韓國 컴퓨터情報學會 論文誌
第10卷 第1號, 2005. 3.

2005-3-1-1-6

시간 의존적인 상품 추천을 위한 지수 평활 시간 연관 규칙

정 경 자*

Exponential Smoothing Temporal Association Rules for Recommendation of Temporal Products

Kyeong-Ja Jeong*

요 약

본 연구에서는 시간 연관 규칙에 지수 평활법을 적용한 상품 추천 알고리즘을 제안한다. 시간 연관 규칙은 기존의 연관 규칙에 시간 개념을 적용한 연관 규칙이다. 본 연구에서는 과거 데이터 보다 최신의 데이터에 가중치를 더 부여한 지수 평활 시간 연관 규칙을 제안한다. 제안한 알고리즘은 시간 의존적인 데이터에 적용하여 시뮬레이션을 한 결과 지수 평활법을 적용한 시간 연관 규칙이 기존의 시간 연관 규칙보다 실행시간 면에서 다소 오래 걸리지만 상품 추천 측면에서 더 효과적이다.

Abstract

We proposed the product recommendation algorithm mixed the temporal association rule and the exponential smoothing method. The temporal association rule added a temporal concept in a commercial association rule. In this paper, we proposed a exponential smoothing temporal association rule that is giving higher weights to recent data than past data. Through simulation and case study in temporal data sets, we confirmed that it is more precise than existing temporal association rules but consumes running time.

▸ Keyword :Exponential Smoothing Temporal Association Rules, Association Rules, Exponential Smoothing Method

• 제1저자 : 정경자
• 접수일 : 2005.01.12, 심사완료일 : 2005.02.28
* 충정대학교 컴퓨터학부 교수
본 연구는 충정대학 교내연구비 지원에 의해 수행되었음.

Ⅰ. 서 론

인터넷의 활성화는 많은 산업 분야를 인터넷으로 이끌어내는 역할을 하고 있다. 그 중에서 대표적인 것으로 기존의 쇼핑몰을 인터넷으로 이끌어내는 전자상거래 시스템의 등장은 세계를 단일 시장권으로 묶고 있다. 초기 전자상거래 시스템에서는 소비자의 전자상점에 대한 신뢰도가 매우 낮고 시스템 측면에서도 취약한 부분이 많아 인터넷을 통한 실제 거래가 매우 적었다.

최근에는 인터넷 기반 시설의 안정화와 인터넷의 저변 확대 등으로 인터넷을 통한 상거래가 나날이 늘어가는 추세이다. 전자 상점들의 수가 나날이 증가하는 현 상황에서 다른 전자상점과의 차별화를 위한 많은 연구가 이루어지고 있다. 실제로 국내외의 많은 전자 상점들에서 원투원 마케팅 개념을 적용한 CRM(Customer Relationship Management)을 도입하여 서비스를 제공하고 있다[1]. 일반적으로 대부분의 전자상점들은 CRM을 위해 고정된 시점의 다수의 고객에 거래 데이터에 연관 규칙(Association Rule), 클러스터링(Clustering), 신경망(Neural Networks)등을 적용한다[1]. 연관 규칙은 장바구니 분석을 통해서 얻어지는 것으로, 현재까지 가장 활발하게 연구 및 활용되고 있는 데이터 마이닝 중의 하나이다[2].

효과적인 CRM을 위하여 많은 전자상점에서는 고객군을 분류하여 관리하는 고객기반 모델과 상품에 대한 아이템 기반 모델을 적용한다. 아이템 기반 모델은 아이템이 유행과 시간에 따라 매우 민감하게 반응하므로, 이를 고려하면 효과적일 것이다. 연관 규칙은 계산과 이해가 편하기 때문에 전자상거래와 같이 많은 거래 데이터를 분석하여 크로스 셀링(Cross Selling)을 하는 경우에 매우 활발히 이용되고 있지만, 고정된 시간에 대한 분석이므로 적시성(Just-in-Time)이 떨어지는 단점이 있다. 즉, 전자상점의 매출에 영향을 크게 미칠 수 있는 계절상품 또는 특별상품의 거래 데이터베이스에서 중요한 속성인 시간을 고려하지 못하기 때문에 좋은 추천 성공률을 높게 가질 수 없다. 예를 들어, {발렌타인 데이, 초콜릿}, {화이트 데이, 사탕}, {크리스마스 이브, 케익}, {추석, 송편}, {추수감사절, 칠면조} 등과 같은 것은 시간과 밀접하게 관련되어 있는 상품군이다.

그러므로 많은 연구에서 시간을 고려한 다양한 연관 규칙들이 연구되었다. 타임스탬프(time-stamp)된 거래에서 특정 시간 간격 동안의 주기적 연관규칙(Cyclic Association Rules), 달력에 기반한 연관 규칙(Calendar-based Association Rules), 연관 규칙이 적용되는 시간 구간을 정한 후, 구간에서 연관 규칙의 주기적 패턴을 발견하여 시간적 특성을 갖는 연관 규칙을 찾아내는 방법 등이 제안되었다[5,8,9].

본 연구에서는 데이터웨어하우스에 장기간에 걸쳐 모아진 데이터를 최근 데이터일수록 높은 가중치를 주고 오래된 데이터일수록 낮은 가중치를 주는 지수 평활법을 적용한 변형된 시간 연관 규칙을 정의하고, 이를 탐색하는 알고리즘을 제안한다. 또한, 지수 평활법을 적용한 시간 연관 규칙이 분할된 데이터 셋에 대해 증감의 추세를 갖는 데이터 셋, 추세가 없는 데이터 셋에 대하여 어떠한 결과를 보이는지 실험을 통하여 보이고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구와 관련된 관련 연구들을 살펴보고, 3장에서는 지수 평활법을 적용한 시간 연관 규칙을 정의하고 탐사 알고리즘을 제안한다. 그리고 4장에서는 실험 및 평가를 통하여 지수 평활법을 적용한 시간연관 규칙과 탐사 알고리즘의 효율성을 보인다. 마지막으로 5장에서는 결론 및 향후 연구에 대하여 논의한다.

Ⅱ. 관련 연구

2.1 연관 규칙

연관 규칙은 거래 데이터베이스에서 아이템간에 발생하는 연관성을 표현하는 것으로[2,10,11], 어떤 사건이 발생할 때 그 다음 사건의 관련성을 의미한다. 연관 규칙 $X \rightarrow Y$은 데이터베이스의 거래 중 X라는 항목 집합을 포함하는 거래는 항목 집합 Y도 함께 포함하는 경향이 있음을 의미한다. 연관 규칙의 척도는 지지도(*support*)와 신뢰도(*confidence*)를 이용하여 타당성을 판단한다[2,4].

2.2 시간 연관 규칙

타임 스탬프된 거래 데이터로부터 의미 있는 지식을 탐사하기 위하여 기존의 데이터 마이닝 모델에 대한 여러 연구가 진행되고 있다[8,9].

여기에 속한 기법으로는 달력으로 표현된 시간 패턴을 가지는 달력 연관 규칙(Calendric Association Rules)[9], 주기적으로 반복되는 주기적 연관 규칙(Cyclic Association Rules)[4], 분할 셋을 기반 연관 규칙(Partitioned Association Rules)[8], 누진적 가중 연관 규칙(Progressive weighted Association Rules)[5] 등이 있으며 이를 요약하면 다음과 같다.

2.1.1 달력 연관 규칙

달력으로 표현 가능한 특정 일에 발생하는 아이템간의 관계성을 표현한 것으로 달력 스키마 식(1)에 대하여 연관 규칙 식(2) 가 있다. 예를 들어 월드컵 기간의 "Be the Red!" 셔츠 등이 될 수 있다.

$$R=(G_n:D_n, G_{n-1}:D_{n-1}, \cdots, G_1:D_1) \quad \cdots\cdots (1)$$

$$(A \rightarrow B) \langle d_n, d_{n-1}, \cdots, d_1 \rangle \quad \cdots\cdots (2)$$

2.1.2 주기적 연관 규칙

주기적 특성을 갖고 발생하는 아이템간의 관계성을 규칙으로 표현한 것으로, 주기 $c=(l, o)$, $0 \le o \le l$ 을 갖는 주기적 연관규 칙 $(A \rightarrow B)$이 있다. 예를 들어 추수감사절의 칠면조, 발렌타인데이의 초콜릿 등을 들 수 있다.

2.1.3 분할 셋 기반 연관 규칙

동일기간으로 분할하여 연관 규칙을 탐색하는 방법으로 분할 셋 i에서 k-아이템 셋의 지역 빈발 집합 L_k^i, 지역 후보 집합 C_k^i, 전역 빈발 집합 L_k^G, 전역 후보 집합 C_k^G에 대해 연관 규칙을 생성하는 것으로 시간분할 효과 추가가 있으나, 가중치 개념은 없다. 메인 메모리가 크게 요구되며, 분할 셋의 크기를 메인 메모리 크기로 제한한다.

2.1.4 누진 가중 연관 규칙

동일기간으로 분할하여 가중함수 W(·)로 최근에 더 큰 가중치를 부여하며, 계산량 감소 식(3)인 누진적 가중 연관 규칙 $(A \bigcup B)^{PW}$이 존재한다.

$$s^{PW}(A \bigcup B)=\frac{S^S(A \bigcup B)}{\sum |P_i| \times W(P_i)} \quad \cdots\cdots (3)$$

가중함수 W(·)의 선택이 임의롭고, 식(4) 의 경우 거래수가 클 경우 적당치 않다. 또한 임의 가중 함수에 의해 결과해석에 대해서도 해석이 모호하다.

$$W(t)=1+\frac{t}{n} \quad \cdots\cdots (4)$$

2.3 지수 평활법

일반적인 지수평활 모형은 다음 식(5)와 같이 정의된다[7].

$$S_t=\alpha y_t+(1-\alpha)S_{t-1}, \quad 0<\alpha \le 1 \quad \cdots\cdots (5)$$

(단, y_t는 시점 t 에서의 총 관측값, 초기값 S_0는 평균값, α는 데이터의 영향 정도를 결정하는 지수평활 상수)

이 방법은 α값이 클수록 최근 데이터에 상대적으로 큰 가중치를 준다. 지수평활 상수 α 값의 결정은 0~1 사이에서 경험적으로 가장 효과적인 값을 이용한다. 이러한 지수평활법은 계산이 간단하고 최신자료를 많이 반영하는 적시성 때문에 많은 공학 분야에 활발히 활용되고 있으며, 최근에 [7]에서는 협력적 여과 추천 알고리즘에 이를 적용한 이론을 제시하였으나, 전자상거래 데이터에는 아직 적용사례가 없다. 따라서 본 연구에서는 최신 정보에 더 많은 가중치를 부여하는 지수 가중치 시간 연관 규칙을 제안하고 그에 대한 알고리즘을 제시한다.

III. 지수 평활 시간 연관 규칙

3.1 지수 평활법을 적용한 시간 연관 규칙

본 논문에서 제안하는 지수평활 시간 연관 규칙은 분할 셋을 기반한 연관 규칙과 누진적 가중 연관 규칙과 같이 데이터 셋을 분할함으로써 정의된다. 타임 스탬프된 거래 데이터 셋 D를 k개의 서브셋으로 분할한다. 이 때 식(6)을 만족한다.

$$D=\bigcup_{t=1}^{k} DS_t, \quad \emptyset=\bigcap_{t=1}^{k} DS_t \quad \cdots\cdots (6)$$

(단, DS_t, $t=1, 2, \cdots, k$, 상품집합 $\{p_1, \cdots, p_m\}$)

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.

본 논문에서 제안하는 지수 평활 시간 연관 규칙은 지수평활 전역 지지도(*global support*) s^{ES}, 지수평활 신뢰도(*global confidence*) c^{ES}이 각각 사용자가 미리 정한 최소 지지도 ms, mc보다 큰 모든 연관 규칙으로 다음과 같이 정의와 한다.

정의 어느 시점 t 에 대하여, 상품 크기가 $j \le m$ 일 때, 지지도와 신뢰도 s^{ES}, c^{ES}를 갖는 식(7)은 지수평활 연관 규칙이다

$$(p_1, \cdots, p_{j-1}, p_{j+1}, \cdots, p_m) \rightarrow p_j (s^{ES}\%,\ c^{ES}\%) \quad \cdots\cdots (7)$$

(단, $0 < \alpha \le 1$ 에 대하여 s^{ES}, c^{ES}는 각각 식(8)을 만족한다.)

$$\frac{\sum_{i=1}^{k} count(\{p_{n}, \cdots, p_{tm}\}, DS_i) \cdot \alpha(1-\alpha)^{k-t+1}}{|D|} \ge ms\%,$$

$$\frac{\sum_{i=1}^{k} [count(\{p_{n}, \cdots, p_{tm}\}, DS_i)}{\sum_{i=1}^{k} [count(\{p_{n}, \cdots, p_{t(i-1)}, p_{t(i+1)}, \cdots, p_{tm}\}, DS_i)} \times \frac{\alpha(1-\alpha)^{k-t+1}]}{\alpha(1-\alpha)^{k-t+1}] \ge mc\%} \quad \cdots\cdots (8)$$

3.2 지수 평활 시간 연관 규칙 탐색 알고리즘

다음 (그림 1)은 3.1절 정의에 의한 지수평활 시간 연관 규칙의 탐색 알고리즘이다.

```
procedure ES_Apriori(D,N,k,α,ms)
 if D≠∅ and N> k
    sort(D) by transaction_time
   for t = 1 to k
       read_in_partition(DSt∈D)
      ES_count(t) = average(DSt)
      Lt=gen_large_itemsets(DSt)
    for(i=2; Lit≠∅; i++)
      C_i^G = U_{i=1,2,···,k} L_i^j
    for all candidates c∈CG
      ES_count(t+1)
←α*gen_count(c,DSt+1)+(1-α)*ES_count(0,t)
    end
   end
  end
    s^EW = c.count/N
    ES_Apriori = {c∈CG| s^EW≥ms}
end if
return ES_Apriori
```

그림 1. 지수 평활 법을 적용한 연관 규칙 알고리즘
Fig 1. Association Rule Algorithm for Exponential Smoothing

이 때, N은 D의 거래 총 수, k는 분할 셋 수, DSt는 t 시점에서의 분할 데이터 셋, α는 지수평활 상수, ES_Apriori은 지수평활 시간 연관 규칙을 나타낸다.

Ⅳ. 실험 및 평가

4.1 실험

본 절에서는 제안한 지수 평활법을 적용한 시간 연관 규칙과 기존의 시간 연관 규칙, 분할 셋에 대한 누진가중치 기법과 비교하였다.

인터넷 쇼핑몰 A쇼핑몰에서 취급하는 32개 항목에 대한 2000년 1월~2003년 3월까지 296,784개의 거래 데이터를 분기별 시점(k=12)으로 구분한 후, 연관 규칙을 사전 계산한다. 구해진 연관 규칙을 대상으로 회귀분석(Regression Analysis)에 의해 시간에 따른 추세 변화를 갖는 구매 상품과 추세가 없는 상품 집합으로 구분하여 지지도에 대한 평균오차제곱합(MSE:Mean Squared Error)과 계산시간을 비교 실험을 하였다. 추세 유무별 상품 집합을 분류한 후, 각 집합에 대해 75%의 레코드를 훈련용(training)으

로, 나머지 25%는 검정용(testing)으로 최소 지지도 값 0.1~0.9에 대해 누진적 가중 시간 연관 규칙과 지수평활 시간 연관 규칙을 JAVA로 구현하여 실험하였다.

다음은 추세 유무에 따른 MSE, 데이터 셋에 따른 MSE을 연관규칙(AR), 누적가중법(PWAR)과 지수 평활 시간 연관 규칙(ES)로 최소 지지도를 0.2, 0.5, 0.75에 대해 시뮬레이션한 결과로 모두 유사한 결과를 보였다. 그중 추세 유무 상품군에 따른 지수 평활법 시간 연관 규칙과 누진 가중 시간 연관 규칙에 대하여 검정용 셋의 최소 지지도 0.2에 대한 MSE가 (그림 2)와 같다.

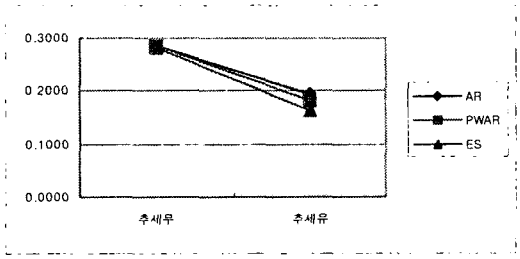

그림 2. 시간 추세 유무군에 따른 MSE(ms=0.2) 비교
Fig 2. MSE(ms=0.2) Comparison for Temporal trend

추세가 없는 상품군에 대해서는 Apriori법(AR), 누진 가중법(PWAR:Progressive Weighted Association Rules), 지수평활 시간 연관법(ES:Exponential Smoothig)들 간에 차가 매우 적으나, 추세가 있는 상품군에 대해서는 지수평활법이 가장 작은 MSE를 갖음으로써 가장 우수한 정확도를 보였다.

DS1(2002년 7월~2002년 12월), DS2(2002년 1월~2002년12월), DS3(2001년 7월~2002년 12월), DS4(2001년 1월~2002년 12월)로 누적 서브셋으로 나누어 누적 데이터 서브셋에 따른 시간 연관 규칙의 매칭 성공에 대한 MSE 비교결과는 (그림 3)과 같다.

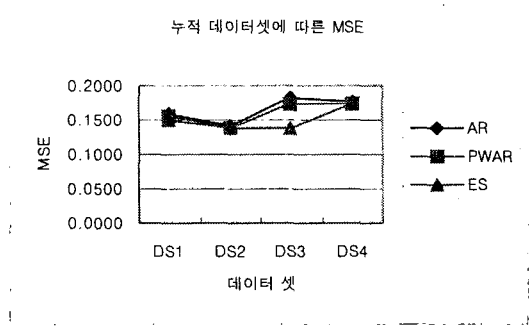

그림 3. 누적 데이터 셋에 따른 MSE
Fig 3. MSE for Accumulated Data Set

(그림 3)은 (그림 2)의 결과와 유사하게 시간 연관 규칙이 전체 데이터 셋에 대해서도 여전히 다른 연관 규칙보다 적은 MSE를 가지므로써 가장 효과적임을 보였다.

데이터 셋의 크기에 따른 지수평활 시간 연관 규칙과 기존 연관 규칙의 실행시간은 (그림 4)와 같이 나타났다.

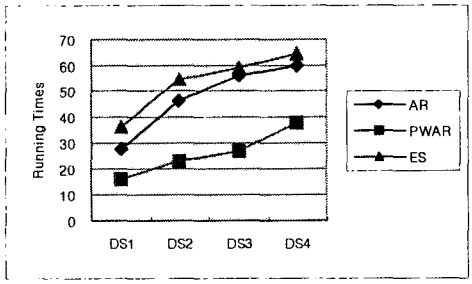

그림 4. 데이터 셋에 따른 실행시간 비교
Fig 4. Runtime Comparison for Data Set

따라서, 본 연구에서 정의한 지수평활 시간 연관 규칙이 기존의 시간 연관 규칙보다 효과적이나, 계산 비용은 다소 높게 나타났다. 이는 분할 셋에 대한 지수평활 필터의 추가 계산으로 인한 결과이다. 최근 컴퓨터들의 발달로 인해 계산 비용의 중요성을 고려하지 않는다면, 지수 평활 시간 연관 규칙은 시간을 고려하지 않는 경우 탐색되지 않을 수도 있는 연관 규칙을 정확하게 추가 탐색할 수 있다. 또한 이렇게 얻어진 연관규칙을 일반 연관규칙 탐색결과와 비교해 봄으로써, 시간을 고려한 직관적 해석도 용이하며 효과적이라 할 수 있다.

4.2 적용 예

본 연구에서 제안하는 지수평활 연관 규칙과 연관규칙을 2002년 1월~2002년 3월(12주)까지 인터넷 B쇼핑몰에서 취급하는 식품 및 잡화, 의류품목 30개 품목에 대한 거래 데이터를(k=12)로 분할하여 지수 평활법 시간 연관 규칙을 계산하였다. 총167,234개 레코드를 125,425개의 훈련용과 41,809개의 검정용으로 분할하여 기존의 연관 규칙과 지수평활 시간 연관 규칙에 의해 각각 상품을 추천한 후, MSE를 〈표 1〉과 같이 나타났다.

총 13개의 연관 규칙이 구매 상품에 대해 회귀분석을 실시하여 추세를 갖는 경우 8개와 추세가 없는 5개의 연관 규칙에 대하여 MSE을 제시하였다. 〈표 1〉과 같이, 추세가 있는 구매상품의 경우에 지수평활 시간 연관 규칙이 기존 연관 규칙보다 추천 성공률을 보다 높일 수 있음을 보여준다.

표 1. 시간 추세에 대한 연관 규칙의 MSE(단위%)
Table 1. MSE of Association Rule for Temporal Trend

추세	연관 규칙	연관 규칙		지수평활 시간 연관 규칙	
		지지도	신뢰도	지지도	신뢰도
무	p50→p11	1.29	3.01	1.39	3.25
	p47→p11	0.74	1.23	0.83	1.44
	p12,p1→p11	7.60	2.19	7.63	2.27
	p5,p1→p11	7.69	1.60	7.86	2.03
	p7,p5,p1→p12	6.63	2.06	6.49	1.63
유	p12→p11	3.65	1.95	3.19	1.32
	p5→p11	3.97	1.06	3.79	0.79
	p7,p12→p11	4.27	1.98	3.83	1.18
	p5,p12→p11	4.30	0.89	3.91	0.13
	p5,p11→p12	4.30	2.24	3.91	1.50
	p7,p5→p11	4.65	1.34	4.42	0.87
	p7,p5→p12	5.25	2.89	4.72	1.80
	p7,p5,p11→p12	4.73	2.52	4.37	1.63

V. 결 론

본 연구에서는 전자상점의 매출에 영향이 크게 미칠 수 있는 계절상품이나 기획 상품과 같이 시간의 변화에 민감한 추세를 갖는 상품들에 대하여 지수평활 시간 연관 규칙을 정의하고, 이를 탐색하는 알고리즘을 제안하고 시뮬레이션을 통하여 제안된 방법이 기존의 시간 연관 규칙에 비해 더 효율적임을 알 수 있었다. 그러나 제안한 방법은 실행시간 측면에서 다소 다른 기법보다 더 많은 실행 시간이 요구되는 점은 있지만 그 차이가 아주 작다.

본 연구에서 제안한 지수평활 시간 연관 규칙은 추세가 없는 상품의 경우보다 추세가 있는 항목의 경우에 더 작은 오류를 갖는 것으로 나타났으며, 이는 적용 사례를 통해서도 같은 결과를 얻었다.

그러므로 시간의존적인 상품을 취급하는 전자상점의 추천 시스템 구축 시 본 연구에서 제안하는 지수평활 시간 연관 규칙을 적용함으로써 소비자에게 적중률이 높은 상품 추천이 가능하므로 쇼핑몰의 매출증대를 높일 수 있을 것이다. 향후 연구로는 제안한 알고리즘의 실행 시간을 감소시키는 것이다.

참고문헌

[1] 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, Washington D.C., pp207-216, 1993.

[2] Agrawal R. and Strikant R., "Fast Algorithms for Mining Association Rules", Proceedings of the 20th International Conference on Very Large Databases, Satiago, Chile, 1994.

[3] Alex Berson, Stephen Smith, Kurt Thearling, Building Data Mining Application for CRM, McGraw-Hill.

[4] Banu Ozden, Sridhar Ramaswamy, Abraham Silberschatz, "Cyclic Association rules", Proceeding of International Conference on Data Engineering, pp412-421, 1998.

[5] Chang-Hung Lee, Jian Chih Ou, Ming-Syan Chen, "Progressive Weighted Miner:An Efficient Method for Time-Constraint Mining," 7th Pacific-Asia Conference, pp449-460, PAKDD 2003.

[6] Hunter,J.S., "The Exponentially Weighted Moving Average", Journal of Quality Technology, Vol.18, pp.203-210, 1986.
[7] Margaret H. Dunham, Data Mining Intoductory and Advanced Topics, Prentice Hall.
[8] X.Chen and I.Petrounias, "A framework for temporal data mining", In Proc. Ninth International Conference on Database and Expert Systems Applications,DEXA, 1998.
[9] Y.Li, P. Ning, X.S. Wang, and S. Jajodia " Discovering Calendar-based Temporal Association Rules", Proceedings of the 8th International Symposium on Temporal and Reasoning, 2001.
[10] 하창승, 윤병수, 류길수,"연관규칙 탐사기법을 이용한 해양전문 검색 엔진에서의 질의어 처리에 관한 연구", 한국컴퓨터정보학회 논문지, 제8권, 제2호, 2003.
[11] 안성욱, 오기욱,"모바일 에이전트를 이용한 상품거래 서비스에 관한 연구",한국컴퓨터정보학회 논문지, 제6권, 제3호, 2001.

저 자 소 개

정경자
1991년 충북대학교 대학원 전자계산학 이학석사
1998년 충북대학교 대학원 전자계산학 이학박사
1995년 ~ 현재 충청대학 컴퓨터학부 교수
〈관심분야〉 시간지원 데이터베이스, 시공간데이터베이스, e-Learning