중요도를 고려한 가중치 그래프에서의 빈발 순회패턴 탐사

이성대° 박휴찬
한국해양대학교 컴퓨터공학과
{omega°, hcpark}@hhu.ac.kr

Discovery of Frequent Traversal Patterns on Weighted Graph with Priority

Seongdae Lee° Hyuchan Park
Dept. of Computer Engineering, Korea Maritime University, Korea

요 약

그래프를 사용하는 데이터 표현법은 직·간접적으로 실세계를 표현하는 다양한 데이터 모델 중에서 가장 일반화된 방법으로 알려져 있다. 기본적으로 그래프는 정점과 간선으로 구성되며, 정점과 간선은 그 중요도나 운영 목적에 따라 다양한 가중치가 부여될 수 있다. 특히, 이러한 그래프를 순회하는 트랜잭션들로부터 중요한 순회패턴을 탐사하는 것은 흥미로운 일이다. 본 논문에서는, 정점과 간선에 가중치가 있고 방향성을 가진 기반 그래프가 주어졌을 때, 그 그래프를 순회하는 트랜잭션들로부터 가중치를 고려하여 빈발 순회패턴을 탐사하는 방법을 제안한다. 또한, 이렇게 탐사한 결과에 가중치를 고려한 중요도를 평가하여 빈발 순회패턴들 간의 우선순위를 결정할 수 있도록 한다. 이 과정에서 발생할 수 있는 트랜잭션 노이즈는 기반 그래프의 간선 가중치의 평균과 표준편차를 이용하여 제거함으로써 보다 신뢰성 있는 빈발 순회패턴을 탐사할 수 있다. 제안한 논문은 웹 로그 마이닝 등 그래프를 이용하는 다양한 응용 분야에 적용할 수 있을 것이다.

1. 서 론

최근 다양한 분야에서 데이터 마이닝(data mining)에 관한 연구가 활발히 이루어지고 있다. 데이터 마이닝은 대량의 실제 데이터로부터 이전에 잘 알려지지 않았지만, 묵시적이고, 잠재적으로 유용한 정보를 추출하는 작업이라고 정의할 수 있다[1]. 기본적인 데이터 마이닝 알고리즘은 기존의 기계 학습 알고리즘에서 대규모의 데이터에 응용 가능하도록 변형되어 사용되는 경우가 많으며, 이 밖에 통계학을 토대로 개발된 알고리즘, 독자적으로 데이터베이스에서 패턴을 찾아내기 위해 개발된 알고리즘 등이 있다. 이러한 알고리즘들은 다양한 데이터 구조나 데이터베이스에서 사용되며, 최근에는 그래프(graph) 기반의 데이터 마이닝 연구가 활발히 진행되고 있다.

그래프를 사용하는 데이터 표현법은 직·간접적으로 실세계를 표현하는 많은 데이터 모델 중에서 가장 많이 사용되고 있는 일반화된 방법이다. 특히, 네트워크나 도로망의 설계 등에서 많이 사용하고 있다. 최근에는 그래프와 데이터 마이닝을 접목시킨 여러 방법들이 연구되고 있으며, 특히 대표적인 것이 웹 마이닝(web mining)이다.

웹 마이닝은 웹 구조 마이닝(web structure mining), 웹 내용 마이닝(web content mining), 웹 로그 마이닝(web log mining)으로 분류된다. 특히, 웹 로그 마이닝은 웹의 구조를 그래프로 표현하고, 웹 로그를 그래프를 순회하는 트랜잭션(transaction)으로 가공하여, 가장 빈발한 페이지 접근 경로(web page access path)를 찾는 문제이다. 따라서 이런 문제는 그래프에서 발생할 수 있는 다양한 순회 트랜잭션에서 최대 빈발경로(large path)를 찾는 문제로 치환하여 해결될 수 있다[2,3].

하지만 기존 연구에서는 웹 페이지들 간의 링크(link)를 방향 그래프로 표현하고, 그 그래프의 방향성(direction)만을 고려하여 최대 빈발경로(frequently large path)를 찾고 있다[1,2]. 이러한 방법들은 그래프의 정점(vertex)이나 간선(edge)에 부여될 수 있는 가중치(weight) 정보를 탐사 과정이나 결과에 반영하지 않는다. 따라서 본 논문에서는 가중치가 부여된 기반 그래프가 주어지고, 이러한 그래프를 순회하는 트랜잭션으로부터 마이닝 패턴을 탐사하는 방법을 제안한다. 특히, 기반 그래프의 간선 가중치를 이용하여 탐사 과정 중에 발생할 수 있는 트랜잭션의 노이즈(noise)를 빈발경로에서 제외하는 방법을 제시하고자 한다. 이 때 부여되는 기반 그래프의 정점 및 간선 가중치는 그래프의 응용 분야에 따라 다양한 형태로 주어질 수 있다. 예를 들면, 웹 로그 마이닝의 경우 문서의 정보량은 정점의 가중치로, 각 문서들 간의 이동 시간은 간선의 가중치로 부여될 수 있다.

2. 관련 연구

데이터 마이닝은 인공지능(artificial intelligence)의 한 분야인 기계 학습(machine learning)이나 데이터베이스로부터의 지식 발견(knowledge discovery in database)과 같이 대규모 데이터 내에 숨겨져 있는 고급 정보나 패턴을 추출해서 의사 결정, 예측, 예보 등에 응용하고자 하는 기술이다. 데이터 마이닝에서 획득할 수 있는 지식으로는 연관 규칙(association rules), 순차 패턴(sequential patterns), 분류 규칙(classification rules), 일반화/요약 규칙(generalization/summarization rules), 클러스터링(clustering) 등 여러 가지가 있다[3].

기존의 데이터 마이닝과 관련된 연구를 그래프를 기준으로 분류하면 트랜잭션만 주어진 경우와 트랜잭션과 기반 그래프가 주어진 경우로 나눌 수 있다. 첫 번째, 사용자의 트랜잭션만 주어진 경우는 연관 규칙 탐사나 순차 패턴 탐사 등이 있을 수 있다. 연관 규칙 탐사는 데이터베이스에 존재하는 항목들의 친화도나 패턴을 찾아내는 방법으로서 "빵을 구매하는 고객의 40%는 우유도 함께 구매한다."와 같이 트랜잭션에 있는 항목간의 연관성을 찾아내는 방법이다. 연관 규칙을 찾는 알고리즘 중에서 가장 많이 사용되고 있는 알고리즘은 Apriori 알고리즘이다[4]. 이 알고리즘은 두 단계로 구성된다. 우선, 각 아이템의 빈도수를 계산하여 최소 지지도(minimum support) 이상을 만족하는 항목들의 집합인 빈발 항목 집합(large itemsets)을 찾는다. 그 다음, 빈발 항목 집합으로부터 최소 신뢰도(minimum confidence) 이상을 만족하는 항목을 구한다. 이때 찾아진 항목을 후보 항목(candidate items)이라고 하며 결과 패턴에 포함될 가능성이 많은 집합이다. 이외에 연관 규칙 탐사에 사용되는 알고리즘으로는 DHP(Direct Hashing and Pruning) 알고리즘[4], Partitioning 알고리즘[5] 등이 있다. 순차 패턴은 연관 규칙에 시간적인 개념을 도입한 것이다.

두 번째는 기반 그래프가 주어지고, 기반 그래프의 간선을

따라서 순회하는 트랜잭션으로부터 최대 빈발경로를 탐사하는 것이다. 이는 순차 패턴과 유사하지만, 간선을 따라서 패턴이 존재한다는 점이 다르다. 예를 들면, 웹 마이닝에서 웹의 구조는 기반 그래프로, 웹 로그는 전처리(preprocess) 과정을 거친 후 트랜잭션으로 표현할 수 있다[2,3].

3. 가중치 그래프에서의 순회패턴 탐사

그림 1은 본 논문에서 제안하는 알고리즘을 구현한 시스템의 구성도이다. 시스템은 크게 전반부와 후반부의 두 부분으로 구성된다. 전반부는 본 알고리즘의 입력인 기반 그래프와 그 기반 그래프를 순회하는 트랜잭션들을 생성하는 부분이다. 후반부는 기반 그래프를 순회하는 트랜잭션들로부터 가중치를 고려하여 빈발경로를 탐사하고, 탐사한 빈발경로의 가중치를 기반으로 그들 간의 중요도를 결정하는 부분이다. 전반부는 3.1에서, 후반부는 3.2와 3.3에서 설명한다.

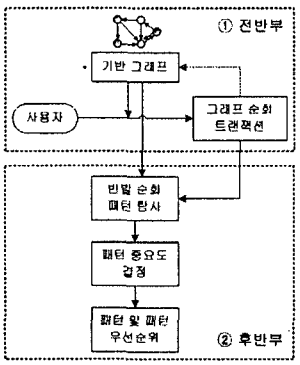

그림 1. 알고리즘 구성도

3.1 순회패턴 탐사를 위한 가중치 그래프

본 논문에서는 그래프를 순회하는 트랜잭션들로부터 최대 빈발경로를 탐사하기 위하여 가중치가 있는 그래프를 기반으로 한다. 먼저, 가중치를 포함한 그래프의 정의는 아래와 같다.

[정의 1] 그래프는 유한한 정점과 간선의 집합이다. 간선은 방향성과 가중치를 가지며, $<v_i, v_j, w_{ij}>$로 표현될 수 있다. (단, v_i, v_j는 그래프를 구성하고 있는 정점, w_{ij}는 v_i에서 v_j로 향하는 간선의 가중치, $v_i \neq v_j$)

정의 1은 가중치와 방향성이 존재하는 그래프의 정의이며, 정점과 간선에 부여된 가중치를 포함한 그래프는 인접 리스트(adjacent list)를 사용하여 구현하였다.

[정의 2] 기반 그래프를 순회하는 트랜잭션들은 $T = \{t_i | t_i = (<v_{i0}, v_{i1}, ..., v_{in}>, <tw_{i1}, tw_{i2}, ..., tw_{in}>)\}$로 표현한다. (단, t_i는 각 트랜잭션의 식별자, $<v_{i0}, v_{i1}, ..., v_{in}>$는 기반 그래프를 순회한 경로, $<tw_{i1}, tw_{i2}, ..., tw_{in}>$는 순회 중 발생하는 간선의 가중치)

정의 2는 기반 그래프를 순회하는 사용자의 트랜잭션들을 나타내며, 각 트랜잭션은 트랜잭션 식별자와 정점과 간선 정보를 포함한다. 트랜잭션의 간선 가중치 tw_{i1}은 i번째 트랜잭션의 0번째 정점에서 1번째 정점으로 가는 간선의 가중치이다. 본 논문에서는, 대부분의 자연적인 측정치가 정규분포(normal distribution) 형태를 지니므로, 트랜잭션의 간선 가중치가 정규분포를 갖도록 값을 할당하였다.

[정의 3] 기반 그래프의 간선 가중치는 정의 2의 트랜잭션들로부터 계산된 각 간선의 평균(average, μ)과 표준편차(standard deviation, σ)의 쌍으로 부여하며, $w_{ij} = (a_{ij}, s_{ij})$로 표현한다.

그림 2는 정의 1, 2, 3에 따라서 방향성과 가중치가 존재하는 기반 그래프이다. 기반 그래프를 완성하기 위한 전처리 단계로서 트랜잭션들에 있는 각 간선 가중치의 평균과 표준편차를 먼저 계산한다. 기반 그래프의 간선 가중치는 정의 3에 따라 평균과 표준편차 순으로, 정점의 가중치는 임의로 부여하였다.

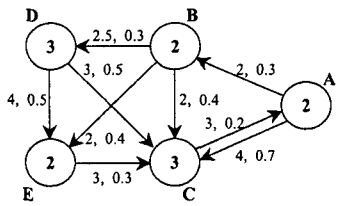

그림 2. 기반 그래프

[정의 4] 그래프에서 간선의 방향성을 고려한 연속적인 정점들의 집합을 경로(path)라고 하며, 경로에서 간선의 방향성을 고려한 부분 집합을 부분경로(subpath)라고 한다.

그림 2에서 임의의 경로 $P = <A\ B\ D\ E\ C>$가 존재할 경우, 경로 P에서 길이가 4가 되는 부분경로는 $<A\ B\ D\ E>$와 $<B\ D\ E\ C>$의 2개가 존재한다. 또한 P의 순서를 고려한 부분 집합 $<A\ B\ D>$, $<A\ B>$, $<B\ D>$ 등은 역시 부분경로이다.

3.2 가중치를 고려한 빈발 순회패턴 탐사

본 논문에서는 기반 그래프의 간선 가중치와 트랜잭션의 가중치를 고려하여 노이즈의 성격이 짙은 부분경로를 제거하면서 최대 빈발 순회패턴을 찾는 알고리즘을 제안한다. 노이즈 부분경로는 그것의 가중치가 평균에 비하여 현격히 크거나 작은 경로이다. 이것을 제거하기 위하여 정의 3의 기반 그래프의 간선의 평균과 표준편차 정보를 이용한다. 또한, 후보경로를 생성할 때 정점의 수가 1인 후보경로는 간선이 존재하지 않으므로 가중치는 고려하지 않고 빈발 횟수만을 고려한다.

```
begin
  C1 ← 정점의 수가 1인 초기 집합
  k = 1
  while (Ck ≠ ∅)
  begin
    for all subpath p ∈ T
    begin
      S = {s|s ∈ Ck, s is subpath of p}
      ∀s ∈ S s.count++
    end
    // 후보경로에서 노이즈 부분경로 제외
    if (k >= 2) Ck ← pruneCandidates(Ck, G)
    // 최소 지지도를 만족하는 후보경로는 빈발경로에 추가
    Lk = {s|s ∈ Ck, s.count ≥ minSupport}
    // 다음 단계 후보경로 생성
    Ck+1 ← genCandidates(Lk, G)
    k++
  end
end
```

그림 3. 빈발 순회패턴 탐사 알고리즘

그림 3은 가중치를 고려하여 빈발 순회패턴을 탐사하는 알고리즘이다. 각 후보경로가 트랜잭션에서 나타나는 빈발 횟수를 구한 후 이들 중에서 노이즈를 포함하는 후보경로의 빈발 횟수를 감산한다. pruneCandidates() 함수는 기반 그래프의 평균과 표준편차를 이용하여 각 간선의 신뢰구간(confidence interval)을 설정한 후, 설정된 신뢰구간 외부의 간선 가중치는 노이즈 경

로로 판별하여 빈발 횟수를 재조정한다. 신뢰구간은 추정하고자 하는 모수가 포함되어 있을 구간과 그 구간 안에 모수가 포함되어 있을 가능성을 의미하며, 본 논문에서는 신뢰구간을 95%로 설정하였다. 이때 임의의 간선의 가중치 x가 신뢰구간에 포함될 확률은 식 1과 같다.

$$P(\mu - 1.96\sigma \leq x \leq \mu + 1.96\sigma) = 0.95 \quad \cdots\cdots\cdots \text{식 1}$$

즉, 95%의 신뢰구간을 벗어나는 트랜잭션의 간선 가중치는 노이즈 경로라고 가정하였으며, 신뢰구간은 사용자의 의도에 따라 조정이 가능하다.

이후, 최소 지지도 이상을 만족하는 후보경로를 빈발경로에 포함시키고, 길이가 k인 빈발경로는 상호 조인(join)을 통하여 경로의 길이가 k+1이 되는 후보경로를 생성하기위해 다음 단계로 진행하며, 후보경로들의 집합이 공집합일 때까지 계속하여 반복한다.

3.3 순회패턴 중요도 결정

그림 3의 알고리즘의 수행 결과는 간선의 가중치를 고려한 최대 빈발경로이다. 이전 연구에서는 동일한 길이를 가지는 빈발경로의 중요도는 발생된 빈발 횟수에만 의존하였다. 본 논문에서는 빈발 횟수뿐만 아니라 패턴에 포함된 정점의 진입차수와 정점 및 간선의 가중치까지 포함하여 패턴의 중요도를 계산한다. 식 2는 본 논문에서 제안하는 빈발 순회패턴의 중요도를 계산하는 식이다.

$$PP(L_n) = \frac{Sup(L_n)}{MinSup} + \sum \frac{IN(V_i)}{T(E)} + \sum\sum \frac{W(< V_i, V_{i+1} >)}{W(V_i)} \quad \cdots \text{식 2}$$

식 2에서 $PP(L_n)$은 빈발경로의 중요도를 나타내며, $MinSup$는 최소 지지도, $Sup(L_n)$은 빈발경로 L_n의 발생 횟수, $T(E)$는 그래프의 전체 간선 수, $IN(V_i)$는 빈발경로 L_n 내의 각 정점 V_i의 진입차수, $W(<V_i, V_{i+1}>)$는 빈발경로 L_n의 정점 V_i에서 V_{i+1}로 가는 트랜잭션 내의 간선 가중치, $W(V_i)$는 정점 V_i의 가중치를 나타낸다. 즉, 빈발경로의 지지도와 빈발경로에 존재하는 정점의 진입차수와 가중치, 트랜잭션의 가중치를 모두 고려하여 패턴의 중요도를 결정하도록 하였다.

3.4 실험

본 논문에서 제안한 알고리즘을 적용한 예제는 그림 4와 같으며, 기반 그래프는 그림 2이다. 그림 4에서처럼 트랜잭션 데이터베이스로부터 먼저 정점의 수가 1이 되는 후보를 생성한다. 생성된 후보로부터 최소 지지도 이상을 만족하는 후보는 빈발경로에 포함된다. 그림 4에서 후보경로 <*A B*>가 생성되는 것을 보면, 그림 2의 기반 그래프에서 평균과 표준편차를 이용하여 식 1에 적용한 경로 <*A B*>의 신뢰구간은 1.412 ~ 2.588이다. 실제 발생한 빈도수는 5지만 7번째 트랜잭션에서의 가중치가 1.3으로 이 신뢰구간을 벗어나므로 1 감소하여 적용되는 빈도수는 4이다. 후보 생성에서 제외된 간선은 다음 단계의 후보 생성에서도 제외되므로 결과 최대 빈발경로는 그림 4의 L_3과 같다. L_3의 패턴 중요도는 식 2를 적용하여 계산하였으며, 결과에 따라 패턴 <*D E C*>가 <*C A B*>보다 중요함을 알 수 있다.

4. 결 론

데이터 마이닝은 데이터베이스 및 인공 지능의 연구 분야에서 각광받고 있는 분야이며, 최근에는 그래프와 데이터 마이닝을 접목시킨 연구가 활발하게 진행되고 있다.

본 논문에서는 가중치가 있는 기반 그래프에서 그래프를 순회하는 트랜잭션으로부터 마이닝 패턴을 탐사하기 위해 가중치를 비교하여 노이즈가 되는 경로를 제거하는 알고리즘과 알고리즘의 결과가 되는 빈발 순회패턴의 중요도를 결정할 수 있는 방법을 제안하였다. 노이즈 경로는 평균과 표준편차를 이용하여 신뢰구간을 설정한 후 신뢰구간을 벗어나는 경로를 의미한다. 이러한 방법을 통해 최종 단계에서 찾아진 최대 빈발 순회패턴은 보다 나은 신뢰성을 확보할 수 있다.

향후 연구 과제로는 가중치 그래프 기반의 클러스터링과 웹 로그 마이닝과 같이 그래프를 이용하는 응용분야에 접목시킬 수 있는 방법에 대한 연구가 필요하다.

Transaction Database

ID	Path	Weight
1	<A B C>	<2.2 2.0>
2	<B D E C A>	<3.1 4.4 3.5 3.3>
3	<C A B D>	<2.7 1.5 2.3>
4	<D C A>	<4.0 3.0>
5	<B C A>	<2.2 2.9>
6	<A B E C>	<2.1 3.4 3.2>
7	<A B D E C>	<1.3 2.3 4.4 3.2>
8	<B E C>	<2.3 3.4>
9	<B D C>	<2.7 3.1>
10	<D C A B E>	<3.8 2.8 2.2 1.9>

C_1

candidate	Support
<A>	8
<B>	9
<C>	10
<D>	6
<E>	5

L_1

Large	Support
<A>	8
<B>	9
<C>	10
<D>	6
<E>	5

C_2

candidate	Support
<A B>	4
<A C>	0
<B C>	2
<B D>	3
<B E>	2
<C A>	5
<D C>	2
<D E>	2
<E C>	4

L_2

Large	Support
<A B>	4
<B C>	2
<B D>	3
<B E>	2
<C A>	5
<D C>	2
<D E>	2
<E C>	4

C_3

candidate	Support
<A B C>	1
<A B D>	1
<A B E>	1
<B C A>	1
<B D C>	1
<B D E>	1
<B E C>	1
<C A B>	2
<D C A>	1
<D E C>	2
<E C A>	1

L_3

Large	Support	Pattern Priority	Rank
<C A B>	2	5.35	2
<D E C>	2	8.06	1

그림 4. 가중치를 고려한 빈발 순회패턴 탐사 예제
(최소 지지도 = 2)

참 고 문 헌

[1] U. M. Fayyad, G. P. Shapiro, P. Smyth, and R. Uthurusamy, *Advances in knowledge discovery and data mining*, The MIT Press, 1996.

[2] A. Nanopoulos and Y. Manolopoulos, "Mining Patterns from Graph Traversals", *Data and Knowledge Engineering(DKE)*, vol.37, no.3, pp.243-266, Jun. 2001.

[3] A. Nanopoulos and Y. Manolopoulos, "Finding Generalized Path Patterns for Web Log Data Mining", *Technical report*, Aristotle University, 2000.

[4] J. S. Park, M. S. Chen, and P. S. Yu, "An Effective Hash-Based Algorithm for Mining Association Rules", In Proc. of *ACM SIGMOD*, pp.175-186, San Jose, May, 1995.

[5] A. Savasere, E. Omiecinski and S. B. Navathe, "An Efficient Algorithm for Mining Association Rules in Large Databases", In Proc. of *21st Int. Conf. on VLDB*, pp.432-444, Zurich, Switzerland, Sep. 1995.