

# 데이터 중심 저장 환경에서 소실 데이터 보정 기법을 이용한 인-네트워크 병합 질의 처리

## (In-network Aggregation Query Processing using the Data-Loss Correction Method in Data-Centric Storage Scheme)

박 준 호 †      이 효 준 †      성 동 옥 †      유 재 수 ††  
 (Jun Ho Park)      (Hyo Joon Lee)      (Dong Ook Seong)      (Jae Soo Yoo)

**요 약** 센서 네트워크에서 발생하는 데이터를 저장하고, 효율적으로 질의를 처리하는 기법에 대한 많은 연구가 이루어지고 있다. 대표적인 연구로 데이터 중심 저장 기법이 있다. 데이터 중심 저장 기법의 경우 질의를 효과적으로 처리하기 위해 수집한 데이터 값에 따라 저장 될 센서 노드를 지정하고, 질의 처리를 위해 질의에 해당하는 데이터를 저장하는 노드에서만 데이터를 수집한다. 하지만 노드의 결함이 발생하면 결함 노드에 저장 되어 있는 전체 데이터가 소실 됨에 따라 질의 결과 정확도가 저하 되는 문제점이 발생한다. 이러한 문제를 해결하기 위해, 본 논문에서는 데이터 중심 저장 기법에서 노드 결함에 따른 데이터 소실이 발생하여도 높은 정확도를 보이는 인-네트워크 병합 질의 처리 기법을 제안한다. 데이터 소실이 발생 하였을 경우 선형 회귀 분석 기법을 이용하여 소실 된 영역에 해당하는 보정 모델을 생성하고, 이를 통해 가상의 데이터를 포함한 질의 결과를 반환한다. 제안하는 기법의 우수성을 보이기 위해 시뮬레이션을 통해 소실 데이터 보정 기법을 적용한 KDDCS(E-KDDCS) 기법과 기존의 데이터 중심 저장 기법과 성능을 비교하였다. 그 결과 기존 기법에 비해 질의 결과 정확도가 향상되었고, 질의 처리 시 에너지 소모를 감소시켰다.

키워드 : 센서네트워크, 데이터중심저장기법, 인-네트워크질의처리, 결함, 선형회귀분석

**Abstract** In Wireless Sensor Networks (WSNs), various Data-Centric Storages (DCS) schemes have been proposed to store the collected data and to efficiently process a query. A DCS scheme assigns distributed data regions to sensor nodes and stores the collected data to the sensor which is responsible for the data region to process the query efficiently. However, since the whole data stored in a node will be lost when a fault of the node occurs, the accuracy of the query processing becomes low. In this paper, we propose an in-network aggregation query processing method that assures the high accuracy of query result in the case of data loss due to the faults of the nodes in the DCS scheme. When a data loss occurs, the proposed method creates a compensation model for an area of data loss using the linear regression technique and returns the result of the query including the virtual data. It guarantees the query result with high accuracy in spite of the faults of the nodes. To show the superiority of our proposed method, we compare E-KDDCS (KDDCS with the proposed method) with existing DCS schemes without the data-loss correction method. In the result, our proposed method increases accuracy and reduces query processing costs over the existing schemes.

Key words : Wireless Sensor Networks, Data-Centric Storage Scheme, In-Network Query Processing, Faults, Linear Regression

· 이 논문은 2010년 교육과학기술부의 지원(지역거점연구단육성사업/충북BIT연구중심대학육성사업단)과 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(No. 2009-0080279)의 결과임

† 학생회원 : 충북대학교 정보통신공학과  
 arionfit@naver.com  
 reverse999@gmail.com  
 Seong.do@gmail.com

†† 종신회원 : 충북대학교 정보통신공학과 교수  
 yjs@chungbuk.ac.kr  
 (Corresponding author)

논문접수 : 2010년 8월 23일  
 심사완료 : 2010년 9월 27일

Copyright©2010 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.  
 정보과학회논문지: 데이터베이스 제37권 제6호(2010.12)

## 1. 서론

최근 컴퓨팅 기술의 비약적인 발전과 통신분야의 하드웨어 기술 발전으로 인해 센서 네트워크는 자연 현상 관측, 군사 응용, 산업 현장, 재난 감시, 의료뿐만 아니라 실생활에서까지 폭넓게 활용되고 있다[1]. 센서 네트워크는 일반적으로 적게는 수백 개에서 많게는 수백만 개의 센서 노드들의 협업으로 구성이 되며, 센서 노드는 다양한 주변의 데이터를 측정하기 위해 다수의 센서 모듈을 탑재하고 있다. 이러한 센서 노드는 일반적으로 온도, 습도, 조도 등과 같은 다양한 데이터를 측정하여, 사람이 직접 수집하기 어려운 다양한 환경에 설치 되어 현상에 대한 감시, 정보의 전달, 그리고 이웃 노드와의 협동 작업 등을 수행한다[2].

기본적으로 센서 네트워크 응용은 센서에서 수집 된 데이터를 저장하고 질의를 처리하는 것을 바탕으로 한다. 센서 네트워크 내 질의는 일정한 주기로 데이터를 수집하여 많은 데이터 전송을 유발시키므로 통신 비용을 줄이는 것이 중요하다. 인-네트워크(In-network) 병합 질의 처리는 데이터 라우팅 도중에 네트워크 내에서 데이터를 병합하여 전송함으로써 메시지 전송 횟수를 줄임으로 기지국으로 전송되는 동안에 발생하는 통신비용을 줄여준다. 이러한 인-네트워크 질의 처리 기법[3]이 이용되기 위해선 각 센서가 가지고 있는 중간 결과를 합쳤을 때 원래의 결과와 같아야 하기 때문에 MIN, MAX, SUM, AVG, COUNT와 같은 질의를 처리할 때 효과적이다. 센서 네트워크에서 측정된 데이터를 저장하는 대표적인 방법으로는 센서 노드에서 측정된 데이터를 외부 저장소에 저장하는 외부 저장 기법(External Storage: ES)과 각 노드가 측정한 데이터는 센서 노드 자신이 저장하는 지역 저장 기법(Local Storage: LS), 그리고 데이터의 값에 따라 데이터를 저장할 센서 노드를 결정하는 데이터 중심 저장 기법(Data-Centric Storage: DCS)이 있다[4].

센서 네트워크의 질의 처리 성능을 향상시키기 위해 감지된 데이터를 센서 네트워크상에서 효과적으로 저장하고 처리하는 데이터 중심 저장 기법에 대한 연구가 활발하게 진행되고 있다[4]. 데이터 중심 저장 기법은 수집한 데이터를 데이터의 값을 기반으로 하는 해시 함수나 지역적인 위치에 의해 분류 되어 특정 센서 노드에 저장한다. 그러므로 질의 생성 시, 네트워크 전체에 데이터를 요구하는 것이 아니라, 질의에 해당하는 데이터를 저장하는 노드만 데이터를 수집하기 때문에 질의 처리의 관점에서 효과적이다.

하지만 센서 네트워크는 다양한 환경에 배포가 되기 때문에 통신 채널, 하드웨어 잡음과 같이 데이터 송수신

에서 발생할 수 있는 오류에 노출되어 있고, 외부 요인에 의한 노드 파손도 쉽게 발생한다. 뿐만 아니라, 데이터 중심 저장 기법의 특성에 따라 저장소 집중과 질의 집중에 의한 에너지 소진에 따른 결함도 발생한다[5-9]. 이러한 센서 네트워크의 배포 환경과 데이터 중심 저장 기법의 특성을 고려 할 때, 노드 결함이 발생하면 특정 범위의 해당하는 전체 데이터가 소실되고, 소실 된 데이터의 누락에 의한 질의 정확도 저하 문제가 발생한다. 그러므로 데이터 중심 저장 기법에 기반을 둔 센서 네트워크에서는 센서 노드의 결함에 의해 데이터가 소실 되어도 높은 정확도를 보장할 수 있는 질의 처리 기법에 대한 연구가 필요하다.

본 논문에서는 이와 같은 문제를 해결하기 위해 센서 노드나 네트워크 결함이 발생하는 데이터 중심 저장 기법에서의 결과 데이터의 정확도를 고려한 인-네트워크 병합 질의 처리 기법을 제안한다. 제안하는 기법은 질의 결과에서 에너지 소진을 포함한 노드 결함 및 네트워크 오류로 인해 데이터 소실이 발생 했을 경우 가상의 데이터를 생성한다. 높은 정확도를 갖는 가상 데이터를 얻기 위해서 통계 분야에서 사용되는 선형 회귀 분석 기법[10]을 이용하여 누락 범위에 해당하는 수식 모델링을 생성하고, 모델링에서 얻어진 가상의 데이터를 이용하여 질의 결과를 생성함으로써 데이터 누락에도 불구하고 높은 정확도를 가지는 질의 결과를 생성한다. 또한 기존의 데이터 중심 저장 기법에서 고려되지 않았던 인-네트워크 병합 처리를 수행 함으로써 질의 처리 비용을 감소시켰다. 결함 노드를 다른 노드로 대체하는 기법을 제안함으로 결함이 발생하는 환경에서도 네트워크의 수명을 지속적으로 유지하기 위한 기법을 제시한다.

본 논문의 구성은 다음과 같다. 제2장의 관련 연구에서는 기존에 제안된 센서 네트워크에서의 데이터 중심 저장 방식의 분석을 통해 문제점을 설명하고, 문제점을 해결하기 위한 선형 회귀 분석 기법에 대해 서술한다. 제3장에서는 본 논문에서 제안하는 기법과 알고리즘에 대해 기술한다. 제4장에서는 제안하는 기법을 적용한 센서 네트워크의 성능을 분석하고 결과를 보여준다. 마지막으로 제5장에서는 결론 및 향후 연구에 대해 언급한다.

## 2. 관련 연구

### 2.1 데이터 중심 저장 기법

기존에 데이터 중심 저장 기법으로 많은 연구가 진행되었다. GHT[5]는 공간 해싱을 하여 데이터를 저장할 좌표를 임의로 산출해서 그 좌표 값에 인접한 센서 노드로 데이터를 저장하는 기법이다. 이 기법은 데이터를 균등하게 분포를 못 할 뿐 아니라 데이터의 지역성을 보장하지 못한다. 그렇기 때문에 범위 질의를 수행할 경우

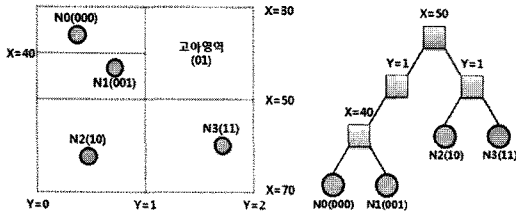


그림 1 DIM의 구조

분산되어 있는 노드로 질의를 나누어 전송해야 하므로 많은 라우팅 비용이 든다.

DIM[6]은 센서 네트워크를 공간 중심으로 영역을 분할한다. 분할된 영역은 일정 범위의 데이터를 담당하여 저장하게 되고 K-D 트리를 이용하여 분할된 영역정보와 노드 정보를 관리한다. 발생한 데이터는 GPSR 기법으로 데이터 저장 노드로 전송한다. 하지만 DIM은 데이터 중심으로 영역을 분할하지 않기 때문에 분할된 영역을 담당하는 노드가 없는 '고아 영역'이 발생하여 이 영역을 처리하기 위한 추가적인 처리비용이 필요하다.

KDDCS[7]는 DIM의 고아 영역 발생에 따른 문제점을 해결하기 위해 고안된 기법이다. KDDCS는 각 센서에게 분할된 영역을 할당하고 K-D 트리를 통하여 분할된 영역 정보와 노드 정보를 유지한다. 각 분할된 영역에 한 개 이상의 노드가 배치 되도록 영역을 분할 하도록 분할 축을 옮기기 때문에 고아 영역이 발생하지 않는다.

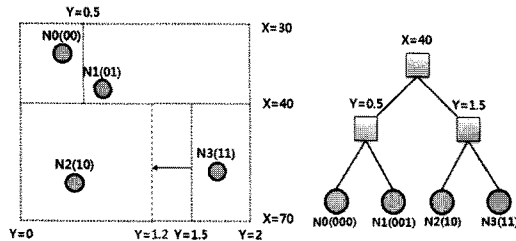


그림 2 KDDCS의 구조

그 외에도 다양한 DCS기법들이 제안되었다. ZS[8]과 DBAS[9]는 데이터 중심 저장 기법에서 문제가 되는 데이터 집중을 분산을 통하여 해결하는 기법이다. ZS은 데이터 집중이 생기는 영역을 다른 노드들이 공유하여 부하를 분산하여 데이터 집중을 감소시키는 방법이다. DBAS는 주기적으로 센서 네트워크가 기지국으로 데이터 집중에 대한 정보를 전송하여 부하를 균등하게 분산한다.

하지만 기존의 데이터 중심 저장 기법에서는 센서 노드의 에너지 고갈, 외부 충격에 의한 고장 그리고 네트워크 문제와 같은 예상치 못한 이유로 결함이 발생하게 되는 상황에 대한 고려가 이루어지지 않았다. 데이터 중

심 저장 기법에서는 일정 범위의 데이터를 모두 하나의 센서 노드에 저장을 하기 때문에 데이터 저장 노드가 결함이 발생 할 경우 특정 범위의 데이터는 모두 소실하게 되고, 결과적으로 질의 결과의 낮은 정확도를 보이게 된다. 따라서 노드 결함에 따른 데이터 소실에도 결과 데이터의 정확도를 높일 수 있는 기법에 대한 연구가 필요하다.

### 2.2 선형 회귀 분석 기법

선형 회귀 분석 기법[10]은 자료에 포함된 변수와 또 다른 하나 이상의 변수들 간의 연관 관계를 적절한 함수 식을 표현하여 자료 분석을 간편하게 하는 통계 기법이다. 센서 네트워크에서는 기존에 수집 된 데이터를 분석하여 임의의 구간에 대한 분포 형태를 예측하는 것이 가능하다.

선형 회귀 분석 기법은 함수관계를 통계적으로 찾아보는 분석방법으로 방법에 따라 다양한 수식 모형으로 만들어질 수 있다. 선형 회귀 분석 기법의 분류로는 단순 선형 회귀(Simple linear regression) 분석 기법과 다중 선형 회귀(Multiple linear regression) 분석 기법으로 나눌 수 있다. 단순 선형 회귀 분석 기법은 독립변수가 한 개이고 종속변수가 한 개인 경우의 분석에 사용되고, 독립변수가 여러 개 존재하는 경우의 분석에는 다중 선형 회귀 분석 기법이 사용된다. 선형 회귀 분석 기법의 모형은 기존 존재하는 데이터 계열을 가장 잘 나타내는 선에 대한 식 (1)과 같은 형식을 사용하고, 선형 회귀 분석 기법을 다이어그램으로 표현하면 그림 3과 같다.

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon \quad (1)$$

식 (1)에서  $\beta_0$ 는 절편(상수항)이고,  $\beta_i$ 는 각 독립 변수의 계수이며,  $p$ 는 독립 변수의 개수이다. 다이어그램의 각 데이터 요소에는 회귀 모델로부터 일정한 오류가 존재한다. 그렇기 때문에 선형 회귀 분석 기법에서는 회

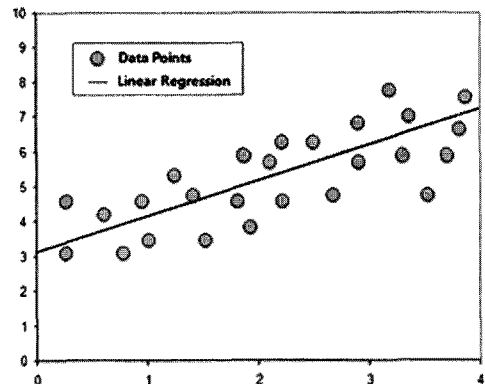


그림 3 독립변수와 종속변수에 따른 선형 회귀의 예

귀 모델은 각 계수와 상수 항을 조절하여, 모든 요소와 관련된 오류의 수의 합계가 최소가 되도록 하는 것이 정확도의 최대 관건이다. 일반적으로 센서 네트워크에서 발생하는 데이터나 응용에서는 수집한 정확한 수치를 요구하기보다는 어느 정도의 오류를 허용하는 근사한 수치를 통한 현상의 발견을 필요로 하기 때문에, 센서 네트워크에서의 선형 회귀 분석 기법을 이용한 소실된 데이터를 보정하는 것은 큰 의미를 가질 수 있다.

**3. 제안하는 인-네트워크 병합 질의 처리 기법**

본 장에서는 제안하는 데이터 소실을 고려한 인-네트워크 병합 질의 처리 기법을 위해 필요한 데이터 수집과 데이터 취합 및 다중 회귀 분석 기법을 통한 질의 결과 보정 기법에 대해서 기술한다.

**3.1 질의 배포 및 데이터 전송**

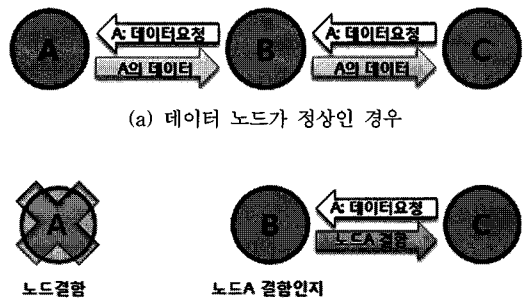
제안하는 기법의 질의를 배포하고 데이터를 취합 단계는 기본적으로 기존의 연구에서 적용하였던 기법을 사용한다. 임의의 노드에 질의가 배포되고, 질의에 해당하는 특정 데이터의 저장을 담당하는 센서 노드에 질의에 해당하는 결과 데이터를 요청한다.

예를 들어, 기상청에서 통계를 작성하기 위해 “지난 밤 최저 기온을 기준으로 25°C 이상인 열대야가 발생한 지역의 수를 알려달라”는 질의를 배포했다고 가정해보자. 그림 4에서 질의의 결과를 반환하기 위해, 25°C 이상의 데이터를 저장하고 있는 노드  $N_{Data} = \{N1, N3, N4, N5, N6, N7\}$ 에 데이터를 요청한다. 질의의 대상이 되는 데이터 저장 노드가 결함이 없이 데이터를 모두 전송하였을 경우는 정상적으로 수집된 결과를 반환하고 질의 처리를 종료한다. 하지만 질의의 대상이 되는 노드 중에서 결함이 발생하면, 그림 4에서의 N4 노드와 같이 결함 노드에서 저장하고 있던 전체 데이터가 소실되기 때문에, 결과에서 누락되고, 질의의 정확도는 낮아진다.

질의가 발생하면 질의에 해당하는 데이터 저장 노드로 데이터 요청 메시지가 전송이 된다. 기존의 전체 원

본 데이터를 질의 노드로 전송 하였던 데이터 중심 저장 기법과 달리 본 기법에서는 인-네트워크 병합 질의 처리를 수행하여 요약 정보만을 전송한다. 뿐만 아니라, 본 논문에서는 병합 질의 처리의 높은 효율성 및 데이터 보정 기법의 수행을 위해 중간 집계 노드를 생성한다. 중간 집계 노드는 질의를 배포하기 위해 생성되는 라우팅 트리의 상위 노드로서 그림 4와 같이 질의를 다수의 노드로 분배하는 최초의 노드가 선정된다. 인-네트워크 병합 질의를 수행함으로써, 기존 데이터 중심 저장 기법에 비해 큰 에너지 효율성을 보이게 되고, 이는 결과적으로 네트워크 생존 시간을 증가시킨다.

일반적으로 센서 네트워크에서는 인접 노드와의 주기적인 통신(e.g. Heartbit)을 통해 노드의 결함을 인지할 수 있기 때문에, 수집해야 할 데이터 저장 노드의 가장 근접한 노드가 해당 노드의 결함을 인지했을 경우, 데이터 요청 메시지를 전송하지 않고, 노드 결함 메시지를 생성하여 전송한다. 노드 결함 메시지는 최종적으로 중간 집계 노드(Aggregator)로 전송이 된다. 중간 집계 노드는 노드 결함 메시지를 전달 받았을 경우 취합한 데이터를 기반으로 누락 데이터에 대한 가상의 데이터를 생성한다.



(a) 데이터 노드가 정상인 경우

(b) 데이터 노드에 결함이 발생한 경우

그림 5 질의 배포와 노드의 결함

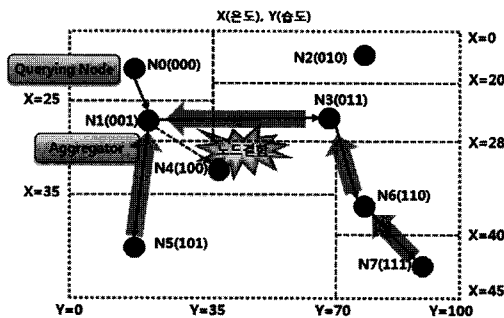


그림 4 질의 배포와 노드의 결함

**3.2 누락 데이터 보정**

중간 집계 노드가 결함 메시지를 받았을 경우, 질의 배포 및 데이터 전송 과정을 통해서 취합한 데이터에 기반한 선형 회귀 분석 기법을 이용하여 보정 모델을 생성한다. 그 후 누락 데이터에 해당하는 가상 데이터를 질의 결과에 반영하여 보정을 수행한다.

선형 회귀 분석 기법을 수행하는 첫 단계는 수집 데이터의 산포도를 분석하는 것이다. 데이터 수집 과정을 통해서 수집된 데이터를 취합하여 각 변수에 따른 데이터 산포도를 생성한다. 앞서 언급한 예제에서는 온도와 습도 데이터를 수집한 후, 그림 6과 같은 데이터 산포도를 생성한다. 생성된 산포도를 바탕으로 선형 회귀 분석

기법을 적용하여 누락 데이터 보정 모델인 식 (2)를 구할 수 있다. 데이터가 완전히 소실 된 경우에는 선형 회귀 분석 기법을 수행하기 위한 데이터 포인트가 존재하지 않게 된다. 그러므로 이러한 경우에는 소실되지 않은 전체 데이터 기반의 선형 회귀 분석 기법과 소실 영역의 인접한 데이터 기반의 선형 회귀 분석 기법을 수행한다. 소실되지 않은 전체 데이터 기반의 선형 회귀 분석 기법의 경우, 데이터 소실 영역의 모든 데이터 점을 대상으로 선형 회귀 분석을 수행한다. 그림 6의 예제에서 산포도를 바탕으로 전체 데이터 점 집합  $P = \{(0, 0), (1.1), \dots, (10, 8), \dots, (43, 2), (44, 0)\}$ 을 추출한 후, 선형 회귀 분석 기법을 수행하여 오차가 가장 적은 직선의 방정식을 구하게 된다. 회귀 분석 기법을 통하여 생성된 보정 모델은 그림 7과 같다. 그림 7에서 실제 수집 데이터와 비교하여 오차를 최소화하여 생성한 보정 모델로 가상의 데이터를 생성하고, 이를 질의 결과에 반영함으로써 높은 정확도의 질의 결과를 기대할 수 있다.

$$Y = 32.4 - 0.3X + \epsilon \quad (2)$$

전체 데이터 기반의 선형 회귀 분석 기법의 경우, 누락 데이터의 보정을 위해 수집한 전체 데이터를 분석하여 생성한 보정 모델도 기존의 기법에 비해 향상된 정

확도를 보인다. 하지만 선형 회귀 분석 기법을 이용한 누락 데이터 보정 모델은 직선 형태를 가지게 되므로, 전체 데이터를 고려하여 모델을 생성하였을 경우, 질의 결과의 높은 정확도를 보이는 것은 불가능하다.

이를 해결하기 위해 본 제안하는 기법에서는 센서 데이터의 특성을 이용한다. 일반적으로 센서 네트워크에서 수집하는 자연 현상은 선형 분포를 가진다. 다시 말해, 임의의 구간의 데이터 분포는 인접한 구간의 데이터와 유사한 분포를 보이게 된다. 그림 7은 센서를 통해 수집한 데이터의 특성을 나타낸다. 센서에서 수집한 데이터는 환경적인 특성상 인접한 물리량과 선형적인 연관성을 지닌다. 그림 7(a)는 자동차의 디스크, 패드, 브레이크 액에 센서를 부착하여 주행 중 브레이크 페달을 밟는 순간 각 부위의 온도 변화를 나타낸다. 이때, 온도는 시간이 변함에 따라 연속적으로 변화한다. 이와 같이 센서를 통해 측정하는 대부분의 물리량은 급격하게 변화하기 보다는 연속적인 값을 가진다. 센서 데이터는 그림 7(b)와 같이 공간적으로도 연관성이 있다. 그림 7(b)는 우리나라 주변의 해수온의 분포를 나타낸 그림이다. 일반적으로 해수온의 분포는 지리적 연관성에 근거하여 인접한 지점과 유사한 온도 분포를 가진다. 이는 공간적으로 인접한 지역에 배치되어 있는 센서들이 유사한 값을 수집했다는 것을 나타내고, 이에 대한 분포도를 작성할 경우, 특정한 물리량이 인접한 물리량에 유사한 패턴을 지니게 될 것이라는 것을 알 수 있다. 이러한 센서 데이터의 특성을 활용하여 소실된 데이터 구간의 인접한 구간만을 고려한 보정 모델을 생성하면 더욱 정확도 높은 보정 모델을 생성할 수 있다. 그러므로 제안하는 기법에서는 전체 데이터 기반 보정 모델뿐만 아니라, 소실 영역 인접 구간의 수집 데이터만을 고려한 보정 모델도 생성한다.

소실 영역 인접 구간을 고려한 데이터 보정 기법은 데이터 소실 영역이 시작 되는 데이터 점과 데이터 소

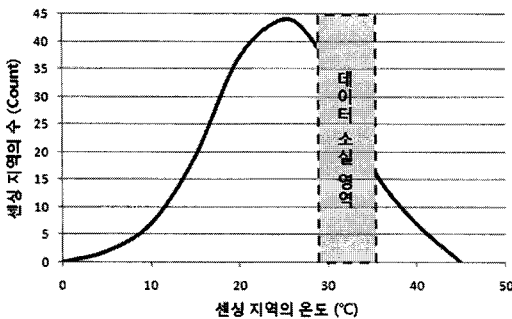
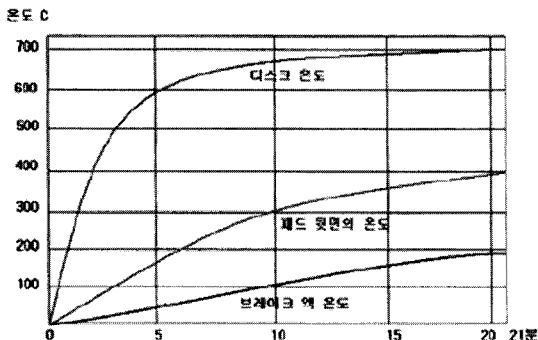
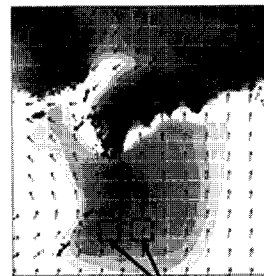


그림 6 데이터 취함에 따른 산포도



(a) 시간적인 연관성



(b) 공간적인 연관성

그림 7 센서 데이터의 특징

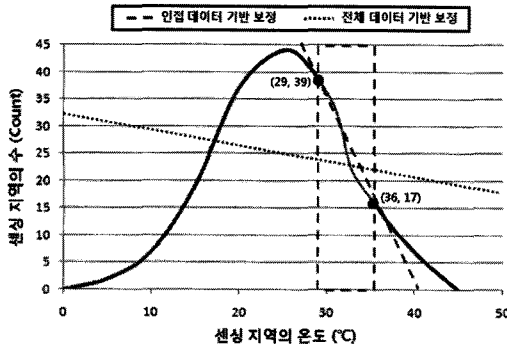


그림 8 선형 회귀 분석 기법을 이용한 보정 모델

실 영역이 종료되는 마지막 데이터 점을 대상으로 선형 회귀 분석 기법을 수행한다. 예를 들어, 그림 6의 예제는 데이터 소실 영역의 모든 데이터가 소실 된 상태이므로, 데이터 소실 영역이 시작 되는 데이터 점 (29, 39)와 데이터 소실 영역이 종료되는 마지막 데이터 점 (36, 17)을 기반으로 선형 회귀 분석을 수행하여 누락 데이터 보정 모델인 식 (3) - <두 데이터 점을 포함하는 선형 방정식>를 구할 수 있다. 그림 7은 전체 데이터 기반 보정 모델과 인접 데이터 기반 보정 모델을 나타낸다. 그림 8에서 전체 데이터 기반 보정 모델은 전체의 산포도를 고려하여 가상의 데이터를 생성하기 때문에 실제 데이터와 비교했을 경우, 결과 정확도가 낮아지는 것을 확인 할 수 있다. 하지만 인접 데이터 기반의 보정 기법은 실제 데이터와 약간의 오차는 존재하지만 매우 높은 수준의 가상의 데이터를 생성하는 것이 가능하다.

$$Y = 130.14 - 3.142X + \varepsilon \quad (3)$$

데이터 중심 기법의 특성상 특정 데이터 영역의 전체 데이터가 소실되는 경우도 발생하지만 그림 9와 같이 일부 데이터만 소실되는 경우도 발생한다. 예를 들어, 그림 4에서 N4 노드가 결함이 발생했을 경우, 28~35°C

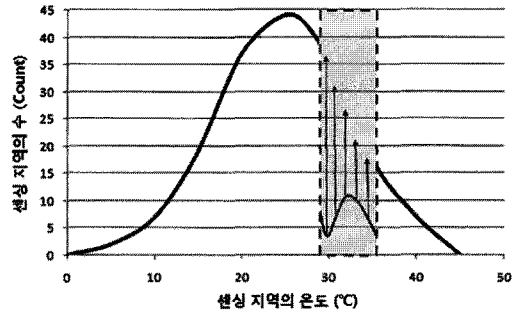


그림 9 발생 빈도를 고려한 보정 모델의 최적화

에 해당하는 데이터가 모두 소실되는 것이 아니라 습도에 따라 N6 노드에는 일부의 데이터가 존재하게 된다. 그러므로 제안하는 기법에서는 일부 데이터가 소실되지 않고 취합되었을 경우, 데이터 보정 모델의 높은 정확도를 보장하고자 기존 현상의 발생 빈도를 고려하여 데이터 보정 모델을 최적화한다. 앞서 설명한 소실 영역 인접 구간을 고려한 보정 모델의 생성에서와 같이 소실 영역의 인접한 데이터를 기반으로 선형 회귀 분석 기법을 수행하여, 선형 방정식을 생성한다. 소실되지 않고 남아있는 데이터 분포에 해당하는 데이터 모델과 보정 모델을 연립하여, 기존의 데이터 발생 패턴을 고려함으로써 최적화된 데이터 모델을 얻을 수 있다.

### 3.3 대체 노드의 선정

제안하는 기법에서 결함이 발생한 인접 노드에서 결함 여부를 파악할 수 있고, 결함 노드 메시지를 중간 집계 노드에서 수신하여 데이터 보정 기법을 수행한다. 데이터 중심 기법에서는 같은 분할 축으로 분할되게 되는 노드가 인접 노드 중에서 반드시 존재하게 된다. 그러므로 제안하는 기법에서는 대체 노드를 선정하기 위한 비용을 최소화하고, 기존의 구조를 유지하기 위해 그림 10과 같이 인접 노드 중에서 동일한 분할 축으로 기준으로 분리된 노드가 결함 노드를 대체하고, 그에 따라 트

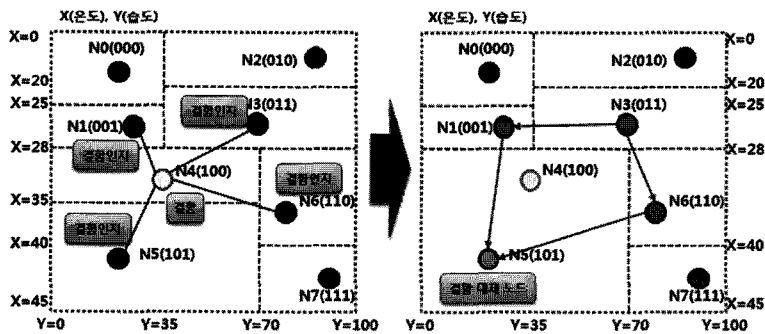


그림 10 결함 노드 대체 기법

리 구조를 수정하게 된다. 하지만 센서 네트워크의 특성 상 지리적으로 인접한 노드가 동시에 유실되는 경우가 발생하고, 결함을 대체할 이웃 노드가 존재하지 않게 된다. 이 때는 결함이 발생한 영역을 커버할 수 있는 노드로 KD-트리 상의 부모 노드가 선정되고, 트리 구조가 수정된다.

#### 4. 성능평가 및 분석

##### 4.1 성능평가 환경

제안하는 기법의 우수성을 보이기 위해 기존의 기법들과 시뮬레이션을 수행하였다. 기존에 연구된 데이터 중심 저장 기법인 GHT, KDDCS와 KDDCS에 제안하는 인-네트워크 질의 처리 기법을 적용한 E-KDDCS(Extended-KDDCS)를 비교 평가하였다. DIM의 경우, KDDCS와 유사한 질의 처리 성능을 보이므로, 본 성능평가에서 제외하였다. 시뮬레이션은 표 1과 같은 성능평가 환경을 구성하여 진행하였다. 또한 본 성능평가에서는 100개의 센서 노드를 센싱 영역 내에 균등하게 분포시키고, KDDCS 기법을 적용하여 동적으로 KD-트리를 생성하여 데이터 영역을 할당하였다. 성능평가의 정확도에 대한 신뢰도 높은 평가를 위해, 센서 네트워크에 배포되는 질의는 임의의 노드 결함에 의해 발생하는 데이터 소실 영역과 같은 범위의 데이터를 반환하는 질의로 한정하였다.

시뮬레이션에서 사용된 데이터[11]는 분포 형태에 따른 가상 데이터의 정확도를 평가하기 위하여, 미국 워싱턴 주에서 측정된 실제 온도 데이터를 사용하였으며, 데이터는 그림 11과 같은 형태의 분포를 가진다.

표 1 성능 평가 환경

파라미터	값
센서네트워크의 크기 (m×m)	100m × 100m
배포된 센서의 수 (개)	100
센서 통신 반경 (m)	10
데이터 패킷의 크기 (Bytes)	4 Bytes
센서 식별자의 크기 (Bytes)	4 Bytes
노드 결함 발생률 (%)	2%

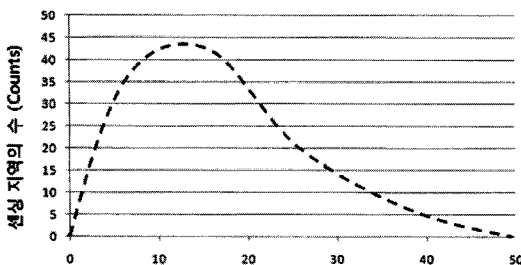


그림 11 시뮬레이션에 사용한 데이터 모델의 분포도

##### 4.2 성능평가 결과

###### 4.2.1 센서 노드 결함에 따른 질의 결과 정확도

노드 결함에 따른 질의 결과 정확도 실험에서는 100m × 100m의 센싱 영역에 100개의 센서 노드를 배치한 후, 임의의 노드에 결함을 발생시키며 결과 보정 기법을 적용하지 않은 기존 데이터 중심 저장 기법과의 성능을 비교 평가 하였다. 그림 12는 센서 노드의 결함 발생에 따른 제안하는 결과 보정 기법을 적용한 기법과 기존 기법과의 질의 결과 정확도를 평가한 결과이다. 센서 노드 결함에 따라 제안하는 기법은 평균 94% 이상의 높은 정확도를 보였다. 기존의 기법 중 GHT의 경우, 제안하는 기법에 비해 높은 저장 비용을 사용하여 데이터가 이웃 노드들에 복사가 되므로, 저장 노드에 결함이 발생하여도 높은 질의 정확도를 보였다. 한편 KDDCS는 노드가 결함이 발생할 경우, 결함 노드의 데이터 전체가 소실이 되므로 질의 결과의 정확도가 매우 낮았다.

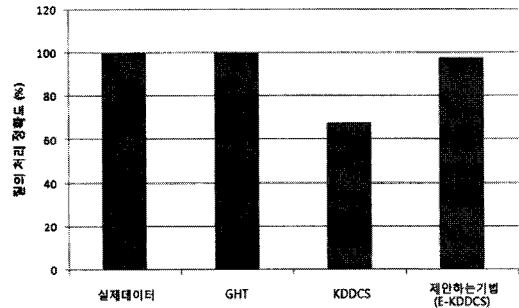


그림 12 센서 노드 결함에 따른 질의 결과 정확도

###### 4.2.2 질의 처리 시의 에너지 소모량

그림 13은 기존 기법과 제안하는 기법 간의 질의 처리 시의 에너지 소모량을 평가한 결과이다. 제안하는 기법의 경우 기존에 제안된 데이터 중심 저장 기법과 달리 질의에 따른 데이터 요청과 수집 시 인-네트워크 질의 처리 기법을 수행함으로써 기존 기법에 비해 평균

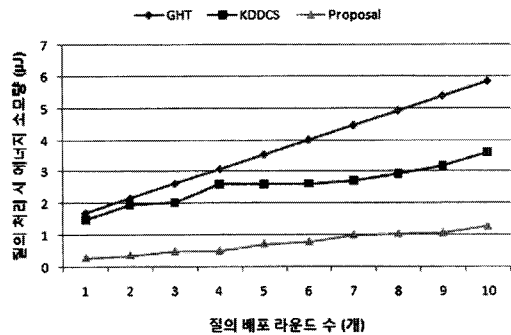


그림 13 질의 처리 시의 에너지 소모량

80% 이상의 질의 처리 비용을 감소시켰다. 기존에 제안된 기법 중 KDDCS의 경우 영역 질의가 배포되면, KD-트리를 이용하여 영역에 해당하는 노드에서만 데이터를 요청하기 때문에 비교적 에너지 효율적이었다.

## 5. 결론 및 향후 연구

본 논문에서는 기존에 제안된 데이터 중심 저장 기법의 문제점을 분석하고, 노드 결합에 따른 데이터 소실을 고려한 인-네트워크 병합 질의 처리 기법을 제안하였다. 센서 네트워크에서의 데이터 중심 기법의 경우 질의를 효과적으로 처리하기 위해 수집한 데이터 값에 따라 저장될 센서 노드를 지정하고, 질의 처리를 위해 질의에 해당하는 데이터를 저장하는 노드에서만 데이터를 수집한다. 그렇기 때문에 데이터를 저장하고 있는 노드에 결합이 발생하면 결합 노드에 저장되어 있는 전체 데이터가 소실됨에 따라 질의 결과 정확도가 저하되는 문제점이 발생한다. 하지만, 기존의 연구에서는 노드 결합이나 네트워크 오류 발생에 데이터 소실 문제가 전혀 고려되지 않았다. 이러한 문제점을 해결하기 위하여 노드 결합에 따른 데이터 소실이 발생 하였을 경우, 소실되지 않은 데이터를 분석하여 누락된 가상의 데이터를 생성하고, 이를 질의 결과에 반영함으로써 높은 질의 결과를 보이는 기법을 제안하였다. 질의에 따라 수집된 데이터를 취합하여 선형 회귀 분석 기법을 이용함으로써 결과 보정 모델을 생성한다. 수집된 데이터의 발생 빈도 및 분포를 고려한 결과 모델을 최적화 시킴으로써 높은 정확도를 가지는 질의 처리 결과를 얻을 수 있다. 성능평가 결과, 평균 94% 이상의 질의 결과 정확도를 보였고, 질의 처리 시 80% 이상의 에너지 소모를 감소시켰다. 향후 연구는 다중 선형 회귀 분석 기법을 적용한 결과 보정 모델을 접목하여 더욱 높은 수준의 질의 결과 정확도를 얻는 기법을 제안하는 것이다.

## 참고 문헌

- [1] Cerpa, A., Elson, J., Estrin, D., Girod, L., Hamilton, M. and Zhao, J., "Habitat Monitoring: Application Driver for Wireless Communications Technology," *Proc. of ACM Workshop on Data Communications in Latin America and the Caribbean*, pp.20-41, 2001.
- [2] Culler, D., Estrin, D., and Srivastava, M., "Guest Editors' Introduction: Overview of Sensor Networks," *IEEE Computer*, vol.37, issue 8, pp.41-49, 2004.
- [3] Madden, S., Franklin, M. J., Hellerstein, J. M., and Hong, W., "TAG: a Tiny AGgregation service for ad-hoc sensor networks," *Proc. of the 5th Symposium on Operating Systems Design and Implementation*, pp.131-146, 2002.
- [4] Ratnasamy, S., Karp, B., Shenker, S., Estrin, D.,

Govindan, R., Yin, L., Yu, F., "Data-Centric Storage in Sensornets with GHT, a geographic hash table," *Mobile Networks and Applications*, vol.8, issue 4, pp.427-442, 2003.

- [5] Ratnasamy, S., Karp, Yin, L., Yu, F., Estrin, D., Govindan, R., Shenker, S., "GHT: A Geographic Hash Table for Data-Centric Storage," *Proc. of the 1st ACM international workshop on Wireless sensor networks and applications*, pp.78-87, 2002.
- [6] Li, X., Kim, Y. J., Govindan, R., Hong, W., "Multi-Dimensional Range Queries in Sensor Networks," *Proc. of the 1st international conference on Embedded networked sensor systems*, pp.63-75, 2003.
- [7] Aly, M., Pruhs, K., Chrysanthis, P.K., "KDDCS: A Load-Balanced In-Network Data-Centric Storage Scheme for Sensor Networks," *Proc. of the 15th ACM international conference on Information and knowledge management*, pp.317-326, 2006.
- [8] Aly, M., Morsillo, N., Chrysanthis, P.K., Pruhs, K., "Zone Sharing: A Hot-Spots Decomposition Scheme for Data Centric Storage in Sensor Networks," *Proc. of the 2nd international workshop on Data management for sensor networks*, pp.21-26, 2005.
- [9] Lai, Y., Chen, H., Wang, Y., "Dynamic Balanced Storage in Wireless Sensor Networks," *Proc. of the 4th workshop on Data management for sensor networks: in conjunction with 33rd International Conference on Very Large Data Bases*, pp.7-12, 2007.
- [10] Cheng, M.-Y., Peng, L., "Simple and Efficient Improvements of Multivariate Local Linear Regression," *Journal of Multivariate Analysis*, vol.97, issue.7, pp.1501-1524, 2006.
- [11] Live from Earth and Mars (LEM) Project, <http://www-k12.atmos.washington.edu/k12/grayskies>



박 준 호

2008년 2월 충북대학교 정보통신공학과 공학사. 2010년 2월 충북대학교 정보통신공학과 공학석사. 2010년 3월~현재 충북대학교 정보통신공학과 박사과정. 관심분야는 데이터베이스 시스템, 무선 센서 네트워크, RFID, 차세대 웹, LMS/LCMS, 바이오인포메틱스 등



이 효 준

2009년 2월 충북대학교 정보통신공학과 공학사. 2009년 3월~현재 충북대학교 정보통신공학과 석사과정. 관심분야는 무선 센서 네트워크, 데이터베이스 시스템, 위치 기반 서비스 등





성 동 옥

2005년 2월 충북대학교 정보통신공학과  
공학사. 2007년 2월 충북대학교 정보통  
신공학과 공학석사. 2007년 3월~현재  
충북대학교 정보통신공학과 박사과정. 관  
심분야는 무선 센서 네트워크, 데이터베  
이스 시스템, FLASH 메모리 저장 시스

템, LCMS/LMS, 위치 기반 서비스 등

유 재 수

정보과학회논문지 : 데이터베이스 참조  
제 37 권 제 1 호 참조