데이터 중심 저장구조에 기반한 스카이라인 질의 처리 기법

Skyline Query Processing Method based on Data Centric Storage

여명호**, 성동욱**, 송석일*, 유재수**
충주대학교*, 충북대학교**

Myung-Ho Yeo**, Dong-Ook Seong**,
Seok-Il Song*, Jae-Soo Yoo**
Chungju National University*,
Chungbuk National University**

요약

센서 네트워크를 위한 데이터 중심 저장구조는 다차원의 범위질의는 물론 정합질의를 효과적으로 처리하기 위해서 제안되었다. 하나의 센서 네트워크는 범위질의 하나만 처리하거나 스카이라인 질의 하나만을 처리하기 위해서 설치될 수도 있지만 일반적으로 다양한 질의를 같이 처리하기 위해서 사용된다. 따라서, 데이터 중심 저장구조에서 다차원의 범위질의 뿐만 아니라 스카이라인 질의도 효과적으로 처리될 필요가 있다. 기존에 제안된 스카이라인 질의 처리 알고리즘들은 데이터 중심 저장구조의 존재를 고려하지 않고 있다. 일부 대표적인 데이터 중심 저장구조는 유사한 데이터를 지리적으로 인접한 센서 노드에 저장하는 특징을 갖는다. 이 논문에서는 이러한 특징을 고려하여 데이터 중심 저장구조에서 보다 효과적으로 동작할 수 있는 스카이라인 질의 처리기법을 제안한다.

Abstract

Data centric storages for sensor networks have been proposed to efficiently process multi-dimensional range queries as well as exact matches. Usually, a sensor network does not process only one type of the query but supports various types of queries such as range queries, exact matches and skyline queries. Therefore, a sensor network based on a data centric storage for range queries and exact matches should process skyline queries efficiently. However, existing algorithms for skyline queries have not considered the features of data centric storages. Some of the data centric storages store similar data in sensor nodes that are placed on geographically similar locations. Consequently, all data are ordered in a sensor network. In this paper, we propose a new skyline query processing algorithm that exploits the above features of data centric storages.

Ⅰ. 서론

무선 통신 기술, 컴퓨팅 기기 제조 기술 및 센서 기술의 발전으로 컴퓨팅 능력과 무선 통신 및 센싱이 가능한 소형 센서 노드들이 개발되었다. 무선 센서 노드들은 다양한 환경에 설치되며 무선 통신 능력을 통해서 서로 네트워크를 이룬다. 센서 노드들로 구성되는 센서 네트워크는 다양한 형태의 데이터들을 센싱 및 처리하고 필요에 따라 서로 전송하여 사용자들에게 유용한 정보를 제공한다. 센서 네트워크가 응용될 수 있는 분야는 매우 다양하며 이에 따라 많은 관심을 받게 되었다. 대표적인 센서 네트워크의 응용분야로는 육지 및 수중의 동물 서식지 관찰, 건물의 안전성 모니터링, 대기 및 수질의 오염도 모니터링, 산불 및 홍수와 같은 재난 발생 모니터링, 건물내 환경 모니터링, 전장에서의 침입자 모니터링 등 매우 다양하다[1].

스카이라인 질의란 다차원의 속성을 갖는 데이터의 최대값이나 최소값등을 찾아내는 질의라고 할 수 있다. 스카이라인 질의를 처리하기 위해서는 다차원의 속성을

이 논문은 2009년 교육과학기술부로부터 지원받아 수행된 연구임(지역거점연구단육성사업/ 충북BIT연구중심대학육성사업단)

갖는 데이터 전체 집합 중에서, 모든 차원에 대해서 다른 데이터들에 의해 지배되지(dominated) 않는 데이터들을 찾아낸다[2]. 스카이라인 질의의 결과들은 다양한 응용의 의사 결정을 지원하는데 사용된다. 센서 네트워크의 각 센서 노드는 일반적으로 여러 센서들(온도, 습도, 조도, SO_2, CO_2, CO 등)을 장착하고 있어서 다차원의 속성을 갖는 데이터들을 수집한다. 따라서, 센서 네트워크에서 스카이라인 질의를 효과적으로 처리하는 기술을 개발하는 것은 매우 중요하다. 센서 네트워크에서의 스카이라인 질의로는 수질이나 대기 오염이 가장 심한 지역을 찾는 질의나 반대로 수질이나 대기가 가장 청정한 지역을 찾는 질의 등을 들 수 있다.

이 논문에서는 지리적으로 유사한 지역의 센서 노드들이 유사한 데이터를 저장하는 데이터 중심 저장구조에서 에너지 효율적인 스카이라인 질의 처리 기법을 제안한다. 데이터 중심 저장 기법의 특성을 이용하여 에너지를 효율적으로 사용하고, 질의 처리 성능을 높이기 위해서 논리적인 이웃노드 개념과 스카이라인 질의 처리를 위한 메시지 전송규칙을 함께 제안한다.

이 논문의 구성은 다음과 같다. II장에서는 이 논문에서 제안하는 스카이라인 질의 처리 기법과 관련된 데이터 중심 저장구조 기법들과 기존에 제안된 센서 네트워크를 위한 스카이라인 질의 처리기법을 설명한다. III장에서는 이 논문에서 제안하는 스카이라인 질의처리 기법을 설명하고 IV장에서 제안한 질의처리기법을 시뮬레이션을 통해서 평가하고 분석한다. 마지막으로 V장에서 결론을 맺는다.

Ⅱ. 관련연구

센서 네트워크에서의 스카이라인 질의를 처리하기 위한 기술 역시 효과적으로 전력소모를 할 수 있도록 설계되어야 한다. 하지만, 아직 이에 대한 연구는 시작단계에 불과하며 소수의 연구가 진행된 바 있다. [3, 4]에서 제안된 스카이라인 질의 처리 기법들은 질의 결과의 후보 집합을 찾을 수 있는 필터를 센서에 설치한다. 그리고, 각 센서들이 정해진 경로를 통해 데이터를 싱크 노드로 전송하면 각 센서 노드에 설치된 필터들이 데이터들을 걸러내어 데이터 전송 량이 줄어들게 되고 자연스럽게 전력 소모가 줄어든다. 이 방법들은 데이터 중심 저장구조와 같은 특별한 형태의 하부 저장시스템을 고려하지 않고 설계 되었다.

어떤 센서 네트워크가 설치되어 응용될 때는 한 가지 종류의 질의만을 처리할 수도 있지만 여러 형태의 질의를 같이 처리하는 경우도 있다. 공기 오염이 가장 심한 지역을 모니터링 하면서 동시에 CO_2 및 SO_2 가 특정 범위안에 있는 지역을 찾는 범위 질의를 처리하거나 CO_2 가 특정 값인 지역을 찾는 정합질의를 처리할 필요도 있을 것이다. 최근에 센서 네트워크상에서 다차원 데이터에 대한 범위질의(range search) 및 정합질의(exact match)를 효과적으로 처리하기 위한 방법으로 다양한 데이터 중심 저장구조(Data Centric Storage, 이하 DCS)가 제안되었다[5, 6, 7]. DIM[5], KDDCS[6], GDCS[7]은 유사한 데이터를 지리적으로 인접한 센서 노드들에 저장하는 접근방법을 사용한다. 이들 데이터 중심 저장구조상에서 데이터들은 그 값에 의해서 정렬된다.

이와 같은 데이터 중심 저장구조를 기반으로 하는 센서 네트워크에서 동시에 스카이라인 질의를 처리할때는 데이터 중심 저장구조가 유사한 데이터를 지리적으로 인접한 지역에 저장한다는 점을 이용한다면 보다 효과적인 스카이라인 질의를 처리할 수 있을 것이다. 즉, 센싱된 데이터들이 센서 네트워크상에 지리적으로 정렬되어 저장되기 때문에 어떤 차원에 대해서 가장 큰 데이터나 가장 작은 데이터를 찾는 것이 보다 용이하다.

Ⅲ. 제안하는 DCS 기반의 스카이라인 질의 처리 기법

이 논문에서는 DCS를 기반으로하는 센서 네트워크에서 에너지 효육적인 스카이라인 질의 처리 기법을 제안한다. 기존에 제안된 스카이라인 질의 처리기법들을 그대로 DCS에서 사용할 수도 있다. 하지만, 유사한 데이터들을 인접한 지역의 센서 노드에 저장하는 DCS에 기존에 제안된 방법들을 그대로 적용하면 보다 효과적으로 질의를 처리할 수 있는 기회를 놓치게 된다. 이 논문에서는 유사한 데이터를 인접한 지역의 센서 노드에 저장하는 DCS의 특성을 고려하여 보다 효과적으로 스카이라인 질의를 처리할 수 있는 기법을 제안한다.

제안하는 질의 처리 기법은 다음과 같은 가정을 전제

한다. 첫 번째, 모든 센서 노드는 자신의 고유한 ID(Identification)를 가지고 있다. 두 번째, 모든 센서 노드는 자신의 위치와 이웃 노드들의 ID 와 위치를 알고 있다. 이웃 노드들이란 자신과 한 홉(hop)에 의해 통신할 수 있는 노드들을 말한다. 세 번째, 각 노드는 이웃 노드들이 해당 DCS에서 저장을 담당하고 있는 데이터의 범위(NDR)를 알고 있다. 이런 가정을 전제로 스카이라인 질의는 다음과 같이 3단계에 걸쳐서 처리된다.

첫 번째 단계에서는 베이스 스테이션(BS)이 사용자가 발생시킨 스카이라인 질의를 분석하여 가장 효과적으로 질의처리가 가능하다고 예상되는 센서 노드로 전송한다. 이 센서 노드로부터 실제적인 스카이라인 질의 처리가 시작되며 이를 시작노드(SN) 라고 한다. BS는 SN을 찾기 위해서 스카이라인 질의가 각 속성값에 대해서 최대 값을 찾는 것인지 최소값을 찾는 것인지 분석한다. 분석결과에 따라 가장 유력한 후보집합을 저장하고 있다고 예상되는 노드를 SN으로 정한다.

두 번째 단계는, SN이 스카이라인 질의 처리 메시지(SQM)를 생성하고 이를 인접한 노드로 전파하여 스카이라인 질의에 대한 후보 집합을 저장하고 있는 후보 노드(CN)들을 찾아내는 단계이다. SN이 생성하고 전파하는 SQM는 (Q, CSR, SIDs, TIDs) 으로 구성된다. Q는 스카이라인 질의의 내용이고, CSR은 현재까지 찾아낸 결과 후보 집합의 각 차원별 데이터 범위를 뜻한다. SIDs는 SQM을 전송하는 노드와 이 노드까지 SQM이 전송되면서 거쳐온 노드들의 ID이다. 메시지를 전송하는 노드의 ID이고 TIDs는 메시지를 수신할 노드들의 ID 이다. SN 이 생성한 최초의 SQM의 CSR, SIDs, TIDs 는 모두 NULL값을 갖는다.

CN을 찾아내기 위해 SQM 을 전송하는 과정은 다음과 같다. SN을 포함한 모든 노드는 SQM의 CSR을 읽고 자신이 저장하고 있는 데이터를 최소한으로 포함하도록 CSR을 변경한다. 당연히 노드가 아무런 데이터도 저장하고 있지 않다면 CSR을 변경할 필요가 없을 것이다. 변경후에도 CSR이 NULL 이라면 모든 이웃노드들에게 SQM을 전송한다. CSR이 NULL이 아니라면 노드가 저장하고 있는 모든 이웃노드들의 NDR들 중에서 CSR에 의해 완전히 지배되지 않는 NDR을 갖는 노드들을 선택한다. 물론, 노드들 중에서 현재 노드에게 SQM을 전송한 노드는 제외한다. 그리고, 선택한 노드들을 SQM의 TIDs에 포함시켜서 메시지를 전송한다. 이 메시지는 TIDs의 노드들에게 전달되며 메시지를 전달받은 노드들은 위의 과정을 반복한다.

SQM을 전송한 노드는 변경된 CSR을 저장하여 메시지 중복 송신을 피한다. 각 노드는 동일한 Q를 갖는 SQM을 여러 노드로부터 중복해서 수신할 수도 있다. 이때 수신하는 모든 SQM에 대해서 위와같은 절차를 반복하게 되면 질의처리에 소요되는 메시지의 크기가 증가하여 전력소모를 가중시킬 것이다. 노드는 자신이 송신한 SQM의 CSR을 저장하고 있다가 같은 질의에 대한 SQM이 수신되면 CSR을 비교하여 저장하고 있던 CSR이 수신된 SQM의 CSR을 지배하면 수신한 SQM을 무시한다. 그렇지 않다면 위의 과정을 반복하고 저장하고 있는 CSR을 새로운 SQM의 CSR로 변경한다. 메시지 전송은 어떤 노드가 더 이상 완전히 지배되지 않는 NDR을 갖는 이웃노드를 찾지 못하면 종료된다.

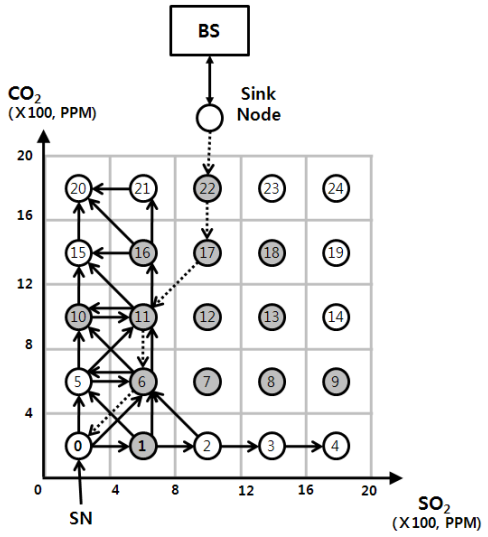
마지막 세 번째 단계는 BS가 선정한 CN들로부터 후보 집합을 수신하여 최종 질의 결과를 만들어내는 단계이다. SQM을 수신하는 노드들중에서 실제로 데이터를 저장하고 있었던 노드들이 CN이 된다. CN들은 자신이 저장하고 있던 데이터를 BS로 전송하고 BS는 CN들이 전송한 후보 집합으로부터 최종 질의 결과를 찾아낸다.

지금까지 설명한 스카이라인 질의처리 과정을 예를 통해서 설명한다. 그림 1의 센서 네트워크는 25개의 센서 노드가 그리드 형태로 배치되어 있고, 한 노드는 최대 8개의 이웃노드를 갖는다. 색이 칠해져 있는 노드는 값을 저장하고 있는 노드이고, 색이 없는 노드는 값을 저장하고 있지 않은 노드이다. BS는 CO2 와 SO2 가 가장 낮은 지역을 찾는 "오염도가 가장 낮은 지역을 찾아라." 라는 스카이라인 질의를 사용자로부터 받았고 이를 분석해서 0번 노드를 SN으로 선정하고 질의를 전송한다.

0번 노드는 (Q, NULL, NULL, NULL) 과 같은 SQM을 생성한 후에 자신이 값을 저장하고 있는지 확인한다. 0번 노드는 값을 저장하고 있지 않은 노드이므로 CSR은 NULL값을 유지한다. 이웃노드인 1, 5, 6번 노드의 NDR은 CSR에 의해서 완전히 지배되지 않으므로 이들 노드에 SQM (Q, NULL, 0, {1, 5, 6})을 전송한다. SQM을 수신한 1번 노드는 자신이 저장하고 있는 값을 최소한으로 포함하도록 CSR을 {{4, 8}, {0, 4}}로 변경한다. 변경된 CSR 과 1번 노드의 이웃노드인 2, 7, 6, 5, 0의 NDR을 비교하여 완번히 지배되지 않는 NDR을 갖는 이웃노드를 선택한다. 0, 2, 5, 6 번 노드가

선택되지만 0번 노드는 1번노드로 SQM을 전송한 노드이므로 제외한다. 1번 노드는 2, 5, 6번 노드에게 SQM (Q, {{4, 8}, {0, 4}}, {0, 1}, {2, 5, 6})을 전송한다.

1번 노드로부터 SQM을 수신한 6번 노드는 자신이 저장하고 있는 값을 포함하도록 CSR을 {{4, 8}, {0, 8}} 로 변경한다. 그리고 6번의 이웃노드들인 0, 1, 2, 5, 7, 10, 11, 12 번의 NDR과 CSR을 비교하여 완전히 지배되지 않는 NDR을 갖는 노드들을 선택한다. 그 결과로 0, 1, 2, 5, 10, 11 번이 선택되지만 0, 1번은 SIDs에 포함되어 있으므로 제외된다. 6번은 1번 노드로부터 SQM을 수신한 후에 다시 0번 노드로부터도 SQM을 수신한다. 하지만, 0번으로 부터의 SQM에 포함된 CSR이 NULL 이므로 6번 노드에 저장되어 있는 CSR에 의해 지배된다. 따라서 0번 노드로부터의 SQM은 무시된다. 이와 같은 방법으로 SQM을 센서 네트워크에 전파하면서 SQM을 수신한 노드 중에 실제로 데이터를 저장하고 있던 1, 6, 10, 11, 16 번 노드가 CN이 되며 이들은 SQM을 수신한 후 바로 자신이 저장하고 있는 데이터를 BS로 전송한다. BS는 수신한 후보 집합 들 중에서 최종 결과를 선정한다.



▶▶ 그림 1. 스카이라인 질의 처리 과정

Ⅵ. 성능평가

이 논문에서 제안하는 스카이라인 질의처리 기법은 GDCS를 기반으로 하여 성능을 평가하였다. 시뮬레이션 파라미터는 표 1과 같다. 노드의 개수는 400개로 하였으며 센서 네트워크가 설치되는 전체 영역은 가로 세로 각각 200m 로 가정하였다. 각 센서 노드의 초기 에너지 레벨은 1000 으로 하였고, 메시지를 전송하는데 0.1 이 소모되고 메시지를 수신하는 데는 0.05 가 소모된다고 가정하였다. 마지막으로 각 센서 노드가 한 홉에 통신할 수 있는 거리는 15m 로 가정하였다.

모든 노드는 정규분포의 데이터를 100개씩 수집하여 GDCS 방식으로 센서 네트워크 내부에 저장한다. 한 노드가 수집하는 전체 데이터의 80%는 20%의 범위에 분포하며 전체 노드의 5%(약 20개 노드)는 데이터를 저장하고 있지 않다. 시뮬레이션에서는 GDCS에 의해서 데이터를 저장하고 있는 센서 네트워크에 "현재 각 차원의 값이 최대가 되는 지역을 찾아라" 라는 스카이라인 질의가 전송되었다고 가정하였다. SN은 우측상단의 노드가 되며 BS와 SN까지의 통신비용은 시뮬레이션에서 제외하고 각 후보노드가 BS까지 자신들의 데이터를 전송할 때까지의 통신비용을 측정하여 에너지 레벨을 계산하였다. 각 후보노드는 자신이 저장하고 있는 데이터 중 가장 최신의 데이터 하나를 BS로 전송한다.

표 1. 시뮬레이션 파라미터

노드 개수	400
센서 네트워크 설치 영역	200 x 200 m2
노드의 초기 에너지	1000 units
메시지 송신에 필요한 에너지 소모	0.1 units
메시지 수신에 필요한 에너지 소모	0.05 units
저장 용량	100 units
노드의 한 홉 통신 거리	15m

표 2는 시뮬레이션을 수행한 후 질의처리에 참여한 노드들의 수와 각 노드들의 평균 에너지 레벨, GDCS를 유지하기 위한 노드의 평균 에너지 소모, GDCS 유지와 하나의 스카이라인 질의처리를 위한 노드의 평균 에너지 소모를 보여주고 있다. 표 2에서 보는 것처럼, 질의에 참여한 노드는 55개로 전체 400개 노드중 약 14%에 불과하다. 또한, GDCS를 구축하고 유지하는데 소모된 에너지가 노드 평균 752.38 units 인데 반해, 질의처리를 포함한 에너지 소모를 측정하면 노드 평균

752.93 units으로 약 0.5 units 을 소모하는데 그쳤다.

표 2. 질의처리에 소모되는 에너지 측정

척도	측정 값
질의처리에 참여한 노드의 수	55개 노드 (13.75%)
질의처리를 위해 소비한 전체 에너지	220.55 units
질의처리에 참여한 노드의 평균 에너지 소모	4.01 units
GDCS를 유지하는데 소요된 노드의 평균에너지 소모	752.38 units
전체 노드의 평균 에너지 소모 (GDCS 유지 + 질의처리)	752.93 units

V. 결론

이 논문에서는 지리적으로 유사한 지역의 센서 노드들이 유사한 데이터를 저장하는 데이터 중심 저장구조에서 에너지 효율적인 스카이라인 질의 처리 기법을 제안하였다. 또한, 에너지 사용의 효율을 높이고, 질의 처리 성능을 향상시키기 위해서 논리적인 이웃노드 개념과 스카이라인 질의 처리를 위한 메시지 전송규칙을 함께 제안하였다. 그 결과, 부분 노드의 참여만으로도 스카이라인 질의에 대한 처리를 수행하였으며, 에너지 소모가 감소하였다. 향후 연구는 메시지의 중복을 줄이고, 다양한 네트워크 모양을 구성하여 제안하는 기법을 추가적으로 실험하는 것이다.

■ 참 고 문 헌 ■

[1] R. Szewczyk, A. Mainwaring, J. Polastre, and D. Culler, “An Analysis of a Large Scale Habitat Monitoring Application”, Proc. ACM Conf. Embedded Networked Sensor Systems (SenSys ‘04) Nov. 2004.

[2] S. Borzsonyi, D. Kossmann, and K. Stocker, “The skyline operator”, Proc. International Conference on Data Engineering, pp.421-430, 2001.

[3] K. Yoon, J. Choi, Y. Chung, and S. Lee, “In-Network Processing for Skyline Queries in Sensor Networks”, IEICE Trans. COMMUN., vol. E90-B, no. 12, Dec. 2007.

[4] 성동욱, 여명호, 유재수, “에너지 효율적인 연속 스카이라인 질의 처리 기법”, 한국정보과학회논문지: 컴퓨팅의 실제 및 레터, 제15권, 제4호, 2009년 4월

[5] X. Li, Y. J. Kim, R. Govindan, and W. Hon, "Multi-dimensional Range Queries in Sensor Networks," Proc. of SenSys, pp. 63-75, 2003.

[6] M. Aly, K. Pruhs, and P. K. Chrysanthis, "KDDCS: a Load-balanced in-network Data-centric Storage Scheme for Sensor Networks," Proc. of CIKM, pp. 317-326, 2006.

[7] 신재룡, 유재수, 송석일, “GDCS: 센서네트워크를 위한 에너지 효율적인 그리드 기반 데이터 중심 저장 시스템”, 한국콘텐츠학회논문지, 제9권, 제1호, 2009년 1월.