖고 있기 때문에 그들의 효과적인 선택은 무엇보다 중요하다. 이 장에서는 대표적인 데이터마이닝의 방법론들을 간단히 소개한다 [Groth, 2000].

의사결정나무 (Decision Trees)

Decision Tree 는 추출된 모델이나 지식들을 아주 잘 표현할 수 있는 장점이 있기 때문에 데이터마이닝 의 방법론 중에 가장 많이 활용되는 방법중 하나이다. 반면에 데이터가 복잡해 질 때 나무구조가 복잡해지는 점이 단점으로 지적된다. Decision Tree 의 대표적인 알고리즘은 Hunt 의 Concept Learning Systems을 기반으로 하는 Quinlan 의 ID3 와 ID3 알고리즘을 발전 시킨 C4.5, 그리고 통계학자들에 의해 개발된 CART 와 CHAID (Chi-Square Automatic Interaction Detection) 등이 있다.

신경망모형 (Neural Networks)

Rosenblastt 와 Widrow 등 몇몇이 패턴 인식 분야에서 신경망 유니트를 기반으로 하는 분류기(Classifier)들을 개발한 이후에 Nillson 이 인공 지능 기법을 그 패턴 인식분야에 접목시킴으로

지금의 학습형 신경망모형 같은 형태의 개발이 시작됨으로 신경망모형이 데이터마이닝과 유사한 분야에 최초로 적용되었다. 신경망모형은 인간이 경험으로부터 학습해 가는 두뇌의 신경망 활동을 흉내내어 자신이 가진 데이터로부터의 반복적인 학습 과정을 통해 패턴을 찾아내고 이를 일반화 한다. 복잡한 문제에서도 정확한 결과를 주고 신경망 모형의 구조가 매우 유연한 장점이 있으나 설명력이 부족하여 종종 블랙박스 (Black box)로 불리는 큰 약점이 있다. 신경망모형의 앞으로 데이터마이닝을 위한 연구 계획의 방향은 크게 두 가지로 분류된다 [Fu, 1999]. 첫 번째는 지식 (Knowledge) 추출을 위한 더욱 효과적인 알고리즘 개발이고 두 번째는 추출된 지식들을 신뢰성 있게 해독하여 설명력 있게 제공해 주는 것이다. 이런 기능을 개발한 대표적인 예는 CFN (Certainty Factor Network) [Fu,1998] 이다.

Bayesian Belief Networks

Bayesian Belief Networks 는 자체 연구 규모에 비해 데이터마이닝에서는 신경망 모형이나 의사결정나무에 비해 활발히 사용되고 있지는 않다. 그러나 최근 데이터마이닝 분야에 응용되기 위해 연구의 방향이 변해야 한다는 주장이 나오고 있다. 특히 마이크로소프트는 이 분야에 집중적인 투자를 하고 있으며 윈도우 기반인 Belief Network Tool을 이미 가지고 있다. 1763년에 Reverend Thomas Bayes에 의해 이론이 나왔으며 쉽게 이해되고 결과를 잘 예측한다는 장점이 있다.

Genetic Algorithms

Genetic Algorithm 은 데이터마이닝에서 데이터들의 가장 효율적인 모델을 찾아내는데 사용되며 종종 신경망모형과 결합하여 데이터들을 모델화 하는데 사용된다. Genetic Algorithm의 근원은 1859년 찰스 다아윈 (Charles Darwin)의 종의 기원 (Origin of Species)으로부터 시작하며 1957년에 G.E.P. Box에 의해 쓰여진 Evolutionary operation:a method of increasing industrial productivity 가 Genetic Algorithms 와 비즈니스 문제를 연결시키는데 큰 공헌을 했다.

통계학 방법 (Statistical Methods)

통계학 방법 중에 대표적인 것들은 On Discriminant Analysis 와 On Regression Modeling 등이 있다. 통계학 방법들은 결과가 정확하며 잘 이해되는 장점이 있는 반면 데이터마이닝 분야에서 효과적으로 사용하는데 어려움이 있다는 것이 단점으로 지적된다.

기타 다른 방법들

위에 설명한 데이터마이닝 방법론 이외에도 Fuzzy Sets, Rough Sets, Case-Based Reasoning, 그리고 Michalski의 AQ와 같은 Rule Induction 방법이 있다.

4. 데이터마이닝의 활용분야

Information Technology 기술이 점점 발달함에 따라 데이터마이닝 의 활용 분야는 점점 커지고 있다. 이들 새로운 활용 분야들은 다음과 같이 크게 세 가지로 나누어 질 수 있다 [Cahlink, 2000].

a. Business & E-commerce Data: 백 오피스, 프런트 오피스 와 기타 네트워크 소프트웨어들을 통해 비즈니스 처리 과정에 대한 많은 양의 데이터들이 생성된다. 이들 데이터를 통한 효과적인 의사 결정을 위해 데이터마이닝 이 사용된다.

Business Transactions: IT의 발달로 인해 비즈니스의 고객들의 수와 그들의 정보를 처리하는 수효는 기하급수적으로 늘어나고 있다. 고객들의 방대한 정보를 통해 마케팅의 전략 등과 같은 비

즈니스에 필요한 의사결정에 중대한 역할을 하게 된다. 한 예로서 고객 관계 관리 (Customer Relationship Management)를 들 수 있다.

Electronic Commerce: E-commerce에서는 마케팅 패턴이나 리스크 패턴 (risk pattern)들의 분석들을 위한 방대한 데이터를 다루는 문제도 중요하지만 고객들의 온라인 요청에 신속히 대처하기 위해 데이터를 실시간에 처리해야 하는 문제들이 대두해 있다.

b. Scientific, Engineering & Health Care Data: 과학용 데이터는 비즈니스 데이터 보다 훨씬 복잡하고 방대하다. 요즘 여러 분야 과학자들은 이런 데이터들의 효과적인 사용을 필요로 하고 있다.

Sensor Data: 인공위성, 항공기, 과학용 풍선 등 다양한 관측기기로부터 지구의 대기 환경, 해양, 지질 등에 관한 방대한 데이터들이 생성되고 있다. 이들 데이터로부터 각 데이터들간의 관계성을 발견하는 것은 매우 중요하다. 예를 들면 산업 공해가 지구 온난화와 관계가 있는 가를 알아보기 위해 여러 관측기로부터 얻은 데이터들을 사용할 수 있다.

Simulation Data: Simulation 작업은 이론과 실험을 뒷받침 해주는 아주 중요한 분야로 부각되었다. 오늘날, 실험뿐만 아니라 Simulation 통해서도 방대한 데이터들이 만들어지고 있다. 그러므로 데이터마이닝 은 이론, 실험과 Simulation 사이를 연결해 주는 중요한 기능이 될 수 있다.

Health Care Data: 병원, 보건 센터 (health care organization), 정부 당국 (예를 들면 미 연방 정부)에서는 환자들에 대한 모든 정보 와 그들을 처방한 방법과 과정들 그리고 그 처방 결과들과 같은 수많은 데이터들을 보관하고 있다. 이 데이터들의 관계성을 이해하는 것은 폭 넓고 다양한 문제들을 해결하는데 큰 도움이 된다.

Genomic Data: 유전자 인식과 분석을 위해 유전자에 대한 정보가 방대하게 데이터 베이스로 구축되어있고 각종 질병, 세포의 기능 그리고 다양한 의약들의 반응 등의 정보 또한 온라인 데이터베이스로 저장되어 있는 것이 지금의 현실이다. 그러기에 그 방대한 데이터들간의 아직 밝혀지지 않은 관계성을 찾아내는 작업이 데이터마이닝으로 가능하다.

c. Wed Data: 웹을 통한 데이터들은 점점 복잡해지며 방대해 지고 있다. 이들 웹 데이터들은 텍스트와 이미지와 음성과 같은 멀티미디어 데이터들을 포함한다.

Multi-media Documents: 디지털 라이브러리나 웹상의 오디오 비디오, 이미지 등과 같은 멀티미디어 데이터들의 양은 빠른 속도로 증가하고 있다. 그러나 여러 다른 사이트에 저장되어 있는 멀티미디어 데이터들로부터 의미 있는 정보나 지식 등을 추출하기는 아직 어려운 실정이다. (그들은 텍스트 데이터에 비해 분석하기가 매우 어렵기 때문이다.)

The Data Web: XML(Extensible Markup Language) 는 웹 환경에서의 차세대 언어이다. 그 언어의 특징중의 하나는 네트워크로 연결되어있는 여러 환경에 있는 데이터를 함께 다루는 점이다. 이런 인프라가 점점 형성이 되면 데이터마이닝 은 더욱 더 중요한 기술이 될 것이다.

5. 데이터마이닝의 대표적인 연구

개요에서도 기술했듯이 데이터마이닝은 여러 학문이 관련된 분야이므로 기존에 수행되어왔던 연구기반을 토대로 새로운 연구 분야에 적용되는 경향이 있다. 이 장에서는 데이터마이닝이 활발

히 수행되고 있는 대학이나 연구소들과 그들의 연구 내용을 소개한다.

5.1. 카네기 멜론 대학교

카네기 멜론 대학교은 전 세계에서 데이터마이닝의 연구를 제일 활발하게 하는 대학중 하나이다. 오래 전부터 Machine Learning 같은 인공 지능의 기초적인 연구가 활발했고 그 기초학문을 토대로 데이터마이닝이라는 새로운 추세 속에 여러 가지 분야에 두각을 나타내고 있고 Automated Learning and Discovery 센터에서는 데이터마이닝에 대한 다양한 연구를 하고 있다. 대표적인 연구 프로젝트들은 다음과 같다.

a. Mining the World Wide Web: Machine Learning 의 대표적인 방법인 Version Space Method 의 창시자이며 "Machine Learning" 의 공동 저자인 Tom Mitchell 이 주도하는 연구 프로젝트이다. 이 프로젝트의 목적은 WWW (World Wide Web) 이나 뉴스 정보들 혹은 다른 인터넷 소스 (Internet Source)로 부터 유용한 지식이나 사실들을 추출하여 방대하면서 구조적인 데이터 베이스를 구축하는 것이다. 이 프로젝트가 성공적으로 수행하게 되면 우리는 WWW이 마치 한 데이터 베이스처럼 Database Query Language 나 연역적 추론 방법으로 데이터 베이스를 다룰 수 있게 된다.

b. Searching by Content and Data Mining in Multimedia Databases: 이 프로젝트는 멀티미디어 데이터 베이스로부터 내용을 찾아내는 것과 방대한 데이터 베이스를 위한 새로운 데이터마이닝 방법들을 개발하는 것이다.

c. Data Mining for Manufacturing: 이 프로젝트의 목적은 산업 현장에서 제품 결함 발견, 생산 스케줄링 (production scheduler), 공정 과정 조정기 (Process controller) 등과 같은 자동 결정 시스템 (autonomous decision-making systems)을 위한 기초적인 알고리즘을 연구하는 것이다.

d. Visualization of Massive Datasets: 이 프로젝트는 통계학자, 인지 심리학자 (cognitive psychologists), 컴퓨터 학자 등 여러 분야에서의 전문가들이 협력하여 수행하는 연구이다. 방대한 데이터 세트를 모두 처리하는 것은 시간이 매우 많이 걸리는 것뿐 아니라 표준 계산 방법으로는 적절치 못하므로 비주얼라이제이션을 통해 거대한 데이터 세트들을 직관적으로 분석할 수 있도록 하는 것이 이 프로젝트의 목적이다.

5.2 IBM

IBM은 데이터마이닝 에 대한 연구 초창기부터 주도적인 역할을 담당해 왔다. 이미 IBM은 데이터마이닝 툴인 Intelligent Miner for Data 와 Intelligent Miner for Text 들을 상용화시켰다. Intelligent Miner for Data 는 여러 가지 문제들을 해결하기 위해 여러 다양한 데이터마이닝 알고리즘을 사용하고 있으며 Intelligent Miner for Text 는 온라인 뉴스 서비스, 이 메일, 웹 등과 같은 텍스트 소스로부터 정보를 해석할 수 있는 기능을 제공하고 있다. 그리고 IBM T.J. Watson 연구소에서는 계속해서 여러 데이터마이닝 에 대한 연구를 하고 있다. 대표적인 연구 분야들은 다음과 같다.

a. BioInformatics & Pattern Discovery: IBM 의 Computational Biology Center에서 수행하고 있다. 분자 생물학(molecular biology)으로부터 문제들을 해결하기 위해 여러 계산기술 (computational techniques) 들을 사용하고 있다.

b. Data Mining & Decision Support Technologies: IBM 의 Almaden Research Center에서 수행하고 있다. 이 연구의 목적은 방대한 데이터 베이스로부터 유용한 패턴을 발견하는 데 목적이 있

다. 사용하는 테크놀로지는 연관성 측정 법칙 (association rules), 순차적 패턴 (sequential patterns), 분류 (classification),와 시간 변동 데이터 군집 분석 (time-series clustering) 등이고 이 기술들은 Intelligent Miner for Data을 향상시키는 데 사용된다.

c. Image Information Systems: IBM 의 Networked Data Systems department에서 수행하고 있다. 이 연구의 관심은 다음 분야 (주식 변동, 이자율 같은 시간 변동 데이터, 사진 이미지들과 비디오, 인공 위성 이미지) 들을 위해 진보적인 content-based query 와 데이터마이닝 프레임워크 들을 개발하는 것이다.

5.3 마이크로 소프트

마이크로 소프트 리서치의 Machine Learning and Applied Statistics 그룹에서 데이터마이닝 에 대한 연구를 하고 있다. 그들 연구의 초점은 크게 두 가지로 나뉜다. 첫 번째는 자동으로 학습할 수 있는 소프트웨어를 만드는 것이다. 다시 말해 소프트웨어가 스스로 모델들을 구축하며, 패턴들을 추출하며, 변하는 환경에 적응 하는 것을 의미한다. 두 번째는 데이터 분석가들이 데이터를 아주 쉽게 이해하고 탐색할 수 있도록 도와주는 툴을 만드는 것이다.

그들의 관심은 기계학습기법과 통계학적인 기법을 통해 빠르고 대량의 데이터들을 처리할 수 있는 알고리즘을 개발하는 것이며 다른 소프트웨어와 차별을 두기 위해 예측 (prediction), 데이터 분석, 데이터마이닝 등에 대한 새로운 테크닉들의 개발하는 것이다. 또한 데이터마이닝의 폭넓은 응용을 위해 자연언어처리 (natural language processing) 와 음성인식 (speech recognition), 필기체 인식 (handwriting recognition) 분야 등의 알고리즘 개발은 물론 데이터 비쥬얼라이징을 위한 새로운 파라다임을 제시하고 있다.

5.4 California Institute of Technology

칼텍 (California Institute of Technology) 의 Jet Propulsion 연구소에 소속된 Machine Learning Systems (MLS) 그룹에서는 주로 NASA 와 JPL instruments 들로부터 제공된 과학용 데이터 (scientific data) 들의 자동 분석 (automated analysis) 에 중점을 두고 있다. 그들은 데이터마이닝, 패턴 인식, 그리고 자동 분류 및 군집 분석 (Automated Classification and clustering) 등과 같은 테크닉을 이용하여 학습 능력을 갖춘 시스템들을 개발하는데 지적되는 어려운 문제들을 해결하는데 중점을 두고있다. 그들의 데이터마이닝에 관련된 연구는 다음과 같다.

DiamondEye : 이 테스크의 목적은 차세대 이미지 데이터마이닝 시스템을 위한 알고리즘과 인프라스트럭쳐를 개발하는데 있다. 이 시스템은 과학자들이 로컬화 되었거나 분산화 된 이미지 데이터로부터 그들이 원하는 지질학적인 특징들 발견할 수 있도록 해 준다.

Neural Net Data Mining: 이 프로젝트에서는 신경망모형 (neural network)의 기술을 이용하여 시간에 따라 변하는 데이터들 안에서 미묘하지만 아주 중요한 변화들을 발견하는 방법론을 개발하고 있다. 예를 들면 매일 해양 표면의 온도를 측정한 데이터들로부터 엘리뇨 현상 같은 특징적인 변화들을 발견하는 것이다.

Temporal Data Mining: 이 태스크의 목적은 원격 센서들로부터 얻은 이미지 (remote sensing imagery) 들로부터 중요하고 유동적인 과학적 사건들이나 과정들을 자동적으로 찾아내는 알고리즘과 소프트웨어를 개발하는 데 있다. 이런 알고리즘은 과학자들이 우주선의 센서로부터 얻은 고 화질의 이미지로부터 중요한 행성들의 사건들을 소프트웨어와의 상호작용을 통해 발견할 수 있도록

해준다.

Oceans and Atmospheres Data Mining: 이 태스크의 목적은 여러 다양한 지구를 관찰하는 우주선과 지상에 있는 센서들에 의해 얻어진 방대한 공간과 시간상(spatio-temporal)의 데이터들을 과학자들이 이해하는데 도움을 주는 기계학습과 데이터마이닝 테크닉들을 적용하는 것이다.

6. 국내의 데이터마이닝

6.1 국내 현실과 문제점

국내에서는 특히 올해(2000년)에 고객관계관리(customer relationship management)에 붐을 입어 데이터마이닝이 각광을 받기 시작했다. 그러나, 국내에서 아직 데이터마이닝의 성공을 좌우하는 사회적인 인프라가 국내 IT 업계에 전반적으로 형성이 안되어 있으며 데이터마이닝의 자체 연구 및 개발(R&D)이 미비한 상태이다.

먼저 국내에서는 기업이윤을 증대하기 위해 자금을 투자할만한 여력이 있는 우량 대기업들만이 데이터마이닝 툴들을 사용하고 있는 형편이고 대부분은 외국산 데이터마이닝 툴들을 사용하고 있다는 점이 가장 큰 문제점으로 지적되고 있다. 그들은 자체적으로 나름대로 데이터마이닝의 활용을 성공적으로 평가하고 있으며 그 예들은 다음과 같다.

1. SAS 코리아는 현대 증권, LG투자증권, 한화 증권의 CRM 프로젝트를 SAS의 데이터마이닝의 툴을 이용하여 수행했거나 수행하고 있음.
2. LG캐피탈은 IBM의 데이터 마이닝 솔루션인 인텔리전트 마이너를 사용하고 있음.
3. BC카드사의 분실/도난 카드에 대한 조기검색시스템은 SAS사의 데이터마이닝 기법 중의 하나인 인공신경망(Neural Network Application) 방법을 이용하여 개발하여 사용하고 있음.

그러나 그 외국산 데이터마이닝 소프트웨어 가격은 평균 이만 달러에서 삼만 달러 정도이며 비싼 것은 오십만 달러에 이른다. 이는 국내 IT 분야에 많은 외화 낭비를 의미한다. 또한 대기업들의 외국산 데이터마이닝 툴들의 선호는 데이터마이닝에 관련된 국내 벤처기업 활성을 막는 것을 의미하며 그에 따른 국내 자체 연구의 위축을 초래한다는 지적이 나오고 있다. 또한 데이터마이닝 사용의 시작 단계이므로 경험이 매우 부족한 상황이다. 그러므로 데이터마이닝 전문가들의 수효는 매우 적어 각 기업에서 값비싼 데이터마이닝 툴을 제대로 활용 못하고 있는 실정이다.

국내에서 데이터마이닝의 활용의 최대 장애는 각 기업들의 데이터들의 신뢰도가 미국이나 다른 선진국들에 비해 매우 떨어진다는 것이다. 다시 말해 어떤 데이터베이스에는 일관성이 없고 불완전하며 오류가 있는 데이터들이 존재한다는 것을 의미한다. 아무리 데이터마이닝 툴이 훌륭해도 이러한 데이터베이스를 사용하게 되면 그 활용도가 낮아질 수밖에 없다.

또한 국내에서 새로운 연구 분야에 아직 미흡하다. 예를 들면 분산 마이닝 (Distributed 데이터마이닝) 이나 WWW 환경 하에서의 마이닝 (Web Mining) 등은 아직 극히 초보단계다. 그리고 데이터마이닝 연구는 그 성격상 인공 지능, 데이터 베이스 통계학, 응용 수학, 컴퓨터 그래픽, 비쥬얼라이제이션 등 같은 여러 다양한 학문들을 바탕으로 이루어져야 하는데 아직 국내에서는 통계학자 위주의 데이터마이닝 연구만이 활발하게 이루어 지고있다. 그리고 CRM과 같은 당장 수익성이 있는 듯한 분야에만 응용되고 있어 더욱 다양한 분야에 적용될 수 있는 사회적인 문화가 절실하다.

6.2. 데이터마이닝 발전을 위한 제안

미국의 경우 데이터마이닝의 기술들의 대부분이 미 연방정부의 연구 후원으로 이루어 졌으며 Federal 데이터마이닝 Conference 등을 개최하면서 데이터마이닝의 필요성을 사회 전반에 알리며 기술들을 보급하고 있다. 국내의 경우에도 데이터마이닝 전반적인 발전을 위해서는 정부의 책임이 제일 크다고 볼 수있다. 데이터마이닝의 발전을 위한 정부의 역할을 정리하면 다음과 같다.

1. 데이터마이닝의 필요성을 사회 전반에 홍보하기 위한 프로그램을 만든다. 예를 들면 그 프로그램을 통해 정부 각 부처 내에서나 기업들에서 데이터마이닝의 활용을 적극 권장한다. 혹은 세계적인 수준인 데이터마이닝 컨퍼런스나 심포지엄 주최를 위해 후원한다. 또는 미국과 같은 선진국에서 데이터마이닝의 다양한 경험을 갖춘 과학자들을 초빙하여 세미나 주최를 후원한다.

2. 선진국에서 활동하고 있는데이터마이닝 전문가들과 국내 수학자, 통계학자, 컴퓨터 학자들이 어떤 특정한 분야 (예를 들면 바이오 인포메틱스) 과학자들과 한 팀이 되어 적절한 데이터들을 만들어 내고 그 데이터들을 분석하는 알고리즘이나 소프트웨어들을 개발할 수 있도록 여러 학문들이 관련된 대형 연구 프로젝트들을 지원한다.

국내 기업 또한 중요한 역할을 감당해야 할 위치에 놓여있다. 먼저 데이터마이닝의 응용 분야의 연구를 지원하며, 데이터마이닝 연구자들이 연구를 위해 이용할 수 있는 실제 데이터들을 제공하며, 대학과 산업간의 합동 연구 프로젝트들을 지원하는 등이다. 그리고 데이터마이닝 분야 관련 연구자들이나 엔지니어들의 역할도 매우 중요하다. 예를 들면 각 분야에서 새로운 데이터들이 모아지고 저장될 때 미래에 그 데이터들이 데이터마이닝에 의해 분석될 때 더욱 더 용이하게 하기 위해 국내 실정에 맞는 프로토콜 (protocol) 이나 컴퓨터 언어 그리고 표준 등을 만들어 나가고 그 데이터들을 장기간 사용할 목적으로 데이터들의 관리나 그들을 사용할 수 있는 시설들을 구축하는 것 등이다.

7. 결론

데이터는 점점 복잡해지고 방대해 지고 그 데이터들을 분석하는 전문가들의 수효는 한정되어 있는 상황에서 데이터마이닝이 자연스럽게 출현했다. 데이터마이닝의 활용은 비즈니스 , 과학, 산업, 교육, 의료, 인터넷 등 많은 분야에서 이루어지며 심지어는 오락, 스포츠, 여행정보 등에서도 사용되고 있다. 그리고 앞으로도 계속해서 더욱더 다양한 분야에 적용될 것이라고 예견되고 그 시장성의 규모는 전세계적으로 점점 커질 것이다. IT 인프라가 잘 형성된 북미와 유럽 대륙의 선진국들에서 특히 데이터마이닝의 활용도가 높으며, 한국을 비롯해 일본, 중국, 싱가포르, 대만, 말레이시아, 태국 등 아시아 국가들에서도 데이터마이닝에 대한 관심이 높아지고 있다.

데이터마이닝은 인공 지능, 데이터베이스, 통계학, 응용수학, 컴퓨터 그래픽, 데이터비쥬얼라이제이션 등과 같은 여러 학문이 관련된 분야에서 나온 여러 기술들이 다양한 응용 분야에서 사용되고 있다. 그러므로 데이터마이닝에 필요한 각 분야들의 지속적인 발전과 그들의 효과적인 협력이 데이터마이닝의 분야를 더욱 더 활발하게 할 것이다.

참고 문헌

[1] Cahlink, George, Data Mining Taps the Trends, Government Executive Magazine September 18, 2000

[2] Fayyad, Usama M., Advances in Knowledge Discover and Data mining, 1998

[3] Fu, L.M., Knowledge Discovery, Communications of the ACM, No. 11, Vol. 42, pp 47 - 50

Nov., 1999.

[4] Fu, L.M., A neural network model for learning domain rules based on its activation function characteristics, IEEE Trans. Neural Networks, 9, 5 (1998), pp 787-795; see also www.cise.u.edu/fu

[5] Goebel, Michael, A survey of data mining and knowledge discovery software tools. SIGKDD Explorations, June 1999.

[6] Groth, Robert, Data Mining - building Competitive Advantage, Prince Hall, 2000.