Journal of Korea Multimedia Society Vol. 7, No. 2, February 2004(pp. 145-155)

연관 규칙 기반의 상품 검색 데이터베이스 최적화 연구

황현숙†, 박규석††

요 약

인터넷 쇼핑몰을 구성하는 관리자 입장에서는 사용자 중심의 편리한 검색 기능과 시스템 중심의 빠른 검색 기능을 가지는 것이 매우 중요하다. 전자는 사용자의 다양한 요구를 만족시킬 수 있는 최적화된 입력 매개 변수를 찾아내는 것이며, 후자는 속성이 다른 다양한 입력 변수들을 효과적으로 정규화하여 빠른 검색해를 찾아내는 것이다. 본 연구에서는 기본적으로 사용자의 다양한 요구를 최대한 반영하기 위해 다중 속성을 가진 검색 기능은 물론 보다 빠른 검색 기능을 가지기 위한 데이터베이스 최적화 구성에 초점을 두고 있다. 이를 위해 인터넷 쇼핑몰의 검색 특성을 반영할 수 있는 연관 규칙의 척도인 지지도와 신뢰도를 고려한 수정된 연관 알고리즘을 제시하며, 빠른 검색 기능을 가지기 위한 모델관리시스템을 제안한다. 수행된 시뮬레이션 결과에 의하면 고객의 검색 트랜잭션 수가 증가할수록 전체 평균 검색 시간은 상대적으로 줄어든다.

A Study on the Product Searching Database Optimization Based on Association Rules

Hyun-Suk Hwang†, Kyoo-Seok Park††

ABSTRACT

It is very important for Internet searching systems to have user-friendly and rapid searching functions at the managers' point of view. The former finds optimized input parameters to support the various searching requirements of user. The latter has fast searching results which are effectively normalized to various input parameters having different attributes. In this paper we basically focus on optimized database construction not only to have searching functions with multiple attributes to support maximal various input requirements of the user but also to have more rapid searching functions. For this research, we suggest a modified association algorithm that takes into consideration to the support and confidence that is the criteria of the association mining rule in order to reflect the searching characteristics of internet shopping malls. We also propose the model management systems for rapid searching functions. The following results are from a processed simulation: the more the number of searching transactions of the users increase, the less the total relative average searching time becomes.

Key words: Database Optimization(데이타베이스 최적화), Association Rules Algorithm(연관 규칙 알고리즘), Product Searching Systems(상품 검색 시스템), Model Management Systems(모델 관리 시스템)

※ 교신저자(Corresponding Author) : 황현숙, 주소 : 부산광역시 남구 대연3동 599-1(608-737), 전화 : 051)620-6394, FAX : 051)620-6390, E-mail : hhs@mail1.pknu.ac.kr
접수일 : 2003년 4월 16일, 완료일 : 2003년 7월 29일
† 정회원, 부경대학교 경영정보학과 졸업(박사)
†† 종신회원, 경남대학교 정보통신공학부 교수
(E-mail : kspark@kyungnam.ac.kr)
※ 이 논문은 2001년 한국학술진흥재단의 지원에 의하여 연구되었음(KRF-2001-037-CA0047).

1. 서 론

인터넷 사이버 쇼핑몰의 주된 목적은 전 세계에 흩어져 있는 데이터베이스를 근간으로 사용자의 요구 사항에 가장 적합한 빠른 상품 검색 시스템을 구축하여 쇼핑몰의 이용률을 증대시키는 것이다. 통계청은 사이버 쇼핑몰 사업체수가 2003년 2월 3,082개로 전년도 대비 35.4% 증가하였으며, 거래액은 5,637

억 원으로 3.9% 감소하였음을 발표하고 있다[1]. 이는 정보 제공자의 참신한 아이디어와 새로운 사업 영역의 확대로 사업체수가 증가하는 현상으로 파악할 수 있으며, 거래액이 감소하는 것은 여러 가지 원인이 있지만 사용자의 편의성과 시스템의 안정성 및 신속성에도 상당한 영향을 받고 있음을 알 수 있다. 사이버 쇼핑몰이 저렴하고 다양한 질(quality) 좋은 상품을 아무리 많이 가지고 있다 할지라도 사용자가 구매하기 위해 불편한 입력조건을 요구하거나 느린 검색 결과를 제공하면 사용자의 관심으로부터 멀어질 수밖에 없다. 본 연구에서는 인터넷 쇼핑몰의 외형적인 성장보다는 사용자의 편의성과 빠른 검색 기능에 초점을 맞추고 있기 때문에 사용자의 검색 유형과 입력 매개변수들 사이의 데이터 연관성을 고려한 최적화된 데이터베이스 구축에 있다.

인터넷이 급속히 보급되면서 수많은 데이터베이스에서 필요한 정보를 검색하는 것은 복잡한 검색 기능과 느린 검색 결과가 항상 문제가 되었다. 이러한 문제점을 해결하기 위해 대량의 데이터베이스에서 알려지지 않은 데이터 규칙 및 패턴을 발견하기 위한 연구들이 수행되었으며, 이러한 연구들 중 가장 많이 알려진 기법이 데이터 마이닝(data mining)[2-4]이다. 데이터 마이닝과 관련된 기법으로는 이웃한 K 근사 방법(K-nearest Neighbor Method), 의사결정 트리, 연관 규칙(association rule), 신경망 이론과 유전자 알고리즘 등[5]이 있다. 본 연구에서는 방대한 데이터베이스에 저장되어 있는 데이터로부터 유용한 정보 및 지식을 추출하는데 가장 잘 알려져 있고, 많은 연구자들이 적용하고 있는 연관 규칙을 기반으로 하고 있다. 연관 규칙은 데이터베이스의 트랜잭션에서 항목 간에 발생하는 규칙을 표현하는 것으로 Agrawal & Srikant[6]에 의해 처음 소개되었다. 이는 어떤 사건이 발생될 때 그 다음 사건의 관련성을 나타내기 위해 X⇒Y 규칙의 형태로 표현하였다. X⇒Y의 규칙은 데이터베이스의 트랜잭션 중 X라는 항목 집합을 포함하는 트랜잭션은 Y라는 항목집합도 함께 포함하는 경향이 있음을 의미한다. 이러한 연관 규칙에서는 지지도(support)와 신뢰도(confidence)라는 척도로 그 타당성이 판단된다. 지지도는 전체 트랜잭션에 대해 트랜잭션 항목 집합이 차지하는 비율을 의미하고, 신뢰도는 조건부 트랜잭션 항목 집합에 대해 규칙에 포함되는 모든 항목 집합이 차지하는 비율을 의미한다. 연관 규칙을 이용한 대부분의 연관 알고리즘은 두 가지 단계로 분할하여 해결한다[7-9]. 첫 번째는 최소 지지도 이상을 갖는 빈발 항목 집합을 발견하는 것이고, 두 번째는 발견된 빈발 항목집합에 대하여 최소 신뢰도를 만족하는 규칙을 생성하는 것이다[10,11]. 그러나 기존의 연관 알고리즘은 단일 항목 필드 구조에서 데이터 연관성을 찾는 방법들이 대부분이다.

본 연구는 기존의 연관 규칙 기법을 활용하는 것은 동일하지만 데이터베이스 저장 매카니즘의 특성을 고려할 때 한 개의 레코드에 다수의 항목을 가지고 있는 가지고 있는 다중 항목 필드의 입력 데이터 구조로 변경한 연관 알고리즘을 제안한다. 수정된 다중항목 연관 마이닝 알고리즘은 트랜잭션별 검색 유형과 각 항목의 속성간 검색 유형을 종합적으로 고려한 자주 검색되는 검색 유형 데이터베이스를 구성하게 된다. 이러한 검색 유형 데이터베이스는 사용자의 입력 파라메터에 대해 정의된 규칙과 일치하면 전체 상품 데이터베이스에서 검색하는 것이 아니라 검색 유형 데이터베이스에서 필요한 자료를 찾을 수 있기 때문에 빠른 검색 결과를 제공한다.

본 논문의 구성은 제 2장에서 연관 규칙을 이용한 상품 검색 시스템에 대해 설명하고 제 3장에서는 상품 검색 데이터베이스를 구축하는 단계를 제시한다. 제 4장에서는 실제 상품 데이터를 가지고 검색 유형 데이터베이스를 생성하여 검색 성능을 분석한다. 마지막으로 제 5장에서는 연구 결과 및 문제점을 제시한다.

2. 관련 연구

2.1 단일항목 구조의 연관 알고리즘

기존의 대부분 연관 알고리즘들은 기본적으로 연관 규칙의 지지도와 신뢰도를 기반으로 연구되었다[10-12]. 연관 알고리즘의 초기 연구로 Agrawal & Srikant[6]은 장바구니 데이터를 사용하여 고객이 구매한 상품간에 연관성이 있는 집합을 발견하는 AIS (Association Item Sets) 알고리즘을 제시하였다. AIS 알고리즘은 연관 규칙 생성에 대한 기준으로 최소 지지도와 신뢰도를 기반으로 전체 데이터베이스에 대해 후보 항목 집합을 발견하는 방법을 제안하였다. Houstma & Swami[13]는 AIS 알고리즘을 기반으로 데이터베이스 쿼리문을 사용하여 연관 집합을

발견하는 SETM(SETs Mining) 알고리즘을 제시하고 있는데, AIS와 SETM 알고리즘은 후보 집합을 생성할 때 데이터베이스를 여러 번 접근하여 생성하기 때문에 메모리 관리와 성능에서 비효율적인 문제점을 가지고 있다. 이러한 문제점을 해결하기 위해 Agrawal et al.[9]은 이전 단계의 빈발 함수 집합을 이용하여 후보집합을 생성하는 Apriori 알고리즘을 제안하였다. 제안된 알고리즘의 두 단계로 구분되어 수행된다. 첫 번째 단계는 최소 지지도 이상을 갖는 빈발 항목 집합(frequent itemset)을 발견하는 단계이다. 빈발 항목 집합(F_k)은 이전 단계(k-1)의 빈발 항목 집합에서 k개의 가능한 항목 집합을 생성하여 F_k의 부분 집합이 아닌 경우를 제거하여 후보 항목 집합(candidate itemset: C_k)으로 한다. 이때, 생성된 후보항목 집합에서 최소 지지도 이상을 만족하는 빈발 항목 집합을 생성한다. 두 번째 단계는 생성된 빈발 항목 집합으로부터 최소 신뢰도를 만족하는 연관 규칙을 생성하기 위해 Apriori_gen이라는 알고리즘을 제시하고 있다. 이후 후보항목 집합의 생성을 효율적으로 계산하기 위해 해쉬-트리를 이용한 여러 가지 방법들이 제안되었다[11]. 그러나 위의 방법들은 지지도를 효율적으로 계산하기 위해 해쉬-트리 데이터 구조를 사용하였지만 항목 집합의 제거 과정과 트랜잭션 개수를 계산할 때 쿼리(query)문을 사용하지 않았다.

Sarawagi[12]는 후보 집합 생성과 트랜잭션 개수를 계산하기 위해 Apriori 알고리즘을 기반으로 데이터베이스 시스템과 통합한 알고리즘을 제시하였다. Sarawagi[12]가 제시한 알고리즘의 입력 데이터 구조는 트랜잭션을 구분하는 트랜잭션 번호(T_{id})와 데이터 항목을 나타내는 한 개의 항목(item) 필드로 구성되어 있다. 이러한 단일 항목을 가지는 입력 데이터 구조에서는 한 개의 레코드에 한 개의 항목이 저장되어 있으므로 레코드의 수가 많아지면 알고리즘의 수행 속도가 떨어지고, 트랜잭션의 개수를 계산하기 위해 k개의 트랜잭션 데이터와 조인을 수행해야 하기 때문에 수행 속도가 많이 떨어진다. 따라서 단일 항목 데이터 구조는 장바구니 데이터를 분석할 때 효율적인 구조로서 활용 가능하지만 실제 쇼핑몰을 구성할 때는 활용 범위가 좁은 제한점이 있다.

2.2 다속성 보정 모델

인터넷 쇼핑몰에서 고객이 선호하는 상품 속성에 따라 상품 선정을 도와주려고 할 경우에 다속성(multi-attributes) 의사결정 방법이 일반적으로 많이 알려져 있다. 다속성 의사결정 문제를 해결하기 위해서는 여러 가지 방법들이 제시[9]되고 있지만 크게 무보정 모델과 보정 모델로 구분할 수 있다. 무보정 모델은 속성간에 선호하는 보정을 고려하지 않고 속성 그 자체의 평가로 의사결정을 하는 것이다. 보정 모델은 어떤 속성이 다른 속성에게 영향을 주는 속성간의 상호보완을 가정하고 의사결정자의 선호에 관한 정보를 사용하여 의사결정 문제를 해결하는 방법이다. 보정 모델 중에서 많이 알려진 방법은 가중치 개념을 적용한 점수 모델 방법으로 의사결정자가 상대적인 가중치를 입력하여 고객이 선택한 각 속성들의 정규화 값을 곱해서 점수를 구한 후, 가장 큰 점수를 얻는 대안을 선택하는 것이다.

본 연구에서는 다속성 보정 모델의 의사결정을 위해 모델 기반 검색 모듈에서 다음과 같은 두 가지 단계를 제시하고 있다.

(1) 가중치법 점수 보정 모델 제안

본 연구는 다속성 보정 모델을 위해 단순 가중치법을 적용하고 있다. 상품 검색 시스템에서 고객의 선호도를 고려하기 위해 본 연구에서는 가중치와 선호기준 규칙을 제시하고 있다. 고객이 입력한 가중치의 전체 합은 1이며, 각 속성의 선호 기준은 이윤 속성과 비용 속성으로 구분하는데 속성 데이터의 값이 높을수록 이윤 속성, 낮을수록 낮은 선호를 가지는 비용 속성으로 구분하였다. 데이터의 정규화(normalization)는 선호 속성에 따른 선형 변환 방법을 사용하여 계산하였으며, 전체 선호 점수는 고객이 입력한 속성별 가중치와 선호기준에 따라 정규화한 데이터를 곱해서 계산된다.

본 연구의 다속성 보정 모델 검색을 위한 테이블 구성은 고객이 키워드 입력한 데이터를 저장하는 키워드 입력 결과 테이블이 필요하다. 이를 위해 점수 필드의 값은 고객이 가중치와 선호기준을 입력하였을 때 계산되며, 가중치 및 선호 기준 테이블은 점수 계산에 사용될 고객이 입력한 선호 가중치와 선호기준을 저장하기 위해 필요하다. 가중치별 점수 테이블은 민감도 분석에서 사용하기 위해 선호 점수 계산에서 나온 다수의 검색에 대한 점수를 저장하기 위한 필요하다. 그림 1은 고객의 선호도에 따른 점수 보정 알고리즘을 나타내고 있다.

Algorithm Jumsu_Computation
JUMSU(i) : 상품별 선호 점수,
D(i,j) : 레코드의 항목 속성값
ND(i,j) : 각 속성의 정규치
PRE(j) : 각 속성의 선호기준
(1:이윤속성, 0:비용속성)
Max[D_j(i,j)] : 각 속성의 최대값,
Min[D_j(i,j)] : 각 속성의 최소값
W(j) : 각 항목의 가중치

$$(0 \le W(j) \le 1,\ \sum_{j=1}^{m} W(j) = 1)$$

```
for(i=1, n ; i++)  do begin  /* i번째 레코드
  for(j=1, m; j++) do begin
       /* i번째 레코드의 j번째 항목
     if (PRE(j)=1) then
       ND(i,j) = D(i,j) / Max[Dj(i,j)]
     else
       ND(i,j) = Min[Dj(i,j)] / D(i,j)
     endif
    JUMSU(i) = JUMSU(i)+W(j)*ND(i,j)
  end
end
```

그림 1. 고객 선호도에 따른 점수 보정 알고리즘

(2) 민감도 분석

고객은 선호하는 속성에 대해 가중치의 반영 비율을 변화하면서 모델 기반 검색을 수행할 수 있다. 이러한 검색 결과를 분석하기 위해 민감도 분석을 수행한다. 민감도 분석은 고객이 입력한 가중치 변화에 따라 상품 순위의 변화를 분석한다. 따라서 고객은 다양한 가중치를 제시함으로써 적합한 상품을 선정하는 데 도움을 받게 된다. 민감도 분석 업무에서는 가중치 및 선호기준 테이블과 가중치별 점수 테이블을 입력받아 고객에게 점수 변화에 따른 상품의 정보를 제공하게 된다.

3. 프로시쥬어 기반의 다중항목 연관 알고리즘

3.1 다중항목 데이터 구조의 특성

현재 사용되고 있는 데이터베이스 구성은 다중 필드들로 구성되어 있다. 그러나 연관 알고리즘과 관련된 기존 연구들 대부분은 트랜잭션을 구분하는 트랜잭션번호와 데이터 항목을 나타내는 항목 필드만을 가지고 연관 규칙을 적용하고 있다. 이는 한 개의 데이터 항목만을 가지는 단일 항목 데이터 구조를 기반으로 후보 항목집합과 빈발 항목집합을 생성하는데 초점을 맞추고 있다. 이러한 단일항목 구조의 지지도와 신뢰도 계산은 특정 항목이 차지하는 비율과 규칙에 포함되는 모든 항목 집합의 비율을 전체 트랜잭션에 대해 반복적으로 계산하는 과정이 요구된다. 이러한 문제점을 개선하기 위해 본 연구에서는 데이터베이스 구성의 특성을 반영하는 다중항목 데이터를 기준으로 지지도와 신뢰도 계산을 빠르게 하기 위해 SQL 프로시쥬어 기반의 다중 항목 연관 알고리즘을 제안한다. 표 1은 장바구니 분석 데이터에서 단일 항목 구조를 다중 항목 구조로 변경하는 것을 나타내고 있다. 다중 항목의 데이터는 단일 항목 구조의 동일한 트랜잭션 번호에 대해 한 개의 레코드가 생성된다.

그림 2는 단일 항목과 다중 항목의 연관 집합 생성 과정을 나타낸 것으로 Trans_tbl은 입력되는 트랜잭션 테이블이고, Tot_C_k는 k단계의 항목 집합 카운트를 가진 임시 후보 항목이고, C_k, F_k는 각 k 단계의 후보 및 빈발 항목 집합을 의미한다. 단일 항목 필드 구조에서 후보 항목 집합은 이전 단계 F_k-1의 조인으

표 1. 단일 항목과 다중 항목 필드 구조

단일 항목 구조

T_id	Item
1	사과
1	콜라
1	빵
2	콜라
2	빵
2	오징어
3	사과
3	콜라

⇒

다중 항목 구조

T_id	$Item_1$	$Item_2$	$Item_3$
1	사과	콜라	빵
2	콜라	빵	오징어
3	사과	콜라	

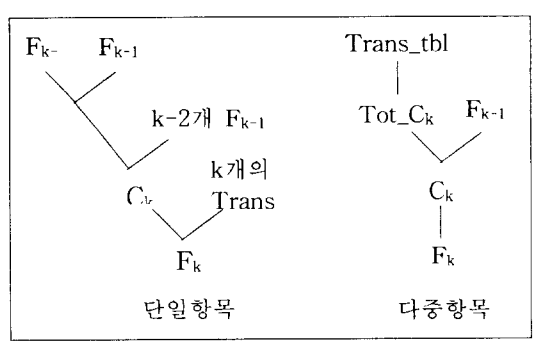


그림 2. 단일 및 다중 항목 연관 집합 생성 과정

로 생성되고, 빈발 항목 집합은 C_k와 k개의 트랜잭션 테이블과의 조인으로 생성된다. 다중 항목 필드 구조에서 후보 항목 집합은 트랜잭션 테이블의 조회로 생성되고 빈발 항목 집합은 C_k와 k개의 F_k-1 테이블의 조인으로 생성된다. 이러한 구조에서 보면 단일 항목 구조에서는 레코드 개수가 많은 트랜잭션 테이블과의 k번 조인으로 실행 시간이 많이 걸리게 되지만 다중 항목 구조에서는 트랜잭션 테이블의 조회에서 후보 항목을 바로 생성할 수 있기 때문에 수행 시간이 단축된다.

3.2 N-항목 연관 규칙 알고리즘 설계

최근의 인터넷 쇼핑몰은 다수의 상품에 대해 각각의 속성별 키워드 검색 화면을 생성하기 보다는 고객이 검색하려는 속성을 직접 선택하여 해당 속성별로 키워드 검색문이 자동적으로 생성되게 하는 검색방법을 많이 사용하고 있다[4]. 이러한 상품 검색 시스템은 기본적으로 다중항목에 기반을 두고 있다. 본 연구도 한 개의 레코드에 다수의 항목을 가지는 n-항목 입력 데이터에 대해 효율적인 연관 규칙을 탐지하고자 한다. 다중 항목 구성은 두 가지 유형이 존재할 수 있다. 한 가지는 동일 항목 필드가 동일한 속성을 가지고 있지 않을 경우로 장바구니 분석 시스템이 해당되며, 다른 한 가지는 동일 항목 필드에서 동일한 속성을 가지고 있는 경우로 상품 검색 시스템이 여기에 속한다[14,15].

본 장에서는 위의 두 가지 n-항목 데이터 유형에 대해 연관 규칙을 생성하는 프로시쥬어 기반의 다중 항목 연관 알고리즘을 제안한다. 다중 항목 연관 알고리즘을 적용하기 위해서는 상품 검색 시스템에서 고객이 선택한 속성에 따라 키워드를 생성하여 속성별로 키워드에 대한 테이블이 구성된다. 이와 같이 구성된 입력 데이터에서 연관성 있는 집합을 생성하기 위해 후보 항목 집합을 생성한 후, 최소 지지도 이상을 가지는 빈발 항목 집합과 최소 신뢰도 이상을 가지는 연관 집합을 생성하는 과정이 필요하다. 이를 위해 본 연구에서는 지지도와 신뢰도 계산을 빠르게 하기 위해 SQL 프로시쥬어 기반의 다중 항목 연관 알고리즘을 다음과 같이 제안한다. 그림 3은 연관 집합을 생성하는 전체 알고리즘을 나타낸다. 여기서 매개변수 C_k, F_k는 각 단계의 집합 테이블이고, @qn은 상품 속성에 해당하는 항목의 수, @k는 집합의 생성

```
procedure asso_ming(@qn)
 '상품 속성의 개수를 매개변수로 가짐
 begin
   while(@k<=@qn) begin '후보 집합 생성
     exec candi_gen @k, @qn
          '후보 집합 생성 프로시저
     select @supp=supp_value
          '단계별 지지도 입력
     exec freq_gen @k, @supp
          '빈발 집합 생성 프로시저
     select @k=@k+1
   end

   select @k=2          '빈발 집합 생성
   while(k<=@qn) begin
     select @conf=conf_value
          '단계별 신뢰도 입력
     exec rule_set_gen @k, @confi
          '규칙 집합 생성 프로시저
     select @k=@k+1
   end
 end
```

그림 3. 연관 집합 생성 알고리즘

단계를 나타내며, @supp는 지지도, @confi는 신뢰도를 의미한다. 첫 번째로 호출되는 프로시저로 candi_gen()은 각 단계의 후보 집합을 생성하기 위해 그림 4에 있는 프로시쥬어를 호출한다. 여기서 query_stmt()는 후보 항목 집합을 생성할 때 count()와 그룹 함수를 사용하여 각 연관 집합의 레코드 개수를 계산한다. 다음 단계로 freq_gen()는 단계별로 지지도에 따라 빈발 항목 집합을 생성하는 프로시쥬어를 호출한다. rule_set_gen은 최소 신뢰도 이상을 가지는 집합을 생성하기 위해 규칙 집합 생성 프로시쥬어를 호출한다. 그림 4는 후보 항목 집합 생성 알고리즘을 나타낸 것으로 k 단계와 상품 속성의 개수를 매개변수로 받아서 단계별 후보 집합을 생성하기 위해 후보 집합 생성에 대한 쿼리문을 호출한다. 프로시쥬어 qury_stmt()는 각 단계별 후보 집합을 생성하는 쿼리문으로 구성되어 있다.

그림 5는 빈발 및 규칙 집합을 생성하는 프로시쥬어를 나타낸 것이다. 프로시쥬어 freq_gen()은 입력 매개변수로 지지도를 입력받아서 이전에 생성된 후보 항목 집합을 이용하여 최소 지지도 이상을 가지는 빈발 항목 집합을 각 단계별로 찾아낸다. 프로시쥬어 rule_set_gen()은 입력 매개변수로 신뢰도를 입력받아서 이미 생성된 빈발 항목 집합을 이용하여 최소

```
procedure candi_gen(@k, @qn)
'k 단계와 상품 속성의 개수를 매개변수로 가짐
 begin
   select @i1=1
   while(@i1 <= @qn) begin
        '단계별로 후보 집합 생성
     exec qury_stmt @k, @col, @hv
        '후보 집합 생성에 관한 쿼리문 수행
     select @i1=@i1+1
   end
end

procedure qury_stmt(@k, @col, @hv)
 '단계별 쿼리문 수행
 'k단계,@col : k개 그룹으로 구성된 항목 필드
 '@hv : 모든 레코드가 가지는 k개의 필드
begin
 select @col_stmt='insert into c'+@k+' select '
  +@col +', count(*) cnt from st group by '
  +@col+ ' having not ('+@hv+')'
  +' order by '+@col
  ' 단계별 후보집합 쿼리문 생성
  execute(@col_stmt)
end
```

그림 4. 후보 집합 생성 알고리즘

```
procedure freq_gen(@k, @supp)
  'k 단계와 지지도를 매개변수로 가짐
 begin '빈발 집합 생성
  select @f_stmt='insert into cf'+@k+'
         select * from cf'+@k+' where
         cnt >= ' + @supp
        'k 단계의 빈발 집합 생성 쿼리문
   execute(@f_stmt)
 end
 procedure rule_set_gen(@k, @confi)
  'k 단계와 신뢰도를 매개변수로 가짐
 begin  '규칙 집합 생성
  while(@i<=@kc/2) begin
    while(@j<=2) begin
      exec rqrystmt @k, @i, @t
        'k 단계별 규칙 집합 생성 쿼리문
      select @f_stmt='select '+@itemq2+'
       ,cnt,'+'cnt/25 as supp,
       cnt/('+@fqry+') as confidence
       from f'+@kc
      execute(@f_stmt)
         'k 단계별 신뢰도 계산
     end
   end
 end
```

그림 5. 빈발 및 규칙 집합 생성 알고리즘

신뢰도 이상을 가지는 규칙 항목 집합을 생성한다.

3.3 다중 항목 검색 유형 데이터베이스 구성

연관 규칙 알고리즘으로 생성된 연관 집합은 키워드 검색에서 수집한 데이터에서 최소 지지도와 최소 신뢰도를 만족하는 실제 데이터 집합으로 구성되어야 한다. 그러나 앞에서 생성한 연관 알고리즘은 빠른 계산을 수행하기 위해 각 속성에 따른 데이터가 간단한 코드로 변경되어 수행되기 때문에 이를 실제 데이터로 변환하는 과정이 필요하다. 예를 들어 n-항목을 가지는 어떤 상품이 있을 경우 고객의 요구는 다양하게 키워드를 선택할 수 있고, 이러한 키워드를 이용하여 고객이 요구하는 연관 규칙 집합은 여러 가지의 속성들이 포함되는 다양한 형태의 검색 유형이 구성될 수 있다. 이와 같이 구성된 키워드 기반의 연관 규칙 집합은 실제 상품 데이터베이스에서 검색 속성과 일치하는 데이터를 검색 유형 데이터베이스에 저장하는 과정이 필요하다.

그림 6은 연관 집합 생성에서부터 검색 유형 데이터베이스 생성 과정을 나타내고 있다. 고객이 입력한 키워드 검색 데이터로부터 상품 속성별 연관 집합을

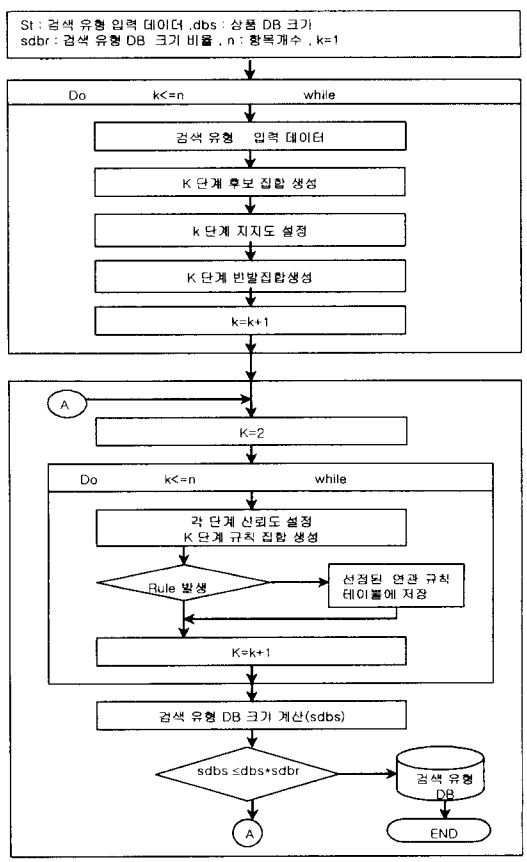


그림 6. 검색 유형 데이터베이스 생성 흐름도

생성한다. 검색 유형 데이터베이스 구성은 전체 상품 통합 데이터베이스에서 지지도와 신뢰도를 기반으로 최종적으로 상품 속성별 연관 집합 데이터들에 대해 빠른 검색을 위해 별도의 데이터베이스가 구성된다. 이러한 구성은 상품이 차지하는 전체 비율에 따라 휴리스틱한 방법으로 지지도를 결정하고, 신뢰도는 상품군별 데이터 크기와 이에 대한 검색 유형 데이터 크기 비율에 따라 결정하도록 하였다. 이와 같은 조건에서 규칙 집합이 생성되면 선정된 규칙 집합은 별도의 테이블에 저장된 후 실제 데이터베이스에서 일치하는 데이터를 검색 유형 데이터베이스에 저장하게 된다. 검색 유형 데이터베이스의 크기는 신뢰도의 설정 값에 의존하기 때문에 설정한 검색 유형 데이터베이스보다 큰 경우는 신뢰도를 다시 설정하여 연관 집합을 생성하도록 하였다.

4. 실험적인 검색 유형 데이터베이스 구현 및 성능 분석

4.1 웹 기반의 상품 검색 시스템 구성

본 연구에서는 인터넷 쇼핑몰에서 상품 검색 속도를 빠르게 위한 방법으로 다중 항목 연관 규칙과 검색 유형 데이터베이스를 고려하여 고객의 의사결정을 지원하는 상품 검색 시스템을 그림 7과 같이 제안한다. 상품 검색 시스템 구성은 상품 검색을 수행하는 사용자, 상품을 제공하는 제공자, 사용자/제공자 인터페이스, 검색 및 모델 에이전트, 데이터관리 시스템, 모델관리 시스템 등으로 구성되어 있다. 특히, 이러한 구성에서 데이터관리 시스템은 빠른 상품 검색을 지원하기 위해 두 종류의 데이터베이스를 운영한다. 첫째는 상품 전체의 데이터를 가지고 있는 상품 통합 데이터베이스이고 두 번째는 전체 상품 데이터베이스에서 빠른 검색을 수행하기 위해 고객이 검색하는 유형 데이터를 수집한 후 연관규칙을 이용하여 생성된 데이터를 가지고 있는 검색 유형 데이터베이스를 제시하고 있다.

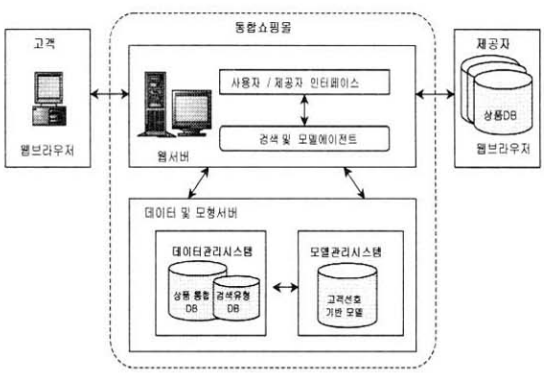


그림 7. 상품 검색 시스템 구성

본 연구에서는 그림 7과 같이 제안된 상품 검색 시스템 구성을 근간으로 하여 프로시쥬어 기반의 연관 집합 생성 알고리즘이 실제 검색 유형 데이터베이스 구성과 일치하는지 점검하고, 또한 성능 분석을 위해 프로토타입의 상품 검색 시스템을 구현하였다. 이를 위해 예제 시스템은 고객의 상품 검색 유형을 추출하기 위해 웹 기반의 키워드 상품 검색 시스템을 구축하였다. 상품 검색 시스템은 Windows NT 기반의 클라이언트에서 수행되고 웹서버는 마이크로 소프트사의 IIS4.0와 썬 마이크로시스템의 Tomcat3.2를 사용하였으며, 데이터베이스는 MSSQL Server2000을 사용하여 개발하였다. 개발 툴로는 Java 언어를 사용하였으며, 데이터베이스와의 연동을 위해 JDBC 드라이버(aveConnect JDBC2.3)를 활용하였다[16,17]. 그리고 웹서버와 데이터베이스 연결 및 동적 데이터를 처리하기 위해 Tomcat 엔진을 통한 JSP(Java Server Page)를 사용하였다[18].

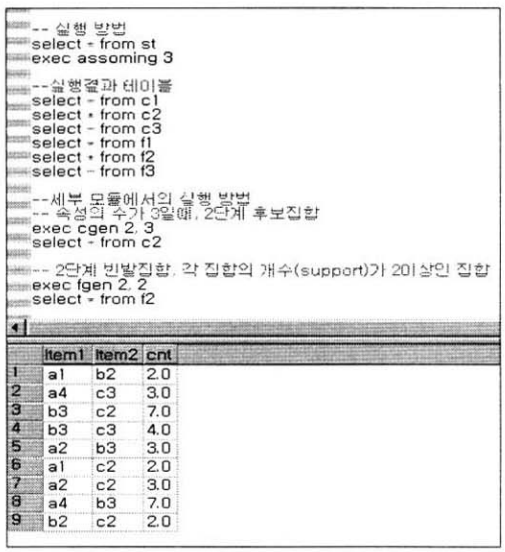
이와 같은 환경에서 그림 8은 키워드 기반의 각 상품을 선택할 수 있는 항목과 다른 속성들이 선택될 수 있도록 화면 구성을 하였다. 이러한 키워드 입력은 미리 정의된 코드에 따라 다중 항목 테이블에 저장하게 된다. 저장된 다중 항목 데이터들은 상품별, 속성별 지지도와 신뢰도에 따라 최종적인 연관 규칙 집합이 생성되면, 이러한 결과는 검색 유형 데이터베이스에 저장되게 된다. 따라서 고객이 입력한 키워드가 이미 검색 유형 데이터베이스에 있으면 빠른 검색

순번	제조회사	모델번호	부피	가격	제조일자	특징
6	삼성	SR-L5278	522	1,030,000	2000년 1월	서랍식냉동실
8	삼성	SR-518HZ	514	998,000	2000년 8월	웰보상도어
26	LG	R-B51CZ	514	1,098,000	2000년 8월	알뜨냉각
28	LG	R-B50CF	500	928,000	2000년 1월	고광택
41	대우	FRB-5270KB	520	1,058,000	2000년 8월	냉각속도2배
42	대우	FRB-5070SB	500	1,124,000	2000년 1월	냉각속도2배

그림 8. 키워드 입력에 의한 상품 검색

결과를 제공하게 된다. 지지도와 신뢰도의 임계값 구성은 3.3절에서 간단히 설명하였지만, 본 연구에서는 빠른 검색 결과를 찾기 위한 연구이기 때문에 구체적인 임계값 설정 기준은 제외한다.

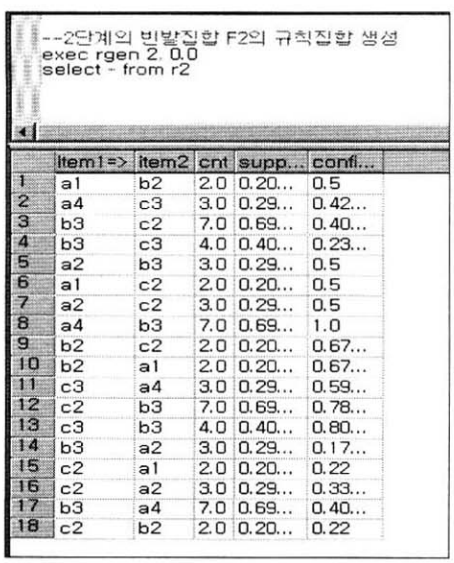
그림 9는 3.3절에서 제시한 검색 유형 데이터베이스 생성 흐름도에 따라 속성의 수가 3일 경우 단계 2의 후보 항목 집합과 빈발 항목 집합을 생성하는 과정을 나타낸 것이며, 그림 10은 2단계의 빈발 집합



```
-- 실행 방법
select * from st
exec assoming 3

--실행결과 테이블
select * from c1
select * from c2
select * from c3
select * from f1
select * from f2
select * from f3

--세부 모듈에서의 실행 방법
-- 속성의 수가 3일때, 2단계 후보집합
exec cgen 2, 3
select * from c2

-- 2단계 빈발집합, 각 집합의 개수(support)가 2이상인 집합
exec fgen 2, 2
select * from f2
```

	item1	item2	cnt
1	a1	b2	2.0
2	a4	c3	3.0
3	b3	c2	7.0
4	b3	c3	4.0
5	a2	b3	3.0
6	a1	c2	2.0
7	a2	c2	3.0
8	a4	b3	7.0
9	b2	c2	2.0

그림 9. 후보 및 빈발 집합 생성



```
--2단계의 빈발집합 F2의 규칙집합 생성
exec rgen 2, 0.0
select * from r2
```

	item1 =>	item2	cnt	supp...	confi...
1	a1	b2	2.0	0.20...	0.5
2	a4	c3	3.0	0.29...	0.42...
3	b3	c2	7.0	0.69...	0.40...
4	b3	c3	4.0	0.40...	0.23...
5	a2	b3	3.0	0.29...	0.5
6	a1	c2	2.0	0.20...	0.5
7	a2	c2	3.0	0.29...	0.5
8	a4	b3	7.0	0.69...	1.0
9	b2	c2	2.0	0.20...	0.67...
10	b2	a1	2.0	0.20...	0.67...
11	c3	a4	3.0	0.29...	0.59...
12	c2	b3	7.0	0.69...	0.78...
13	c3	b3	4.0	0.40...	0.80...
14	b3	a2	3.0	0.29...	0.17...
15	c2	a1	2.0	0.20...	0.22
16	c2	a2	3.0	0.29...	0.33...
17	b3	a4	7.0	0.69...	0.40...
18	c2	b2	2.0	0.20...	0.22

그림 10. 규칙 집합 생성

에서 규칙 집합을 생성하기 위해 실행되어진 결과값이다.

4.2 연관 규칙 테이블 생성

고객이 입력한 상품의 키워드 검색에서 수집된 데이터는 코드로 변경되어 저장되는데, 본 연구에서 제안한 알고리즘은 이 코드를 이용하여 연관 규칙 집합을 생성하게 된다. 그림 11은 그림 8의 검색 화면에서 선정된 속성을 기준으로 속성별 키워드 코드와 코드 이름을 가진 테이블 예제를 나타내고 있다. 속성별 코드 테이블의 코드 "ALL"은 상품 데이터에서 해당 항목 속성의 모든 레코드가 대상이 된다는 의미이고, 코드 이름에 있는 "all"은 그 속성 데이터 모두가 대상이 됨을 의미한다. 수치 데이터의 코드 값은 경계를 구분하기 위해 데이터의 상한과 하한의 값을 부여하였고, 상한은 데이터의 최대치를 하한은 데이터의 최소치를 의미한다.

그림 11과 같이 구성된 속성별 코드 테이블 내용은 그림 6에서 제시된 과정에 따라 설정된 지지도와 신뢰도에 의해 선정된 연관 규칙 테이블이 생성되는데, 이러한 테이블의 구성은 그림 12와 같은 코드로

```
select * from it1code
code cdname
---- ----------
a1   대우
a2   삼성
a3   동양
a4   LG
ALL  all

select * from it2code
code cdname1     cdname2
---- ----------- --------
b1   0           400
b2   400         500
b3   500         600
b4   600         2000
ALL  0           2000

select * from it3code
code cdname1              cdname2
---- -------------------- --------
c1   0                    900000
c2   900000               1200000
c3   1200000              1400000
c4   1400000              5000000
ALL  0                    5000000
```

그림 11. 속성별 코드 테이블

```
item1 item2 item3
----- ----- ------
ALL   b3    ALL
a4    b3    ALL
ALL   b3    c2
ALL   b3    c3
a1    b2    c2
a2    b3    c2
a4    b3    c3
```

그림 12. 연관 규칙 테이블 생성

구성된다. 생성된 다중 항목 item1, item2, item3은 각 상품별 속성을 나타내고, 첫 번째 생성된 연관 집합은 코드 b3를 가진 데이터를 의미한다. 즉, 속성 코드 테이블을 참조하면 냉장고 상품에서 부피가 500에서 600인 모든 상품 데이터가 포함된다.

4.3 성능 분석

다중 항목 기반의 키워드 검색 입력 데이터들에 대해 제안된 다중 항목 연관 알고리즘을 이용하여 자주 검색하는 검색 유형을 별도의 데이터베이스에 저장해 두어 검색 속도를 높이는 실험을 수행하였다. 이를 위해 최소 지지도와 신뢰도에 따라 생성되는 데이터의 크기를 비교하였으며, 인터넷 쇼핑몰에서 가장 중요한 검색 시간을 측정하기 위해 검색 유형 데이터 크기에 따라 고객에게 검색 결과를 응답하는 검색 시간을 측정하였다. 실험에 사용한 데이터는 냉장고 데이터를 대상으로 하였으며, 냉장고에 관한 전체 상품 데이터 테이블은 500개의 레코드로 구성하였고, 각 레코드의 크기는 92바이트로 하였다.

(1) 검색 유형 데이터베이스 크기 분석

고객의 키워드 입력 데이터는 50에서 250까지의 트랜잭션 수를 수집하여 실험에 사용하였다. 수집된 데이터는 앞에서 제시한 다중 항목 연관 알고리즘을 사용하여 검색 유형 데이터베이스를 위한 연관 규칙 테이블을 구축하였다. 구축된 연관 규칙 테이블에서 지지도와 신뢰도의 임계값을 변화시킬 경우 데이터 크기가 어떻게 변하는지 실험하였다. 이 때 검색 유형 데이터베이스 크기를 구하기 우해 데이터 크기는 생성되는 검색 유형 데이터 테이블의 수와 각 테이블당 레코드 수, 레코드의 전체 크기를 모두 곱해서 계산하였다. 그림 13의 (a)는 지지도가 0.3일 경우 (b)는 지지도가 0.7일 경우에 각각 신뢰도를 0.5, 0.7, 0.9로 변화함으로써 데이터 크기를 비교하였다. 실험 결과에서도 지지도와 신뢰도가 낮을수록 연관 집합이 많아져 상대적으로 검색 유형 데이터베이스의 크기는 커지며, 지지도와 신뢰도가 높을수록 연관 집합이 적기 때문에 검색 유형 데이터베이스의 크기는 적어짐을 보여주고 있다. 그리고 연관 규칙 집합에 의해 새롭게 구성해야 할 검색 유형 데이터베이스 크기는 실험 대상에 사용된 전체 상품 데이터가 10KB인 경우 약 22% 정도가 추가적으로 필요하다.

(2) 상품 검색 응답 속도 분석

인터넷 쇼핑몰의 상품 검색 시스템은 내부적으로 성능이 좋은 알고리즘을 가지고 있더라도 고객에게는 빠른 응답 결과를 제공하지 못하면 의미가 없다. 본 절에서는 고객이 검색하고자 하는 키워드 입력에 대해 전체 상품 데이터베이스 또는 검색 유형 데이터베이스에 접근하여 검색 결과를 응답해 줄 때까지의 누적 시간을 측정한 것이다. 고객이 입력한 다중 항목 키워드 입력에 따라 검색 유형 데이터 크기를 5KB에서 20KB까지 변화시키면서 검색 유형 데이터베이스를 구축한 후 검색 수행 시간을 실험하였다. 그림 14의 결과를 보면 검색 유형 데이터 크기가 클수록 검색 수행 시간은 감소되는데, 이는 사용자가 입력한 키워드의 많은 부분이 검색 유형 데이터베이스에서 검색되기 때문이다. 그리고 고객으로부터 입력되는 트랜잭션의 횟수가 증가할수록 검색 시간은

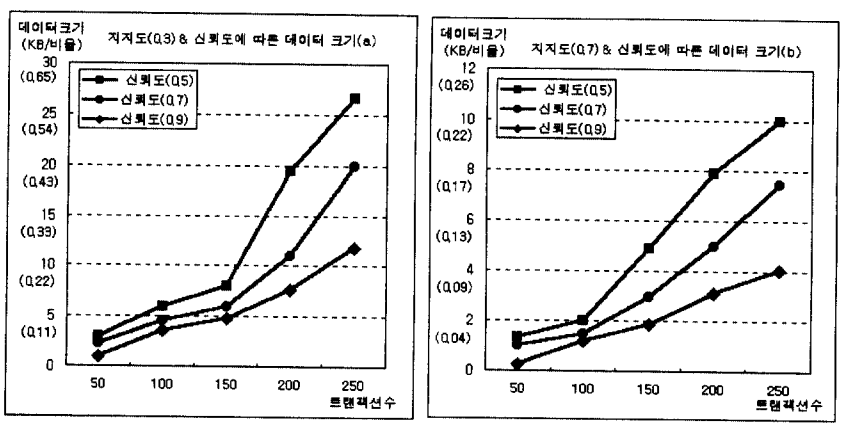


그림 13. 검색 유형 데이터베이스 크기 비교

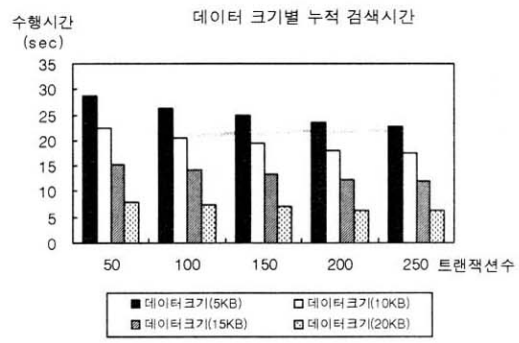


그림 14. 검색 회수별 누적 시간 비교

점차 줄어들을 보여주고 있다. 예를 들면 데이터 크기가 15KB일 경우 트랜잭션 수가 50보다 250일 경우 약 3초 감소됨을 보이고 있다. 이는 트랜잭션의 수가 늘어나면 연관 규칙 항목들이 증가하는 경우가 많을 것이며, 이는 추가적인 검색 유형 데이터베이스를 만들게 되어 전체적으로 검색 시간이 줄어들게 됨을 의미한다.

5. 결 론

인터넷 쇼핑몰의 가장 중요한 요구 조건은 사용자에게 편리한 검색 기능과 빠른 검색 결과를 제공하는 것이다. 이를 위해 대부분 인터넷 쇼핑몰은 편리한 검색 기능을 제공하기 위해 키워드 검색 기반 기능을 제공하고 있으며, 빠른 검색 결과를 제공하기 위해 응용된 연관 마이닝 기법을 적용하고 있다.

본 연구에서는 현재 사용하고 있는 쇼핑몰의 검색 기능과 기존의 데이터베이스 구조를 최대한 활용하여 빠른 검색 기능을 제공하는데 목적을 두었다. 이를 위해 사용자의 다양한 요구를 반영하기 위해 다중 속성 키워드 입력과 빠른 검색 기능을 가지기 위해 연관 규칙 알고리즘이 적용된 검색 유형 데이터베이스 구성 방법을 연구하였다. 본 연구는 현존하는 상품 검색 데이터베이스 시스템을 활용하기 위해 프로시쥬어 기반의 다중 항목 연관 알고리즘을 제안하였으며, 제안된 알고리즘에 의해 최종적으로 생성되는 연관 규칙 집합은 다중항목 검색 유형 데이터베이스에 구성되도록 설계하였다. 그리고 설계된 시스템의 성능 분석을 위해 웹 기반의 키워드 입력에 의한 상품 검색 시스템을 프로토타입으로 구현하였으며, 수행 속도를 개선하기 위해 속성별 코드 테이블을 별도로 구성하였다. 제안된 키워드 기반의 다중 항목 연관 알고리즘은 트랜잭션의 수가 늘어날수록 그리고 검색유형 데이터베이스의 크기가 증가할수록 검색 응답 속도는 줄어들고 있으며, 단점으로는 검색 유형 데이터베이스를 별도로 구성하는 것과 주기적으로 다중항목 연관 알고리즘을 수행해야 한다는 것이다.

참 고 문 헌

[1] 서찬일, 변효섭, "2003년 2월 사이버쇼핑몰 통계조사 결과," 통계청 서비스업 통계과, 2003. 4.

[2] 김충석, 김경창, "데이터 마이닝 질의 처리를 위한 질의 처리기 설계 및 구현," 정보처리학회논문지, 제8-D권 제2호, pp 117-124, 2001. 4.

[3] 김진옥, 황대준, "멀티미디어 데이터의 재발생 항목 마이닝을 위한 연관규칙 연구," 멀티미디어학회 논문지, 제5권 3호, pp 281-289, 2002. 6.

[4] S. Noel, V. Raghavan and C. Chu, "Visualizing Association Mining Results through Hierarchical Clusters," Advanced Issues of E-Commerce and Web-based Information System(WECWIS 2002), Proceeding Fourth IEEE International Workshop, pp 425-432, 2002.

[5] 황현숙, "연관마이닝 방법을 이용한 상품 검색 의사결정지원시스템 설계 및 구현," 부경대학교 박사학위 논문, 2001. 1.

[6] R. Agrawal, and R. Srikant, "Fast Algorithms for Mining Association Rules," Proceeding of the 20th VLDB Conference, 1994.

[7] N. Megiddo, and R. Srikant, "Discovering Predictive Association Rules", Proceedings of the 4th International Conference on KDD and DM, 1998.

[8] N. Megiddo, and R. Srikant, "Discovering Predictive Association Rules", Proceedings of the 4th International Conference on Knowledge Discovery in Databases and Data Mining, 1998.

[9] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," Proceeding of ACM SIGMOD Conference on Management of Data, pp. 207-216, 1993.

[10] 허용도, 이광형, "데이터 마이닝에서 IRG에 의한 효율적인 빈발항목 생성방법," 멀티미디어학회 논문지, 제5권 제1호, pp 120-127, 2002. 2.
[11] 이재문, "대화형 환경에서 효율적인 연관 규칙 알고리즘," 정보처리학회논문지, 제8-D권 제4호, pp 339-346, 2001. 8.
[12] S. Sarawagi, S. Thomas and R. Agrawal, "Integrating Association Rule Mining with Relational Database Systems: Alternatives and Implications," Data Mining and Knowledge Discovery, Vol.4, No.2, pp. 89-125, 2000.
[13] Houtsma, M. and Swami, "A Set-oriented Mining of Association Rules," Proceedings 11th International Conference on Data Engineering, 1998.
[14] 황현숙, 어윤양, "연관 마이닝과 고객 선호도 기반의 인터넷 상품 검색 시스템 설계 및 구현", 경영정보학회 논문지, 제12권 1호, pp 1-16, 2002.
[15] Jeon, T.G., Hwang, H.S., Kim, C.S., Shim, K.B. and Shim, D.S, "A Study on the Association Mining Algorithm for Intrusion Detection," Proceedings of International Conference on EALPIIT, pp 26-31, 2000.
[16] Art, T., JDBC Develpoer's Resource, Prentice-Hall Inc, 1997.
[17] H.W. Ian and F.K. Eibe, "Data Mining - Practical Machine Learning Tools and Techniques with Java Implementations", Morgan Kaufmann Pushlishers, 1999.
[18] L. Pekowsky, Java Server Pages, Addison-Wesley, 2000.

황 현 숙

1995년, 2001년 부경대학교 전산학 석사, 경영정보 박사
2001년 11월~2002년 10월 경남대학교 국내 박사 후 연수과정 수행
2003년 8월~현재 University of Missouri, Kansas City(UMKC)에서 국외 박사 후 연수과정을 수행
관심분야 : Data Mining, 의사결정지원시스템, 인터넷 비즈니스, 경매 시스템 등이다.

박 규 석

1980년 중앙대학교 대학원 전자계산학과(석사)
1988년 중앙대학교 대학원 전자계산학과(박사)
1982년~현재 경남대학교 정보통신공학부 교수
1992년~1996년 경남대학교 전산정보원 원장
1995년~1996년 한국 정보과학회 이사, 영남지부장
1999년~2002년 경남대학교 정보통신연구소 소장
2002년~현재 경남대학교 산업대학원 원장
2002년~현재 한국 멀티미디어 학회 회장
관심분야 : 분산처리 시스템, 정보통신 소프트웨어, 멀티미디어 시스템