한국지능정보시스템학회논문지 제7권 제1호 2001년 6월(pp. 97~121)

기업과 소비자간 전자상거래에서의 웹 마이닝을 이용한 상품관리

임광혁
한국과학기술원 산업공학과
(gunni@major.kaist.ac.kr)
홍한국
동의대학교 경영정보학과 전임강사
(honghk@hyomin.dongeui.ac.kr)
박상찬
한국과학기술원 산업공학과 부교수
(sangpark@kaist.ac.kr)

본 연구에서는 웹 마이닝을 이용하여 기업과 소비자간 전자상거래(Business-To-Customer Electronic Commerce) 환경에 기초한 가상상점(Cyber market)의 상품 관리자 입장에서 효율적인 상품관리를 가능케 하는 시스템적 접근방법을 통한 상품관리 방법론을 제시하고자 한다. 또한 이 상품 관리 방법론을 실제 웹 상에서 운영되고 있는 가상상점에 직접 적용하여 봄으로써 실증적인 예를 보여주고자 한다.

1. 서 론

상품관리는 인기 상품과 비 인기 상품을 구분하고, 필요한 상품과 물량을 결정하며, 사업자를 선정하고, 상품을 분배하는 활동을 포함한다. 상품 판매를 기초로 하고 있는 소매(Retail) 산업에 있어서 상품관리는 가장 중요한 활동중의 하나이다. 현재, 물리적 상점을 통한 소매 산업에서는 상품관리를 상품관리자라는 사람에 의존하고 있다. 그 이유는 상품판매 동향 및 고객의 구매 성향에 대한 정보의 제약으로 인하여 시스템적으로 자동화하지 못하고 상품관리자의 경험에 의한 접근을 위주로 진행하고 있기 때문이다. 그러나, 이러한 물리적 상점에 비교하여 가상 상점은 상거래가 이루어 지고 있는 모든 정보의 검색이나 의사결정 과정을 정보 기술에 의해 지원 받을 수 있다는 장점을 가지고 있다. 즉, 고객의 상점이용 상태와 구매성향, 상품판매 동향이 디지털화 된 자료로 제공되므로 이를 이용하여 상품관리를 체계적으로 수행할 수 있다.

웹 상에 존재하는 디지털화 된 자료를 이용하여 가장 효율적으로 필요한 자료를 추출해내고자 하는 연구 분야가 웹 마이닝이다. 가상 상점에 축적되어 있는 디지털화 된 자료를 웹 마이닝 기법을 이용하여 분석하면, 기존 데이터 베이스 자료만을 이용하는 다른 분석 기법들로는 추출할 수 없었던 유용한 정보를 추출해 낼 수 있다. 웹 마이닝은 어떻게 하면 웹 상에 존재하는 데이터를 효율적으로 분석하여 고객의 요구에 맞는 자료를 제공하는가를 연구하는 웹 내용 마이닝

(Web Content Mining) 분야와 웹 서버 로그(Web server logs), 조회 로그(referral logs), 등록 데이터(Registration data) 등을 이용하여 사용자 경로 분석(path analysis), 가상상점 이용 패턴 분석(pattern analysis)를 분석하는 웹 사용 마이닝(Web Usage Mining) 분야로 구분할 수 있다.

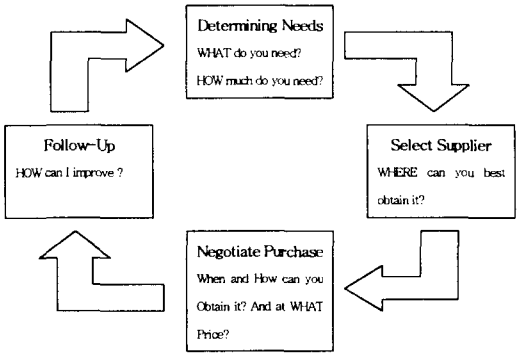
가상 상점에 관련된 지금까지의 연구는 주로 고객을 위한 지능화된 서비스와 구매 의사 결정 지원을 중심으로 한 웹 내용 마이닝을 중심으로 진행되고 있다. 즉, 가상 공간상에 산재 되어 있는 자료를 효율적으로 검색하고 분석하여 어떻게 하면 고객의 요구에 가장 잘 맞는 자료를 제공해 줄 수 있는 가에 관련된 연구가 활발하게 진행되고 있다. 이에 반하여, 가상상점 관리자 입장에서 상품의 효율적인 관리를 위한 체계적인 분석 기법과 관련된 연구는 미흡한 상태이다. 이러한 연구는 웹을 사용하는 사용자의 경로 분석, 이용패턴 분석을 포함하는 웹 사용 마이닝에 기초한다. 가상상점은 물리적 상점에 비교하여 고객의 구매 이력과 행동, 상품의 판매동향을 모니터링할 수 있으므로 관리자에서 매우 큰 장점을 가져다 줄 수 있다. 시장 선점이 가속화되고 있는 현 시점에서, 경쟁에서 살아 남기 위해서는 차별화 전략이 필요하며 그 해법은 효율적인 상품관리로부터 시작된다. 그러므로 지금은 효율적인 상품관리를 위한 체계적인 분석 기법과 관련된 연구가 필요한 시점이다.

본 연구에서는 웹 마이닝을 이용하여 기업과 소비자간 전자상거래 환경에 기초한 가상 상점의 상품 관리자 입장에서 효율적인 상품관리를 가능케 하는 시스템적 접근방법을 통한 상품관리 방법론을 제시하고자 한다. 또한 이 상품 관리 방법론을 실제 웹 상에서 운영되고 있는 가상상점에 직접 적용하여 봄으로써 실증적인 예를 보여주고자 한다.

2. 기존 연구

2.1 상품 관리(Merchandise Management)

소매 비즈니스 프로세스는 크게 상품 관리(merchandise management), 상품처리 관리(merchandise process management), 매장 관리(store management)와 같이 세 개의 그룹으로 나누어 진다[Lee, W. J., 1998]. 그 중에서 상품관리는 Mason, et al.에 의해 제안된 구매 사이클에 의해 잘 표현된다[Mason, et al., 1991]. 구매 사이클은 <그림 1>에 보여지는 것과 같이 4개의 프로세스로 구성되어 있다.



<그림 1> 구매 사이클(Buying Cycle)

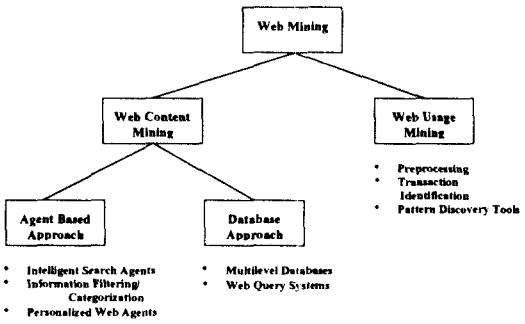
■ 필요 상품 및 물량 결정(Determining Needs) : 이 단계는 인기 상품과 비인기 상품을 구분하고, 필요한 상품과 물량을 결정하여 재고 물량을 가장 가능한 낮은 레벨로 유지하는 것을 목표로 한다. 그러나 고객이 선택할 수 있는

색상, 사이즈, 모델은 가능한 한 다양하게 유지해야 한다.

- 제공 업체 선정(Select Supplier) : 상품의 종류와 물량을 결정한 후에 상품을 제공받을 수 있는 업체를 검색하고, 검색한 업체 중에서 품질과 가격을 고려하였을 때 가장 좋은 조건의 업체를 선정해야 한다.
- 구매 협상(Negotiate Purchase) : 이 단계에서 바이어는 최적의 거래를 위해 상품들의 가격, 할인율, 배달기간, 배달비용 등을 협상해야 한다.
- 지속적인 조사 및 향상 (Follow-Up) : 더 필요한 업체의 검색, 필요한 상품과 물량의 결정, 상품 구매와 상품 처리 활동 등을 계속적으로 조사하여 더 좋은 상품관리로의 향상을 도모하는 활동을 의미한다.

2.2 웹 마이닝(Web Mining)

웹 마이닝은 대체적으로 World Wide Web으로부터 유용한 정보를 발견하고 분석하는 것으로서 정의할 수 있다. <그림 2>에서 살펴볼 수 있듯이 웹 마이닝은 어떻게 하면 웹 상에 존재하는 데이터를 효율적으로 분석하여 고객의 요구에 맞는 자료를 제공하는가를 연구하는 웹 내용 마이닝(Web Content Mining) 분야와 웹 서버 로그(Web server logs), 조회 로그(referral logs), 등록 데이터(Registration data) 등을 이용하여 사용자 경로분석(path analysis), 사용자이용 패턴 분석(pattern analysis)를 중심으로 하는 웹 사용 마이닝(Web Usage Mining) 분야로 나눌 수 있다[R. Cooley, B. Mobasher, and J. Srivastava, 1997].



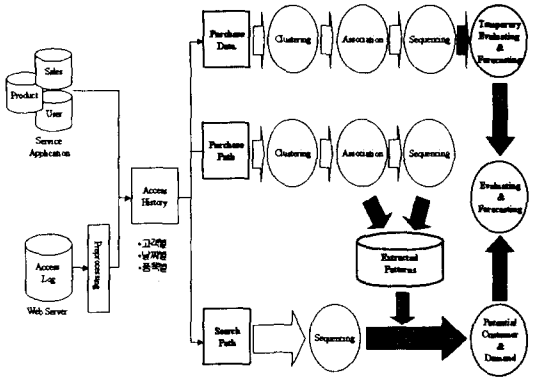
<그림 2> 웹 마이닝의 분류

웹 내용 마이닝은 야후, 알타비스타, 라이코스와 같은 전통적인 검색 엔진들이 제공하지 못하는 구조화된 정보를 제공할 수 있도록 하기 위해 두 가지 접근 방식을 취하고 있다. 첫 번째 접근 방식은 에이전트-기반 접근 방식(Agent-based Approach)으로서 인공지능 웹 에이전트와 같은 더 지능적인 툴을 개발하여 구조화된 정보를 제공하고자 하는 접근 방식이다. 또 다른 접근 방식은 데이터베이스 접근방식(Database Approach)로서 웹 상에서 사용 가능한 정보의 구조를 더 높게 하기 위하여 데이터 마이닝 기술을 확장한 접근 방식이다.

웹 사용 마이닝은 웹 로그에서 불필요한 정보를 제거하고 필요한 정보만을 추출하는 사전처리(Preprocessing)와 다양한 로긴(login) 정보에서 한 사용자의 트랜잭션을 규정하는 트랜잭션 검증(Transaction Identification)을 수행한다. 이 과정이 끝나면, 로그 데이터 상에 나타나는 사용자의 검색 또는 구매 패턴을 발견하는 패턴 발견(Pattern Discovery)을 수행하게 되며, 발견된 패턴을 사용 목적에 맞게 분석하는 패턴 분석(Pattern Analysis)을 수행한다.

3. 방법론

3.1 개 요



<그림 3> Framework of Analysis

본 연구에서 구현하고자 하는 상품관리 시스템(Merchandise Management System Using Web Mining: MeMSWeM)은 웹 마이닝 기법을 이용하여 가상 상점의 고객과 상품의 검색/구매 패턴을 분석하고, 이 결과를 이용하여 상품관리의 내부적 기능인 수요 예측(Demand Forecasting)과 평가 및 선정(Evaluating & Selecting) 기능을 수행하는 시스템이다.

웹 마이닝의 최대 장점은 기존의 분석 기법과 달리 가상상점을 이용하는 고객의 검색/구매 경로, 상품의 검색/구매 패턴에 대한 웹 서버 로그 정보를 이용할 수 있다는 점이다. 이 시스템은 이러한 웹 마이닝 기법의 장점을 최대한 이용하여 고객의 검색/구매 경로에 대한 체계적인 분석을 통한 상품 간의 상관 관계와 선후 관계를 추출하고, 또한 잠재적 구매 고객과 상품을 추출함으로써 인기 상품과 비인기 상품을 구별하고, 가까운 미래의 판매 경향을 예측하는 것을 목적으로 한다.

분석을 수행하기에 앞서 우선, 웹 로그에서 불필요한 데이터를 선택하여 제거하는 사전처리(Preprocessing)가 선행되어야 한다. 사전처리는 웹 로그를 검색하여 에러가 발생한 페이지 데이터 및 이미지 관련 데이터를 추출하여 제거하게 되며 이렇게 정제된 웹 로그 데이터와 데이터베이스 자료를 이용하여 사용자의 사용 기록(Access History)을 작성하게 된다.

사용 기록은 웹 로그 데이터와는 구별되게 사용자가 검색한 웹 페이지를 해당 상품 코드로 변환하여 관리하고, 각 상품을 사용자가 검색만 수행하였는지 그 검색이 구매까지 도달하였는지를 나타내는 구매/검색 구분 필드를 포함하게 된다. 즉, 사용자의 사용 데이터를 검색만 이루어진 검색경로와 검색이 이루어 짐과 동시에 구매까지 발생한 구매경로를 구분하여 관리하게 된다. 또한 필요에 따라 사용 데이터를 고객별, 날짜별, 시간별, 품목별로 구분하여 검색이 가능하다.

이렇게 작성된 사용기록을 이용하여 상품관리를 위한 분석을 수행하게 된다. 분석은 사용기록을 구매데이터(Purchase Data), 구매경로(Purchase Path), 검색경로(Search Path)로 나누어 수행하게 된다. 구매데이터는 고객이 구매한 상품정보를 포함하고, 구매경로는 고객이 구매하기 까지 상점을 이용한 이용 경로를 포함하며, 검색경로는 아직 구매가 이루어지지 않은 고객의 상점이용 패턴을 포함한다. 각 경로를 거치면서 다음과 같은 분석이 수행된다.

첫째, 구매데이터를 체계적으로 분석하여 기존 구매고객과 상품판매 경향을 추출함으로써 상품관리를 위한 잠정적인 평가 및 예측(Temporary Evaluating & Forecasting)을 수행한다. 분석 기

법으로는 클러스터링(Clustering), 연관성 분석(Association)과 연속성 분석(Sequencing)을 이용한다. 클러스터링은 상품과 고객간의 연관성을 추출하는 기능을 수행하며, 그 결과는 차후에 DM, 공동 마케팅, 상점상품 재배열에 이용될 수 있다. 연관성 분석은 클러스터링과는 달리 상품과 상품간의 연관성을 추출하며, 이러한 상품간의 관계는 시간을 고려하지 않은 정적인 관계성을 나타낸다. 연속성 분석은 연관성 분석과 같이 상품과 상품간의 연관성을 추출하지만, 연관성 분석과 달리 시간을 고려한 동적인 관계성을 추출한다. 잠정적인 평가 및 예측을 수행하기 위해서는 연속성 분석과 연관성 분석 결과를 이용하게 된다. 우선, 연속성 분석 결과에 의한 차후 구매가능 상품과 물량을 결정하며 선정된 상품과 정적인 연관성이 높은 상품을 연계하여 선정함으로써, 인기 상품을 선정하고 물량을 선정할 수 있다.

둘째, 고객의 구매경로를 체계적으로 분석하여 기존 구매고객의 상품 검색 패턴을 추출하여 패턴 데이터베이스에 저장한다. 분석 기법으로는 구매데이터를 이용한 분석과 마찬가지로 클러스터링, 연관성 분석과 연속성 분석을 이용한다. 이 단계에서의 클러스터링은 잠재적 상품과 고객간의 연관성을 추출하는 기능을 수행하고, 연관성 분석은 상품과 상품간의 잠재적 연관성을 추출하며, 이러한 상품간의 관계는 시간을 고려하지 않은 정적인 관계성을 나타낸다. 연속성 분석도 상품과 상품간의 잠재적 연관성을 추출하며, 이러한 상품간의 관계는 시간을 고려한 동적인 관계성을 나타낸다. 연속성 분석과 연관성 분석결과를 이용하여 추출된 상품과 상품간의 정적, 동적 패턴은 차후 잠재적 고객과 요구를 분석하기 위해 사용될 수 있으므로 패턴 데이터베이스에 저장되어 관리된다.

셋째, 고객의 검색 경로를 체계적으로 분석하여 패턴 데이터베이스에 저장된 패턴을 소유한 잠재적 고객과 요구를 파악한다. 분석 기법으로는 연속성 분석을 이용한다. 연속성 분석을 통하여 고객의 검색 패턴을 추출하고, 추출된 패턴이 패턴 데이터베이스에 저장된 패턴과 일치하면 이 고객을 잠재적 고객으로 선정할 수 있다. 선정된 잠재 고객의 연속 패턴과 일치한 패턴 데이터베이스의 패턴 정보를 이용하면, 해당 패턴을 소유한 상품을 추출해 낼 수 있다. 바로 이 상품이 잠재 고객이 차후에 구매할 가능성이 있는 잠재 구매상품이 되고, 잠재 구매 물량도 결정할 수 있다. 이렇게 잠재 구매상품이 선정되면, 이를 통하여 이 상품과 연관성이 높은 상품을 잠재 구매 상품으로 추가할 수 있다.

넷째, 잠정적인 평가 및 예측 결과와 잠재적 구매상품과 물량 예측을 통합하여 최종 평가 및 예측을 수행한다. 잠정적인 평가 및 예측에서 선정된 상품과 잠재적 구매 상품을 통합하여 최종적으로 재고로 관리해야 할 상품을 선정하고, 잠정적인 평가 및 예측에서 결정한 상품 별 물량과 잠재적 구매 상품별 물량을 종합하여 최종 재고로 관리해야 할 상품별 물량을 결정한다.

3.2 방법론과 도구 (Methodology & Tools)

3.2.1 사전처리(Preprocessing) : Data Cleaning and Transformation

Common Logfile Format(CLF)은 NCSA 계열의 웹 서버에서 사용하는 파일형식으로, 현재 대부분의 웹 서버에서 지원하고 있다. CLF는 <그림 4>와 같은 구조로 되어 있으며 다음과 같은 정보를 포함하고 있다.

<그림 4> CLF의 구조

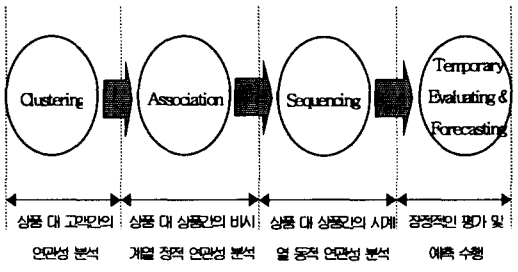
<그림 4>의 로그 데이터는 다음과 같은 사실을 기록하고 있다. 157.55.85.138 이라는 IP주소를 가진 사용자가 doug라는 이름으로 1996년 6월 7일 오후 5시 39분 4초, 그리니치 표준시로부터 8시간 30분 떨어진 곳에서 POST의 방법으로 iisadmin/ default.htm을 요청하였으며, 이는 HTTP 버전 1.0 프로토콜에 의해 성공적으로(200) 이루어졌고 이동한 총 데이터량은 3,401Byte이다. <그림 4>와 같은 형식의 로그 파일의 데이터를 분석에 이용할 수 있는 분석 데이터로 사용하기 위해서, 우리는 먼저 원시 데이터를 정리(clean) 하고 변환(transform) 해야 한다.

데이터 정리와 변환은 크게 두 가지의 프로세스로 구성되어 있다[Osmar R., 1998]. 첫번째 프로세스는 로그 파일의 모든 엔트리를 하나도 빠짐없이 분석데이터로 매핑하는 것이다. 즉, 이 과정에서는 기록된 서버 요청 실패, 입증 실패, 페이지 그래픽 등 어떠한 데이터도 잃어버리는 것은 없다는 것이다. 모든 엔트리는 분석의 목적과 질문의 종류에 따라서 잠재적인 유용성이 있다고 믿는다. 일단 이렇게 모든 정보를 분석데이터로 만들었다면, 우리는 항상 분석의 필요성에 따라 적절한 정보를 추출하여 사용할 수 있다. 첫번째 프로세스가 끝난 변환된 데이터 엔트리는 기존 필드 이외에 1 개의 필드가 더 포함되어 있다. 상품코드 필드는 사용자가 검색한 페이지에 전시되고 있는 상품의 코드를 나타낸다.

두번째 프로세스는 사용자의 행동을 추론하는 것이다. 이 프로세스에서 우리는 로그 데이터와 변환된 데이터들의 집합을 사용자 행동의 집합으로 매핑한다. 즉, 사용자가 수행한 CGI Script명과 각종 파라미터를 이용하여 사용자의 행동을 간단한 표시(flag)를 이용하여 나타낸다. 이것이 더 자세한 사용자의 행태를 분석하기 위한 가장 기초적인 작업이다. 두 번째 프로세스 후에 변환된 데이터 엔트리는 기존 필드이외에 1 개의 필드가 더 포함되어 있다. 구매/검색(Purchase/Search) 구분 필드는 지금 사용자가 검색한 상품을 검색만 하는 것인지, 아니면 구매까지 하였는지를 구분하는 필드이며, 이는 차후에 구매 경로를 추출할 때 이용된다.

3.2.2 잠정적인 평가 및 예측 : 구매 데이터를 이용하여

잠정적인 평가 및 예측 과정은 고객과 상품의 연관성을 추출하는 클러스터링, 상품과 상품의 정적인 연관성을 추출하는 연관성 분석과 상품과 상품의 동적인 연관성을 추출하는 연속성 분석을 수행함으로써, 궁극적으로 상품관리를 위한 잠정적인 재고 상품과 물량을 설정하는 것을 목적으로 한다. 전체적인 잠정적인 평가 및 예측을 수행하는 분석과정은 <그림 5>와 같으며, 각 분석 단계는 다음과 같다.



<그림 5> 잠정적인 평가 및 예측 분석 과정

3.2.2.1 상품 대 고객간의 연관성 분석 : 클러스터링(Clustering)

클러스터링은 전체 상품군을 고객들이 비슷하게 구매하는 상품들로 그룹화하게 된다. 클러스터링을 수행하기 위해서는 사전처리된 데이터를 직접 이용할 수 없기 때문에 클러스터링을 수행하기 전에 클러스터링 데이터를 만드는 작업이 수행되어야 한다. 즉, 상품을 고객이 구매하였으면 "1", 구매하지 않았으면 "0" 을 할당하여 상품 대 구매 고객간의 행렬을 구성하게 된다. 이러한 클러스터링을 수행한 결과에서 각각의 클러스터는 고객 대 구매상품의 연관성을 나타낸다. 즉, 고객들이 상품을 구입할 때 비슷한 구매 성향을 가진 고객들이 구매하는 상품들이 하나의 클러스터로 구분되게 된다. 그러므로 같은 클러스터로 분류되는 상품군들은 구매 고객들이 비슷한 계층이므로 하나의 그룹군으로 파악하여 공동 마케팅, DM 발송, 같은 매장 내 상품진열과 같은 이벤트를 공동으로 수행할 수 있다.

클러스터링을 위한 툴로는 SOM(Self-Organization Map)을 이용한다. SOM은 신경회로망(Neural Network)의 일종으로 Kohonen에 의해 일반화되었기 때문에 Kohonen net 이라고도 한다[Sabrina Sestito, Tharam S Dillon., 1994]. 신경회로망의 중요한 특징 중의 하나는 환경으로부터 학습을 하고, 학습을 통하여 수행능력을 개선시키는 능력을 가지고 있다는 것이다. SOM은 외부의 피드백이나 지도가 없이 스스로 학습하여 입력자료에서 의미 있는 패턴이나 특징을 발견하는 시스템이다. 기업 조직의 목표를 달성하기 위해 여러 변수를 한꺼번에 고려하면서 중요한 고객군/상품군과 비슷한 고객/상품들을 선정할 때 적합한 시스템이다.

3.2.2.2 상품 대 상품간의 비시계열 정적 연관성 분석 : Mining Association Rules

클러스터링을 통해 상품과 고객간의 연관성을 확인했다면, 여기서 수행하는 연관성 분석은 상품과 상품간의 연관성을 추출한다. 상품과 상품간의 연관성을 추출하기 위해서는 상품과 상품간의 상관계수(Correlation)를 이용한 상관계수 행렬(correlation matrix)을 이용할 수 있다. 또한 전체 상품군에 대한 상관 행렬을 구성하기는 매우 어려우므로 클러스터링 결과를 이용하여 비슷한 고객 성향을 나타내는 상품들 중심으로 상관 계수 행렬을 구성한다. A 상품과 B 상품의 상관계수는 다음과 같이 정의할 수 있다. *correlation(A, B)* = (A상품 구매고객 중 B상품 구매한 고객 수) / (A상품 구매고객 수) 즉, 상관계수가 높게 나타난다는 것은 A상품을 산 고객이 B상품을 구매할 확률이 매우 높다는 것을 나타내므로 A, B상품의 관계성이 매우 높다는 것을 나타낸다. 이렇게 상관계수 행렬을 통한 연관성을 추출한 결과는 다음과 같이 이용할 수 있다.

첫째, 관계성 있는 상품간의 재고량을 조절할 수 있으므로 재고관리를 포함한 상품관리에 이용할 수 있다. 즉, A상품과 B상품이 매우 높은 상관계수를 가지고 있다면, 재고상품 선정에 있어서 A, B상품 모두를 선택하여야 하며, 재고물량을 결정할 때도 A, B상품의 재고물량을 비슷하게 유지하여야 한다는 정보를 추출할 수 있다. 이러한 정보를 이용하여 재고, 상품관리를 수행할 수 있으므로, 재고비용을 줄일 수 있고, 재고물량 확보를 통한 고객까지의 배송시간을 단축할 수 있으므로 고객 만족도를 향상 시킬 수 있다.

둘째, A상품을 구매한 고객이 있다면, 상관계수가 매우 높은 B상품에 대한 정보를 DM(Direct

Mailing)을 통하여 고객에게 제공할 수 있다. A상품을 구매한 고객은 B상품에 대한 관심이 높을 가능성이 매우 크므로 B상품 정보를 입수한 고객은 B상품을 구매할 확률이 상대적으로 높아지게 된다. 그러므로 연관성 분석 결과를 고객에 대한 DM에 이용하여 고객서비스를 향상시킬 수 있을 뿐 아니라 매출을 촉진할 수 있다.

셋째, A상품을 구매한 고객은 B상품을 구매할 확률이 높다는 연관성 분석 결과를 이용하여 A상품과 B상품의 진열을 같은 페이지에 함으로써 고객의 관심을 유도할 수 있다. 즉, A상품을 구매한 고객이 다시 B상품을 찾기 위해 가상 시장을 검색할 필요 없이 바로 옆에 B상품이 진열되어 있다면, 고객의 구매 확률은 상대적으로 증가할 것이다.

넷째, 연관성 분석을 이용하여 상품간의 관계성을 파악할 수 있으므로 이러한 관계성을 데이터베이스로 관리함으로써 조기경고시스템을 구축할 수 있다. 즉, 새로운 관계성이 발생하거나 또는 지금까지 관리해 온 관계성이 의미가 없어지면 이러한 정보를 상품관리자에게 제공해 줌으로써 상품관리에 참고할 수 있도록 한다.

3.2.2.3 상품 대 상품간의 시계열 동적 연관성 분석 : Mining Sequence Patterns

클러스터링과 연관성 분석은 시간은 고려하지 않은 정적인 정보를 이용하여 고객 대 상품, 상품 대 상품의 관계성을 규정하는 방법론이다. 그러나 연속 패턴은 상품 대 상품의 관계성을 시간을 고려하여 선후관계로서 규정하는 방법론이다. 연속성 분석을 통하여 상품 구매의 선후 관계가 정의되고 구매까지 걸리는 시간을 추출하면 다음과 같은 이점을 얻을 수 있다.

첫째, A상품의 구매량을 추출하면 구매까지 걸리는 시간 이후의 B상품의 구매량을 추정할 수 있다. 즉, A상품과 B상품이 1개월의 기간 사이로 30%의 연속 패턴을 가지고 있다면, A상품이 이번 달에 100개의 매출을 올렸다면, 1개월 후 B상품은 최소한 30개의 매출을 올릴 수 있다는 것을 추정할 수 있다.

둘째, 이러한 추정을 이용하여 재고 물량과 상품관리를 체계적이고 효율적으로 할 수 있다. 즉, 시간적으로 선후 관계가 있는 상품들 간의 재고 상품 선정 및 재고물량을 구매시간의 차이를 두고 관리할 수 있으므로 적정한 수준으로 상품의 재고를 관리할 수 있다.

셋째, 이런 연속 패턴을 이용하여 A상품을 구매한 고객에게 B상품을 구매할 시간이 되면 DM을 통해 상품정보를 제공해 주는 일대일 마케팅(One-To-One Marketing)이 가능하다. 클러스터링과 연관성 분석은 고객군 및 상품군에 대한 분석 기법이지만 연속성 분석은 고객 개개인에 대한 분석 기법이므로 일대일 마케팅이 가능하다. 연관성 규칙(Association Rule)을 추출하고 연속 패턴(Sequence pattern)을 찾아내기 위해 IBM Almaden Research Center의 AprioriAll Algorithm을 사용한다[R. Agrawal and R. Srikant, 1994].

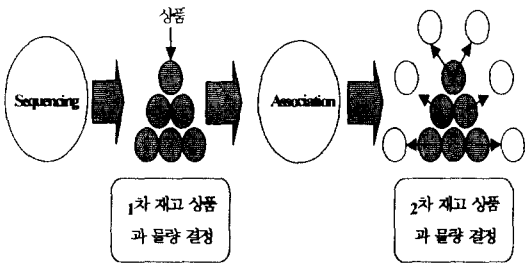
3.2.2.4 잠정적인 평가 및 예측(Temporary Evaluating & Forecasting)

클러스터링을 통한 상품 대 고객간의 연관성, 연관성 분석을 통한 상품 대 상품의 비시계열 연관성 분석, 연속패턴 분석을 통한 상품 대 상품의 시계열 연관성 분석 결과를 이용하여, 최종적으로 상품관리를 위한 수요 예측에서 도출 할 수 있는 결과는 다음과 같다(Stock, J. R. and Lambert, D. M., 1987; Lo, T., 1994).

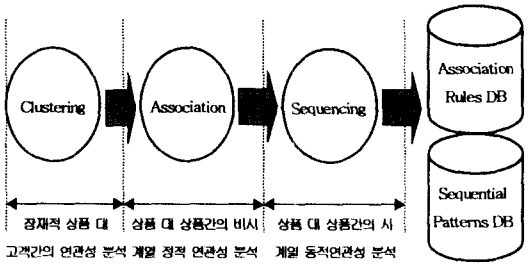
- 연속패턴 분석의 결과를 이용하여 다음 기간에 재고를 확보해야 할 상품과 물량을 결정한다. 즉, 연속패턴에 근거하여 이번 기간에 판매된 상품들 다음에 구매가 이루어 질 상품 리스트와 물량을 결정할 수 있다.
- 연속 패턴 분석을 통해 선정된 상품과 물량을 기초로 하여, 선정된 상품과 연관성이 높은 상품과 상관계수를 추출해 낼 수 있다. 상관계수를 이용하면 관계된 상품의 물량도 결정할 수 있다. 이 데이터를 이용하면 재고로 관리해야 할 최종 상품 리스트와 물량을 구해낼 수 있다.

3.2.3 구매고객의 구매경로 패턴 추출

구매고객의 구매경로 패턴 추출 과정은 고객과 잠재적 상품의 연관성을 추출하는 클러스터링, 상품과 상품의 정적인 연관성을 추출하는 연관성 분석과 상품과 상품의 동적인 연관성을 추출하는 연속성 분석을 수행하여 추출된 패턴을 패턴 데이터베이스에 저장함으로써, 궁극적으로 상품관리를 위한 잠재적 고객과 상품의 요구를 추출하기 위한 기반을 마련하는데 그 목적이 있다. 전체적인 분석과정은 <그림 7>과 같으며, 각 분석단계의 상세내용은 다음과 같다.



<그림 6> 잠정적인 평가 및 예측 절차



<그림 7> 구매고객의 구매경로 패턴 추출

3.2.3.1 잠재적 상품 대 고객간의 연관성 분석 : Clustering

구매 경로를 이용한 클러스터링은 웹 서버 로그의 구매한 고객이 구매하기까지 검색한 상품들의 경로 정보를 이용하여 클러스터링을 수행한다. 즉, 구매 고객이 검색을 하였으면 "1", 검색하지 않았으면 "0"을 할당하여 구매검색 상품 대 고객 구매 경로간의 행렬을 구성하게 된다. 이러한 클러스터링을 수행한 결과에서 각각의 클러스터링은 고객 대 잠재적인 구매 상품의 연관성을 나타낸다. 즉, 구매한 고객들이 비슷하게 검색하는 상품들이 하나의 클러스터로 구분되므로 같은 클러스터로 분류되는 상품군들은 구매 고객들이 비슷하게 검색하는 상품군들이므로 차후에 구매할 확률이 매우 높은 상품군들이다. 그러므로 구매한 상품 주위에 이러한 상품군들을 진열함으로써 고객의 구매를 촉진할 수 있다. 또한 사용자 군별 또한 구매상품 별로 검색하는 상품군이 다르므로 각 사용자 별로 로그인 시 맞춤형 검색경로를 제공해 줄 수 있다. 클러스터링을 위한 툴로는 SOM(Self-Organization Map)을 이용한다.

3.2.3.2 상품 대 상품간의 비시계열 정적 연관성 분석 : Mining Association Rules

이 단계에서의 연관성 분석도 3.2.2.2와 같이 상품과 상품간의 연관성을 추출하기 위해서 상품과 상품간의 상관계수를 이용한 상관계수 행렬(correlation matrix)을 이용할 수 있다. 상관계수가 높게 나타난다는 것은 A상품을 검색한 고객이 B상품을 검색할 확률이 매우 높다는 것을 나타내므로 A, B상품의 관계성이 매우 높다는 것을 나타낸다. 이렇게 상관계수행렬을 통한 연관성을 추출한 결과는 차후에 잠재적 구매 상품과 물량을 결정하는데 이용될 수 있다. 그러므로 연관성 분석 결과는 패턴 데이터베이스에 저장되어 관리되어야 한다.

3.2.3.3 상품 대 상품간의 시계열 동적 연관성 분석 : Mining Sequence Patterns

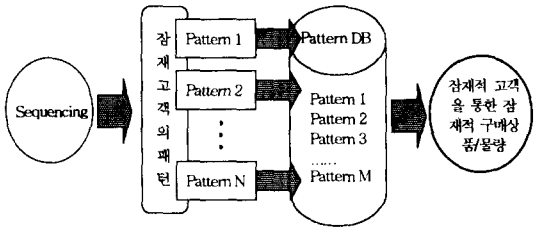
이 단계에서 수행하는 연속성 분석은 고객의 구매 경로 상에 검색되는 상품의 선후관계와 선후관계 사이에 걸리는 시간을 추출한다. 추출된 연속패턴은 패턴 데이터베이스에 저장되어 차후에 고객의 검색경로를 분석하여 추출된 패턴과 비교하여 고객이 잠재적 고객으로 발전가능성이 있는가를 판단하는 자료로 이용된다. 연관성 규칙을 추출하고 연속패턴을 찾아내기 위한 알고리즘으로 AprioriAll Algorithm을 사용한다.

패턴 데이터 베이스는 연관성 분석으로 추출된 상품과 상품간의 정적 연관성 패턴과 연속성 분석으로 추출된 상품과 상품간의 선후관계를 나타내는 동적 패턴을 저장하게 되며, 패턴이 발견될수록 계속적으로 축적하여 관리한다. 축적된 패턴 정보는 고객의 검색경로 패턴과 비교하여 잠재적 고객을 추출하고 잠재적 구매상품을 추출하는 데 이용되며, 차후에는 패턴 정보의 변화를 감지하여 상품관리자에게 변화에 대한 정보를 제공하거나, 경고 메시지를 줄 수 있는 조기 경보 시스템 구축의 기반 데이터를 제공해 줄 수 있다.

3.2.4 검색경로 패턴 추출을 통한 잠재적 고객과 요구 추출 : 검색 경로를 이용하여

3.2.4.1 상품 대 상품간의 시계열 연관성 분석 : Mining Sequence Patterns

이 단계에서 수행하는 연속성 분석은 고객의 검색경로 상에 나타나는 상품의 연속패턴을 추출한다. 이렇게 추출된 패턴과 패턴 데이터베이스에 저장된 패턴과 비교하여 패턴이 일치하면 해당 고객을 잠재적 고객으로 선정하고 해당 패턴에 해당하는 상품을 잠재적 상품으로 선정한다. 모든 고객의 검색패턴에 대하여 이러한 과정을 반복적으로 수행하여 전체 잠재적 고객과 구매상품을 추출해 낸다. 잠재적 고객을 통한 잠재적 구매 상품 및 물량을 추출해 내는 과정은 <그림 8>과 같다.



<그림 8> 잠재적 구매 상품 및 물량 추출

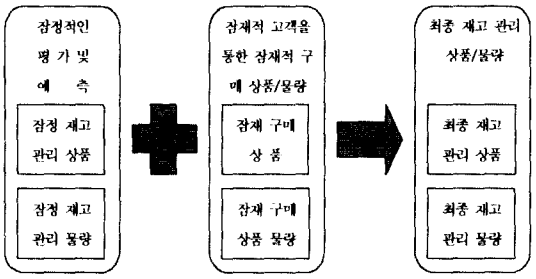
3.2.5 통합 평가 및 예측 : 잠정적인 평가 및 예측, 잠재적 상품/물량 정보를 이용하여

구매 데이터 분석을 통하여 추출한 잠정적인 평가 및 예측 결과와 구매경로와 검색경로 분석을 통하여 추출한 잠재적 구매 상품 및 물량 정보를 종합하여 최종적으로 재고로 관리해야 할

상품과 물량을 결정한다.

- Evaluating(재고로 관리해야 할 상품 결정): 재고로 관리해야 할 상품은 Temporary E & F에서 선정한 상품과 잠재적 구매 상품으로 선정된 상품을 합하여 결정한다. 중복적으로 선정된 상품은 하나로 취급하여 전체적인 상품 목록을 작성한다.
- Forecasting(선정된 상품의 재고 물량 결정): Evaluating에서 선정된 각 상품에 대해서 Temporary E & F에서 결정한 물량과 잠재적 구매 물량을 합하여 결정한다.

여기서 결정된 상품과 물량은 재고로 관리해야 할 상품과 물량의 최저 수준을 제공한다. 즉, 최소한 판매가 이루어질 상품과 물량을 결정한 것이므로 다른 기법을 통한 Evaluating & Forecasting을 수행한 결과를 이용하여 좀 더 많은 상품과 물량을 추가할 수 있을 것이다.

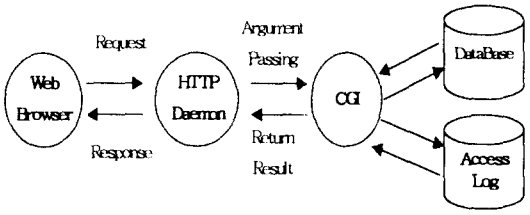


<그림 9> 최종 재고 관리 상품/물량 결정

4. 응용 사례

4.1 상품관리 시스템의 개요

웹 마이닝을 이용한 상품관리 시스템(MeMSWeM)은 World Wide Web과 Database의 연동을 통해서 일반 웹 브라우저를 통해 접속을 하여 분석할 수 있도록 설계되었다. 즉, 웹 브라우저에서 상품관리자의 요구를 웹 서버로 보내면, 웹 서버에서는 요구 사항을 CGI(Common-Gateway Interface)를 호출하여 처리 하도록 하고, 처리 결과를 받아서 사용자에게 전달해 주는 역할을 하게 된다. CGI는 표준이며, 데이터베이스 연동을 구현할 수 있는 효율적인 방법 중의 하나이므로 CGI로 구현하였다. 또한 상품관리 시스템을 웹으로 구현하면, 상품 관리자가 사내 네트워이 아닌 다른 장소에서도 상품관리 시스템을 접근할 수 있으므로 접근이 매우 용이한 장점을 가질 수 있다.



<그림 10> 상품 관리 시스템의 구조

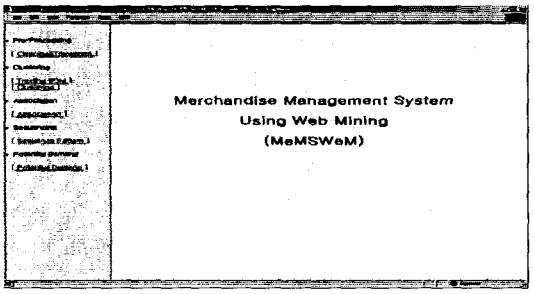
분석하고자 하는 기업 대 고객간의 가상 상점의 시스템 환경은 <표 1>과 같다.

<표 1> 시스템 환경

항 목	해 당 값	항 목	해 당 값
운영체제	Windows NT 4.0	등록 사용자 수	45,000~50,000명
데이터 베이스	Oracle 8.1.5	등록 상품 수	5,000종
웹 서버	MS IIS	하루 웹 로그 축적량	20 - 30 MB

4.2 실 데이터 분석 및 결과

<그림 11>은 웹 마이닝을 이용한 상품관리 시스템의 여러 기능을 선택할 수 있는 메인 화면이다. 메인 화면은 분석 메뉴를 디스플레이 하는 메뉴 프레임과 분석을 수행하는 실행 프레임으로 구성되어 있다. 상품 관리 시스템은, 다음과 같은 기능으로 구성되어 있다.

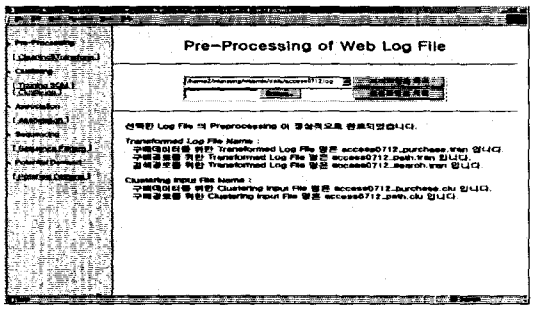


<그림 11> 메인 화면

- 원시 웹 로그 데이터를 사전 처리하는 Preprocessing(Cleaning & Transformation)
- 상품 대 고객간의 연관성 분석을 위한 Clustering: Clustering은 SOM을 학습시키는 기능과 학습된 SOM을 이용하여 Clustering 하는 기능으로 구성 된다.
- 상품 대 상품간의 정적인 연관성 분석을 위한 Association
- 상품 대 상품간의 동적인 연관성 분석을 위한 Sequencing
- 잠재적 구매 고객과 상품 분석을 위한 Potential Demand

4.2.1 사전 처리(Preprocessing)

<그림 12>는 사전처리 화면으로 원시 웹 서버 로그 파일을 서버에 있는 파일로 설정하거나 클라이언트에 있는 로컬파일로 설정할 수 있도록 구현되어 있다. 사전 처리 할 파일을 설정한 후 처리 메뉴를 선택하면 사전 처리를 수행한다. <그림 12>는 7월부터 12월까지 6개월간의 access0712.log 파일을 사전처리 한 것이며, 사전처리 결과로는 구매데이터, 구매경로, 검색경로 별로 나누어 Transformed log file을 생성하고, 클러스터링 입력 데이터로 사용하는 Clustering Input file을 생성한다.



<그림 12> 사전처리 수행 화면

분석하는 가상상점에서 사용하는 MS IIS - Microsoft Internet Information Server - 는 Windows NT에서 가장 많이 사용되는 웹 서버로서 자체적으로 분석도구를 제공하고 있다. 로그파일 형식은 NCSA 계열의 로그파일과는 다르며, 파일의 기록기간을 단위 별로(일별, 월별) 설정할 수 있고, 이는 IIS 관리자에서 실행할 수 있다. <그림 13>은 IIS 로그파일의 예이다.

<그림 13> MS IIS 로그 파일의 구조

이는 다음과 같은 사실을 기록하고 있다 : 10.75.176.21의 IP주소를 가진 일반 사용자가

1997년12월11일 오전 7시 55분 20초에 이름이 TREY1이며 IP가 10.107.1.121인 서버에게 웹 서비스를 요청하였다. 서비스는 4,502msec(약 4.5초)동안 진행되었고, 163바이트의 명령어를 사용하였으며, 그 결과 3,223바이트의 데이터를 에러없이(200과 0코드) 사용자에게 전송하였다. 여기에 사용된 HTTP 명령어는 GET이었으며, 요청한 파일의 이름은 small.gif, 마지막으로 사용된 프로토콜은 Mozilla/3.01 Gold(Win95-1)이었다. 사전처리는 3단계로 구성되어 있다.

- 1단계 : 데이터 클리닝을 하는 단계이다. 즉, Status 필드가 200 또는 0번 이외의 값을 가지고 있는 로그 엔트리는 에러가 발생한 엔트리이다. 그러므로 이러한 에러 엔트리를 제거한다. 또한 분석에 필요치 않는 이미지에 대한 데이터, 사용자 경로에 고려하지 않아도 되는 페이지에 대한 데이터를 걸러내는 데이터 크리닝 과정을 우선 수행한다.
- 2단계 : 데이터 크리닝이 끝난 로그 데이터에서 방문한 상품코드를 추출한다. 즉, 로그 엔트리에 나타난 페이지가 어떤 상품을 디스플레이하고 있는지에 대해 상품코드를 추출하여 매핑한다.
- 3단계 : 로그 엔트리가 구매까지 이루어 진 구매 경로인가? 아니면 검색만 이루어진 검색 경로인가를 구분하는 작업을 수행한다. 즉, 구매/검색 구분 필드에 구매가 이루어지면 "1", 검색만 이루어지면 "0" 으로 매핑한다.

변환 되기 전 원시 로그데이터는 <표 2>와 같은 형태이며, 3 단계의 사전처리를 거친 변환된

<표 2> 원시 웹 로그 파일 형식

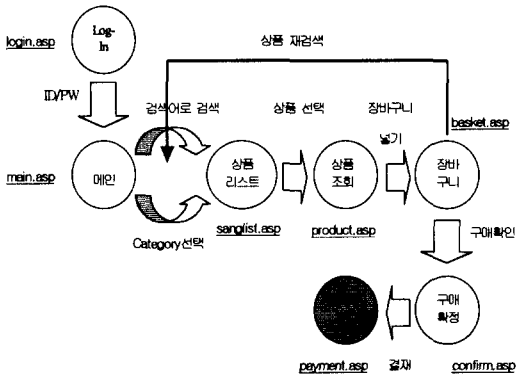
210.219.71.140, -, 99-07-01, 0:04:26, W3SVC2, MAIN_SERVER, 210.219.71.130, 19390, 524, 16851,200, 0, GET,/sanglist.asp,page=1&div1=03&div2=02&div3=03&div1name=패션/잡화 &div2name=신발&div3name=등산화
210.219.71.140, -, 99-07-01, 0:04:26, W3SVC2, MAIN_SERVER, 210.219.71.130, 19000, 410, 1596, 200, 0, GET, /images/product/small/k110701.gif, -,
210.219.71.140, -, 99-07-01, 0:04:26, W3SVC2, MAIN_SERVER, 210.219.71.130, 203, 410, 1578, 200, 0, GET, /images/product/small/k110702.gif, -,
.....

<표 3> 변환된 로그 데이터

IP Address	Date	Time	File	ProdCode	P/Sflag
210.219.71.131	99-07-01	7:35:53	/sanglist.asp	030101	0
210.219.71.131	99-07-01	7:35:56	/product.asp	030101	0
210.219.71.131	99-07-01	7:35:58	/basket.asp	030101	1
210.219.71.131	99-07-01	7:36:02	/log.asp	030101	0
210.219.71.131	99-07-01	7:36:37	/sanglist.asp	030109	0
210.219.71.131	99-07-01	7:36:39	/product.asp	030109	0
210.219.71.131	99-07-01	7:36:42	/basket.asp	030109	1

데이터는 <표 3>과 같다. <표 3>에서 ProdCode 필드는 해당 엔트리가 전시하고 있는 상품의 상품코드이다. P/S Flag 필드는 해당 엔트리가 사용자가 상품을 검색만 한 것인지, 구매가 발생한 것인지를 나타내는 표시이다. 원시 로그 데이터의 변환이 완료되어 변환된 로그 데이터가 생성되면, 이를 이용하여 구매데이터만을 포함한 분석파일, 구매경로를 포함한 분석파일, 검색경로를 포함한 분석파일을 구분하여 3개의 파일로 생성한다.

<그림 14>에서 살펴보면 알 수 있듯이, 로그상에 payment.asp가 나타나면 그것은 구매완료를 나타내므로 해당 상품을 구매한 구매데이터로 구분하면 된다. 구매경로는 로그 상에 payment.asp가 나타난 사용자의 이전 Path를 거슬러 올라가면서 경로를 생성하면 된다. 검색경로는 구매경로를 제외한 경로를 사용자 별로 정리하면 된다.



<그림 14> 가상 상점의 구매 Flow

또한 Preprocessing 과정에서는 클러스터링을 위한 입력 데이터 행렬을 생성하며, 구매데이터, 구매경로를 위한 입력 데이터를 생성한다. 클러스터링을 위한 입력 데이터의 열(row) 데이터는 상품들의 종류이며, 행(column)은 구매데이터/구매경로에 따라 고객이 그 상품을 구매/방문했으면 "1", 구매/방문하지 않았으면 "0"을 대입한 자료이다. 이렇게 Clustering 을 수행하면, 고객들이 비슷하게 구매/방문하는 제품들이 하나의 group으로 나누어 짐으로 차후의 분석에 도움이 된다. 데이터 행렬은 <표 4>와 같다.

<표 4> 클러스터링 입력 데이터 행렬

	Cust 1	Cust 2	Cust 3	Cust 4	.....	Cust m
ProdCode 1	0	1	1	0	.....	0
ProdCode 2	1	1	0	1	.....	0
ProdCode 3	0	0	0	1	.....	1
ProdCode 4	1	1	0	1	.....	1
.....	.....	.....	.....	.....	.....	.....
ProdCode n	0	0	1	1		0

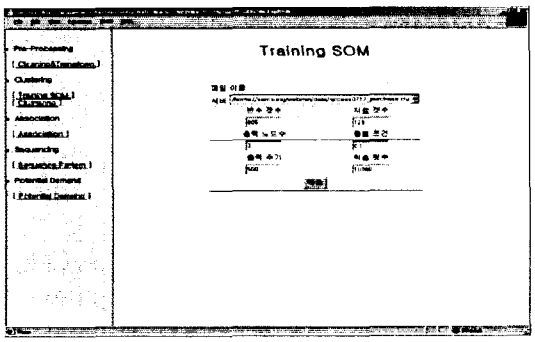
4.2.2 구매 데이터를 이용한 잠정 재고 상품 평가 및 예측

잠정 재고상품 평가 및 예측은 구매데이터를 이용하므로 사전처리에서 생성한 access0712_purchase.tran, access0712_purchase.clu 파일을 이용한다.

4.2.2.1 클러스터링 (Clustering) : 구매데이터를 이용하여

클러스터링을 수행하기 위해서는 우선 SOM을 학습시키는 과정이 필요하다. SOM 의 학습 데이터는 사전 처리 단계에서 생성한 구매데이터의 클러스터링 입력 행렬이다. <그림 15>는 SOM 학습을 위한 화면이며, 각 항목들을 살펴보면 다음과 같다.

- 변수 개수 : 고객의 수를 나타낸다.
- 자료 개수 : 상품의 수를 나타낸다.
- 출력 노드 : SOM 의 출력 노드 수를 설정한다., 3이면 3 X 3 행렬 형태로 출력한다.
- 종료 조건 : SOM의 학습을 어느 정도 정확하게 할 것인가를 설정한다. 낮을수록 많은 학습을 수행하게 된다.
- 출력 주기 : 학습 결과를 어떤 주기로 출력할 것인가를 설정한다. 500이면 500번 학습 후의 SOM의 학습 값들을 출력하게 된다.
- 학습 횟수 : 최대 학습 횟수를 설정한다. 종료 조건에 이르지 못했더라도 최대 학습 횟수를 학습하면 학습을 종료한다.



<그림 15> SOM 학습 수행 화면

SOM의 학습을 수행하였으면, 학습된 SOM을 이용하여 데이터에 대한 클러스터링 수행하게 된다. <그림 16>은 클러스터링을 수행한 결과이다. <그림 16>는 3개의 출력노드를 가진 SOM의 출력으로서 3X3행렬 형태로 나타나게 된다. 화면에 나타난 색의 종류가 Cluster의 수를 나타난다. 화면에서는 9개의 색이 나타났으므로 각각의 Cell이 1개의 클러스터를 나타내고 전체 클러스터의 개수는 9개 이다. 각각의 클러스터에 속한 상품을 확인하기 위해서는 각 Cell의 숫자를 선택하면, 해당 클러스터에 속한 상품의 종류를 출력한다. <그림 17>은 가운데 Cell에 대한 클러스터 결과를 출력한 것이다. <그림 17>의 결과를 살펴보면, 1, 3, 19, 20 와 같은 몇 개의 상품을 제외하면 컴퓨터 관련 상품들이 하나의 클러스터에 포함되어 있는 것을 살펴볼 수 있다. 이 클러스터링 결과를 이용하여 상품 대 고객간의 연관성을 살펴보면 다음과 같은 정보를 추출해 낼 수 있다.

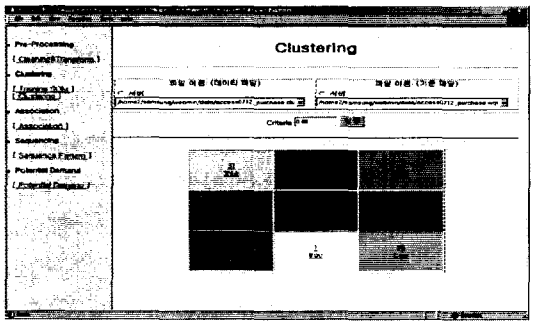
- 고객들이 컴퓨터 관련 상품들을 비슷한 형태로 구매한다. 즉, 컴퓨터 쪽에 관심이 있는 고객들은 관련 상품들을 지속적으로 구매하는 경향을 파악할 수 있다.
- 구매 고객들의 IP Address를 살펴보면, 대부분의 고객들이 천리안을 이용하여 인터넷을 사용하는 천리안 고객들임을 알 수 있다. 천리안을 이용한 IP Address는 다음과 같다.
  * 210.120.77.64 - 210.120.77.127
  * 210.216.92.64 - 210.216.92.95
  * 211.60.76.160 - 211.60.76.191
  * 211.60.61.0 - 211.60.61.63
  * 211.60.81.48 - 211.60.81.63
  * 210.108.250.128 - 210.108.250.159
  * 211.60.76.0 - 211.60.76.31 .

<그림 17>의 클러스터링 결과가 상품 대 고객간의 연관성을 나타내므로, 이 결과를 이용하여 다음과 같이 상품군별 고객에 대한 마케팅을 수행할 수 있다.

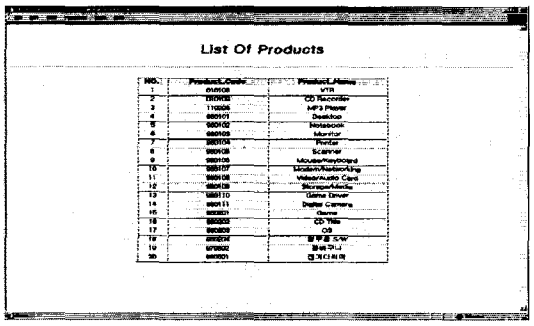
- 천리안을 통해 인터넷을 사용하는 고객이 컴퓨터 관련 상품을 많이 구매하는 경향을 나타

내고 있으므로, 컴퓨터 관련 상품군의 판매를 촉진하기 위한 마케팅은 천리안과 제휴하여 천리안 고객을 중심으로 수행하면 될 것이다.

- 천리안 사용자가 컴퓨터 관련 상품 군을 많이 구매하므로 천리안과 가상 상점 간의 연결 라인을 신설/증설하여 접속 속도를 향상시킨다면 고객 만족도를 높일 수 있을 것이다.



<그림 16> 클러스터링 수행 화면



<그림 17> 클러스터링 한 Cell에 포함된 상품 리스트 출력 화면

4.2.2.2 연관성 분석 (Association) : 구매데이터를 이용하여

클러스터링 된 결과를 참조하여 상품들간의 상관계수를 구하고 서로 연관성이 높은 상품들끼리 동일 상품군으로 구분한다. 상품간의 연관성을 구하기 위하여 상품간 상관계수를 구한 상관행렬을 이용하고, 서로 간의 상관계수가 높은 상품들을 구한다.

<그림 18>은 구매 데이터를 이용하여 연관성 분석을 수행한 결과이다. 출력을 위한 상관계수를 0.2로 입력했으므로 0.2이상의 연관성을 가진 상품들만을 출력한다. <그림 18>의 결과를 살펴보면, 다음과 같은 상품 대 상품간의 정적인 연관성을 추출해 낼 수 있다.

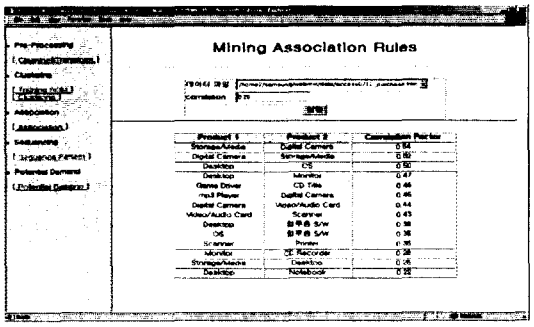
- Storage/Media를 구매한 고객의 54%가 Digital Camera를 구매.
- Digital Camera를 구매한 고객의 각각 52, 44%가 Storage/Media, Video/Audio Card를 구매.
- Desktop을 구매한 고객의 각각 50, 47, 38, 22%가 OS, Monitor, 업무용 S/W, Notebook을 구매.
- Game Driver를 구매한 고객의 46%가 CD Title을 구매.
- Mp3 player를 구매한 고객의 45%가 Digital Camera를 구매.
- Video/Audio Card를 구매한 고객의 43%가 Scanner를 구매.
- Scanner를 구매한 고객의 35%가 Printer를 구매.
- Monitor를 구매한 고객의 28%가 CD Recorder를 구매.

<그림 18>의 Association결과를 이용하여 다음과 같은 상품관리와 마케팅 정책을 수립할 수 있다.

- 상품관리 정책을 수립할 때는 서로 연관성이 있는 상품들은 연관성 정도에 따라 재고 상품

선정과 물량을 결정 할 때 같이 고려한다. 또한 매장 진열에서도 동일 화면에 같이 진열시키는 것이 구매욕을 증대시킬 수 있다. 즉, Storage/Media와 Digital Camera, Desktop 과 OS가 50% 이상의 높은 연관성을 보이고 있으므로, 재고 상품 선정 시에 같이 선정하며, 물량을 결정할 때도 비슷하게 유지해 주는 것이 좋다. 또한 진열 시에 동일 페이지에 같이 진열하는 것이 좋다. 이와 같이 다른 상품들도 서로 간의 연관성 정도에 따라 재고 상품 선정과 물량 결정, 상품 진열 시에 연관성 정도를 고려한다.

- 고객 마케팅 정책을 수립할 때도 서로 연관성이 있는 상품들은 연관성 정도에 따라 공동 마케팅의 시기와 물량을 고려한다. 또한 구매 상품을 통하여 연관성이 높은 상품 정보를 DM을 통하여 고객에게 제공함으로써 구매욕을 증대시킬 수 있다. 즉, Storage/Media와 Digital Camera, Desktop과 OS가 50% 이상의 높은 연관성을 보이고 있으므로, 이 상품들은 서로 공동 마케팅을 수행하는 것이 좋다. 또한 Storage/Media를 구매한 고객에게 Digital Camera 정보를 DM으로 제공해 줄 수 있다.

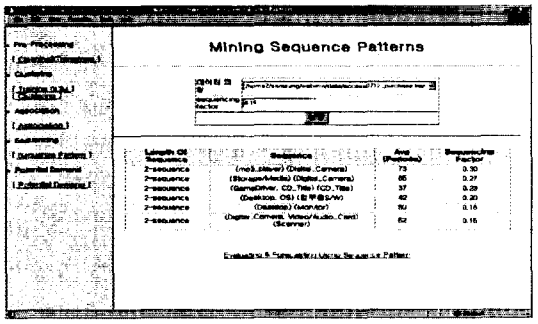


<그림 18> 구매데이터를 이용한 연관성 분석 수행 화면

4.2.2.3 연속 패턴 분석(Sequencing) : 구매데이터를 이용하여

연관성 분석을 수행하여 서로 연관성이 매우 높은 상품들끼리 모았다면, 그 결과를 이용하여 연속 패턴 분석을 통한 트랜드 분석을 수행한다. 연속 패턴 분석은 상품의 판매 경향을 시간을 고려한 선후관계로 정의함으로써, 상품 판매 패턴을 추출해 낼 수 있다.

<그림 19>는 구매데이터를 이용한 연속 패턴 분석의 결과이다. Sequencing Factor를 0.15로 입력했으므로, 0.15 이상의 연속성을 가진 패턴만을 추출한다.



<그림 19> 구매데이터를 이용한 연속 패턴 분석 수행 화면

<그림 19>의 결과는 6개의 길이 2인 연속 패턴을 보여주고 있으며, 각 패턴들이 나타나는 평균 기간과 Sequencing Factor를 나타내고 있다. <그림 19>의 결과를 좀 더 자세히 살펴보면, 다음과 같은 상품 대 상품간의 동적인 연관성을 추출해 낼 수 있다.

- Mp3 player를 구매한 고객의 30%가 평균적으로 73일 이후에 Digital Camera를 추가 구매.
- Storage/Media를 구매한 고객의 27%가 평균

65일 이후에 Digital Camera를 추가 구매.

- Game Driver와 CD Title을 동시에 구매한 고객의 23%가 평균 37일 이후에 또 다른 CD Title을 추가 구매.
- Desktop과 OS를 동시에 구매한 고객의 20%가 평균 42일 이후에 업무용 S/W를 추가 구매.
- Desktop을 구매한 고객의 16%가 평균적으로 80일 이후에 Monitor를 추가 구매.
- Digital Camera과 Video/Audio Card를 동시에 구매한 고객의 15%가 평균적으로 52일 후에 Scanner를 추가 구매.

<그림 19>의 연속 패턴 분석 결과를 이용하여 다음과 같은 상품 관리와 마케팅 정책을 수립할 수 있다. 여기서 주목해야 할 것은, 지금까지의 분석은 고객군 및 상품군에 대한 상품관리 및 마케팅을 수행할 수 있는 분석결과를 도출해 냈지만, 연속 패턴 분석은 개별 고객의 연속 패턴을 조사하여, 잠재적인 상품을 도출해 낼 수 있으며, 일대일 마케팅을 수행할 수 있는 분석결과를 도출해 낼 수 있다.

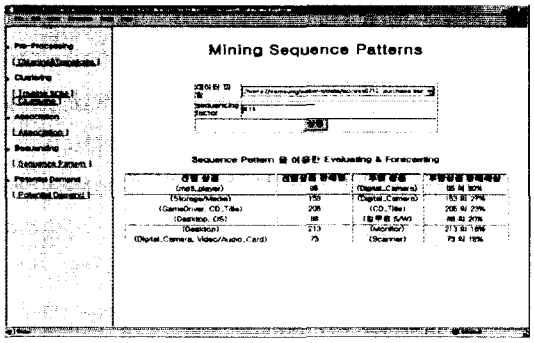
- 상품관리 정책을 수립할 때는 서로 연속 패턴이 있는 상품들은 재고 상품 선정과 물량을 결정 할 때 같이 고려한다. 또한 매장 진열에서도 먼저 구매하는 상품화면에서 다음 구매가 이루어 지는 상품화면으로 직접 이동할 수 있는 링크를 추가함으로써 검색 용이성 및 구매확률을 증대시킬 수 있다. 즉, mp3 player를 구매한 고객은 평균적으로 73일 후에 Digital Camera를 구매하므로 다음 달 Digital Camera 재고 상품 물량 결정시에 이번 달에 판매된 mp3 player의 개수와 Sequencing Factor를 고려하여 선정할 수 있다. 이와 같이 다른 상품들도 서로 간의 연속 패턴과 평균 이동시간, Sequencing Factor를 재고상품 선정과 물량 결정, 상품진열 시에 고려할 수 있다.
- 고객 마케팅 정책을 수립할 때도 상품의 연속 구매 패턴을 이용하여, 고객이 다음에 구매할 상품을 예측할 수 있으므로 DM을 통하여 고객에게 추가 구매할 가능성이 높은 상품 정보를 제공함으로써 구매욕을 증대시킬 수 있다. 즉, mp3 player를 구매한 고객은 약 73일 후에는 Digital Camera를 구매하므로, 60-70일 정도 지났을 때 Digital Camera 정보를 제공해 준다면, 고객 만족도를 향상 시킬 수 있고, 고객 개개인들에게 차별적으로 필요한 정보를 제공해 주는 일대일 마케팅을 시작할 수 있다.

4.2.2.4 잠정적인 평가 및 예측(Evaluating & Forecasting) : 구매데이터를 이용하여

구매데이터를 이용한 분석 결과를 이용하여 잠정적으로 재고로 관리해야 할 상품과 물량을 다음과 같이 결정할 수 있다.

① 연속 패턴 분석을 통한 상품과 물량 결정: 연속 패턴 분석 결과를 이용하여 상품과 물량을 결정하기 위해서는 <그림 20>과 같은 연속 패턴상의 선행 상품의 판매자료가 필요하다. 선행 상품의 해당 기간 판매량을 조사하여 후행 상품의 판매 예상 물량을 결정한다. 여기서 선행 상품의 해당 기간 판매량을 계산할 때 주의할 것은, 선행 상품의 전체 판매량에서 이미 후행 상품을 구매한 고객에게 판매된 판매량을 제외해야 한다는 것이다. 후행 상품의 예상 판매량은 선행 상품의 판매량에

Sequencing Factor를 곱하여 계산한다.



<그림 20> 연속 패턴 분석을 통한 상품과 물량 결정 화면

② 연관성 분석을 통한 연관 상품과 물량 결정: 연속패턴 분석결과를 이용하여 선정된 후행 상품과 판매 연관성이 높은 상품들을 추가적으로 선택한다. 연관성이 높은 상품을 선정하기 위해서는 연관성 분석 결과를 이용한다. 연관성 분석 결과를 이용한 추가 상품 선정 결과는 <표 5>와 같다.

<표 5> 구매데이터 연관성 분석 결과를 이용한 추가 상품 선정

1차 선정 상품	연관 상품	상관계수	연관 상품 판매예상
Digital Camera	Storage/Media	0.54	(95의 30% + 153의 27%)의 54%
Digital Camera	Video/Audio Card	0.44	(95의 30% + 153의 27%)의 44%
CD Title	-		-
업무용 S/W	-		-
Monitor	CD Recorder	0.28	213의 16%의 28%
Scanner	Printer	0.35	73의 15%의 35%

③ 잠정적인 재고 관리 상품 및 물량: 잠정적으로 재고로 관리해야 할 상품과 물량은 <표 6>과 같다.

<표 6> 잠정적인 재고 관리 상품 및 물량

잠정 재고 관리 상품	예상 판매량
Digital Camera	95의 30% + 153의 27%
CD Title	205의 23%
업무용 S/W	88의 20%
Monitor	213의 16%
Scanner	73의 15%
Storage/Media	(95의 30% + 153의 27%)의 54%
Video/Audio Card	(95의 30% + 153의 27%)의 44%
CD Recorder	213의 16%의 28%
Printer	73의 15%의 35%

4.2.3 구매 경로를 이용한 상품 구매 검색 패턴 추출

상품 구매 패턴을 추출하기 위해서는 구매경로를 이용하므로 사전처리에서 생성한 access0712_path.tran, access0712_path.clu 파일을 이용한다.

4.2.3.1 클러스터링 (Clustering) : 구매경로를 이용하여

클러스터링은 구매데이터를 이용할 때랑 입력데이터를 구매경로를 이용한다는 것을 제외하고는 동일하다.

4.2.3.2 연관성 분석 (Association) : 구매경로를 이용하여

클러스터링 된 결과를 참조하여 상품들간의 상관계수를 구하고 서로 구매 검색 연관성이 높은 상품들끼리 동일 상품 군으로 구분한다. 연관성 분석 결과는 패턴 데이터베이스에 저장된다.

<그림 20>은 구매 경로를 이용하여 연관성 분석을 수행한 결과이다. 출력을 위한 상관계수를 0.3으로 입력했으므로 0.3 이상의 연관성을 가진 상품들만을 출력하였으며, 다음과 같은 상품 대 상품간의 정적인 연관성을 추출해 낼 수 있다.

- Desktop을 구매검색 한 고객의 각각 68, 64, 60, 58, 47%가 Storage/Media, Monitor, Notebook, CD Recorder, 업무용 S/W을 구매 검색.
- Mp3 player를 구매검색 한 고객의 65%가 Desktop을 구매 검색.
- Storage/Media를 구매검색 한 고객의 63%가 Digital Camera를 구매 검색.
- Monitor를 구매검색 한 고객의 각각 55, 42%가 Video/Audio Card, Printer를 구매 검색.
- CD Title을 구매검색 한 고객의 52%가 Mouse/Keyboard을 구매 검색.
- Mouse/Keyboard를 구매검색 한 고객의 47%가 Video/Audio Card를 구매 검색.
- 업무용 S/W를 구매검색 한 고객의 40%가 Monitor를 구매 검색.
- OS를 구매검색 한 고객의 35%가 업무용 S/W를 구매 검색.

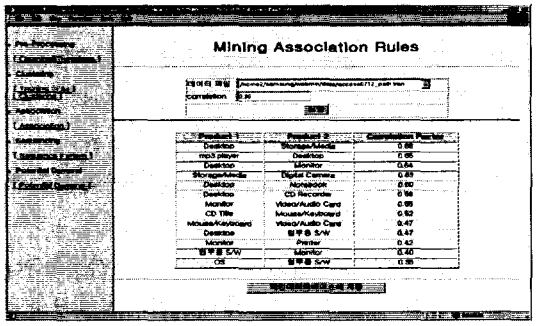
추출된 연관성 분석 패턴을 패턴 데이터 베이스에 저장하기 위해서는 패턴 데이터베이스에 저장 버튼을 선택한다.

4.2.3.3 연속패턴 분석(Sequencing) : 구매경로를 이용하여

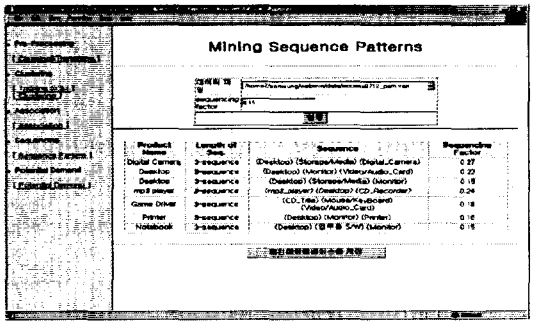
연관성 분석을 수행하여 서로 연관성이 매우 높은 상품들끼리 모았다면, 그 결과를 이용하여 연속패턴 분석을 통한 트랜드 분석을 수행한다. 연속패턴 분석은 상품별 구매검색 경향을 시간을 고려한 선후관계로 정의함으로써, 상품구매 검색 패턴을 추출해 낼 수 있다. 구매경로를 이용한 연속패턴 분석이 구매데이터를 이용한 연속패턴 분석과 다른 점은 각 상품별 구매 경로에 대해서 연속패턴을 추출한다는 것이다. 즉, 각 상품에 대해서 구매경로를 구분하고 그 구매경로에서 나타나는 연속패턴을 추출하여 패턴 데이터베이스에 저장한다.

<그림 21>은 구매경로를 이용한 연속 패턴 분석의 결과이다. Sequencing Factor를 0.15로 입력했으므로, 0.15 이상의 연속성을 가진 패턴만을 추출한다.

<그림 21>의 결과는 각 상품별로 구매 검색 패



<그림 20> 구매경로를 이용한 연관성 분석 수행 화면



<그림 21> 경로를 이용한 연속 패턴 분석 수행 화면

턴을 추출한 것이며, 6개의 길이 3인 연속 패턴과 각 패턴별 Sequencing Factor를 나타내고 있다. <그림 21>의 결과를 좀 더 자세히 살펴보면, 다음과 같은 상품 대 상품간의 동적인 연관성을 추출해 낼 수 있다.

- Digital Camera를 구매한 고객의 27%가 Desktop, Storage/Media, Digital Camera 순으로 구매검색 후 Digital Camera를 구매.
- Desktop을 구매한 고객의 22%는 Desktop, Monitor, Video/Audio Card 순으로 구매검색 후 Digital Camera를 구매하고, 15%는 Desktop, Storage/Media, Monitor 순으로 구매검색 후 Desktop을 구매.
- Mp3 player를 구매한 고객의 24%가 mp3 player, Desktop, CD Recorder 순으로 구매검색 후 mp3 player를 구매.
- Game Driver를 구매한 고객의 18%가 CD Title, Mouse/Keyboard, Video/Audio Card 순으로 구매검색 후 Game Driver를 구매.
- Printer를 구매한 고객의 16%가 Desktop, Monitor, Printer 순으로 구매검색 후 Printer를 구매.
- Notebook을 구매한 고객의 15%가 Desktop, 업무용 S/W, Monitor 순으로 구매검색 후 Notebook을 구매.

연관성 분석과 마찬가지로 추출된 연속패턴을 저장하기 위해서는 패턴 데이터베이스에 저장버튼을 선택한다.

4.2.4 검색 경로를 이용한 잠재적 구매 상품과 물량 추출

잠재적 구매 상품과 물량을 추출하기 위해서는 검색경로를 이용하므로 사전처리에서 생성한 access0712_search.tran 파일을 이용한다.

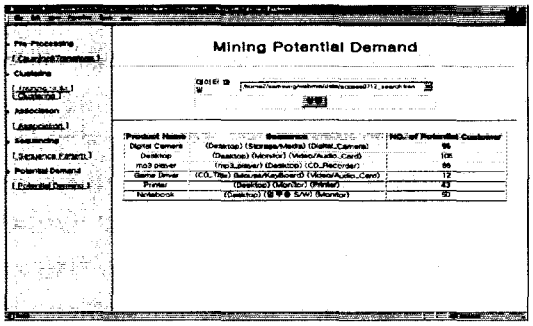
4.2.4.1 연속 패턴 분석(Sequencing) : 검색경로를 이용하여

구매데이터, 구매경로를 이용한 분석과 달리, 검색경로를 이용한 연속패턴 분석은 클러스터링과 연관성 분석을 수행하지 않고 바로 연속패턴을 추출한다. 연속패턴은 각 검색을 수행한 사용자 별로 추출하며, 추출된 연속패턴은 패턴 데이터베이스에 저장되어 있는 패턴들과 비교하여 일치하는 것이 있는가를 검색한다. 일치하는 패턴이 존재한다면, 추출된 연속패턴을 지닌 상품은 잠재적 구매상품으로 선정할 수 있다.

<그림 22>는 검색경로를 이용한 연속 패턴 분석의 결과이다. 여기서는 sequencing factor를 고려하지 않아도 된다. <그림 22>의 결과는 검색경로 연속 패턴 중에서 구매경로를 이용하여 추출한 연속패턴과 일치하는 패턴을 소유한 잠재적 구매 고객을 추출하여 각 패턴별로 나타내고 있다. <그림 22>의 결과를 좀 더 자세히 살펴보면, 다음과 같다.

- Digital Camera 구매 고객의 구매 경로를 소유한 잠재적 고객의 수가 55명이다.
- Desktop 구매 고객의 구매 경로를 소유한 잠재적 고객의 수가 105명이다.
- Mp3 player 구매 고객의 구매 경로를 소유한 잠재적 고객의 수가 66명이다.
- Game Driver 구매 고객의 구매 경로를 소유한 잠재적 고객의 수가 12명이다.
- Printer 구매 고객의 구매 경로를 소유한 잠재적 고객의 수가 43명이다.
- Notebook 구매 고객의 구매 경로를 소유한 잠

재적 고객의 수가 50명이다.



<그림 22> 검색경로를 이용한 연속 패턴 분석 수행 화면

검색 경로를 이용한 연속 패턴 분석을 통하여 잠재적 구매 상품과 물량을 예측하여 보았다. 이렇게 선정된 잠재적 구매 상품과 검색 연관성이 높은 상품들도 잠재적 구매 상품이라고 볼 수 있다. 연관성 분석 결과를 이용한 추가 잠재적 구매 상품 선정 결과는 <표 7>과 같다.

<표 7> 구매경로 연관성 분석 결과를 이용한 추가 상품 선정

1차 선정 상품	연관 상품	상관계수
Digital Camera	-	-
Desktop	Storage/Media	0.68
	Monitor	0.64
	Notebook	0.60
	CD Recorder	0.258
	업무용 S/W	0.47
mp3 player	Desktop	0.65
Game Driver	-	-
Printer	-	-
Notebook	-	-

최종적인 잠재 구매 상품과 물량은 다음 <표 8>과 같다.

<표 8> 잠재적 구매 상품 및 물량

잠재 구매 상품	잠재 구매 물량
Digital Camera	55
Desktop	105
mp3 player	66
Game Driver	12
Printer	43
Notebook	50
Storage/Media	-
Monitor	-
CD Recorder	-
업무용 S/W	-

4.2.5 평가 및 예측(Evaluating & Forecasting)

평가 및 예측은 구매데이터를 이용한 잠정적인 재고관리 상품과 물량예측 결과와 구매경로, 검색경로를 이용한 잠재적 구매상품과 물량예측 결과를 종합하여 수행한다. 평가 및 예측 결과는 <표 7>과 <표 8>의 결과를 종합한 결과이며 <표 9>와 같다.

<표 9> 최종적인 재고 관리 상품 및 물량

재고 관리 상품	예상 판매량
Digital Camera	(95의 30% + 153의 27%) + 55
CD Title	205의 23%
업무용 S/W	88의 20%
Monitor	213의 16%
Scanner	73의 15%
Storage/Media	(95의 30% + 153의 27%)의 54%
Video/Audio Card	(95의 30% + 153의 27%)의 44%
CD Recorder	213의 16%의 28%
Printer	(73의 15%의 35%) + 43
Desktop	105
Mp3 player	66
Game Driver	12
Notebook	50

5. 결 론

본 연구에서는 웹에 기초한 기업과 소비자간 전자상거래 환경 하에서 운영되고 있는 가상상점의 디지털화 된 고객 구매이력과 행동, 상품의 판매동향 정보를 체계적으로 분석할 수 있는 시스템적 접근 방법을 제안하였고, 상품관리에 대한 연구 결과들과 웹 마이닝 기법들을 통합함으로써 가상상점에서의 상품관리를 위한 새로운 방법론을 제시하였다. 또한 제안한 분석 방법을 이론으로만 제시한 것이 아니라 실제 운영되고 있는 가상상점에 적용해 봄으로써 분석 방법의 타당성을 검증하였다.

향후, 본 연구에서 개발된 시스템을 AI tool, Heuristic 을 이용한 Intelligent Agent System의 구축으로 발전시켜 나갈 예정이다.

참고문헌

[1] R. Cooley, B.Mobasher, and J. Srivastava (1997), "Web Mining : Information and Pattern Discovery on the World Wide Web", In Proc. of the 9th IEEE International Conference, (1997) 558-567.

[2] Osmar R. Zalane, Man Xin, Jiawei Han, "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs", In Proc. IEEE Int.Forum, (1998) 19-29.

[3] R. Agrawal and R. Srikant, "Fast Algorithms for mining association rules." In Proc. Of the VLDB Conference, Santiago, Chile, September(1994).

[4] R. Agrawal and R. Srikant, "Mining Sequential Patterns", Research Report RJ 9910, IBM Almaden Research Center, San Jose, California, October (1994)

[5] Lee, W. J., "Retailer's Cost Saving with Quick Response Implementation : Department Store Case", Graduate School of Management, Korea Advanced Institute of Science and Technology, (1998).

[6] Mason, J. B., Mayer, M. L. and Ezell, H. F., "Retailing", Fourth edition, Irwin Inc., (1991).

[7] Stock, J. R. and Lambert, D. M., "Strategic Logistics Management", Irwin Inc., (1987).

[8] Lo, T., "An export system for choosing demand forecasting techniques", International Journal of Production Economics, 33, (1994) 5-15.

[9] Cash, R. P., Wingate, J. W., and Freidlauder, J. S., "Management Retail Buying", John Wiley & Sons, Inc., (1995).

[10] Sabrina Sestito, Tharam S Dillon. "Automated Knowledge Acquisition", Prentice Hall. (1994).

Abstract

Merchandise Management Using Web Mining in Business To Customer Electronic Commerce

Kwang Hyuk Im*
Han Kook Hong**
Sang Chan Park*

Until now, we have believed that one of advantages of cyber market is that it can virtually display and sell goods because it does not necessary maintain expensive physical shops and inventories. But, in a highly competitive environment, business model that does away with goods in stock must be modified. As we know in the case of AMAZON, leading companies already consider merchandise management as a critical success factor in their business model. That is, a solution to compete against one's competitors in a highly competitive environment is merchandise management as in the traditional retail market.

Cyber market has not only past sales data but also web log data before sales data that contains information of path that customer search and purchase on cyber market as compared with traditional retail market. So if we can correctly analyze the characteristics of before sales patterns using web log data, we can better prepare for the potential customers and effectively manage inventories and merchandises. We introduce a systematic analysis method to extract useful data for merchandise management - demand forecasting, evaluating & selecting - using web mining that is the application of data mining techniques to the World Wide Web. We use various techniques of web mining such as clustering, mining association rules, mining sequential patterns.

Key words: Merchandise Management, Merchandizing, Web Mining, Electronic Commerce, Demand Forecasting, Evaluating, Web Access Pattern.

* Dept. of Industrial Engineering, Korea Advanced Institute of Science and Technology
** College of Commerce and Economy, DongEui University