韓國 OA學會誌
第8卷 第2號, 2001. 6

데이터 마이닝 기법의 현황 및 추세

오 승 준*, 송 영 덕**, 오 민 근***

Current Status and Trend of Data Mining Techniques

Seung-Joon Oh*, Young-Duck Song**, Min-Keun Oh***

요 약

최근에 이용 가능한 데이터의 양이 폭발적으로 증가하고 있다. 따라서, 이들 데이터로부터 유용한 지식을 발견하는 자동화된 기법이 주목을 받고 있다. 데이터 마이닝이란 지식 발견의 중요한 단계로서, 데이터로부터 유용한 패턴을 발견하는 방법이다.

본 논문에서는 데이터 마이닝 기법을 조사한다. 이러한 조사과정을 통하여 실세계에서 보다 효율적으로 적용 가능한 데이터 마이닝 기법을 찾아내고, 이들 기법에 대한 적절한 응용 영역과 앞으로의 연구방향을 제시한다.

Abstract

Recent times have seen an explosive growth in the availability of various kinds of data. It has resulted in an unprecedented opportunity to develop automated data-driven techniques of extracting useful knowledge. Data mining, an important step in this process of knowledge discovery, consists of methods that discover interesting, non-trivial and useful patterns hidden in the data.

In this paper, we surveyed data mining techniques. We find effective data mining techniques in applying real world, and suggest appropriate application area for the each techniques. We conclude the paper with some research issues.

* 동원대학 인터넷정보과 겸임교수
** 동원대학 정보통신과 초빙교수
*** 동원대학 사무자동화과 겸임교수

Ⅰ. 서론

최근에 데이터의 양이 폭발적으로 증가하고 다양해짐에 따라 데이터 마이닝에 대한 관심이 높아지고 있다. 데이터 마이닝이란 대용량의 데이터베이스에서 의미 있는 패턴과 규칙을 발견하고 분석하는 작업을 뜻한다[1,2]. 현재 널리 쓰이고 있는 데이터 마이닝 기법들에는 연관 규칙(Association Rule), 클러스터링(Clustering), 분류(Classifications) 및 뉴럴 네트워크 (Neural Network) 등이 있으며, 최근에는 웹 마이닝이 새로운 분야로 주목을 받고 있다.

본 논문에서는 데이터 마이닝에 효율적으로 적용되고 있는 여러 가지 기법들을 선정, 분석하여 데이터 마이닝에 대한 폭 넓은 이해를 도모하고, 이들 기법들이 실세계에서 보다 효율적으로 적용 가능하도록 앞으로의 연구방향을 제시한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 연관규칙에 대해서 알아보고, 제 3장에서는 클러스터링에 대해서 살펴본다. 제 4장에서는 분류 및 뉴럴 네트워크에 대해 조사하고, 제 5장에서는 웹 마이닝에 대해서 살펴본다. 마지막으로 제 6장에서는 결론 및 향후 연구를 제시한다.

Ⅱ. 연관 규칙

연관 규칙을 탐사하는 문제는 기본적으로 미리 결정된 최소 지지도 이상의 트랜잭션 지지도를 가지는 항목집합들의 모든 집합들인 빈발항목집합들을 찾아내어 연관 규칙을 생성하는 단계로 이루어진다.

'시장 바구니 분석(market basket analysis)'을 예로 들어 연관 규칙 문제를 설명하면 다음과 같다. 시장 바구니 분석이란 고객이 소매점에서 상품을 거래한 이력을 저장한 소매점 데이터로부터 고객의 구매 행위를 분석하는 것이다. 여기서 연관규칙 탐사는 사용자가 관심을 나타낼 만큼 소매점 데이터에서 빈번하게 발생하는 상품들의 집합들인 빈발항목집합을 먼저 발견한 후 연관 규칙을 발견하는 것이다.

예를 들면, 맥주(x1)를 산 사람의 80%는 기저귀(y1)도 함께 산다에서 x1→y1인 연관 규칙을 생각할 수 있다. 연관 규칙의 대표적인 알고리즘들은 다음과 같다.

① Apriori 알고리즘[3]

대부분의 연관 규칙 알고리즘은 후보항목집합의 생성 단계와 후보항목집합의 지지수 계산 단계로 구성된 Apriori 알고리즘을 기초로 한다. Apriori에서는 각 패스에서 빈발항목집합들의 후보 항목집합을 구성하고 난 후에 각 후보 항목집합의 발생 빈도수를 계산하고, 사용자가 정의한 최소 지지도를 기초로 하여 빈발항목집합들을 결정한다. Apriori-gen이라는 새로운 후보 항목집합 생성전략을 개발하여 연관 규칙 부분에 중요한 기여를 하였다.

② DHP(direct hashing and pruning)[4]

해시 함수를 사용하여 $|C_k|$을 감소시켰다. $|C_k|$이 크면 클수록 (즉, 후보 k-항목집합의 총 개수가 많으면 많을수록) Lk를 발견하는 데 요구되는 실행시간도 더 증가하기 때문에 $|C_k|$을 감소시키는 것은 매우 중요하다. DHP 알고리즘은 특히 후보 2-항목집합의 총 개수 $|C_2|$을 Apriori 알고리즘에 비해 매우 감소시켰다.

DHP 알고리즘은 지지수를 계산하기 위해서 액세스하는 데이터베이스에 대해서 데이터베이스의 트랜잭션의 수뿐만 아니라 각 트랜잭션 항목의 수 또한 감소시킬 수 있음을 보여 주었다. 이렇게 데이터베이스를 축소하여 데이터베이스를 액세스하는 시간을 단축하고 지지수를 계산하는 시간도 단축하였다.

③ 분할(partition) 알고리즘[5]

메인 메모리에서 처리할 수 있는 단위로 데이터베이스를 충분히 작게 분할함으로써 데이터베이스를 단지 두 번 액세스하여 빈발항목집합을 발견한다. 처음 데이터베이스를 액세스하여서는 메인 메모리에 분할되어 저장된 각 데이터베이스에 대해서 모든 부분 빈발항목집합을 발견한다. 부분 빈발항목집합은 분할된 데이터베이스에 대한 최소지지도보다 큰 지지도를 갖는 항목집합이다. 두 번째 데이터베이스를 액세스하여서는 부분 빈발항목집합을 후보항목집합으로 하여 빈발항목집합을 발견한다.

분할 알고리즘은 데이터베이스를 단지 두 번 액세스하기 때문에 입출력 시간은 단축하지만 메인 메모리에서 연산시간이 증가한다. 빈발항목집합을 발견할 때 입출력 시간보다 메인 메모리에서 연산 시간이 전체 실행시간을 좌우한다. 특히, 코드화된 파일로부터 빈발항목집합을 발견할 경우에는 메인 메모리에서 연산 시간이 더 중요하다.

④ FP-Tree 알고리즘[6]

대부분의 기존 연구들이 Apriori계열의 후보항목집합 생성과 테스트 단계를 채택한다. 그러나 많은 수의 패턴과 긴 패턴들이 존재할 경우에는 후보항목집합 생성에 많은 계산량이 요구된다. 여기서는 FP-tree (Frequent Pattern tree)라는 새로운 구조를 제안한다. FP-tree 구조란 빈발 패턴들에 대한 중요한 정보들을 압축하여 저장하는 확장된 트리구조를 말한다. 이 FP-tree 구조를 사용하여 완전한 빈발항목집합을 찾아내는 새로운 기법을 제안한다.

최근에는 기본적인 연관 규칙을 응용한 순차 패턴, 순회 패턴, 주기적인 연관성 등에 관한 연구가 이루어지고 있다. 연관 규칙을 활용하는 관점에서 사용자와 대화식으로 시스템이 구성되어 보다 유용한 규칙을 찾아내려는 시도도 이루어지고 있다. 연관 규칙의 중요도 및 관심도는 그 규칙의 지지도와 신뢰도가 있고, 최근에는 확신도(conviction)와 개선도(improvement)등이 유용성의 측정단위로 연구되고 있다. 최근의 연구 분야로는 여러 단계에 걸쳐 최소지지도를 만족하는 연관 규칙을 찾는 알고리즘[7], 병렬로 후보항목집합을 효과적으로 찾기 위한 확장 가능한 알고리즘[8] 등이 있다.

Ⅲ. 클러스터링

클러스터링이란 속성들의 값에 의거하여 유사한 속성 값을 가지는 객체들끼리 그룹핑(grouping)하는 작업이다. 현재의 클러스터링 기법들은 크게 분할(partition)방법과 계층적(hierarchical)방법의 두 가지로 나눌 수 있다. 분할 방법은 어떠한 범주 함수를 최적화 시키는 k개의 분할을 결정해 나가는 방법으로 Euclidean distance 측정법에 기반을 둔다. 여기에는 클러스터의 무게중심점을 대표 값으로 분할해 나가는 k-means 방법과 클러스터내에 중심과 가장 가

까운 object로 대표 점을 찾아 가는 k-medoid 방법이 있다.

계층적 방법은 처음에 각 객체를 별개의 클러스터로 설정 한 후, 유사한 객체들을 병합/분할해 나가는 방법으로 모든 객체들이 한 클러스터에 포함 될 때까지 과정을 진행해 나간다. 클러스터링 분야의 대표적인 알고리즘들은 다음과 같다.

① CLARANS(Clustering Large Applications based on RANomized Search)[9]

k-medoid 방법의 대표적인 알고리즘인 PAM(Partitioning Around Methods)[9][10]은 가장 최소로 비용이 발생하는 선택된 객체와 선택되지 않은 객체 쌍(pair)을 결정하여 초기 medoid를 구한다. 이러한 초기 medoid를 구하기까지의 과정에서 과다한 계산량(complexity)으로 인하여 데이터가 큰 경우에는 적합치 못하다. 이러한 단점을 보완하기 위하여 모든 객체들을 교체 대상에서 고려하는 것이 아니라 데이터에서 샘플을 추출하여 medoid를 찾는 CLARA(Clustering LARge Applications)알고리즘이 제안되었다. 그러나 이 방법도 추출된 샘플들에서 최소가 되는 medoid가 없는 경우가 발생할 수 있는 단점이 있다. PAM과 CLARA의 단점들을 보완한 알고리즘이 CLARANS이다.

② CURE(Clustering Using REpresentatives)[11]

CURE는 클러스터당 하나 이상의 대표점을 가지며, 이들은 클러스터의 평균값으로 줄어드는 well -scattered point로 지정된다. 계층적 방법을 적용시킬 때 합병되는 두 클러스터에 대한 대표점은 합병된 클러스터내의 모든 점에 대해서가 아닌 두 오리지널 클러스터로부터 선택되어지며 특히 랜덤 샘플링과 분할, k-d tree와 heap data 구조를 사용함으로써 기존에 알고리즘들이 찾아낼 수 없었던 긴 원형의 클러스터를 발견 가능하게 하는 특징이 있다.

③ ROCK(RObust Clustering using linKs)[12]

시장 바구니 분석과 같이 Boolean이나 범주형(categorical) 속성을 갖는 데이터에 대한 계층적 클러스터링 알고리즘으로 제안되었다. 각각의 클러스터를 합병할 때 데이터간의 유사성 측정 기준으로 거리 대신 링크라는 새로운 개념을 도입하였다. 즉 두 개의 포인트 pair가 유사성 측면에서 특정 thershold 이상인 경우 이웃이라고 하고, 포인트들간의 공통 이웃의 개수를 링크수라고 정의한다. 동일한 클러스터에 속하는 점들은 일반적으로 많은 수의 공통 이웃의 수를 갖고 동시에 많은 수의 링크를 갖는다. 그러므로 클러스터를 합병할 때 가장 많은 수의 링크를 갖는 것끼리 합병하는 것이 의미 있는 클러스터를 생성하게 된다.

클러스터링 알고리즘은 다량의 데이터 셋에 대해 효율성을 증가시키는 방법으로 여러 가지 샘플링 기법이나 경계 최적화 기법, 인덱스 기법, 집중화 기법등을 사용하고 있으며, 향후 대상 데이터 집합의 특성과 클러스터링 목적에 따른 최상의 알고리즘 선택 기준에 대해 지속적인 연구가 필요하다.

Ⅳ. 분류 및 뉴럴 네트워크

1. 분류

분류(Classification)는 마이닝 분야에 있어서 주요 연구분야중 하나로 그 목적은 과거에 알고 있는 DB정보로 부터 새로운 DB 투플을 분류해낼 수 있는 분류 규칙을 생성해 내는 것이다. 따라서 분류 문제는 다음과 같이 기술되어진다.

"트레이닝 셋이라 불리는 입력 데이터가 여러 레코드로 구성되어지고 각 레코드는 애트리뷰트를 갖는다. 각 레코드는 클래스(또는 그룹) 레이블로서 표현되어진다. 분류의 목적은 데이터에 나타난 특징(feature)을 사용해서 입력 데이터를 분석해서 각 클래스에 대한 정확한 모델을 개발하는 것이다. 클래스 값은 앞으로 클래스 레이블을 모르는 테스트 데이터(test data)를 분류하는데 쓰인다[13]." 분류의 대표적인 기법들은 다음과 같다.

① CART(Classification and Regression Trees)

지니 지수(Gini Index) 또는 분산의 감소량을 이용하여 이지분리를 수행하는 알고리즘이다. 여기서 이지분리란 부모마디로부터 자식마디가 2개만 형성되게 한다는 것을 의미한다. CART는 모든 가능한 분리자(splitter)를 조사하고, 그 중에서 가장 뛰어난 레코드를 찾아내는 것이고, 가지치기 할 때에는 최대한의 예측력을 가진 레코드를 찾기 위해서 가지치기를 한다.

② C*.*

범주화된 분리자를 다루고, 서브 트리를 평가할 때에 테스트 집합을 사용하지 않는다. CART의 가지치기와는 달리 끝마디에서 에러율을 조사하여 최소한의 에러율을 얻기 위해서 가지치기를 한다. 고전적인 트리분류기인 ID3의 후계자이다.

2. 뉴럴 네트워크

인간의 사고와 인지에 관심이 있던 인지과학자와 새로운 계산 모형에 관심을 갖고 있던 학자들은 신경해부학적 사실을 토대로 하여 간단한 연산기능만을 갖는 처리기(신경세포)를 고안했다. 그리고 이러한 처리기들을 가중치(weight)를 갖는 데이터 통로(channel)로 연결한 망(network) 형태의 계산 모형을 제안하였다. 이렇게 제안된 모형을 뉴럴 네트워크(neural network)라고 한다.

뉴럴 네트워크에서 정보는 신경세포간을 연결하는 연결가중치의 형태로 저장된다. 이때 연결가중치 W를 조정함으로써 정보를 저장시키는 과정을 학습이라 한다. 학습 방법은 크게 지도 학습(supervised learning)과 비지도학습(unsupervised learning)으로 나눌 수 있다.

지도학습에서는 학습데이터로 입력벡터와 함께 입력이 가해졌을 경우 기대되는 출력을 제시한다. 이때 뉴럴 네트워크에서 출력된 결과가 기대되는 출력과 다르면 그 차이를 줄이는 방향으로 연결가중치를 조정한다. 지도학습을 따르는 대표적인 알고리즘들에는 probabilistic neural network[14], radial-basis function[15] 등이 있다.

반면 비지도 학습에서는 학습데이터가 입력 벡터만으로 구성되며, 지도학습에서처럼 기대되는 출력 벡터는 제시되지 않는다. 비지도 학습을 따르는 대표적인 알고리즘들에는 Kohonen self-organizing map[16], adaptive resonance theory등이 있다.

뉴럴 네트워크는 각 신경세포가 독립적으로 동작하는 처리기 역할을 하기 때문에 병렬성(parallelism)이 뛰어나고, 많은 연결선에 정보가 분산되어 저장되기 때문에 몇몇 신경세포에 문제가 발생하더라도 전체 시스템에는 큰 영향을 주지 않는 결함극복(fault-tolerant) 능력이 있으며, 주어진 환경에 대한 학습능력이 있다. 이러한 특성 때문에 패턴인식, 화상처리, 최적화 문제, 제어(control) 등 여러 분야에서 유용한 도구로서 활발히 연구되고 있다.

V. 웹 마이닝

웹 마이닝은 웹 문서나 서비스로부터 데이터 마이닝 기법을 사용하여 정보를 발견하거나 추출하는 기법[17]으로, 표 1.과 같이 Web content mining, Web structure mining, Web usage mining등 세 분야로 나눌 수 있다[18].

표 1. 웹 마이닝 분류
Table 1. Web mining categories

	Web content mining	Web structure mining	Web usage mining
데이터 관점	Unstructure /Semi structure	Links structure	Interactivity
주요 데이터	텍스트문서, 하이퍼텍스트문서	Links structure	서버 로그, 브라우저 로그
표현	Bag of words, n-grams, Edge-labeled graph (OEM)	그래프	관계형 테이블, 그래프
방법	TFIDF, Machine learning Proprietary algorithms	Proprietary algorithms	Machine learning, 변형된 연관규칙
응용분야	분류, 클러스터링, 웹 사이트 스키마 발견	분류, 클러스터링	사이트 제작 및 운영, 마케팅

Web content mining은 웹 사이트의 컨텐츠, 자료, 정보 등의 관계를 분석하여 사용자의 요구에 가장 잘 부합하는 내용을 보여 줄 수 있도록 자동적으로 찾아주는 기법이다.

Web structure mining은 웹 사이트와 웹 페이지의 하이퍼 링크를 데이터 마이닝 과정을 통해 정보를 구조화, 표준화 시키는 프로세스이다.

Web usage mining은 웹 서버로부터 사용자의 엑세스 패턴을 발견하는 자동화된 마이닝 기법을 말한다. [19]와 같이 기존의 데이터 마이닝 기법들을 이용할 수도 있고 합성 연관 규칙[20]이나 순서 발견 기법[21]처럼 기존 알고리즘을 변형한 방법들도 있다.

웹 마이닝을 통해서 기업은 웹사이트상의 패턴을 의미 있는 정보로 종합해내고, 인터넷상의 고객들과 예상치들을 이해하고 연관시킬 수 있게 된다. 데이터와 웹이 제공하는 방대한 사업지식의 흐름에 근거한 웹 마이닝은 온라인 고객과의 관계를 생성하고 유지시키며 생산성 있는 온라인 상점의 최전선을 구축하는데 있어 결정적 열쇠가 되는 것이다.

Ⅵ. 결론 및 향후 연구

지식 탐사 프로세스의 핵심적인 역할을 담당하는 데이터 마이닝 단계에서는 여러 가지 목적에 따라 알맞은 기법들을 선택하여 사용한다. 최근 통계, 비즈니스, 전자상거래, 의학, 생물학 등의 분야에서 데이터 마이닝 기술이 적극적으로 활용되고 있으며 이를 위해 다양한 기법들이 계속해서 연구, 개발되고 있다.

본 논문에서는 데이터 마이닝중 연관 규칙, 클러스터링, 분류 및 뉴럴 네트워크, 웹 마이닝 분야에 대한 현황 및 여러 기법들을 조사, 분석하였다.

향후에는 본 논문의 기법들에 대하여 실제 데이터를 기반으로 텍스트 마이닝, 공간 마이닝, 이미지 마이닝 등의 구체적인 응용 분야별 데이터 마이닝 기법들의 활용도 및 적용 가능성, 특성 등을 비교, 분석할 연구가 필요하다.

참고문헌

[1] A. berson, S. smith and K. thearling, Building Data Mining Application for CRM, McGraw-Hill,2000

[2] M. berry and G. linoff, Data Mining Techniques for Marketing Sales and Customer Support, John Wiley & Sons, inc, June 1997

[3] Agrawal, R. and R. Snkant, "Fast Algorithms for Mining Association Rules", Proc. of the 20th Int'l conference on Very Large Databases, pp. 487-499, Santiago, Chile, Sep. 1994

[4] J. S. Park, M. S. Chen, and P. S. Yu, "An Effective Hash-Based Algorithm for Mining Association Rules", Proc. of the ACM SIGMOD Int'l Conference on Management of Data, pp. 175-186, San Jose, California, May 1995

[5] A. Savasere, E. Omiecinski, and S. Navathe, "An Efficient Algorithm for Mining Association Rules in Large Databases", Proc. of the 21st Int'l Conference on Very Large Databases, pp. 432-444, Zurich, Switzerland, Sep. 1995

[6] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation", Proc. of the ACM SIGMOD Int'l Conference on Management of Data, Dallas, TX, May, 2000

[7] Bing Liu, Wynne Hsu, and Yiming Ma, "Mining Association Rules with Multiple Minimum Supports", In Proc. of ACM KDD-99, 1999

[8] Eui-Hong Han, George Kerypis, and Vipin Kunmar, "Scalable Parallel Data Mining for Association Rules", In Proc. of the SIGMOD, 1997

[9] R. Ng and J. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining", In Proc. 1994 Int'l Conference on Very Large Databases, pp144-155, Santiago, Chile, Sep. 1994

[10] L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data: an Introdution to Cluster Analysis", John Wiley & Sons, 1990

[11] S. Guha, R. Rastogi, and Kyuseok Shim, "CURE: An Efficient Clustering Algorithm for Large Databases", SIGMOD98, 1998

[12] S. Guha, R. Rastogi, and Kyuseok Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes", IEEE99, 1999

[13] http://www.dpc.or.kr/dbworld/document/9709/spec.html

[14] Specht, D.F., Probabilistic neural networks, Neural Networks, 1990

[15] Mark J. L. Orr, Introduction to Radial Basis Function Networks, Edinburgh Univ., 1996

[16] Kohonen, T., Self-Organizing Maps, Berlin: Springer-Verlag, Second edition, 1997

[17] O. Etzioni, "The world wide web: Quagmire or gold mine", Comunications of the ACM, pp 65-68, 1996

[18] R. Kosals, H. Blockeel, "Web Mining Research: A Survey", ACM SIGKDD, July 2000

[19] J. Srivastava, R. Cooley, M. Deshpande, and P. N. Tan, "Web usage mining: Discovery and

applications of usage patterns from web data", SIGKDD Explorations, 1(2), 2000

[20] J. Borges and M. levene, "Mining association rules in hypertext databases", In Proc. of the Fourth Int'l Conference on Knowledge Discovery and Data Mining, Aug. 27-31, 1998

[21] A. Buchner, M. Baumgarten, S. Anand, M. Mulvenna, and J. Hughes, "Navigation pattern discovery from internet data", In Proc. of the WEBKDD '99, Aug. 1999

□ 著者紹介

오 승 준

94. 2 한양대학교 산업공학과 석사
94. 3 - 2000. 2 대우자동차 기술연구소 주임연구원
2000. 2 - 현재 아이메이트닷넷 선임연구원
2000. 3 - 현재 동원대학 인터넷정보과 겸임교수

송 영 덕

94. 2 아주대학교 컴퓨터공학과 석사
94. 3 - 99. 10 (주)태산정밀 기술연구소 연구원
2000. 3 - 현재 동원대학 정보통신과 초빙교수

오 민 근

93. 2 한국외국어대학교 경영학 석사
95. 1 - 98. 12 제이콥스 정보기술 대표
1999. 3 - 현재 동원대학 사무자동화과 겸임교수
1999. 3 - 현재 BI System 선임연구원, eMulti Korea 대표