韓國컴퓨터情報學會 論文誌
第15卷 第11號, 2010. 11.

2010-15-11-4-4

변형된 FP-Tree를 기반한 상품 추천 시스템

김 종 희*, 정 순 기**

The Goods Recommendation System based on modified FP-Tree Algorithm

Jong-Hee Kim*, Soon-Key Jung**

요 약

연관규칙 마이닝 기법 중에 하나인 FP-트리 알고리즘을 이용하는 추천시스템이 시도되고 있다. 본 논문에서는 트랜적션 데이터베이스로부터 빈발 2-항목집합만을 추출하여 연관규칙을 생성하는 변형된 FP-알고리즘을 사용하는 추천시스템을 제안하였다. 제안된 추천시스템은 전처리 모듈, 학습 모듈, 추천 모듈 및 평가 모듈로 구성되었다. 제안된 추천시스템의 실험을 통하여 상품 추천의정확률과 재현율과 F-Measure와 성공률과 추천실행시간을 수행하였으며, 순차패턴 마이닝 기법을 사용하는 추천시스템과의 성능을 비교분석 하였다. 순차패턴 마이닝기법을 사용하는 추천시스템과 학습 성능, 추천 성능을 비교한 결과 학습 성능은 5배 이상 향상되었으며, 추천 성능은 20%이상 향상 되었다. 결론적으로, 순차패턴 추천시스템과 같은 데이터를 가지고 실험하여 추천시스템 성능의 타당성에는 보다 나은 시스템임을 입증 하였다.

Abstract

This study uses the FP-tree algorithm, one of the mining techniques. This study is an attempt to suggest a new recommended system using a modified FP-tree algorithm which yields an association rule based on frequent 2-itemsets extracted from the transaction database. The modified recommended system consists of a pre-processing module, a learning module, a recommendation module and an evaluation module. The study first makes an assessment of the modified recommended system with respect to the precision rate, recall rate, F-measure, success rate, and recommending time. Then, the efficiency of the system is compared against other recommended systems utilizing the sequential pattern mining. When compared with other recommended systems utilizing the sequential pattern mining, the modified recommended system exhibits 5 times more efficiency in learning, and 20% improvement in the recommending capacity. This result proves that the modified system has more validity than recommended systems utilizing the sequential pattern mining.

• 제1저자 : 김종희 교신저자 : 정순기
• 투고일 : 2010. 06. 03, 심사일 : 2010. 06. 17, 게재확정일 : 2010. 06. 20.
* 선문대학교 IT교육학부 전임강사 ** 충북대학교 컴퓨터공학과 교수

▸ Keyword : Recommendation System, FP-tree, 연관 규칙, 빈발 2-항목집합(L2)

Ⅰ. 서 론

본 논문에서는 기존 순차패턴 마이닝 기법을 이용하는 추천시스템의 처리속도와 추천 정확도 문제점을 개선하기 위하여 FP-트리 알고리즘을 이용하는 개인화 추천시스템을 제안한다. 대용량 트랜잭션 데이터베이스의 처리성능을 향상시키기 위하여 FP-트리에서 빈발 2-항목집합(L2) 만을 대상으로 연관규칙을 생성하는 변형된 FP-트리 알고리즘을 사용하여 메모리 증가를 줄이고 실행 속도를 개선하고자 한다. 빈발 2-항목집합으로부터 추출된 연관규칙들로 구성된 지식베이스와 고객의 과거 상품구매 정보를 이용하여 고객별 상품 추천 리스트를 생성한다. 지식베이스를 기반으로 4가지 관심지표(지지도, 신뢰도, 향상도 및 지지도와 신뢰도) 값을 가중치로 사용하여 추천시스템의 정확률, 재현율, F-Measure, 성공률 및 추천 실행시간을 분석하며, 분석 결과를 순차패턴 기반 추천시스템의 성능과 비교, 평가한다.

Ⅱ. 관련 연구

1. 관련연구

1.1 데이터 마이닝 기법

인터넷 쇼핑몰은 시간과 장소에 관계없이 이용할 수 있다는 장점과 생활패턴의 급속한 변화로 인해 다양한 종류의 상품이 거래되면서 쇼핑몰 이용자 수가 급속히 증가되고 있다. 쇼핑몰 이용자에게 상품 구매를 유도하기 위하여 이용자가 자주 방문하는 상품 카테고리에서 가장 많이 팔리는 상품을 이용자에게 제안해주는 추천시스템이 널리 이용되고 있다. 다양한 종류의 상품들로부터 고객의 선호도를 고려하여 상품을 추천하는 방법으로는 협업 필터링 방식과 내용기반 필터링 방식 등이 있다. 그러나 두 가지 방법 모두가 이용자 수나 상품 수에 비하여 데이터 처리량이 비선형적으로 증가하므로 현재와 같이 대형화되고 있는 인터넷 쇼핑몰의 추천시스템에 적용하는 데는 어려움이 있다.

대용량 데이터베이스로부터 숨겨진 지식을 찾아내는데 이용되고 있는 데이터 마이닝 기법은 기존 온라인 정보처리 시스템에서 데이터 분석시 취약점을 보완할 수 있다. 1980년대부터 모든 조직에서 고객, 경쟁자 및 제품에 관한 데이터를 저장하고 있는 데이터베이스를 정보인프라로 인식하기 시작했으며, 데이터베이스로부터 숨겨진 정보를 검색하기 위하여 데이터 마이닝 기법을 도입하였다. 방대한 데이터를 저장할 수 있는 대용량 데이터베이스의 출현으로 기존 SQL(Structured Query Language)이나 간단한 질의어를 이용한 정보검색에는 한계가 있으므로 데이터 마이닝 기법으로 연관규칙 마이닝과 순차패턴 마이닝등이 이용되고 있다. 본 논문에서 제안하는 개인화 추천시스템은 연관규칙 마이닝을 기반으로 한다.

1.2 연관규칙 마이닝

연관규칙 마이닝은 Agrawal 등이 상품의 신뢰도 및 지지도에 대한 개념과 함께 제안 하였다[1]. 연관규칙은 상품구매 트랜잭션시 항목간의 종속관계로부터 유추할 수 있는 일련의 상품구매 패턴을 의미하다.

예를 들면, “기저귀를 구매하는 많은 고객들은 맥주도 구매한다” 로부터 유추할 수 있는 연관규칙은 “기저귀 ⇒ 맥주 [지지도=2%, 신뢰도=60%]” 로 표현된다. 항목들 간에 의미있는 관계성을 추출하기 위해서는 항목들을 비교할 수 있는 측정단위(척도)가 필요하다. 데이터 마이닝에서 사용하는 규칙의 흥미도를 측정하는 단위로는 지지도, 신뢰도를 사용한다. 연관규칙에서 지지도 2%는 모든 트랜잭션의 2%가 기저귀와 맥주를 함께 구매한다는 것을 의미하며, 신뢰도 60%는 기저귀를 구매한 고객의 60%가 맥주도 함께 구매한다는 것을 의미한다. 일반적으로 연관규칙은 최소 지지도 임계값과 최소 신뢰도 임계값을 둘 다 만족할 때 강한 연관규칙을 갖는다. 이 임계값은 마케팅 관리자에 의해 결정된다. 연관된 항목들 사이의 흥미 있는 통계적 상관성을 알아내기 위해 향상도를 반영한다. 향상도는 연관규칙 {A} ⇒ {B}가 강한 연관성으로 잘못 유도되는 것을 여과하기 위해 이용된다[1][2].

연관규칙 {A} ⇒ {B}에서 A, B가 항목집합 I의 부분집합일 경우 A⊂I, B⊂I 및 A∩B = ∅ 조건을 만족해야 한다. 다음 표 1은 연관규칙의 표현 형식을 나타낸다[1][2].

표 1. 연관규칙 형식
Table 1. Association rule format

{항목집합 A} ⇒ {항목집합 B}
if A then B : 만일 A가 일어난다면 B도 일어난다.

(1) 지지도

지지도는 전체 트랜잭션(거래) 개수 중에서 상품집합A와 상품집합B가 동시에 포함될 수 있는 트랜잭션 개수의 비율을 나타낸다. 지지도 S의 측정값은 다음 식(1-1)과 같이 계산된다[3].

$$\text{지지도 } S(A,B) = \frac{\text{상품집합A와 상품집합B를 포함하는 거래 개수}}{\text{전체 거래 개수}} \quad \text{식 (1-1)}$$

(2) 신뢰도

신뢰도는 상품집합A가 포함된 트랜잭션 개수 중에서 상품집합A와 상품집합B가 동시에 포함된 트랜잭션 개수의 비율을 나타낸다. 신뢰도 C의 측정값은 다음 식(1-2)과 같이 계산된다[3].

$$\text{신뢰도 } C(A,B) = \frac{\text{상품집합A와 상품집합B를 포함하는 거래 개수}}{\text{상품집합A를 포함하는 거래 개수}} \quad \text{식 (1-2)}$$

(3) 향상도

향상도는 지지도와 신뢰도뿐만 아니라 항목집합A와 항목집합B사이의 상관성에 의해서 계산된다. 상품A의 구매를 전제로 상품B를 구매하는 경우와 상품A의 구매를 전제 하지 않고 상품B를 직접 구매하는 경우 보다 얼마나 가능성이 높은지를 나타낼 수 있다. 향상도 L의 측정값은 다음 식(1-3)과 같이 계산된다[3].

$$\text{향상도 } L(A,B) = \frac{\text{상품집합A와 상품집합B를 둘 다 구입하는 확률}}{\text{상품집합A를 구입할 확률} \times \text{상품집합B를 구입할 확률}} \quad \text{식 (1-3)}$$

1.2.1 FP-트리 알고리즘

후보생성 없이 완전한 빈발 항목집합을 발견하는 기법으로 사용하는 빈발 패턴 증가또는 FP-증가라고 한다. FP-증가는 빈발 항목집합을 가지는 데이터베이스를 FP-트리로 라고 하며, FP-트리는 빈발 패턴길이가 길거나 짧은 경우 모두 매우 효율적이다. FP-트리는 압축된 데이터베이스를 하나의 빈발 항목에 대하여 연관되어진 조건부 데이터베이스의 집합으로 분할하게 된다. 분할된 각각의 데이터베이스에 대해서 마이닝 한다. 이 기법은 기존 연관규칙 알고리즘보다 속도가 빠르다고 알려져 있다[1][4][5].

FP-트리 알고리즘은 트리형태로 되어있으며 후보 항목집합 생성 과정없이 완전한 빈발 항목집합을 발견하는 기법으로 첫 번째로 데이터베이스를 스캔하여 빈발 항목집합들과 지지도 개수(발생빈도)를 추출한다. 다음은 데이터베이스 즉, 트랜잭션 데이터를 예로 들면 표 2와 같다.

표 2. 트랜잭션 데이터
Table 2. Transaction data

트랜잭션 ID	항목
T1	I1, I2, I4
T2	I1, I4
T3	I2, I3
T4	I2, I4
T5	I1, I2, I3, I4

트랜잭션은 T1, T2, T3, T4, T5로 5개 이며 각 트랜잭션은 장바구니라고도 불리며 하나의 트랜잭션에는 각각의 항목들이 존재한다.

최소 지지도 개수를 2라고 가정한 경우 빈발 항목집합은 지지도 개수를 기준으로 내림차순으로 정렬한다. 트랜잭션 데이터를 기준으로 내림차순 한 결과는 표 3과 같다. 지지도 개수별로 내림차순으로 정렬한 트랜잭션들은 표 4와 같다.

표 3. 지지도 개수의 내림차순 정렬(최소 지지도 개수=2)
Table 3. Descending sort of support count(minimum support count = 2)

항목	지지도 개수
I2	4
I4	4
I1	3
I3	2

표 4. 지지도 개수별 내림차순한 트랜잭션
Table 4. Transaction descending of support count

트랜잭션 ID	항목
T1	I2, I4, I1
T2	I4, I1
T3	I2, I3
T4	I2, I4
T5	I2, I4, I1, I3

FP-트리는 'null'로 표시된 트리의 뿌리를 생성한 후 데이터베이스를 스캔한다. 각 트랜잭션들의 항목들은 표 4의 순서

로 각 트랜잭션마다 가지가 생성된다. 예를 들면 다음과 같다.

1단계는 트랜잭션 T1은 'I1, I2, I4' 에서 내림차순한 항목 'I2, I4, I1'을 통하여, <I2:1>, <I4:1>, <I1:1>을 가지는 트리의 첫 번째 가지(branch)가 생성된다. I2는 뿌리의 자식으로 연결되고, I4는 I2에 연결되며, I2는 I4와 연결된다.

2단계 트랜잭션 T2는 'I1, I4'에서 내림차순한 'I4, I1'를 통하여, I4는 뿌리의 자식으로 연결되고, 두 번째 가지로 연결되고 I1은 I4의 자식으로 연결된다. 따라서, 노드 <I4:1>, <I1:1>를 새로 생성하여 <I4:1>의 자식으로 연결된다.

3단계 트랜잭션 T3은 'I2, I3'을 포함하고, I2는 뿌리의 자식으로 연결되고, I3은 I2와 연결되도록 가지를 생성한다. 그런데 이 가지에서 <I2>는 T1에서 이미 생성된 경로와 공통된 접두부인 <I2>를 공유하고 있다. 따라서, I2는 노드를 1만큼 증가시키며, 노드<I3:1>를 새로 생성하여, <I2:2>의 자식으로 연결된다.

4단계 트랜잭션 T4는 'I2, I4'를 포함하고, I2는 뿌리의 자식으로 연결되고, I2는 I2의 자식으로 연결된다. 따라서, <I2:2>는 이미 생성된 경로와 공통된 접두부로써 I2와 I4는 노드를 1만큼 증가시켜, 노드<I2:3>, <I4:2>를 갖게 된다.

5단계 트랜잭션 T5는 'I1, I2, I3, I4'에서 내림차순으로 정렬한 'I2, I4, I1, I3'을 통하여, I2는 뿌리의 자식으로 연결되고, I4는 I2의 자식으로 연결되며, I1은 I4의 자식으로 연결되며, I3은 I1의 자식으로 연결된다. 여기서 I2, I4, I1노드를 1만큼 증가시키고, 노드 <I3:1> 새로 생성하여 <I1:2>의 자식으로 연결한다.

일반적으로, 하나의 트랜잭션에 해당하는 가지가 추가되는 경우에는 공통 접두부에 속하는 각 노드의 수는 1씩 증가하고, 접두부 다음에 나타나는 항목에 대해서는 노드를 생성한 후에 그에 맞는 링크를 연결한다.

이러한 방식으로 트랜잭션 T5까지 빈발 항목집합에 대해서 노드를 생성한 후 링크를 연결한다. FP-트리에서 노드를 생성하는 과정은 다음 그림 1과 같다.

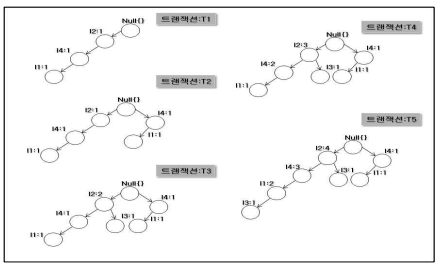

그림 1. 각 트랜잭션에 대한 노드생성과정
Fig. 1. Node generation process in each transaction

빈발 항목집합을 가지는 데이터베이스로 압축된 FP-트리에서 트리 탐색을 하기 위해 항목 헤더테이블을 만든 후 각 항목들은 노드-링크 형태로 트리내의 해당되는 노드와 연결된다. 표 4의 트랜잭션 T1에서 T5까지의 데이터베이스를 스캔하여 유도한 연관된 노드-링크구조는 그림 2와 같다.

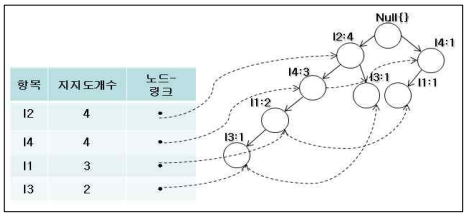

그림 2. 헤더 테이블을 만든 후 노드-링크 연결
Fig. 2. Node-link connection made header table

FP-트리 마이닝은 다음과 같이 진행된다. 초기의 접미부 또는 길이 1인 빈발 패턴에서 시작하여 조건부 패턴 베이스를 생성하고 조건부 FP-트리를 생성한다. 조건부 패턴 베이스는 FP-트리에서 접미부 패턴과 함께 발생되어지는 접두부 경로 집합으로 구성되는 부분 데이터베이스를 의미한다. 이 트리에 대해서 반복적으로 마이닝하는 과정을 트리-프로젝션 과정이라고 한다[1].

그림 2의 항목 헤더 테이블에서 가장 마지막 항목인 I3을 가지고 생성한 경우, 항목 I3의 접두부 경로를 따라가면 {I2, I4, I1:1}과 {I2:1}이며 이 접두부 경로들은 조건부 패턴 베이스를 생성한다. 만일 조건부 FP-트리에서 최소 지지도 임계값보다 작으면 제거된다. 이 단일 경로에서 빈발 패턴의 모든 조합들 {I2, I3 :2} 생성하게 된다.

항목 I1은 두 개의 접두부 경로가 조건부 패턴 베이스 {I2, I4,: 2}, {I4: 1}을 구성하고, 이는 단일 노드의 조건부 FP-트리{I4: 3}를 가지며, 하나의 빈발 패턴 {I4, I1:3}를 생성한다.

마지막으로 항목 I4는 접두부 경로가 조건부 패턴 베이스 {I2: 3}을 구성하고, 조건부 FP-트리는 단지 하나의 노드 {I2: 3}를 가지며, 하나의 빈발 패턴 {I2, I4: 3}을 생성하게 된다. 다음 FP-트리 마이닝 과정은 표 5와 같다.

표 5. FP-트리 마이닝 과정
Table 5. Mining process of FP-tr

항목	조건부 패턴 베이스	조건FP-트리	빈발 패턴 생성
I3	{I2,I4,I1: 1}, {I2: 1}	{I2: 2}	{I2,I3: 2}
I1	{I2, I4,: 2}, {I4: 1}	{I4: 3}	{I4,I1: 3}
I4	{I2: 3}	{I2: 3}	{I2,I4: 3}

위에서 분석한 방법에서 항목 I1의 조건부 FP-트리는 그림 3에서와 같이 1개의 가지를 갖는다. 여기에서 하나의 빈발 패턴 {I4, I1:3}을 생성한다.

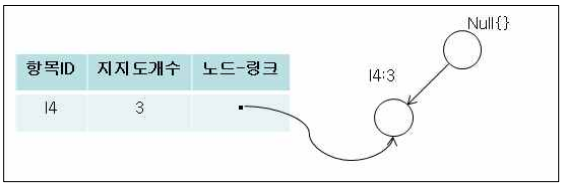

그림 3. 노드 I1와 연관된 조건부 FP-트리
Fig. 3. Node I1 and associated Condition FP-Tree

이 알고리즘에서 생성된 연관규칙의 형태는 그림 4의 예와 같다. I3의 빈발 항목집합은 {{I2, I3:2}이고, I4의 빈발 항목집합은 {I2, I4:3}이며, 빈발 항목집합 I1은 {I4, I1:3}} 를 생성한다. 연관규칙에 대한 예제이며 각각의 지지도, 신뢰도는 다음과 같다.

I2 ⇒ I3, (지지도 = 2/5 = 40%, 신뢰도 = 2/4 = 50%)
I2 ⇒ I4, (지지도 = 3/5 = 60%, 신뢰도 = 3/4 = 75%)
I4 ⇒ I1, (지지도 = 3/5 = 60%, 신뢰도 = 3/4 = 75%)

그림 4. FP-트리 알고리즘에서 생성된 빈발 항목집합
Fig. 4. frequent itemsets generated in FP-tree Algorithm

FP-트리 알고리즘은 데이터베이스를 두 번 스캔한다. 첫 번째 스캔은 길이가 1인 빈발 항목집합과 지지도 개수를 유도하기 위해서 필요하고 두 번째 스캔은 트리의 뿌리를 생성하고 난 후 사용하게 된다. FP-트리 알고리즘은 빈발하지 않은 항목은 제거함으로써 트랜잭션을 검사하는 비용이 절감하게 된다.

FP-트리는 데이터베이스가 크거나 빈발 항목집합이 많을 경우에는 많은 계산을 필요로 하며, 한 트랜잭션에 많은 빈발 항목집합이 존재할 경우에는 노드가 중복되어 메모리 크기가 증가하는 단점을 갖게 된다[1][5-6]. FP-트리를 이용한 트리-프로젝션방법이 Apriori 알고리즘 보다 빠른 성능을 보여주지만 조건부 FP-트리의 생성을 위한 메모리를 사용하기 때문에 많은 메모리양을 차지하게 된다[7].

III. 변형된 FP-트리기반 상품 추천시스템 설계

제안한 추천시스템은 인터넷 쇼핑몰에서 고객이 클릭한 상품들에 대한 정보가 날짜별로 저장된 웹 로그를 이용하는 전처리 모듈과 고객의 구매패턴을 파악하는 학습 모듈과 고객의 선호도를 분석하여 고객에게 최적의 상품을 추천해주는 추천 모듈로 구성된다. 추천 모듈에서는 목표 고객의 구매성향을 파악하여 고객이 선호하는 상품들을 나열하여 강한 연관규칙을 갖는 상품을 추천하는 조합 추천 알고리즘을 제안하여 적용한다.

본 논문에서 실험한 데이터는 가상의 데이터가 아닌 실제 상품을 클릭하고 구매한 고객의 실제 이용한 데이터를 가지고 실험에 적용하였다. 고객이 클릭 한 상품으로 제안한 추천시스템의 실험을 통하여 연관규칙을 생성하고 조합을 수행하는 구현 및 평가는 표 6에 나타나있는 환경에서 수행되었다.

표 6 추천시스템의 구현 환경
Table 6. Application environment of recommendation system

시스템	도구
OS	MS Windows Vista Home Premium K Service Pack1
CPU	Intell Core 2 Duo E6750 2.66G
RAM	4G
Lanaguage	Visual C++ 및 STL (Standard Template Library)

원본 로그 파일 중 추천시스템에서 필요한 형식으로 변환해주는 알고리즘을 구현하는 전처리 모듈과 연관규칙 생성을 위한 변형된 FP-트리 알고리즘을 구현하는 학습 모듈과 추천 데이터를 조합하는 추천 모듈과 추천 상품의 성능을 대용량 데이터베이스 기반의 순차패턴과 비교하는 평가 모듈의 구현 도구는 표 7과 같다.

표 7. 모듈의 구현 도구
Table 7. Tool application of Module

모듈		구현 도구
전처리 모듈	전반 15일	WiseLog Premium, SQL Server 2000
	후반 15일	
학습 모듈		Visual C++
추천 모듈		Visual C++
평가 모듈		SQL Server 2000

3.1 전처리 모듈

전처리 모듈에서는 흔히 사용되는 웹 로그 데이터를 분석하는 WiseLog Premium과 MS SQL Server를 이용하였다. 웹 로그 파일에는 불필요한 데이터들이 많이 포함되어있기 때문에 이를 분석하여 불필요한 데이터를 제거하는 전처리 과정이 필요하다. 본 논문에서는 상품을 추천하기 위한 방법을 제시하기 위해서 대형 온라인 쇼핑몰에서 실제 고객들이 클릭 한 행위 즉, 웹 로그 정보를 분석하여 학습 모듈과 추천 모듈에 사용할 수 있도록 하였다.

실험을 위하여 30일 동안 사용자가 상품을 클릭 한 정보를 가공하여 날짜별로 가져왔다. 본 추천시스템에서는 실험을 위해 사용된 데이터로는 표 8과 같다.

표 8 실험 데이터
Table 8. Test Data

내 용	범 위
로그 수집기간	30일
이용자 수	50만 명 이상
상품 수	100만 개 이상
전반 15일 클릭 한 상품	400만 건 이상

생성된 정보에는 고객ID 와 상품코드를 포함하는 형태로 가공하였다. 로그 파일은 총 400만 건 이상이다.

3.2 추천시스템의 구조

웹 로그에서 추출한 30일간의 데이터를 가지고 전체 추천 시스템의 구조를 그림 5와 같이 표현한다.

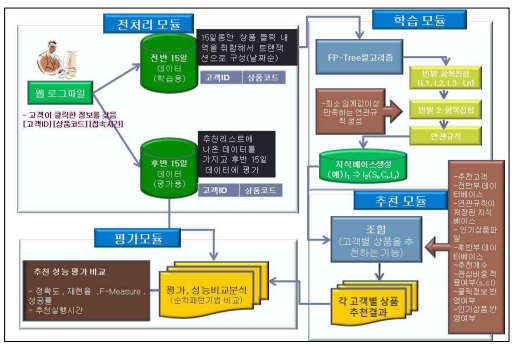

그림 5. 추천시스템 구조
Fig. 5. Architecture recommended system

다음 추천시스템 구조에서 상품을 추천하는 과정은 다음과 같다. 학습 모듈은 클릭한 고객 데이터를 기준으로 구매한 상품들 간의 유용한 연관 패턴을 찾아 추천 상품을 생성하는 모듈이다. 대용량의 데이터를 찾아내는 데이터마이닝의 모든 빈발 항목집합에서 변형된 FP-트리 알고리즘을 이용한 길이가 2(Size 2 Large Itemset)인 빈발 2-항목집합을 사용하여 최소 임계값을 만족하는 연관규칙 지식베이스를 생성한다.

본 논문에서는 그림 6과 같이 트리-프로젝션 과정을 사용하지 않은 빈발 2-항목집합만을 사용하여 연관규칙을 생성한다.

```
FP_T_learning_recommended_good_list();
입력: 학습용 dBT1()
출력: 변형된 FP-트리를 이용한 선호도가 높은 상품 리스트
begin{  // 최소 지지도 임계값 0%~100%, 최소 신뢰도 임계값 0%~100%, 최소 향상도 임계값 1~1,000,000
    tree = FP_Tree_Creation(); //전반부 데이터베이스를 스캔하여 FP-트리구조를 생성
    extract FP_tree_L2_traverse from dBT1(tree) //빈발 2-항목집합 생성
        using support, confidence, lift; //지지도, 신뢰도, 향상도 적용
    mining FP_tree_association_rule(item1, item2);
                        // 최소 임계값 이상 만족하는 길이가 2인 빈발 2-항목집합을 이용하여 연관규칙생성
        recommend good_list(support, confidence, lift);  //선호도가 높은 상품 리스트추천}
```

그림 6. 연관규칙 추출 알고리즘
Fig. 6. Algorithm mining Association rule

FP-트리 알고리즘에서는 트리-프로젝션 과정없이 직접 메모리에서 데이터를 접근하여 추천 리스트를 추출한다. 빈발 2-항목집합을 가지는 압축된 데이터베이스인 FP-트리를 생성한 후 지지도, 신뢰도 및 향상도를 추출하는 연관규칙을 탐색한다.

추천 모듈에서는 대상 고객이 접근하면 조합 추천 알고리즘을 이용하여 지식베이스로부터 검색하게 된다. 이를 통해 고객이 이전에 어떤 상품을 구매했는지를 확인하여 연관규칙 지식베이스를 기반으로 상품을 추천한다. 추천 상품을 선택하는 관심지표(interesting measure)로는 지지도, 신뢰도, 향상도 및 지지도와 신뢰도를 합한 4가지를 기준으로 측정한다. 추천결과 상품이 부족한 경우에는 가장 많이 팔린 상품을 중심으로 추천한다. 그림 7은 조합 추천 알고리즘을 나타낸 것이다. 조합 추천 알고리즘은 추천 고객 파일의 고객 ID와 전반부 데이터베이스 파일을 읽어 들인다. 읽어 들인 추천 고객의 고객ID를 인덱스화 한 후 학습된 연관규칙에서 좌측항목

을 기준으로 AVL 트리로 저장한다. 생성된 패턴들과 상품 인기 분포 파일들을 순위화 하여 읽어 들인다. 읽어 들인 추천 고객 파일에 대하여 고객 ID를 기반으로 고객의 과거 상품 리스트를 추출한다.

후반부의 상품 항목을 좌측 항으로 가지는 규칙을 연관규칙 지식베이스 패턴생성에서 추천결과를 추출한다. 추출한 각 항목은 파라메터를 변경하여 가중치를 계산하였다. 추천 상품 개수가 부족할 경우에는 인기 상품에서 추출한 항목을 포함하여 각 사용자의 추천결과를 생성하였다.

recommend good_list(support, confidence, lift);
입력: 추천 고객, 전반부 데이터베이스, 연관규칙이 저장된 지식베이스, 인기 상품파일, 후반부 데이터베이스, 추천 개수, 관심 비중 적용여부(지지도,신뢰도,향상도), 클릭정보 반영여부, 인기 상품 반영여부
출력: 각 고객 별 상품 추천결과

begin
for(고객데이터가 없을 때 까지)
begin
1. 학습된 연관규칙 지식베이스를 읽어들이고 Left item기준으로 AVL트리로 저장
2. 고객별 과거 구매 리스트를 읽어들인다. 고객 ID를 인덱스(index)화 한다.
3. 주어진 한명의 고객에 대해 고객ID를 기반으로 2번에서 준비된 데이터 구조를 사용한다.
4. 고객 과거 구매 리스트내의 각 항목(item)에 대해 다음을 수행
ⓐ 후반 고객 항목(item)을 Left item으로 가지는 규칙을 1번 결과에서 추출
ⓑ 각 항목(item)의 가중치를 계산하여 Max를 유지
5. 반환개수에 모자를 경우에는 인기상품에서 추출
end

그림 7. 조합 추천 알고리즘
Fig. 7. Alogoithm combination recommendation

평가 모듈에서는 각각의 가치 지표에서 얻은 최종 추천리스트 상품을 평가용 데이터와 비교하는 평가의 척도로 계산한다. 평가용 데이터를 적용하여 추천시스템에 적용해보고 실제 상품과 비교하여 추천의 정확도를 계산한다. 추천 정확도 계산을 하기 위하여 수행되어진 정확률, 재현율, F-Measure, 성공률을 이용한다. 시스템 성능과 처리속도는 대용량 데이터베이스를 기반으로 하는 군집화 및 순차패턴 기반의 대용량 추천시스템과 비교 분석한다.

IV. 성능 비교 분석

본 논문에서는 학습 모듈과 추천 모듈의 처리결과를 바탕으로 연관규칙에서 생성된 지식베이스를 가지고 제안한 조합 추천시스템의 성능을 평가한다. 학습 모듈에서는 전체 고객을 기준으로 생성한 연관규칙 생성 수와 순차패턴 생성 수를 비교 분석하였으며, 평가 모듈에서는 추천시스템의 성능 실험을 통하여 정확률, 재현율, F-Measure, 성공률을 순차패턴 시스템과 비교하여 성능의 우수함을 보여준다.

4.1 학습모듈 비교분석

본 논문에서는 웹 로그 데이터의 전처리 과정을 거친 후, 데이터 마이닝의 연관규칙을 이용하여 최소 지지도에 따른 추천 상품들을 생성하였다. 최소 지지도가 0.05%(0.0005)인 실행결과와 실행시간을 비교 분석하였다[8][9]. 다음의 표 9와 표 10은 패턴 생성수를 비교 분석하였다. 실험결과에서 나타난 패턴 생성 수에서 많은 차이가 나타남을 알 수 있다.

표 9. 변형된 FP-트리 알고리즘의 개인화 추천시스템
Table 9. Personalized recommend system of modified FP-Tree Algorithm

최소 지지도	0.05%(0.0005)	0.03%(0.0003)	0.01%(0.0001)	0.005%(0.00005)
빈발 패턴 수	5,422	16,072	122,002	372,444

표 10. 군집화 와 순차패턴 기반의 개인화 추천시스템
Table 10. personalized recommend system using clustering and sequential pattern

최소 지지도	0.05%(0.0005)	0.25%(0.0025)	0.01%(0.0001)
빈발 패턴 수	160	477	1,714

4.2 추천모듈 비교분석

본 논문에서는 고객 17만 명 이상에 대하여 실험을 하였으며, 추천 상품을 추출하는데 걸리는 시간은 2분을 넘지 않은 실행시간을 보여주었다. K-Means 클러스터링 기법을 사용한 순차패턴 기반의 성능평가와 연관규칙의 성능평가로는 정확률, 재현율, F-Measure, 성공률을 평가 측도로 사용하였다. 그림 8은 추천결과가 얼마나 정확한지를 나타낸다. 추천 상품수를 증가시키게 되면 정확률은 낮아지게 되지만 순차패

턴 기반보다 연관규칙 기반의 정확률이 10%이상 우수한 성능을 보였다.

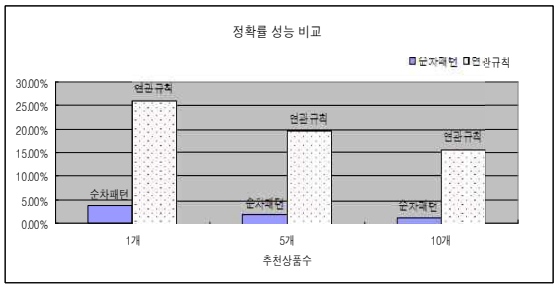

그림 8.순차패턴과 연관규칙의 정확률 성능 비교
Fig. 8. precision rate efficiency measure of sequential pattern and association rule

그림 9의 재현율은 추천 상품 수를 실제로 얼마나 찾았는가를 나타낸다. 정확률과 재현율은 서로 반비례 관계로 정확률이 점점 낮아지면 재현율은 점점 증가하게 되어 추천 상품을 선택할 수 있는 범위가 넓어지게 된다. 순차패턴 기반보다 연관규칙 기반의 재현율이 높게 나타남을 알 수 있다.

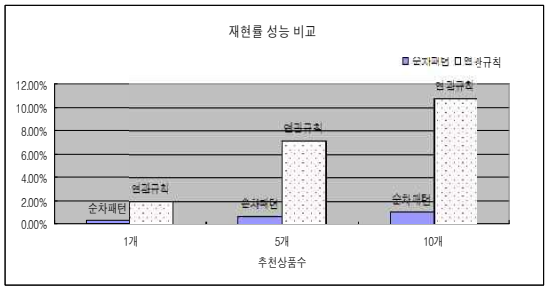

그림 9. 순차패턴과 연관규칙의 재현율 성능 비교
Fig. 9. recall rate efficiency measure of sequence pattern and association rule

그림 10의 F-Measure는 정확률과 재현율 사이에서 서로 보완해주는 상품 추천 성능을 갖는다. 추천 상품수가 1일 경우에는 2%의 성능을 보였으나 추천 상품수가 10개인 경우에는 11%이상의 보다 나은 성능을 보였다.

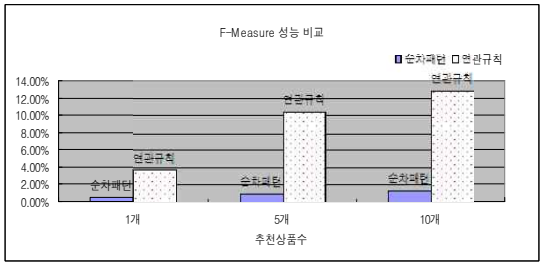

그림 10 순차패턴과 연관규칙의 F-Measure 성능 비교
Fig. 10. F-Measure efficiency measure of sequential pattern and association rule

그림 11을 보면 추천 상품수가 많아질수록 성공률은 증가됨을 보여준다. 이것은 추천 상품에 대한 성공여부를 결정한다고 볼 수 있다. 추천한 상품수가 많아질수록 성공률은 증가하게 된다. 추천 상품수를 10개로 한 경우 순차패턴 기법은 12%의 성공률을 보였고 연관규칙 기법에서는 55%의 성공률을 보였다. 본 논문에서는 추천 상품을 후반부에 클릭한 고객 중에서 과반수이상이 추천해준 상품을 클릭하였다는 것을 의미하게 되며 순차패턴 기법보다 연관규칙 기법이 보다 나은 성능을 나타내고 있음을 알 수 있다.

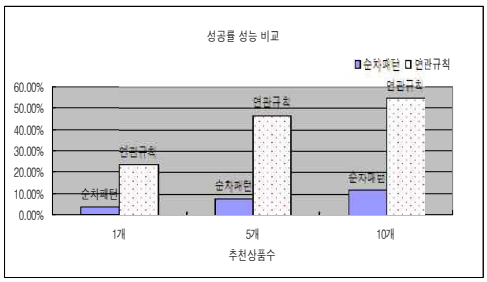

그림 11. 순차패턴과 연관규칙의 성공률 성능 비교
Fig. 11. success rate an efficiency measure of sequence pattern and association rule

IV. 결 론

기존 연관규칙 마이닝에서 사용되는 알고리즘(Apriori, DHP 등)은 빈발 항목집합을 발견하기 위해서 많은 후보 항목집합의 탐색을 필요로 하므로 반복적인 데이터베이스 스캔으로 인한 메모리 량과 처리 속도가 증가된다. 후보 항목집합을 탐색하지 않고 데이터베이스로부터 직접 빈발 항목집합을 생성할 수 있는 FP-트리의 경우도 대용량 데이터베이스에

서는 모든 빈발 항목집합의 생성은 처리속도 면에서 매우 비효율적이다. 따라서 본 논문에서는 빈발 2-항목집합 만을 대상으로 연관규칙을 추출하는 변형된 FP-트리 알고리즘을 사용하는 추천시스템을 제안하였다.

또한 상품 추천 성능을 순차패턴 기반 추천시스템과 비교 분석하였다. 고객 35만 명이상이 클릭한 상품을 처리하는 학습 모듈의 패턴 생성은 순차패턴 기반의 실험 결과 최소 지지도의 자동 수정 기능을 이용하여 순차패턴 생성에 24시간 정도가 소요되었으며, 본 논문에서 제안한 추천시스템의 연관규칙에서는 모두 3분 이내에 생성된 것으로 확인 되었다. 또한, 추천 모듈의 실행속도는 실험환경에서의 순차패턴 기법은 추천 실행 시간이 초당 2천명 이상을 처리하였고, 본 논문에서 제안한 추천시스템은 초당12,000명으로 순차패턴 기법보다 6배의 나은 성능을 보였다. 또한, 추천 모듈의 타당성을 비교하는 정확률은 순차패턴과 비교하여 20%, 재현율은 9%, F-Measure는 11%, 성공률은 42%의 나은 성능으로 나왔다. 따라서, 비교한 순차패턴 기법에서 학습 모듈과 추천 모듈의 정확성과 타당성에서는 보다 나은 추천시스템임을 입증하였다. 향후 연과 과제로는 대용량 데이터베이스에서 활용하기 위해서는 인터넷 환경에 적용 문제나 인기 상품과 함께 상품을 중복 추천하는 방법에 대한 연구가 필요하다. 또한 고객들은 온라인 쇼핑 뿐 만 아니라 모바일 환경에서도 쇼핑을 원하기 때문에 모바일 시스템과 추천시스템과의 연동방법에 대한 연구가 진행되어야 할 것이다.

참고문헌

[1] 강창완 · 강현철 · 박우창 · 승현우 · 용환승 · 이동희 · 이성건 · 이영섭 · 진서훈 · 최종후 · 한상태, "*데이터마이닝-개념과 기법*," 제 2판, 서울, 사이플러스, 2007년.

[2] 용환승 · 나연묵 · 박종수 · 승현우 · 이민수 · 이상준 · 최린, "*데이터마이닝*," 서울, 인피니티북스, 2007년.

[3] 김용 · 문성빈, "학습알고리즘 기반의 하이브리드 개인화 추천시스템 개발에 관한 연구," 한국문헌정보학회지, 제 39권, 제 3호, 75-91쪽, 2005년 9월.

[4] Rakesh Agrawal , Ramakrishnan Srikant, "Fast Algorithms for Mining Association Rules," in Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487-499, 1994.

[5] Jang-Sup Shim, Dong-Ha Lee, Soon-Key Jung, "A Large Real-Time Personalized Recommendation System Using Data Mining Techniques," KMIS International Conference, pp.712-716, 2005.

[6] 심장섭, K-means 군집화와 순차패턴 기법을 사용하는 VLDB기반의 추천 시스템 설계, 충북대학교 대학원 박사학위 논문, 2005년.

[7] 윤상균, 정보포털사이트의 고객관계관리를 위한 통합 데이터 마이닝 모형 연구 및 구현, 홍익대학교 대학원, 석사학위 논문, 2008년.

[8] 심장섭, K-means 군집화와 순차패턴 기법을 사용하는 VLDB기반의 추천 시스템 설계, 충북대학교 대학원 박사학위 논문, 2005년.

[9] 김종희 · 심장섭 · 이동하 · 정순기, "대용량 개인화 실시간 상품 추천 시스템 설계," 한국정보처리학회 학술발표논문집, 109-112쪽, 2006년.

저 자 소 개

김 종 희

1997년 : 호서대 전자계산학과 졸업, 동 대학원 석사(2001), 충북대 컴퓨터공학과 박사(2009),

2003년~2007년 : 강의전담교수

2009~현재 : 선문대학교 IT교육학부 전임강사

관심분야 : 데이터베이스, 실시간시스템

정 순 기

1982년 : Dortmund 대학 전산과, Dipl.Inf.

1994년 : Groningeneogkr 전산과 Dr.

1985년~ 현재 : 충북대학교 컴퓨터공학과 교수.

1994년 : 충북대학교 전자계산소장.

1998년 : 한국 과학재단, 한국기초과학협력위원회, 정보 분과 위원장.

2000년 : 충북대학교 중앙도서관장.

관심분야 : 실시간시스템, 소프트웨어 공학, DBMS.