Journal of the Korea Industrial Information Systems Research Vol. 20 No.2, Apr. 2015 :1-9
http://dx.doi.org./10.9723/jksiis.2015.20.2.001 ISSN:1229-3741

데이터 스트림 마이닝에서 양방향 감쇠 기법을 활용한 고관심 정보 탐색

(Mining highly attention itemsets using a two-way decay mechanism in data stream mining)

장 중 혁[1)*]
(Joong-Hyuk Chang)

요 약 데이터 스트림 마이닝에서 대부분의 정보 중요성 차별화 기법들은 오래된 정보에 비해 최근에 발생한 정보에 보다 큰 가중치를 부여한다. 하지만, 오래 전에 발생한 정보 중에도 매우 중요한 의미를 갖는 정보들이 존재하기도 한다. 예를 들어, 도소매 상점에서 과거에는 단골 고객이었으나 일정 기간 동안 방문하지 않은 경우, 해당 고객의 구매 기록 등이 포함된 오래된 정보들은 집중 마케팅을 통한 판매실적 증대에 매우 중요한 자료가 될 수 있다. 본 논문에서는 하나의 데이터 스트림에서 최근에는 자주 발생되지 않으나 과거에 빈번히 발생했던 것으로서 관심도가 큰 항목집합을 의미하는 고관심 정보 HAI(Highly Attention Itemsets)를 정의하고, 이를 효율적으로 탐색하기 위한 양방향 감쇠 기법 및 데이터 스트림 마이닝 기법을 제안한다.

핵심주제어 : 고관심 정보, 양방향 감쇠 기법, 데이터 스트림, 데이터 스트림 마이닝, 정보 중요성 차별화

Abstract In most techniques of information differentiating for data stream mining, they give larger weight to the information generated in recent compared to the old information. However, there can be important one among the old information. For example, in case of a person was a regular customer in a retail store but has not come to the store in recent, old information with the shopping record of the person can be importantly used in a target marketing for increasing sales. In this paper, highly attention itemsets(HAI) are defined, which mean the itemsets generated in the past frequently but not generated in recent. In addition, a twao-way decay mechanism and a data stream mining method for finding HAI are proposed.

Key Words : Highly attention itemsets, Two-way decay mechanism, Data streams, Data stream mining, Information differentiation

* Corresponding Author
이 논문은 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2012R1A1B4000651)
Manuscript received December 12, 2014 / Revised February 05, 2015 / Accepted February 25, 2015
1) 대구대학교 컴퓨터IT공학부, 교신저자(jhchang@daegu.ac.kr)

1. 서 론

근래 대부분의 컴퓨터 응용 분야에서는 데이터 스트림 형태로 정보를 발생시키고 있으며, 데이터 스트림은 구성요소가 지속적으로 발생되는 무한 집합으로서 시간 흐름에 따른 가변성이 매우 큰 특징을 갖는다 [1,2]. 하나의 데이터 스트림에서는 이를 구성하는 구성요소 및 이의 특성이 시간흐름에 무관하게 유사한 형태를 유지하기도 하지만 시간 흐름에 따라 구성요소 또는 특성이 변화되기도 한다. 하나의 데이터 스트림에서 과거에는 자주 발생되었던 정보가 근래에는 발생 빈도가 현저히 줄어드는 경우도 이러한 변화의 대표적인 예라 할 수 있다.

실제 응용 서비스에서 발생되는 데이터 스트림의 가변성 등을 고려하여 데이터 스트림 내재된 다양한 정보를 탐색하기 위한 연구들이 활발히 제시되어 왔다[3,4]. 또한, 시간 변화에 따른 데이터 스트림의 변화를 마이닝 결과에 효과적으로 반영할 수 있는 데이터 스트림 마이닝 기법들도 활발히 제안되어 왔다[2,5,6,8]. 특히, 하나의 데이터 스트림에 포함되는 구성요소의 중요성을 발생 시점을 고려하여 차별화 함으로써 해당 데이터 스트림의 변화를 효과적으로 탐지하기 위한 연구들도 활발히 연구되고 있으며, 대표적인 것으로 감쇠율 기법[2,5]과 이동윈도우 기법[6,7]이 제안되었다. 해당 기법들에서는 일반적으로 근래에 발생한 정보들은 큰 중요성을 갖는 것으로 간주되는 반면 과거에 발생한 정보들은 매우 낮은 중요성을 갖는 것으로 간주되거나 중요성이 무시되기도 한다. 하지만, 실제 응용 분야에서는 비록 오래 전 과거에 발생된 정보라 할지라도 중요한 의미를 갖는 관심도가 큰 정보들이 존재하기도 한다. ‘Oldies but goodies’라는 말처럼 희소성이나 역사성 측면에서 중요성을 인정받는 경우도 있으며, 현재 또는 최근의 해당 응용 분야 특성을 보다 효과적으로 분석하기 위해 과거의 정보들이 중요한 의미를 갖는 경우도 있다.

본 논문에서는 근래 연구 관심이 증대되고 있는 고유용 관심 정보[8,9]의 하나로서 데이터 스트림에서 최근에는 자주 발생되지 않으나 과거에는 빈번히 발생했던 것으로서 관심도가 큰 항목집합(즉, 과거 단골에 대한 집중형 마케팅 분야 등에서 활용도가 큰 항목집합)을 HAI로 정의하고, 이의 효율적 탐색을 위한 양방향 감쇠 기법(Two-way decay mechanism)을 제시한다. 이어서 해당 기법을 적용하여 데이터 스트림에서 HAI를 효과적으로 탐색하는 데이터 스트림 마이닝 기법을 제안하며, 이는 기존의 다른 데이터 스트림 마이닝 기법들과 마찬가지로 마이닝 수행 과정에서의 메모리 사용량 및 트랜잭션 처리 시간 등의 측면에서 효율적이다. 끝으로, 논문에서 제안되는 HAI의 유용성 및 탐색 기법의 효율성은 일련의 실험을 통해 검증된다.

본 논문의 구성은 다음과 같다. 2장에서는 간략한 관련 연구를 포함한 본 논문에서 다루게 되는 문제를 정의하며, 데이터 스트림에서 구성요소의 가중치를 차별화하기 위한 양방향 감쇠 기법은 3장에서 제시된다. 4장에서는 HAI에 대한 정의 및 이를 효율적으로 참색하기 위한 마이닝 기법을 제시하고, 해당 기법의 유용성을 검증하기 위한 다양한 실험 결과는 5장에서 기술한다. 끝으로 6장에서 논문의 결론을 맺는다.

2. 기본 정리

고관심 정보를 탐색하기 위한 데이터 스트림은 구성요소가 지속적으로 생성되는 무한 데이터 집합으로 간주할 수 있으며, [2]에서의 정의에 따라 다음과 같이 정의된다. 먼저, I는 하나의 응용 서비스에서 단위 정보를 표시 하는데 사용되는 단위항목(item)들의 집합을 나타낸다. 항목집합(itemset) e는 단위항목들의 집합으로서 $e \in (2^I - \{\varnothing\})$를 만족하며, 2^I는 I의 멱집합을 의미한다. 하나의 항목집합 e에 대해서 해당 항목집합을 구성하는 단위항목의 수를 해당 항목집합의 길이라 지칭하고 $|e|$로 나타내며, m개의 단위항목으로 구성되는 항목집합을 m-항목집합이라 한다. 또한 논문에서는 항목집합 $\{a,b,c\}$를 간략히 abc로 표시한다. 트랜잭션(transaction)은 항목집합들로 구성되며, 서로 다른 트랜잭션을 구분하는 식별자 TID를 갖는다. 이때, k 번째 생성된 트랜잭션을 T_k로 나타낸다. 하나의 새로운 트랜잭션 T_k가 생성되었을 때, 현재 데이터 스트림 D_k는 현재까지 생성된 모든 트랜잭션들로 구성된다. 즉, $D_k = \langle T_1, T_2, \cdots, T_k \rangle$ 로 표현되며, 해당 데이터 스트림에 포함된 트랜잭션의 총 개수는 $|D|_k$로 나타낸다.

일반적으로 데이터 스트림 D_k에 새로운 트랜잭션 T_k가 생성되었을 때, 해당 데이터 스트림에서 발생된 하나의 항목집합 e의 출현빈도 수 $C_k(e)$는 D_k에 포

함되는 트랜잭션 중 해당 항목집합 e를 포함하고 있는 트랜잭션 개수를 의미한다. 또한, 하나의 항목집합 e의 지지도를 나타내는 $Supp_k(e)$는 D_k에 포함되는 트랜잭션의 총 개수 대비 해당 항목집합 e를 포함하고 있는 트랜잭션 개수의 비율을 의미하며 $C_k(e)/|D|_k$로 구해진다. 하나의 데이터 스트림 D_k에 대해서 지지도 임계값(0보다 크고 1보다 작거나 같은 범위)이 설정되었을 때, D_k에서 발생된 항목집합 e는 지지도 값이 해당 지지도 임계값 보다 크거나 같은 경우 빈발 항목집합이라 지칭한다. 이와 유사한 정의에 의하여, 분석 대상이 되는 하나의 데이터 스트림과 지지도 임계값이 주어졌을 때 고관심 정보 ***HAI*** (Highly Attention Itemsets)라 함은 해당 데이터 스트림에서 과거에는 빈발 항목집합이었으나 근래에는 발생빈도가 적은 항목집합을 의미하며, 데이터 스트림에서 HAI 탐색이라 함은 분석 대상 데이터 스트림에 존재하는 모든 HAI를 탐색하는 작업을 의미한다.

시간 흐름에 따른 가변성이 큰 데이터 스트림의 특성을 고려하여 데이터 스트림 구성 요소의 중요성을 발생 시간 축을 기준으로 차별화하기 위한 다양한 기법들[2,5,6,7]이 연구되어 왔으며, 이들 기법들은 빈발 항목집합이나 순차패턴 등을 탐색하기 위한 데이터 스트림 마이닝 과정에 적용되어 보다 관심도가 큰 마이닝 결과를 얻는데 활용되어 왔다. 해당 방법들 중에서 대표적인 것으로 이동 윈도우 기법[6,7]과 감쇠 기반 기법[2,5]이 고려될 수 있다.

일반적으로 이동 윈도우 기법 및 감쇠 기반 기법은 최근에 발생된 정보 혹은 최근에 가까운 시점에 발생한 정보의 중요성을 높게 간주하고 이외의 정보는 무효하거나 중요성이 낮은 것으로 간주한다. 따라서, 해당 기법들은 과거 일정 시점에 관심도가 큰 것으로 간주되었던 정보들을 탐색하거나, 특히 집중 마케팅 등을 위해 중요한 정보로 활용될 수 있는 과거 발생 정보들을 효과적으로 탐색하는 데에는 한계가 있다.

3. 양방향 감쇠 기법

본 절에서는 먼저 하나의 데이터 스트림에서 구성요소의 중요성을 시간 축에 따라 차별화 하기 위한 방법으로 기존의 감쇠 기반 기법에서 제안된 감쇠율(본 논문에서는 의미를 명확히 하기 위해서 이를 정방향 감쇠율이라 지칭함)에 대해 간략히 기술한다. 이어서 하나의 데이터 스트림의 구성요소들 중 과거에 발생된 것들이 근래에 발생된 것보다 높은 가중치를 갖는 역방향 감쇠율을 제시한다. 끝으로 이들 두 종류의 감쇠율을 결합하여 새로운 관심 항목집합인 HAI를 효율적으로 탐색하기 위한 데이터 스트림 마이닝 기법을 제시한다.

3.1 감쇠율

감쇠율(decay rate) $d \in (0,1)$는 단위 시간 동안의 가중치 감소 정도를 의미한다. 감쇠율 기법에서는 단순히 사용자가 직접 감쇠율을 정의할 수도 있다. 하지만 이러한 경우 응용 분야의 실제 상황에 맞지 않는 지난치 가중치 감소 현상이 발생될 수 있다. 따라서, 가중치 감쇠 정도를 실제 응용 분야에 부합되도록 유연하게 조절하기 위해서 보다 세밀한 형태로 감쇠율을 정의하고 있다[2]. 즉, 감쇠율을 두 개의 독립적인 변수인 감쇠-기본값 $b(b>1)$와 감쇠-기본주기 $h(h \geq 1)$를 이용하여 나타내어 $d=b^{-(1/h)}$으로 정의한다. 감쇠-기본값 b는 감쇠율 적용을 위한 기본 단위 시간 동안 감소되는 가중치 양을 결정하며, 감쇠-기본주기 h는 감쇠율 적용에 따라 가중치 값이 $1/b$가 되기 위해 필요한 단위 시간의 수를 나타낸다. 이러한 두 변수를 조절함으로써 다양한 형태의 감쇠율을 정의할 수 있다.

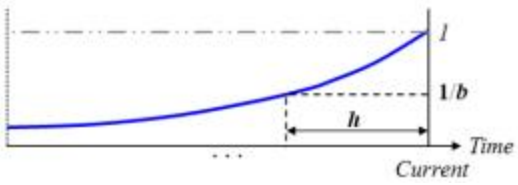


Fig. 1 The changed weight of information in a direct decay mechanism

3.2 정방향 감쇠 기법

기존의 데이터 스트림 마이닝에서 활용된 정방향 감쇠 기법(direct decay mechanism) 에서는 구성요소가 지속적으로 생성되는 하나의 데이터 스트림에서 현재 시점에 발생된 구성요소가 가장 큰 가중치

를 갖고 시간의 흐름에 따라 해당 가중치가 감쇠된다[2]. 감쇠율 $d=b^{-(1/h)}$를 갖는 정방향 감쇠 기법에 있어서 하나의 데이터 스트림에서 현재 시점에 새롭게 발생되는 구성요소의 가중치는 일반적으로 1로 부여되며, 시간 흐름에 따라 n 번의 단위 시간이 경과된 후 해당 구성요소의 가중치는 $1\times d^n$으로 감소된다. <Fig. 1>은 감쇠율 $d=b^{-(1/h)}$를 갖는 정방향 감쇠 기법에서 시간 흐름에 따른 정보의 가중치 변화를 보여준다.

하나의 데이터 스트림 D_k에 대해서 감쇠율 $d=b^{-(1/h)}$를 갖는 정방향 감쇠 기법이 적용될 때 해당 데이터 스트림에 포함된 트랜잭션의 총 개수 $|D|_k^D$는 다음과 같이 구해진다.

$$|D|_k^D = \begin{cases} 1 & \text{if } k=1 \\ |D|_{k-1}^D \times d+1 & \text{if } k\geq 2 \end{cases} = \frac{1-d^k}{1-d}$$

이와 유사한 방법으로 해당 데이터 스트림에서 발생된 하나의 항목집합 e의 출현빈도 수 $C_k^D(e)$는 다음과 같이 구해진다.

$$C_k^D(e) = \begin{cases} W_k^D(e) & \text{if } k=1 \\ C_{k-1}^D(e)\times d+W_k(e) & \text{if } k\geq 2 \end{cases}$$

where $W_k^D(e)=\begin{cases}1 & \text{if } e\subseteq T_k \\ 0 & \text{otherwise}\end{cases}$

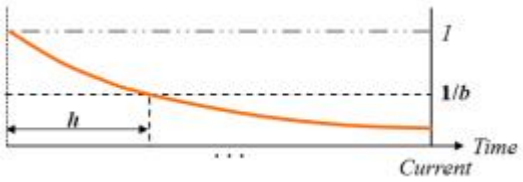


Fig. 2 The changed weight of information in a reverse decay mechanism

3.3 역방향 감쇠 기법

역방향 감쇠 기법(reverse decay mechanism) 에서는 구성요소가 지속적으로 생성되는 하나의 데이터 스트림에서 현재 시점에 발생된 구성요소가 가장 작은 가중치를 갖는 반면 해당 데이터 스트림의 구성요소 중 가장 오래된 것이 가장 큰 가중치를 갖는다. 감쇠율 $d=b^{-(1/h)}$를 갖는 역방향 감쇠 기법에서 하나의 데이터 스트림의 각 시점에서 발생되는 구성요소는 발생 시점에 따라 가중치가 결정되며, 해당 구성요소에 부여된 가중치는 시간이 흐른 뒤에도 변경되지 않고 동일하게 유지된다. 반면 새롭게 발생되는 구성요소에 부여되는 가중치는 감쇠율 d에 의해 감쇠된다. 예를 들어, 현재 시점에서 발생된 구성요소의 가중치를 w라 할 때 n 번의 단위 시간이 경과된 후 해당 구성요소의 가중치는 w로 유지되지만 새롭게 발생되는 구성요소의 가중치는 $w\times d^n$으로 부여된다. 일반적으로 하나의 데이터 스트림에서 맨 처음 발생된 구성요소의 가중치는 1로 부여되며, 감쇠율 $d=b^{-(1/h)}$를 갖는 역방향 감쇠 기법에서 시간 흐름에 따른 정보의 가중치 변화를 보여준다.

감쇠율 $d=b^{-(1/h)}$를 갖는 역방향 감쇠 기법에서 하나의 데이터 스트림 D_k에서 첫 번째 트랜잭션 T_1이 발생되었을 때 해당 데이터 스트림에 속하는 트랜잭션의 총 개수는 명백히 1이 된다. 이어서 새로운 트랜잭션 $T_k(k\geq 2)$가 계속적으로 발생되었을 때 해당 데이터 스트림에 속하는 트랜잭션의 총 개수 $|D|_k^R$는 다음과 같이 구해진다.

$$|D|_k^R = \begin{cases} 1 & \text{if } k=1 \\ |D|_{k-1}^R + d^{k-1} & \text{if } k\geq 2 \end{cases}$$

이와 유사한 방법으로 해당 데이터 스트림에서 발생된 하나의 항목집합 e의 출현빈도 수 $C_k^R(e)$는 다음과 같이 구해진다.

$$C_k^R(e) = \begin{cases} W_k^R(e) & \text{if } k=1 \\ C_{k-1}^R(e)+W_k^R(e) & \text{if } k\geq 2 \end{cases}$$

where $W_k^R(e)=\begin{cases}1\times d^{k-1} & \text{if } e\subseteq T_k \\ 0 & \text{otherwise}\end{cases}$

4. 양방향 감쇠 기법을 활용한 고관심 정보 탐색

본 절에서는 HAI의 개념을 명확히 정의하고, 하나의 데이터 스트림에서 이를 효율적으로 탐색하기 위한 수행 과정을 예제 데이터 스트림을 이용하여 기술한다. 끝으로, 데이터 스트림에서 HAI를 효율적으로 탐색하는 데이터 스트림 마이닝 기법을 소개한다.

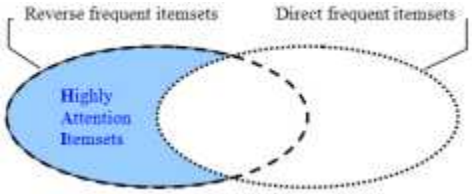


Fig. 3 Highly attention itemsets

4.1 HAI : Highly attention itemsets

하나의 데이터 스트림에서 HAI는 해당 데이터 스트림의 현재 시점에서는 빈번히 발생되지 않으나 과거에는 발생빈도가 컸던 항목집합을 지칭하며, 정방향 감쇠율 및 역방향 감쇠율로 구성되는 양방향 감쇠 기법에 기반하여 다음과 같이 정의된다.

하나의 데이터 스트림 D_k에서 발생된 항목집합 e에 대해서 정방향 감쇠 기법을 적용한 해당 항목집합의 ***정방향 지지도*** $s_k^D(e)$는 해당 항목집합의 정방향 출현빈도 수 $c_k^D(e)$ 값을 해당 데이터 스트림에 속하는 트랜잭션의 총 개수 $|D_k^D$로 나누어 구할 수 있다. 마찬가지로 동일한 항목집합의 역방향 감쇠 기법 적용시 지지도인 ***역방향 지지도*** $s_k^R(e)$는 해당 항목집합의 역방향 출현빈도 수 $c_k^R(e)$ 값을 $|D_k^R$ 로 나누어 구할 수 있다. 본 논문에서는 하나의 데이터 스트림에 발생된 모든 항목집합들 중에서 하나의 항목집합 e의 정방향 지지도 $s_k^D(e)$가 사전에 정의된 지지도 임계값 $S_{min} \in (0,1)$ 보다 크거나 같은 경우 해당 항목집합을 ***정방향 빈발 항목집합***(direct frequent itemsets)라 지칭한다. 또한 해당 데이터 스트림에서 발생한 하나의 항목집합 e의 역방향 지지도 $s_k^R(e)$가 S_{min} 보다 크거나 같은 경우 해당 항목집합을 ***역방향 빈발 항목집합***(reverse frequent itemsets)이라 지칭한다. 분석 대상이 되는 하나의 데이터 스트림에 대해서 지지도 임계값 S_{min}이 주어졌을 때, 해당 데이터 스트림에서 발생된 모든 항목집합들 중에서 역방향 빈발 항목집합에는 해당되나 정방향 빈발 항목집합이 아닌 것들을 본 논문에서는 ***HAI***(Highly attention itemsets)라 정의한다. 즉, 하나의 항목집합 e가 $s_k^R(e) \geq S_{min}$ 및 $s_k^D(e) < S_{min}$를 동시에 만족하는 경우 해당 항목집합은 HAI로 정의된다. <Fig. 3>은 하나의 데이터 스트림에서 속하는 항목집합들 중에서 HAI, 역방향 빈발 항목집합 및 정방향 빈발 항목집합 사이의 관계를 보여준다.

4.2 TDecay 기법

본 절에서는 하나의 데이터 스트림에 대해 HAI를 효율적으로 탐색하기 위한 데이터 스트림 마이닝 기법에 대해서 기술한다. 해당 기법은 정방향 및 역방향 감쇠율을 이용하여 분석 대상이 되는 데이터 스트림에서 발생되는 구성요소의 중요성을 차별화하고 이를 바탕으로 HAI를 탐색하며, 본 논문에서는 해당 기법을 *TDecay*(**T**wo-way **Decay** mechanism for mining data streams) 기법 이라 지칭한다. 즉, *TDecay* 기법은 양방향의 감쇠율을 활용하여 데이터 스트림에 발생한 구성요소의 중요성 차별화를 구현함으로써 HAI를 효율적으로 탐색한다. 대다수 기존의 데이터 스트림 마이닝 기법들[2,10]에서는 지속적으로 확장되는 데이터 스트림에 대한 마이닝 결과 탐색 과정에서 메모리 사용량이 한정적으로 유지되고, 마이닝 결과를 필요로 하는 경우 비교적 짧은 시간에 이를 구할 수 있도록 지원한다. 반면, 각 시점에서 얻어진 마이닝 결과에는 다소간의 오차가 포함될 수 있다[11]. *TDecay* 기법도 마이닝 수행 과정에서의 메모리 사용량 및 처리 시간을 감소시키기 위해서 [2] 및 [10] 등에서 제안한 기법들을 적용하고 있다. HAI 탐색을 위한 양방향 감쇠 기법을 포함한 *TDecay*의 주요 수행 과정은 <Fig. 4>와 같다.

5. 실험 결과 고찰

본 절에서는 양방향 감쇠 기법에 기반한 HAI의

유용성 및 데이터 스트림에서 HAI의 효율적 탐색을 지원하는 TDecay 기법의 효용성을 검증하기 위한 실험 결과를 기술한다. 본 절에서 소개되는 일련의 실험들에서 데이터 집합 DS_aB가 사용되었다. 해당 데이터 집합은 이어진 두 부분 part_A와 part_B로 구성되어 있다. part_A는 단위항목들 집합인 set_A로부터 생성된 트랜잭션들로 구성되어 있으며, part_B는 다른 단위항목들 집합인 set_B로부터 생성된 트랜잭션들로 구성되어 있다. 이때, set_A와 set_B에 공통되는 단위항목은 존재하지 않는다. 각 트랜잭션 집합은 [12]에서 기술된 데이터 집합 생성 방법에 의해 생성되었으며, 각 트랜잭션 집합 생성에 사용된 단위항목 수는 1,000개이다. 데이터 집합 DS_aB는 100,000개의 트랜잭션들로 구성되어 있으며, 처음 10,000개의 트랜잭션은 part_A에 해당되며 나머지는 part_B에 해당된다. 한편, 각 실험에서는 지속적으로 확장되는 데이터 스트림의 특성을 구현하기 위해서 실험 데이터 집합의 각 트랜잭션을 하나씩 순차적으로 처리하여 실험하였다.

양방향 감쇠 기법을 적용한 HAI 탐색 기법 및 데이터 스트림에서 이를 효율적으로 탐색하기 위한 *TDecay* method의 특성을 검증하기 위해서 먼저 정방향 및 역방향 감쇠 기법 적용에 따른 마이닝 결과 집합의 변화를 분석하였다. 이를 위해서 데이터 집합 *DS_aB*에 대한 빈발 항목집합 탐색에서 얻어진 마이닝 결과 집합의 항목집합 수를 분석 하였으며 정방향 감쇠 기법 적용, 역방향 감쇠 기법 적용 및 감쇠 기법 미적용 등 세 가지 경우에 대해서 분석한 결과

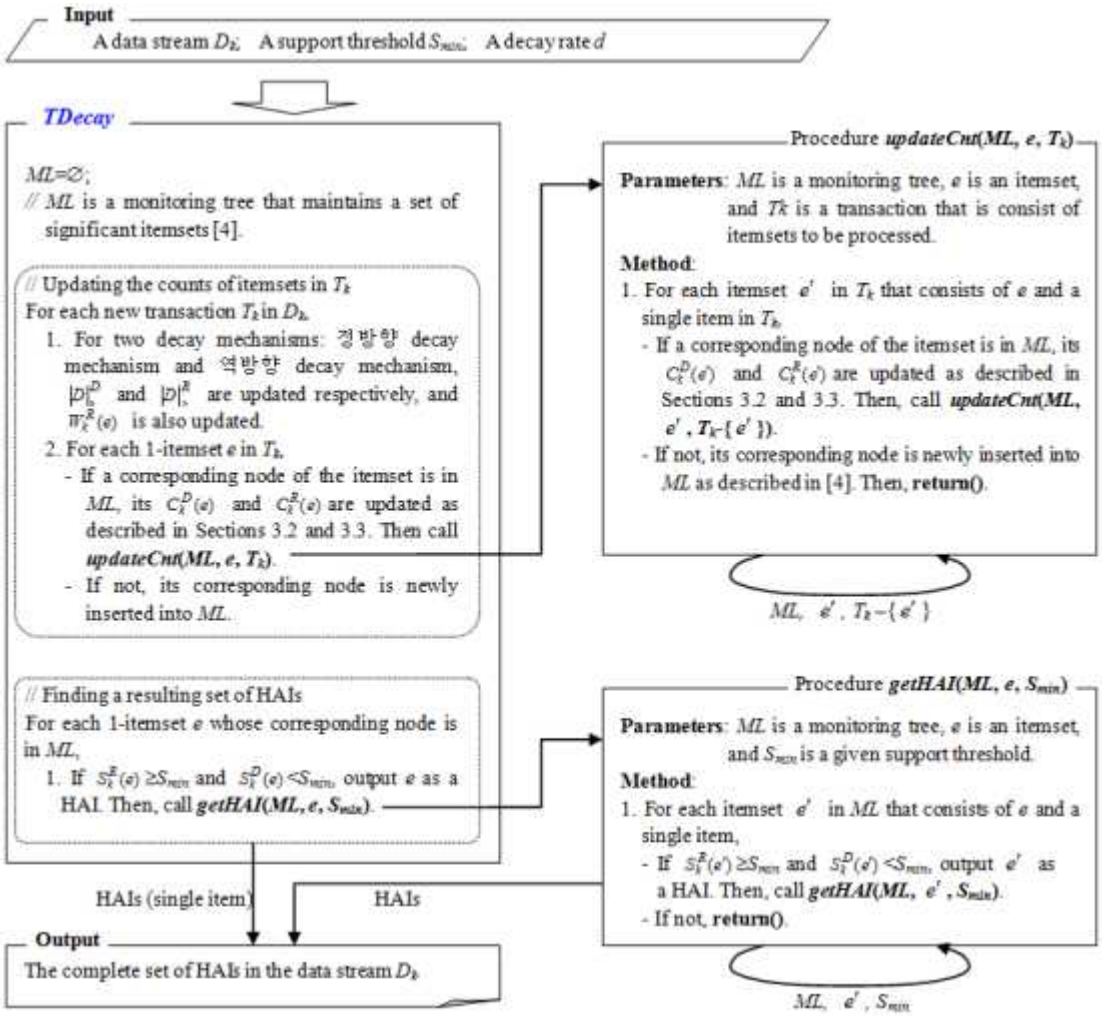
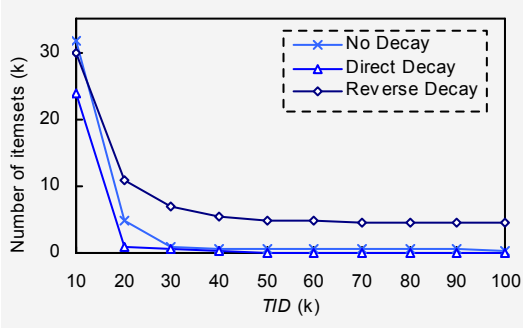


Fig. 4 *TDecay* method: A mining method of HAI

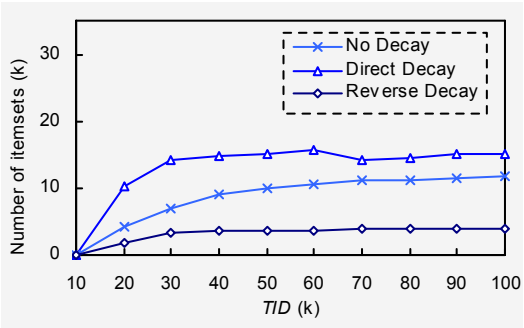
<Fig. 5>에서와 같은 결과를 얻었다. 본 실험에서 빈발항목 집합 탐색을 위한 지지도 임계값 S_{min}은 0.05%로 설정되었으며, b 값이 2이고 h 값이 10,000인 감쇠율 $d=b^{-(1/h)}$을 적용하였다. 본 실험에서 일련의 트랜잭션들은 10,000개의 트랜잭션들로 구성되는 10개의 구간으로 나누어 결과를 구하였으며, 각 구간의 종료 시점에서 구해진 마이닝 결과집합에서 각 단위항목 집합으로부터 유도된 항목집합의 수를 비교하였다.

<Fig. 5>-(a)에서 데이터 집합 *DS_aB*의 첫 번째 구간은 모든 트랜잭션들이 *set_A*로부터 생성된 것들(즉, *part_A*에 속하는 트랜잭션들)로 구성되며, 따라서 해당 구간에서는 세 가지 감쇠 기법 모두 *set_A*로부터 유도된 항목집합들이 마이닝 결과로 다수 탐색된다. 하지만, 두 번째 구간 이후부터는 그 수가 확연한 차이를 보인다. 즉, 정방향 감쇠 기법이나 감쇠 기법 미적용의 경우 *part_B*에 속하는 트랜잭션들이 증가함에 따라 마이닝 결과 집합에서 *set_A*로부터 유도된 항목집합의 수가 급격히 감소됨을 알 수 있다. 반면, 역방향 감쇠 기법의 경우 *part_B*에 속하는 트랜잭션들이 증가하더라도 마이닝 결과 집합에서 *set_A*로부터 유도된 항목집합의 수가 많이 감소되지 않고 다수의 *set_A*로부터 유도된 항목집합들이 지속적으로 탐색된다. 한편, <Fig. 5>-(b)에서 보듯이 두 번째 이후 구간에서는 세가지 경우 모두에서 *set_B*로부터 유도된 항목집합 다수가 지속적으로 탐색된다. 즉, 역방향 감쇠 기법의 경우 과거 발생 정보들 중 관심도가 큰 항목집합(본 실험에서는 *set_A*로부터 유도된 항목집합)뿐만 아니라 근래에 발생된 정보들 중 관심도가 큰 항목집합(본 실험에서는 *set_B*로부터 유도된 항목집합)도 다수가 함께 탐색됨을 알 수 있다. 이러한 결과로부터 과거 단골 고객에 대한 집중형 마케팅 등을 위한 HAI의 효율적 탐색을 위해서는 역방향 감쇠 기법을 단독으로 적용하기보다는 양방향 감쇠 기법과 같이 정방향 및 역방향 감쇠 기법이 복합적으로 적용될 필요가 있음을 알 수 있다. 즉, 본 논문에서 제안한 HAI가 해당 목적을 위해 유용하게 활용 될 수 있을 것이다.

<Table 1>은 <Fig. 5>와 동일한 실험에서 마이닝 결과로 얻어지는 몇 개 항목집합의 지지도 변화를 상세히 분석한 결과를 보여주며, 세 가지 감쇠 기법에 대해서 해당 항목집합들의 지지도 변화를 조사하였다. 해당 항목집합들은 *part_A*에 속하는 트랜잭션에서 발생된 항목집합으로서 정방향 감쇠 기법이나 감쇠 기법 미적용의 경우 *part_B*에 속하는 트랜잭션들이 증가함에 따라 지지도가 급격히 감소된다. 특히, 정방향 감쇠 기법의 경우 해당 항목집합들의 지지도가 매우 작은 수준(거의 0에 가까운 값)으로 감소되었다. 반면, 역방향 감쇠 기법의 경우 동일한 조건에서 지지도가 감소되기는 하나 감소폭이 상대적으로 적은 수준이다. 해당 실험의 S_{min} 값이 0.05%임을 감안할 때, T_{10000}이 처리된 후 얻어진 마이닝 결과 집합에서 네 개의 항목집합들은 HAI로 탐색되지 못한다. 왜냐하면 해당 항목집합들의 지지도가 역방향 감쇠 기법 및 정방향 감쇠 기법 모두에서 S_{min}보다 크기 때문이다. 하지만 T_{100000}이 처리된 후 얻어진 마이닝 결과 집합에서 네 개의 항목집합들은 모두 HAI로 탐색된다. 왜냐하면 해당 항목집합들의 지지도가 역방향 감쇠 기법의 경우는 S_{min}보다 크고 정방향 감쇠 기법의 경우에는 S_{min}보다 작기 때문이다.



(a) From a set of items *set_A*



(b) From a set of items *set_B*

Fig. 5 Number of itemsets derived from each set of items

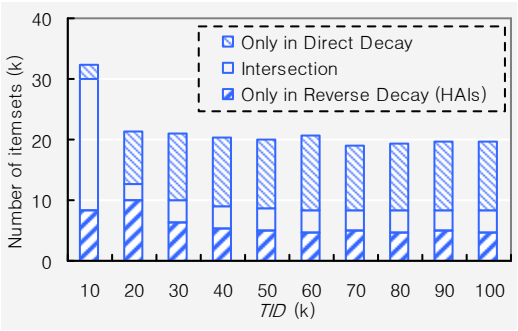


Fig. 6 Number of HAI

<Fig. 6>은 데이터 집합 *DS_aB*에 대한 HAI 탐색 실험에서 역방향 감쇠 기법 적용시 얻어진 빈발 항목집합의 수, 정방향 감쇠 기법 적용시 얻어진 빈발 항목집합의 수 및 이들 사이에 공통으로 포함되는 항목집합의 수에 대한 분석 결과를 보여준다. 본 실험에서 지지도 임계값 등의 실험 조건은 <Fig. 5>의 실험과 동일하며, 각 구간 종료 시점에서 마이닝 결과로 구해지는 빈발 항목집합의 수를 비교하였다. 그림에서 보듯이 모든 트랜잭션들이 part_A에 속하는 것들로 구성되는 첫 번째 구간에서는 많은 수의 항목집합들이 두 감쇠 기법에서 공통으로 탐색되었다. 하지만, 두 번째 이후의 구간에서는 공통적으로 탐색되는 항목집합의 수가 크게 감소된다. 하지만, 해당 구간들에서도 역방향 감쇠 기법에서 탐색된 빈발 항목집합 중 일부가 정방향 감쇠 기법에서도 빈발 항목집합으로 탐색되며, 해당 항목집합들은 분석 대상 데이터 집합 전체에서 지속적으로 빈번히 발생된 것들로서 기존의 일반적인 데이터 스트림 마이닝 기법(즉, 감쇠 기법 미적용 또는 정방향 감쇠 기법만 적용)에서도 탐색된다. 본 실험 결과에서 얻어진 항목집합들 중 정방향 감쇠 기법이나 공통으로 탐색된 항목집합을 제외하고 역방향 감쇠 기법 적용 시에만 탐색된 항목집합을 HIAI라 할 수 있다. 한편, 개별 항목집합의 지지도가 지속적으로 변화되므로 HAI로 탐색되는 항목집합의 수가 비슷한 경우라도 시간 흐름에 따라 이에 속하는 항목집합은 계속 변화된다.

TDecay 기법의 기본적인 성능은 마이닝 수행과정에서의 메모리 사용량 및 수행시간 분석 등을 통해 확인할 수 있으며, [2] 등에서 제시된 기본적 감쇠 기법에서와 크게 다르지 않다. <Fig. 5>와 동일한 실험에서 트랜잭션 수가 증가됨에 따라 메모리 사용량 및 마이닝 수행시간이 다소 증가하였으나, 메모리 최대 사용량은 15MB를 넘지 않으며 마이닝 수행시간의 경우에도 25ms 보다 작음을 확인하였다.

6. 결 론

본 논문에서는 데이터 마이닝 관심 분야 중의 하나인 고유용 관심 정보의 하나로서 HAI를 제시하고 이를 효율적으로 탐색하기 위한 양방향 감쇠 기법을 제안하였다. 또한, 하나의 데이터 스트림에서 HAI를 효율적으로 탐색하기 위한 탐색 기법을 제시하였다. HAI는 하나의 데이터 스트림에서 최근에는 자주 발생되지 않으나 과거에는 빈번히 발생했던 것으로 관심도가 큰 항목집합으로서 과거 단골에 대한 집중형 마케팅 분야 등에서 활용도가 큰 항목집합을 지칭한다. 근래 발생 정보의 중요성은 높게 부여하는 반면 과거 발생 정보는 중요성이 낮은 것으로 간주하거나 무시하는 기존의 정보 중요성 차별화 기법들과 달리 HAI 및 이의 효율적 탐색을 위한 양방향 감쇠 기법은 분석 대상 데이터 스트림에서 시간축을 기준으로 보다 다양한 형태의 정보 중요성 차별화를 지원하여 고유용/고관심 항목집합을 탐색할 수 있도록 지원한다. 결론적으로 본 논문에서 제시된 HAI 및 양방향 감쇠 기법은 데이터 스트림 형태로 정보를 발생 시키는 다양한 컴퓨터 응용 환경에서 효과적으로 적용

Table 1 Comparison of supports

Mechanism / TID / Itemsets	Reverse decay			No decay			Direct decay		
	10000	50000	100000	10000	50000	100000	10000	50000	100000
(1368,1692)	0.00281	0.00145	0.00141	0.00270	0.00054	0.00027	0.00260	x	x
(1227,1722)	0.00268	0.00138	0.00134	0.00260	0.00052	0.00026	0.00253	x	x
(1381,1640,1896)	0.00170	0.00088	0.00085	0.00170	0.00034	0.00017	0.00168	x	x
(1207,1381,1640,1896)	0.00161	0.00083	0.00081	0.00160	0.00032	0.00016	0.00157	x	x

될 수 있으며, HAI 탐색의 결과는 해당 분야 사용자들에게 매우 유용한 정보를 제공할 수 있다.

References

[1] S.K. Tanbeer, C.F. Ahmed, B.-S. Jeong, and Y.-K. Lee, "Sliding window-based frequent pattern mining over data streams," *Information Sciences*, 179(22), pp. 3843-3865, 2009.

[2] J.H. Chang and W.S. Lee, "Finding Recently Frequent Itemsets Adaptively over Online Transactional Data Streams," *Information Systems*, 31(8), pp. 849-869, 2006.

[3] Q. Huang and W. Ouyang, "Mining Sequential Patterns in Data Streams," in Proc. of the 6th Int'l Symposium on Neural Networks, pp.865-874, 2009.

[4] H.T. Lam and T. Calders, "Mining top-K frequent items in a data stream with flexible sliding windows," in Proc. of the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, pp.283-292, 2010.

[5] C.C. Aggarwal and P.S. Yu, "A framework for clustering uncertain data streams," in Proc. of the Int'l Conf. on Data Engineering, pp. 150-159, 2008.

[6] C.-W. Li and K.-F. Jea, "An adaptive approximation method to discover frequent itemsets over sliding-window-based data streams," *Expert Systems with Applications*, 38(10), pp. 13386-13404, 2011.

[7] H.-F. Li and S.-Y. Lee, "Mining frequent itemsets over data streams using efficient window sliding techniques," *Expert Systems with Applications*, 36(2), pp. 1466-1477, 2009.

[8] H.-F. Li, H.-Y. Huang, Y.-C. Chen, Y.-J. Liu, and S.-Y. Lee, "Fast and memory efficient mining of high utility itemsets in data streams," in Proc. of the Int'l Conf. on Data Engineering, pp. 881-886, 2008.

[9] B.-E. Shie, P.S. Yu, V. S.Tseng, "Efficient algorithms for mining maximal high utility itemsets from data streams with different models," *Expert Systems with Applications*, 39(17), pp. 12947-12960, 2012.

[10] J.H. Chang and W.S. Lee, "Efficient Mining Method for Retrieving Sequential Patterns over Online Data Streams," *Journal of Information Science*, 31(2), pp. 420-432, 2005.

[11] N. Gabsi, F. Clerot, and G. Hebrail, "Efficient trade-off between speed processing and accuracy in summarizing data stream," in Proc. of the 14th PAKDD, pp.343-353, 2010.

[12] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," in Proc. of the 20th International Conf. on Very Large Data Bases, pp. 487-499, 1994.

장 중 혁 (Joong-Hyuk Chang)

- 정회원
- 연세대학교 컴퓨터과학과 이학사
- 연세대학교 컴퓨터과학과 공학석사
- 연세대학교 컴퓨터과학과 공학박사
- 대구대학교 정보통신대학 컴퓨터IT공학부 교수
- 관심분야 : 데이터공학, 데이터베이스, 데이터 스트림, 데이터 스트림 마이닝, 정보 중요성 차별화, 빅데이터, SNS 분석