Journal of the Korea Academia-Industrial cooperation Society
Vol. 17, No. 4 pp. 521-528, 2016

http://dx.doi.org/10.5762/KAIS.2016.17.4.521
ISSN 1975-4701 / eISSN 2288-4688

웹 클릭 스트림에서 고유용 과거 정보 탐색

장중혁[1*]
[1]대구대학교 컴퓨터IT공학부

Finding high utility old itemsets in web-click streams

Joong-Hyuk Chang[1*]
[1]**Division of Computer & IT, Daegu University**

요 약 개인용 컴퓨터 및 각종 모바일 기기의 이용 증가로 인해 많은 분야에서 다양한 형태의 웹기반 서비스들이 널리 활용되고 있다. 이에 따라 해당 분야에서 개인 맞춤형 서비스를 지원하기 위한 사용자 이용 로그 분석 등에 대한 연구가 활발히 진행되고 있으며, 특히 사용자 로그 데이터를 구성하는 구성요소의 중요성 차별화에 기반한 분석 기법들이 활발히 연구되었다. 본 논문에서는 웹 클릭 스트림에서 유용하게 적용될 수 있는 고유용 과거 정보 탐색 기법을 제시한다. 해당 기법을 통해 기존의 웹 클릭 스트림 분석 기법에서는 쉽게 탐색하지 못했던 정보인 타겟 마케팅 등에 유용하게 활용될 수 있는 중요 정보를 쉽게 탐색할 수 있다. 본 논문의 연구 결과는 IoT 환경 및 생물정보 분석 등과 같이 데이터 스트림 형태로 정보를 발생시키는 다양한 컴퓨터 응용 분야에도 활용될 수 있을 것이다.

Abstract Web-based services are used widely in many computer application fields due to the increasing use of PCs and mobile devices. Accordingly, topics on the analysis of access logs generated in the application fields have been researched actively to support personalized services in the field, and analyzing techniques based on the weight differentiation of information in access logs have been proposed. This paper outlines an analysis technique for web-click streams, which is useful for finding high utility old item sets in web-click streams, whose data elements are generated at a rapid rate. Using the technique, interesting information can be found, which is difficult to find in conventional techniques for analyzing web-click streams and is used effectively in target marketing. The proposed technique can be adapted widely to analyzing the data generated in a range of computing application fields, such as IoT environments, bio-informatics, etc., which generated data as a form of data streams.

Keywords : Data streams, Data stream mining, High utility old itemsets, Highly attention itemsets, Web-click streams

1. 서론

웹 클릭 스트림 분석은 데이터 스트림 분석의 주요 분야 중 하나로서 대용량의 웹 사용 기록 등을 효과적으로 분석하여 보다 고품질의 서비스 제공을 지원하는 것에 목적을 두고 있다. 예를 들어, 하나의 웹 사이트에서 발생되는 웹 클릭 스트림에 대한 분석을 통해 해당 웹 사이트를 이용하는 사용자의 개인별 선호도 및 이용 성향 등을 탐색하고 이를 반영한 맞춤형 서비스를 제공한다. 웹 클릭 스트림은 구성요소가 지속적으로 확장되는 특성으로 인해 이전의 한정적인 데이터 집합에 대한 분석 기술로는 효율적인 분석에 많은 어려움이 있으며, 이를 보완하기 위해 데이터 스트림 분석 기법[1,2]을 활용한 다양한 분석[3,4]이 연구되었다.

근래 들어 데이터 스트림 마이닝 분야에서 기존의 일반적인 방법으로는 탐색에 어려움이 있었던 새로운 형태

이 논문은 대구대학교 학술연구비지원에 의한 논문임
*Corresponding Author : Joong-Hyuk Chang(Daegu Univ.)
Tel: +82-53-850-6588 email: jhchang@daegu.ac.kr
Received February 12, 2016 Revised March 16, 2016
Accepted April 7, 2016 Published April 30, 2016

의 고유용 정보 탐색에 대한 관심이 크게 증가되어 왔다. 이의 일환으로 데이터 스트림을 구성하는 구성요소의 발생 시간을 기준으로 과거 발생 정보에 높은 가중치를 부여하는 정보 중요성 차별화 기법에 대한 관심이 증가되고 있다. 해당 기법의 유용성은 인터넷 쇼핑몰 사이트 판매 실적에 대한 다음의 사례에서 확인된다.

[*고객_A*] 레이저 프린터, 스캐너, SSD, RAM, DVD 레코더 구매 (2014년 1사분기)→ 구매 기록 없음 (2014년 2사분기 이후)

[*고객_B*] 구매 기록 없음 (2014년 이전)→ 레이저 프린터, 스캐너 구매 (2015년 1사분기)

위 예제에서 *고객_A*는 2014년 1사분기에는 많은 물품을 구매한 고객이었으나 그 이후에는 물품 구매기록이 없는 반면 *고객_B*의 경우 2015년 1사분기에 처음으로 물품을 구매하였으며 그 이전에는 구매 기록이 없다. 해당 구매 기록이 포함된 데이터 스트림에 대해 최근 정보에 높은 가중치를 부여하는 정보 중요성 차별화 기법을 적용하는 경우 2015년 2사분기 이후를 기준으로 *고객_A* 구매기록은 낮은 중요성을 갖는 과거 정보이므로 일반적으로 중요성이 낮게 간주된다. 하지만, *고객_A*와 같은 고객을 대상으로 집중 마케팅 전개하여 해당 고객이 다시 물품을 구매하도록 하는 경우 판매 실적을 증대시킬 수 있을 것이다. 즉, 최근 발생 정보에 집중된 일반적인 기법과 다른 정보 중요성 차별화 기법을 통해 새로운 형태의 고유용 정보(즉, 특화된 마케팅 대상)를 얻게 된다.

빠른 속도록 확장되는 웹 클릭 스트림에서 발생 시점에 따른 적절한 가중치 부여를 통해 특화된 마케팅을 위한 중요 정보로 활용될 수 있는 고유용 정보를 탐색할 수 있으며, 본 논문에서는 이와 관련하여 다음의 내용들을 기술한다. 먼저 2장에서는 웹 클릭 스트림 및 데이터 스트림 분석, 정보 중요성 차별화 등 본 논문의 연구와 관련된 이전 연구들을 기술한다. 3장에서는 웹 클릭 스트림을 명확히 정의하고 고유용 과거 정보의 유용성을 예를 들어 기술하며, 해당 정보의 효율적 탐색을 위한 웹 클릭 스트림 분석 기법을 제시한다. 4장에서는 웹 로그 데이터를 활용한 실험을 통해 본 논문에서 제시된 방법의 성능을 검증하고, 5장에서 논문의 결론을 맺는다.

2. 관련 연구

웹 기반 응용 서비스에서 사용자 맞춤형 서비스를 제공하기 위해 웹 로그 또는 웹 클릭 스트림을 분석하는 다양한 연구들[3-6]이 진행되어 왔다. [4]에서는 검색 키워드에 대한 웹 페이지 사용 행위 및 방문 웹 페이지 리스트를 분석하여 패턴을 추출하여 검색 의도별 방문 웹 페이지의 연결망을 생성하였다. 이를 통해 사용자별 맞춤형 웹 페이지 추천을 지원하였다. [6]에서는 웹 사용자 그룹별 맞춤형 서비스를 제공하도록 웹 방문 데이터를 근거로 유사 특성을 갖는 그룹을 생성하였다. [3]과 [5]에서는 웹 기반 서비스에서 발생되는 정보를 스트림 형태의 지속적인 발생 정보로 간주하여 특성 정보를 추출하여 맞춤형 서비스에 활용할 수 있도록 하는 웹 클릭 스트림에 대한 분석 기법을 제시하였다. 이들 방법을 포함한 다수의 이전 연구들에서는 웹 로그 또는 웹 클릭 스트림을 구성하는 정보들은 동일한 중요성을 갖는 것으로 간주하고 있다.

데이터 스트림 연구에서는 시간 흐름에 따른 가변성이 큰 데이터 스트림의 특성을 고려하여 구성 요소의 중요성을 발생 시간 축을 기준으로 차별화하기 위한 다양한 기법들[7-11]이 연구되었으며, 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 것은 이동 윈도우 기법과 감쇠 기반 기법이다. 이동 윈도우 기법[7-9]은 일정 크기의 시간 윈도우를 정의하여 해당 범위 내에 포함되는 정보만 유효한 것으로 간주하고 범위에 포함되지 않는 정보는 무효한 것으로 간주하여 정보 중요성 차별화를 구현한다. 일반적으로 해당 기법에서는 시간 흐름에 따라 윈도우를 이동하면서 윈도우 크기만큼의 최근 시간 범위를 유효 범위로 정의한다. 감쇠 기반 기법[9-11]은 하나의 데이터 스트림을 구성하는 구성요소들 중에서 최근에 발생한 구성요소는 상대적으로 높은 중요성을 갖는 것으로 간주하고 과거에 발생한 구성요소는 그 중요성이 시간 흐름에 따라 점차적으로 감쇠되도록 하는 기법이다. 이를 통해 일정 시점에서 발생한 정보가 해당 시점에서는 매우 중요한 정보로 간주되지만 시간 흐름에 따라 그 중요성이 감쇠되고 충분히 오랜 시간이 지난 후에는 사실상 무효한 정보로 간주된다.

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 최근 정보에 집중된 분석 결과를 얻고자 하는 경우에는 매우 효과적으로 적용될 수 있으나 과거 일정 시점에 관심도가 큰 것으로 간주되었던 정보들을 탐색하는데 어려움이 있다. 즉, 집중 마케팅 등을 위해 유용하게 활용될 수 있는 과거 발생 정보들을 효과적으로 탐색하는 데에는 한계가 있다.

3. 웹 클릭 스트림에서 고유용 정보 탐색

웹 기반의 실제 응용 분야에서 발생되는 로그 등의 정보로부터 고유용 과거 정보를 탐색하기 위해서는 해당 정보들을 웹 클릭 스트림 형태로 재구성할 필요가 있다. 본 절에서는 먼저 웹 클릭 스트림의 정의에 대해 간략히 기술한다. 다음으로 고유용 과거 정보의 개념을 명확히 기술하고 이의 유용성을 예를 들어 설명한다. 끝으로 웹 클릭 스트림에서 고유용 과거 정보 탐색을 위한 세부 과정을 기술한다.

3.1 웹 클릭 스트림

고유용 과거 정보 탐색 대상이 되는 웹 클릭 스트림은 일반적인 데이터 스트림과 마찬가지로 구성요소가 지속적으로 생성되는 무한 데이터 집합으로 간주할 수 있으며, [5]에서 제시된 웹 정보 매핑 방법과 [12]에서의 기술된 데이터 표기법에 따라 다음과 같이 정의된다.

먼저, I는 하나의 응용 서비스에서 단위 정보를 표시하는데 사용되는 단위항목(item)들의 집합을 나타내며, 분석 대상 웹 사이트를 구성하는 각 웹페이지를 단위항목으로 간주한다. 항목집합 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.

다음으로 사용자가 해당 웹 사이트에 접근하여 연속적으로 접근된 웹 페이지들을 하나의 트랜잭션으로 구성한다. 이때 일정 시간 동안 사용자 입력이 없는 경우나 사용자가 접속을 종료한 경우에도 하나의 트랜잭션이 완성된 것으로 간주한다. 즉, 트랜잭션(transaction)은 웹 페이지를 나타내는 항목집합들로 구성되며, 서로 다른 트랜잭션을 구분하는 식별자 *TID*를 갖는다. 이때, k 번째 생성된 트랜잭션을 T_k로 나타낸다.

이러한 방법을 통해 하나의 웹 사이트에서 접근하는 여러 사용자의 지속적인 접근 기록을 데이터 스트림 형태를 띄는 웹 클릭 스트림으로 변환할 수 있다. 즉, 웹 클릭 스트림은 하나의 웹 사이트에 대한 접근 기록으로부터 생성된 것으로서 [5,9,10]에서 정의된 데이터 스트림의 형태를 갖는다. 따라서 본 논문의 나머지 부분에서 데이터 스트림이라 함은 웹 클릭 스트림을 지칭한다. 한편, 하나의 웹 사이트에서 새로운 트랜잭션 T_k가 생성되었을 때, 현재 데이터 스트림 D_k는 현재까지 생성된 모든 트랜잭션들로 구성되는 웹 클릭 스트림을 의미한다. 즉, $D_k=<T_1,T_2,\cdots,T_k>$ 로 표현되며, 해당 데이터 스트림에 포함된 트랜잭션의 총 개수는 $|D|_k$.로 나타낸다.

3.2 고유용 과거 정보

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 포함되는 트랜잭션 중 해당 항목집합 e를 포함하고 있는 트랜잭션 개수를 의미한다. 또한, 하나의 항목집합 e의 지지도를 나타내는 $S_k(e)$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.

하나의 데이터 스트림에서 ***고유용 과거 정보***는 해당 데이터 스트림의 현재 시점에서는 빈번히 발생되지 않으나 과거에는 발생빈도가 컸던 항목집합을 지칭한다. 즉, 분석 대상이 되는 하나의 데이터 스트림과 지지도 임계값이 주어졌을 때 고유용 과거 정보라 함은 해당 데이터 스트림에서 과거에는 빈발 항목집합이었으나 근래에는 발생빈도가 적은 항목집합을 의미하며, 정방향 감쇠율[11] 및 역방향 감쇠율[5]을 활용하여 정의된다. 정방향

감쇠율은 최근 발생 정보에 보다 높은 중요성을 갖고 과거 발생 정보는 시간 흐름에 중요성이 감쇠되도록 한다. 반면 역방향 감쇠율은 과거 발생 정보에 보다 높은 중요성을 부여하고 새로 발생되는 정보는 시간 흐름에 따라 감쇠된 낮은 중요성을 갖도록 한다. 하나의 데이터 스트림 D_k에서 발생한 항목집합 e에 대해서 정방향 감쇠 기법 적용시 지지도 및 역방향 감쇠 기법 적용시 지지도를 각각 정방향 지지도 $DS_k(e)$ 및 역방향 지지도 $RS_k(e)$라 하고 해당 데이터 스트림에 대한 고유용 과거 정보 탐색을 위한 임계값 S_{min}이 주어졌을 때, $RS_k(e) \geq S_{min}$ 및 $DS_k(e) < S_{min}$를 동시에 만족하는 경우 해당 항목집합은 고유용 과거 정보로 탐색된다.

예제 데이터를 활용해 고유용 과거 정보의 개념을 설명하면 다음과 같다. Fig. 1에서 제시된 예제 데이터 스트림 D_k는 4개의 단위항목으로 구성되는 총 4개의 트랜잭션으로 구성된다. 이때, 감쇠율 $d=2^{-1}$(즉, 하나의 트랜잭션이 새로 발생될 때마다 트랜잭션의 가중치가 2^{-1}만큼 변화됨)을 적용하는 경우 각 트랜잭션의 가중치는 Table 1에서와 같이 구해지며, 따라서 D_k에서 발생된 몇 개 항목집합의 지지도를 각 감쇠율에 대해 구하면 Table 2에서와 같다.

해당 결과에서 항목집합 *ab* 및 *abc*의 경우 정방향 지지도는 낮으나 역방향 지지도는 매우 높다. 반면, 항목집합 *bd* 및 *bcd*의 경우 정방향 지지도는 높으나 역방향 지지도는 상대적으로 낮다. 본 예제에서 지지도 임계값 S_{min}이 0.5로 설정된 경우 항목집합 *bd* 및 *bcd*의 경우 근래에 빈번히 발생한 빈발 항목집합은 될 수 있으나 고유용 과거 정보는 되지 못하는 반면 항목집합 *ab* 및 *abc*는 고유용 과거 정보로 탐색된다. 한편, 고유용 과거 정보의 정의에 따라 항목집합 *bc*와 같이 정방향 지지도 및 역방향 지지도 모두가 지지도 임계값 이상인 경우에도 고유용 과거 정보가 되지 못한다. 고유용 과거 정보 탐색은 항목집합 *ab* 및 *abc* 등에서 보는 바와 같이 데이터 스트림 초기의 과거 트랜잭션에서는 빈번히 발생되었으나 근래에 생성된 트랜잭션에서는 발생빈도가 낮은 것들을 탐색하여 과거 단골에 대한 타겟 마케팅 등에 유용하게 활용될 수 있는 정보를 탐색하는데 목적을 두고 있기 때문이다.

TID	Transaction
1	{a, b, c}
2	{a, b, c, d}
3	{b, c}
4	{b, c, d}

Fig. 1. An example data stream D_k

Table 1. The weights of transactions

TID	Weight	
	Direct decay	Reverse decay
1	0.125	1
2	0.25	0.5
3	0.5	0.25
4	1	0.125

Table 2. The supports of several itemsets

Itemset e	Support	
	$DS_4(e)$	$RS_4(e)$
ab	0.20	0.80
bc	1.00	1.00
bd	0.67	0.33
abc	0.20	0.80
bcd	0.67	0.33

3.3 웹 클릭 스트림에서 고유용 과거 정보 탐색

분석 대상인 하나의 데이터 스트림과 고유용 과거 정보 판단의 기준이 되는 지지도 임계값 S_{min} 및 감쇠율 $d=2^{-(1/h)}$이 주어졌을 때, 분석 결과로 얻어지는 고유용 과거 정보는 Fig. 2에서와 같이 다음의 과정을 거쳐 얻어진다. 먼저, 분석 대상 데이터 스트림에 포함된 하나의 트랜잭션을 읽는다. 이어서 해당 트랜잭션에 출현한 모든 항목집합(예를 들어, 트랜잭션 T_1={*a*, *b*, c}에 대해서는 항목집합 *a*, *b*, *c*, *ab*, *ac*, *bc*, *abc*)의 출현빈도 수를 두 가지 경우에 대해 모두 갱신한다. 즉, 역방향 감쇠 기법을 적용한 경우의 출현빈도 수 및 정방향 감쇠 기법을 적용한 경우의 출현빈도 수를 모두 갱신한다. 웹 클릭 스트림을 비롯한 데이터 스트림의 경우 하나의 데이터 스트림을 구성하는 단위항목 수가 많고 트랜잭션 지속적으로 생성되므로 관리해야 할 항목집합이 크게 증가될 수 있다. 이러한 상황에서 항목집합의 출현빈도 수를 효율적으로 관리하기 위한 방법은 [11] 등에서 자세히 제시된 바 있으며, 본 논문에서는 이에 대한 설명은 생략한다. 이상의 과정들은 새로운 트랜잭션을 읽을 때마다 계속 반복된다. 즉, 일정 시점에서 탐색 결과 집합을 얻고자 하는 경우가 아니면 이러한 과정들이 반복된다. 일정 시점에 존재하는 모든 고유용 과거 정보로 구성되는 분

석 결과를 얻고자 하는 경우 출현빈도 수를 관리하고 있는 모든 항목집합에 대해서 해당 항목집합의 역방향 지지도 및 정방향 지지도를 S_{min}과 비교하여 분석 결과를 구한다. 예를 들어, 출현빈도가 관리되고 각 항목집합의 e에 대해서 $RS_k(e) \geq S_{min}$을 만족하는 동시에 $DS_k(e) < S_{min}$을 만족하는 경우 해당 항목집합 e는 고유용 과거 정보로 탐색된다. 이와 더불어 기존의 데이터 스트림 분석에서와 같이 정방향 감쇠율만 적용한 분석 결과도 쉽게 구할 수 있다. 즉, 앞서 기술한 고유용 과거 정보 탐색 단계 설명에서 정방향 지지도만 고려하여 그 값이 S_{min}보다 크거나 같은 항목집합을 탐색하면 해당 결과를 얻게 된다.

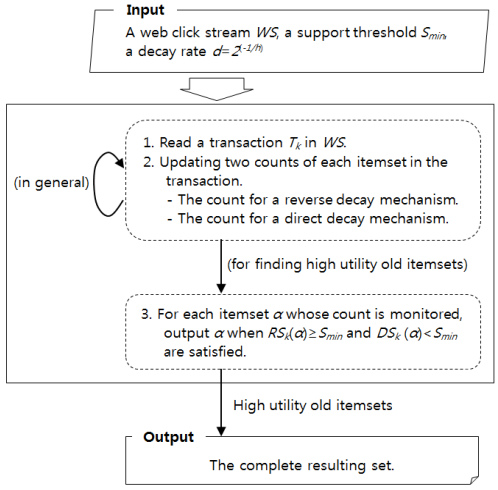

Fig. 2. Steps for finding high utility old itemsets

4. 실험 결과 고찰

본 절에서 소개되는 일련의 실험들에서는 데이터 집합 *DS_Web*이 사용되었으며, 해당 데이터 집합은 국내 유명 포털 사이트의 사용자 접근 기록으로부터 생성되었다. 웹 사이트에 접속하는 하나의 사용자에 의해 연속적으로 접근된 웹 페이지들을 의미적 묶음으로 간주하여 트랜잭션을 구성한다. 즉, 하나의 트랜잭션은 한 번의 접속을 통해 함께 탐색되는 웹 페이지를 분석하는데 중요한 의미를 갖는다. 한편, 일정 시간 동안 사용자 입력이 없는 경우에도 하나의 트랜잭션이 생성된 것으로 간주되고, 이후 접근되는 웹 페이지들은 새로운 트랜잭션으로 간주된다. 실험 데이터 집합 *DS_Web*을 구성하는 단위 항목(즉, 웹 사이트를 구성하는 웹 페이지)의 수는 545이며, 총 트랜잭션 수는 260,385개이다. 트랜잭션의 최소 길이는 2이고 최대 길이는 30이며 평균 길이는 5이다. 또한, 동일 사용자가 60초 동안 다른 웹 페이지를 접근하지 않는 경우 하나의 트랜잭션이 종료된 것으로 간주하였다. 웹 클릭 스트림 환경에서는 분석 대상 트랜잭션을 순차적으로 처리하게 되며, 이를 위해 본 실험에서는 각 데이터 집합을 구성하는 순차정보를 하나씩 차례로 탐색하여 처리한다.

일반적인 데이터 스트림 분석에서와 마찬가지로 고유용 과거 정보 탐색에서도 수행 과정에서의 메모리 사용량 및 트랜잭션 처리 시간 등의 기본 성능에 대한 검증이 필요하다. 이를 위해서 본 논문에서는 기본 감쇠율을 $d=2^{-(1/h)}$로 설정하고 h값을 변화 시켜 실험하였다. h값은 트랜잭션의 가중치 감쇠 속도를 결정하는 값으로서, 새로운 트랜잭션이 h개 발생될 때마다 트랜잭션의 가중치가 절반으로 감쇠됨을 의미한다. 예를 들어 h값이 5인 경우 매 다섯 개의 트랜잭션이 새로 발생될 때마다 가중치가 절반으로 감쇠된다. 기본 성능 검증 실험에서 고유용 과거 정보 탐색을 위한 임계 지지도는 0.5%로 설정되었다. 이하 실험에서는 실험 데이터 집합 *DS_Web*을 각각 50,000개의 트랜잭션으로 구성되는 5개의 구간으로 나누고 각 구간 끝 시점의 결과를 구하여 비교하였다.

Fig. 3은 고유용 과거 정보 탐색 과정에서의 메모리 사용량을 보여주며, 각 구간별 최대 메모리 사용량을 의미한다. 그림에서 보듯이 트랜잭션 수가 증가됨에 따라 메모리 사용량이 다소 증가되었으나 최대 사용량은 15MB를 넘지 않으며, 특히 충분히 많은 수의 트랜잭션이 처리된 후(즉, 4번째 구간 이후)에는 메모리 사용량이 거의 증가되지 않고 일정 수준으로 유지되고 있다. 이는 h 값이 변화되는 경에도 유사한 경향을 보여준다. Fig. 4는 각 구간에서 트랜잭션 처리 시간의 평균값을 보여준다. 메모리 사용량에서와 마찬가지로 트랜잭션 수가 증가됨에 따라 트랜잭션 처리시간이 다소 증가되나 25msec(즉, 25/1,000초)보다 작으며, 충분히 많은 수의 트랜잭션이 처리된 후(즉, 4번째 구간 이후)에는 트랜잭션 처리시간이 거의 증가되지 않고 일정 수준으로 유지되고 있다. 이러한 결과에서 보듯이 기존의 데이터 스트림 마이닝 기법들과 마찬가지로 고유용 과거 정보 탐색도 메모리 사용량 및 트랜잭션 처리 시간 측면에서 효율

적으로 분석 결과를 얻게 된다.

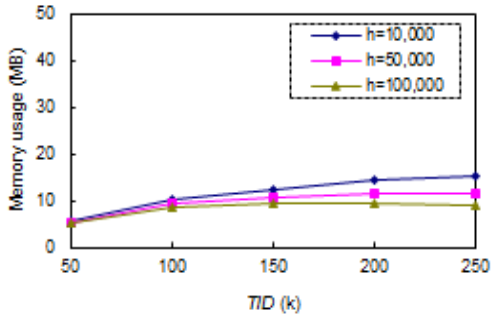

Fig. 3. Memory usage

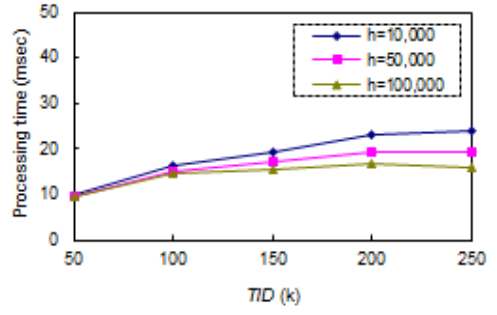

Fig. 4. Processing time

Table 3은 정방향 감쇠 기법 및 역방향 감쇠 기법 적용에 따라 탐색된 항목집합의 수를 보여주며, 이들 사이에 공통으로 탐색된 항목집합 수 또한 제시하고 있다. 이때, *h* 값은 10,000으로 설정되었다. 표에서 보듯이 정방향 감쇠 기법에서만 탐색된 항목집합의 수는 매우 적은 반면 역방향 감쇠 기법에서만 탐색된 항목집합은 그 수가 상당히 많은 편이다. 이러한 결과로부터 해당 데이터 집합 *DS_Web*에서는 시간 흐름에 따라 트랜잭션을 구성하는 항목집합이 크게 변화되었음을 알 수 있다. 즉, 데이터 집합 초창기에 빈번하게 발생된 항목집합들 중 다수가 시간 흐름에 따라 새롭게 생성된 트랜잭션에서는 매우 적게 포함된 것으로 판단된다. 이와 같은 특성을 갖는 데이터 집합의 경우 정방향 감쇠 기법만을 적용하는 기존의 가중치 감쇠 기법을 적용하는 경우 최근 데이터에만 집중된 탐색 결과를 얻게 되고, 고유용 과거 정보 등과 같은 중요한 결과를 얻지 못할 수 있다.

Table 3. Number of itemsets

TID (k)	Number of itemsets		
	Only in reverse decay	Intersection	Only in direct decay
50	1624	1994	2
100	1794	1766	5
150	2607	948	6
200	498	3057	382
250	2021	1534	15

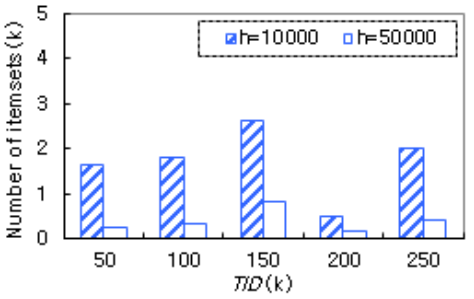

Fig. 5. Experimental results in various h

Fig. 5는 *h* 값 변화(즉, 가중치 감쇠 속도의 차별화)에 따른 탐색 결과 변화를 보여주며, 각 구간이 종료되는 시점에서 탐색되는 고유용 과거 정보의 수를 보여준다. 그림에서 보는 바와 같이 *h* 값이 크게 설정된 경우 탐색되는 고유용 과거 정보가 크게 감소됨을 알 수 있다. *h* 값은 커질수록 가중치 감쇠 속도가 낮아진다. 즉, 발생 시간에 따른 정보의 중요성 차이가 적어진다. 따라서 고유용 과거 정보와 같이 근래 시점에 발생이 적는 항목집합은 탐색될 가능성이 낮아지므로 그림에서와 같은 결과를 얻게 된다. 반면 h 값을 지나치게 작은 값으로 설정하는 경우 감쇠 가중치 감쇠 속도가 크게 증가되어 분석 대상이 되는 웹 클릭 스트림(즉, 사용자 접근 패턴)의 작은 변화에도 지나치게 민감하게 반응할 수 있다.

5. 결론

시간 흐름에 따른 가변성이 큰 일반적인 데이터 스트림과 마찬가지로 웹 클릭 스트림의 경우에도 시간 변화에 따른 구성요소 등의 가변성이 큰 특징을 갖는다. 이러한 특징을 고려하여 하나의 데이터 스트림에서 구성요소의 중요성을 발생 시점에 따라 차별화하기 위한 여러 기법들이 제안되어 왔다. 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 관심도가 큰(혹은 중요한 의미를 갖는) 정보들이 존재하기도 하며, 희소성이나 역사성 측면에서 중요성을 인정받는 경우도 있다. 즉, 과거 발생 정보이나 빈번히 발생되거나 중요한 의미를 갖는 정보를 효과적으로 탐색하는 것은 분석 결과의 효용성 및 활용성을 높이는데 기여할 수 있다.

본 논문에서는 웹 클릭 스트림 분석 결과의 효용성을 높일 수 있는 방법으로 고유용 과거 정보 탐색 기법에 대해 기술하였다. 해당 기법에서는 분석 대상이 되는 웹 클릭 스트림에 대해 정방향 및 역방향 감쇠 기법을 동시에 적용하여 각 항목집합의 출현빈도 수를 관리하며, 각 항목집합의 지지도와 사전에 정의된 지지도 임계값을 비교하여 고유용 과거 정보를 탐색한다. 성능 검증을 통해 과거와 최근 발생 정보간의 변화가 큰 경우에는 더욱 유용한 정보를 효과적으로 탐색할 수 있음을 확인하였으며, 웹 클릭 스트림 처리를 위한 기본 성능을 충분히 만족함을 확인하였다. 즉, 웹 클릭 스트림 뿐만 아니라 데이터 스트림 형태로 정보를 발생시키는 다양한 컴퓨터 응용 환경에서 효과적으로 활용될 수 있을 것으로 판단된다.

한편, 과거 발생 정보의 경우 응용 분야의 특성에 따라 비유용 정보를 포함할 수 있으며, 이 경우 본 논문에서 제안한 고유용 과거 정보 탐색 기법을 적용하여 해당 정보들을 탐색하더라도 활용 가치가 낮은 경우가 된다. 본 논문에서는 고유용 정보 탐색을 위한 기본적인 접근법을 제시하고 과거 발생 정보의 경우에도 손쉽게 탐색할 수 있음을 검증한 것으로서 해당 과거 정보들이 실제 유용한 정보로 활용되기 위해서는 응용 분야의 특성 등이 반영된 분석 후처리 과정 등이 필요하다. 향후 해당 내용에 대한 연구를 통해 흥미롭고 가치있는 결과를 얻을 것으로 판단된다. 더불어, 과거 정보에만 집중된 분석이 아니라 과거 및 근래 발생 정보의 중요성을 응용 분야 특성에 따라 다양하게 차등화 함으로써 보다 흥미로운 분석 결과를 얻는 것에 대한 연구도 흥미로운 주제가 될 것으로 판단된다.

References

[1] L. Chen and Q. Mei, "Mining frequent items in data stream using time fading model," Information Sciences, 257(1), pp. 54-69, 2014.
DOI: http://dx.doi.org/10.1016/j.ins.2013.09.007

[2] B.-E. Shie, P.S. Yu, and 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.
DOI: http://dx.doi.org/10.1016/j.eswa.2012.05.035

[3] C. Zhang, F. Masseglia, and Y. Lechevallier, "The anti-bouncing data stream model for web usage streams," Information Sciences, 278(1), pp. 757-772, 2014.
DOI: http://dx.doi.org/10.1016/j.ins.2014.03.089

[4] T. Yoon and J.-H. Lee, "Adaptive web search based on user web log," Journal of the Korea Academia-Industrial cooperation Society, 15(11), pp. 6856-6862, 2014.
DOI: http://dx.doi.org/10.5762/KAIS.2014.15.11.6856

[5] J.-H. Chang, "Mining interesting sequential pattern with a time-interval constraint for efficient analyzing a web-click stream," Journal of the Korea Industrial Information Systems Research, 16(2), pp. 19-29, 2011.
DOI: http://dx.doi.org/10.9723/jksiis.2011.16.2.019

[6] H.-K. Lee, "A study on web-user clustering algorithm for web personalization," Journal of the Korea Academia-Industrial cooperation Society, 12(5), pp. 2375-2382, 2011.
DOI: http://dx.doi.org/10.5762/KAIS.2011.12.5.2375

[7] 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.
DOI: http://dx.doi.org/10.1016/j.eswa.2011.04.167

[8] 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.
DOI: http://dx.doi.org/10.1016/j.eswa.2007.11.061

[9] H. Chen, L.C. Shu, J. Xia, and Q. Deng, “Mining frequent patterns in a varying-size sliding window of online transactional data streams,” Information Sciences, 215(1), pp. 15-36, 2012.
DOI: http://dx.doi.org/10.1016/j.ins.2012.05.007

[10] 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.
DOI: http://dx.doi.org/10.1109/icde.2008.4497423

[11] 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.
DOI: http://dx.doi.org/10.1016/j.is.2005.04.001

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

장 중 혁(Joong-Hyuk Chang) **[정회원]**

- 1998년 8월 : 연세대학교 대학원 컴퓨터과학과 (공학석사)
- 2005년 8월 : 연세대학교 대학원 컴퓨터과학과 (공학박사)
- 2006년 1월 ~ 2008년 7월 : UIUC, WSU 박사후연구원
- 2008년 9월 ~ 현재 : 대구대학교 컴퓨터IT공학부 교수

<관심분야>
데이터 스트림, 데이터 스트림 마이닝, 데이터 마이닝, 웹 정보 시스템