수질센서 데이터 마이닝을 위한 클러스터링 적용 모델

권대현, 조수선
충주대학교 컴퓨터과학과
e-mail:kfsura@nate.com

An Application Model for Clustering in Water Sensor Data Mining

Daehyeon Kweon, Soosun Cho
Dept of Computer Science, Chungju National University

요 약

센서 데이터의 마이닝 기술은 의사결정을 위한 통합정보 및 예측정보를 제공하는 USN 지능형 미들웨어의 주요 구성 요소이다. 본 논문에서는 수질 센서 데이터 마이닝 시스템을 개발하기위해 대표적인 데이터 마이닝 기법인 클러스터링의 적용 모델을 소개한다. 적용 모델의 클러스터링을 통해 중간노드에서의 데이터 이상치 검출과 호스트에서의 시간대별 데이터 변화 검출이 가능하다.

1. 서론

수자원의 보호 및 관리가 환경 문제의 중요한 이슈로 부각된 것은 이미 오래된 일이며 근래에는 이를 자동화하기 위한 여러 가지 시스템의 개발이 이루어지고 있다. 특히 수질 측정을 위한 다양한 센서의 보급과 함께 이 센서들을 USN(Ubiquitous Sensor Network)으로 연결하여 자동 모니터링하려는 시도가 활발하게 이루어지고 있다[1, 2]. 이와 같은 수질 모니터링 시스템을 이용하면 채수 및 수질검사를 담당자가 직접 하는 방식에 비해 실시간으로 수질 데이터의 모니터링 결과를 측정할 수 있어서 이상 데이터 발생에 즉시 대처할 수 있는 이점이 크다. 본 논문에서는 이와 같은 수질 센서 데이터의 모니터링에 필요한 데이터 마이닝 시스템을 개발하기위해 대표적인 데이터 마이닝 기법인 클러스터링의 적용 모델을 소개한다.

2. 연구의 필요성

기존의 수질 데이터의 모니터링 시스템에서는 수질 센서 데이터(pH, DO, 탁도, 온도, 전도도, 암모늄)를 수집하기 위한 센서노드의 컨트롤러 및 ZigBee 모듈의 구현, CDMA 기능이 포함된 중계 노드의 적용, 센서노드로부터 수집된 정보를 유선망으로 전송하기 위한 인터페이스, 데이터의 기준치 초과 시 관리자에게 SMS로 경보 전송 등의 기능을 포함하고 있다. 그러나 모니터링을 위한 하부 시스템의 구축에 집중한 반면 수집된 데이터의 분석을 위한 지능형 미들웨어 기술의 개발 및 적용은 아직 부진한 것이 사실이다.

USN 미들웨어 기술은 센서노드로부터 수집된 데이터를 필터링하고 통합 및 분석하여 의미 있는 상황정보를 추출, 저장, 관리 및 검색하고 응용서비스로 전달하여 서비스간 연계하는 역할을 수행한다[3]. 이 중 지능형 미들웨어의 일종인 센서 데이터의 마이닝 기술은 의사결정을 위한 통합정보 및 예측정보를 제공하는 역할을 하는 주요 구성 요소이다.

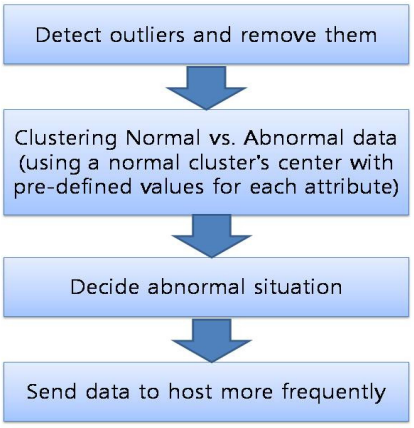
3. 센서 데이터 클러스터링의 특징

센서 네트워크에서 발생되는 데이터는 광범위한 네트워크의 크기와 연속적으로 흘러가는 스트림 데이터의 특성상 기존의 중앙집중식 DBMS에 기반한 방법으로는 처리가 어렵다. 스트림 데이터 마이닝 기법 중 하나인 클러스터링에 있어서도 연속적으로, 대부분 빠른 속도로 전송되는 새로운 데이터를 반영하는 클러스터 구조의 적용이 필요하다. 또한 데이터 흐름상의 변화 검출(change detection)과 오래된 데이터의 무시(forgetting outdated data)를 고려하는 것이 최근의 추세이다. 이러한 환경을 고려할 때, 스트림 데이터의 클러스터링은 다음과 같은 특징[4, 5]을 가져야 한다. 1)간결한 표현, 2)빠르고 점진적인 새 데이터의 처리, 3)전체 데이터 집합에 대한 one-pass 처리, 4)처리 중인 정보, 잔여 시간 등 가치 있는 반응의 제공, 5)점증하는 프로세싱을 위한 정지, 중단, 및 재개 가능, 6)기 정의된 모델에 새 데이터를 통합할 수 있는 능력, 7)제한된 RAM 버퍼 내에서의 처리 등이 그것이다. 본 논문에서는 이와 같은 특징들을 고려하는 센서 데이터 클러스터링 모델을 소개한다.

4. 센서 데이터 클러스터링 적용 모델

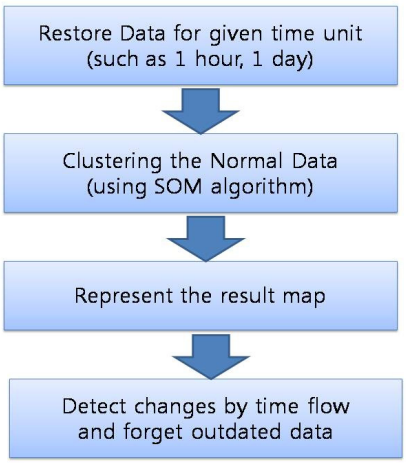
본 논문에서는 USN 시스템 환경으로 수질 측정을 위한 다수의 수질 센서 노드가 호수나 하천에 설치되어 있고, 지역별로 데이터의 수집 및 전송을 위한 중간 노드가 있으며, 전체 시스템을 관장하는 호스트가 있다고 가정한다. 수질 센서 데이터를 분석하는 방법은 여러 가지가 있지만 여기서는 중간노드에서의 데이터 이상치 검출과 호스트에서의 시간대별 데이터 변화 검출을 위해 클러스터링 방법을 적용할 수 있는 두 가지 모델을 소개한다.

먼저 중간 노드에서 센서 노드들로부터 수집된 데이터를 클러스터링을 통해 정상 데이터와 비정상 데이터로 구분해주는 기능을 생각해보자. 이 때, 비정상 데이터는 데이터 자체의 에러와는 구분되며, 온도, 탁도, pH 등 수질 측정 데이터에서 관심을 가져야할만한 이상 값이 관측되었음을 의미한다. 센서 네트워크의 부하를 줄이고 저전력 설계를 하기 위해서는 중간노드에서 호스트로의 데이터 전송이 평상시에는 일정한 시간 간격으로 이루어지고, 이러한 이상치가 발생되었을 때 더 자주 일어나게 하는 전략이 필요하다. 이를 위해 중간 노드에서의 클러스터링이 효과적이다. 중간 노드에서의 클러스터링 적용 모델을 단계별로 표현하면 다음과 같다.



<그림 1> 중간노드에서의 클러스터링 적용 모델

다음으로 호스트에서의 적용 모델을 살펴보면, 이것은 센서 데이터의 클러스터링을 적용하여 시간대별로 데이터 분포의 변화를 감지하기 위한 것이다. 호스트에 전송된 수질 센서 데이터는 일정한 시간 간격을 두고 클러스터링함으로써 시간대별로 데이터의 분포가 어떻게 변하는지 확인할 수 있다. 이를 위해 다차원 데이터의 2차원 그래픽 표현이 가능한 클러스터링 기법인 SOM (Self-organizing Map) 알고리즘을 사용한다면 마치 기상도에서 구름의 분포가 변하는 것을 시간대별로 볼 수 있는 것처럼 클러스터링의 결과를 시각적으로 확인할 수 있게 된다. 이와 같이 시간대별로 변화를 감지할 수 있는 시각적 클러스터링 결과는 여러 가지 의사 결정에 유용하게 쓰일 수 있을 것이다. 호스트에서의 클러스터링 적용 모델을 단계별로 표현하면 다음과 같다.



<그림 2> 호스트에서의 클러스터링 적용 모델

5. 결론 및 향후 연구

본 논문에서는 수질 센서 데이터 마이닝 시스템을 개발할 목적으로 중간노드에서의 데이터 이상치 검출과 호스트에서의 시간대별 데이터 변화 검출을 위한 클러스터링 적용 모델을 소개하였다. 연구의 최종 목표가 수질 센서 데이터 마이닝 시스템의 개발이므로, 후속 작업으로는 적용 모델을 바탕으로 시스템을 설계하고 구현해야 한다. 이 때, 본 연구팀의 또 다른 연구 성과물인 수질 센서 데이터 발생기를 이용하여 여러 가지 상황 별 시뮬레이션을 통해 개발되는 시스템의 성능과 안정성을 높일 수 있을 것으로 기대한다.

참고문헌

[1] 한국정보사회진흥원, “USN 기반의 소양강 상류천 수질관리를 위한 정보수집 시스템 구축”, 유비쿼터스 서비스 모델 발굴을 위한 USN 현장시험 연구과제, 정보사회진흥원 홈페이지 (http://www.nia.or.kr/), 2006

[2] 한국정보사회진흥원, “u-태화강 테스트베드 구축사업”, u-City 테스트베드 구축과제 종합 결과보고서, 2008

[3] 김석우, “센서 네트워크 연구개발 및 상용화 사례”, 정보통신연구진흥원 주간기술동향 통권 1325호, 2007. 12

[4] D. Barbara, "Requirements for clustering data streams," SIGKDD Explorations, 3(2), Special Issue on Online, Interactive, and Anytime Data Mining, 2002

[5] P. Bradle, U. Fayyad, C. Reina, "Scaling clustering algorithms to large databases," In proc. of the 4th international conference on Knowledge Discovery and ata Mining, 1998