데이터 스트림에서의 확률기반 빙산 질의 처리

서대홍*, 이원석*
*연세대학교 컴퓨터과학과

e-mail : comafire@database.yonsei.ac.kr
leewo@database.yonsei.ac.kr

Probability-based Iceberg Query Processing Over Data Streams

Dae Hong Seo*, Won Suk Lee *
*Dept. of Computer Science, Yonsei University

요 약

실시간에 지속적으로 방대하게 생성되는 데이터 스트림에서의 데이터 마이닝은 빠른 처리시간 및 낮은 메모리 사용량을 요구한다. 이러한 데이터 스트림에서의 데이터 마이닝은 전체 데이터에 대한 분석 보다는 사용자가 관심을 갖는 영역에 대한 마이닝에 초점이 맞추어져 있어, 사용자 관심영역에 대한 분석 데이터 탐색을 필요로 한다. 이에 본 논문에서는 기존의 분석 데이터 탐색 기법인 빙산 질의 및 상위-k 질의에 대하여 알아보고, 이를 보완하기 위한 확률에 기반한 데이터 탐색법인 확률기반 빙산 질의를 제안한다.

1. 서론

유비쿼터스 환경이 도래함에 따라서 실시간에 지속적으로 방대하게 생성되는 데이터 스트림에서의 데이터 마이닝에 대한 연구가 활발히 진행되고 있다. 데이터 스트림에서의 데이터 마이닝은 빠른 처리시간 및 낮은 메모리 사용량을 요구하기 때문에 데이터 스트림 전체에 대한 마이닝은 이런 요구사항을 만족시키기 힘들다. 또한, 사용자들은 전체 질의 결과 보다는 자신이 정의한 임계 값 이상의 결과에 더욱 관심을 가지기 때문에 데이터 스트림에서의 데이터 마이닝 및 마이닝을 위한 분석 데이터 탐색 논문[2, 3, 4, 5, 6]들은 데이터 스트림 전체에 대한 분석 보다는 사용자 관심 부분을 빙산 질의 및 상위-k 질의 개념을 통한 데이터 탐색을 이용함으로써, 빠른 처리시간 및 낮은 메모리 사용량 요구에 대응하고 있다.

* 본 논문은 2007 년도 정부(과학기술부)의 재원으로 한국과학재단의 국가지정 연구실사업으로 수행된 연구임 (No.M10600000225-06J0000-22510).

하지만, 데이터 스트림의 경우 데이터의 분포를 미리 알 수 없으며, 수시로 데이터의 분포가 변하며, 한번 지나간 데이터는 다시 분석할 수 없기 때문에 사용자 관심 영역을 객관적으로 평가하기 위한 임계 값 선정에 많은 어려움을 겪게 된다. 예를 들어 사용자가 정의한 임계 값 이상인 데이터 항목 집합을 결과로 반환하는 질의인 빙산 질의의 경우 사용자 임계 값이 낮다면, 아주 많은 수의 결과를 반환하게 되어, 데이터 마이닝을 처리하기 위해 느린 처리시간 및 많은 메모리를 요구하게 된다. 그와는 반대로 사용자가 정의한 임계 값이 높다면, 아주 적은 수의 결과를 반환하게 되어, 데이터 마이닝을 하더라도 의미 있는 결과를 도출하기 어려운 상태가 되어 버린다. 이런 상황은 사용자가 정의한 상위 k 개의 데이터를 반환하는 상위-k 질의에서도 나타난다. 데이터의 분포가 비교적 고를 경우 사용자가 정의한 상위 k 개와 나머지 데이터는 큰 차이를 나타내지 않게 되며, 사용자는 정확한

k 의 기준을 잡기가 모호해지기 때문이다.이런 문제점을 해결하기 위하여, 사용자가 통계학적인 기준을 가지고 임계 값을 적용할 수 있도록 확률기반 빙산 질의 방법을 제안하였다.

본 논문의 구성은 다음과 같다. 2 장에서는 앞서 언급한 기존의 사용자 중심적 데이터 추출법인 빙산 질의 및 상위-k 질의에 대하여 알아보고, 3 장에서는 본 논문에서 제안한 통계학적인 기준을 이용한 데이터 추출 법인 확률기반 빙산 질의에 대하여 기술하고, 4 장에서는 데이터 분포 변화를 가진 데이터에서의 각 질의 비교 실험 및 결과에 대하여 논의하고, 마지막으로 5 장에서는 결론을 기술한다.

2. 관련 연구

2.1 빙산 질의

빙산 질의(Iceberg Query)는 대용량 데이터들에 대하여 GROUP-BY 집단 함수 적용 후, 집계 함수의 결과가 사용자가 정의한 임계 값 이상인 결과들을 반환하는 질의이다. 보통 사용자들은 전체의 질의 결과 보다는 특정 임계 값 이상의 결과에 관심을 갖는다는 점에 착안하여 고안된 질의로서 [1]에서 처음으로 제안되었다. 기본적인 빙산 질의 형태를 보면 (그림 1)과 같다.

```
SELECT    d1, d2, .... , dn, COUNT(*)
FROM      R
GROUP BY  d1, d2, ... , dn
HAVING    COUNT(*) >= T
```

(그림 1) 기본적인 빙산 질의 형태

(그림 1)과 같이 테이블 R 과 임계 값 T 가 주어졌을 때, 빙산 질의는 테이블 R 의 속성 d_1, d_2, ..., d_n 각각의 속성 값들에 대한 GROUP-BY 집단 함수 적용 후, 결과 튜플들에 대한 카운트나 합 또는 평균을 계산하는 집계 함수를 수행하고, 결과가 임계 값 T 이상인 데이터들만 반환하게 된다. 데이터 스트림에서 분석 데이터 탐색 및 데이터 마이닝을 수행하는 논문[3, 4, 5]에서 빙산질의 방법을 이용하고 있다.

2.2 상위-k 질의

상위-k 질의(Top-k Query)는 데이터들에 대하여 GROUP-BY 집단 함수 적용 후, 집계 함수의 결과를 정렬하여, 사용자가 정의한 상위 k 개의 결과를 반환하는 질의이다. 기본적인 상위-k 질의 형태를 보면 (그림 2)와 같다.

```
SELECT      d1, d2, .... , dn, COUNT(*)
FROM        R
GROUP BY    d1, d2, ... , dn
ORDER BY    COUNT(*)
STOP AFTER  k
```

(그림 2) 기본적인 상위-k 질의 형태

(그림 2)와 같이 테이블 R 과 임계 값 k 가 주어 졌을 때, 상위-k 질의는 테이블 R 의 속성 d_1, d_2, ..., d_n 각각의 속성 값들에 대한 GROUP-BY 집단 함수 적용 후, 결과 튜플들에 대한 카운트나 합 또는 평균을 계산하는 집계 함수를 수행하고, 결과를 정렬한 후에 상위 k 개의 데이터를 반환하게 된다. 데이터 스트림에서 분석 데이터 탐색 및 데이터 마이닝을 수행하는 논문[5, 6]에서 상위-k 질의 방법을 이용하고 있다.

3. 확률기반 빙산 질의

확률기반 빙산 질의(Probability-based Iceberg Query)는 빙산질의 및 상위-k 질의에서 사용하는 임계 값인 T 나 k 의 결정에 있어, 사용자 관심 영역을 객관적으로 정의하기 쉬운 확률 값 P 를 임계 값으로 사용한다.

확률기반 빙산 질의는 데이터들에 대하여 GROUP-BY 집단 함수 적용 후, 집계 함수의 결과에 대하여 [정의 1] 및 [정의 2]의 방법을 적용한 확률기반 평가함수인 CI 를 적용하여 임계 값인 확률 P 이상의 확률을 가지는 결과를 반환 하게 된다. 확률기반 평가함수에서 적용된 [정의 1]및 [정의 2]는 다음과 같다.

[정의 1] 이산형 확률분포

이산형 확률분포(Discrete probability distribution)[8]는 확률변수로 가능한 모든 값 각각에 확률을 계산한 것을 말하며, (식 1)의 조건을 만족해야 한다.

$$0 \le P(x) \le 1, \quad \sum P(x) = 1$$

$P(x)$: 확률 변수 x 에 대한 확률

(식 1) 이산형 확률분포 조건

각 확률변수의 확률은 0 과 1 사이에 존재해야 하며,전체 확률변수의 합은 1 이 되어야 한다. 이러한 이산형 확률분포에 대한 평균인 μ 그리고 표준편차인 σ 는 (식 2)를 통하여 계산이 가능하다.

$$\mu = \sum xP(x), \quad \sigma = \sqrt{\sum (x-\mu)^2 P(x)}$$

(식 2) 이산형 확률분포의 평균과 표준편차

(식 2)에서 평균 μ 는 각 확률변수와 확률변수의 곱의 합으로 나타내어지고, 표준편차 σ 는 구해진 평균과 확률변수의 차를 제곱하고, 확률변수의 확률을 곱한 것의 합에 대한 제곱근이다.

[정의 2] 체비쇼프의 부등식

체비쇼프 부등식(Chebyshev's inequality)[9]은 체비쇼프에 의하여 증명된 부등식으로 확률변수 x 에 대하여 x 의 평균을 μ , 표준편차를 σ 라 하면, 이는 임의의 양수 k 에 대하여 다음 (식 3)의 부등식을 만족하게 된다.

$$P(|x-\mu| \ge k\sigma) \le \frac{1}{k^2}$$

(식 3) 체비쇼프 부등식

즉, 자료가 평균으로부터 $k\sigma$ 이상 떨어질 확률은

$1/k^2$ 보다 적다는 것이다. 그러므로, 자료가 평균으로부터 $k\sigma$ 이내의 범위 $\mu - k\sigma < x < \mu + k\sigma$ 내에 있을 확률은 최소한 $1-1/k^2$ 보다 크게 된다.

[정의 1]과 [정의 2]를 적용한 확률기반 평가 함수인 CI 의 값이 임계 값인 확률 P 이상인 결과를 반환하는 확률기반 빙산 질의의 기본적인 형태를 (그림 3)과 같이 정의 한다.

SELECT	d_1, d_2, ..., d_n, *COUNT*(*)
FROM	R
GROUP BY	d_1, d_2, ..., d_n
HAVING	*CI(COUNT(*))* >= P

(그림 3) 기본적인 확률기반 빙산 질의 형태

확률기반 빙산 질의는 (그림 3)과 같이 테이블 R 과 임계 값 확률 P 가 주어졌을 때 테이블 R 의 속성 d_1, d_2, ..., d_n 각각의 속성 값들에 대한 GROUP-BY 집단함수 적용 후, 결과 튜플들에 대한 카운트를 [정의 1]및 [정의 2]를 적용한 CI 함수를 통하여 계산한 확률이 P 이상 되는 범위의 결과를 반환한다.

(그림 3)에서 주어진 테이블 R 에서 GROUP-BY 집단 함수의 결과로 생성된 튜플들은 확률 변수로 재정의 될 수 있으며, 각 확률 변수에 대한 카운트를 이용하여, 확률 변수에 대한 확률을 생성할 수 있게 된다. 예를 들어, 테이블 R 에 대하여 d_1, d_2, d_3 의 속성들이 각 v_1, v_2 의 속성 값을 갖는 다면, d_1, d_2, d_3 의 속성에 대한 GROUP-BY 집단 함수의 카운트 결과를 확률 변수 x 로 재 정의하여 변환한 테이블은 (표 1)과 같다.

(표 1) GROUP-BY 결과에 대한 확률변수 적용

d_1	d_2	d_3	COUNT	x	P(x)
v_1	v_1	v_1	2	1	0.02
v_1	v_1	v_2	4	2	0.04
v_1	v_2	v_1	17	3	0.17
v_1	v_2	v_2	26	4	0.26
v_2	v_1	v_1	28	5	0.28
v_2	v_1	v_2	15	6	0.15
v_2	v_2	v_1	5	7	0.05
v_2	v_2	v_2	3	8	0.03
SUM			100	SUM	1

(표 1)에서 확률변수 x 로 재 정의 된 결과 튜플들의 확률 값은 0 보다 크고 1 보다 작으며, 전체 확률의 합은 1 과 같다. 따라서, (표 1)은 [정의 1]의 조건 (식 1)을 만족하게 되므로, 테이블 R 은 이산형 확률분포 이며, [정의 1]의 (식 2)를 이용하여 확률변수 x 에 대한 평균 μ 및 표준편차 σ 를 구할 수 있으며, 계산한 확률변수 x 의 평균 μ 및 표준편차 σ 를 [정의 2]에 적용하여 임계 값 확률 P 이상의 확률을 가지는 확률변수 x 의 범위를 계산 할 수 있다.

[정의 2]의 (식 3)을 계산하기 위하여 필요한 양의 정수 k 값은 임계 값인 확률 P 와 같으므로 $P = 1 - 1/k^2$ 식을 k 값에 대하여 풀이한 (식 4)를 통해 계산할 수 있다.

$$k = \sqrt{\frac{1}{1-P}}$$

(식 4) [정의 2]의 k 값 계산

(표 1)에 대하여 임계 값인 확률 P 가 0.7 이상인 확률기반 질의를 한다면, [정의 1]의 (식 2)를 이용하여 계산한 평균 μ 는 4.54 이고 표준편차 σ 는 1.42 이다. 또한, [정의 2]의 (식 3)에 필요한 양의 정수 k 값은 (식 4)를 이용하여 1.58 로 계산된다. 지금까지 구한 (표 1)의 확률 변수 x 의 평균 μ 및 표준편차 σ 그리고 양의 정수 k 값을 [정의 2]에 적용하여 확률이 임계 값인 확률 P 이상을 가지는 확률변수 x 의 범위를 구하면, 확률변수 x 의 범위는 $2.28 < x < 6.79$ 이 된다. 이를 (표 1)에 적용한 확률기반 빙산 질의의 결과는 (표 2)와 같다.

(표 2) P = 0.7 인 확률기반 빙산 질의 결과

d_1	d_2	d_3	COUNT
v_1	v_2	v_1	17
v_1	v_2	v_2	26
v_2	v_1	v_1	28
v_2	v_1	v_2	15

(표 2)는 (표 1)의 d_1, d_2, d_3 의 속성에 대한 GROUP-BY 집단 함수의 결과 튜플에 대한 카운트 집계 결과에 대하여 적어도 0.7 이상의 확률을 가지게 된다.

4. 실험 결과

데이터는 지프 분포(Zipf Distribution)[7]을 기반으로 생성하였다. 지프 분포는 (식 5)를 만족하는 분포로서 i 개의 이산형 변수에 대하여, 각 변수의 빈도수 f 는 α 값에 따라 변하게 된다. 지프 분포에서는 α 값이 커질수록 데이터는 더욱 편향된 분포를 가지게 된다.

$$f_i \propto \frac{1}{i^{\alpha}} \qquad (i=1,\ldots,N)$$

(식 5) 지프 분포(Zipf Distribution)

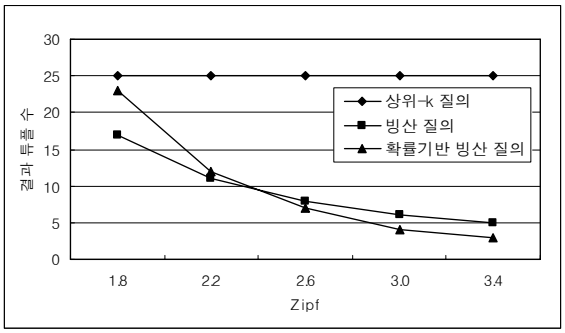
지프 분포는 인구수, 단어사용의 빈도수, 책의 판매 등, 많은 실 세계 데이터들의 모델링이 가능하다. 본 실험에서는 (표 3)의 인자를 조합하여 지프 분포를 가지는 데이터를 랜덤 생성하여 사용하였다.

(표 3) 지프 분포 데이터 생성 인자

Zipf (Zipf Distribution)	RV (Random Variable)	T (Tuple)
1.8~3.4	100~2500	100000~500000

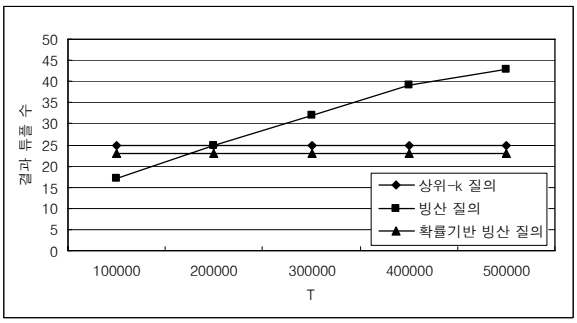
실험은 각 지프 분포, 확률변수, 튜플의 변화에 따른

결과 튜플수의 변화에 대하여 실험하였다. 각 질의에 대하여 임계 값은 상위-k 질의는 25, 빙산 질의는 300, 확률기반 빙산 질의는 0.75 로 하였다.



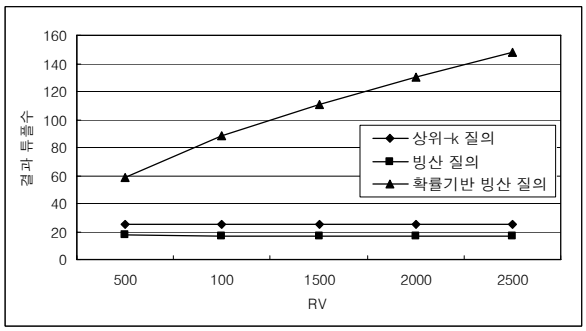
(그림 4) 지프 분포에 따른 결과 튜플 수 변화

(그림 4)는 지프 분포에 따른 결과 튜플 수 변화를 보인다. (그림 4)에서 지프 분포에 따라 데이터의 분포가 한 곳으로 몰릴 때, 상위-k 질의의 경우 데이터가 데이터 분포 특성을 반영하지 못하는 것을 볼 수 있다.



(그림 5) 튜플 수 변화에 따른 결과 튜플 수 변화

(그림 5)는 데이터 스트림에서 들어온 튜플 수의 변화에 따른 각 질의에 대한 결과 튜플 수에 대한 실험이다. (그림 5)에서 빙산 질의의 경우, 튜플 수가 증가함에 따라 결과 튜플 수가 늘어나는 결과를 보이고 있다. 데이터 스트림의 특성상 지속적으로 방대하게 생성된다는 특성을 볼 때, 빙산 질의는 데이터 스트림에 적합하지 않다는 것을 알 수 있다.



(그림 6) 확률변수 변화에 따른 결과 튜플 수 변화

(그림 6)은 확률변수로 재 정의된 GROUP-BY 튜플 수의 변화에 따른 결과 튜플 수의 변화 실험이다. (그림 6)에서 확률변수로 재 정의된 GROUP-BY 튜플 수가 증가했음에도 불구하고, 상위-k 질의 및 빙산 질의는 GROUP-BY 튜플 수 증가에 따른 데이터의 특성을 반영하지 못하고 있다.

5. 결론

실시간에 지속적으로 방대하게 생성되는 데이터 스트림의 특성상 데이터의 분포를 미리 알 수 없으며, 수시로 데이터의 분포가 변하고, 한번 지나간 데이터는 다시 분석할 수 없기 때문에 전체 데이터에 대한 마이닝 보다는 사용자 관심영역을 객관적으로 탐색하는 것이 필요하다. 본 논문에서는 기존의 상위-k 질의나 빙산 질의 방법에서는 만족 시켜 줄 수 없었던, 객관적 확률적 기준을 이용한 데이터 탐색법인 확률기반 빙산 질의를 제안하였다. 또한, 비교 실험을 통하여 데이터 스트림에서의 확률기반 빙산 질의의 유용성을 입증하였다.

참고문헌

[1] Min Fang, Narayanan Shivakumar, Hector Garcia-Molina, Rajeev Motwani, Jeffrey D. Ullman *"Computing Iceberg Queries Efficiently"* Proceeding of the 24th VLDB Conference New York, USA, 1998

[2]Raymond Chi-Wing Wong and Ada Wai-Chee Fu *"Mining top-K frequent itemsets from data streams"*, Springer Netherlands, Volume 13, Number 2 / September, 2006

[3] Ahmed Metwally, Divyakant Agrawal, Amr El Abbadi, *"Efficient Computation of Frequent and Top-k Elements in Data Streams"*, ICDT 2005: 398-412

[4] Wong, R.C.-W. and Fu, A.W.-C. 2005b. *"Mining top-K itemsets over a sliding window based on zipfian Distribution"*. In SIAM International Conference on Data Mining.

[5] C. Silvestri and S. Orlando. *"Approximate Mining of Frequent Patterns on Streams."*, Proceedings of the Second International Workshop on Knowledge Discovery in Data Streams in conjunction with 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), 2005, Porto, Portogallo.

[6] G S Manku and R Motwani *"Approximate Frequency Counts over Data Streams"* VLDB 2002 (28th VLDB), p 346-357, August 2002

[7] M.E.J. NEWMAN, *"Power laws, Pareto distributions and Zipf's law"* Contemporary Physics, Vol. 46, No. 5, September–October 2005, 323 – 351

[8]Wikipedia The Free Encyclopedia, Discrete probability distribution,http://en.wikipedia.org/wiki/Discrete_probability_distribution

[9]Wikipedia The Free Encyclopedia, Chebyshev's inequality http://en.wikipedia.org/wiki/Chebyshev_inequality