Journal of Digital Contents Society Vol. 9 No. 2 June. 2008(pp. 251-260)

효율적인 클러스터링을 이용한 관심 정보 추출을 위한 웹 마이닝

김성학*, 안병태**

요 약

인터넷의 전자상거래에서 대규모 정보 저장소에 있는 원하는 정보를 신속하게 검색하기란 어렵다. 대부분의 전자상거래 사이트에 있어서 정보를 제공하는 방법으로는 통계적 분석이나 분류별 지향의 간단한 과정을 통해 생성된다. 그러나 이러한 것은 생성 정보들 사이의 다양한 상호관계를 표현할 수 없고 사용자의 정확한 구매 패턴을 반영하기 어렵다. 본 논문에서는 전자상거래에서 효과적인 클러스터링을 이용한 다양한 관심정보 추출을 위한 효율적인 웹 마이닝을 제안한다. 이러한 방법은 분류별 항목에서 순차 패턴과 상관 규칙을 이용하여 생성 정보들 사이의 보다 적합한 상관관계를 구성하고 제안된 방법을 통해 효율적인 실험 결과를 나타낸다. 그리고 효과적인 클러스터링을 이용하여 신속한 검색을 제안한다.

Web Mining for Discovering Interesting Information using Effective Clustering

Sung-Hark Kim*, Byeong-Tae Ahn**

Abstract

In internet being a repository of massive information, we easily may not find our desired information, this issue also exists in e-commerce which gets rapid growth. In most of e-commerce sites, the methods furnishing information have been made use of statistical analysis or simple process by category-oriented, but these can't represent diverse correlation among products information and also hardly reflect users' purchasing patterns precisely. In this thesis, we propose more efficient web mining ways for discovering interesting information using effective clustering in e-commerce, which get achieved more suitable relationship among products information using both sequential patterns and association rules in category-independent, and experiments show the efficiency of our proposed methods. And we propose search using effective clustering rapidly.

Keywords : Web Mining, Clustering

1. 서론

인터넷을 통한 정보교류와 정보검색이 증가하면서 웹(web)은 방대한 정보를 포함하게 되는 반면, 사용자는 적절한 시간 내에 원하는 정보의 획득이 힘들고 또한 제공되는 많은 정보는 그 가치가 저하되는 등의 문제가 발생하게 된다. 이를 해결하기 위해 웹 서버에 저장되어 있는 로그(log) 데이터와 데이터베이스를 분석하여 사용자의 행동패턴을 파악하고, 효율적인 웹 사이트의 구성과 마케팅방법 등을 개선하기 위한 시도가 진행되고 있다[2][4][5][9][10].

현재 전자상거래에서 인터넷마케팅에 대한 관심이 고조되면서 사용자의 요구를 충족시킬 수 있는 다양한 마케팅의 필요성을 인식하게 되었고, 사용자에게 더 많은 정보와 편리한 사용자 인터페이스를 제공함으로써 보다 많은 고객을 확보하려고 노력하고 있다. 이러한 인터페이스 중의 하나는 사용자가 관심을 갖는 상품과 관련 있는 정보를 추천해주는 서비스이며, 이는 구매

※ 제일저자(First Author) : 김성학
접수일자:2008년02월01일, 심사완료:2008년02월24일
* 유한대학 컴퓨터정보과, saint@yuhan.ac.kr
** 유한대학 경영정보과(교신저자)

정보, 행동 등 사용자로부터의 특정 행동패턴을 추출하여 분석하고 이를 바탕으로 빠른 시간 내에 사용자가 원하는 정보를 제공하는 것이다[4][10]. 이를 위해 사용자가 원하는 정보를 서버에서 저장 및 관리하기 위해서는 효율적 검색이 가능하도록 새로운 클러스터링 기법이 필요하다[11].

따라서 본 논문은 전자상거래에서 카테고리에 독립적으로, 제공되는 상품정보들 간의 보다 적합하고 다양한 연계성을 나타내기 위해서 순차패턴(sequential patterns) 탐사방법[8]에 연관규칙(association rules)[1]을 적용하고, 또한 가중치를 사용하여 다양한 정보들 간의 연관도를 구체적으로 나타내어 보다 효율적으로 관심정보를 추출할 수 있는 웹 마이닝 방법을 제안하고, 실험을 통해 그 효율성을 보인다. 그리고 사용자가 관심 있는 항목에 대해서 카테고리별 상관 저장이 가능하도록 새로운 클러스터링 방법을 제안한다. 사용자의 관심 있는 정보는 메타데이터가 가능한 XML 문서로 저장되며 저장된 문서는 메타데이터의 정보 추출로 보다 효율적 정보 검색 및 관리가 향상된다.

본 논문의 구성은 다음과 같다. 제2장에서 관련연구로서 연관규칙, XML 문서 저장 기법에 대하여 살펴보고, 제3장에서는 연관리스트를 생성하여 고객에게 관련 상품을 추천하는 에이전트에 대해 설명한다. 제 4장에서는 카테고리별 상관관계가 있는 데이터를 저장하는 새로운 클러스터링 방법을 제안하고 제5장에서 본 연구에 사용된 시스템의 실험 및 분석을 다룬다. 끝으로 제6장에서 결론 및 향후과제를 제시한다.

2. 관련 연구

2.1 연관규칙

데이터베이스에 숨겨져 있는 패턴을 탐사하는 연구 중에서 연관규칙에 대해 가장 많은 연구가 이루어졌다[1][3][6][7]. 연관규칙 탐사과정은 크게 두 단계로 진행이 된다. 첫 번째는 높은 지지도(support)를 갖는 즉, 항목간의 연관성이 높다고 가정되는 항목의 집합(itemset)인 빈발 항목집합(frequent or large itemset)를 식별하는 작업이고, 두 번째 단계는 이러한 빈발 항목집합을 이용하여 높은 신뢰도(confidence)를 갖는 연관규칙을 도출하는 작업이다. 여기서 지지도와 신뢰도는 매우 중요한 개념으로 빈발 항목집합과 연관규칙을 찾아내는데 있어서 논리적 타당성을 제공하는 큰 역할을 한다.

- 지지도(support) : 생성된 연관규칙이 전체 항목에서 차지하는 비율을 말한다. 즉, 데이터베이스에 속한 전체 트랜잭션의 개수 중 그 연관규칙을 지지하는 트랜잭션의 비율을 의미하며, 전체 거래 중 X와 Y를 포함하는 거래의 정도를 나타내는 식으로 표현된다.

sp(X∪) = X와 Y를 포함하는 거래 수/전체 거래 수

- 신뢰도(confidence) : 연관규칙의 강도를 의미하며, 전제부를 만족하는 트랜잭션이 결론부까지를 만족하는 비율을 말한다. 즉, X를 포함하는 거래 중에서 Y 가 포함된 거래의 정도를 의미하며 다음과 같이 표현된다.

$$cf(R) = p(Y \subseteq T \mid X \subseteq T) = \frac{p(Y \subseteq T \wedge X \subseteq T)}{p(X \subseteq T)} = \frac{sp(X \cup Y)}{sp(X)}$$

연관규칙을 발견하기 위한 알고리즘으로는 Apriori[1], DHP[6], FP-growth[3] 등이 연구되었는데, 이들 대다수는 많은 프로세싱 시간 및 메모리가 요구되는 빈발 항목집합을 빠르게 발견하여 알고리즘의 성능을 높이고자 하였다.

2.2 XML 문서

XML 문서를 효과적으로 관리하기 위해서는 XML 문서를 효율적으로 저장할 수 있는 방법으로써 순수 내장형 XML 데이터베이스 기술을 이용할 수 있다[12]. 이때 XML 데이터의 특성에 적합한 클러스터링 방법의 지원이 필요하다.

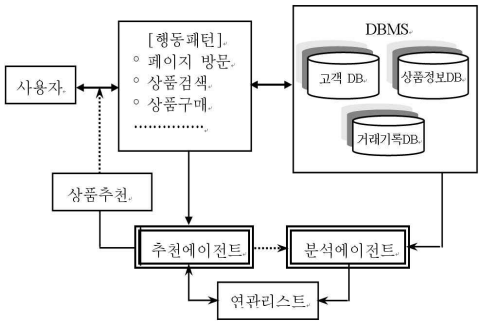
본 논문에서는 XML 문서를 효율적으로 저장할 수 있는 XML 데이터 클러스터링 방법을 기반으로 순수 내장형 XML 데이터베이스 시스템인 버클리 DB XML[13]을 이용하여 XML 문서 관리 시스템을 클러스터링하여 저장한다. XML 데이터베이스 관리는 XML 문서 데이터베이스 관리 기술을 활용할 수 있다. 이러한 이유로 최근 XML 데이터베이스 관리에 대한 연구는 주로

XML 문서 데이터베이스 관리 기술을 적용하는 방법으로 진행되고 있다[14]. 이에 따라 XML 데이터베이스 관리 시스템에 대한 연구는 크게 두 가지로 진행되고 있다. 즉, XML 특성이 지원되도록 XML 문서 관리 기능을 갖춘 기존 데이터베이스 관리 시스템(주로 관계형과 객체 관계형 데이터베이스 시스템)을 확장하는 접근[15][16]과 순수(Native) XML 데이터베이스 관리 시스템을 확장하는 접근[17][18]이 있다.

최근 XML 문서 데이터베이스 관리의 필요성이 높아지면서 XML DBMS에 대한 연구가 활발해졌다. 이에 따라 다양한 형태의 XML DBMS가 개발되고 있다. 대표적인 XML DBMS로는 Tamino[17], InfonyteDB[20] 등이 있다. 이들은 대부분 XML 문서를 작은 단위인 엘리먼트 수준에서 데이터의 갱신을 가능하게 하고 있다. 그러나 전통적인 DBMS의 핵심 기능인 트랜잭션 관리, 병행성 제어, 회복 관리 같은 기능은 제공하지 않고 있다. 그런데 최근에 Sleepycat사에서 데이터베이스 시스템으로 XML을 지원하는 버클리 DB XML이 개발되었다[13]. 이는 공개 소프트웨어로써 이미 기업용 제품으로 데이터베이스 시스템의 기본 기능이 인정된 버클리 DB[21] 위에 구축되었다. 무엇보다 다양한 기기를 위해서 기능을 자유롭게 선택하여 DBMS를 갖출 수 있도록 지원하고 있으며, 라이브러리로 제공되어 클라이언트 응용과 직접 연결되므로 프로세스나 시스템들 사이에 통신을 크게 줄여 우수한 성능을 제공한다.

3. 시스템 설계

본 논문에서 제안하고 있는 시스템의 전체적인 구성도는 (그림 1)과 같으며, 분석에이전트가 연관리스트를 생성하고 이를 추천에이전트가 이용하면서 사용자에게 효율적으로 정보를 제공하게 된다.



(그림 1) 시스템의 전체 구성도

3.1 분석에이전트

분석에이전트는 관련 상품에 대한 정보를 얻는데 이용하며 분석방법은 연관규칙과 순차패턴 탐사기법을 모두 사용한다. 하나의 트랜잭션이 자연스럽게 정의되는 실세계의 바코드 분석과는 다르게 웹 환경하에서는 연관규칙을 찾기 위한 트랜잭션의 정의가 자연스럽지가 않다. 그리고 인터넷 쇼핑몰에서의 거래는 한 번에 많은 물건을 사는 것이 아니라 보통 적은 수의 물품을 구매하기 때문에 시간개념이 들어간 순차패턴 탐사기법이 필요하다.

3.1.1 연관규칙의 적용

모든 트랜잭션을 T, 특정고객과 관련된 모든 트랜잭션을 Ti라고 하고 Ti의 j번째 트랜잭션을 Tij 라고 하면 트랜잭션은 다음과 같은 항목을 갖는다.

T_{ij}.code : 상품코드

T_{ij}.uid : 사용자 ID

T_{ij}.time : 트랜잭션이 발생한 시간, T_{ij} T_i

트랜잭션에는 다른 항목도 있지만 목적을 위해 필수적인 항목만을 선택하였다. 임의의 시간 t에서 발생한 트랜잭션을 Tij라고 하고 다음에 일어난 트랜잭션을 Tij+1, 그리고 δ를 0에 가까운 수라고 한다면 임의의 최소시간 안에 발생한 트랜잭션을 묶은 클러스터는 다음과 같다.

$$Cluster = \{(T_{ij}) \mid 0 \le T_{ij+1}.time - T_{il}.time \le \delta, i, j : \text{정수}\}$$

여기서, 클러스터의 의미는 각 트랜잭션의 발생시간이 보다 작은 것들의 모임 또는 독립 트랜잭션으로 한정한다. 보다 작은 Ti 의 모든 트랜잭션을 클러스터로 묶어 하나의 트랜잭션으로 변환한 것을 ti 라고 하면, ti 는 다음과 같은 형태를 갖는다.

t_i = (T_{ij}, T_{ij+1}, T_{ij+2}) + (T_{ij+3}) + (T_{ij+4}, T_{ij+5}) + (T_{ij+6}, T_{ij+7}) + …

ti 에 속하는 클러스터는 C(ti), 각 클러스터는 Ck(ti)이므로, 연관규칙을 찾기 위해 각각의 클러스터 내에서 고려해야 할 전체조합 CSA는 다음과 같다.

$$CSA = \sum_{i=1}^{|D|} \sum_{k=1}^{|C(t_i)|} \sum_{m=1}^{|C_k(t_i)|} {}_{|C_k(t_i)|}C_m$$

여기서, C : combination, |D| : 상품을 구매한 전체 구매자의 수, |C(ti)| : 특정 구매자에 대한 전체 클러스터의 수, |Ck(ti)| : 특정 클러스터 내의 트랜잭션의 수를 의미한다. 위의 클러스터는 관련된 트랜잭션의 모임이므로, A→B 의 규칙을 찾을 때는 적합하지 않다. 따라서 클러스터의 트랜잭션을 상품으로 변환해야 하며, 또한 이 과정에서 발생하는 중복된 품목을 제거해야 한다. 이렇게 중복이 제거된 클러스터를 |Ck(ti)'|라고 가정하면, CSA' 는 아래와 같다.

$$CSA' = \sum_{i=1}^{|D|} \sum_{k=1}^{|C(t_i)'|} \sum_{m=1}^{|C_k(t_i)'|} {}_{|C_k(t_i)'|}C_m$$

0에 가까운 이내의 트랜잭션들을 하나의 트랜잭션으로 취급해서 분석하기 위해 모든 트랜잭션의 수 |X|를 구하면 아래와 같고,

$$|x| = \sum_{i=1}^{|D|} |C(t_i)|$$

생성되는 규칙 A→B에서의 지지도, 신뢰도 및 개선도(lift)는 다음과 같다. 여기서, 개선도는 임의로 B가 구매되는 경우에 비해, A와의 관계가 고려되어 B가 구매되는 정도를 의미한다.

지지도: $$P(A \& B) = \frac{|A \cap B|}{|X|}$$

신뢰도: $$P(B \mid A) = \frac{|A \cap B|}{|A|}$$

개선도: $$P(B \mid A) / P(B) = \frac{|A \cap B \,||\, X|}{|A \,||\, B|}$$

여기서, A∈S ‘, B∈S ‘, (A∩B)∈S ‘ 이고 |X|는 전체 트랜잭션의 수 및 |A∩B|는 A와 B가 속한 거래가 발생한 횟수를 의미한다. 즉. A∩B가 S 에서 나온 횟수를 뜻하는 것이다.

3.1.2 순차패턴의 적용

규칙 A→B를 만족시키는 집합을 (A, B)라고 하면, ‘A의 발생시간 ≤ B의 발생시간’이고 B는 단일 품목이지만 A는 다중 품목일 수 있다. 시간의 순서에 따라 발생된 Ti = {Tij, Tij+1, Tij+2, Tij+3}라는 트랜잭션이 있다고 가정할 때, 규칙 A→B를 만족시키는 경우의 순차집합은 아래의 11개 조합이 존재하게 된다.

(T_{ij}, T_{ij+1}), (T_{ij}, T_{ij+2}), (T_{ij}, T_{ij+3}), (T_{ij+1}, T_{ij+2}), (T_{ij+1}, T_{ij+3}), (T_{ij+2}, T_{ij+3}), ((T_{ij}, T_{ij+1}), T_{ij+2}), ((T_{ij}, T_{ij+1}), T_{ij+3}), ((T_{ij}, T_{ij+2}), T_{ij+3}), ((T_{ij+1}, T_{ij+2}), T_{ij+3}), ((T_{ij}, T_{ij+1}, T_{ij+2}), T_{ij+3})

따라서, 고려해야 할 순차집합은 $\sum_{m=1}^{|T_i|} {}_{|T_i|}C_m$ 이 되며, 전체고객에 대해 순차집합으로 고려해야 할 전체조합 CSS는 다음과 같다.

$$CSS = \sum_{i=1}^{|D|} \sum_{m=2}^{|T_i|} {}_{|T_i|}C_m$$

여기서 |Ti|는 특정 고객의 모든 트랜잭션 수이고 |D|는 상품을 구매한 전체 구매자의 수이다.

순차패턴 탐사방법에서의 지지도는 순차집합을 지지하는 전체 고객들의 수이므로 A→B에서의 지지도, 신뢰도 및 개선도는 다음과 같다.

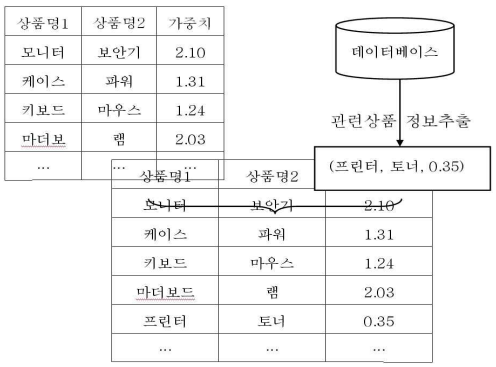
지지도 : $$P(A \& B) = \frac{|A \cap B|}{|D|}$$

신뢰도 : $$P(B \mid A) = \frac{|A \cap B|}{|A|}$$

개선도 : $$P(B \mid A) / P(B) = \frac{|A \cap B|}{|A \| B|}$$

여기서 |A|는 A라는 순차집합을 지지하는 전체 고객들의 수, |B|는 B라는 순차집합을 지지하는 전체 고객들의 수, |A∩B|는 (A,B)라는 순차집합을 지지하는 전체 고객들의 수 (A∩B)∈R, |D|는 전체고객의 수를 나타낸다.

3.2 연관리스트



(그림 2) 연관리스트

연관리스트란 서로 관련 있는 품목들을 묶어 놓은 것으로 분석에이전트에 의해서 구성되며, 상품정보를 제공할 때 사용된다. 이러한 연관리스트를 이용한 정보제공은 고객에 대한 과거의 구매기록을 모두 찾아보지 않고도 실시간으로 적합한 상품정보를 제공할 수 있다는 장점이 있다. 관련되는 상품의 정보를 추출하는 방법은 다음과 같다.

[단계 1] : 모든 사용자의 트랜잭션을 가져온다. 모든 트랜잭션을 T, 특정고객과 관련된 모든 트랜잭션을 Ti 라고 가정한다.

[단계 2] : 연관규칙을 찾기 위해 Ti의 모든 트랜잭션을 임의의 작은 시간 보다 작은 트랜잭션을 하나의 클러스터로 묶고 이것을 ti라 한다. ti에 속하는 모든 클러스터는 C(ti)이고 각 클러스터를 Ck(ti)라고 하면, 최소지지도(minimum support)와 최소신뢰도(minim-um confidence)보다 큰 클러스터의 집합을 CSA' 에서 추출하여 X 에 놓는다.

[단계 3] : 순차패턴을 찾기 위해서, Ti에서 2의 크기를 가진 모든 조합을 선택하고 이것을 순차집합이라고 하면 최소지지도와 최소신뢰도보다 큰 순차집합을 CSS에서 추출하여 Y에 놓는다.

[단계 4] : X, Y 에 속한 Xi , Yi 에 대해, '가중치 = (순차패턴에서의 지지도) (1+연관규칙에서의 지지도)'에 의해 가중치를 구한다.

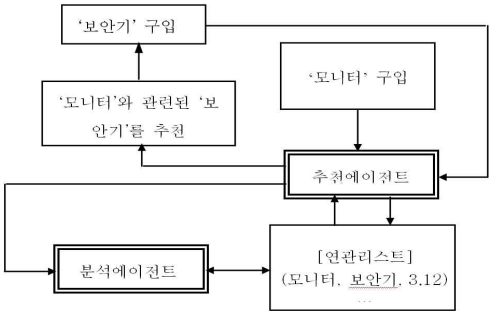
연관리스트는 분석에이전트에 의해 생성되며, 규칙 AB에 대한 연관리스트의 구성은(A, B, 가중치)의 형식을 갖는다. 여기서 A, B는 단일 품목이고 가중치(weight)는 AB를 만족시키는 정도를 나타내는데, 가중치가 클수록 규칙 AB는 적합하다는 것을 의미한다. 규칙이 적절해지려면 충분한 지지도와 신뢰도를 만족해야 하기 때문에, 최소신뢰도(Cmin)와 최소지지도(Smin) 이상의 규칙에 대해서만 가중치를 가질 수 있다. 즉, 규칙 R : AB 에서 conf(R)Cmin, sup(R)Smin 인 경우에만 가중치를 갖게 된다. 또한 순차패턴과 연관규칙에서의 최소신뢰도와 최소지지도 각각의 값들은 서로 다르다.

조건을 만족하는 규칙 R : AB이 있고, 연관리스트에 존재하지 않는다면 이 규칙은 (A, B, 가중치)의 형태로 연관리스트에 삽입된다. [단계 4]에서의 가중치를 구하는 식의 의미는, 규칙 R : AB 에서 conf(R)Cmin, sup(R)Smin 인 경우, AB라는 규칙이 순차패턴에서 나왔고 연관규칙에서도 나왔다면 순차패턴에서만 나타난 것보다 규칙의 신뢰도가 높으므로 가중치를 높여준다는 것이다. 예를 들면, 순차패턴에서의 조건은 conf(R) = 0.4, sup(R) = 0.2이고 연관규칙에서의 조건은 conf(R) = 0.4, sup(R) = 0.1일 때, '모니터를 사면 보안기를 산다'라는 규칙이 순차패턴에서는 conf(R) = 0.45, sup(R) = 0.28, 연관규칙에서는 conf(R) = 0.40, sup(R) = 0.05라고 가정하자. 이 때의 가중치는 0.451 = 0.45이다. 만약 연관규칙에서 conf(R) = 0.40, sup(R) = 0.19가 나왔다면 가중치는 0.45(1+0.4) = 0.63이 된다. (그림 2)는 새로운 관련 상품 정보를 얻은 후, 이를 기존의 연관리스트에 삽입하는 과정을 보이고 있다.

3.3 추천에이전트

분석에이전트가 추천과정에서 back-end의 기능을 담당한다면 추천에이전트는 front-end를 담당한다고 할 수 있으며, 고객의 행위를 모니터링하는 역할을 한다. 고객이 인터넷 쇼핑몰에서 상품을 구매하거나 장바구니에 넣는 행위 등의 경우에, 추천에이전트는 연관리스트를 검색하여 고객이 구매하려는 상품과 관련된 상품정보를 추천하게 된다. 그리고 추천한 상품을 구매하는 경우, 이 사실을 분석에이전트에게 통보해서 연관리스트가 적절하게 유지되는데 도움을 준다.

(그림 3)은 어떤 상품을 구매한 경우에 추천에이전트가 구매된 상품과 관련된 상품정보를 제공해주는 과정을 보이고 있다.



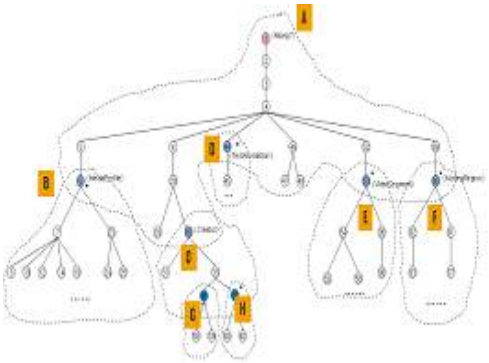
(그림 3) 에이전트들의 수행과정

4. XML 데이터 클러스터링

앞에서 언급된 버클리 DB XML은 일반적인 XML 문서를 관리하는 데이터베이스 시스템으로 개발된 것이다. 따라서 이를 본 시스템의XML 문서 관리에 활용하기 위해서는 XML 고유의 특성을 반영한 새로운 관리 방법이 요구된다. 따라서 본 논문에서는 XML 문서에 적용할 수 있는 새로운 데이터 클러스터링 방법을 제안하였다. 문서의 저장에는 XML 데이터베이스 시스템을 이용한다. 설명을 위해 실제 사용자가 구매한 모니터 항목 정보와 이와 상관 관련성이 있는 정보들을 추출하여 데이터베이스에 XML문서로 저장 후 XML 스키마와 이에 맞게 발생된 XML 문서 예를 사용하였다.

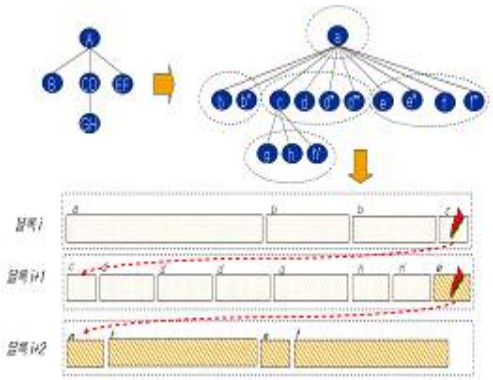
XML 데이터 클러스터링 방법에서는 먼저 각 요소 엘리먼트들 사이에 밀접한 관계가 있음을 나타내는 관련성을 0에서 99까지의 레벨 숫자 값으로 표시한다. 이는 R-CT (Relationship-ClusTer) 속성이라 명명되었으며, 이의 표시를 위해 기존 XML 스키마에 R-CT 속성이 추가되었다. 이 R-CT 속성의 레벨 추출을 위해서 다양한 XML 응용들이 분석되었다. 그리고 이 분석들을 바탕으로 크게 3가지 기준이 설정되었다. 먼저 부모 엘리먼트가 같은 경우에 같은 레벨로 정해지고 R-CT 속성 값은 1로 지정되었다. 또한 부모 엘리먼트가 다른 경우에 다른 레벨로 정해지고 R-CT 속성 값은 2에서 9까지로 지정되었다. 마지막으로 사용자가 XML 스키마를 생성할 때 별도로 정의한 경우를 사용자 정의 레벨로 하고 R-CT 속성 값은 10에서 99까지 지정되었다. 이렇게 재 정의된 XML 스키마는 시스템에 의해 분석되어 의미적으로 함께 다룰 수 있다고 판단되는 엘리먼트를 묶는 의미 블록을 구성한다. 의미 블록을 구성하기 위한 규칙은 두 가지이다 (규칙 S1) 엘리먼트가 R-CT 속성 값을 가지고 자식 노드를 가질 때 동일 의미 블록으로 묶는다. (규칙 S2) 엘리먼트가 카디널리티를 가지고 자식 노드를 가질 때 동일 의미 블록으로 묶는다.

먼저 규칙 S1이 적용되어 각 엘리먼트가 분석된다. 다음으로 규칙 S2가 적용되어 카디널리티를 가지는 엘리먼트를 찾는다. 앞의 예제에 대해서는 R-CT 속성 값이 1에서 3까지를 가지고 있다. (그림 4)는 의미 블록 구성 규칙들을 이용하여 실제 XML 스키마 트리를 의미 블록들로 나눈 최종 형태이다. 크게 8개의 의미 블록이 구성되었다. 그리고 이들 중 R-CT 값에 따라 C와D, E와F, G와H가 통일 의미 블록으로 취급된다.



(그림 4) 의미 블록 스키마 그래프

마지막으로 이 의미 블록 트리를 깊이 우선 순회(DFS)를 하여 같은 의미 블록에서 발생된 문서 인스턴스들이 물리적으로 가깝게 위치되도록 저장한다. 즉, 먼저 루트 노드를 찾고 그에 따르는 자식 노드를 깊이 우선 순회 방식으로 하나씩 방문하여 해당 자식 노드가 지정된 블록 내에 저장 가능한지 확인함으로써 최대한 같은 블록 내에 저장되게 한다. (그림 5)는 앞의 예제에 대해 적용된 최종 블록별 저장 형태를 보이고 있다.



(그림 5) 블록별 저장 형태

XML 데이터 클러스터링 저장 방법은 XML 엘리먼트들 간의 관계성을 고려하였기 때문에, 클러스터링 적용 전과 비교하여 제한된 시스템에서 저장 공간을 줄여주고, 응용에 가장 적합한 의미 블록을 구성해 준다. 하지만 다양한 응용들을 분석해 미리 R-CT값을 재정의 해주어야 하는 선행 작업이 필요하다.

5. 실험 및 분석

이 장에서는 연관규칙과 순차패턴 탐사방법과의 비교실험을 통해 관련상품의 정보를 효율적으로 추출하는 가에 대한 분석을 한다.

5.1 규칙의 생성

실험용 인터넷 쇼핑몰은 100명의 회원을 보유하고 있고 모든 회원들이 각각 5건씩의 거래를 했다고 가정하면, 데이터베이스에는 500건의 트랜잭션이 들어있게 된다. 그리고, 100명의 회원 가운데 몇 명은 특정한 패턴을 가진 거래를 했다고 가정하고 이 패턴을 연관규칙과 순차패턴 탐사방법을 사용하여 알아낼 수 있는 지를 실험하였다. 먼저, 실험을 위한 데이터를 다음과 같이 2가지로 구분한다.

[실험 데이터1] : 30개의 상품가운데 M명의 회원들은 상품을 거래한 순서가 A1→B1→C1→D1→E1인 패턴1에 속하고 다른 M명은 A2→B2→C2→D2→E2인 패턴2에 속하며 나머지 (100-2M)명은 20개의 상품을 랜덤하게 5개씩 구입한 데이터.

[실험 데이터2] : 실험데이터1과 같은 내용이지만, 나머지 (100-2M)명이20개가 아닌 30개의 상품을 랜덤하게 5개씩 구입하고, 패턴1의 A1→B1→C1→D1→E1의 순서에서 A1→B1과, 패턴2의 A2→B2→C2→D2→E2의 순서에서 A2→B2를 같은 장바구니에 담은 상품이라고 가정한 데이터. 즉, A1→B1과 A2→B2의 시간간격은 매우 적어서 하나의 클러스터로 묶이게 된다.

<표 1> 연관규칙을 적용하여 생성한 규칙

	비율(%)	실험데이터1	실험데이터2
'1 < lift'인 규칙 수	10	0	0
	20	0	1
	30	0	3
	40	0	4
	50	2	7
평균지지도(%)	10	0	0
	20	0	1.14
	30	0	1.34
	40	0	1.73
	50	1.06	1.71
평균신뢰도(%)	10	0	0
	20	0	22.02
	30	0	22.37
	40	0	24.71
	50	19	22.24

<표 2> 순차패턴을 적용하여 생성한 규칙

	비율(%)	실험데이터1	실험데이터2
'1 < lift'인 규칙 수	10	32	54
	20	48	76
	30	82	103
	40	129	123
	50	184	114
평균지지도(%)	10	4.76	3.45
	20	6.44	4.18
	30	5.95	4.28
	40	5.15	4.71
	50	4.38	5.51
평균신뢰도(%)	10	71.04	21.44
	20	54.39	23.80
	30	39.22	23.35
	40	30.69	23.37
	50	25.59	26.82

유용한 규칙을 판별하기 위해서 앞에서 설명한 개선도(lift)를 사용하였는데, '1 < lift'인 규칙만이 규칙으로서의 가치가 있다. <표 1>과 <표 2>는 연관규칙과 순차패턴 탐사방법을 적용하여 나온 결과이다. 모든 표에서의 '비율(%)'은 패턴 A, B가 전체에서 차지하는 비율을 의미한다.

5.2 패턴의 인식

데이터베이스에 패턴1 (A1→B1→C1→D1→E1)과 패턴2 (A2→B2→C2→D2→E2) 등이 있을 때, 연관규칙과 순차패턴 탐사방법이 도출해내는 규칙가운데 이 패턴에 관련된 것이 얼마나 있는지의 여부를 살펴봄으로써 유용한 규칙을 얼마나 추출해낼 수 있는 지를 판단할 수 있다. <표 3>과 <표 4>는 실험데이터1(D1)과 실험데이터2(D2)에 연관규칙과 순차패턴 탐사방법을 적용했을 때 나온 패턴A와 패턴B에 관련된 규칙의 수를 보이고 있다.

<표 3> 연관규칙을 적용한 실험결과

	비율(%)	패턴A		패턴B		패턴C (랜덤한규칙)	
		D1	D2	D1	D2	D1	D2
'1 < lift'인 규칙 속에서 차지하는 수	10	0	0	0	0	0	0
	20	0	0	0	1	0	0
	30	0	1	0	2	0	0
	40	0	1	0	3	0	0
	50	1	3	1	4	0	0

<표 4> 순차패턴을 적용한 실험결과

	비율(%)	패턴 A		패턴 B		패턴 C (랜덤한규칙)	
		D1	D2	D1	D2	D1	D2
'1 < lift'인 규칙 속에서 차지하는 수	10	10	10	10	10	12	34
	20	10	10	10	10	28	56
	30	10	10	10	10	62	83
	40	10	10	10	10	109	103
	50	10	10	10	10	164	94

5.3 분석

실험데이터1을 사용하여 실험한 결과, <표 1>과 <표 2>에서 나타난 것처럼 순차패턴 탐사방법이 연관규칙보다는 많은 유용한 규칙을 찾아낼 수 있었으며, 또한 <표 3>과 <표 4>의 결과에서 패턴A, 패턴B 에 대한 모든 규칙을 찾아냈듯이 연관규칙보다는 특정패턴에 속하는 매우 많은 규칙을 찾아낼 수 있었다. 이러한 이유는 단일한 상품의 트랜잭션에서는 연관규칙을 사용하여 규칙을 생성할 수 없기 때문이다.

그러나 실험데이터2를 이용했을 때, 여러 트랜잭션이 하나의 클러스터로 묶일 수 있게 되면 연관규칙은 유용한 규칙을 생성할 수 있게 된다. 인터넷 쇼핑몰에서 이러한 경우는 여러 상품을 장바구니에 담은 후 한 번에 처리하게 되면 각 상품의 시간간격은 줄어들게 되고 연관규칙의 적용이 가능해 진다. 이 때, 장바구니에 담는 순서를 상품을 구매하는 순서로 해석하게 되면 순차패턴 탐사방법이 적용되는 것이다. 따라서, 연

관규칙을 적용하기 위해서는 시간간격이 적은 트랜잭션들을 하나의 클러스터로 묶고 이러한 클러스터를 하나의 트랜잭션으로 취급하여 사용해야 한다.

실험결과에서 보인 것과 같이, 많은 유용한 규칙을 발견할 수 있는 순차패턴탐사와 발견된 규칙의 유용성을 높일 수 있는 연관규칙을 같이 사용하게 되면 더욱 효율적인 패턴탐사를 할 수 있다.

6. 결론 및 향후 과제

본 논문에서는 웹 사용의 증가에 따른 방대한 정보에서 사용자의 관심 정보를 효율적으로 추출할 수 있는 웹 마이닝 방법을 연구하여, 전자상거래 영역에서 사용자에게 보다 효율적으로 관심 정보가 제공될 수 있도록 하였다. 연관규칙은 한 트랜잭션에 많은 상품을 포함하는 바코드 분석과 같은 경우에는 적합하지만, 한 트랜잭션 내에 하나의 상품만을 포함하는 인터넷 쇼핑몰에서의 거래방식에는 적합하지 않으며 보편적으로 순차적인 정보를 갖는 순차패턴 탐사방법이 더 적합하다. 그러나, 장바구니를 이용하는 경우에는 연관규칙이 유용하지만, 상품에 대한 구매시점의 차이를 반영하지 못하기 때문에 연관규칙과 순차패턴 탐사방법을 같이 적용하는 것이 보다 더 효율적이다. 따라서 본 논문에서는 순차패턴 탐사방법에 연관규칙을 적용하여 보다 유용한 정보를 추출하고 또한 가중치를 사용하여 다양한 상품정보들간의 연관도를 보다 적합하고 구체적으로 나타낼 수 있는 웹 마이닝 방법을제안하고, 실험을 통해 그 효율성을 보였다. 이는 다양한 품목군간의 간섭을 줄이면서 효과적으로 품목간에 연관도를 측정할 수 있는 방법이다. 그리고 상품 정보들간의 연관도를 바탕으로 상품정보에 대한 연관 정보를 XML 문서화하여 새로운 클러스터링 기법을 적용한 데이터베이스에 적용함으로써 보다 효율적인 상품 검색이 가능하였다.

향후 과제로는 단위시간에 많이 판매되는 인기품목을 반영하기 위한 가중치 필터링에 대한 연구와 품목분류 체계 등 에이전트의 성능을 향상시키기 위한 연구가 필요하다.

참 고 문 헌

[1] Agrawal R, and Srikant R., "Fast Algorithms for Mining Association Rules in Large Databases", In Proc. Of the 20th Int. Conf. on Very Large Databases, 1994.

[2] Cooley R., Mobasher R., and Srivastava J., "Web Mining: Information and Pattern Discovery on World Wide Web, In Proc. 9th IEEE Int. Conf. On Tools with Artificial Intelligence, 1997.

[3] Han J.,Pei J., and Yin Y., "Mining Frequent Patterns without Candidate Generation", SIGMOD '00, pp. 1-12, Dallas, TX., May 2000.

[4] Ling C.X., and Li C., "Data Mining for Direct Marketing: Problems and solutions", In Proc. 4th Int. Conf. On KDD, pp. 73-79, 1998.

[5] Michael J. A., and Linoff B. G., "Data Mining Techniques For Marketing, Sales, and Customer Support", WILEY COMPUTER PUBLISHING, 1997.

[6] Park J.S., Chen M.S., and Yu P.S., "An Effective Hash-Based Algorithm for Mining Association Rules", In Proc. Of ACM SIGMOD, pp. 175-186, 1995.

[7] Pei J., Han J., Mao R., "CLOSET: An efficient algorithm for mining frequent closed itemsets", In Proc. ACM-SIGMOD Int. Workshop Data Mining and Knowledge Discovery, pp.11-20, 2000.

[8] Srikant R., and Agrawal R., "Mining Sequential Patter-ns: Generalizations and Performance Improvements", In Proc. of 5th Int. Conf. on Extending Database Technology, pp. 3-17, 1996.

[9] Zhu T., Greiner R., Häubl G., "Predicting Where a Web User Wants to Go", CHI2003, Workshop on Best Practices and Future Visions for Search User Interfaces, April 5-10, Florida, USA, 2003.

[10] Zhu T., Greiner R., Häubl G., "An effective complete-web recommender system", In the Int. World Wide Web Conf., Budapest, Hungary, May 2003.

[11] A. Guillaume and F. Murtagh, "Clustering of XML Document," Computer Physics Communications, Vol. 127, pp. 215-227, May 2000.

[12] T. Fiebig, et al., "Anatomy of a Native XML Base Management System," VLDB, Vol. 11, No. 4, pp. 292-314, 2002.

[13] Sleepycat, "버클리 DB XML, Berkeley DB," http://www.sleepycat.com

[14] U. Westermann and W. Klas, "An Analysis of XML

Database Solutions for the Management of MPEG-7 Media Descriptions," ACM Computing Surveys, Vol. 35, No. 4, pp. 331-373, December 2003.
[15] MicroSoft, " XML for SQL Server 2000," http://www.microsoft.com/technet/prodtechnol/sql/2000/evaluate/xmlsql.mspx
[16] Oracle, "Oracle 9i," http://www.oracle.com
[17] Software AG, "Tamino," http:www1.softwareag.com/corporate/products/tamino
[18] X-Hive, "X-Hive/DB," http://www.x-hive.com/products/db/index.html
[19] K. Eunjung, "The Design of Effective Storage Structure and Index Model in XML Document" KISC, Vol. 29, No. 02, pp. 0110-0118, 2001.
[20] Infonyte,"Infonyte DB 3.0," http://www.infonyte.com/
[21] Sleepycat, "Berkeley DB," http://www.sleepycat.com/products/bdb.html

김 성 학

2004년 : 건국대학교 컴퓨터공학 공학박사
현 재 : 유한대학 컴퓨터정보학과 교수
관심분야 : 웹 마이닝, 데이터 마이닝, 전자상거래

안 병 태

2006년 : 국립경상대학교 컴퓨터과학부 공학박사
1999년~2006년 : Best Click Computer 대표
2006년~현 재 : 유한대학 경영정보과 교수
관심분야 : 멀티미디어 데이터베이스(Multimedia Database), XML, MPEG-7, 유비쿼터스(Ubiquitous), 모바일(Mobile)등