공통 조인 작업 공유를 통한 다중 연속 질의 처리*

박홍규, 이원석
연세대학교 컴퓨터과학과

Processing Multiple Continuous Queries by sharing common join operations

Hong Kyu Park, Won Suk Lee
Dept. of Computer Science, Yonsei University
E-mail : {gladiator11, leewo}@database.yonsei.ac.kr

요 약

데이터 스트림이란 제한 없이 끊임없이 흘러 들어오는 일련의 많은 양의 데이터 객체들을 의미하며, 센서 데이터 처리, 인터넷 트래픽 분석, 웹 서버 로그와 같은 다양한 트랜잭션 로그 분석등과 관련된 수많은 응용 분야에 적용 가능하기 때문에 이들을 처리 하기 위해 많은 연구가 진행되었다. 데이트 스트림을 처리하기 위해서는 미리 등록된 질의들(연속 질의)을 새롭게 들어오는 스트림 데이터들로 계산하여 그 결과를 계속적으로 생성하여야 하므로 연속 질의들은 스트림 데이터가 들어올 때마다 반복적으로 수행되며, 데이터 스트림은 매우 빠르게 입력되는 특성을 가지고 있기 때문에 보다 빠르게 질의를 처리하여야만 한다. 본 논문에서는 다수의 조인 연속 질의들이 시스템에 등록되어 있을 때, 이들을 보다 빠르게 처리할 수 있도록 여러 개의 질의에 반복적으로 적용되는 조인 연산들을 공유함으로써 최적의 질의 계획을 생성하는 기법을 제안한다.

1. 서 론*

최근 산업발달과 고도 사회로 접어들면서 과학 기술이나 공학 분야 이외에 경제 사회 등의 다양한 분야에서도 각종 데이터들의 중요성이 강조되고 있으며, 데이터의 관리방식도 일정한 저장공간에 데이터를 저장한 후 관리·활용하던 체계에서 데이터 종류의 다양화, 용량의 대형화, 그리고 연속적인 발생특성으로 인한 한정된 공간을 효율적으로 처리하는 방식으로 점차 변화하고 있다. 이러한 종류의 데이터는 유비쿼터스 센서 네트웍, 주식거래, 금융거래, 전화통화, 교통관리 및 교통상황수집자료 등에서 찾아볼 수 있으며, 실시간에 연속적으로 발생되어 빠른 속도로 관리 시스템에 입력되어 들어오는 이러한 데이터를 데이터 스트림(Data Stream)라 한다. 데이터 스트림의 경우 시스템 입력 속도가 매우 빠르고 연속적이어서 일반적으로 데이터 처리에 사용되는 DBMS 에서 데이터를 저장한 후 처리하는 방식을 적용할 수 없다. 이미 많은 연구자들이 오랜 동안 데이터 스트림의 처리기술을 연구해 오고 있으며 특히 데이터 처리의 효율성을 높이기 위하여 질의 최적화 연구가 활발히 이루어지고 있다.

본 논문에서는 지역 최적화 방법을 기초로 하지만, 다중 질의에서의 공유가 최대한 많이 일어날 수 있도록 하는 다중 그리디 최적화 방법(*Multiple*

* 이 논문은 2008 년도 정부(교육과학기술부)의 재원으로 한국과학재단의 국가지정연구실사업으로 수행된 연구임 (No.R0A-2006-000-10225-0)

Greedy)을 제안한다. 즉, 여러 개의 질의에 동일한 조인 조건들이 있을 때 공유 활용하게 함으로써 불필요하게 반복수행 될 수 있는 질의 수행과정을 최소화시켜 연산처리의 효율성을 높이도록 하였다. 즉, 사전에 공통적으로 사용되는 조인조건들을 검색해 질의 구성 시 조건들을 공유할 수 있도록 질의 조건 가운데 공유가 가능한 부분을 우선적으로 탐색·수행시킴으로써 조인 공유율을 최대화하여, 결과적으로는 질의 수행 비용을 최소로 감소 시킬 수 있는 질의 수행 최적화 기법을 제안하고자 한다.

2. 관련 연구

데이터 스트림 처리에 관한 연구는 최근 몇 십년 간 지속적으로 수행되어 왔으며, 다양한 연구성과를 거두고 있는데, Aurora[2], Telegraph[17], STREAM[18] 같은 스트림 쿼리 프로세싱을 수행할 수 있는 시스템들이 개발 되어 흥미로운 연구 결과들을 보여주었다. 그리고 최근에는 대용량 스트림 데이터의 증가에 따라 센서 모니터링 데이터와 같이 상당한 수준의 연속질의가 연속적으로 처리되어야 하는 데이터 처리시스템에서는 응용 프로그램들이 한꺼번에 대량의 데이터 스트림을 처리할 수 있는 능력이 추가적으로 요구 된다. 또한, CACQ[3][18], PSoup[19]에서는 데이터 스트림을 처리하기 위한 다중 연속질의를 공유할 수 있는 방법을 제안하였는데, 두 가지 모두 가장 큰 윈도우 사이즈를 사용하여 다중 질의에서의 공유 윈도우 조인을 수행한다. 다중 연속질의에서 연산 공유의 문제는 더 이상 새로운 연구 대상은 아니다. 이미 전통적인 관계형 데이터베이스분야의 연구에서도 최적화된 공유 질의 계획을 찾는 연구들이 수행된바 있다[22]. 최근의 다중 연속질의에 대한 연구에서 Mistry[7], Prasan[14] 등은 질의 집합에서 최적화 계획을 찾는 공간을 줄이기 위한 방법으로 휴리스틱 방법을 사용하였다. 그리고 Hammad[12], Wang[16] 등은 연속질의에서 연산 공유에 초점을 맞춘 연구를 수행 하였으며, Wang 은 서로 다른 크기의 윈도우를 가진 조인을 공유하여 수립 할 수 있는 다양한 질의 계획들을 제안한바 있다.

3. 공유 기반의 그리디 최적화 기법

데이터 스트림 처리에 관련된 연산은 여러 가지가 있으나 그 가운데 시스템에 가장 큰 부하를 주는 연산은 조인 연산이다. 예를 들어 다중 연속 질의에서 2개 이상의 조인 연산이 동시에 수행되는 경우, 탐색 영역의 크기는 질의 숫자가 늘어날 때 마다 제곱이 되므로 연산처리를 위한 시스템부하는 배가 될 것이다. 그러므로 연속 질의 조인연산 수를 최소화 시키는 것은 조인 비용을 감소시키는 데 매우 유효한다.

다중 질의에 사용 된 조인 조건 가운데 공통으로 사용된 조인 조건을 찾아 다중 연속 질의의 전체 질의 수행 계획 수립 시 조인 연산을 개별적으로 실행 하지 않고 공통된 조인 조건의 일회 실행을 통해 생성되는 중간 연산 결과를 재사용할 수 있다. 따라서 조인연산의 실행 횟수를 줄일 수 있다. 중간 결과를 공유해 조인 연산을 수행하기 위해서는 몇 가지 정의사항들이 필요하며, 이를 정의하면 다음의 2 가지와 같다.

정의 1. 공유 조인

다수의 질의에서 공통적으로 적용되는 조인 조건을 수행하는 조인 연산자를 공유 가능한 조인, **공유 조인**이라고 한다.

정의 2. 공유수(Task-Number)

공유 조인이 n 개의 질의에서 사용되어 공유할 수 있는 횟수를 공유수라 한다.

본 연구에서 조인 비용은 2 개의 스트림을 조인하는 조인에 대한 단위 시간당 처리해야 할 일로서 정의한다. 다음 <표 1>는 조인 비용 모델을 정의하기 위한 용어들을 정리한 표로서 <표 1>에서 Kang 이 사용한 용어를 기초로 하여 다중 조인 질의 비용을 계산 위한 과정에 필요한 기호들을 들을 추가하여 정의 하였다.

Kang[3]의 연구에서는 스트림 데이터를 삽입하는 연산과 삭제하는 연산은 조인결과를 생성하는 연산에 비해 비용발생 규모가 상대적으로 매우 작은 것으로 평가했다.

<표 1> 다중 조인 질의의 비용계산 관련 용어

λ_R, λ_S	스트림 R과 S의 입력속도
W	윈도우의 크기
$\|R\|$	윈도우 R의 해시 버켓의 수
$\triangle R$	스트림 R에서 새로 들어온 데이터
$\sigma R,S$	스트림 R과 S의 조인 선택도
$Task_R \bowtie_S$	스트림 R과 S 조인의 공유수
$C_i(R,S)$	새로운 스트림이 들어왔을 때 필요한 연산 비용
$C_o(R,S)$	조인된 결과를 처리하는 비용
$C(R,S)$	스트림 R과 S의 조인 비용

본 논문에서도 삽입과 삭제 연산에 대해서는 계산과정의 효율성을 위해 비용산출에 고려하지 않고 상대 스트림을 조사하는 연산만을 고려하였다. 따라서 두 개의 스트림 R 과 S 의 조인비용 계산식은 식(1)와 같이 정의 내려질 수 있다.

$$C_i(R,S) = \lambda_R \times \frac{W \times \lambda_S}{|S|} + \lambda_S \times \frac{W \times \lambda_R}{|R|} \quad \cdots\cdots (1)$$

위 식(1)의 우변은 각각 스트림 R 에서 새로운 데이터가 들어왔을 때 스트림 S 를 조사하는 비용과 스트림 S 에서 새로운 데이터가 들어왔을 때 스트림 R 을 조사하는 비용의 합을 의미한다. 그리고 조인을 수행하면서 생성된 결과들을 다음 조인들을 위해 저장되어야 하기 때문에 그 비용도 고려되어야 하는데 이는 스트림들의 입력속도와 조인 선택도를 이용해 구할 수 있다. 조인 선택도란 단위 시간당의 조인의 카티션프러덕트(Cartesian Product)와 실제로 생성된 조인 결과의 크기로 결정된다

$C_o(R,S)$는 조인된 결과의 처리 비용이며, 조인 결과의 크기, 즉 조인 결과의 개수와 같으며 식(2)와 같다.

$$C_o(R,S) = \lambda_R \lambda_S (2W-1)\sigma_{R,S} \quad \cdots (2)$$

따라서, 스트림 R 과 스트림 S 를 조인하는 글로벌조인 비용은 식(1)와 식 (2)의 합으로 구할 수 있으며, 식(3)와 같다.

$$C(R,S) = C_i(R,S) + C_o(R,S) = \lambda_R \lambda_S W\left(\frac{|R|+|S|}{|R||S|}\right) + \lambda_R \lambda_S (2W-1)\sigma_{R,S} \quad (3)$$

식(3)에서 구한 조인비용에 본 연구에서 제안하는 방법을 적용시키기 위해서는 다중 질의 중에서 조인의 공유가 많이 되는 조인이 수행 될 수 있도록 하여야 한다. 따라서 조인 비용에 조인 연산을 공유할 수 있도록 조인 비용에 공유수를 반영하도록 하였다.

$$\frac{C(R,S)}{Task(R \bowtie S)} \quad \cdots\cdots (4)$$

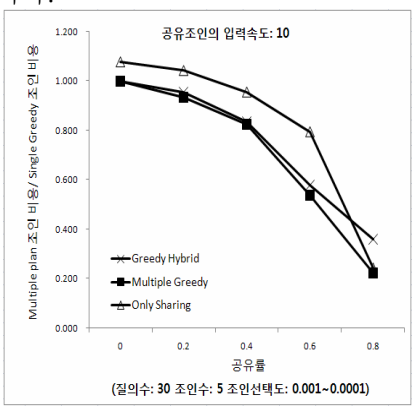
앞에서도 언급했듯이 다중 그리디 최적화 방법(*Multiple Greedy*)은 지역 최적화 방법을 토대로 하지만 다중 질의에서 공유가 최대한 많이 될 수 있도록 실행 계획을 생성하는 방법이다. 이를 하기 위하여 최적화를 수행하기 전에 우선 모든 질의에 대한 정보를 파싱하여 어떤 조인 조건들이 몇 개의 질의에서 사용되었는지를 파악한다. 파악된 정보를 이용하여 식(4)와 같이 공유가 발생하는 조인들은 조인 비용을 사용되는 질의의 개수로 나눔으로써, 실제 해당 조인을 수행할 때 드는 비용보다 작은 비용으로 책정되도록 만든다. 그리고 난 후, 지역 최적화 방법을 그리디 방법을 이용하여 각 질의 실행 계획들을 생성한 후에 각 실행 계획들을 합치는 방법이다. 이 방법은 일반적인 그리디 알고리즘을 질의 최적화 과정에 적용시킨 것이며, 공유가 되는 조인들의 비용들은 낮춘 후에 가장 비용이 작게 소요되는 조인을 선택하도록 함으로써 공유가 될 수 있는 조인들이 먼저 수행될 수 있도록 최적화된 질의 수행 계획을 생성한다. 이러한 최적화 과정은 다중 연속 질의 내 모든 조인이 수행될 때까지 계속 반복돼 최종 다중 연속 질의의 질의 수행 계획을 생성하게 된다.

4. 실험

제안된 알고리즘의 성능을 평가하기 위하여 다양한 스트림 환경을 시뮬레이션하여 실험하였으며, 비교 대상으로는 [11]에서 제안한 방법과 그리디 지역 최적화 기법(Greedy Hybrid)을 사용하였다. [11]에서 사용된 방법(Only_Sharing)은 조인 비용과는 상관 없이, 다중 질의에서 공유가 가장 많이 될 수 있는 조인 연산들을 우선적으로 선택하여 최적화를 하는 방법이며, 그리디 지역 최적화 방법은 각 질의들의 최적 실행 계획을 각각 그리디 방법을 적용하여 생성한 후, 각 질의들의 실행 계획을 합치는 방법이다. 조인 조건의 공유 효과 결과인 조인 비용 감소를 다중 질의 계획의 조인 비용에 반영되게 하기 위해서는 질의 셋에 있는 조인 조건의 공유율을 조절할 수 있어야 한다. 공유율은 다음 식(5)을 통해 계산할 수 있다.

$$\frac{Task\_Join(R \bowtie S)}{\sum_{i=1}^{n} C_{(Join)i}} \quad \cdots (5)$$

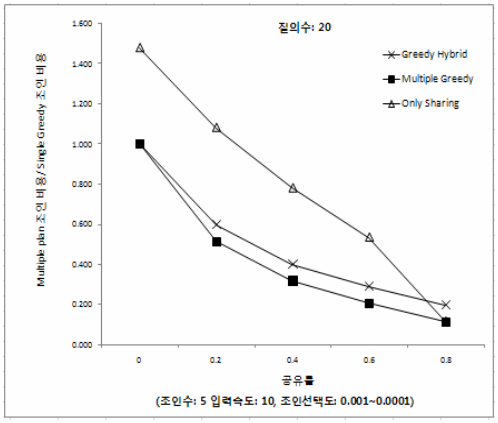
여기서, $Task\_Join(R \bowtie S)$ 는 스트림 R과 S의 조인이 다중 연속 질의에 사용된 총 횟수이고, $\sum_{i=1}^{n} C_{(Join)i}$ 는 다중 연속 질의내의 모든 조인 조건개수 이다.



<그림 1> 공유율 변화에 따른 조인 비용 비교

<그림 1>을 보면 왼쪽 실험은 공유 조인의 입력 스트림의 입력속도를 공유되지 않는 조인의 입력 스트림의 입력속도 보다 작거나 같게 하였다. 공유 조인의 입력 스트림의 입력 속도가 공유되지 않는 조인의 입력 스트림의 입력 속도 보다 작거나 같기 때문에 그리디 하이브리드 알고리즘과 다중 그리디 알고리즘 모두 공유

효과를 보았다.



<그림 2> 공유율에 따른 조인 비용 비교

그림 2은 질의수가 변화 하여도 공유율에 따른 질의 계획의 조인 비용은 같은 비율로 변화함을 보여준다.

5. 결론

최근 데이터 베이스 연구는 한정된 데이터 보다는 실시간으로 무한히 흘러 들어오는 데이터 스트림을 처리하는 연구가 활발히 진행 되고 있다. 데이터 스트림은 연속성, 시변성, 데이터의 무한 순서라는 특징을 가지고 있다. 데이터 스트림을 처리하는 질의 중에서 조인 조건을 갖는 질의를 처리하는 것은 다른 연산에 비해 복잡하고 많은 비용이 필요한 연산이다. 어떤 순서로 조인을 하느냐에 따라 데이터의 처리 비용과 수행 시간이 달라질 수 있기 때문에 조인 질의의 최적화가 필요하다. 본 논문은 공유가 많이 되는 조인 조건을 질의 계획에 우선적으로 선택될 확률을 높이기 위해 다중 연속 질의 내의 모든 조인의 비용에 공유율이 반영 될 수 있도록 기존 연구에서 제시한 조인 비용에 공유 수를 나누어 주는 방법을 제시하고 효용성을 검증하였다.

[참고 문헌]

[1] Babu S. and Widom J., Continuous Queries Over Data Streams, ACM SIGMOD Record archive volume 30, Issue 3, 2001, 109-120.

[2] B. Babcock, Babu S., M. Datar, R. Motwani,and Widom J., Models and Issue in Data Stream Systems, In Proc. of PODS, 2002.

[3] J. Kang, J. F. Naughton, and S. Viglas, Evaluating window joins over unbounded streams, In Proc. of the 2003 Intl. Conf. on Data Engineering, Mar. 2003.

[4] Lukasz Golab M. Tamer Ozsu, Processing Sliding Window Multi-Joins in Continuous Queries over Data Streams, Proc. VLDB, 2003

[5] Moustafa A. Hammad, Michael J. Franklin, Walid G. Aref, Ahmed K. Elmagarmid, Scheduling for shared window joins over datastreams, Proc. VLDB, 2003.

[6] M. H. Ali, W. G. Aref, R. Bose, A. K. Elmagarmid, A. Helal, I. Kamel, and M. F. Mokbel. NILE-PDT:A phenomenon detection and tracking framework for data stream management systems. In VLDB, p1295-1298, 2005

[7] Song Wang, Elke Rundensteiner and Ganguly S., Bhatnagar S., StateSlice:New Paradigm of Multi-query Optimization of Window-based Stream Quaries, Proc. VLDB, 2006.

[8] The STREAM groups STREAM : The Stanford Stream Data Manager IEEE Data Engineering Bulletin, March 2003.

[9] TIMOS K.SELLIS, Multiple-Query Optimization, ACM Transactions on Database Systems, 1988.

[10] Yali Zhu, Elke A Rundensteiner and George T. Heineman, Dynamic Plan Migration for Continuous Queries Over Data Streams, ACM, SIGMOD June 13-18, 2004.

[11] 박연경, 이원석, 데이터 스트림에서 다중 조인 질의의 최적화 기법, 한국정보처리학회 춘계학술발표대회 논문집 제 14 권 제 1 호 (2007. 5)