수질 모니터링 시스템에서의 K-means 클러스터링 모델

권대현, 조수선
충주대학교 컴퓨터정보공학과
e-mail : kfsura@nate.com, sscho@cjnu.ac.kr

A K-means Clustering Model on a Water Quality Monitoring System

Daehyeon Kwon, Soosun Cho
Dept. of Computer Science & Information Engineering, ChungJu National University

요 약

본 논문에서는 USN환경에서 수질 모니터링 시스템의 일부인 싱크노드에서의 클러스터링 모델을 설계하였다. 싱크노드에서 수집된 많은 데이터 중 핵심 데이터만을 전송하기 위해서 많은 연구들이 진행중에 있다. 본 논문에서 사용된 K-means 클러스터링 모델은 비슷한 속성들로 이루어진 K개의 클러스터로 데이터들을 묶어 불필요한 중복을 줄이고 위험 요소로 판단되는 데이터들을 추출하는 모델이다. 실험을 통해서 제안한 시스템의 성능을 다른 시스템과 비교하여 얼마나 더 효과적으로 데이터를 축약하였는지 확인할 수 있었다.

1. 서론

수질 모니터링 시스템의 대표적인 기능은 아래와 같다. 첫 번째는 데이터 수집 및 분석이다. 이는 인-네트워크 단에서 이루어지고 있는 자료 수집을 뜻한다. 두 번째는 모아진 자료를 분석하고 분류하여 사용자가 원하는 정보를 뽑아내는 마이닝(Mining)기능이다. 마이닝은 서비스가 이루어지고 있는 서버에서 주로 이루어졌으나 점점 서버에서 미들웨어로, 미들웨어에서 싱크노드로 내려가고 있는 추세이다[1]. 세 번째는 사용자가 원하는 자료를, 원하는 시점에서, 원하는 모양으로 보여 주는 뷰어기능이다. 이는 서비스에서 이루어진다. 네 번째는 알람(Alarm)기능인데 이것은 수질모니터링 시스템에서 가장 중요하며 최종적인 기능이라고 할 수 있다. 특히 데이터 마이닝 기능은 데이터베이스로부터 과거에는 알지 못했지만 데이터 속에서 유도된 새로운 데이터 모델을 발견하여 미래에 실행 가능한 정보를 추출해 내고 의사결정에 이용하는 과정을 말한다. 즉, 데이터에 숨겨진 패턴과 관계를 찾아내어 광맥을 찾아내듯 정보를 발견해 내는 것이다. 따라서 많은 센서 데이터들 중에서 이상한 패턴을 발견하고 경고를 하거나 사용자에게 보여주기 위해서는 데이터 마이닝 기술이 꼭 필요하며 매우 중요하다.

본 논문에서는 USN환경에서 수질 모니터링 시스템의 일부인 싱크노드에서의 클러스터링 모델을 설계하고, 시스템에서 센서 데이터를 효율적으로 전송하기 위한 다양한 시도를 소개한다. 이어지는 2절에서는 관련연구로 싱크노드에서 데이터 통신의 양을 줄이는 연구에 대해 알아보고, 3절에서는 본 논문에서 제안하는 싱크노드에서의 클러스터링 모델을 자세히 서술한다. 4절에서는 3절에서 소개한 클러스터링(Clustering) 적용 모델의 실험과 성능 평가를 서술하며, 5절에서는 결론을 맺는다.

2. 관련연구

센싱된 데이터를 보다 효과적으로 전송하기 위해 네트워크의 전송방식에 대한 연구가 많이 이루어지고 있다.

연구[2]에서는 모든 센서노드들이 서로 다른 시간 슬롯을 할당 받고, 시간 슬롯에 따라 자신의 데이터를 싱크노드로 전송하는 방법을 택한다. 센서노드들은 자신의 시간 슬롯을 기다리는 동안 이웃 노드들이 싱크노드로 전송하는 데이터를 도청한다. 이때 도청한 모든 값에 대한 평균값을 계산하고, 자신의 값과 이 평균값을 비교하여 동일한 경우 데이터를 전송하지 않는다. 이 연구는 데이터 전송의 횟수를 줄였으나 주변의 환경을 기준으로 전송 유무를 결정하기에 본 논문의 연구와는 차이가 있다.

연구[3]는 센서로부터 수집된 데이터를 손실 없이 저장하는 동시에 시간에 따라 누적 되는 센서 데이터의 이력을 효율적으로 관리할 수 있는 기법이다. 이 방법은 센서의 측정값이 변경된 시점을 기준으로 그 변경된 센서의 값과 센서노드의 ID만을 전송하고 있다. 이러한 방법으로 모든 측정값을 정확하게 전송 할 수 있다 특히 중복 데이터를 전송하지 않기 때문에 전송량의 축소를 할 수 있다. 이 방법은 각각의 데이터 값을 싱크노드에서 마지막 값과 비교하여 변경된 값이 있을 때에만 그 값만을 전송한다.

하지만 이 기법은 단순히 하나의 센서노드에서 여러 개의 센서가 있다면 그 센서별로 분할을 하여 전송을 하기 때문에 단순히 변화가 있다고 모든 값을 전송하진 않는다. 4절의 실험에서 이연구의 중복제거 기법과 본연구의 클러스터링 기법을 비교 분석하여 성능을 입증하였다.

3. 수질 모니터링 시스템의 구조적 기능

USN환경에서의 수질 모니터링 시스템의 기본이 되는 싱크노드에서는 많은 센서노드에서 들어오는 방대하고 끊임없는 스트림 데이터를 모두 호스트로 전송하지 않고 꼭 필요한 데이터들만 찾아내어 전송하는 지능적인 행동을 할 수 있어야 한다. 또한 서비스 서버에서 지원하게 될 시간대별 혹은 요일별 모니터링이나 특정 위치의 현재 센싱 정보를 제공해 주기 위해서는 단순히 특이한 값만을 전송해서도 안된다. 따라서 데이터양을 줄이면서도 특이한 값들은 그대로 보존하기 위한 방법으로 대표적인 데이터마이닝 기법인 클러스터링을 접목하여야 한다[4].

클러스터링의 목적은 많은 센서노드들로부터 짧은 시간동안 대량의 스트림 데이터가 들어 올 때, 수집된 전체 데이터를 호스트로 전송하는 대신 주의가 필요한 이상치 데이터는 모두 전송하고 그렇지 않은 일반 데이터는 클러스터의 센터값만 전송함으로써 전송양을 줄이고자 하는 것이다.

싱크노드의 대략적인 실행 알고리즘은 다음과 같다.

- *센싱 데이터들을 일정한수(예를들면 20개)의 클러스터로 K-means 기법을 사용해 클러스터링한다.*
- *전체 센싱 데이터들의 중심을 찾아 이로부터 각 클러스터와의 거리를 산출한다.*
- *거리가 먼 클러스터부터 이 클러스터에 포함된 센싱 데이터를 전송한다. 이때, 전체 센싱 데이터의 일부(예를 들면, 20%)만 전송한다.*
- *센싱 데이터가 전송되지 않은 클러스터들은 클러스터의 정보(중앙값)와 각 클러스터에 포함된 센서노드 정보(센서노드 ID)만을 전송한다.*

이와 같은 클러스터링 기능을 사용하면 클러스터링된 센싱 데이터 중 일부(예를 들면, 20%)만을 전송함으로써 전송데이터의 양을 줄이고, 의심스러운 데이터는 누락 없이 호스트로 전송할 수 있다. 또한 클러스터 정보와 각 클러스터에 포함된 센서노드의 정보를 함께 보냄으로써 오차가 존재하는 근사치 값을 호스트에서 유추해 낼 수 있게 해준다.

본 연구팀은 연구[4]에서 USN환경에서의 수질 모니터링 시스템을 위한 센서 데이터의 선택적 전송방법에 대하여 서술하였다. 그 연구에서는 센서노드에서의 센서 매니지먼트 시스템과 싱크노드에서의 클러스터링모델을 제안한 것이다. 이때 싱크노드에서의 클러스터링 모델은 최소한의 데이터만을 보내는 목적으로 설계가 되어 대량의 데이터가 한 번에 들어올 때 유용하지만 200개미만의 센서노드를 가지고 있는 시스템에서는 너무 적은 양의 데이터만 전송되므로 데이터의 신뢰성이 낮아지는 단점이 있었다. 본 연구에서는 데이터 전송비율을 20%로 확장하여 더 많은 실데이터를 보내면서도 타 연구[3]와의 비교에서는 더 우수한 성능을 보이는 클러스터링 기법을 구현하였다

4. 실험과 평가

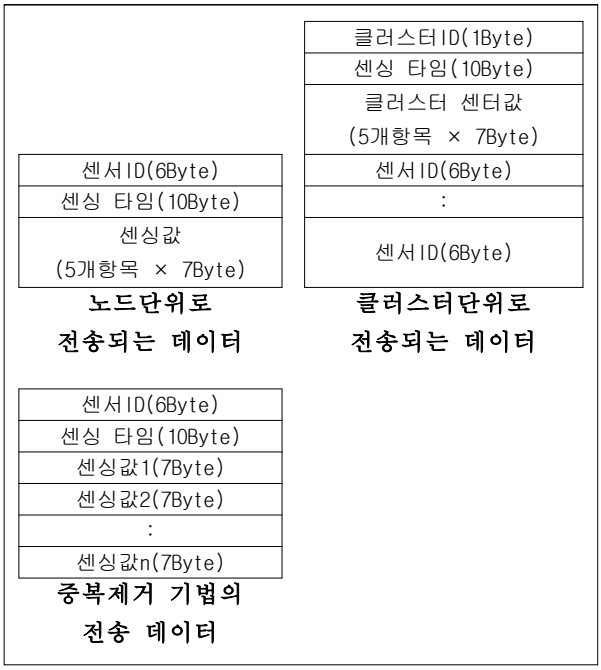
클러스터링 모델로는 K-means 기법을 사용하였다. K-means 기법은 n개의 객체들의 집합을 K개의 군집으로 나누기위해 거리에 기반을 둔 클러스터링 기법이다. 본 연구팀에서는 미국 지질조사국(U.S. Geological Survery ; USGS)에서 제공하는 실시간 수질 모니터링 시스템[5]의 자동 수집 자료 중 캔사스 주의 데이터를 이용하여 제안하는 클러스터링 모델을 평가하였다.

실험을 위해 캔사스 주에 위치한 2군데의 강에서 수집한 데이터(용존산소량, pH, 전기전도도, 온도, 엽록소)를 사용하였다. 실험에서는 2010년 5월 5일부터 7월 6일까지 15분 단위로 수집된 데이터 중 새벽 12시 15분부터 2시 30분까지 10회 측정한 자료를 사용하였다. 이때 자료측정 기간인 126일 동안 측정 시간이 같은 자료를 같은 시간에 126개의 장소에서 수집된 자료라고 가정하였다. 이와같은

<표 1> 클러스터링을 이용한 데이터 전송량

Time	# of total data (A)	# of clusters including top 20% data	# of sensing value transmissions	# of center value transmissions	# of total transmissions (B)	Ratio(%) (B/A)
00:15	126	5	28	14	42	33.33%
00:30	126	5	28	14	42	33.33%
00:45	126	5	28	14	42	33.33%
01:00	126	5	26	14	40	31.75%
01:15	126	5	28	14	42	33.33%
01:30	126	5	28	14	42	33.33%
01:45	126	5	28	14	42	33.33%
02:00	126	5	28	14	42	33.33%
02:15	126	5	28	14	42	33.33%
02:30	126	5	28	14	42	33.33%

측정자료를 사용하여 비교분석 대상인 중복제거 기법[3]과 본 논문에서 제안하는 클러스터링 기법과의 데이터 전송량을 비교 분석하였다.



(그림 1) 전송 데이터의 구조

중복제거 기법은 각각의 데이터가 직전의 데이터와 비교하여 다른 부분만을 분리해서 전송하였다. 15분 전의 데이터와 온도, pH의 값만 다르다면 센서ID와 온도, pH값만을 전송하는 방법이다. 또한 싱크노드의 Time_segment와 Time_point를 저장한 후 다음 센서 데이터가 들어오면 다시 비교를 하여 변경사항만을 추출하여 전송한다.

제안한 K-means 클러스터링 기법은 3절에서 설명한 알고리즘에 따라 센서 값의 전체 평균을 A라고 할 때, 20개의 클러스터를 기준 A에서의 거리가 큰 순으로 정렬한 후 총 데이터의 수의 상위 20%의 데이터를 포함하는 클러스터의 모든 데이터를 전송한다. 따라서 정확하게 20%가 되기는 어렵고 많은 경우 20%를 초과하게 된다. 그 외의 데이터를 가지고 있는 클러스터에서는 각각의 센터 값과 클러스터 ID, 해당 클러스터에 포함되어 있는 센서노드 ID와 센싱 타임 정보만을 전송한다.

<표 1>은 수집된 전체 데이터를 전송하는 것에 비해 클러스터링을 이용함으로써 전송 수량을 33%내외까지 줄일 수 있음을 보여준다. 이때 20%는 실제 센싱값이 전송됨으로 센싱값의 손실도 적은 편이라 할 수 있다. <표 2>는 중복제거 기법의 센싱값 전송량을 나타낸 표로서 <표 1>과 비교해보면 약 13% 정도 더 많은 데이터의 전송을 한 것을 알 수이다.

또한 중복제거 기법의 전송비율이 <표 2>에서는 46~50%의 범위이지만 상황변화가 클수록 중복제거 효과가 떨어질 것이므로 이 비율은 증가할 것으로 예상된다.

〈표 3〉 실제 데이터 전송량 비교

Time	중복제거 기법		클러스터링 기법	
	전송량	전송비율	전송량	전송비율
00:15	3332	51.85%	2660	41.39%
00:30	3402	52.94%	2660	41.39%
00:45	3304	51.42%	2660	41.39%
01:00	3339	51.96%	2570	39.99%
01:15	3262	50.76%	2660	41.39%
01:30	3227	50.22%	2660	41.39%
01:45	3234	50.33%	2660	41.39%
02:00	3248	50.54%	2660	41.39%

(그림 1)은 제안한 클러스터링 기법과 중복제거 기법의 데이터 전송 구조이며, <표 3>은 두 기법의 실제 데이터 전송량을 비교한 것이다. <표 3>에서 두 기법의 전송 비율이 <표 1> 및 <표 2>의 전송비율보다 높은 것은 센서 정보인 메타 데이터를 포함하기 때문이다. <표 3>에서 확인한 바와 같이 제안한 클러스터링 기법에서 실제 전송 데이터량이 10%정도 더 적은 것으로 나타났다.

5. 결 론

본 논문에서는 USN환경에서의 수질모니터링 시스템에

<표 2> 중복제거 기법의 센싱값 전송량

Time	# of total data(A)	Chlorophyll	DO	pH	EC	Temperature	Turbidity	# of total transmissions (B)	Ratio(%) (B/A)
00:15	756	119	41	9	68	44	87	368	48.68%
00:30	756	114	39	9	77	43	96	378	50.00%
00:45	756	114	35	8	67	48	92	364	48.15%
01:00	756	109	32	7	70	48	103	369	48.81%
01:15	756	112	33	3	73	46	91	358	47.35%
01:30	756	113	26	4	69	48	93	353	46.69%
01:45	756	109	31	7	70	43	94	354	46.83%
02:00	756	103	30	12	72	42	97	356	47.09%
02:15	756	114	25	5	77	51	91	363	48.02%
02:30	756	108	29	10	70	46	98	361	47.75%

서 핵심적인 역할을 하는 싱크노드에서의 K-means 클러스터링 모델을 제안하고 타 연구와의 비교 실험을 통해 더 우수한 성능을 확인하였다. 제안된 기법은 싱크노드에서 전달받은 센싱 데이터들 중에서 필요 없는 많은 데이터들을 제거하고 필요한 것들만을 걸러내어 전송하기 때문에 배터리의 사용량, 데이터의 전송량, 호스트에서의 처리량 감소까지 많은 부분에서 이익을 얻을 수 있는 효율적인 모델이다

참고문헌

[1] 김민수, 이용준, 박종현, USN미들웨어 기술개발동향, 텔레매틱스, RFID/USN, GIS 융합기술 동향 특집 논문, 제22권, 제3호, 한국전자통신연구원, 2007년 6월.

[2] X. Meng, L. Li, T. Nandagopal and S. Lu, “Event contour: An Efficient and Robust Mechanism for Tasks in Sensor Networks”, Technical Report, UCLA, 2004.

[3] 이양구, 류근호, “센서데이터의 시간 정보를 이용한 이력 정보 관리”, 한국공간정보시스템학회 논문지, 제10권 4호, 2008. 12.

[4] 권대현, 조수선, “수질 모니터링 시스템을 위한 센서 데이터의 선택적 전송방법, 인터넷정보학회논문지, 제11권 4호, 2010. 8.

[5] USGS(U.S. geological Survey), USGS Water Data for Kansas homepage (http://waterdata.usgs.gov/ks/nwis)