다차원 모델링 기반의 거래분석 시스템 설계 및 구현

이성운*, 최진영*
*고려대학교 컴퓨터정보통신대학원 소프트웨어공학과
e-mail:unius1004@paran.com

Design and Implementation of Trading Analysis System based on Multi-Dimensional Modeling

Sung-Wun Lee*, Jin-Young Choi*
*Dept of Software Engineering, Korea University

요 약

한국증권선물거래소의 유가증권 매매체결시스템은 안정적이고 신속한 데이터 처리에 초점을 둔 시스템이다. 인터넷과 HTS(Home Trading System)의 대중화로 인해 대량의 데이터로부터 적시에 정보를 추출하고 분석하고자 하는 요구가 증가하고 있다. 그러나 현재의 통계정보시스템은 이와 같은 요구를 수용하기 어려우며 개발자의 별도 노력이 요구된다. 또한 목표성능에 대한 요구가 매우 높아짐에 따라 시스템 및 어플리케이션의 증설과 개선작업이 빈번하지만 그 효과를 예측하기 어려우며 정량화된 근거자료의 부재로 의사결정을 지연시킨다. 따라서 이와 같은 요구사항들을 해결하기 위해 기존의 통계정보시스템을 활용하고 추가적인 데이터들을 다양한 차원에서 분석 가능하도록 웨어하우스 데이터베이스를 구축하며 성능예측을 위한 요소들을 추출하고 데이터마이닝을 수행하여 의사결정에 도움을 줄 수 있는 다차원 모델링 기반의 거래분석 시스템을 제안한다. 거래분석 시스템의 구축으로 사용자는 웹상에서 적시에 다차원 분석보고서를 생성할 수 있다. 또한 관리자는 외부적 환경변화에 따른 향후 시스템 성능 감소를 예측할 수 있으며 내부적 요인을 제어하여 이를 상쇄할 수 있는 방안을 찾을 수 있게 된다.

1. 서론

1.1 배경

한국증권선물거래소의 유가증권 매매체결시스템은 tightly-coupling 하게 구성된 두 대의 메인프레임을 통해 일평균 500 만건 전후의 데이터를 수신하여 처리하며 2.5배에 이르는 파생 데이터를 만들어 내고 있다.

1990년 말부터 보급된 초고속 인터넷과 HTS(Home Trading System)의 급속한 보급으로 주식투자가 대중화되었으며, 이로 인해 최근 몇 년 사이에 주문건수가 급격하게 증가되었을 뿐만 아니라 성능 및 거래데이터에 대한 정보 요구가 급격히 늘어나고 있다. 따라서 시스템의 증설 시기가 빨라지고 규모도 커지고 있지만 사용자의 요구는 훨씬 빠른 속도로 증가되고 있는 것이 현실이다.

하루 평균 1,000만 건이 넘게 생성되는 거래 데이터에 대한 백업 및 복원은 테이프 장치에 기록, 보관되고 간단한 집계 데이터 정도만 데이터베이스화 하여 관리하고 있다. 이로 인해 빠르게 변하는 사용자들의 성능 요구 및 거래 데이터에 대한 분석이 필요한 시기에 빠르고 정확하게 산출되지 못함에 따라 관리자의 중요한 의사결정이 지연되고 다소 부정확한 데이터를 근거로 시스템 증설 및 소프트웨어 개선이 이루어지고 있다.

1.2. 대용량 데이터 분석을 위한 요구사항

현재 거래 데이터의 분석을 위해서는 테이프 장치로부터 일 단위로 저장된 데이터를 텍스트 형식으로 로드하고 사용자가 요구하는 세부적인 분석을 위해 개발자가 부가적인 노력을 들여야 원하는 자료를 얻을 수 있으며, 따라서 데이터의 수준이나 품질은 제한적일 수밖에 없다.

또한 시스템 증설이나 소프트웨어 개선을 위한 중요한 지표가 되는 각종 성능 자료들은 요약된 정보만 데이터베이스화 하여 일별로 저장되므로 단순한 정보만을 조회할 수 있다. 요약된 데이터의 수준이 매우 낮기 때문에 데이터의 내용에 대한 세부적인 정보나 현상에 대한 원인을 파악하기에는 부족하다. 따라서 의사결정의 지연과 목표성능에 대한 예측치가 많은 차이를 보이게 된다.

사용자들은 대량의 데이터에 대해서 필요할 때 원하는 정보를 빠르고 쉽게 얻기를 바라며 현상에 대해 다양한 각도에서 분석하기를 원한다. 또한 관리자들은 성능에 관한 정확하고 세밀한 정보를 적시에 제공받기를 원하며 많은 비용이 요구되는 시스템 증설 등을 위한 명확한 근거자료와 예측자료를 요구한다.

본 논문에서는 위와 같은 요구사항을 해결할 수 있는 방안을 제시한다.

첫째, 현재의 개별적인 요약 데이터들의 중복을 제거하고 정규화하여 데이터의 불일치를 제거한다. 그리고 항시적이고 빈번하게 요구되는 측정값들을 기초로 다양한 관점을 갖도록 차원을 정의하고 상세 데이터로 드릴다운

(Drill-down) 가능하도록 데이터베이스를 재설계한다.
둘째, 구축된 데이터마트로부터 미리 집계된 큐브를 통해 다차원적인 분석을 빠르게 수행하고 엑셀(Excel)형태로 제공되던 보고서를 웹에서 접근 가능하도록 쉽고 편리하게 구성한다.
셋째, 성능 예측을 위한 중요한 지표로 사용할 수 있도록 관련된 특성과 측정값들을 추출하여 데이터마이닝을 통해 보다 신뢰성 있고 정량적인 예측치를 제공한다.

본 논문의 구성은 다음과 같다. 제 2절에서는 다차원 정보의 분석과 관련된 기존의 연구들에 대해 간략한 설명을 제시한다. 데이터웨어하우스의 배경과 개념을 살펴본다. 또한 활용 측면에서 OLAP (Online Analytical Processing)와 데이터마이닝을 살펴본다. 제 3절에서는 기존의 통계정보시스템을 최대한 활용하면서 다양한 관점에서 분석을 수행하고 예측할 수 있는 거래분석 시스템의 설계 및 구현을 제시한다. 제 4절에서는 본 연구의 효과를 설명하고 한계점과 추가적인 연구방향을 제시한다.

2, 관련 연구

기업은 운영환경에서 대량의 데이터가 시스템 상에 축적되지만 축적된 데이터 그 이상의 부가가치를 더하지 못해 왔으며, 축적된 데이터를 효과적으로 활용할 수 있는 방안을 찾게 되었으며 이로 인하여 데이터웨어하우스라는 개념이 나타났다[1]. 데이터웨어하우스는 의사결정 지원환경의 전단계인 데이터 통합과 관리, 인프라 구축의 측면을 강조한다. 반면 OLAP는 데이터의 접근과 활용, 어플리케이션 구축 측면을 강조한다. 즉, OLAP는 최종 사용자의 분석적 요구사항과 모델링 측면을 다루게 된다. 데이터마이닝은 데이터웨어하우징의 초점이 구축에서 활용 측면으로 급속히 이동하면서 대규모 데이터베이스로부터 정보를 추출하는 문제를 해결하기 위해서 등장하였다[1].

최근에는 데이터의 구축에서 활용단계로 급속히 발전함에 따라 OLAP가 데이터웨어하우징 환경으로 통합되고, OLAP와 상호보완적으로 사용될 수 있는 데이터마이닝 또한 통합되고 있다.

2.1. 데이터웨어하우스와 OLAP

Inmon(1996)은 데이터웨어하우스를 기업의 의사결정 과정을 지원하기 위한 주제 중심적, 통합적, 시간성을 갖는 비휘발성 자료의 집합으로 정의한다[2]. Kelly(1994)는 전사적 데이터웨어하우스를 기업내의 의사결정 지원 어플리케이션들을 위한 정보기반을 제공하는 하나의 통합된 데이터 저장 공간으로 정의하고 있다[3]. 또 Poe(1994)는 운영시스템과 연계하여 의사결정에 효과적으로 사용될 수 있도록 다양한 운영시스템으로부터 추출, 변환, 통합되고 요약된 읽기 전용 데이터베이스로 정의하고 있다[4]. OLAP는 이와 같이 추출 및 변환과정을 통해 주제별로 통합된 대량의 저장 공간에 접근하여 분석하는 방식으로 조재희와 박성진(2000)은 최종사용자가 다차원 정보에 직접 접근하여 대화식으로 정보를 분석하고 의사결정에 활용하는 과정으로 정의하고 있다[1]. 이를 지원하기 위해 다차원 모델을 구축하게 된다.

다차원 모델은 다차원 데이터베이스를 기반으로 큐브 방식으로 표현된다[5]. 큐브는 축과 좌표로 구성되며 축은 차원으로, 좌표는 차원항목으로 이해할 수 있다. 차원은 사용자가 분석하기를 원하는 각각의 관점을 나타낸다.
차원과 차원항목은 일차적으로 사용자의 입장에서 분석목적에 적합하게 설계된다[1]. 차원을 구성하는 항목들의 특성을 나타내는 정보를 애트리뷰트(Attribute) 혹은 프라퍼티(Property)라고 한다.

이와 같은 다차원 모델의 구축 방식과 과정은 각 OLAP 제품마다 다르다. 다차원 자료구조를 직접 지원하는 전용 OLAP서버인 MDDBMS (Multidimensional DBMS)를 통해 다차원모델을 쉽게 구축할 수도 있으며 RDBMS (Relational DBMS)에서도 일반적으로 1개의 사실테이블과 다수의 차원테이블들로 구성되는 '스타스키마'와 '스노우플레이크스키마'가 주요 모델링 방법으로 사용된다[6,7,8]. 스타스키마는 '사실테이블'을 중심으로 '차원테이블'이 둘러싼 형태를 갖는다. 이와 같은 형태의 모델링을 통해 사용자는 쉽게 이해할 수 있으며, 원하는 보고서를 얻기 위해 필요한 조인의 횟수를 줄여 질의에 빠르게 응답할 수 있는 장점을 갖는다.

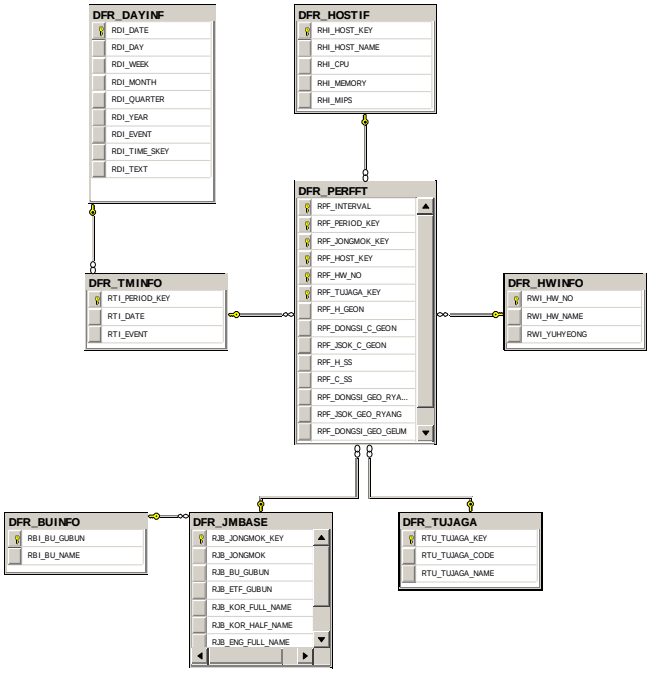
2.2 데이터마이닝

데이터마이닝은 방대한 양의 데이터 속에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 일련의 과정이다[9]. 요즘과 같은 데이터의 홍수 속에서 정보를 얻기란 쉬운 일이 아니다. 이미 인간의 두뇌는 한계를 넘어선 것이다. 이에 컴퓨터로 하여금 방대한 데이터 속에 숨겨진 정보를 발견하는 방법이 필요하게 되었으며 그것이 바로 데이터마이닝이다[10]. 예를 들면, 매출액에 영향을 미치는 요소들을 자동적으로 추출해주게 되며 추정이나 회귀분석 등의 작업을 통해 예측을 가능하게 한다[1].

일반적으로 사용되는 마이닝 기법들로는 연관성탐사(Association), 군집탐사(Clustering), 의사결정수(Decision Tree), 신경망모형(Neural network), 연속성탐사(Sequence) 등이 있다. 연관성탐사와 연속성탐사는 거래기록 데이터로부터 상품간의 연관성과 시간개념을 포함한 연속적 구매패턴 등을 발견하여 연관성 있는 상품들을 그룹화하여 타겟마케팅이나 시장바구니분석(Market Basket Analysis) 문제에 활용될 수 있다. 또한 의사결정수나 신경망 모형은 결과변수에 영향을 주는 요소들을 찾아내고 상호관계를 파악하여 이를 근거로 미래를 예측하는데 활용될 수 있다.

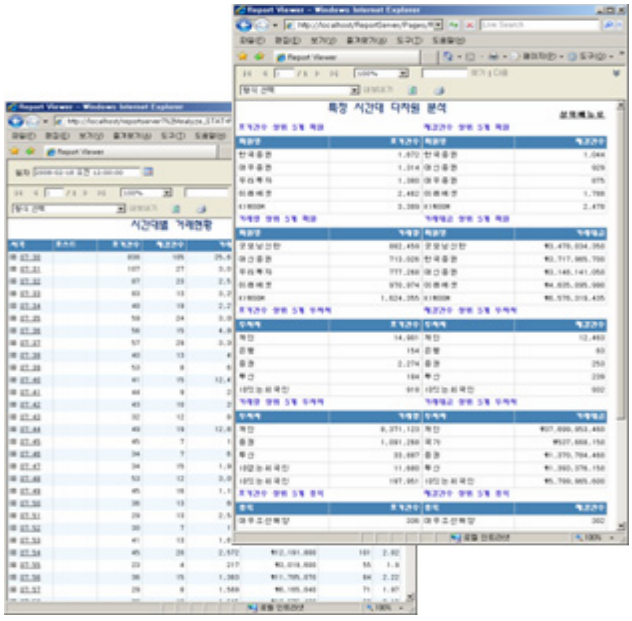
3. 다차원 모델링 기반의 거래분석 시스템 설계 및 구현

본 논문에서 설계된 거래분석 시스템은 초기에는 거래 및 성능 데이터를 시간대, 종목, 시스템, 회원, 투자자의 5개 관점에서 분석할 수 있도록 스타스키마 형태로 설계하였다. 하지만 기존의 집계테이블들을 활용 하면서 다차원 분석이 가능하도록 하기위해서 차원테이블들과 외래키 관계를 가지도록 하여 (그림 1)과 같이 스노우플레이크 스키마를 혼용하였다. DBMS로는 마이크로소프트의 SQL Server 2005 Enterprise 평가판을 사용하였다.



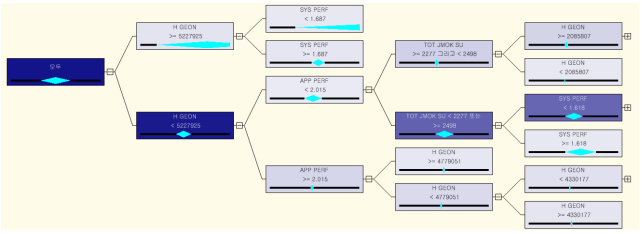
(그림 1) 스노우플레이크 스키마가 혼용된 거래분석 개체-관계 다이어그램

매매체결시스템으로부터 웨어하우스 데이터베이스로의 데이터 추출 및 로딩은 매매체결시스템의 보안상 이유 등으로 ODBC(Open Database Connectivity) 등의 미들웨어가 지원되지 않아 클라이언트/서버 방식으로 구성하였다. 다수의 클라이언트 프로세스들은 당일 주식시장 종료 후 생성된 거래기록들을 처리하여 하나의 멀티스레딩 서버로 전송한다. 멀티쓰레딩 서버는 DBMS에 데이터를 저장하고 SQL Server의 스케쥴링 작업을 통해 다차원 큐브를 구성하도록 구현하였다. 사용자에게 제공되는 분석리포트는 마이크로소프트의 Reporting Service를 이용하여 다차원 거래분석 및 성능분석을 수행할 수 있도록 설계하였다. 따라서 사용자는 웹상에서 보고서간 상호연결을 통해 상세 데이터로 드릴다운(Drill-down)이 가능하다. 즉, 웹상에서 특정 요약데이터로부터 분석을 원하는 차원의 상세 데이터로 직접 접근할 수 있게 된다. (그림 2)는 시간대별 거래현황 보고서에서 특정 시간대 다차원 분석 보고서로 이동하여 분석을 수행하는 화면이다.



(그림 2) 드릴다운(Drill-down)을 통한 상세데이터 접근

미래의 성능을 예측하기 위해 마이닝 기법들 중에서 의사결정트리(Decision Tree)기법을 사용하였다. 성능에 영향을 주는 요소들을 추출하고 상관관계를 파악하여 시스템의 현재 성능분석 뿐만 아니라 향후 시스템 증설 및 어플리케이션 개선 등에 따른 효과도 예측 가능 하도록 데이터마이닝을 수행하였다. (그림 3)는 마이닝을 통해 생성한 의사결정트리를 보여준다.



(그림 3) 의사결정트리

마이닝 과정을 통해 성능(Turnaround Time)에 영향을 주는 외부적 요인으로 호가건수와 종목수를 발견하였다. 추가적으로 시스템 증설과 애플리케이션 개선에 따른 과거 경험치를 수치화하여 내부적 요인으로 제공하고 이를 기반으로 모델링을 수행하였다. 시스템과 애플리케이션 성능 지수는 현재의 통계정보시스템이 가동된 2006년 6월을 기준으로 하였다. 이후 3번의 시스템의 증설 및 애플리케이션 개선시의 효과를 과거 성능 데이터를 분석하고 개발자의 경험을 고려하여 수치화 하였다.

생성된 모델을 통해 사용자는 제어할 수 없는 외부적 요인의 변화로 인한 현재 시스템의 성능 저하를 예측할 수 있다. 또한 이를 상쇄하기 위해 필요한 하드웨어적, 소

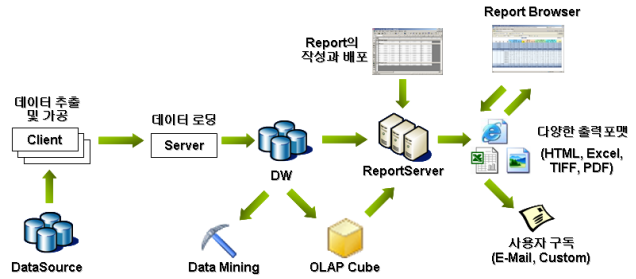
프트웨어적 노력 정도를 판단할 수 있고 이에 따른 성능 개선 효과도 예측할 수 있게 된다. <표 1>은 이와 같은 상황을 시뮬레이션 한 결과를 요약한 표이다. 시스템의 현재 성능에서 주문건수가 65% 정도 증가할 때 1.8배 정도의 성능저하를 예측할 수 있으며, 이를 상쇄하기 위한 하드웨어적 노력정도(시스템 성능지수), 소프트웨적 노력 정도(응용성능지수)를 1.5~2배로 조절할 때의 성능 예측치를 보여준다. 이와 같은 성능 예측치를 활용하여 관리자는 비용을 고려하여 성능 개선의 방향을 결정할 수 있게 되며 정량적인 예측을 할 수 있다.

<표1> 예측쿼리 실행 결과

내부적 요인		외부적 요인		성능예측 (TAT)	비고
시스템 성능지수	응용 성능지수	주문 건수	종목 수		
2.03	2.45	4500000	3000	3.578483	현재성능
2.03	2.45	7000000	3000	6.350174	외부적 요인 변화에 따른 성능 예측
4.06(2배)	2.45	7000000	3000	3.361113	내부적 요인 변화에 따른 성능개선 예측
2.03	3.68(1.5배)	7000000	3000	4.612990	
4.06(2배)	3.68(1.5배)	7000000	3000	1.623929	
2.03	4.9(2배)	7000000	3000	2.889930	

4. 결론

지금까지 본 논문에서는 다차원 모델링 기반의 거래분석 시스템의 설계과정과 구현에 대해 연구하였으며 (그림 4)는 시스템의 전체 구성도를 보여준다.



(그림 4) 거래분석 시스템 구성도

거래분석 시스템의 구축으로 몇 가지 연구결과를 도출할 수 있었다.

첫째, 기존 데이터베이스를 정규화하고 다차원 모델링을 통해 다양한 관점에서 상세 데이터로 드릴다운(Drill-down) 가능하게 되었다.

둘째, 미리 집계된 큐브를 통해 사용자가 웹을 통해 빠르고 편리하게 조회할 수 있게 되었으며 다양한 출력 형식을 지원하여 개발자의 도움 없이 원하는 형태의 보고서를 작성할 수 있으므로 엔드 유저 컴퓨팅(End-User Computing)을 실현하였다.

셋째, 외부적 요인의 변화에 따른 현재 시스템의 성능을 정량적으로 예측하고 이를 바탕으로 시스템이 목표 성능에 도달하기 위해 요구되는 내부적 요인을 판단할 수 있다.

본 연구의 한계점중 하나는 성능예측을 위하여 내부적 요인으로 제공한 시스템 및 어플리케이션 성능지수 간의 상관관계를 간과하고 있는 점이다. 시스템의 상황을 파악하기 위한 여러 지표들에 관한 데이터의 확보가 용이하지 못하여 내부적인 요인들을 각기 독립적인 요인으로 전제하고 마이닝을 수행하였다.

향후 CPU사용률, I/O률 등의 데이터들을 확보하여 시스템 및 어플리케이션 개선지수 간의 상관관계를 파악하고 어플리케이션 개선이 시스템 성능에 미치는 영향을 반영하여 데이터마이닝을 수행한다면 좀 더 정확한 성능예측이 가능할 것이다.

참고문헌

[1] 조재희 외, "OLAP 테크놀로지", 시그마컨설팅그룹, 2000

[2] Inmon, W.H., "Building the Data Warehouse(2nd Ed.)",John Wiley & Sons, Inc., 1996

[3] Kelly, S., Data Warehousing : The Route to Mass Customization, John Wiley & Sons, 1994

[4] Poe, V., Guidelines for Warehouse Development, Database Programming & design, September 1994

[5] Crandall, Richard L., "Multi-Dimensionality in a Decision Support System", White Paper, 1983

[6] Kimball, Ralph, "A Dimensional Modeling Manifesto", DBMS, 1997

[7] Youness, Sakhr, Professional Data Warehousing with SQL Server 7.0 and OLAP Service, Wrox Press, Birmingham (UK), 2000

[8] Red Brick System, "Star Schemas and STARjoin Technology", white paper, 1996

[9] 강현철, 한상태 외 3명, "데이터마이닝 방법론 및 활용", 자유아카데미, 2002

[10] Kirt Thearling, "Increasing customer value by integrating data mining with campaign management software", Exchange Applications Inc.