2001 년 한국정보처리학회 춘계 학술발표논문집 제 8 권 제 1 호

웹 로그에 대한 온라인 연관 규칙 기법

박은주, 권혜련, 김은주, 이일병
연세대학교 컴퓨터과학산업시스템공학과
e-mail : ejpark@csai.yonsei.ac.kr

Online Association Rule Technique for Web Access Log

Eun-Joo Park, Hye-Ryun Kwon, Eun-Joo Kim, Yill-Byung Lee
Dept. of Computer Science & Industrial System Engineering, Yonsei University

요 약

본 논문에서는 웹에서 온라인상으로 발생되는 기록 데이터들의 연관 규칙을 구성할 수 있는 효과적인 기법을 제안하고 있다. 기본적으로, 온라인상에서 연관 규칙을 추출하는 방법은 Carma 알고리즘을 바탕으로 하였기 때문에 최대 데이터의 scan 회수를 2 회로 유지하였다. 각 사용자가 방문한 웹 사이트의 수에 대하여 정규 분포를 따르는 가중치를 Phase I 알고리즘의 지지도 관련 변수에 영향을 줌으로써, lattice 의 크기를 조절하는 요소로 사용하여 처리 시간을 단축시키고 있다. 기존의 Carma 알고리즘과 제안하는 W-Carma(Weighted-Carma) 알고리즘과 처리 시간을 비교하였으며, 대량의 데이터일 경우 좋은 성능을 보이고 있다.

1. 서론

지금까지 누적된 다양한 종류의 데이터들을 이용하여, 다양한 데이터마이닝 기법이 제시되어 왔고, 실제로 경영, 마케팅, 금융등 다양한 분야에서 활용되고 있다. 특히, 연관 규칙(Association Rule)은 Market Basket Analysis(MBA)로 불리기도 하며, 임의의 항목집합과 다른 항목 집합 사이에 연관성을 분석하는데 이용되는 기법이다. 최종적으로 연관 규칙은 구매자들의 구매 패턴 및 기타 유용한 행동 패턴을 알아내어 Cross-Selling 을 하거나, 시장성 예측, 의사 결정 시스템등을 지원하고 있다. 하지만, 구매자의 수와 물품들의 수가 증가할수록 연관 규칙을 구하기 위한 계산량은 기하급수적으로 증가하게 된다. 이와 같은 계산량을 감소시키기 위하여, 지금까지 많은 연구가 진행되어 왔다. 최근에는 온라인 상태에서 사용자의 요구에 실시간으로 서비스하기 위한 연구가 이루어지고 있다. [10][12]

본 논문에서는 기존에 제시된 Carma [10] 알고리즘을 바탕으로 웹 데이터의 패턴을 정리하고, 트랜잭션(transaction)의 카디날러티(Cardinality)에 따라 정규 분포를 따르는 가중치를 계산한다. 가중치는 Carma 의 Phase I 단계에서 maxSupport 값을 갱신하여, 이후 단계인 전지(prune) 단계에서 lattice 의 크기를 줄이고, 빈발하지 않는 항목집합(small itemset)을 제거하므로, 결과적으로 처리 시간을 단축시키게 된다.

논문의 구성은 다음과 같다. 2 장은 연관 규칙 탐사의 기본 개념을 설명하였고, 3 장에서는 지금까지 연구되어온 연관 규칙 탐사 기법과 W-Carma 의 기본 알고리즘으로 사용하고 있는 Carma 에 대해서 살펴보도록 하겠다. 4 장에서는 제안하는 알고리즘, W-Carma 를 소개하고, 5 장에서는 실험을 통하여 제안한 알고리즘을 분석하였다. 끝으로 6 장에서는 W-Carma 의 고찰 및 향후 연구 방향으로 결론 맺었다.

2. 연관 규칙

2.1 빈발 항목집합의 정의

트랜잭션이 빈번하게 발생하는 소매점의 물품 판매에서 만들어지는 트랜잭션 데이터베이스[1][2]를 고려해 보자. 항목들의 집합 $I=\{i_1,i_2,\ldots,i_m\}$이 주어진다고 할 때, 트랜잭션 T 는 I 의 부분집합으로 정의된다. 집합과 같이 트랜잭션들은 중복된 항목을 허용하지 않는다. 그러나 우리는 순수한 집합의 개념을 확장하고 트랜잭션과 다른 모든 항목집합들 내에 있는 항목들은 정렬된 것으로 가정한다. 데이터베이스 D 를 n 개의 트랜잭션들의 집합이라 하고 각 트랜잭션은 고유한

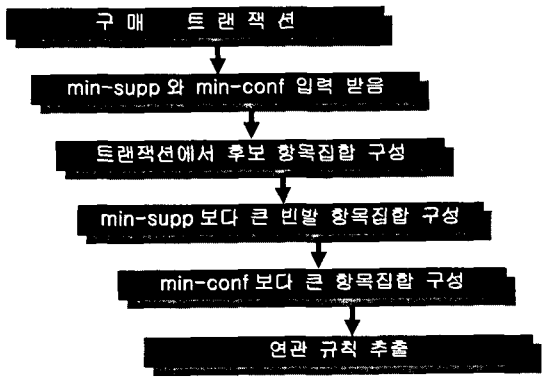
트랜잭션 번호(TID)가 부여된다. 만일 트랜잭션 T 가 X 의 모든 항목들을 포함한다면, T 가 집합 X 를 지지한다(support)라고 한다. 우리는 X 의 지지도를 생략된 형태 supp(X)로 정의하며 이는 X 를 지지하는 D 에 있는 모든 트랜잭션들의 개수를 의미한다. 만일 사용자가 정한 최소 지지도 s_{min}에 대하여 $supp(X) \geq s_{min}$이라면, 집합 X 는 빈발하다. (이런 경우에 항목들의 집합 X 를 일반적으로 large itemset 이라 하고 또는 frequent itemset 이라고도 한다, 빈발하지 않은 항목집합들을 small itemset 이라고 한다.)라고 한다. 최소 지지도를 사용하는 이유는 D 에 대하여 관심있을 정도로 빈발하게 나타나는 항목만을 고려하기 위함이다. 항목집합 X 의 개수를 k=|X|로 나타내고 이를 k-항목집합이라 부른다.

2.2 연관 규칙의 정의

X 와 Y 를 항목들의 집합이라 하자. 연관 규칙은 R:X=>Y 형식의 함축이고, 이때 X 와 Y 는 서로 같은 원소를 갖지 않는 항목집합이다. $X, Y \subseteq I$ 이고, $X \cap Y = \emptyset$ 이고, $Y \neq \emptyset$여야 한다. X 를 규칙의 조건부(antecedent)라 하고, Y 를 결과부(consequent)라 한다. 만일 한 트랜잭션이 X 를 지지한다면, 또한 어떤 확률에 의해 Y 도 지지할 것이라는 예측으로 이해될 수 있는 것이 연관 규칙이다. 이런 확률을 이 규칙의 신뢰도(confidence, conf(R)로 표시)라 한다. R 의 신뢰도는 X 를 지지하는 T 에 대하여 Y 또한 지지할 조건부 확률로 정의된다. 즉, $conf(R) = supp(X \cup Y) / supp(X)$가 된다.

D 에 있는 규칙 R 에 대한 지지도는 $supp(X \cup Y)$로 정의한다. 규칙이 데이터베이스에서 의미가 있기 위해선 충분한 지지도와 신뢰도를 가져야한다. 그러므로 어떤 주어진 최소 신뢰도 c_{min} 과 최소 지지도 s_{min} 에 대하여 $conf(R) \geq c_{min}$ 이고, $supp(R) \geq s_{min}$ 하면, 규칙 R 은 D 에 대하여 성립한다.

이와 같은 연관 규칙 탐사의 접근 방식은 다양한 알고리즘이 존재하지만 대부분 기본적 스키마를 사용한다.[8]



[figure 2-1] 연관 규칙의 흐름도

단계 1. 빈발 항목집합을 찾아낸다.
단계 2. 빈발 항목집합을 사용하여 최소 신뢰도를 만족하는 연관 규칙을 생성한다.

3. 기존 연구

연관 규칙의 문제는 1993 년 AIS 알고리즘에 의해 제안되어졌으며[1], Apriori 에서 후보 항목집합(candidate itemset)을 효과적으로 구하는 방법이 발표되었다. [2] 1995 년도에 제안된 Partition 알고리즘은 전체 트랜잭션의 scan 수를 2 회로 줄이기는 하였으나, 데이터들의 특성상 기대 이상의 성능 향상을 가져오기는 힘들었다.[3] 샘플링을 이용한 기법은 연관 규칙의 탐사 시간을 줄이는 방법을 제안하였다.[6] DIC 은 기존의 방법보다 더 작은 패스 회수로 빈발 항목집합들을 찾아낼 수 있는 자료 구조를 제안하였으며, 규칙의 유용성 측정 방법에도 확신도(conviction)를 기초로 한 함축 규칙(implication rule)을 제시하였다.[7]

본 논문에서 기본 알고리즘으로 응용하고 있는 Carma(Continuous Association Rule Mining Algorithm)은 1998 년 Christian Hidber 에 의해 제안된 알고리즘으로, 크게 Phase I 과 Phase II, 2 개의 알고리즘으로 구성되어 있다. 빈발 항목집합으로 구성되는 lattice V 와 V 에 속해있는 항목집합 v 에 대하여, count(v), firstTrans(v), maxMissed(v)가 알고리즘의 주요 변수들이다.

count(v)는 v 를 lattice 에 삽입하면서 누적되는 값으로, 각 집합들의 발생 빈도를 알 수 있다. firstTrans(v)는 항목집합 v 가 lattice 안으로 처음 삽입 될 경우, 트랜잭션상의 인덱스 값이 저장된다. maxMissed(v)는 Carma 의 핵심이 되는 부분으로, v 가 lattice 에 삽입하기 전 상태에 v 의 발생 빈도수의 상한지를 저장하게 된다. 그밖에도, lattice 안의 항목집합들은 현재 읽는 트랜잭션의 인덱스를 i 라 할 경우, count(v)/i 값을 가지는 minSupport 와 (maxMissed(v)+count(v))/i 의 값을 가지는 maxSupport 값이 유지된다.

연속 트랜잭션의 i 번째 트랜잭션에 대한 지지도 임계치인 ρ_i 는 다음과 같다. 단, $avg(\sigma_i)$는 i 까지의 평균 지지도 임계치이다.

$$\rho_i = avg(\lceil \sigma_i \rceil) + (|v|+1)/i$$

Phase I 은 lattice 관리를 위해 3 단계로 구성이 된다.

1) 트랜잭션을 읽어 항목집합 v 의 지지도를 계산하기 위하여 증가시키고, 2) lattice 안에 항목집합 v 가 이미 존재하지 않을 경우에는 lattice 에 삽입하고, firstTrans(v)를 트랜잭션상의 v 의 현재 인덱스로 설정한다. 3) v 의 카디날러티가 2 보다 같거나 크고, maxSupport(v)가 지지도 임계치보다 작은 경우의 항목집합을 lattice 에서 전지한다. 이 때, lattice 에서 1-항목집합은 전지하지 않는다.

```
Function Phase I ( transaction sequence(t1,.....tn),
                   support sequence σ = (σ1,...., σn)
                   ) : support lattice;
support lattice V;
begin
  V:={0};
  MaxMissed(v):= 0; firstTrans(v):=0; count(v):=0;
  for i from 1 to n do
  //1) Increment
  for all v∈V with v⊆ti do count(v)++; od;
  //2) Insert
  for all v⊆ti with v∉ V do
  if ∀w⊂v: w∈V and maxSupport(w) ≥ σi then
  V:=V U {v}; FirstTrans(v) := i; count(v) := 1;
  maxMissed(v) :=
  min{⌊(i-1)avgi-1(⌈σ⌉i-1)⌋+|v|-1,maxMissed(w)+count(w)-
1|w⊂v};
  if |v|==1 then maxMissed(v):=0; fi;
  od;
  //3)Prune
  if (i%max{⌈i/σi⌉,500}) == 0 then
  V:={ v∈V | maxSupport(v)≥ σi or |v|==1};
  fi;
  od;
  return V;
end;
```

[figure 3-1] Carma 의 Phase I

```
Function Phase II ( support lattice V,
                    transaction sequence(t1,.....tn),
                    support sequence σ )
                  : support lattice;
integer ft, i=0;
begin
  V:=V\{ v∈V | maxSupport(v) < σn};
  While ∃v∈V:i < firstTrans(v) do
    i++;
    for all v∈V do
      ft := firstTrans(v);
      if v ⊆ ti and ft<i then
        count(v)++; maxMissed(v)--;
      fi;
      if ft==i then
        maxMissed(v) := 0;
        for all w∈V:
          v⊂w and maxSupport(w) > maxSupport(v) do
          maxMissed(w) := count(v) – count(w);
        od;
      fi;
      if maxSupport(v) < σn , then V:=V\{v}; fi;
    od;
  od;
  return V;
end;
```

[figure 3-2] Carma 의 Phase II

Phase II 에서는 lattice V 로부터 빈발하지 않는 항목집합들을 모두 전지하고, 남아 있는 모든 항목집합의 정확한 지지도를 계산한다.

예를 들어, 지지도의 임계치보다 적은 maxSupport 값을 가진 항목집합을 V 에서 삭제하고, 모든 항목집합을 계산하여 firstTrans 가 될 때까지 수행하여, 각 항목집합의 정확한 지지도를 계산한다.

마지막으로 Carma 는 앞서 설명한 Phase I 와 Phase II 를 순차적으로 호출함으로써 수행되며, 결과로 전체 빈발 항목집합을 가진 lattice 를 구성할 수 있다

```
Function Carma ( transaction sequence T,
                 support sequence σ)
               : support lattice;
support lattice V;
begin
  V := Phase I (T, σ);
  V := Phase II (V, T, σ);
  return V;
end;
```

[figure 3-3] Carma

4. W-Carma

W-Carma(Weighted-Carma)는 각 웹사이트를 방문한 사용자별 트랜잭션에 대한 가중치와 Carma 를 연결한 연관 규칙 기법이다.

먼저, 웹 사이트의 뷰(view)에 대해서 정리하여, 각 뷰와 관련한 가중치를 정의한다.

전제 1. scanning view

서버에 접근한 한 사용자에 의해 단시간 동안 전체 사이트에 가까운 수의 뷰가 있을 경우에 해당 뷰는 낮은 가중치를 갖는다.

전제 1 은 웹이라는 익숙한 인터페이스로 인한 행동 패턴으로 의도적 또는 우연하게 웹사이트에 방문할 경우, 해당 사이트에 링크 된 부분을 습관적으로 모두 클릭하여 발생하는 트랜잭션으로 전체적인 사이트의 뷰를 가지고 있을 경우에는 연관규칙을 구성, 또한 전반적인 사이트의 발생 빈도를 가지게 되므로 규칙을 구성하기위한 항목으로써의 가중치가 적어진다.

전제 2. error view

짧은 시간에 한사이트만을 확인하고, 다른 사이트로 이동하는 경우의 뷰는 낮은 가중치를 갖는다.

전제 2 는 자신의 의도한 내용과 웹사이트와 완전히 일치할 경우도 비슷한 행동 패턴을 보일 수 있지만, 잘못으로 인한 뷰일 경우의 확률이 높다. 여기서는 극히 적은 수의 뷰는 잘못된 뷰로 정의할 때, 정의 1 과 정의 2 에 의해서 극소와 극대의 뷰를 제외한 뷰만이 의미 있는 데이터가 될 가능성이 높게 된다.

따라서, 전체 사이트의 수는 T, 사용자의 뷰 또는 항목집합 v 의 카디날러티를 x 라 할 경우, 임계치 μ는 $0 < \mu \le T$ 이고, 정규 분포를 따른다고 할 때, 사용자의 뷰 x 는 μ에 가까울 때 가장 높은 가중치를 가지게 된다. 따라서, Carma 의 maxSupport 에 정규 분포의 가중치항을 추가함로써, 다음과 같이 다시 정리할 수 있다.

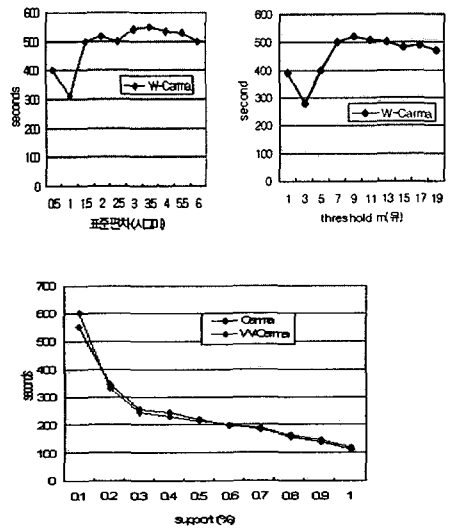
$$\text{maxSupport}(v) = \frac{\text{maxMissed}(v)+\text{count}(v)}{i} + \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2}$$

Phase I 에서 전지 단계에서 가중치가 높은 경우에는 lattice 에 빈발 항목집합으로 유지 되고, 정의 1 과 정의 2 와 같은 뷰는 빈발하지 않은 항목집합은 전지가 되므로, 최종적으로 lattice 의 수를 조절 할 수 있는 가중치항을 제시한 것이다.

5. 실험

본 실험의 OS 는 Solaris 7, UltraSPARC 167MHz 의 CPU, 64MB Ram 에서 수행하였다. 실험하기 위한 dataset 은 UCI dataset 으로 "Anonymous web data from www.microsoft.com"이다. 본 dataset 의 전체 트랜잭션은 3453 개이며, 어트리뷰트(attribute) 수는 294 개이다. 본 알고리즘에서는 연관 규칙에서 일반적인 형태인 비트열로 변경하여 연산하였다.

5%의 고정된 지지도에 대하여, μ값과 표준 편차인 σ를 임의 값으로 주어서 실험을 수행하였다. 먼저, μ의 평균인 3.95 전후의 값을 위주로 실험하였으며, 연관 규칙에서 가장 많은 영향을 미치는 요소인 지지도에 대한 실험 하였다.



6. 결론 및 향후 계획

기존의 Carma 를 웹 환경에 맞도록 가중치항을 제시함으로써, 빈발 항목집합을 줄이고, 이것은 곧 lattice 의 수를 줄임으로써, 계산 시간을 줄이는 결과를 실험을 통하여 알 수 있었다.

현재 사용한 dataset 은 공개된 dataset 으로 일반적으로 웹서버에 저장되어 있는 log 의 형식이 아닌, 각 사용자의 id 와 방문 사이트에 대한 정보만을 유지한 dataset 이다. 앞으로, 시간 및 다양한 정보를 가진 웹 데이터에 대하여, 더욱 신뢰성이 부여된 빈발 항목집합의 lattice 구성 및 계산 성능을 향상시키는 방법을 모색하고자 한다.

또한, Carma 의 규칙 생성 부분에서 확신도를 기준으로 조건부와 결과부에 충실한 규칙을 찾고자 한다.

7. 참고문헌

[1] R. Agrawal, T. Imielinski and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", Proc. of the 1993 ACM SIGMOD Conference,1993

[2] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", Proc. of the 20th VLDB Conference Santiago, Chile, 1994

[3] A. Savasere, E. Omiencinsky and S. Navathe, "An efficient algorithm for mining association rules in large databases", 21th VLDB Conference, pp. 432-444, Zurich, Swizerland, 1995

[4] R. Srikant and R. Agrawal, "Mining Generalized Association Rules", IBM Research Report

[5] J. S. Park, M. S. Chen, and P. S. Yu, "An effective hash-based algorithm for mining association rules", ACM SIGMOD, pp.175-186, 1995

[6] H. Toivonen, "Sampling Large Databases for Association Rules", Proc. of the 22nd VLDB Conf. India, 1996

[7] S. Brin, R. Motwani, J. D. Ulman and S. Tsur, "Dynamic Itemset Counting and Implication Rules for Market Basket Data", ACM SIGMOD '97 AZ, USA, 1997

[8] 박종수외, "연관 규칙 탐사와 그 응용", 한국정보과학회 SIGDB 춘계튜토리얼, 1998

[9] Charu C. Aggarwal and Philip S. Yu, "Online Generation of Association Rules", 1998. Proc. of 14th Int'l Conf. on Data Engineering, pp. 402 -411, 1998

[10] Christian Hidber, "Online Association Rule Mining", Technical Report TR-98-033, Int'l Computer Science Institute Berkeley, CA, 1998

[11] R. J. Bayardo Jr., R. Agrawal, D. Gunopulos, "Constraint-Based Rule Mining in Large, Dense Databases", Proc. of the 15th Int'l Conf. on Data Engineering, pp.188-197, 1999

[12] C. Lo and V. Ng, "Discovering web access orders with association rules", Proc. of IEEE Int'l Conf. on Systems, Man and Cybernetics, vol.4, pp.99 -104, 1999

[13] W. Wang, J. Yang and P. S. Yu, "Efficient Mining of Weighted Association Rules", Proc. of the 6th ACM SIGKDD Int'l Conf. on Knowledge discovery and data mining 2000, pp. 270 - 274, 2000