通信情報保護學會誌
第7卷 第4號, 1997. 12

데이터 탐사와 보안성

On the Data Mining and Security

심 갑 식*

요 약

웨어하우스나 다른 데이터베이스에 있는 데이터를 어떤 유용한 정보로 변환하는 기술은 데이터 탐사이다. 즉, 데이터 탐사는 데이터베이스의 많은 데이터에서 이전에는 몰랐던 정보를 추출하기 위해 일련의 적당한 질의들을 취하는 과정이다. 데이터 탐사 기술은 통계, 기계 이해(machine learning), 데이터베이스 관리, 병렬 처리(parallel processing)등을 포함한 다양한 기술들의 혼합이다.

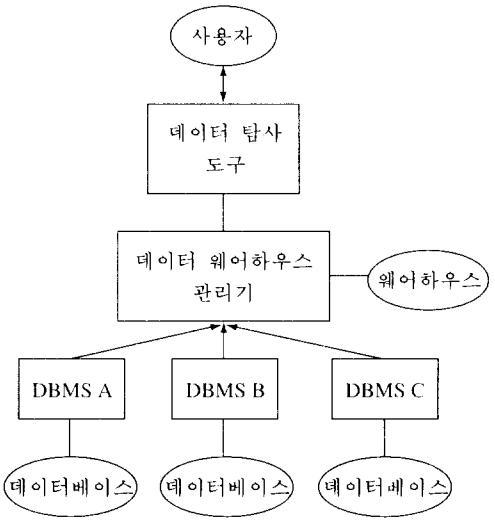
본 연구에서는 데이터 탐사에서 기인될 보안 위협, 이런 위협을 처리하기 위한 기법, 보안 문제점을 처리할 도구로서 데이터 탐사의 이용 등을 알아볼 것이다.

1. 개 요

데이터 웨어하우스에서는 이질형 데이터베이스로부터 데이터를 모으므로(assemble) 사용자는 단일 항목(single point)만 질의한다. 질의에서 사용자가 얻은 응답은 데이터 웨어하우스의 내용에 의존한다. 즉, 일반적으로 데이터 웨어하우스에서는 웨어하우스에 있는 데이터로부터 정보 추출을 시도하지 않는다. 밀접하게 관련된 기술(웨어하우스에 있는 데이터를 유용한 정보로 바꾸는데 사용되는)이 데이터 탐사이다. 즉, 데이터 탐사(data mining)는 데이터베이스에 있는 많은 데이터로부터 예전에 몰랐던 정보를 추출하기 위해 일련의 적절한 질의들을 처리하는 것이다. 그림 1은 데이터 웨어하우징과 데이터 탐사간의 관련성을 나타내고 있다[THUR97a]. 웨어하우스를 가진다고 해서 반드시 탐사를 하는 것은 아니다. 즉, 데이터 탐사는 데이터베이스에도 역시 적용될 수 있다. 그러나, 웨어하우스는 질의 처리를 용이하게 하는 방법으로 데이터를 구조화한다.

기본적으로, 많은 조직체에서 데이터 탐사의 목적은 시장 기능을 향상시키고, 비 정상적인 패턴을 찾고, 과거 경험과 현재 추이를 바탕으로 미래를 예측하는 것이다. 이런 기술에 대한 분명한 필요성이 있다. 저장되어야 할 현재나 과거의 대량 데이터가 있다. 그러므로, 데이터베이스가 더 커질수록 의사결정을 지원하기가 점점 더 어렵게 된다. 또한, 데이터는 여러 원천(source)들이나 도메인(domain)에서 나올 수 있다. 기업체의 계획이나 그밖의 기능을 지원하기 위해서는 데이터를 분석할 필요가 있다.

* 진주산업대학교 조교수



[그림 1] 데이터 탐사 대 데이터 웨어하우징

데이터 탐사를 의미하는 다양한 용어들이 사용되었다. 그것들은 지식/데이터/정보 발견(discovery)이나 지식/데이터/정보 추출(extraction)이다. 어떤 사람은 데이터 탐사를 이전에 몰랐던 정보를 추출하는 과정이라고 정의하고, 지식 발견은 추출된 정보에서 의미를 부여(make sense)하는 과정이라고 정의한다. 본 연구에서 우리는 데이터 탐사와 지식 발견을 구별하지 않을 것이다. 어떤 특정한 기법이 데이터 탐사 기법인지 아닌지를 결정하기는 어렵다. 예를 들면, 어떤 사람들은 통계적 분석 기법이 데이터 탐사 기법이라고 주장한다. 다른 사람들은 그것이 아니라고 하며, 데이터 탐사 기법은 간단하지 않은 관련성을 밝혀내야 한다고 주장한다. 예를 들면, 데이터 탐사에서, 의료 공급 회사는 제품을 살만한 의사를 지향하여 광고함으로써 판매를 신장시킬 수 있고, 신용 회사는 지불 이행을 하지 않을 후보자를 선택함으로써 손실을 덜 수 있다. 그런 실세계 경험은 [GRUP95]에 발표되었다. 또한, 데이터 탐사에서는 비 정상적인 행위(abnormal behavior)를 찾을 수 있다. 예를 들면, 지능적인 대리점은 이런 기술을 사용하여 그들 고용인의 비 정상적인 행위를 결정할 수 있다.

어떤 데이터 탐사 기법에는 비 정규 집합(rough set), 귀납적 논리 프로그래밍(inductive logic programming), 기계 학습(machine learning), 그리고 뉴럴 네트워크(neural network) 등을 바탕으로한 기법들이 포함된다. 데이터 탐사 문제에는 분류(classification : 데이터를 그룹으로 분할하는 규칙을 찾는다), 연관성(association : 데이터 사이에 연관을 지을 규칙을 찾는다), 그리고 순차(sequencing : 데이터를 순서화 하는 규칙을 찾는다)가 포함된다. 기본적으로 예제나 패턴을 관찰함으로써 정보가 추출된다는 가정에 도달하게 된다. 이들 패턴은 일련의 질의들을 제시함으로써 관찰된다. 각각의 질의는 이전 질의에서 얻은 응답에 의존한다. 데이터 탐사에서 몇 가지 개발이 있었다. Lockheed Martin Inc.에 의한 RECON과 같은 도구가 이들에 포함된다. 데이터 탐사의 논의와 여러 가지 상용 도구는 [GRUP95]에 나타나 있다.

본 연구에서는 데이터 탐사에서 기인될 보안 위협, 이런 위협을 처리하기 위한 기법, 보안 문제점을 처리할 도구로서 데이터 탐사의 이용 등을 제시할 것이다.

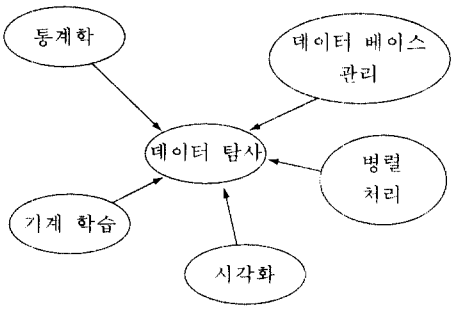
2. 데이터 탐사 기술

데이터 탐사는 그림 2에 나타난 것처럼 여러 가지 기술이 통합된 것이다. 여기에는 데이터베이스 관리, 기계 학습, 시각화, 통계학, 그리고 고성능 컴퓨팅이 포함된다.

데이터 탐사 연구는 여러 분야에서 수행되고 있다. 데이터베이스 관리 연구자들은 연역적(deductive) 질의 처리나 지능형(intelligent) 질의 처리에서 나온 연구성과를 데이터 탐사에 이용하고 있다. 흥미 있는 분야 중 하나는 데이터 탐사가 용이하도록 질의 처리 기법을

확장하는 것이다.

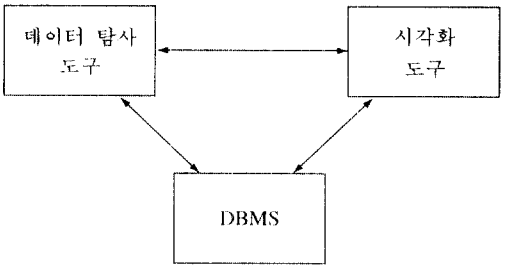
기계 학습이 나온 지 얼마 안되었다. 여기서의 아이디어는 기계가 패턴을 관찰함으로써 여러 가지 규칙을 학습하게 한 후, 문제를 해결하는데 이들 규칙을 적용하는 것이다. 기계 학습과 데이터 탐사에서 사용되는 원리가 유사하지만, 데이터 탐사에서는 탐사할 대량의 데이터를 고려한다. 그러므로, 데이터베이스 관리와 기계 학습 기법의 통합은 데이터 탐사를 위해 필요하다.



[그림 2] 데이터 탐사 기술

컴퓨터 시각화(visualization) 분야의 연구자들은 또다른 관점에서 데이터 탐사를 접근하고 있다. 그 분야 중 한 초점은 데이터 탐사 과정에 도움이 되도록 시각화 기법을 사용하는 것이다. 다른 말로, 대화식(interactive) 데이터 탐사가 시각화 연합체의 목적이다. 그림 3은 대화식 데이터 탐사를 나타낸다. 여기서, 데이터베이스 관리 시스템, 시각화 도구, 그리고 기계 학습 도구 모두는 데이터 탐사를 위해 서로 상호작용한다.

통계 분석 연구자들은 데이터 탐사를 위한 더 정교한 통계 기법을 개발하려고 그들의 기법과 기계 학습 기법을 통합하고 있다. 일찍이 언급한 것처럼, 통계 분석 기법이 역시 데이터 탐사 기법이 되는지에 대한 논쟁이 있다.



[그림 3] 대화식 데이터 탐사

마지막으로, 데이터 탐사 알고리즘이 탄력성이 있도록 고성능 컴퓨팅 분야의 연구자들 역시 적절한 기법 개발에 착수하고 있다. 고성능 데이터 탐사를 위한 적절한 하드웨어를 개발하기 위하여 하드웨어 연구자도 역시 교감을 하고 있다.

2.1 데이터 탐사 유형

다양한 유형의 데이터 탐사가 있다. 우리는 데이터를 탐사하는데 사용되는 실제적인 기법을 나타내지 않고 결과만을 보여줄 것이다. 몇 가지 유형이 [AGRA93]에서 논의되고 있다. 우리는 여기서 몇 가지만을 서술할 것이다.

데이터 탐사의 첫 번째 유형은 "분류(Classification)"이다. 레코드들을 어떤 의미있는 서브클래스(subclass)나 클러스터(cluster)로 그룹짓는 것이다. 그래서, 데이터에서 패턴을 밝혀서 클래스들을 설정한다. 분류의 예제는 다음과 같다. 목록에서 도시 X에 사는 모든 사람들은 20K 이상 가격의 차를 소유한다는 정보를 어떤 자동차 판매 회사가 가지고 있다고 하자. 그러면, 도시 X에는 살지만 목록에는 없는 사람조차도 20K 이상 가격의 차를 소유할 수 있다고 그들은 가정할 수 있다. 이런 방법으로 회사는 도시 X에 살고 있는 사람들을 분류한다.

두 번째 유형의 데이터 탐사는 "순차 탐지

(Sequence detection)"이다. 즉, 데이터 패턴을 분석함으로써, 순차를 결정한다. 순차 탐지의 예제를 보자. John은 은행을 가고난 다음에, 일반적으로 식료품점에 간다.

세 번째 유형의 데이터 탐사는 "데이터 의존 분석(Data dependency analysis)"이다. 여기서는 데이터 항목들 사이에 잠재적으로 흥미있는 의존성(dependency), 관련성(relationship), 혹은 연관성(association)이 탐지된다. 예를 들어, 만일 John, James, 그리고 William이 회의를 한다면, Robert도 역시 그 회의에 참석할 것이다. 이러한 유형의 탐사는 많은 사람들에게 매우 흥미있을 것이다.

네 번째 유형의 탐사는 "이탈 분석(Deviation analysis)"이다. 예를 들면, John은 토요일에 은행에 가지만, 그 후로 식료품점에 가지 않는다. 대신에 그는 풋볼 경기 보러 갔다. 이런 유형으로, 이상(anomalous) 인스턴스나 불일치(discrepancy)를 찾는다.

앞에서 언급한 것처럼, 다양한 기법들이 이런 다양한 유형의 데이터 탐사를 위해 사용된다. 이런 기법들은 비 정규 집합(rough set), 퍼지 논리(fuzzy logic), 귀납적 논리 프로그래밍(inductive logic programming), 혹은 뉴럴 네트워크(neural network) 등에 기반을 둘 수 있다. 상용 제품 역시 이들 기법과 데이터 탐사 유형에 기초해서 개발되었다.

3. 데이터 탐사와 보안성

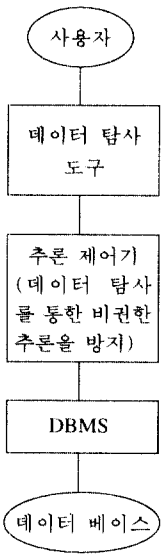
최근에 데이터 탐사와 보안성 간의 관련성을 해결하려는데 많은 관심이 고조되고 있다. 일부의 예비적 아이디어들이 1995년 "the Ninth IFIP 11.3 Working Conference on Database Security"에서 개최된 데이터 탐사 특별 세션에서 논의되었다[LIN95]. 이 주제에 대한 더 세밀한 것은 [MARK96]에 있다. 데이터 탐사와 보안성에는 두가지 특징이 있다[THUR97b]. 하나는 데이터 탐사 기법들이 침투(intrusion) 탐지와 데이터베이스 감사(auditing)에서의 문제점을 처리하는데 적용될 수 있다는 것이다. 감사의 경우에서는, 탐사될 데이터가 많은 량의 감사 데이터라는 것이다. 우리는 데이터 탐사 도구를 비정상 패턴을 탐지하는데 적용할 수 있다. 예를 들면, 한 고용인이 특정 국가를 과도하게 여러번 여행을 한다는 사실이 어떤 질의들을 제시함으로써 알았다고 하자. 제시된 다음 질의는 그 고용인이 그 국가의 어떤 사람과의 연관성 여부이다. 만일 해답이 긍정적이라면, 그 고용인 행위는 프래그(flag)된다.

현재의 데이터 탐사 도구들은 충분히 발전이 되어서 침투나 비 정상적인 행위를 탐지하는데 그런 도구들을 적용할 수 있다. 그러나 이 도구들은 관계형 데이터베이스와 같은 구조화된 데이터베이스에서 작동된다. 그러므로, 이 도구들을 적용하기 위해서는 검사될 데이터가 구조화된 형식으로 먼저 변환되어야 한다. 최근에 [GRIN96] 연구에서의 아이디어는 탐사될 네트워크 침투 데이터를 여러 저장소에 배치시키는 것이다. 이것은 연구자나 개발자들이 이들 공통의 저장소에서 의심스런 행위를 알아보기 위한 알고리즘이나 도구들을 테스트할 수 있게 한다. 다른 말로, 조사될(예를 들면, 탐사될, 시각화될 등등) 네트워크 침투 데이터 집합은 연구자들이 데이터 탐사에 다양한 접근방법들을 비교할 수 있게 할 것이다. 데이터 탐사, 시각화, 다른 도구들 또는 인간 전문가도 이 과정에 이용될 수 있다. 목적은 실시간으로 의심스런 행위를 발견하는데 도움이 되는 도구들을 결정하는 것이다. 텍스트나 이미지와 같은 비 구조화된 데이터에서의 데이터 탐사 연구가 시작되고 있다. 개발이 될 때, 우리는 비 구조적 감사 데이터에 응용될 도구를 기대할 수 있다.

데이터 탐사와 보안에서의 두 번째 특징은 추론 문제이다. 다시 말해서, 이전의 예제는

침투나 비 정상 행위(behavior)를 탐지하는데 데이터 탐사 도구들이 어떻게 사용되는가를 보여주었다. 다음 예제는 데이터 탐사 도구가 어떻게 적용되어 보안 문제를 일으키는지를 보여준다. 데이터 탐사 도구를 이용할 수 있는 사용자를 고려해 보자. 이 사용자는 다양한 질의를 제시하여 비밀의 가정(hypothesis)을 추론할 수 있다. 다시 말해서, 데이터 탐사로 인하여 추론 문제가 발생한다. 이런 문제를 다룰 다양한 방법들이 있다. 한 접근방법은 다음과 같다. 데이터베이스와 데이터 탐사 도구 집합이 주어진다면, 미분류 정보로부터 연역되는 기밀 정보를 정당하게 얻을 수 있는 지를 알아보기 위해 도구들을 적용하라. 그렇다면, 추론 문제가 있다. 데이터베이스가 갱신될 때, 정기적으로 그런 접근방법을 수행할 수 있다. 이 접근방법에는 어떤 논점이 있다. 하나는 우리가 도구들의 제한된 집합만을 적용하고 있다는 것이다. 실제로, 사용자는 그가 이용할 수 있는 다른 데이터 탐사 도구도 갖을 수 있다. 더욱이, 추론 문제가 발생할 수 있는 모든 방법을 망라하기는 불가능하다.

성취하기가 훨씬 더 어려운 또다른 접근방법은 그림 4에 나타난 것처럼 실행 시간 동안에 데이터 탐사-기반 추론 제어기를 적용하는 것이다. 이 의미는 사용자가 질의를 제시할 때, 결과를 내줌으로써 추론 문제가 발생하는지를 결정하는 것이다. 이 접근 방법에서 추론 제어기는 분류, 연관성, 그리고 순차와 같은 데이터 탐사 기법들에 기초할 것이다. 예를 들면, 철수가 서울을 여행할 때마다, 영희도 그렇다는 사실을 보호하고자 한다고 하자. 이것은 철수가 분류된 프로젝트에서 일하고 있고, 영희 역시 동일 프로젝트에서 일하고 있다는 사실을 숨기고자 한다는 사실에서 기인할 수도 있다. 철수와 영희가 항상 서울로 여행한다는 유형을 관찰함으로써, 연관성을 통해서 기밀 사실을 추론할 수 있다. 사용자가 이런 기밀 정보를 추론할 수 있으며 그 사용자에게 어떤 결과를 넘겨줄 수 없다는 것을 추론 제어기가 탐지해야 한다.



[그림 4] 추론 제어기

데이터 탐사에 대한 이론과 기초가 여전히 개발되고 있기 때문에, 두 번째 접근방법에 기초한 추론 제어기를 만들기는 매우 어렵다. 귀납적 논리 프로그래밍과 데이터 탐사 사이의 관련성에 대한 연구가 있지만, 아직은 초기 단계이다. 현재의 데이터 탐사 기법들은 오히려 임기 응변적이므로, 그런 추론 제어기를 만든다는 것은 거의 불가능하다. [THUR95]에서 보고된 연구는 추론 문제를 다루기 위해 유사한 접근방법을 취하고 있다. 그러나 연역적 추리(reasoning)에만 초점을 두고 있다. 데이터 탐사 기법들은 연역적 추리보다 훨씬 더 복잡하다.

[CLIF96]에서 보고된 연구는 첫 번째 접근방법에 기초한 추론 문제를 다루기 위한 기법 개발에 많은 가망성을 보여 주었다. 예를 들

면, 현존하는 다양한 데이터 탐사 도구들을 적용함으로써, 어떤 잠재적 기밀 정보를 연역할 수 있다는 것을 보여 주었다. 해 볼만한 것은 이런 문제을 다룰 기법을 개발하는 것이다. 연구되고 있는 방법들 중에는 질의에 대한 부분적 해답을 주는 것, 부가적 정보나 허튼 정보를 응답에 도입시키는 것, 그리고 관련 질의와 다른 해답을 주는 것 등이다. 이 분야에서의 연구는 매우 초보적인 것이다.

4. 결 론

이 연구에서는 데이터 탐사에 대한 총괄을 알아보고 몇몇 보안 논점들을 서술하였다. 데이터 탐사에서는 데이터 탐사의 두 가지 보안 관련사항 뿐만 아니라, 보안 문제를 다루기 위한 데이터 탐사 도구의 이용을 논의했다.

이 연구에서 논의된 데이터 탐사에서의 보안 논점들은 매우 초보적인 것이다. 데이터 탐사에서의 보안 관련사항에 대한 문제점은 훨씬 더 어렵다. 의미있는 해결책이 개발되기 전에 많은 연구가 필요하다. 그러나, 보안 문제를 설명하기 위한 데이터 탐사 도구의 이용에 대한 연구는 기존 기술로도 시작할 수 있다. 요약하면, 데이터 웨어하우징에서의 보안성, 데이터 탐사와 보안성 사이의 관련성은 많은 연구 거리가 된다.

참 고 문 헌

[AGRA93] Agrawal, A., Imielinski, T., and Awami, A., "Database Mining a Performance Perspective," IEEE Transaction on Knowledge and Data Engineering, Vol. 5, December 1993.

[CLIF96] C. Clifton, and D. Marks, "Security and Privacy Issues for Data Mining", Proceedings of the ACM SIMOD Conference Workshop on Data Mining, Montreal, Canada, June 1996.

[GRIN96] G. Grinstein, "Data Exploration through Mining and Visualization", To be published in the Proceedings of the IEEE Visualization '96 Conference, SanFrancisco, CA, October 1996.

[GRUP95] Grupe, F. and Owrang, M., "Database Mining Tools," in the Handbook of Data Management Supplement, Auerbach Publications, 1995 (Ed: B. von Halle and D. Kull).

[LIN95] T. Y. Lin, D. Marks, T. Hike, and B. Thuraisingham, "Data Mining and Security", Special Session at the 9th IFIP 11.3 Database Security Workshop, N.Y. August 1995.

[MARK96] D. Marks, "Inference in MLS

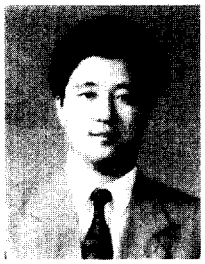
Databases", IEEE Transactions on Knowledge and Data Engineering, February 1996.

[THUR95] B. Thuraisingham and W. Ford, "Security Constraint Processing in a Multilevel Secure Distributed Database System", IEEE Transactions on Knowledge and Data Engineering, April 1995.

[THUR97a] B. Thuraisingham, Data Management Systems Evolution and Interoperation, CRC Press, 1997.

[THUR97b] B. Thuraisingham, "Security Issues for Data Warehousing and Data Mining", Database Security Volume X : Status and Prospects, Chapman & Hall, 1997.

□ 著者紹介



심 갑 식

1985년 2월 전남대학교 계산통계학과(학사)
1987년 2월 전남대학교 대학원 계산통계학과(석사)
1993년 8월 전남대학교 대학원 전산통계학과(박사)
1993년 11월 ~ 현재 진주산업대학교 교양과정부 조교수

※ 주관심분야 : database security, data warehousing and mining