실시간 추천을 위한 분할셋 기반 Up-to-Moment 선호모델 탐색☆

Mining the Up-to-Moment Preference Model based on Partitioned Datasets for Real Time Recommendation

한 정 혜* 변 루 나**
Jeong-Hye Han Lu-Na Byon

요 약

최근 들어 유비쿼터스 컴퓨팅에 대한 많은 연구들이 활발히 시작되고 있는데, 특히 모바일을 활용한 실시간 추천 모델에 대한 요구는 점차 커지고 있다. 본 연구에서는 기존 대용량 데이터베이스에서 실시간 추천을 위하여 Up-To-Moment 연관규칙 탐색 알고리즘이 있는데, 보다 더 정교하게 과거의 거래 세부정보까지 고려할 수 있도록 Up-To-Moment 데이터 셋의 과거 데이터 셋 부분을 (k-1)개로 분할-조합규칙을 적용하는 연관규칙 선호모델을 제안하였다. 제안된 모델은 전자상점 뿐만 아니라 유비쿼터스 컴퓨팅에 적용 가능한 레스토랑 음식 추천 데이터에 대하여, 전통적인 Up-To-Moment 연관규칙 탐색 모델 EM_{past}, 데이터 셋 크기값을 가중 조합한 EM^{w}_{past}, 그리고 시간에 따른 지수평활법 분할-조합규칙을 적용한 EM^{ES}_{past}을 비교하여 보았다. 특히 EM^{ES}_{past}의 지수평활 상수 α 값의 변화에 따른 세 알고리즘의 연관규칙 계산에 대한 민감도도 비교함으로써, 실제 데이터 적용 시에 보수적 또는 진보적 실시간 추천의 선택이 가능하도록 하였다. 세 알고리즘의 비교 시뮬레이션 결과를 보면, 데이터 셋 크기값을 가중 조합한 EM^{w}_{past}이 가장 효율이 떨어지는 것으로 나타났으며, 누적된 과거 데이터 셋의 크기가 클수록 EM^{ES}_{past}의 정확성이 높은 추천을 하는 것으로 나타났다.

Abstract

The up-to-moment dataset is built by combining the past dataset and the recent dataset. The proposal is to compute association rules in real time. This study proposed the model, EM_{past}, and algorithm that is sensitive to time. It can be utilized in real time by applying partitioned combination law after dividing the past dataset into(k-1). Also, we suggested EM^{ES}_{past} applying the exponential smoothing method to EM^{p}_{past}. When the association rules of EM_{past}, EM^{w}_{past}, and EM^{ES}_{past} were compared, The simulation results showed that EM^{ES}_{past} is most accurate for testing dataset than EM_{past} and EM^{w}_{past} in huge dataset.

☞ 키워드 : 실시간 추천(Real-Time Recommendation), 시간 연관 규칙(Temporal Association Rules), 실시간 데이터셋(Up-to-Moment Dataset), 분할-조합규칙(Partitioned Combination Law), 지수평활법(Exponential Smoothing Method)

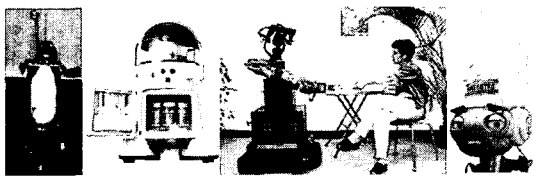
1. 서론

연관 규칙(Association Rule)은 장바구니 분석(Market-basket Analysis)을 통해서 얻어지는 것으로, 계산과 이해가 편하기 때문에, 전자상거래가 활성화됨에 따라 고객 개인의 관심에 부합하는 개인화된 정보나 상품 서비스를 제공하기 위하여 활발하게 연구되고 있다. 대부분 전자상점들은 CRM을 위해 다수의 고객에 대한 거래 데이터 웨어하우스의 분석을 통한 고정된 시점에서 연관규칙 등을 적용한다[2]. 그러나 전자상거래와 같이 많은 거래 데이터를 분석하여 크로스 셀링

* 정 회 원 : 청주교육대학교 컴퓨터교육과 조교수
hanjh@cje.ac.kr
** 정 회 원 : 보건복지부 사무관
lnbyon@korea.kr
[2007/01/10 투고 - 2007/01/11 심사 - 2007/01/31 완료]
☆ 이 논문은 2004년도 정부(과학기술부)의 재원으로[한국과학재단의 지원을 받아 수행된 연구임(D00573)]

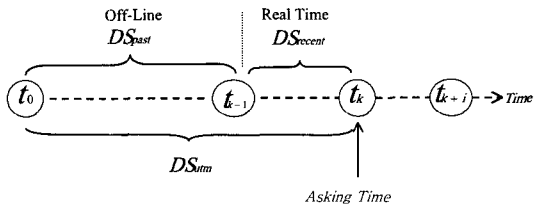
(Cross Selling)을 하는 경우에 매우 활발히 이용되고 있지만, 시간을 고려하지 않기 때문에 적시성(Just-in-Time)이 떨어지는 단점이 있었다. 이를 극복하기 위해 다양한 시간 연관규칙들이 제안되고 있다.

또한 유비쿼터스 컴퓨팅(Ubiquitous Computing)에 대한 연구와 산업적 응용이 활발한 경우 실시간 추천에 대한 전자거래 요구는 점차 커지고 있다. 예를 들면 로봇을 활용한 레스토랑의 음식 주문 및 배송 등도 활발히 연구되고 있는데, PDA뿐만 아니라 조만간 웨이터 로봇 역시 실시간 추천의 중요한 응용서비스 하드웨어가 될 것이다[15,16,17].



(그림 1) 음식을 추천 및 배송하는 로봇

실시간 추천은 매우 빈번히 바뀌는 전자상점 또는 유비쿼터스 컴퓨팅의 비즈니스 패턴에 대하여, 새로운 데이터 변화(고객이 특정 제품에 대해서 선호가 급격히 증감)가 즉시 반영되어야 할 것이다. 이를 개선하기 위하여 다양한 시간 연관규칙(Temporal Association Rule)들이 등장하고 있지만, 현행 대부분의 과거 데이터 셋들은 수 기가바이트(GB)에 이르는 등 파일 양이 매우 커서, 실시간으로 고객 선호 모델을 계산하여 추천하기가 매우 어렵다. 결국 오프라인에서 배치(Batch)파일로 구축되고 주기적으로 시간 연관규칙이 계산되므로, 기존의 과거 데이터 셋 기반 알고리즘으로는 비교적 안정적인(stable) 고객의 선호도만 반영할 수 있다[1,2,5,9,11]. 이에 Shen et al(2004)은 전자상거래에서 자주 변하는 소비자들의 선호를 파악하는 것을 목표로 하여 과거자료만이 아닌 최신 자료와 과거자료를 조합한 그림2의 Up-to-Moment 데이터셋 기반 모델을 제안하였다[9].



(그림 2) Up-to-Moment dataset의 구조

즉, 과거와 현재 진행 중인 데이터까지 포함한 Up-to-Moment 데이터 셋를 정의한 후, 이 2분할 셋을 조합하여 연관규칙을 계산하는 알고리즘을 제안하였다. 이 알고리즘은 과거자료에서 덜 빈번하더라도 Up-to- Moment 자료에서 꽤 빈번하게 발생하는 경우, 이 패턴들을 Up-to-Moment 자료가 반영할 수 있으며 계산양이 적어 실시간 추천이 가능하다는 매우 현실적 적용의 장점이 있다. 이러한 장점으로 Shen et al(2004)의 2분할을 (k-1)개로 분할하고, 데이터 셋 크기(결정치)를 고려한 지지도와 신뢰도에 따른 가중치를 이용하여 실시간 추천이 가능한 EM^{G}_{past}가 연구되었다[8].

본 논문에서는 Up-to-Moment 데이터 셋을 (k-1)개로 일반화 분할하고, 지지도와 신뢰도를 분할 셋에 대하여 지수평활법을 적용한 모델을 제시하였다. 4장에서 이 모델이 기존의 Shen et al(2004)와 EM^{G}_{past}와 어떠한 성능차이를 보이는지를 실시간 추천의 대표적인 적용 사례가 될 수 있는 레스토랑 데이터로 시뮬레이션을 통해 보이고자 한다. 마지막으로 5장에서는 결론 및 향후 연구에 대하여 논의하였다.

2. 관련 연구

시간 연관규칙이란 기존 연관규칙에 시간 개념을 추가하여 타임 스탬프된 데이터로부터 시간의 미와 시간관계를 가지는 유용한 지식을 탐사는 기법이다[6]. 타임 스탬프된 거래 데이터로부터 의미 있는 지식을 탐사하기 위하여, 과거 셋에 대

한 기존의 연관규칙과 현재 셋까지 고려한 여러 연구가 진행되고 있다.

2.1 과거 데이터 셋 기반 시간연관규칙

시간 속성을 가지는 연관규칙에 대한 이전 연구들은 크게 주어진 시간간격 동안 주기적으로 발생되는 현상, 즉 시간간격에서의 완전한 주기성을 만족하는 연관규칙을 탐사하는 주기적 연관규칙 탐사[3]와 캘린더로 표현된 시간패턴을 가지는 연관규칙을 탐사하는 캘린더 기반 연관규칙(calendar-based association)탐사[13]가 있다. 또한 분할 셋을 기반 연관 규칙(Partitioned Association Rules)[9], 누진적 가중 연관 규칙(Progressive weighted Association Rules)[5], 과거 셋에 기반한 분할 셋의 최적 크기 선정 등이 있다.

주기적 연관규칙이란 트랜잭션이 발생한 전체 시간을 사용자에 기반한 시간단위(년, 월, 일)에 따라 시간간격의 집합으로 나누고, 시간구간에서의 완전한 주기성을 갖는 모든 연관규칙들을 탐사한다[6]. 그러나 실제로 주어진 시간간격 동안에 완전하게 유지되는 규칙은 존재하지 않으며 대부분 불완전한 주기를 이루고 있다. 또한 주기적 연관규칙에서는 다양한 시간단위를 표현하지 못하고 단 하나의 시간단위만을 다룰 수 있다. 따라서 "매달 첫 번째 월요일"과 같은 실제 응용분야에서 적용되는 시간표현은 불가능하다.

캘린더 연관규칙(calendaric association)은 시간패턴의 명시를 위해서 캘린더 대수(calendar algebra)를 사용한다. 이 기법은 주기적 연관규칙보다 유용하고 실용적인 시간 연관규칙을 생성할 수 있으나 시간패턴 탐사를 위해 사용자 기반의 캘린더 대수 표현이 요구되어지며 이는 사용자로 하여금 탐사될 시간패턴에 대한 정확한 이전 지식을 필요로 하는 단점을 가진다.

[11]은 많은 이전 지식을 필요로 하는 캘린더 대수 표현들을 사용하는 것 대신에 시간패턴 탐사를 위한 프레임워크로 캘린더 스키마를 사용하였다. 이 접근방법은 주어진 캘린더 스키마에 대한 모든 가능한 캘린더 패턴들을 고려하고 [4]에서의 완전한 주기성을 탐사하는 것 대신에 캘린더 스키마에 의해 주어진 시간간격 동안 충분히 만족되는 불완전 또는 부분적인 주기성을 탐사한다. 따라서 일정 시간간격에서 만족되는 잠재적인 시간 연관규칙들의 탐사가 가능하다. 캘린더 패턴에 의한 시간패턴 탐사를 위해 시스템의 달력을 이용한 [12,13]에서는 시간 연관규칙을 위한 문제 제시, 마이닝언어(DMQL)와 규칙탐사를 위한 프레임워크를 제안하였고 각 시간단위와 시간간격을 이용하여 기존의 Apriori 알고리즘에 시간적인 표현을 추가한 형태로 규칙을 탐사한다. 또한 이 기법에서는 연관규칙의 효율적 탐사를 위해 주어진 시간간격들에서 일정기간 동안만 유지되는 규칙을 탐사할 수 있도록 최소 발생도를 정의하여 발생도를 고려한 연관규칙 탐사를 가능하게 하였다. 그러나 시간표현을 위한 방법으로 캘린더 시스템에 의존한 특정 시점이나 시간간격의 데이터를 분석하는데 그쳤으며, 시간에 따른 다양한 상호관련성을 제시하는 측면에서는 한계성을 가진다.

[1]에서는 전자상점의 매출에 영향이 크게 미칠 수 있는 계절상품이나 기획상품과 같이 시간의 변화에 민감한 추세를 갖는 상품들의 과거 데이터 셋에 대하여 최신 정보를 더 많이 고려하는 지수평활 시간 연관 규칙과 탐사하는 알고리즘을 제안하였다. 그리고 시뮬레이션을 통하여 제안된 방법이 기존의 시간 연관 규칙에 비해 실행시간이 다소 소요되는 것으로 나타났지만, 보다 좋은 예측력을 가짐을 보였다. [9]에서는 [8]의 실행시간을 줄이기 위한 최적의 과거 데이터 셋을 정하는 알고리즘을 제시하였다. 그러나 이러한 연관규칙은 모두 과거 셋 기반으로 실시간 추천은 어려운 단점을 갖고 있다. 이외에도 최근에는 사회경제학적이고 지리정보의 다양한 변화를 GIS기반에서도 공간-시간연관규칙과 각 변수에 대한 계층 분류기술의 다중 연관규칙 마이닝이 연구되었다[7].

2.2 Up-to-Moment 데이터셋 기반 시간연관 규칙

과거 데이터 셋 기반 모델은 빈번한 변화를 효과적으로 따라잡을 수 없으므로, 고객 선호도에 따른 제품 추천 성공률이 떨어지게 되는 것이다. 더욱이 기존의 방법은 과거 데이터 셋 모델의 룰(rule)집합과 새로운 데이터 셋 모델의 룰 집합의 합집합을 통하여 연관 규칙을 업데이트하여 생성을 하므로, 비록 모델을 업데이트하는 것이 새로운 모델을 구축하는 것보다 빠를 수는 있을지라도 여전히 상당한 시간을 요구하게 된다. 따라서 실시간 추천은 불가능하여, 적시성이 떨어졌다.

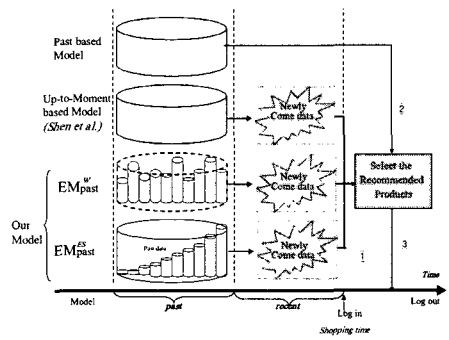
Shen et al(2004)은 이러한 기존의 시간 연관규칙의 단점을 극복하기 위하여, Up-to-Moment 데이터 셋을 정의하여 이를 분할하여 연관규칙을 가중 지지도(support)를 이용한 EM_{past}(Expanded past preference model)을 계산하였다[10]. EM_{past}은 과거자료에서 덜 빈번하더라도 Up-to-Moment 자료에서 꽤 빈번할 수 있는 패턴들을 표현하는 확장된 연관규칙을 포함하는 SM(standard preference model)에 대응하는 개념이다. 즉, 과거자료에서 빈번하게 발생하지 않는 몇몇 패턴들이 Up-to-Moment 자료에서는 빈번하게 발생하는 경우, 이 패턴들을 Up-to-Moment 자료가 반영할 수 있을 것이라는 접근에서 제안되었다. 또한 Up-to-Moment 데이터를 (k-1)개로 분할하고, 데이터 셋 크기(결정치)를 고려한 지지도와 신뢰도에 따른 가중치를 이용하여 실시간 추천이 가능한 일반화모델 EM^{G}_{past}가 연구되었다[8].

Shen의 알고리즘은 전자상점에서 고객이 로그인 하면 실시간 추천이 가능하다는 장점을 갖고 있는데, 과거 데이터 셋의 크기나 최신 정보에 대한 민감도는 고려되지 않아 과거 데이터 셋이 클 경우에는 민감도가 현저히 줄어드는 적용상의 한계점이 있으며, 분할 셋의 크기에 따라 다른 가중치를 주는 일반화 모델 $EM^{C}_{past} = EM^{w}_{past}$은 실행시간에 대한 시뮬레이션만 제시되어 효율성 비교가 되지 못하였다. 따라서 본 연구의 3장에서는 실시간 추천이 가능하면서도 Shen 알고리즘의 단점을 없애면서, 제안되었던 일반화 모델 EM^{w}_{past}와 보다 시간에 민감한 지수평활법 적용한 EM^{ES}_{past}의 정의 및 시간연관규칙 탐사 알고리즘을 제안하고 성능 비교를 하고자 한다.

3. 분할셋 기반 Up-to-Moment 연관 규칙과 탐색

3.1 분할셋 기반 Up-to-Moment 시간연관규칙

본 절에서는 Shen et al(2004)의 일반화된 형태로 Up-to-Moment 데이터 셋을 분할하여 지지도 가중치를 계산하는 방법에 따라, 지수평활필터를 적용한 EM^{ES}_{past}을 정의하고자 한다. 그림 3은 모델간의 데이터 셋 구조를 나타낸 것으로써 시간에 더 민감하게 반응하기 위하여 과거기반 데이터 셋을 k-1개의 부분 셋으로 분리한 그림이다. 분할 셋의 지지도에 따른 가중치를 이용한 선호모델 EM^{w}_{past}과 본 논문의 지지도에 지수평활 법을 적용한 EM^{ES}_{past}의 개념도이다



(그림 3) , SM_{past} EM_{past}, EM^{w}_{past}, EM^{p}_{past}.간의 개념도

지지도에 지수평활 법을 적용한 EM^{ES}_{past}를 정의하기 위하여 각 용어를 다음과 같이 정의한다. 타임 스탬프된 거래 데이터 셋 D를 동일 구간을 갖

는 k개의 서브 셋으로 분할한 DS_t, $t=1, 2, \cdots, k$이 있다고 하자. 이 때 $D=\bigcup_{t=1}^{k} DS_t$, $\varnothing=\bigcap_{t=1}^{k} DS_t$을 만족하며, 상품집합을 $\{p_1, \cdots, p_m\}$라 하자. 각 DS_t에 대하여 지역 지지도(*local support*) s_t, 지역 신뢰도(*local confidence*) c_t, D에 대하여 전역 지지도(*global support*) s, 전역 신뢰도(*global confidence*) c가 존재한다. 각각 사용자가 미리 정한 최소 지지도 ms, 최소 신뢰도 mc보다 큰 모든 연관 규칙으로 다음 정의와 같다. 거래한 기록 데이터 셋을 $|DS_{past_i}|=n_i$ $|DS_{recent_i}|=n_r$, $i=1, \cdots, k-1$ 라 표현하면 $|DS_{utm}|=\sum_{i=1}^{k-1} n_i+n_{recent_i}$ 라고 표현할 수 있으며, DS_{past_i}의 지지도와 신뢰도는

$$s_i=\frac{count(p_1, \cdots, p_m)}{|DS_{past_i}|} \geq ms\%,$$

$$c_i=\frac{count(\{p_1, \cdots, p_m\}, DS_{past_i})}{count(\{p_1, \cdots, p_{j-1}, p_{j+1}, \cdots, p_m\}, DS_{past_i})}$$

$\geq mc\%$이다. 이때 $N_{\max i}$는 i번째 분할 셋의 어느 날 한 가지 상품의 최고매출을 나타낸다고 하자.

Definition EM_{past}^{ES}에 대한 정의는 다음과 같다.

어느 시점 t에 대하여, 상품 크기가 $j \leq m$일 때, 분할 셋의 크기 k→∞로 간다면, 지지도와 신뢰도 s^{ES}, c^{ES}를 갖는

$(p_1, \cdots, p_{j-1}, p_{j+1}, \cdots, p_m) \rightarrow p_j(s^{ES}\%, c^{ES}\%)$

은 지수평활 시간 연관규칙이다. 단, $0<\alpha \leq 1$에 대하여 s^{ES}, c^{ES}는 각각 다음을 만족한다. 지지도

$$s^{ES}=\frac{\sum_{t=1}^{k} count(\{p_{t1}, \cdots, p_{tm}\}, DS_t) \cdot \alpha(1-\alpha)^{k-t+1}}{|D|} \geq ms\%$$

이고, 신뢰도 $c^{ES}=$

$$\frac{\sum_{t=1}^{k}[count(\{p_{t1}, \cdots, p_{tm}\}, DS_t)\ \alpha(1-\alpha)^{k-t+1}]}{\sum_{t=1}^{k}[count(\{p_{t1}, \cdots, p_{t(i-1)}, p_{t(i+1)}, \cdots, p_{tm}\}, DS_t)\ \alpha(1-\alpha)^{k-t+1}]} \geq mc\%$$

□

지수평활모델에 대하여 특정시점 t에서 k→∞일 때, $\lim_{j \to k} S_{jt}=\lim_{j \to k}[\alpha y_{jt}+(1-\alpha)S_{jt-1}]=\frac{\alpha(1-\alpha)^{k-t+1}]}{\alpha(1-\alpha)^{k-t+1}]}$

이므로, 최소 지지도와 최소 신뢰도 부등식의 양변에 극한 값을 취하여 정의하였다.

Theorem 지수평활법 일반화 모델 EM_{past}^{ES}에 대한 분할 조합 규칙 분할 셋 크기가 충분히 크다면 $EM_{past}^{ES}(SM_{utm}^{ES})$ 은 연관규칙 $p_1, \cdots, p_m \rightarrow p(s_3\%, c_3\%)$을 갖는다.

{i) 시간 연관규칙 집합 R_{past}에는

$p_1, \cdots, p_m \rightarrow p(s^{ES}\%, c^{ES}\%)$, R_{recent}에는

$p_1, \cdots, p_m \rightarrow p(s_2\%, c_2\%)$들이 존재한다.

이때,

$$s_3\%=\frac{s^{ES}\% * |DS_{past_i}|+s_2\% * |DS_{recent}|}{|DS_{utm}|} \geq ms\%,$$

$$c_3\%=\frac{s^{ES}\% * |DS_{past_i}|+s_2\% * |DS_{recent}|}{\frac{s^{ES}\% * |DS_{past_i}|}{c_1}+count_{p_1-p_m}} \geq mc\%$$

{ii) 시간 연관규칙 집합 R_{past}에는

$p_1, \cdots, p_m \rightarrow p(s^{ES}\%, c^{ES}\%)$, R_{recent}에는

$p_1, \cdots, p_m \rightarrow p(s_2\%, c_2\%)$들이 존재한다.

이때,

$$s_3\%=\frac{s^{ES}\% * |DS_{past_i}|}{|DS_{utm}|} \geq ms\% \text{ 이고,}$$

$$c_3\%=\frac{s^{ES}\% * |DS_{past}|}{\frac{s^{ES} * |DS_{past}|}{c^{ES}}+count_{p_1-p_m}} \geq mc\%.$$ □

proof. 일반화 분할 셋 모델 EM_{past}^{w}의 분할 셋 크기 k→∞로 하면 위의 Definition에 의해 EM_{past}^{ES} 가 정의된다. 이 때 [10]의 조합규칙(combination law)를 적용하면 위 증명 완료.

3.2 분할셋 기반 Up-to-Moment 시간연관규칙 탐색

방대한 양의 과거 데이터 셋을 모두 사용하는 것은 오히려 정보의 질을 떨어뜨리는 것이므로,

지식 추출자는 과거 데이터 셋의 크기를 적당히 조절하여 데이터 마트(Data Mart)를 구성해야한다. 따라서 본 절에서는 과거 데이터 셋의 정보량에 의하여 제안된 연관규칙 탐색을 제안하고자 한다. 소비자들이 현재 쇼핑 카트에 물건들의 조합을 $SC=\{p_1,\cdots p_m\}$ 라고 표현할 때 SC의 모든 상품들에서 Up-to-Moment 선호를 계산함으로써 소비자들을 위해 온라인상에서 상품 주문을 만들 수 있다. 즉, DS_{utm}에 있는 $s\% \geq ms\%$와 $c\% \geq mc\%$의 모든 연관규칙 $p_1,\cdots,p_m \rightarrow p(s\%,c\%)$의 EM_{past}와 DS_{recent}로부터 탐색한다. 분할 셋이든 분할 셋이 아니든 거대한 과거 데이터 셋을 고려할 경우 많은 계산양이 요구되는 점이 매우 취약한 점이라고 할 수있다. 따라서 이러한 계산양을 줄이기 위한 최적의 데이터 셋을 선정하는 것도 중요한 탐색절차에 고려되어야한다. [8]의 연구결과를 보면 분할된 Up-to-Moment 셋에 대한 EM^{W}_{past}와 EM^{ES}_{past}의 탐색 알고리즘은 분할(k)에 대한 추가 계산량이 요구되는 단점이 있으므로, [9]에서 제시된 방법으로 과거 데이터 셋을 최적화 하여야 한다. 따라서 분할된 Up-to-Moment 셋에 대한 EM^{W}_{past}와 EM^{ES}_{past}의 탐색 알고리즘은 충분히 큰 k값으로 분할한 후 최적의 과거 데이터 셋을 선정이 선행되어야 하며 다음 표 1과 같이 요약할 수 있다.

(표 1) 분할된 시간 연관규칙 탐색 단계

Given $D(All\ Datasets)$, $PL(Item\ List)$, $PR(Item\ Association\ Rules)$
Step0: Calculate the optimal size of past dataset (Procedure I~Procedure IV)
Step1: Partition D into $k-1$ subsets, $DS_1,\cdots,DS_{k-1}$
Step2: Calculate EM_{past} from each DS_i
Step3: Combine EM^{ES}_{past} with each weight function by *Definition* and Theorem (Procedure V).
Step4: Derive R_{past} and $count_{p1-pm}$ for DS_{recent} based on R_{past} Derive R_{recent} to obtain PL_{utm}.

Step0의 Procedure 정의는 [9]를 참고하면 된다. 즉, 과거 데이터 셋의 정보를 최대로 확보하면서 탐색비용을 최소로 줄이는 최적 데이터 셋 선정을 위하여 Procedure I과 Procedure II에서 시간 방향에 따른 정적 누적 셋 $db^{1,j}$과 반대방향의 부정적 누적 셋 $db^{j,k}$으로 분할한다. 이때 $db^{j,k}$는 DSI에서 주어진 시점 k까지 대하여 얻어진 누적 셋을 말한다. 그리고 Procedure III에서 1>k인 분할된 셋을 기반으로 정보 함수 값을 계산하여, 두 함수 값이 만나는 데이터 셋을 Procedure IV에서 찾는다. 그리고 Procedure V에서 찾아진 최적의 과거 셋에 대하여 분할하여, 표 1과 같이 지수가중 함수 모델 EM^{ES}_{past}을 적용하고, Procedure VI에서 소비자들의 현재 쇼핑 카트에 물건들의 조합에서 Up-to-Moment 선호를 계산함으로써 소비자들을 위해 온라인상에서 상품 주문을 추천 목록을 계산한다.

앞 절에서 정해진 최적의 과거 데이터 셋의 선정에 따라 지지도의 지수가중 함수 모델 EM^{ES}_{past}을 탐색하는 절차는 다음 표 2와 같다.

(표 2) 분할된 시간 연관규칙 탐색 단계

Given $D(All\ Datasets)$, $PL(Item\ List)$, $PR(Item\ Association\ Rules)$
Step0: Calculate the optimal size of past dataset (Procedure I~Procedure IV)
Step1: Partition D into $k-1$ subsets, $DS_1,\cdots,DS_{k-1}$
Step2: Calculate EM_{past} from each DS_i
Step3: Combine EM^{ES}_{past} with each weight function by *Definition* and Theorem (Procedure V).
1 *for* $i=1$ *to* $k-1$ *begin;* 2 $read\_in\_partition(DS_i)$ 3 $L^i = gen\_large\_itemset(DS_i)$ 4 $N_{\max_i} = max\_count(DS_i)$ 5 *for* $j=1$ *to* m *begin* 6 $C_j^W = \bigcup_{j}^{k-1} L_j^i$ 7 $EM_{past,} = \left\{ c \in C_j^W \mid \frac{count(C_j^W) + N_{\max i}}{N_i + N_{\max i}} \geq ms_i \right\}$

8 *end*
9 $R^{W}_{past} = \bigcup_{i=1}^{k-1} EM_{past_i}$
10 $PL_{utm} = \emptyset$
11 *While* $(GR_{past} \neq \emptyset)$ *do begin*
12 *for* $i = 1$ *to* $k-1$
13 *Remove from* EM_{past} *rule* $(p_1, \cdots, p_m) \rightarrow p(s_i\%, c_i\%)$
14 *If* R_{recent} *has a rule* $(p_1, \cdots, p_m) \rightarrow p(s_k\%, c_k\%)$ *then do*
15 $s^{ES} = \dfrac{\sum_{t=1}^{k} count(\{p_{t1}, \cdots, p_{tm}\}, DS_t) \cdot \alpha(1-\alpha)^{k-t+1}}{|D|} \geq ms\%$

$$c^{ES} = \frac{\sum_{t=1}^{k} [count(\{p_{t1}, \cdots, p_{tm}\}, DS_t) \times \alpha(1-\alpha)^{k-t+1}]}{\sum_{t=1}^{k} [count(\{p_{t1}, \cdots, p_{t(i-1)}, p_{t(i+1)}, \cdots, p_{tm}\}, DS_t)\ \alpha(1-\alpha)^{k-t+1}]} >= mc\%$$

16 $s_k\% = \dfrac{s^{ES}\% * |DS_{past}| + s_2\% * |DS_{recent}|}{|DS_{utm}|} \geq ms\%$

$$c_k\% = \frac{s^{ES}\% * |DS_{past}| + s_2\% * |DS_{recent}|}{\frac{s^{ES}\% * |DS_{past_i}|}{c_1} + count_{p_1 - p_m}} \geq mc\% \cdot$$

17 $PL_{utm} = PL_{utm} \bigcup \{p(s^{ES}, c^{ES})\}$ for EM^{ES}_{past}
18 *else goto* 3
19 *end*
20 *end*
21 *Return* PL_{utm} *which is sorted by confidence*

Step4 : Derive R_{past} and $count_{p1-pm}$ for DS_{recent} based on R_{past} Derice R_{recent} to obtain PL_{utm}(Procedure V).

4. 실험 및 평가

4.1 실험환경 및 실험데이터

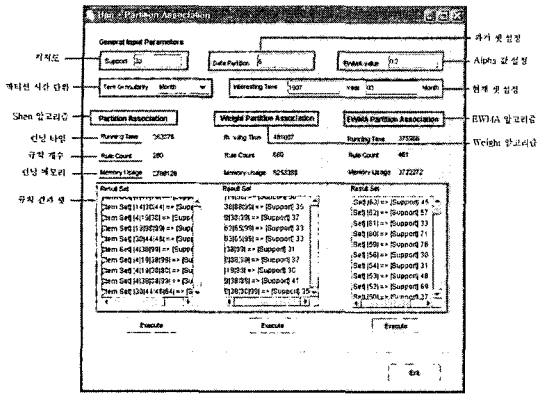
본 절에서는 Shen 모델과 기존의 연구에서 제안된 데이터 셋 크기별 지지도 가중 모델 EM^{w}_{past}, 그리고 본 논문에서 제안된 지수가중 모델 EM^{ES}_{past}에 대한 성능비교를 하고자 한다. 실험 평가를 위한 적용 도메인으로 전자상점 데이터 외에 연관규칙의 실시간 추천 개념이 요구되는 유비쿼터스 컴퓨팅 환경을 고려하여 보았다. 즉 웨이터의 PDA를 통하여 요리를 추천할 수 있는 레스토랑 어플리케이션 활용하는 상황을 선정하였다. 실험 데이터는 KDD CUP'99 Entree Chicago Recommendation Data(1996년 9월~1999년 4월까지 24,036개)를 사용하여 데이터의 공신력을 높였다. 다음 표 3은 레스토랑 원시 데이터를 나타낸다.

(표 3) 레스토랑 원시 데이터

TID	Date Time	Menu List
…	…	…
00006126	05/Apr/1997:11:54:58	07 30 35 20 13 25 94 07 80 09
00006127	05/Apr/1997:11:57:11	63 05 88 76 66 32 15 92 38 01 22
00006128	05/Apr/1997:12:01:40	19 07
00006129	05/Apr/1997:12:08:34	38 99 88 19 04 58 35 59 39
…	…	…

누적 데이터 셋을 Shen과 마찬가지로 DS1(1997년 2월~1997년 3월), DS2(1997년 1월~1997년 3월), DS3(1996년 12월~1997년 3월), DS4(1996년 11월~ 1997년 3월), DS5(1996년 10월~1997년 3월)로 나누어 실험 및 평가를 진행하였다. 또한 Shen에서는 단순히 룰의 수로만 알고리즘의 효과를 제시했으나, 이는 시간연관규칙의 적합성에 대한 결과라고 보기 어려우므로, 본 논문에서는 요리 주문 데이터의 과거 50%를 훈련용(training)으로 연관규칙을 계산한 후, 나머지 50%를 검정용(testing)으로 사용하여 MSE를 계산하여 정확도를 비교하였다.

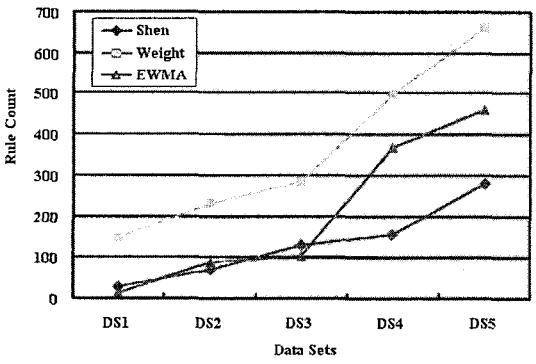
그림 5와 같이 WindowsXP 환경에서 Java1.4.2을 이용하여 구현하였으며, 데이터베이스는 Postgresql 7.4.5을 사용하였다. 실험에 사용된 시스템은 Pentinum PC 2.8GHz 512Mbyte이며, Postgresql 7.4.5과의 연동을 위해 JDBC3.0 드라이버를 사용하였다. 구현된 프로그램 UI에는 Shen 모델, 데이터 셋 크기별 가중 모델 EM^{w}_{past}, 지수가중 모델 EM^{ES}_{past}을 환경 설정 변수 값과 결과 제시를 포함하였다.



(그림 5) 구현된 시뮬레이션 프로그램

4.2 성능 분석

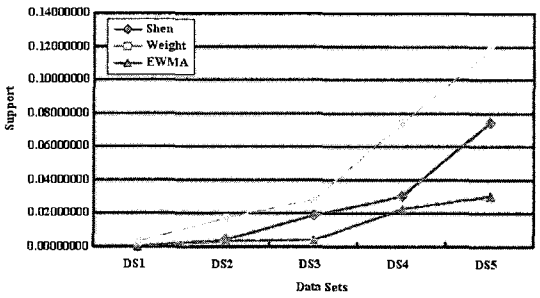
먼저 지지도가 30일 때 각 과거 서브 셋에 대해서 3가지 알고리즘의 누적 분할 규칙 탐사 개수를 측정하였다.



(그림 6) 누적 분할 별 규칙 탐사 개수

그림 6을 보면 직관적으로 누적 데이터 셋이 커질수록 탐사된 규칙의 개수가 세가지 모두 증가하는 것으로 나타났으며, EM^{ES}_{past}(EWMA)알고리즘이 EM_{past}(Shen)알고리즘에 비해 다소 많은 규칙을 탐사한 것으로 나타났고 EM^{w}_{past}(Weight) 알고리즘이 가장 많은 규칙을 탐사한 것으로 나타났다. 이는 Weight 알고리즘은 각 분할 셋의 크기에 따른 가중치 값을 주기 때문에 평균 지지도 값을 갖는 Shen의 알고리즘에 비해 탐사된 규칙의 수가 많게 되기 때문이다. 다른 지지도에 대해서도 유사한 경향을 보이므로 더 제시하지 않았다.

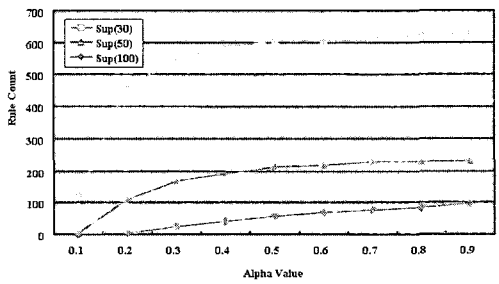
탐사된 연관규칙의 수가 연관규칙의 질적 적합을 의미하지는 않는다. 따라서 본 연구에서는 탐사된 연관규칙의 질적 평가를 위하여, 훈련 데이터 셋에서 추정한 지지도의 값이 검정 데이터 셋의 지지도 값과 얼마나 일치하는지를 보는 것으로 연관규칙의 질적 적합을 제시하고자 한다. 왜냐하면 지지도 값은 탐사되는 연관규칙의 개수를 변화시킬 수 있는 가장 중요한 모수로서 이 값이 달라지면 해당 지지도에 따른 연관규칙 탐사의 의미가 없어지기 때문이다. 세 알고리즘에 대하여 각 누적 셋별로 평균오차제곱(Mean Squared Error)를 계산하고 변화를 그림 7과 같이 보였다. 즉 각 누적 과거 셋의 단계마다 최소 지지도를 만족하는 탐사된 규칙들의 지지도 차의 제곱의 합에 대해 실험한 결과, EM^{ES}_{past}알고리즘이 Shen 알고리즘보다도 연관규칙의 추천 정확도가 높음을 보여주고 있다.



(그림 7) 탐사된 규칙의 지지도 평균오차제곱(MSE)값

또한 오래된 데이터 셋을 포함할수록 MSE는 커지므로, 적정한 누적 데이터 셋을 활용해야 함을 시사하고 있다. 계산해야 하는 과거 데이터 셋의 크기가 클수록 EM^{ES}_{past} 알고리즘의 성능이 더욱 높아짐을 알 수 있었고, 계산에 사용된 CPU 시간이나 메모리는 거의 차이가 없었다. 따라서 실험을 통하여 본 연구에서 정의한 EM^{ES}_{past} 알고

리즘이 기존의 알고리즘들에 비해 지수 평활법 상수 α값에 의해서 최신자료에 대한 정보를 더 잘 반영함으로 더 유용하고 정확한 정보를 실시간으로 탐사할 수 있다는 것을 알 수 있다.
그러나 제안된 이 알고리즘이 최근에 입력된 데이터를 실시간으로 바로 반영하여 점진적인 기법으로 탐사된 규칙들을 업데이트하여 갱신함으로써 사용자들에게 보다 유용하고 정확한 정보들을 신속하게 제공할 수 있는 장점을 갖지만, 구체적으로 지수 평활법 상수 α 값을 어떤 값을 주었을 때 가장 적당한지는 제시하지 못하고 있다. 일반적으로는 α =0.2를 선택하는 것이 보통이나 아직 전자상거래에서는 적용된 사례나 연구가 없다. 따라서 그림 8과 같이 EM^{ES}_{past} 알고리즘의 3가지 최소 지지도 30, 50, 100에 대해서 α 값을 증가함에 따라서 탐사된 규칙의 개수를 평가하였다. 그림 8에서 보여주듯이 전자상거래에서는 α 값을 0.2 내외로 적용하는 것이 가장 적당하다.



(그림 8) α값에 따른 탐사된 규칙 개수

즉, α값이 증가할수록 급격한 지수적 가중치 값이 부여됨으로 탐사된 규칙의 수도 급격하게 증가하게 된다. 따라서 0.2를 중심으로 엄격하고자 하는 경우는 좀 작은 값을 주는 등 데이터의 성격을 보며 세팅할 수 있을 것이다.

5. 결 론

시간에 민감한 상품을 취급하는 전자상점이나 실시간 추천이 요구되는 레스토랑의 요리추천 등은 최근 발전하고 있는 유비퀴터스 컴퓨팅 응용산업과 향후 밀접한 관계를 가질 것이다. 즉, 과거 데이터를 토대로 정확한 실시간 추천은 고객만족의 극대화할 뿐만 아니라 매출증대를 높일 수 있기 때문이다.

본 연구에서는 과거자료와 최근자료를 동시에 조합한 Up-To-Moment 데이터 셋을 이용하여, 추천 성공률을 높이면서도 실시간 계산이 가능한 Shen의 연관규칙 탐색 알고리즘을 보다 더 정교하게 과거의 거래 세부정보까지 고려할 수 있도록 Up-To-Moment 데이터 셋에 대하여 과거 데이터 셋 부분을 (k-1)개로 분할하여 지수평활법을 적용한 조합규칙(Partitioned Combination Law)을 적용하는 연관규칙 선호모델을 제안하였다. 즉, 기존의 과거 데이터 셋을 기반으로 하는 전형적인 과거모델 SM_{past}과 EM_{past}을 개선하기 위한 과거 데이터 셋을 $(k-1)$개 분할 셋의 결정치 가중값을 준 일반화 모델 EM^{C}_{past}(분할 셋 지지도 가중값을 준 EM^{W}_{past})를 개선한 지수 평활법을 적용 모델 EM^{ES}_{past}을 제안하였다.

제안한 지수 평활법을 적용한 시간연관 규칙 EM^{ES}_{past}(EWMA 알고리즘)을 기존의 분할 셋 연관규칙 EM_{past}(Shen 알고리즘) 및 분할 셋에 대한 지지도 가중 연관규칙 EM^{W}_{past}(Weight 알고리즘)과 비교하기 위한 실험에서는 적용 도메인으로 연관규칙의 개념이 가장 직관적인 웨이터의 PDA를 통하여 전자적으로 요리주문 프로그램으로 전송되어 처리될 수 있는 레스토랑 음식 추천 애플리케이션을 선정하였다. 이를 위해 KDD CUP'99 레스토랑 데이터를 활용하여 연관규칙을 계산하였다.

기존의 EM_{past}, EM^{W}_{past}와 제안된 EM^{ES}_{past} 세 모델에 대한 시뮬레이션 결과, 각 누적 과거 셋의 단계마다 최소 지지도를 만족하는 탐사된 규칙들의 지지도 오차의 제곱의 합에 대해 실 EM^{ES}_{past}의

경우가 가장 적은 것으로 나타났다. 즉, 최근 자료 셋의 지지도에 가장 가까우므로, 실제 ms가 주워지고 시간 연관 규칙을 탐색했을 때 가장 최신 자료를 잘 반영한다고 할 수 있다. 또한 세 모델 중 EM_{past}^{ES}이 EM_{past}에 비해 다소 많은 규칙을 탐사한 것으로 나타났고, EM_{past}^{W}이 가장 많은 규칙을 탐사한 것으로 나타났다. 따라서 지지도 오차의 제곱합의 결과와 함께 고려할 때 EM_{past}^{W}는 불필요한 과거의 시간 연관 규칙이 많이 생성되어 최신자료 반영이 잘 되지 않는데 비해, EM_{past}^{ES}는 EM_{past}과 시간 연관규칙의 수가 유사했으나 지지도 오차 제곱 합은 오히려 적어 보다 민감하게 반응되는 효과적인 모델임을 실증하였다. 그리고 실제 전자상점 데이터에 적합한 지수평활 지수 값($0 \leq \alpha \leq 1$)을 탐색하여 보았으며, 메모리와 실행시간에 대하여 비교평가 하였는바 기존의 모델과 거의 유사했다.

이와 같이 본 연구에서 정의한 EM_{past}^{ES}이 기존의 실시간 추천 모델에 비해 최신자료에 대한 정보를 더 잘 반영함으로 더 유용하고 정확한 정보들을 탐사할 수 있다는 것을 알 수 있었으며, 최적 데이터 셋의 결정 알고리즘과 경험적 지수평활 지수 값 $\alpha = 0.2$을 활용하면 매우 안정되고 빠른 계산이 가능함을 보였다. 제안된 이 알고리즘은 전자상거래의 추천시스템과 같은 분야에서 최근에 입력된 데이터를 실시간으로 바로 반영하여 점진적인 기법으로 탐사된 규칙들을 업데이트 하여 갱신함으로써 사용자들에게 보다 유용하고 정확한 정보들을 신속하게 제공할 수 있는 장점을 갖는다.

참 고 문 헌

[1] 변루나, 박병선, 한정혜, 정한일, 임춘성, "지수평활법을 적용한 시간연관규칙", 정보처리학회논문지D, 제11-D권 제3호, pp.741~746, 2004.

[2] Agrawal R., Imielinski T. and Swami A., "Mining Association Rules between Sets of Items in Large Database," Proceedings of ACM SIGMOD Conference on Management of Data, pp.207~216, 1993.

[3] Banu Ozden, Sridhar Ramaswamy, Abraham Silberschatz, "Cyclic Association rules," Proceeding of International Conference on Data Engineering, pp.412~421, 1998.

[4] B. Ozden, S. Ramaswamy and A. Silberschatz, "Cyclic association rules," 11th International Conference on Data Engineering Orlando, May. 1998.

[5] Chang-Hung Lee., "Mining Association Relationship in a Temporal 'Database," Ph. D. Thesis, Department of Electrical Engineering Nat'l Taiwan Univ. Taipei, 2003.

[6] J. F. Roddick and M.Spiliopoulou, "Temporal data mining: survey and issues," Research Report ACRC-99-007, Univ. of South Australia, 1999.

[7] Jeremy. M. and Jun W.L., "Mining Association Rules in Spatio-Temporal Data: An Analysis of Urban Socioeconomic and Land Cover Change", Transaction in GIS, vol.9, Blackwell Publications Inc., pp.5~18, 2005.

[8] Luna, Byon, "The Modified EM_{past} Model for Mining Association Rules and its Statistical Application", Korea Statistical Studies, 9(1), pp.103-122, 2004

[9] Luna B. and J.H. Han, "Fast Algorithms for Temporal Association Rules in Large Database", Key Engineering Materials, vol.277, Trans Tech Publications Inc., pp.287~292, 2005.

[10] Shen, Y., Yahg, Q., Zhang, Z., and Lu, H., "Mining the Customer's Up-To-Moment

Preferences for E-Commerce Recommendation," Lecture Notes in Computer Science, Springer-Verlag, pp.166~177, 2004.

[11] S. Ramaswsmy, S Mahajan and A. Silberschatz, "On the discovery of interesting patterns in association rules," the VLDB Conference, New York City, Sep. 1998.

[12] X. Chen and I. Petrounias, "A Framework for temporal data mining," 9th International Conference on Database and Expert System Applications, 1998.

[13] X. Chen, I Petrounias and H. Heathfield, "Discovering temporal association rules in temporal database," International Workshop on Issues and Applications of Database Technology, 1998.

[14] Y. Li, P. Ning, X. S. Wang, and S. Jajodia, "Discovering Calendar-based Temporal Association Rules," Proceedings of the 8th Internat'l Symposium on Temporal and Reasoning, 2001.

[15] http://www.cs.swarthmore.edu/~meeden/

[16] http://www.newscientist.com/article.ns?id=dn8642

[17] http://www.spacedaily.com/news/robot-04zza.html

◐ 저 자 소 개 ◑

한 정 혜(Jeong-Hye Han)

1992년 충북대학교 통계학과 졸업(학사)
1994년 충북대학교 통계학과 졸업(석사)
1998년 충북대학교 전자계산학과 졸업(박사)
1998년~1999년 연세대학교 산업시스템 공학과 포닥연구원
1999년~2001년 행정자치부 국가전문행정연수원 통계연수부 전산교육 전임교수
2001년~현재 청주교육대학교 컴퓨터교육과 조교수
관심분야 : 데이터마이닝, 전자상거래, 인간과 로봇 상호작용
E-mail: hanjh@cje.ac.kr

변 루 나(Lu-Na Byon)

1992년 충북대학교 통계학과 졸업(학사)
1995년 충북대학교 컴퓨터공학과 졸업(석사)
2004년 충북대학교 통계학과 졸업(박사)
2006년~현재 보건복지부 사무관
관심분야 : 전산통계, 데이터마이닝, 보건정책
E-mail: lnbyon@korea.kr