일 반 관 리 분 야

산업재해 관리 시스템 구축을 위한 데이터 웨어하우스 마이닝 기법의 활용

한정훈, 유 훈, 이원근, 심종칠, 김창은
명지대학교 산업공학과

ABSTRACT

데이터 마이닝은 대용량 데이터베이스의 데이터 사이에 묻혀 있는 '패턴'을 발견하여 규칙을 추론함으로 여러 가지 유용한 지식을 캐내는 기법이다. 본 논문에서는 효과적인 재해관리 시스템을 구축하기 위해서 재해를 분석하고 대책을 마련할 수 있는 데이터 마이닝을 적용한 '데이터베이스 웨어하우스 마이닝 재해관리 데이터베이스 시스템'을 제시하고자 한다. 데이터 웨어하우스 마이닝은 다차원 데이터베이스로 구축되며 재해데이터간의 상호관련성, 특성요인별 패턴을 찾고 재해발생 가능성을 예측함으로써 재해예방의 의사결정을 지원할 수 있다.

1. 서론

재해발생 이후 저장되는 대량의 재해 데이터 내에서 '재해패턴'을 발견하여 그 발생 추이를 예측한다는 것은 현재 불가능하다고 할 수 있다.[1][2]

이를 가능하게 하기 위해서는 재해 데이터를 정보로 구조화하여야 한다. 그리고 정보로 모델링된 데이터를 데이터 웨어하우스로 구축하고, 데이터 마이닝을 적용하여 그 대량의 데이터 사이에 묻혀 있는 패턴을 발견하여 규칙을 추론하여 재해발생을 예측함으로써 재해분석의 의사결정을 지원하고, 재해예방의 효과를 이룰 수 있다.[3][9]

본 논문에서는 재해를 분석하고 대책을 마련할 수 있는 데이터 웨어하우스 마이닝 기법을 이용한 재해관리 데이터베이스 시스템을 제시한다.

2. 데이터 웨어하우스와 데이터 마이닝 그리고 데이터 웨어하우스 마이닝

데이터 마이닝(DM : Data Mining)은 데이터베이스로부터의 지식발견(KDD : Knowledge Discovery in Databases)이라고 하는데, 대규모의 데이터 내에 숨겨져 있는 고급 정보를 추출해서 의사결정, 예측, 예보에 응용하고자 하는 기법이다.[11]

데이터 관련 기술이 발전해온 과정을 <표1>에서 살펴볼 수 있다.[10]

제1단계는 1960년대로 데이터의 수집에 중점을 둔 시대이다. 주로 테이프나 디스크 등이 사용되었다. 제2단계는 데이터 액세스 단계로 1980년대의 기술이며 이는 RDBMS로 대표된다. 제3단계는 1990년대의 기술로서 데이터 웨어하우징 단계이다. 이 단계에서는 과

거의 데이터를 다이나믹하게 여러 레벨로 처리해준다.

단 계	특 징
Step 1.(1960s) Data Collection	-retrospective, static data delivery -computers, tapes, disks -IBM, CDC
Step 2.(1980s) Data Access	-retrospective, dynamic data delivery at record level -RDBMS, SQL, ODBC -ORACLE, INFORMIX, IBM
Step 3.(1990s) Data Warehousing	-retrospective, dynamic data delivery at multiple levels -OLAP, multimensional DB -Readbrick, Pilot
Step 4.(2000s) Data Mining	-prospective proactive information delivery -advanced algorithms massive databases -Lockheed IBM(nascent industry)

<표1> 데이터마이닝으로의 발전

제4단계가 바로 데이터 마이닝 단계로 2000년대의 기술이다. 기존의 기술과는 달리 미리 예측적인 정보전달을 하는 것이 그 특징이다.

2.1 데이터 웨어하우스

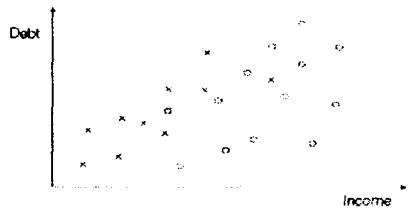
데이터 웨어하우스(Data Warehouse)는 의사결정에 필요한 정보처리 기능을 효율적으로 지원하기 위한 통합된 데이터를 가진 양질의 요약된 읽기 전용 데이터베이스로서, 가장 큰 특징은 운영시스템과 분리되어 있다는 것이다.[5]

데이터 웨어하우스를 구축하는 가장 기본적인 이유는 사용자들이 의사결정을 하기 위해 필요로 하는 정보들을 한곳에 저장해 놓고 이러한 정보는 이를 필요로 하는 사람에게 적시에 적절한 형태로 제공하는 것이 데이터 웨어하우스의 구축 목적이다.[7]

2.2 데이터 마이닝의 개념

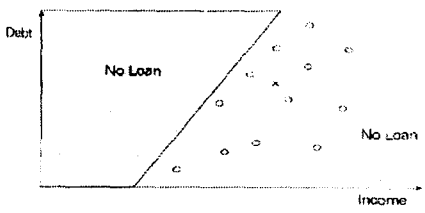
데이터 마이닝은 대규모 데이터베이스에 존재하는 감추어진 지식을 찾아내는 작업으로서, 현실세계에서 쌓여가고 있는 수많은 데이터로부터 유용한 정보를 캐내어 응용하고자 하는 요구에 부응하기 위한 기술이다.[4][13]

데이터 마이닝의 개념을 좀더 쉽게 이해할 수 있도록 예를 들어 설명하면,



<그림 1) 간단한 데이터 셋

<그림1>은 23경우를 가진 이차원의 인공 데이터 셋이다. 데이터 셋은 두 클래스로 나뉜다. 'x'는 대부에 결함이 있는 사람이고 'o'는 대부를 받을 수 있는 좋은 신용상태의 사람을 나타낸다. 이러한 간단한 예로부터 이 데이터 셋은 은행이 대부를 결정할 때 도움을 주는 유용한 지식을 포함하고 있음을 알 수 있다.



<그림 2) 선형분류 경계

<그림 2>는 이 데이터 셋에 분류 개념을 적용한 것으로 데이터 셋이 두 지역으로 나누어 있다. 은행은 이러한 분류 지역을 사용해서 차후 대부가 허용될 것인지 아닌지를 자동적으로 결정할 수가 있다.[8][14]

2.3 데이터 웨어하우스 마이닝

데이터 웨어하우스를 구축한 후 데이터 마이닝 기법을 수행하는 것을 데이터 웨어하우스 마이닝이라 한다.[12]

데이터 마이닝이 데이터 웨어하우스 상에서 수행해야 하는 이유는 첫째, 마이닝을 위한 적합한 데이터가 운영시스템과 분리되어 있고, 둘째로 데이터 웨어하우스는 여러 소스로부터의 데이터를 포함하므로 여러 관계를 마이닝 할 수 있다는 것이고, 셋째는 데이터 웨어하우스는 일차 필터링하여 깨끗한 데이터가 저장되므로 마이닝 정확도를 위한 데이터 질과 일관성이 보장된다는 것이다.

이러한 데이터 웨어하우스의 구축은 다차원적 모델링과 다차원 데이터베이스 기법에 기반을 두고 있다.[13]

3. 데이터 마이닝을 위한 다차원 데이터 모델링

다차원적 데이터베이스(Multidimensional database)란 대규모의 데이터를 쉽게 저장하고 검색할 수 있도록 설계된 데이터베이스 시스템을 말한다.[12]

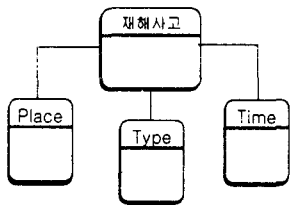
다차원적 데이터베이스 모델링의 가장 큰 장점은 단순성이다. 대량의 데이터를 관계형 데이터베이스로 모델링 한다면, 수십 개 혹은 수백 개의 테이블로 구성되므로 분석을 위한 설계를 제대로 한다는 것은 어려운 일이다.[6] 그러나, 다차원적 데이터베이스는 매우 단순하게 설계된다. 더욱이, 잘된 개념적 다차원적 데이터베이스 모델링은 관계형 DB로 구현될 수도 있고 객체지향 DB로 구현될 수도 있다.[11]

다차원적 데이터베이스 모델링에 사용되는 대표적인 스키마는 스타스키마이다. 스타스키마(Starschema)는 데이터의 새로운 뷰를 제공하는 스키마이다. 스타스키마의 기본 전제는 정보가 Fact와 dimension(차원)으로 분류될 수 있다는 것이다. Fact는 핵심 데이터 요소이고, 차원은 Fact에 관한 특성들이다.

대부분의 경우에 있어 분석은 이러한 차원에 근거하므로 이를 차원 분석이라 한다. 따라서 <그림3>과 같이 스타스키마는 하나의 Fact 테이블과 다수의 차원 테이블로 구성된다.

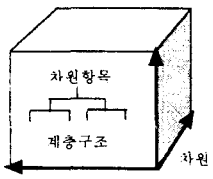
<그림3>은 재해사고에 관한 DB 테이블로서 이는 Place, Type, Time 등 3개의 차원을 가진다. 각 차원은 다음과 같은 애트리뷰트 계층을 가지고 있다.

Place 차원 : 시 → 도 → 지역
Type 차원 : 재해원인 → 카테고리
Time 차원 : 일 → 월 → 년



<그림3> 스타스키마의 개념도

다차원 데이터베이스는 <그림4>와 같이 스타스키마에 의한 차원과 그에 따른 세부적인 차원항목을 갖는다.



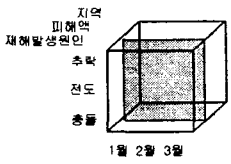
<그림4> 다차원 데이터베이스의 구조

차원 내의 차원항목은 계층구조를 갖으며 소·중·대 분류로 나뉜다.[7]

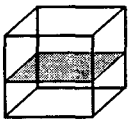
다차원 질의란 사용자가 큐브의 어떤 부분을 볼 것인지를 정의하는 것을 말한다.[5] 이는 사용자가 큐브의 일부분을 절단하여 자신이 원하는 부분을 살펴보는 것과 같다. 따라서 다차원 질의를 슬라이싱 앤 다이싱 (Slicing and Dicing) 이라고 부른다.

다차원 질의의 결과로 사용자는 하나의 셀 값을 얻을 수도 있고, 2차원 혹은 3차원 이상의 서브 큐브(Sub-Cube)를 얻을 수도 있다.

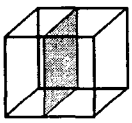
다차원 질의의 결과로 특정 차원들은 보고서의 열(Column)을 형성하고 특정 차원들은 보고서의 행(Row)을 형성한다. 보고서의 열과 행을 형성하지 않는 차원을 페이지(Page)차원이라 한다. 보고서의 열과 행은 하나의 차원들로 구성될 수도 있다.



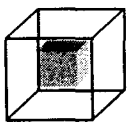
피해액	1월	2월	3월
추락			
전도			
충돌			



전도	1월	2월	3월
재해발생원인			
피해액			
지역			



2월	지역	피해액	재해발생원인
추락			
전도			
충돌			



		2월	3월
추락	피해액		
	재해발생원인		
전도	피해액		
	재해발생원인		

<그림5> 다차원 질의 결과

이처럼 3차원 이상의 차원을 사용자가 볼 수 있는 평면상에 나타내기 위해 차원의 중첩(Nesting)이 이루어진다.

사용자는 <그림5>처럼 보고서의 열과 행, 그리고 페이지 차원을 무작위로 바꾸어 볼 수 있으며 이러한 조작을 피보팅이라 한다.

전통적인 어플리케이션들이 고정된 화면상에서 정형화된 보고서를 단순히 보여 주는 것과는 달리 다차원적 데이터베이스는 사용자에게 대화식 어플리케이션을 제공한다.[5]

다차원 데이터베이스는 이러한 활용 방법에 초점을 맞춘 데이터베이스로서 사용자 관점에서 설계되며, 사용자는 피보팅과 드릴-다운, 드릴-업 등을 통하여 다양한 각도에서 문제를 분석할 수 있으며, 이제까지 함께 고려하지 않았던 관점들을 함께 비교 검토함으로써 관리현황과 미래에 대한 새로운 관점을 획득하고 전략적인 대안을 얻을 수 있다.

4. 데이터 웨어하우스를 기반으로한 재해관리 시스템

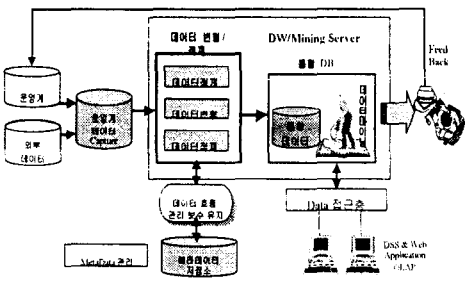
데이터 웨어하우징을 한마디로 표현한다면, 사용자가 쉽게 접근할 수 있고, 다양한 분석을 펼칠 수 있는 분석용 데이터베이스를 운영 데이터베이스와 별도로 구축해야 한다는 사상이다.[7]

정보 시스템 전문가라 할지라도 기존의 트랜잭션 시스템 외에 재해관리 시스템에 데이터 웨어하우스가 왜 필요한지 이해하지 못할 수도 있다.

<그림6>에서 보는 바와 같이 데이터 웨어하우징 모델의 각 구성요소들은 재해관리 시스템의 운영 소스와 함께

존재하는, 전체 시스템 구성의 일부이다. 재해관리 시스템의 웨어하우스 구조는 소스 시스템의 변화와 끊임없이 변화하는 접근 기술들을 수용할 수 있도록 융통성을 지녀야 한다.[7]

본 논문에서 제시하고자 하는 재해관리 시스템은 재해데이터가 접근되고 유용한 '정보'로 바뀔 수 있는 것은 최종 재해관리 시스템 사용자 도구들을 통해서만 가능하기 때문에 , 최종 사용자 도구들이 최종사용자와 데이터베이스간의 인터페이스를 원활히 이루어지게 함으로 재해관리 시스템의 성공의 핵심을 이룰 수 있다.



<그림6> 재해관리 시스템의 개념도

재해관리 시스템의 응용 프로그램은 사용자의 이해를 증진시키는데 도움을 주기 위한 그래픽 사용자 인터페이스(Graphic User Interface). 전문 데이터 조작 및 탐구(drill-down)기능, 그리고 광범위한 사용자 보고서를 제공할 수 있는 전문적인 기능들을 지원하는 사용자 도구들 즉, 응용 프로그램은 범용 클라이언트 / 서버 Tool을 통해 구축할 수 있다. [6]

데이터 웨어하우스 마이닝 기법을 적용한 재해관리 시스템은 다음과 같은 효과를 기대할 수 있다.

-전사적 데이터 통합과 접근 기초마련
-재해관련 프로세스 개선 또는 창출
-재해 발생원인에 대한 이해 증진
-분석 속도와 유연성 향상
-새로운 재해 관리 영역 창출

6. 결론 및 추후 연구과제

본 논문은 데이터 마이닝을 통해 연관규칙, 분류규칙, 요약규칙, 클러스터링 등 여러 가지 지식들을 이용하여 대용량의 재해 데이터베이스에 존재하는 패턴을 찾으므로 재해발생의 의사결정, 예측, 예보, 평가 등에 광범위하게 적용될 수 있음을 제시하고자 하였다.

하지만 무엇보다도 기본이 되는 것은 올바른 데이터의 수집이다. 현 시점에 있어서 가장 시급히 이루어져야하는 문제는 재해조사 항목의 표준화이다.

올바른 데이터 취합이야 말로 재해관리시스템 구축에 첩경이며, 아무리 좋은 분석시스템이라도 불확실한 데이터로부터 나올 분석결과는 불확실할 것이라는 것은 자명한 사실이 아닐 수 없다.

데이터는 데이터 마이닝 과정에서 필수적이며, 데이터 웨어하우스는 이런 일련의 과정을 지원하는 최선의 구조이다.

슬라이싱 앤 다이싱을 할 수 있는 다차원 데이터베이스는 데이터뿐만 아니라 어플리케이션이 포함되었다고 할 수 있는데 RDBMS와 달리 이는 전문가가 아니라도 분석이 용이하다는 장점이 있다.

재해관리 시스템의 구축에 있어서 데이터 웨어하우스 마이닝 기법의 적용은 재해예방에 실현을 앞당길 수 있고,

Web을 기반으로 하는 응용 프로그램을 통하여 새로운 관리체제의 정착을 이룰 수 있을 것이다.

모든 관리기법의 최종사용자는 작업자이다. 아무리 훌륭한 데이터 관리기법이라 하여도 질의의 응답하는 작업자들이 어렵게 느낀다면 소용없는 것이다. 최종사용자들과 데이터베이스간의 인터페이스가 원활히 이루어지게 하는 것이 관리 시스템의 성공의 핵심이 될 것이다.

추후에도 계속적으로 연구되어야 할 분야는 재해관리 시스템의 데이터 웨어하우스로 업데이트 방법, 데이터 웨어하우스와 OLAP(On-Line Analytical Processing)의 연계 등이다.

또한 분산 마이닝이나 WWW 환경하에서의 마이닝 등은 아직 극히 초보단계이므로 많은 연구가 이루어져야 할 것이다.

<참고문헌>

[1] 김용수, 김창은, 심종칠 공저, '안전공학론', 한올출판사, 1997.

[2] 허성관 저, '산업안전관리론', 보성각 ,1995.

[3] 김병석 저, '신산업재해방지론', 형설출판사, 1998.

[4] 'Data Mining', Pieter Adriaans, Dolf Zantinge : Addison-Wesley 1997.

[5] 조재희 박성진 저, '데이터 웨어하우징과 OLAP', 대청, 1996-1998.

[6] 나민영 편저, '데이타베이스 설계', 기한재, 1997.

[7] 레이먼 C. 바킨, 'Planning and designing the DATA Warehouse', 데이터 웨어하징 연구소, 1997.

[8] Agrawal, R., Imielinski, T., and Swami, A., "Mining Association Rules between Sets of items in Large Databases", Proceedings of the ACM SIGMOD Conference, 1993.

[9] Agrawal, R, and Srikant, R., "Mining Swquential Patterns", Proceedings of the 11th Data Engineering, 1995.

[10] Chen, M. S., Han, J., and Yu, P. S., "Data Mining : An Overview from Database Perspective", IEEE TKDE, (to appear) 1997.

[11] Han, J. et al., "DBMiner : A System for Mining Knowledge in Large Relational Databases", 2nd International Conf. on Knowledge Discovery and Data Mining(KDD' 96), 1996.

[12] Holsheimer, M., Siebes, A. P. J. M., "Data Mining : The Search for Knowledge in Databases", Technical Report CS-R9406, CWI, Netherlands, 1994.

[13] Korth, H. F., and Silberschatz, A., Database System Concepts, McGraw-Hill, 1991.

[14] Mehta, M., Agrawal, R., and Rissanen, J., "SLIQ : A fast Scalable Classifier for Data Mining", EDBT, 1996.