순차패턴 마이닝에서 발생 간격 기반 가중치 부여 기법

장중혁, 신무종
대구대학교 컴퓨터 IT 공학부
e-mail : {jhchang, mjshin}@daegu.ac.kr

A Gap-based Weighting Approach in Mining Sequential Patterns

Joong-Hyuk Chang, Mu-Jong Shin
Dept. of Computer & Information Technology, Daegu University

요 약

순차패턴 마이닝에서 관심도가 큰 순차패턴을 얻기 위해서 구성요소의 단순 발생 순서뿐만 아니라 구성요소의 가중치를 추가로 고려할 수 있다. 본 논문에서는 순차패턴 마이닝에서 가중치 순차패턴을 탐색하기 위한 가중치 계산 기법으로 발생 간격에 기반한 순차패턴 가중치 부여 기법을 제안한다. 발생 간격 기반 가중치는 사전에 정의된 별도의 가중치 정보를 필요로 하지 않으며 순차정보를 구성하는 구성요소들의 발생 간격으로부터 구해진다. 즉, 순차패턴의 가중치를 구하는데 있어서 구성요소의 발생 순서와 더불어 이들의 발생 간격을 고려하며, 따라서 보다 관심도가 크고 유용한 순차패턴을 얻도록 지원한다.

1. 서론

일반적으로 순차패턴 마이닝[1,2]에서는 순차패턴이나 이를 구성하는 단위항목들의 중요성이 동일한 것으로 간주된다. 하지만 실제 응용 분야에서 이들 단위항목(즉, 단위항목들이 나타내는 실제 응용 분야에서의 단위 정보)들은 서로 다른 중요성을 가지며, 따라서 순차패턴 마이닝에서 이들의 차별화된 중요성을 고려하는 경우 보다 흥미도나 관심도가 큰 순차패턴을 얻을 수 있다. 이러한 상황을 고려하여 가중치 순차패턴 마이닝에 대한 연구들이 활발히 진행되어 왔다. 단순 지도를 기반으로 하는 일반적인 순차패턴 마이닝과는 달리 가중치 순차패턴 마이닝에서는 순차정보(sequence) 구성요소별로 차별화된 가중치를 고려하여 보다 관심도가 큰 순차패턴을 탐색한다[3,4].

한편, 순차정보 또는 순차패턴에 있어서는 이를 구성하는 단위항목들의 발생 순서뿐만 아니라 각각의 발생 시간이나 발생 간격 등도 중요한 정보를 제공한다. 예를 들어 두 개의 순차패턴이 서로 동일한 단위항목들로 구성되며 이들의 발생 순서가 서로 동일한 경우에도 하나의 순차패턴을 구성하는 단위항목들의 발생 간격이 다른 하나에 비해 짧은 경우 발생 간격이 짧은 순차패턴을 보다 중요한 순차패턴으로 간주할 수 있다.

이러한 가중치 순차패턴 마이닝의 효용성 및 순차패턴 탐색에서 구성요소의 발생 간격의 중요성 등을 바탕으로 본 논문에서는 순차패턴 마이닝에서 효율적으로 적용될 수 있는 발생 간격 기반 가중치 부여 기법을 제안하고자 한다. 더불어 해당 기법을 데이터 스트림에서 순차패턴 마이닝에 적용하여 발생 간격 기반 가중치 순차패턴을 구하고, 이의 효용성을 검증한다.

2. 발생 간격 기반 가중치

발생 간격 기반 가중치를 정의하기 위해서는 먼저 순차패턴(sequential pattern) 및 순차정보(sequence)를 명확히 정의할 필요가 있다. 하나의 **순차패턴**(sequential pattern) s 는 단위항목들이 순차적으로 정렬된 리스트(ordered list)로서 $<e_1 e_2 \dots e_l>$와 같이 나타낸다. 여기서 e_j ($1 \le j \le l$)는 단위항목을 나타내며, 분석 대상 데이터 집합에서 현재까지 사용된 단위항목들의 집합으로서 응용 분야에서 발생한 개별 정보를 의미한다. **순차정보**(sequence) S 는 하나 이상의 단위항목들이 정렬된 집합으로서 각 순차정보는 다른 순차정보와 구별되는 **식별자**(sequence identifier) SID 를 갖는다.

일반적으로 하나의 순차패턴에서 발생 간격이라 함은 해당 순차패턴을 구성하는 구성요소들의 순차정보 내에서의 발생 순서 차이를 의미하며, 본 논문에서는 인접한 구성요소들의 발생 순서 차이를 해당 순차패턴의 발생 간격으로 정의한다. 즉, 순차 데이터 스트림을 구성하는 하나의 순차정보 $S=<a_1 a_2 \dots a_m>$와 이에 출현한 하나의 순차패턴 $s=<b_1 b_2 \dots b_n>$ 사이에는 $b_1=a_{j1}$, $b_2=a_{j2}$, … , $b_n=a_{jn}$ 관계를 만족하면서 $1 \le j_1 < j_2 < \dots < j_n \le m$ 관계를 만족하는 정수 $j_1, j_2, \dots, j_n$ 이 존재한다. 이때, 해당 순차패턴 s 에서 인접한 두 개의 단위항목 b_p 및 b_q ($1 \le p \le n-1$, $q=p+1$) 사이의 **발생 간격** G_{pq}는 다음과 같이 정의되며 $n-1$ 개의 발생 간격이 정의된다.

$$G_{pq} = j_q - j_p$$

이와 같이 정의되는 발생 간격을 활용하여 순차패턴의 가중치를 정의하기 위해서는 크게 두 가지 과정을 필요로 한다. 하나는 서로 다른 크기로 구해지는 발생 간격에 대한 정규화 과정이며, 다른 하나는 하나의 순차패턴에 다수가 정의되는 발생 간격(또는 이의 정규화된 값)을 통합하여 해당 순차패턴을 대표할 수 있는 가중치를 정의하는 것이다. 먼저 0 이상의 다양한 정수 값으로 구해진 발생 간격들간의 공평한 비교를 위한 정규화의 과정으로 순차패턴에 존재하는 하나의 발생 간격에 대한 가중치를 [정의 1]에서와 같이 정의한다.

[정의 1. 발생 간격의 가중치] 발생 간격 기반 가중치 부여 기법에서 가중치를 설정하는 기준이 되는 *단위 발생 간격*을 $u(u>0)$라 하고, 단위 발생 간격 u 마다 감소되는 가중치의 양을 결정하는 *감쇠기본값*을 $b(0<b<1)$라 하며, 또한 단위항목들 사이에 발생 간격이 존재하더라도 가중치가 감소되지 않는(즉, 발생 간격이 없는 것으로 간주되는) 최대 발생 간격을 의미하는 *허용 발생 간격*을 G_a 라 하자. 이때, 하나의 순차패턴 $s=<b_1\ b_2\ \dots\ b_n>$의 인접한 두 항목 b_p 및 b_q $(1\le p\le n-1,\ q=p+1)$ 사이의 발생 간격 G_{pq} 에 대한 가중치 $w(G_{pq})$는 다음과 같이 정의된다.

$$w(G_{pq}) = b^{(G_{pq}-G_a)/u}$$

■

이어서 하나의 순차패턴에 존재하는 다수의 인접한 단위항목들의 조합으로부터 구해진 발생 간격 기반 가중치를 통합하여 해당 순차패턴에 대한 발생 간격 기반 가중치를 정의하며 [정의 2]에서와 같이 정의된다.

[정의 2. 순차패턴의 발생 간격 기반 가중치] 하나의 순차정보 S_k에 출현한 순차패턴 $s=<b_1\ b_2\ \dots\ b_n>$에 대해서 해당 순차패턴에 존재하는 다수의 인접한 항목들 간의 발생 간격 기반 가중치들 중에서 최소값을 해당 순차정보에서 순차패턴 s 의 발생 간격 기반 가중치 $W_k(s)$라 정의하며, 다음과 같이 구해진다. 이때, $n=0$ 인 경우는 해당 순차정보에서 s 가 출현하지 않은 경우를 의미한다.

$$W_k(s) = \begin{cases} \min_{1\le p\le n-1, q=p+1}(w(G_{pq})) & (n \ge 2) \\ 1 & (n=1) \\ 0 & (n=0) \end{cases}$$

■

순차패턴에 대한 발생 간격 기반 가중치가 정의되면 이를 바탕으로 순차 데이터 스트림에서 발생한 순차패턴의 발생 간격 기반 가중치 출현빈도 수 및 지지도를 정의할 수 있다. 단순 지지도 기반의 순차패턴 출현빈도 수 및 지지도에 대한 정의와 유사하나 해당 순차패턴의 출현빈도 수 계산시 각 순차정보에서 해당 순차패턴의 발생 간격 기반 가중치가 고려되어 [정의 3]에서와 같이 정의된다.

[정의 3. 발생 간격 기반 가중치 출현빈도 수 및 지지도] 순차 데이터 스트림 D_k 에서 발생한 하나의 순차패턴 s 에 대해서 발생 간격 기반 가중치 출현빈도 수 $gwC_k(s)$는 D_k 에 포함되는 k 개의 순차정보에서 해당 순차패턴의 발생간격 기반 가중치를 구하고 이를 전부 더한 값으로서 다음과 같이 정의된다.

$$gwC_k(s) = \sum_{S_i:(s\subseteq S_i)\wedge(S_i\in D_k)} W_i(s)$$

따라서 D_k 에서 해당 순차패턴 s 의 발생 간격 기반 가중치 지지도 $gwS_k(s)$는 다음과 같이 정의된다.

$$gwS_k(s) = \frac{gwC_k(s)}{|D|_k} = \frac{\sum_{S_i:(s\subseteq S_i)\wedge(S_i\in D_k)} W_i(s)}{|D|_k}$$

■

순차패턴의 발생 간격 기반 가중치 지지도가 정의되면, 이를 바탕으로 순차 데이터 스트림에서 가중치 순차패턴 마이닝을 위한 빈발 발생 간격 기반 가중치 순차패턴이 정의된다. 즉, 순차 데이터 스트림 D_k 에 대해서 최소 지지도 S_{min} 이 주어졌을 때, 해당 데이터 스트림에서 발생한 하나의 순차패턴 s 의 발생 간격 기반 가중치 지지도 $gwS_k(s)$가 S_{min} 보다 크거나 같은 값을 가질 때 해당 순차패턴을 빈발 발생 간격 기반 가중치 순차패턴이라 정의한다.

SID	순차정보
1	<a, b, a, d, e>
2	<c, d, f>
3	<a, b, c, d, f>
4	<c, d>

(그림 1) 예제 데이터 집합

이이서 예제 데이터 집합을 이용하여 앞서 기술한 발생 간격 기반 가중치 적용에 대해서 설명한다. 먼저 예제 데이터 집합은 그림 1 에서 보는 바와 같이 4 개의 순차정보로 구성되며, $u=1$, $b=0.9$ 및 $G_a=1$ 을 갖는 발생 간격 기반 가중치를 적용한다. 여기서 두 개의 순차패턴 $s_1=<a\ b>$와 $s_2=<a\ d>$에 대해서 발생 간격 기반 가중치 적용에 따른 지지도 변화를 살펴보자. 먼저, 각 순차정보에서 두 순차패턴의 발생 간격 기반 가중치를 구하면 다음과 같다.

$$W_1(s_1)=0.9^{(1-1)/1}=1.0,\ W_2(s_1)=0,$$

$$W_3(s_1)=0.9^{(1-1)/1}=1.0,\ W_4(s_1)=0$$

$$W_1(s_2)=0.9^{(3-1)/1}=0.81,\ W_2(s_2)=0,$$

$$W_3(s_2)= 0.9^{(3-1)/1} =0.81,\ W_4(s_2)=0$$

따라서 두 순차패턴의 발생 간격 기반 지지도 $gwS_k(s_1)$ 및 $gwS_k(s_2)$는 다음과 같이 구해진다.

$$gwS_k(s_1) = (1+0+1+0)/4 = 0.5$$

$$gwS_k(s_2) = (0.81+0+0.81+0)/4 = 0.405$$

두 순차패턴의 단순 지지도는 0.5 로 서로 동일한 값을 갖는 반면 발생 간격 기반 가중치 지지도에서는 0.5 와 0.405 로 차별화된 값을 갖는다. 만약 해당 데이터 집합에 대한 빈발 순차패턴 탐색에서 최소 지지도가 0.5 로 설정되었다면, 단순 지지도를 적용하는 경우 모두가 빈발 순차패턴으로 탐색되나 발생 간격 기반 가중치를 적용하는 경우 s_1 은 빈발 순차패턴이 되지만 s_2 는 빈발 순차패턴이 되지 못한다. 그 이유는 그림 1 에서 보는 바와 같이 s_1=<*a b*>를 구성하는 두 단위항목들 사이의 발생 간격은 작은 반면에 s_2=<*a d*>를 구성하는 두 단위항목들 사이의 발생 간격은 상대적으로 크기 때문이다. 즉, 순차패턴을 구성하는 단위항목들의 발생 간격에 따라 해당 순차패턴의 중요성이 차별화되어 서로 다른 지지도 값을 갖는다.

<표 1> 결과 집합의 길이별 순차패턴 개수 [G_a=5, b=0.9]

u	L_1	L_2	L_3	L_4	L_5	L_6 이상	계
미적용	735	1177	20	6	1	0	1939
5	735	695	20	6	1	0	1457
10	735	843	20	6	1	0	1605
20	735	916	20	6	1	0	1678

<표 2> 결과 패턴의 지지도 변화 [u=20, G_a=10, b=0.9]

패턴	단순 지지도	가중치 적용 지지도	지지도 차이
<186,308>	0.00526	0.00482	0.00044
<223,991>	0.00530	0.00489	0.00041
<970,862>	0.00518	0.00481	0.00037
<378,991>	0.00508	0.00472	0.00036
<446,228>	0.00511	0.00476	0.00035
<223,542>	0.00514	0.00479	0.00035
<403,780>	0.00523	0.00488	0.00035
<72,42>	0.00530	0.00497	0.00033
<479,348>	0.00501	0.00468	0.00033
<657,533>	0.00527	0.00494	0.00033

3. 실험 결과 고찰

본 절에서는 발생 간격 기반 가중치 부여 기법의 유용성을 검증하기 위한 실험 및 분석 결과를 기술한다. 실험에 사용된 데이터 집합은 *SD_IBM* 데이터 집합으로서 순차패턴 마이닝 방법의 효율성 검증을 위한 실험용 데이터 집합 생성에 널리 활용되는 IBM 데이터 생성기(IBM data generator)[6]를 이용하여 생성되었으며, 1,000 개의 단위항목으로부터 생성된 100,000 개의 순차정보로 구성된다. 아래의 실험들에서 S_{min} 값은 0.005 로 설정되었다.

표 1 은 발생 간격 기반 가중치 적용에 따른 가중치 순차패턴 마이닝 결과 집합에서 길이별 순차패턴의 상세 변화 내용을 제시하고 있으며, *SD_IBM* 데이터 집합을 구성하는 100,000 개의 순차정보가 모두 처리된 후의 결과를 분석하였다. 본 실험에서 감쇠기본값 b 및 허용 발생 간격 G_a 는 각각 0.9 및 5 로 설정되었으며, 단위 발생 간격 u 의 변화에 따른 순차패턴 수의 변화를 분석하였다. 표에서 L_k(k=1,2,…)는 k 개의 단위항목으로 구성된 빈발 순차패턴을 의미한다. 하나의 단위항목으로 구성되는 L_1 의 경우는 순차패턴 내에 발생 간격이 존재하지 않으므로 발생 간격 기반 가중치 적용 여부에 무관하게 얻어지는 결과가 동일하다. L_3 이상의 경우에도 발생 간격 기반 가중치를 적용하는 경우에도 미적용시와 동일한 결과를 보인다. 반면 L_2 의 경우는 발생 간격 기반 가중치에 크게 영향을 받음을 알 수 있다. 실험에 사용된 *SD_IBM* 데이터 집합은 실제 응용 분야 데이터 집합이 아니라 인위적으로 생성된 데이터 집합으로서 각 순차패턴의 지지도 및 발생 간격 등이 적절히 분포되도록 생성된다. 또한 일반적으로 큰 발생 간격을 갖는 순차패턴이 자주 발생되어 빈발 순차패턴이 되는 경우는 거의 존재하지 않는다. 따라서 길이가 긴 순차패턴에 있어서는 발생 간격 기반 가중치 적용 여부에 무관하게 서로 동일한 결과를 보며, 상대적으로 짧은 길이의 순차 패턴인 L_2의 경우에만 큰 차이를 나타낸다.

발생 간격 기반 가중치 적용에 따른 순차패턴의 지지도 변화를 보다 명확히 파악하기 위하여 표 2 에서와 같이 동일한 순차패턴에 대해서 단순 지지도와 발생 간격 기반 가중치 지지도를 비교하였다. 본 실험에서 단위 발생 간격 u, 감쇠기본값 b 및 허용 발생 간격 G_a 는 각각 20, 0.9 및 10 으로 설정되었다. 표 2 에서 제시된 순차패턴은 상대적으로 발생 간격이 큰 순차패턴으로서 발생 간격 기반 가중치 적용시 지지도가 감쇠되는 것들이다. 표에서 제시된 것보다 많은 수의 순차패턴들이 발생 간격 기반 가중치 지지도가 감소되었으며, 표에서는 지지도 감소폭이 큰 10 개를 제시하였다. 표에서 알 수 있듯이 각 순차패턴들은 단순 지지도 기반으로 순차패턴 마이닝을 수행하는 경우 S_{min}=0.005 보다 큰 지지도를 가지므로 전부 빈발 순차패턴으로 탐색된다. 하지만 발생 간격 기반 가중치 순차패턴 탐색에서는 해당 최소 지지도보다 작은 지지도를 갖게 되므로 빈발 순차패턴이 될 수 없다. 해당 순차패턴들은 상대적으로 큰 발생 간격을 갖는 것들로서, 발생 간격 기반 가중치 적용시 가중치가 감쇠되기 때문이다.

4. 결론

분석 대상이 되는 데이터 스트림이나 데이터 집합에 내재된 지식이나 정보를 찾는데 있어서 구성요소의 발생 순서까지 고려하는 순차패턴 마이닝은 이전의 한정적인 데이터 집합뿐만 아니라 근래의 변화된

컴퓨터 응용 환경에서 생성되는 데이터 스트림의 형태의 정보를 분석하는데도 효율적으로 적용되고 있다. 하지만 기존의 단순 지지도 기반 순차패턴 마이닝은 관심도나 중요도가 작은 순차패턴들까지 포함된 마이닝 결과 집합을 구한다. 이를 보완하기 위한 순차패턴 마이닝 관련 연구들 중에서 주목받는 연구 분야 중의 하나인 가중치 순차패턴 마이닝은 구성요소의 발생 순서뿐만 아니라 구성요소나 순차정보의 가중치를 산정할 수 있는 부가적인 정보를 활용하여 보다 관심도가 큰 순차패턴을 마이닝 결과로 제공함으로써 탐색된 순차패턴의 활용도를 높일 수 있다.

이러한 흐름을 고려하여 본 논문에서는 발생 간격 기반 가중치 부여 기법을 제안하였다. 하나의 순차패턴에 있어서 발생 간격은 분석 대상이 되는 데이터 스트림의 특성이나 사용자의 관심도에 따라 다양한 기준으로 정의될 수 있다. 본 논문에서 제안한 발생 간격 기반 가중치 부여 기법에서는 하나의 순차패턴을 구성하는 여러 단위항목들에 대해서 인접한 두 단위항목의 순차정보에서의 발생 순서 차이를 발생 간격으로 정의하고, 이를 활용하여 발생 간격 기반 가중치를 부여한다. 먼저 가중치 함수를 이용하여 하나의 순차패턴에 존재하는 각 발생 간격에 대한 가중치를 구하고 이로부터 해당 순차패턴의 가중치를 구한다. 이어서 분석 대상이 되는 하나의 데이터 스트림에서 각 순차패턴의 가중치를 종합하여 해당 순차패턴의 지지도를 구하고, 이로부터 해당 순차패턴이 빈발 순차패턴이지를 판단한다.

참고문헌

[3] S. Lo, “Binary Prediction based on Weighted Sequential mining method,” *Proc. of the 2005 Int’l Conf. on Web Intelligence*, pp. 755-761, 2005.

[4] U. Yun, “A New Framework for Detecting Weighted Sequential Patterns in Large Sequence Databases,” *Knowledge-Based Systems*, 21(2), pp. 110-122, 2008.

[1] M.-Y. Lin, S.-C. Hsueh, and C.-W. Chang, “Fast Discovery of Sequential Patterns in Large Databases using Effective Time-Indexing,” *Information Sciences*, 178(22), pp. 4228-4245, 2008.

[2] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.-C.- Hsu, “Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach,” *IEEE Transactions on Knowledge and Data Engineering*, 16(11), pp. 1424-1440, 2004.

[5] J.H. Chang and W.S. Lee, “Efficient Mining Method for Retrieving Sequential Patterns over Online Data Streams,” Journal of Information Science, Vo. 31, pp. 420-432, 2005.

[6] R. Agrawal and R. Srikant, “Mining Sequential Patterns,” in Proc. of the 1995 Int’l Conf. on Data Engineering, pp. 3-14, 1995.