요 약
그래프를 사용하는 데이터 표현법은 직·간접적으로 설명을 표현하는 다양한 데이터 모델 중에서 가장 일반화된 방법으로 알려져 있다. 기존의 데이터 표현 방식은 단순한 관계를 표현하고 관계의 조함성은 그 중요도에 영향을 미치는 단계에 따라 다양한 가중치가 부여될 수 있다. 특히 이러한 그래프 표현의 트랜잭션들로부터 중요한 순환패턴을 탐색하는 과정은 흔히 가중치가 있는 단순 환경에서 문제가 발생한다. 본 논문에서는, 그래프의 가중치가 있고 방향성을 가진 그래프의 주어졌을 때, 그 그래프는 순환을 구현하는 트랜잭션들로부터 차원의 수를 고려하여 반복 순환패턴을 탐색하는 방법을 제안한다. 이 과정에서 발생할 수 있는 트랜잭션 노이즈는 기존 그래프의 간선 가중치의 평균과 표준편차를 이용하여 제한되며, 또한 신조생성 반복 순환 패턴을 탐색할 수 있다. 제안한 논문은 웹 로그 데이터 등 그래프를 이용하는 다양한 응용 분야에 적용할 수 있을 것이다.

1. 서론
최근 다양한 분야에서 데이터 마인딩(data mining)에 관한 연구가 활발히 이루어지고 있다. 데이터 마인딩은 대량의 실제 데이터를 분석하여 이와 같은 작업에 사용할 수 있는 데이터 모델을 생성하고, 특별히, 통계적으로 유의한 정보를 추출하는 작업에 중요하게 생각될 수 있다. 기존의 데이터 마인딩 알고리즘은 기존의 기계 학습 알고리즘에서 대규모의 데이터에 적용 가능하도록 변형되어 사용되는 경향이 많기 때문에, 이 밖에 통계학을 토대로 개발된 알고리즘 등, 특이적으로 데이터베이스에서 패턴을 찾아내기 위해 개발된 알고리즘 등이 있다. 이러한 알고리즘들이 다양한 분야에서 데이터베이스에서 사용되고 있으며, 최근에는 그래프(graph) 기반의 데이터 마인딩 연구가 활발히 진행되고 있다.

그래프를 사용하는 데이터 표현법은 직·간접적으로 설명을 표현하는 데이터 모델 중에서 가장 많이 사용되고 있는 일반화된 방법이다. 특히, 네트워크나 도로망의 설계 등에서 많이 사용하고 있다. 최근에는 그래프와 데이터 마인딩을 테마로 총 여 البعض의 연구가 있으며, 특히 대표적인 것이 웹 마인딩(web mining)이다.

웹 마인딩은 웹 구조 마인딩(web structure mining), 웹 내용 마인딩(web content mining), 웹 로그 마인딩(web log mining)으로 분류된다. 특히, 웹 로그 마인딩은 웹의 구조를 그래프로 표현하고, 웹 로그는 트랜잭션(transaction)으로 가공하여, 가장 일반적인 메이징 접근 경로(web page access path)를 찾아내는 문제로, 이러한 문제는 그래프에서 발생할 수 있는 다양한 순환 패턴에서 최대 빈도 경로를 찾는 문제로 치료하여 해결할 수 있다.[3]

하지만 기존 연구에서는 웹페이지와 웹 검색 엔진의 링크(link)를 바탕으로 그래프로 표현하고, 그 그래프의 방향성(direction)을 고려하여 최대 빈도 경로(frequently large path)를 찾아내는 방법들이다.[1,2] 이러한 방법들은 그래프의 정점(vertex)이나 간선(edge)에 부여될 수 있는 가중치(weight) 정보를 탐색하거나 결과에 영향을 미치지 않는다. 따라서 본 논문에서는 기존의 사례를 유의미한 그래프가 주어지면, 이러한 그래프로 표현하는 트랜잭션로부터 마인딩 패턴을 탐색하는 방법을 제안한다. 특히, 기존의 그래프의 간선 가중치를 이용하여 탐사 과정 중에서 발생할 수 있는 트랜잭션의 노이즈(noise)를 바탕으로 데이터를 구현하는 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 마인딩의 경우 동의 부여방식에는 첫 번째의 가중치로, 각 문서들 간의 이동 시간은 간선의 가중치로 부여될 수 있다.

2. 관련 연구
데이터 마인딩은 인공지능(artificial intelligence)의 한 분야인 기계 학습(machine learning)이나 데이터베이스에서의 지식 발굴(knowledge discovery in database)과 같은 구조적인 데이터 내에서 숨겨진 알고리즘과 패턴을 추출하여 의사결정, 예측, 예측 등에 응용하고자 하는 기술이다. 데이터 마인딩에서 획득할 수 있는 정보로는 연관 규칙(association rules), 순환 패턴(sequential patterns), 분류 규칙(classification rules), 일반화/요약 규칙(generalization/summarization rules), 클러스터링(clustering) 등 여러 가지가 있다.[3]

기존의 데이터 마인딩과 관련된 연구를 그래프 기반으로 분류하면 트랜잭션들 주어진 경우 트랜잭션 및 기반 그래프가 주어진 경우로 나눌 수 있다. 첫 번째, 사용자의 트랜잭션을 주어진 경우는 연관 규칙 탐사, 순환 패턴 탐사 등이 있을 수 있다. 연관 규칙 탐사는 데이터베이스에서 존재하는 항목들의 간 결이나 패턴을 찾아내는 방법으로서, 연관 규칙은 각 항목들에 아웃편성(support)의 정점을 찾아내는 방법이다. 연관 규칙을 찾는 알고리즘 중에서 가장 많이 사용되고 있는 알고리즘은 Apriori 알고리즘 이다.[4] 이 알고리즘은 두 단계에 구성된다. 우선, 각 아이템의 빈도수를 계산하여 최소 지지도(minimum support) 이상을 만족하는 항목들의 집합을 빈도 항목 집합(large itemsets)을 찾는다. 그 다음, 빈도 항목 집합으로부터 최소 신뢰도(minimum confidence) 이상을 만족하는 항목을 구한다. 이때 최첨 항목을 추출(candidate items)하고 하위 항목에 포함될 가능성이 있는 집합이다. 이하 연관 규칙 탐사에 사용되는 알고리즘으로는 DHP(Direct Hashing and Pruning) 알고리즘, Partitioning 알고리즘 등이 있다. 순환 패턴은 연관 규칙에서 서식관련 개념을 도입한 것이다.

두 번째는 기반 그래프가 주어지고, 기반 그래프의 간선을
32회 추계학술발표회 논문집 Vol. 32, No. 2(II)

따라서 순회하는 트랜잭션으로부터 최대 빈발경로를 탐색하는 것이다. 이는 순차 펼개와 유사하지만, 간선을 따라 펼개가 존재한다는 점에 달라진다. 예를 들어, 합 이상에서 경의 구조는 기반 그래프, 특히 그래프는 pred里斯(preprocess) 과정을 거친 후 트랜잭션으로 표현될 수 있다(2,3).

3. 가중치 그래프에서의 순회패턴 탐색

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

그림 1. 알고리즘 구성도

3.1 순회패턴 탐색을 위한 가중치 그래프

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

[정의 1] 가중 그래프는 유한한 정점과 간선의 집합이다. 간선은 방향성과 가중치가 있으며, $<v_i, v_j, c_{ij}>$로 표현할 수 있다. ($v_i, v_j$는 그래프를 구성하고 있는 정점, $c_{ij}$는 $v_i$에서 $v_j$로 향하는 간선의 가중치, $c_{ij} 
eq 0$)

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

[정의 2] 가중 그래프를 순회하는 트랜잭션들은 $T = \{t_i | t_i = <v_{i0}, v_{i1}, ..., v_{in}, c_{i1}, c_{i2}, ..., c_{in}>\}$로 표현한다. ($i$, $i'$는 각 트랜잭션의 식별자, $v_{i0}, v_{i1}, ..., v_{in}$는 기반 그래프를 순회한 경로, $c_{i1}, c_{i2}, ..., c_{in}$은 순환 중 발생하는 간선의 가중치)

정의 2는 가중 그래프를 순회하는 사용자의 트랜잭션들로 나타내며, 각 트랜잭션은 트랜잭션 식별자와 정점과 간선 정보를 포함한다. 트랜잭션의 간선 가중치 $c_{in}$을 1번에 트랜잭션의 0번 경로 정점에서 1번 경로 정점으로 가는 간선의 가중치이다. 본 논문에서는, 대부분의 자연적인 측정치가 정규분포(normal distribution) 형태를 지니므로, 트랜잭션의 간선 가중치가 정규 분포를 갖도록 랜덤을 했다.

[정의 3] 가중 그래프의 간선 가중치는 정의 2의 트랜잭션들로부터 계산된 각 간선의 평균(average, $\mu$)과 표준편차(standard deviation, $\sigma$)의 성으로 부여하며, $w_{ij} = (\mu_{ij}, \sigma_{ij})$로 표현한다.

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

그림 2. 가중 그래프

[정의 4] 그래프에서 간선의 방향성을 고려한 연속적인 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 $P$에서 임의의 경로가 $<A, B, D, E>$의 $<B, D, E>$의 2개가 존재한다. 또한 $P$의 순서를 고려한 부분 집합 $<A, B >, <A, D>, <A, B, D>, <A, B, D, E>$ 등은 역시 부분경로이다.

3.2 가중치를 고려한 빈발 순회패턴 탐색

본 논문에서는 기반 그래프의 간선 가중치와 트랜잭션의 가중치를 고려하여 노이즈의 영향이 적은 부분경로를 제거하면서 최대 빈발 순회패턴을 찾는 알고리즘을 제안한다. 노이즈 부분 경로는 그것의 가중치가 평균에 비해 현저히 크거나 적은 경로이다. 이들을 제거하기 위하여 정의 2의 기반 그래프의 간선의 평균과 표준편차 정보를 이용한다. 또한, 노이즈 경로는 선택할 때 정점의 수가 1인 노이즈 경로는 간선이 존재하지 않으므로 가중치는 고려하지 않고 빈발 횟수만을 고려한다.

begin
$G = $ 정점의 수가 1임 초기 집합
$k = 1$
while ($G_k \neq \emptyset$)
begin
for all subpath p ∈ T
begin
$S = \{s ∈ G_k, s$ is subpath of $p\}$
$v_s \in S$ s.count++
end
end
// 후보경로에서 노이즈 부분경로 제외
if ($k \geq 2$) $G_{k-1}$ pruneCandidates($G_k$)
// 최소 지지도를 만족하는 후보경로는 빈발경로에 추가
$L_k = \{s ∈ G_k, s.count \geq minSupport\}$
// 다음 단계 후보경로 생성
$G_{k+1}$ genCandidates($L_k$, $G_k$)
$k++$
end
그림 3. 빈발 순회패턴 탐색 알고리즘

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

본 논문에서 제안한 알고리즘을 적용한 예제는 그림 4와 같이, 기반 그룹으로 구성된 20가. 그림 4에서처럼 트랜잭션 데이터베이스로부터 인지된 모든 항목의 수가 10이 되는 후보를 생성한다. 생성된 후보들은 최소 지도도 이상을 만족하는 후보는 빈발도로서 포함한다. 그림 4에서 후보지정 <A B>가 생성되는 것을 보면, 그림의 기본 그룹에서 빈발도된 후보지정에 적용하여 그림 1의 신뢰도를 (1.12 - 5.588)로 개선된다. 실질적 방법은 5진한 7번의 트랜잭션에서의 가중치가 1.9으로 이 신뢰도를 벗어나므로 1 경계치는 빈도수 40이다. 후보 생성된 후보지정은 다음과 같은 항목의 부합도가 있다고 생각되어, 결과에 따라 패턴 <C D E>가 <C A B>보다 부합도를 갖고 있다. 

5. 결 문

데이터 마인딩은 데이터베이스 및 인공 지능의 연구 분야에서 활발하고 있는 분야이며, 최근에는 그래프와 데이터 마이닝을 핵심으로 연구가 활발하게 진행되고 있다. 본 논문에서는 단일 노드의 이론이 있는 기반에서 그래프를 사용하는 트랜잭션으로부터 마이닝 패턴을 탐색하기 위해 가중치를 비교하여 노이즈가 없는 경우를 제거하는 알고리즘과 알고리즘의 결과가 되는 빈발 순회패턴의 중요도를 결정할 수 있는 방법을 제안하였다. 노이즈 경우는 평균과 표준편차를 이용하여 신뢰도를 결정한 후 신뢰도를 평가하는 경우를 의미한다. 이러한 방법을 통해 최종 단계에서 찾아진 최대 빈발 순회패턴을 보다 나은 신뢰도를 확인할 수 있다.