센서네트워크 데이터베이스를 위한 새로운 조인 연산자 정의*

이승재, 김창화, 김상경
강릉대학교 컴퓨터공학과
e-mail : silveree, kch, skkim98@kangnung.ac.kr

A New Join Operator Definition for Sensor Network Databases

Seungjae Lee, Changhwa Kim, Sangkyung Kim
Dept of Computer Science & Engineering,
Kangnung National University

요 약

최근 센서네트워크에서 수집되는 방대한 양의 데이터를 효율적으로 처리하기 위하여 관계형 데이터베이스를 이용한 센서네트워크가 활발히 연구되고 있다. 센서네트워크에서는 제한된 에너지를 사용한다는 점, 스트림 데이터를 처리할 수 있어야 한다는 점 등에서 기존 데이터베이스와는 다른 연구가 필요하다. 정확히 일치하는 키 값에 대하여만 조인이 발생하는 조인연산 또한 센서네트워크에서 사용하기 위해서는 새로운 정의가 필요하다. 온도센서와 습도센서가 일정영역에 무작위로 뿌려져 있는 센서네트워크를 가정해 보자. 데이터베이스 관점에서는 온도릴레이션과 습도릴레이션이 존재하게 된다. 이 때 위치에 따른 온도와 습도의 상관관계를 얻기 위하여 좌표를 키 값으로 하여 릴레이션을 조인하면 결과는 공집합이거나 아주 적은 수의 튜플만 얻게 되어 사용자가 원하는 결과를 얻을 수 없다. 그 이유는 동일한 좌표를 가지는 서로 다른 종류의 센서쌍이 존재할 확률이 매우 적기 때문이다. 본 논문에서는 이러한 문제를 해결하기 위하여 새로운 범위조인연산자를 제안한다. 이 범위조인연산자를 센서네트워크에 적용하면 좀 더 효율적인 데이터관리가 가능하고 데이터베이스에서 응용계층에 표준화된 인터페이스를 제공할 수 있다.

키워드 : 센서네트워크, 데이터베이스, 조인, 범위조인

1. 서론

센서네트워크의 응용범위가 점점 넓어지면서, 수집되는 방대한 양의 데이터를 효율적으로 처리 및 관리하기 위하여 센서네트워크에 데이터베이스 개념을 이용하려는 시도가 늘고 있다[1,2,3]. 그러나 이와 같이 기존의 데이터베이스를 직접적으로 센서네트워크에 적용할 경우 센서네트워크에 대한 요구사항을 충족할 수 없는 상황이 발생한다.

예를 들어 온도센서와 습도센서가 일정 영역 내에 무작위로 뿌려져 있는 센서네트워크가 있다고 가정하자. 이 때 온도센서들은 온도릴레이션을 구성하고 습도센서는 습도릴레이션을 구성한다. 데이터베이스 관점에서 이 센서네트워크를 이용하여 위치에 따른 온도와 습도의 상관관계는 두 릴레이션을 조인연산함으로써 얻을 수 있다. 하지만 동일한 위치에 서로 다른 종류의 센서가 위치할 가능성의 매우 낮으므로 결과는 공집합이거나 매우 적은 수의 튜플이 될 것이고 이 결과는 사용자의 의도와는 동떨어진 것이다.

이러한 문제는 응용계층에서 해결할 수 있으나 이는 응용프로그램을 복잡하게 하며 표준 인터페이스를 제공하지 못한다.

본 논문에서는 이러한 문제점을 해결하기 위하여 일정 범위 내의 키 값에 대하여도 조인이 가능한 새로운 범위조인연산자를 제안한다. 범위조인연산자가 도입된 데이터베이스를 사용하면 목적 달성을 위한 유연한 조인연산이 가능하고 응용계층에 표준화된 인터페이스를 제공할 수 있다.

* 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음 (IITA-2006-C1090-0603-0044)

본 논문의 구성은 다음과 같다. 제 2 장에서는 관련연구를 살펴보고, 제 3 장에서는 범위조인연산자를 자세히 소개하며 제 4 장에서 결론을 맺으며 본 논문을 마친다.

2. 관련 연구

센서네트워크는 방대한 양의 데이터를 다루는 특성상 효율적인 데이터 관리를 위하여 데이터베이스 시스템의 사용이 필요하며 관계형 데이터베이스가 가장 널리 사용되고 있다. 센서네트워크는 그 자체를 센서넷 데이터베이스라 불리는 하나의 데이터베이스로 생각할 수 있다. 센서넷 데이터베이스에서 각각의 센서가 센싱하는 센싱값과 시간, 위치 등은 하나의 튜플이 되고, 동일한 종류의 센서가 센싱한 값들의 집합은 릴레이션에서 하나의 속성값에 해당되며, 동일한 센서들이 생성한 튜플들의 집합이 하나의 릴레이션을 구성한다[3].

센서넷 데이터베이스는 전통적인 데이터베이스와 비교하여 연산이 네트워크 내부에서 수행된다는 점, 사용할 수 있는 에너지가 제한적이라는 점 등 여러 가지 차이점을 가지고 있으며 이런 차이점을 극복하기 위한 여러 연구가 진행되고 있다. 특히 센서넷 데이터베이스의 연산자와 관련하여 현재 진행되고 있는 연구들은 다음 두 가지로 나눌 수 있다.

첫 번째는 집계연산 시 센서네트워크 및 센서노드의 에너지 소모량을 줄이는 방법에 관한 연구이다. MIN, MAX, AVERAGE 등과 같은 집계연산 시 연관된 모든 센서노드로부터 센싱값을 요구하지 않고 일부 센서노드에만 센싱값을 요구하거나 요구 횟수를 줄여 전달받은 센싱값을 이용하여 근사한 결과를 계산한다. 이러한 방법은 집계연산 시 통신에 소모되는 에너지를 줄여 네트워크 전체의 생존 시간을 늘릴 수 있다[4,5,6,7,8,9].

두 번째는 스트림 데이터에 대한 질의 처리 방법에 대한 연구이다. 고전적인 데이터베이스에서는 질의가 수행될 시점의 릴레이션 스냅샷을 대상으로 질의를 수행하며 결과 또한 정적인 릴레이션이다. 하지만 센서넷 데이터베이스의 경우 센서노드가 지속적으로 센싱을 수행하면 릴레이션에 튜플이 계속해서 추가되며 이러한 동적인 릴레이션에 대한 연산은 매우 빈번하게 요구된다. 즉 센서넷 데이터베이스의 연산자는 블록킹되지 않은 릴레이션에 대한 연산도 수행할 수 있어야 한다. 물론 이 연산의 결과 또한 블록킹되지 않은 릴레이션이 된다. 이러한 블록킹되지 않은 릴레이션과 이러한 릴레이션의 조인 문제를 해결하기 위한 다양한 연구가 있다[10,11,12,13].

3. 범위조인연산자

본 논문에서 제안하는 범위조인연산자는 자연조인연산자와 매우 유사하며 크게 세 가지 측면에서 차이가 난다. 첫 번째는 조인속성의 도메인이 실수의 부분집합이어야 한다. 두 번째는 조인속성집합은 조인 릴레이션 속성집합에 대한 교집합의 부분집합이다. 세 번째는 릴레이션의 카티시언 곱에서 조인속성값이 동일하지 않은 튜플도 선택된다는 것이다.

자세한 범위조인연산자 정의를 위해 아래와 같이 튜플벡터에 대한 정의를 하겠다.

정의.

릴레이션 R의 스키마가 A이고 $A \supset A'$를 만족하는 속성집합 $A' = \{A_1', A_2', \cdots, A_{a'}'\}$가 있고 A'를 구성하는 모든 속성의 도메인은 실수의 부분집합이면, 가능한 모든 속성값의 조합 $(a_1', a_2', \cdots, a_{a'}')$은 R^n의 부분집합이다. 이 때 릴레이션 R의 임의의 튜플 t에 대하여 $\boldsymbol{t(A')} = (t.a_1', t.a_2', \cdots, t.a_{a'}')$을 A'에 대한 튜플 t의 튜플벡터라 한다.

범위조인연산자는 자연조인연산자와 유사하게 $\bowtie_\rho$를 사용하며 두 릴레이션에 대한 범위조인연산은 다음과 같다.

스키마가 각각 R과 S인 두 릴레이션 r과 s가 있고 $C \subset R$와 $C \subset R$를 만족하는 두 릴레이션의 범위조인속성집합 $C = \{C_1, C_2, \cdots, C_c\}$가 있을 때 r과 s의 범위조인은 식(1)과 같다.

$$r \bowtie_\rho s(C) \text{ 또는 } r \bowtie_\rho s(C_1, C_2, \cdots, C_c) \qquad (1)$$

이 때 C의 모든 원소의 도메인의 실수의 부분집합이며 ρ는 범위상수로 $r \times s$에서 선택될 튜플의 조인속성값의 최대 거리를 나타낸다. 결과 릴레이션의 스키마는 자연조인과 동일하게 $R \cup S$이다.

균형조인연산 과정은 다음과 같다. 먼저 $r \times s$에서 유효한 튜플을 선택한다. 유효한 튜플이란, r에서 임의의 튜플 t_r과 s에서 임의의 튜플 t_s에 의하여 조합된 튜플 t_v가 식(2)를 만족할 때 해당 튜플 t_v는 유효한 튜플이다. 즉 유효한 튜플 t_v는 속성집합 C에 대한 튜플벡터의 거리가 ρ 이하인 두 튜플로 조합된 튜플을 의미한다.

$$|t_r(C) - t_s(C)| \le \rho \quad (2)$$

그 다음 단계로 결과 릴레이션에 속성집합 C를 추가하고 각각의 속성값에 식(3)과 같이 $r.C_i$와 $s.C_i$의 평균값을 할당한다.

$$t_v.c_i = \frac{t_r.c_i + t_s.c_i}{2},\ (1 \le i \le c) \quad (3)$$

마지막으로 결과릴레이션에서 r과 s의 속성집합 C를 제거하고 범위조인연산을 마친다.

범위조인연산에서 $\rho = 0$이고 $R \cap S = C$이면 결과는 자연조인과 동일하다.

예를 들어 <표 1>과 같이 온도와 습도에 대한 두 릴레이션 *temp*와 *hum*이 있다고 하자. 속성 X와 Y는 센서의 좌표를 나타내고 T와 H는 온도와 습도를 각각 나타낸다. 이 때 좌표 X와 Y를 조인속성으로 하여 자연조인을 수행하면 결과는 공집합이지만 거리 10 이하의 센서쌍에 대하여 조인이 가능하도록 하는 범위조인연산 $temp \bowtie_{10} hum(X, Y)$의 결과는 <표 2>와 같이 6개의 튜플을 가진다.

다음으로 3개 이상의 릴레이션을 대상으로 하는 범위조인연산에 대하여 살펴보겠다. 자연조인에서는 $(r_1 \bowtie r_2) \bowtie r_3 = r_1 \bowtie (r_2 \bowtie r_3)$과 같이 여러 릴레이션을 조인할 때 연산순서에 관계없이 동일한 결과를 얻을 수 있지만 범위연산에서는 그러하지 못하다. 따라서 복수의 범위조인연산은 하나의 조인연산으로 취급한다. n개의 릴레이션 r_1, r_2, …, r_n에 대한 범위조인연산은 식(4)와 같이 표기하며 자연조인과의 혼돈을 피하기 위해 ρ는 첫 번째 조인연산자와 함께 표기한다.

$$r_1 \bowtie_\rho r_2 \bowtie \cdots \bowtie r_n(C) \text{ 또는}$$
$$r_1 \bowtie_\rho r_2 \bowtie \cdots \bowtie r_n(C_1, C_2, \cdots, C_n) \quad (4)$$

식(4)는 복수의 조인연산이 아니라 하나의 범위조인연산이므로 유효한 튜플은 $r_1 \times r_2 \times \cdots \times r_n$으로부터 추출한다. 이 때 유효한 튜플은 다음과 같다.

<표 1> *temp*와 *hum* 릴레이션

temp

노드 ID	*X*	*Y*	*T*
TS1	62	48	24
TS2	54	70	23
TS3	56	74	25
TS4	78	90	23
TS5	93	34	26
TS6	99	65	22

hum

노드 ID	*X*	*Y*	*H*
HS1	34	68	70
HS2	65	45	60
HS3	73	90	77
HS4	56	73	89
HS5	90	25	56
HS6	80	85	86

<표 2> $temp \bowtie_{10} hum(X, Y)$ 결과

노드쌍	X	Y	T	H
(TS1, HS2)	64	47	24	60
(TS2, HS4)	55	72	23	89
(TS3, HS4)	56	74	25	89
(TS4, HS3)	76	90	23	77
(TS4, HS6)	79	88	23	86
(TS5, HS5)	92	30	26	56

튜플 t_{r_1}, t_{r_2}, …, t_{r_n}가 각각 r_1, r_2, …, r_n에 포함된 튜플이고 t_v는 이 튜플로 조합된 튜플이라고 하자. 이 때 t_v를 조합하는 모든 튜플이 식(5)를 만족하면 t_v는 유효한 튜플이다. 이는 유효한 튜플을 구성하는 모든 튜플의 조인속성 C에 대한 튜플벡터들의 거리가 ρ 이하임을 의미한다.

$$|t_{r_i}(C) - t_{r_j}(C)| \le \rho,\ \forall i,j (1 \le i,j \le n) \quad (5)$$

유효한 튜플을 선택한 후 앞에서와 마찬가지로 C를 추가하고 속성값을 할당한다. 두 릴레이션에 대한 범위조인과 마찬가지로 유효한 튜플 t_v의 속성값은 t_v를 구성하는 모든 튜플들의 속성값에 대한 평균으로 식(6)과 같이 정의할 수 있다.

$$t_v.c_i = \frac{1}{n}\sum_{j=1}^{n} t_{r_j}.c_i,\ 1 \le i \le c \quad (6)$$

속성값을 할당한 후 각각의 $r_i (1 \le i \le n)$의 C를 삭제함하고 범위조인연산을 마친다.

이와 같은 여러 릴레이션에 대한 범위조인에서도 $\rho = 0$이고 $R_1 \cap R_2 \cap \cdots R_n = C$인 경우 결과는 자연조인 $r_1 \bowtie r_2 \bowtie \cdots \bowtie r_n$과 동일하다.

4. 결론

여러 종류의 센서로 구성된 센서네트워크에서 노드를 무작위로 뿌렸을 경우 서로 다른 종류의 센서 노드가 동일한 위치에 배치될 확률은 매우 적다.

따라서 대부분 센서의 위치는 동일하지 않으므로 서로 다른 종류의 센싱 값들을 조인하기 위하여 좌표를 조인속성값으로 이용할 경우 실제로 조인결과는 공집합에 가깝게 된다.

이러한 문제를 해결하기 위하여 본 논문에서는 동일한 키 값이 아닌 일정한 범위 내에서 유사한 키 값을 가지는 튜플도 선택할 수 있는 범위조인연산자를 제안하였다.

이 범위조인연산자를 센서네트워크에 적용할 경우 좀 더 유연한 조인연산을 수행할 수 있고 데이터베이스계층에서 응용계층으로 표준화된 인터페이스를 제공할 수 있다.

범위조인연산자를 센서네트워크에 적용함에 있어 튜플 전송에 따른 에너지 효율적인 연산 및 질의처리 방법을 다음 연구과제로 남기며 본 논문을 마친다.

참고문헌

[1] Philippe Bonnet, Johannes Gehrke, and Graveen Seshadri “Towards sensor database systems” In Mobile Data Management (2001) 3-14

[2] M. Srivastava, R. Muntz, and M. Potkonjak.: Smart Kindergarten “Sensor-Based Wireless Networks for Smart Developmental Problem-Solving Environments” In Proceedings of the Seventh Annual ACM/IEEE International Conference on Mobile Computing and Networking (Mobicom2001)

[3] R. Govindan, J. Hellerstein, W. Hong, S. Madden, M. Franklin, and S. Shenker “The Sensor Network as a Database” Technical Report 02-771, Computer Science Department, University of Southern California (2002)

[4] J. M. Hellerstein, W. Hong, S. Madden, and K. Stanek.: Beyond average “Toward sophisticated sensing with queries” In Information Processing in Sensor Networks: 2nd Intl. Workshop, Springer-Verlag, LNCS 2634 (2003) 63-72

[5] S. Madden, M.J. Franklin, J. Hellerstein, and W. Hong. “Tag: a tiny aggregation service for ad-hoc sensor networks” In Proc. of OSDI '02 (2002)

[6] N. Shrivastava, C. Buragohain, D. Agrawal, and S. Suri. “Medians and beyond: New aggregation techniques for sensor networks” In Proc. of Sensys'04 (2004)

[7] Joseph M. Hellerstein, Peter J. Haas, and Helen J. Wang. “Online Aggregation” In Proc. ACM SIGMOD International Conference on Management of Data (1997)

[8] T. Friedman and D. Towsley. “Multicast Session Membership Size Estimation” In Proc. of IEEE Infocom (1999)

[9] W. Hou, G. Ozsoyoglu, and B. Taneja. “Statistical estimators for relational algebra expressions” In Proc. Seventh ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems(PODS) (1988) 276-287

[10] Johannes Gehrke, Samuel Madden “Query Processing in Sensor Networks” IEEE CS and IEEE ComSoc (2004)

[11] Annita N. Wilschut and Peter M. G. Apers “Dataflow query execution in a parallel main-memory environment” Distributed and Parallel Databases, 1(1) (1993) 103-128

[12] Peter J. Hass and Joseph M. Hellerstein “Ripple Joins for Online Aggregation” In Proc. ACM-SIGMOD International Conference on Management of Data (1999) 287-298

[13] S. Madden and M. Franklin. “Fjording The Stream: An Architecture for Queries over Streaming Sensor Data” In Proceedings of the 18th International Conference on Data Engineering, San Jose, CA (2002)