FP-Tree를 기반으로 한 웹 사용 패턴에 대한 순차적 연관성 탐색 기법

김영희[O†] 강우준[‡] 김응모[†]
[†]성균관대학교 정보통신공학부 컴퓨터공학과, [‡]그리스도신학대학교 경영정보학부
pink77hee@skku.edu[O†], wjkang58@hanmail.net[‡], umkim@yurim.skku.ac.kr[†]

A Sequential Association Rules Searching Methods for Web-Usage Patterns Based On Frequent-Pattern Tree

Y.H. Kim[O†] W.J. Kang[‡] U.M. Kim[†]

[†] Dept. of Computer Science & Engineering, SungKyunKwan University
[‡] Dept. of Management information Technology, Korea Christian University

요 약

대용량 웹 데이터베이스로부터 필요한 관련 정보를 탐색하고, 다양한 형태의 정보로부터 지식을 창출하는 일은 매우 어려운 일이다. 본 논문은 복잡하고 다양한 형태의 패턴이 존재하고, 연속된 입력을 갖는 웹 데이터베이스에서 발생되는 빈발 패턴들을 효과적으로 저장할 수 있는 FP-Tree를 기반으로 하여 변화된 정보들을 능동적으로 유지하고 새로운 정보들에 대해 FP-Tree를 재구성하여 웹 페이지에 대한 유용한 패턴 정보와 사용자의 웹 사용 패턴 분석을 용이하게 한다. 그 결과 새로이 발견된 웹 사용 패턴들을 통해 웹 페이지의 구조적 정보와 구조적 연관 정보를 효과적으로 얻을 수 있다.

1. 서 론

빈발 패턴(Frequent Pattern)마이닝은 연관 규칙, 순차 패턴, 에피소드, 구조 패턴, 분류화, 군집화 등과 같은 데이터 마이닝 작업에서 필수적인 역할을 할 뿐만 아니라, 빙산 입방체 계산(iceberg cube computation), 입방체 변화율 분석(cube gradient analysis)과 같은 다른 많은 문제 해결에도 확장하여 사용될 수 있는 마이닝 기법이다. 빈발 패턴 마이닝에 사용되고 있는 알고리즘으로는 Apriori, AprioriAll, AprioriSome, Dynamic-Some, DHP와 같은 알고리즘들이 있다. 그러나 이러한 알고리즘의 대부분은 빈발 항목들을 생성하는 과정에서 많은 양의 후보 항목 집합들을 생성한 후, 각 후보 항목들에 대한 지지도를 계산하여 이들 가운데 최소 지지도를 만족하는 빈발 항목 집합들을 결정하기 위해 여러 번의 데이터베이스 스캔이 반복 되어진다. 따라서, 다양한 형태의 패턴이 존재하거나, 긴 패턴들이 존재할 때는 비용이 많이 드는 문제들을 가지게 되고, 또한 빈발 항목 집합을 결정하는 과정에서 최소 지지도를 만족하지 못하는 후보 항목들은 빈발 패턴에서 무시하게 되는데, 정적 트랜잭션 데이터 집합과 비교할 때 시계열 데이터의 경우는 정보에 대한 경로의 변화가 크고 복잡하여 빈발 패턴에서 무시되어진 후보 항목들이 후에 유용한 데이터가 될 수 있다. 따라서 이러한 데이터를 무시한다면 중요한 정보를 잃을 수 있다.

본 논문에서는 대용량의 데이터를 다루고, 시간적 변이에 따라 빈발 패턴의 형태가 다양하게 변화되는 웹 데이터에 대해 효율적인 저장구조를 가지고 있는 FP-Tree (Frequent-Pattern Tree)[3]를 기반으로 하여 계속적으로 변화되는 입력 데이터들의 새로운 정보들에 대해서 빈발 패턴 히스토리를 능동적으로 유지하고, 웹 데이터에 대한 유용한 순차 패턴 탐색, 웹 페이지간의 구조에 대한 연관 구조 정보를 빠르게 분석, 유지할 수 있도록 한다.

본 논문의 구성은 다음과 같다. 제 2장의 관련 연구로 웹 사용 마이닝과 빈발 패턴, 순차 패턴등의 마이닝 기법에 대해 알아보고 제 3장에서는 FP-Tree(Frequent-Pattern Tree)의 구성과 새로운 입력 정보에 따른 FP-Tree의 재구성에 대해 살펴 본 후 제 4장에서는 웹 페이지간의 구조에 대한 순차적 연관 구조 정보를 분석한다. 끝으로 제 5장에서는 결론 및 향후 과제로 끝을 맺는다.

2. 관련 연구

웹 마이닝[2]은 웹 컨텐츠 마이닝, 웹 구조 마이닝, 웹 사용 마이닝의 3가지 영역으로 분류될 수 있다. 웹 사용 마이닝은 웹 서버 로그로부터 사용자들의 접속 유형을 자동적으로 찾아주어 사용자들이 자주 방문한 웹 페이지와 평균 접속 시간, 자주 발생되어지는 오류 정보, 웹 트래픽과 같은 정보 제공과 사용자들의 웹 페이지 방문 순서를 통해 웹 페이지간의 순차적 구조 정보를

제공하므로 웹을 사용할 때 그 안에 잠재된 유용한 정보를 효율적으로 제공할 수 있다. 웹 사용 마이닝은 전처리 단계, 패턴 발견, 패턴 분석의 3단계로 처리 되어진다. 전처리(preprocessing)단계에서는 웹 서버 로그에 자동으로 발생된 많은 양의 트랜잭션들에 대해 불필요하고 관련이 없는 데이터를 정제 및 클리닝하는 단계이다. 둘째 단계는 패턴 발견을 위해 정제된 자료 및 정보를 이용하여 연관 규칙, 규칙 기반, 순차 패턴, 군집 분석 등의 마이닝 기법을 이용하여 특정한 패턴을 찾는 작업을 수행하다. 이 단계에서 중요한 문제가 빈발 항목 패턴을 효과적으로 찾는 방법으로 Apriori-like 알고리즘과 DIC, DHP, TreeProjection 알고리즘이 주로 사용되어지고 있다. 빈발 패턴 마이닝(Frequent Pattern Mining)은 항목 집합 $I = \{i_1, \cdots, i_m\}$ 일 때, 항목집합 $X \subseteq I$ 인 항목 집합 X는 항목집합 I 의 부분집합이다. 따라서, $X=i_{j1} \cdots i_{jn}$로 나타내고, ℓ개의 항목을 갖는 항목집합을 ℓ-항목집합이라 표현한다. 트랜잭션들의 집합을 데이터베이스 TDB라 하고 만약 $Y \subseteq X$ 이면 $T=(tid, X)$는 항목집합 Y를 포함한다. TDB에서 항목 집합 X의 지지도를 $supTDB(X)$ 또는 $sup(X)$로 나타낼 때 TDB에서 X를 포함하는 트랜잭션의 수는 다음과 같다.

$sup(X) = |\{(tid, Y) | ((tid, Y) \in TDB) \wedge (X \subseteq Y)\}|$

이들 가운데서 주어진 최소지지도 min_sup에 대해 $sup(X) \geq min_sup$을 만족하는 X를 빈발 항목, 빈발 패턴이라 한다.

순차 패턴[1]은 항목 집합 i를 $(i_1, i_2, \cdots, i_m)$으로 표현고 이때 i_j는 하나의 항목이고, 시퀀스 s는 $<s_1, s_2, \cdots, s_n>$으로 표현하며 s_j는 하나의 항목 집합을 나타낸다. 시퀀스는 항목 집합들의 순서화된 형태로 만일 $a_1 \subseteq b_{i1}$, $a_2 \subseteq b_{i2}$, $\cdots$, $a_n \subseteq b_{in}$을 만족하는 정수 $i_1<i_2,<\cdots<i_n$ 이 있을 때, 시퀀스 $<a_1, a_2, \cdots, a_n>$은 다른 시퀀스 $<b_1, b_2, \cdots, b_m>$에 포함된다고 말한다. 순차 패턴 마이닝은 사용자가 정의한 최소 지지도를 갖는 시퀀스인 빈발 시퀀스를 추출하고 이들 가운데 어떤 다른 빈발 시퀀스에도 포함되지 않는 최대 시퀀스를 찾는다. 마지막 단계는 발견된 패턴들을 이용하여 쿼리, 필터링등을 통해 흥미로운 규칙이나 유용한 패턴을 추출하기 위해 패턴을 분석하여 패턴의 유용성 여부를 결정하게 된다.

3. 연속된 데이터에 따른 FP-Tree의 구성

3.1 Frequent-Pattern Tree의 구성

FP-Tree는 빈발 패턴에 대한 중요한 정보와 압축된 데이터의 저장을 위해 확장된 전위 트리 구조로 정적 데이터베이스의 마이닝에 쉬운 기본적 구조를 제공한다.

FP-Tree를 구성하는 과정에 대한 간단한 예를 살펴보면 원시 트랜잭션 데이터베이스에 대하여 빈발 항목집합을 생성 후 최소 지지도를 만족하는 빈발 항목들이 표 1과 같다고 했을 때 FP-Tree는 그림 1과 같이 생성된다.

표 1. 내림차순 빈발 항목 데이터베이스

tid	X
100	{a, b, c, d, e, f}
200	{a, b, c, d, e}
300	{a, d}
400	{b, d, f}
500	{a, b, c, e, f}

FP-Tree를 생성하는 자세한 내용은 [3]을 참조한다.

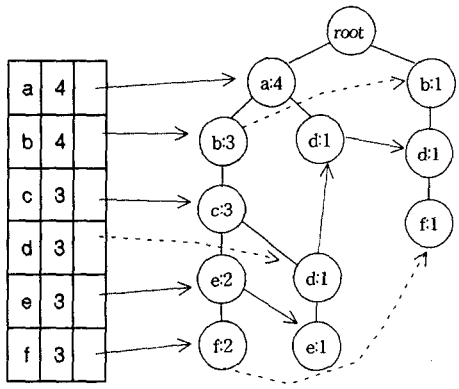

그림1 FP-Tree

3.2 히스토리와 연속된 정보에 대한 FP-Tree구성

연속된 데이터가 발생되는 경우 Apriori-like 알고리즘은 조인 연산을 수행하고 조인 연산은 블록킹 연산자로 연속된 데이터의 형태에서는 수행되기 어렵다. 또한, 비빈발 항목들이 연관된 정보로 이후에 수행될 때 히스토리 정보를 위한 비빈발 정보를 모두 주 기억장치에 저장하기 어려운 문제를 가지고 있다. 따라서, 빈발 히스토리를 가진 패턴 모임은 FP-Tree와 유사한 형태로 들어오는 트랜잭션에 증가적으로 갱신되어 저장될 수 있는 방법을 통해 웹 데이터 환경에서 빈발 패턴의 추출과 갱신을 효과적으로 수행할 수 있다. 그림 2는 연속된 빈발 패턴들의 정보를 오버랩(overlap)하여 데이터베이스 테이블에 계속적으로 업데이트 하고 이들 테이블 정보를

통해 그림3, 그림4와 같이 T_1, T_2, … T_n 각각에 대한 FP-Tree를 구성한다.

T_1	
a	4
b	4
c	3
d	3
e	3
f	3

T_2	
b	5
a	4
c	4
d	3
e	3
f	3

…

T_n	
a	4
b	4
c	4
e	3
f	3
g	3

그림 2. 연속데이터 빈발 패턴 데이터베이스

그림 3과 그림4의 FP-Tree에서 보면 연속된 정보들이 오버랩되면서 T_1 FP-Tree에서 빈발 패턴으로 발견되지 않았던 항목 g에 대해 T_n 에서는 빈발 패턴으로 발견됨을 알 수 있고, 연속된 데이터들 사이의 새로운 정보들이 발견되어짐을 알수 있다. 이와 같이 대용량의 웹 데이터들의 빈발 패턴을 발견하는 과정에서 변화되는 정보들에 관해 히스토리 정보를 유지하고, 새로운 정보을 업데이트 하며 FP-Tree를 구성한다면 비빈발 패턴의 모든 정보를 유지하기 위해 주기억장치에 저장할 필요가 없고 변화된 빈발 패턴의 모든 정보를 발견할 수 있다.

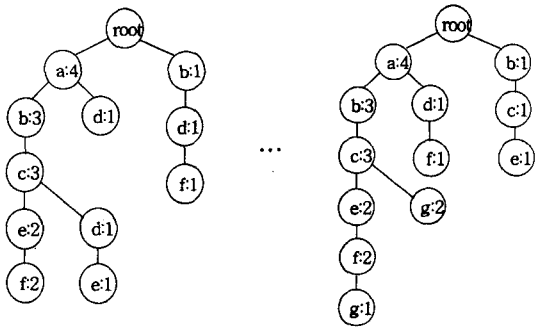

그림 3 T_1 FP-Tree　　그림 4 T_n FP-Tree

4. 순차적 연관 구조 분석

연속된 새로운 정보들의 입력에 따라 구성된 FP-Tree를 기반으로 하여 마이닝한 결과로 Maximal Frequence Sequence가 a→b→c→e→f에서 a→b→c→e→f→g로 변화됨과 새로운 빈발 항목 g를 얻으므로 기존에 유용하지 못했던 빈발 패턴이 중요한 패턴이 됨을 알 수 있다. 반면, 항목d는 변화된 빈발 패턴 항목집합에서 제거됨을 알 수 있다. 그리고, 새롭게 구성된 FP-Tree를 통해 사용자들이 빈발하게 접근하고 있는 웹 페이지와 각각의 항목 a,b,c,d,e,f,g의 빈발 패턴의 지지도를 지속적으로 알 수 있을 뿐만 아니라, 사용자들의 웹 사용 행동 패턴이 b→d→f에서 b→c→e의 형태로 변화된 새로운 정보를 알 수 있다. 또한, 이렇게 새로이 발견된 웹 사용 패턴들을 통해 웹 페이지의 구조적 정보와 구조적 연관 정보를 효율적으로 얻을 수 있다.

5. 결론 및 향후 과제

본 논문에서는 연속된 정보를 가진 대용량 웹 데이터베이스에서 빈발 패턴을 발견하기 위해 FP-Tree구조를 기반으로 하여 연속된 빈발 항목 히스토리 테이블에 따른 새로운 정보를 가진 FP-Tree를 구성하여 웹 페이지에 대한 유용한 패턴 정보와 새로운 정보들을 효과적으로 발견할 수 있도록 하였다.

향후 과제로 대용량의 데이터베이스에서 빈발 패턴을 보다 빠르고 능동적으로 발견하기 위한 방법과 발견된 패턴들을 효율적으로 분석할 수 있는 방법들이 앞으로 연구되어야 할 것이다.

참고 문헌

[1] R. Agrawal and R. Srikant. Mining sequential patterns. In Proc. 1995 Int. Conf. Data Engineering (ICDE'95), pages 3-14, Taipei, Taiwan, Mar. 1995.

[2] Robert Cooley, Pang-Ning Tan, and Jaideep Srivastava. Discovery of interesting usage patterns from web data. Technical Report TR 99-022, University of Minnesota, 1999.

[3] Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining. Knowledge, Discovery. 8(1), 53-87, 2004.

[4] Ming-Yen Lin and Suh-Yin Lee. Improving the Efficiency of Interactive Sequential Pattern Mining by Incremental Pattern Discovery. Proceedings of the 36th Hawaii International Conference on System Sciences(HICSS'03), 2002.