대용량 교통 데이터의 자료처리 과정과 시스템의 개발

정수정*, 송수경*, 이민수*, 남궁성**
*이화여자대학교 컴퓨터정보통신공학과
**한국도로공사 도로교통기술원 교통연구그룹
e-mail : bloom01@ewhain.net

Development of data processing method and system for huge Highway Data

Sujeong Cheong *, Sookyung Song*, Minsoo Lee *, Sung Namgung**
*Dept. of Computer Science and Engineering, Ewha Womans University
**Korea Expressway Corporation & Transportation Technology

요 약

교통 관련 검지기 시스템에 의해 수집된 교통량, 점유율, 속도와 같은 교통 정보 데이터는 품질 평가, 오류판단, 결측보정의 자료처리를 거치게 되며 이러한 전처리 후 다양한 목적에 의해 연구자들에게 활용된다. 신속하고 정확한 자료처리와 보다 편리하고 효과적인 웹 UI 의 제공은 매우 중요하다. 본 논문에서는 품질평가, 오류판단, 결측보정에 해당하는 세 단계의 자료처리 알고리즘을 개발하고 사용자에게 자료처리의 과정을 제공하는 웹 UI 시스템을 구현한다.

1. 서론

차량 검지기 시스템으로부터 수집된 교통량, 점유율, 속도와 같은 교통 정보 데이터는 각종 자료처리의 과정을 거친 후 대용량 데이터베이스 시스템에 저장되어 연구자들에게 활용된다. 현재 막대한 양의 교통 정보 데이터는 무분별하게 데이터베이스에 적재되어 그 데이터의 활용이 어려운 상태이다. 특히 정확한 자료처리의 과정을 거치고 있지 않으며 사용자를 위한 웹 UI 가 전혀 제공되지 않고 있다. 보다 효율적인 데이터의 활용을 위해서는 체계적이고 정확한 자료처리의 과정이 요구되며 사용자에게 자료처리의 과정을 웹으로 접근 가능하게 하여 편리하고 신속하게 사용자의 의견을 반영하고 도출된 결과 데이터를 활용할 수 있도록 하여야 한다.

본 논문에서는 데이터의 완전성과 유효성을 판단하는 품질평가와 오류 데이터를 식별할 수 있도록 하는 오류판단, 유효하지 않은 자료의 값을 보정하는 결측보정의 자료처리 기법을 연구하고 사용자에게 전 자료처리의 과정을 웹으로 제공하여 의견을 반영하고 그 과정을 인지할 수 있도록 하며 결과 데이터를 획득하여 다양한 목적을 위한 활용이 이루어질 수 있도록 하였다.

2. 관련 연구

캘리포니아의 교통부에서 운영하는 교통 데이터 관리 시스템인 PeMS(Performance Measurement System)는 루프 검지기에 의해 수집된 교통 정보 데이터를 처리하고 저장하는 시스템이다. PeMS 는 1997 년, 교통 관리 시스템 운영에 대한 효과를 평가하기 위한 목적으로 UC 버클리와 CalTrans 의 합작으로 개발되었다. 2003 년에 오늘날의 PeMS 버전 6.2 와 유사한 PeMS 버전 4.0 이 배포되었다. 웹 기반 서비스를 제공하는 이 시스템의 이점은 검지기 상태 정보를 제공하는 기능을 갖추고 있고 사용자가 인터넷을 통하여 쉽게 요청한 정보로 접근할 수 있다는 것이다. 간단하게 사용자가 원하는 자료에 해당하는 구간, 시간 부분을 선택하면 시스템은 데이터를 검색하여 사용자에게 텍스트 파일과 엑셀 파일로 검색 결과를 제공하고 저장 가능할 수 있게 한다. 사용자는 시스템의 사용자 인터페이스의 도움을 받아 스스로 데이터를 해석할 수 있고 편리하게 이용할 수 있다. PeMS 는 ADMS 시스템의 좋은 예이다.

캘리포니아는 단일 루프 검지기를 통하여 수집된 자료를 이용하여 속도를 예측하기 위해 g-factor 를 개발하여 자료 처리 과정에 적용하며 5 분 단위 자료로 집계하여 속도를 계산할 뿐 아니라 outlier 를 필터링하고 누락된 결측 자료는 과거의 정상 자료를 이용하여 보정하고 이렇게 처리된 자료는 1 시간, 1 일 단위로 재 집계된다. PeMS 는 교통 이력 자료 뿐 아니라 실시간 생성 자료에 대해 쉬운 이해를 제공할 수 있도록 그래프 및 도표와 통계 측정치를 표출할 수 있고 사용자가 선택한 지역의 과거 30 일 간의 자료를 9 가지 항목으로 간단 명료하게 나타내어 주는 서비스를 제공한다. 보다 신속하고 정확한 자료처리의 과정을 거쳐 사용자에게 시각화된 정보를 제공하여 쉬운 이해를 가능하게 함으로써 우리나라 자료처리 시스템에도 선별하여 적용시킬 필요가 있다.

3. 자료처리 과정의 개발

캘리포차량 검지기 시스템으로부터 수집된 교통 정보 데이터는 세 단계의 자료처리 과정을 거치게 된다. 그 자료처리 과정은 품질평가와 오류판단, 결측보정의 단계이다.

품질평가란 완전성과 유효성을 체크하는 데이터의 질에 대한 수준을 수치화하여 표출하도록 하는 자료처리의 과정이다. 완전성이란, 루프· 지점별로 수집된 전체의 원시자료인 교통량(volume; vol), 점유율(occupancy; occ), 속도(speed) 데이터 중에서 누락되지 않은, 즉 결측 데이터를 제외한 자료의 비율을 말한다. 완전한 자료란 누락되어 있지 않은 이용 가능한 자료이다. 결측 자료(missing data)는 차량 검지기 시스템의 고장이나 차량의 부재에 의해 발생하며 논문에 적용한 원시자료의 결측 데이터는 -111 로 처리하여 품질평가의 과정을 거치게 된다. 완전성이 높다는 것은 이용 가능한 자료의 비율이 높음을 의미한다. 완전성은 전체 원시 데이터의 개수 중 결측 되지 않은 데이터의 개수를 퍼센트화하여 계산된다. 유효성이란, 오류가 없는 자료의 비율을 말하며 이 때 오류판단 기준에 의거하게 된다. 유효성이 높다는 것은 정상 데이터의 비율이 높음을 의미한다. 유효성은 결측 되지 않은 데이터의 개수 중 오류가 없는 데이터의 개수를 퍼센트화하여 계산된다.

오류판단이란 오류 데이터를 구별하도록 하는 자료처리의 과정이다. 오류판단 기준에 의해 오류로 판단되는 데이터는 -999 로 변경되어 쉽게 정상 데이터와 구분할 수 있다. 오류 판단 기준으로는 임계값 검사(Data Threshold)와 관계 검사(Data Relation)가 있다. 이 때 오류 판단 기준은 소정의 한국도로공사 FTMS 오류 판단 기준 개선안에 따르게 된다.

결측보정이란 원시 자료 중 결손 자료와 오류자료를 다른 시공간 유효 자료를 참조하여 새로운 값으로 보정하는 자료처리의 과정이다.

본 논문에서는 동적 PL/SQL 을 이용하여 품질평가, 오류판단, 결측보정의 전 자료 처리 과정에 입력 /출력 테이블을 'v_from_table'과 'v_temp_table'로 파라미터화하여 사용자가 원하는 대상 테이블에 대한 처리가 이루어지도록 하고 결과 데이터가 저장될 수 있도록 하였다. 또한 지역별 차선 변경에 대한 FIND_COUNT_OF_LOOP1 함수를 도입하여 차선 개수를 차선 id 에 의해 융통적으로 적용할 수 있도록 하였고 다양한 변수를 활용하여 각 자료처리의 간단한 계산식이 도입될 수 있도록 하였으며 필요한 조건을 파라미터화하여 원하는 정보를 얻을 수 있도록 하였다.

품질평가에서는 완전성과 유효성의 계산식을 간단명료히 하여 연구자의 알고리즘 분석의 이해를 쉽게 할 수 있도록 하였으며 정확한 측정이 이루어지도록 하였다.

<표 1> 데이터 품질평가 구현 코드 부분

```
// v_statement is the number of occupancies that are
// not '-111'.   There are 23 loop detectors.

// identify the ones that are not -111
    if v_from_table.occupancy0 != -111 then 1
    … if v_from_table.occupancy23 != -111 then 1
//sum of the identified ones
    v_statement = count(v_from_table.occupancy0) +…+
count(v_from_table.occupancy23)
 v_loop_count_o = v_statement
 …

//Check Data Threshold and Data Relation conditions
    if ((v_from_table. loop_v0 != -111) > 0 or >30)
or…((… *(v_from_table. loop_o23 != -111)/12) >18 *
(v_from_table. loop_v23 != -111)) then 1
  …

// calculate results
v_loop_count_invalid = v_statement
v_validity=100-
((v_loop_count_invalid/v_loop_count_v)*100);
…
```

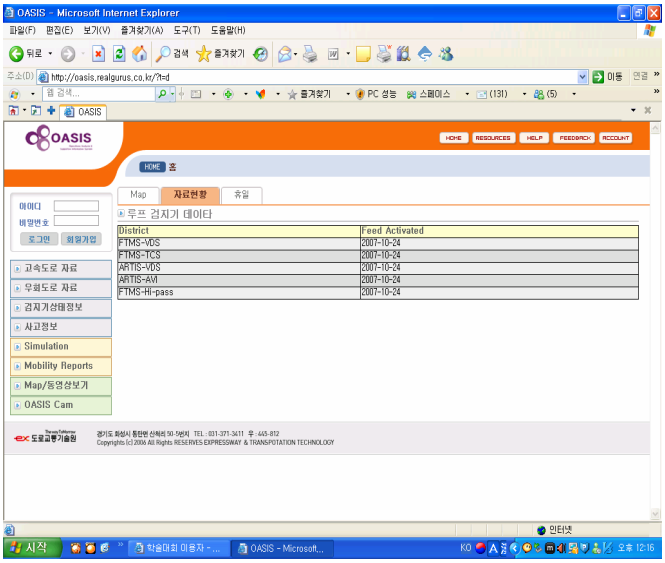
오류판단에서는 오류판단 기준식 중 최대 교통량값을 'v_max_v'로 최대 점유율 값을 'v_max_o'로 쌍루프 간의 교통량 차를 'v_difference_v'로 파라미터화하여 사용자가 원하는 조건 값을 적용할 수 있도록 하였으며 디폴트 값으로 각각 30, 100, 2 를 주석처리하여 인지할 수 있도록 한다. 또한 정확한 오류판단 기준식을 적용하도록 하여 올바른 오류 데이터를 -999 로 변경한다.

결측보정의 자료처리는 크게 두 가지의 기법으로 구현되었는데 인접 지점 참조 기법과 이력 자료 활용 기법이 그것이다. 인접 지점 참조 기법은 동일한 시간 정보와 인접한 공간 정보를 가지는 교통 자료를 참조하는 방법으로 차량의 유입과 출입이 없는 폐쇄된 구간 내의 교통 자료는 서로 유사성을 가지는 특성을 이용한다. 인접 지점 참조 기법으로는 전후 지점 동일 주기 적용, 이전 지점 동일 주기 적용, 지점간 이동 소요 주기 적용 기법이 있다. 이력 자료 활용 기법은 인접한 시간 정보와 동일한 공간 정보를 가지는 교통 자료를 참조하는 방법으로 동일 시간대의 교통 자료는 서로 유사성을 가지는 특성을 이용하여 이력 자료의 유효 자료를 참조하는 기법으로 이력 자료 동일 주기 적용, 차로별 이용률 적용 기법이 있다.

4. 자료처리 시스템의 개발

고속도로 교통 정보 데이터 자료처리를 위한 사용자 인터페이스는 두 가지 주요한 구성을 갖는다. 데이터 자료처리 부분과 데이터 분석 부분이다. 데이터 자료처리 부분은 사용자가 차량 검지기 시스템으로부터 제공된 원시 데이터의 전 처리를 가능하게 하고 전처리된 결과 데이터의 검증과 필터링의 기능을 제공

한다. 데이터 분석 부분은 사용자가 다양한 질의를 요구할 수 있도록 하며 몇 시간 내의 신속한 자료처리 과정을 거쳐 질의에 대한 응답이 이루어지도록 한다. 데이터 분석은 GIS 를 기반으로 한 인터페이스를 갖추며 (그림 1)과 같은 질의를 반영할 수 있다.



(그림 1) 교통 데이터 자료처리 시스템의 GIS 인터페이스

5. 결론

방대한 양의 자료를 저장하고 활용하게 되면서 그 자료를 처리하고 관리하는 과정이 매우 중요하게 되었다. 특히 유효한 자료를 효과적으로 활용하도록 하여야 할 것이다. 올바르지 못한 자료처리는 데이터를 사용하고자 하는 연구자에게 잘못된 결과 값을 제공하여 이 후 결과 데이터를 이용하고자 하는 목적에의 손실을 야기할 수 있다.

본 논문에서는 대량의 교통 데이터를 사용하여 품질평가로 완전성과 유효성을 체크하여 활용 대상 데이터인 원시 데이터의 질을 수치로 확인할 수 있고 오류판단으로 정상데이터를 추출하여 오류데이터와 구별할 수 있도록 한다. 결측보정에 의해 다양한 기법으로 결측 데이터를 유효한 값으로 보정하여 이 후 활용 가능하게 한다. 또한 인터넷을 통한 사용자의 자료처리 과정 접근이 가능하도록 웹 UI 를 개발하여 데이터의 사용을 범용화한다. 이 자료처리 시스템에서는 전 자료처리의 과정에 입력 파라미터를 적용할 수 있도록 함으로써 사용자의 요구를 반영하도록 하고 결과 값을 그래프로 출력하여 이해를 쉽게 하며 GIS 인터페이스를 구축, 실재 지리적 환경을 제공한다.

향 후, 변화하는 교통 환경 조건에 대한 자료처리 알고리즘을 변경하도록 하며 웹 UI 의 기능을 더욱 다양화하여 사용자의 적극적인 활용을 도모하도록 한다.

● 이 연구는 BK21 지원을 받아 수행되었음.

6. 참고 문헌

[1] 김정연, 이영인, 백승걸, 남궁성, “차량 검지자료 결측 보정처리에 관한 연구(이력자료 활용방안을 중심으로)”, 대한교통학회지, 제 24 권 제 7 호, pp.27-40, 2006.

[2] ITS 사업실, 도로교통기술원 교통연구그룹, ‘ITS 구축· 운영 편람’, 한국도로공사, 2005.

[3] Smith, B., and S. Babiceanu. Investigation of Extraction, Transformation, and Loading Techniques for Traffic Data Warehouses. In Transportation Research Record 1879, TRB, National Research Council, Washington D.C., 2004. pp. 9-16.

[4] Smith, B. D. Lewis, R. Hammond. Design of Archival Traffic Databases: Quantitative Investigation into Application of Advanced Data Modeling Concepts. In Transportation Research Record 1836, TRB, National Research Council, Washington D.C., 2003. pp. 126-131.

[5] Al-Deek, H. M., C. Chandra. New Algorithms for Filtering and Imputation of Real-Time and Archived Dual-Loop Detector Data in I-4 Data Warehouse. In Transportation Research Record 1867, TRB, National Research Council, Washington D.C., 2004. pp. 116-126.