

고품질 데이터를 지원하는 교통데이터 웨어하우스 구축 기법

(An Integrated Framework for Data Quality Management of Traffic Data Warehouses)

황재일* 박승용** 나연묵***
(Jae Il Hwang) (Seung Yong Park) (Yun Mook Nah)

요약 본 논문에서는 교통데이터 웨어하우스에서 데이터 품질 관리를 위한 통합기법을 제안한다. 고속도로 교통관리시스템(FTMS)과 우회도로 교통정보시스템(ARTIS) 으로부터 대용량 교통데이터를 수집하여 데이터 웨어하우스를 구축하기 위한 방안을 기술하고, 다양한 분석을 위한 고품질 교통데이터를 제공하기 위한 통합 데이터 품질관리 기법을 제안하고 구현 평가 한다.

제안된 통합 데이터 품질관리 기법을 활용하면 연구자들에게 검증된 고품질 교통데이터를 제공할 수 있고, 데이터처리와 평가를 위한 별도의 비용을 절감할 수 있을 것으로 기대된다.

키워드 : 데이터 웨어하우스, 데이터 품질 관리, 지능형 교통 체계

Abstract In this paper, we propose an integrated techniques for managing data quality in traffic data warehousing environments. We describe how to collect and construct the traffic data warehouses from the operational databases, such as FTMS and ARTIS. We explain how to configure the traffic data warehouses efficiently. Also, we propose a quality management techniques to provide high quality traffic data for various analytical transactions. Proposed techniques can contribute in providing high quality traffic data to the traffic related users and researcher, thus reducing data preprocessing and evaluation cost.

Key words : Data warehouses, Data cleansing, Data quality management, Intelligent transport system

1. 서론

현재 고속도로와 우회도로에서 취득되는 매일 수 GB에 달하는 교통 원시 데이터는 연구 활용 가치가 매우 높은 데이터다. 이렇게 수집되는 교통 원시데이터는 주로 실시간 교통정보제공의 용도로 사용되고 있으며, 축적되는 이력데이터에 대한 활용이 미약하다. 아울러 교통 데이터가 대부분 실시간 정채상태 파악에 활용되는 것에 그치고 있어, 시스템의 용량 범위 내에 저장되어 있는 데이터도 그 활용도가 매우 낮은 실정이다. 최근 교통 데이터를 위한 저장 기법 및 이를 활용하기 위한 알고리즘 등에 관한 연구가 활발하게 진행되고 있다[1, 2, 3, 4, 5].

본 논문에서는 대상 운영계 시스템을 고속도로 교통관리시스템과 우회도로 교통정보 시스템으로 보고 교통데이터 웨어하우스를 구축하는 방안을 기술하고 사용자들에게 고품질의 교통데이터를 제공하기 위한 통합 품질관리 프레임워크를 제안한다. 본 논문에서 제안된 통합 품질관리 기법은 효율적인 교통데이터의 품질관리를 위해서 교통공학 관점과 데이터베이스 관점의 통합 데이터처리 과정과 데이터평가 방안을 제시한다. 또한, 데이터 값과 데이터 구조 관점을 통합하여 처리하기 위한 통합 데이터처리 프로세스를 기술한다. 데이터평가 부분에서는 데이터 값의 경우에 교통공학 관점과 데이터베이스 관점 각각에서 평가할 수 있는 평가항목 기준을 제시하고 데이터 구조 관점과 통합하여 평가할 수 있는 통합 데이터평가 프로세스를 제시한다. 또한, 제안된 교통 데이터 웨어하우스의 품질평가를 위한 방법의 유용함을 보이기 위해서 시스템을 구현하고 평가한다.

본 논문의 구성은 다음과 같다. 2장에서는 국외의 교통데이터 웨어하우스에 대해 소개하고, 현재 국내의 교통관리 시스템에 대해 기술하였다. 3장에서는 본 논문에서 제안한 교통 데이터 웨어하우스 시스템에 대해 설명하였다.

본 논문의 구성은 다음과 같다. 2장에서는 국외의 교통데이터 웨어하우스에 대해 소개하고, 현재 국내의 교통관리 시스템에 대해 기술하였다. 3장에서는 본 논문에서 제안한 교통 데이터 웨어하우스 시스템에 대해 설명하였다.

[†] 본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 육성·지원사업 (IITA-2008-C1090-0801-0031) 및 2007년도 정부(과학기술부)의 재원으로 한국과학재단의 지원(No. R01-2007-000-20958-0)을 받아 수행된 연구임.

* 단국대학교 컴퓨터공학과 박사과정, jihwang@dankook.ac.kr

** 단국대학교 컴퓨터공학과 박사과정, piuspark@dankook.ac.kr

*** 단국대학교 컴퓨터공학과 교수, ymnah@dku.edu(교신처자)

4장에서는 교통 데이터 품질 관리 기법에 대해 기술하였다. 5장에서는 제한한 시스템의 구현 및 평가를 하였고, 마지막으로 6장에서는 결론을 제시하였다.

2. 관련연구

미국 National ITS Architecture 에서는 1999년 9월부터 ADMS를 ITS(Intelligent Transport Systems)의 서브시스템으로 정의 하고 있으며, 교통정보 이력데이터 사용자 서비스(Archived Data User Service : ADUS)는 서비스 제공분야 중 한 부분을 차지하고 있다. 미국의 ITS 아키텍처 표준에는 ITS 구성요소에서 수집되는 모든 이질적인 데이터(현장 데이터수집 장비에서 수집되는 원시데이터, 운영관리 중 발생하는 제반 정보 등)를 통합 구성하여 분석결과를 통해 교통정책 관리, 정책결정지원, 교통안전진단, 교통계획, 시스템 운영효과 감시 등의 업무에 지원하는 부 시스템으로 정의하고 있다. 또한 ITS 를 구성하는 모든 부 시스템의 이력데이터는 ADMS로 통합되고 각 부 시스템에서 요청되는 분석, 조회 정보를 제공하는 방식으로 물리적 아키텍처를 구성하고 있다. 정리해보면 ADMS는 교통관리시스템에서 수집되는 데이터를 다양한 목적에 효율적 이용이 가능하도록 데이터의 가공처리, 저장, 공급과정을 체계적으로 구성하여 1, 2차 데이터처리과정(데이터처리 및 데이터검색 기능)과 데이터 저장기능을 갖는 물리적 시스템이고 ADUS는 ADMS 에 대하여 연구자, 관리자, 일반 이용자들에게 조회 및 분석 요구데이터에 대한 질의를 입력하고 질의에 대한 결과를 표출해주는 ITS 아키텍처에서의 서비스제공 분야 중 하나이다. 국외 교통데이터 웨어하우스 구축 사례로는 캘리포니아 PeMS, 메릴랜드 CATT, 버지니아 ADMS 등이 있다.

우리나라의 고속도로 교통 관리 시스템 (FTMS)은 급증하는 교통 수요를 도로의 건설이나 확장으로 대처하기에는 한계가 있어 최소한의 투자비용으로 교통흐름을 관리하여 이용객에게 고속도로 이용 전 또는 이용 중에 실시간 교통정보를 제공함을 목적으로 하는 시스템으로, 운영시스템에는 지난 한달 간의 데이터만이 저장되고 있는 실정이다. 우회도로 교통정보시스템(ARTIS : Alternative Road Traffic Information System)은 고속도로 우회도로의 실시간 교통정보를 수집, 관리, 제공함으로써, 교통시설의 이용효율을 극대화하는 교통체계로 대량의 과거 데이터를 적재하고 관리하지 않고 있다.

3. 교통데이터 웨어하우스

3.1 통합시스템 구조

FTMS와 ARTIS의 정보 수집 시스템에서 제공되는 다양한 교통데이터를 대상으로 교통데이터 웨어하우스를 구축하고자 한다. 교통데이터 웨어하우스는 분산된 데이터 소스로부터 정보를 추출, 변환, 적재(ETL: Extract, Transform and Load)하여 사용자들에게 일관된 정보를

제공하기 위해 중앙 집중적인 저장소에 통합, 저장하고 데이터에 대한 다양하고 복잡한 분석을 가능하게 한다[1].

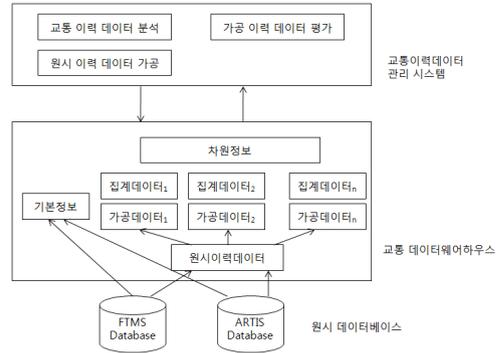


그림 1. 통합 시스템 구조

그림 1은 통합 시스템의 구조를 그린 것이다. FTMS와 ARTIS는 각각 고속도로 통행량 정보와, 우회국도 통행정보를 관리하는 운영계 시스템이다. FTMS와 ARTIS 데이터베이스의 기본정보와 이력데이터는 ETL을 통해 추출되어 교통데이터 웨어하우스에 원시이력데이터로 적재된다. 교통데이터 웨어하우스 관리시스템의 원시이력데이터는 원시이력데이터 가공 프로그램을 이용하여 다양한 정제 알고리즘들을 통해 가공 데이터로 바뀐다. 서로 다른 정제 방법을 통해 가공된 가공데이터는 가공 데이터 평가 프로그램을 이용하여 평가될 수 있다. 각각의 가공데이터는 빠른 시공간 분석업무를 지원하기 위해, 시공간적으로 집계되어 집계데이터가 만들어진다. 이를 위해 데이터 웨어하우스에는 시공간적인 데이터 분석을 위한 차원정보가 존재한다. 가공, 집계 데이터가 모두 만들어지면, 교통 이력데이터 관리시스템의 분석 프로그램을 이용하여, 다양한 분석질의들을 수행할 수 있다.

3.2 통합 스키마

웨어하우스에서 교통이력데이터는 사실 테이블이 되며, 이를 분석하기 위한 시공간 정의의 테이블은 차원 테이블이 된다.

그림 2는 고속도로 공간 차원 정보를 나타내는 스키마이다. 대부분의 분석 질의는 대량의 데이터를 대상으로 수행되므로, 교통데이터 웨어하우스에서는 보다 빠른 분석을 위해 이력데이터를 사전에 집계하여, 실체화 뷰(Materialized View)를 만든다. 그림 3은 실체화 시공간적 차원을 이용하여 만드는 실체화 뷰들과 5분 지점 자료를 저장하는 뷰의 모습이다. 사고나, 기상 조건에 따른 검색을 위해, 사고유무, 강수량, 적설량도 저장한다. 또, 시스템은 다양한 데이터 정제 알고리즘을 원시데이터에 적용하고 가공데이터를 만들게 되는데, 이를 쉽게 관리하기 위한 테이블들이 존재한다. FTMS에서 원시자료는 30초 간격으로, LOOP 검지기 단위로 저장되는데, 이를 쉽게 공간차원으로 분석이 용이하도록 하기 위해 LOOP 검

지기 데이터를 노선, 구간, 지점, 차로 단위로 분할하고 재배치한다. 공간적으로 정리된 데이터는 분석 및 검색 속도 향상을 위해, 다시 시간적으로 30초, 5분, 15분, 1시간, 1일, 1주, 1달 등 7가지 시간차원으로 통합한다.

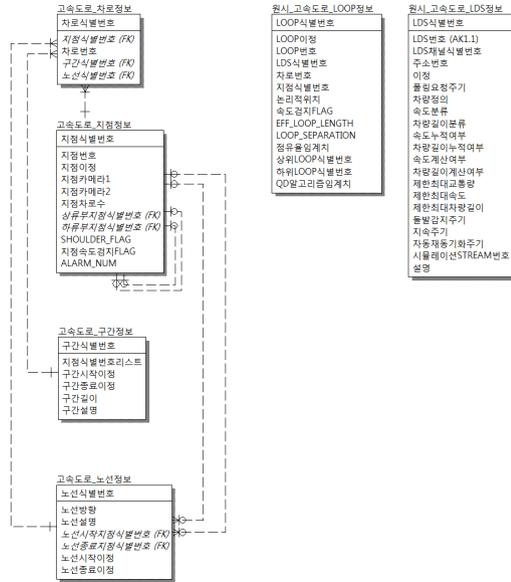


그림 2. 고속도로로 공간 차원 스키마

시간	30초	MV1	MV8	MV15	MV22
	5분	MV2	MV9	MV16	MV23
	15분	MV3	MV10	MV17	MV24
	1시간	MV4	MV11	MV18	MV25
	1일	MV5	MV12	MV19	MV26
	1주	MV6	MV13	MV20	MV27
	1달	MV7	MV14	MV21	MV28
		차로	지점	구간	노선

고속도로_5분_지점

년월일시분초_날짜
지점식별번호

교통량
점유율
속도
사고발생유무
경수량
적설량

그림 3. 시공간 집계에 의한 실제화 뷰 종류 및 고속도로로 5분 지점 뷰의 스키마

4. 통합데이터 품질관리

4.1 데이터 값 관점의 데이터 처리

교통공학적인 관점에서의 오류판단은 차량검지기로부터 수집된 데이터에 대한 논리적 오류 여부를 판단하여 제거하는 과정이다. 일반적으로 대상 원시데이터에 대해 검지기 이상 유무, Data threshold, Data relation의 논리적 유효범위를 근거로 데이터의 오류 여부를 판단하고 제거하고 임시테이블에 저장한다. 오류로 판단된 교통량, 점유율, 속도 모두를 -999로 처리한다. 오류 판단이 이루어지면, 오류보정을 행하는데, 해당지점의 데이터는 공간적

으로 참고지점과 시간적으로 교통상태가 유사성을 갖는다는 전제하에 수행된다. 오류 보정을 위한 방법은 공간적 추세 활용법, 시간적추세 활용법, 이력데이터 활용법을 적용하며, 공간적추세 활용법은 해당지점과 비슷한 교통 특성을 보이는 지점의 데이터를 활용하여 누락데이터를 보정한다[6, 7, 8, 9]. 그림 4는 오류판단 및 오류 보정을 행한 예이다. (a)는 오류판단을 수행하기 이전의 차로별 원시데이터이며, (b)는 교통공학적인 오류판단 알고리즘을 수행한 후의 결과 데이터다. 결측 되었거나, 오류로 판단된 데이터는 -999로 표시된다. (c)는 오류 보정 후 지점별로 그룹화한 30초 원시데이터 예이다.

데이터 타입 오류의 경우 모든 데이터 타입을 기본 데이터타입을 적용하거나 특정 데이터 타입으로 변환하는 변환규칙을 명시한다. 데이터 표현상의 오류의 경우에는 데이터베이스 사전에 명시된 데이터 표현을 사용하여 전체를 표현하거나 혹은 등록된 표현방법으로 표현하기 위한 변환규칙을 명시한다. 단위 오류의 경우에는 기본적으로 적용해야 하는 단위를 전체 단위로 적용하거나 혹은 기본 단위로의 변환 공식을 명시한다. 예를 들어, 기본 단위가 cm인 경우에 inch 데이터에는 $cm\ 값 = 현재데이터값 * 2.5399$ 를 명시한다. 정밀도 오류의 경우에는 표준 정밀도표현 방법을 명시하여 표현한다. 위에서 기술한 변환 규칙이나 공식을 명시하는 경우 반드시 데이터베이스 사전에 등록된 기준으로 변환할 필요는 없으며, 성능이나 응용에 따라서 다른 개별 데이터 타입, 데이터표현, 단위로 변환하는 것도 가능하다. 본 논문에서는 데이터베이스 관점의 오류 판단과 오류 보정을 위한 기준을 별도의 사전 데이터베이스에 설계자 또는 관리자가 미리 명시하는 것으로 가정하고 변환 대상의 원칙으로 기본적으로 사전 데이터베이스에 명시된 것으로 하며, 위에서 언급한 바와 같이 별도의 다양한 요구에 따라서 변환규칙을 생성할 수 있다.

4.2 데이터 구조 관점의 데이터처리

구조 관점에서의 데이터처리 과정은 교통공학적인 측면보다는 데이터베이스 구조적인 측면에서의 데이터처리 과정을 말하는 것이다. 교통데이터 웨어하우스를 구축하면서 다양한 운영계 소스로부터 데이터를 통합하는 과정에서 다양한 원 소스간의 구조적인 상이성에 의한 충돌 문제가 있을 수 있으며 본 논문에서는 데이터 값에서 기술한 바와 같이 여러 소스를 통합하는 기본 통합 스키마 정보를 사전 데이터베이스에 설계자나 관리자가 미리 지정해 놓는 것으로 가정한다. 그림은 데이터 구조 관점에서 오류를 판단하고 오류를 보정하는 전체적인 알고리즘을 나타낸다. 데이터 구조 관점에서의 오류판단은 크게 엔티티(entity), 속성(attribute), 관계(relationship)으로 볼 수 있다. 표 1은 교통데이터를 통합하면서 발생할 수 있는 데이터베이스 관점에서의 데이터값 오류 및, 데이터 구조 오류들의 정의한 것이다. 정의된 오류에 대한 오류 보정 프로세스를 통합 데이터 품질관리 프레임워크에 구현하여 실험결과를 5장에 기술하였다.

SAMPLE_TIME	LDS_ID	VOL1	VOL2	OCC1	OCC2	SPEED1	SPEED2
20061010000000	0010LD0380	1	1	1	1	103	300
20061010000030	0010LD0380			1	1	110	117
20061010000100	0010LD0380	3	3			130	109
20061010000130	0010LD0380	1	1	1	1	133	107

(a) 30초 원시 자료

SAMPLE_TIME	LOOP_ID	LOOP_V0	LOOP_S0	LOOP_O0	LOOP_V1	LOOP_S1	LOOP_O1
20061010000000	380	2	93	1	2	-999	1
20061010000030	380	-999	-999	0	0	100	0
20061010000100	380	0	100	0	-999	100	-999
20061010000130	380	0	100	0	0	100	0

(b) 오류판단 결과

SAMPLE_TIME	VDS_ID	VDS_V	VDS_O	VDS_S
20061010000000	0010VDS00380	5	1	113
20061010000030	0010VDS00380	3	0	88
20061010000100	0010VDS00380	4	0	88
20061010000130	0010VDS00380	2	0	88

(c) 오류보정 결과

그림 4. 고속도로 원시 자료 오류 보정 예

표 1. 오류데이터 정의

품질	정량화 가능한 데이터 품질 특성	오류 데이터 항목
운영데이터 데이터 값	정확성	설계하고 다른 이상치 데이터가 존재
	일관성	용어 정의가 일치하지 않음 규정, 표준이 일치하지 않음 속성 정의가 일치하지 않음 데이터 형식이 일치하지 않음
	완전성	누락된 데이터가 있음
분석데이터 데이터 값	정확성	데이터 타입 오류가 있음 데이터 표현의 오류가 있음 데이터 단위의 오류가 있음 데이터의 정밀도에 오류가 있음
엔티티 데이터구조	완전성	두 개 미만의 속성을 가짐 두 개 미만의 인스턴스를 가짐
	식별성	키 역할을 하는 속성이 없음
	정확성	같은 엔티티에 대한 서로 다른 이름이 존재 서로 다른 엔티티에 대한 같은 이름이 존재 누락된 애트리뷰트가 존재 테이블레벨의 제약조건 오류가 존재
관계 데이터구조	선택성	필수관계의 엔티티 인스턴스가 없음
	관계형태	사전데이터베이스의 기본구조와 다름
	관계명칭(M:N)	M:N 관계에서 관계명칭이 다름
속성 데이터구조	원자성	원자성을 위배하는 데이터가 존재
	일관성	하나의 속성에 여러 데이터유형이 존재
	무결성	참조되는 속성의 데이터와 참조하는 속성의 데이터가 불일치
	정확성	같은 속성에 대해서 다른 이름이 존재 서로 다른 속성에 대해서 같은 이름이 존재 속성레벨의 제약조건 오류가 존재 기본 값 오류가 존재

4.3 통합 데이터 처리

본 논문에서는 운영계 대상 시스템을 고속도로 교통관리시스템과 우회도로 교통정보시스템으로 보고 있으며, 그림 5에서 보는 바와 같이 운영계 데이터베이스의 원시 데이터는 추출, 변환, 적재 과정을 거쳐서 교통데이터 웨어하우스에 축적되던 일별, 주별, 월별로 실체화 뷰 등의 기법을 사용하여 성능 향상 기법을 이용하여 미리 가공을 하고 별도의 데이터베이스에 저장하여 사용할 수 있다.

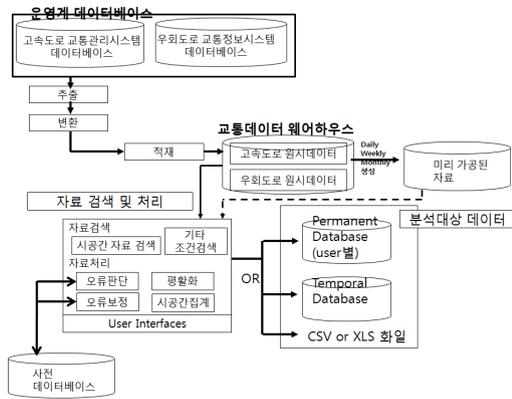


그림 5. 통합 데이터 처리 프레임워크

제안된 품질관리 프레임워크를 통해서 데이터구조 관점에서 효율적인 교통데이터 웨어하우스의 품질관리가 가능하며, 연구자들에게 쉽게 고품질 교통데이터를 제공함으로써 연구환경 개선 및 비용 절감 효과를 볼 수 있다.

교통데이터 웨어하우스에 저장된 원시데이터는 시공간 조건에 의해서 또는 기타 사용자 요구에 의한 조건에 의해서 검색되며, 검색된 결과 데이터에 대해서 데이터처리 과정을 수행한다. 통합 데이터베이스 데이터처리 프로세스는 본 논문에서 제안한 교통공학 관점의 데이터처리 및 데이터베이스 관점에서의 데이터처리 기법을 통합해서 진행할 수 있도록 구성 한 것이다. 데이터처리 결과 데이터는 사용자에게 분석 대상데이터로 제공되어 지며, 사용자의 권한 검사를 통해서 데이터베이스 혹은 파일 형태로 출력된다. 데이터베이스는 기본적으로 입시적인 공간이 할당되며, 적절한 권한을 소유한 특정 사용자의 경우에 영구적인 데이터베이스 공간에 저장하여 계속적으로 분석 활용할 수 있다.

5. 구현 및 평가

5.1 실험 환경

원시데이터 소스인 FTMS와 ARTIS 데이터베이스는 교통정보센터에 ETL 방법에 의해 교통데이터 웨어하우스에 저장된다. 교통데이터 웨어하우스는 HP MSA 1500 모델로 40TB의 공간이 확보되어 있다. 교통데이터 웨어

하우스 관리서버로는 HP 9000 RP7420서버를 사용하며, 운영체제는 HP-UX 11버전을 사용하고, CPU는 4×1GHz이며, 메모리 용량은 64GByte이며, Oracle 10g DBMS가 설치되어 있다. 교통정보센터의 FTMS는 Oracle 7.3 버전이고, ARTIS는 Oracle 8i가 사용되었다. 그림 6은 실험 환경을 나타낸다.

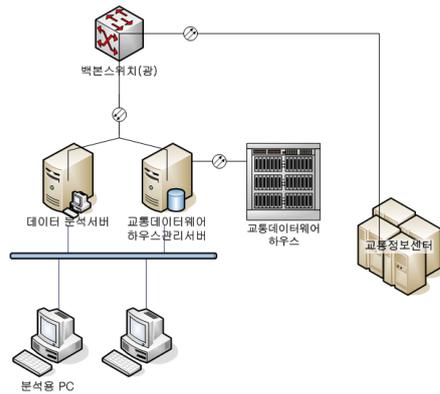


그림 6. 실험 환경

교통데이터 웨어하우스는 FTMS, ARTIS, IDB 3개의 테이블스페이스로 구성된다. FTMS 원시데이터는 FTMS 테이블스페이스에 저장되며, ARTIS 원시데이터는 ARTIS 테이블스페이스에 저장된다. IDB 테이블스페이스는 교통데이터 웨어하우스의 모든 테이블들이 저장되는 공간이다. 원시데이터를 별도의 가공 없이 교통데이터 웨어하우스로 그대로 옮기는 이유는 여러 가지 정제 알고리즘을 사용하여, 새로운 이력데이터를 만들어내는데 가공되지 않은 원시데이터가 필요하기 때문이다.

5.2 실험 및 평가

실험 대상 데이터는 고속도로 경부선의 부산->서울 방향으로 설치된 모든 VDS 검지기 데이터들과 모든 우회국도 VDS 검지기 데이터들에 대한 데이터이다. 데이터는 2007년 4월 1일부터 2007년 4월 30일까지 30일 간의 데이터에 대해 처리한다. 표 2는 실험 대상 데이터의 규모를 나타낸다.

그림 7은 데이터값에 대한 통합 프레임워크의 유용함을 보이기 위해서 기간을 1일, 1주일, 1개월로 다르게 하면서 통합 시 오류가 발생해도 아무런 처리도 하지 않은 경우와 상용제품인 Oracle OWB(Oracle Warehouse Builder)를 이용한 경우 그리고, 본 논문에서 제안한 통합 프레임워크 간의 결과 품질을 비교 평가하였다. 비교 평가 결과 통합 프레임 적용 시 높은 품질을 보였고 기간에 관계없이 일정한 수준의 품질을 제공함을 알 수 있었다.

그림 8은 데이터구조 관점에서의 엔티티에 대한 품질 평가를 수행하고 비교한 내용이며, 그림에서 보이는 바와 같이 제안된 통합프레임워크를 적용 시에는 그래프가 거

표 2. 실험 데이터

System	Item	Number of Records
FTMS (Busan to Seoul)	Line	1
	Interval	43
	VDS	3,00
	LOOP	2,016
	Lane	1,008
	1 day record	2,903,040
	1 week record	20,321,280
	1 month record	87,091,200
ARTIS (All area)	LINE	46
	VDS	724
	Lane	3,044
	1 day record	8,766,720
	1 week record	61,367,040
	Temporal dimension record	86,400
	Historical data	350,092,800

의 수립하는 것으로 보아 데이터의 일정한 수준의 품질을 제공하는 것을 알 수 있다. 또한, 상용 제품인 OWB를 사용하는 경우 현재 상용제품의 특징이 교통데이터에 특화된 정제 기능을 제공하지 못하고 있어서 데이터의 종류에 따라 그리고 오류 처리에 대한 규칙의 기술 정도에 따라서 품질 결과가 많은 변화를 보임을 알 수 있다.

그림 9를 보면 원시데이터의 경우 3가지 경우 모두 비슷하게 관계 면에서 우수한 품질을 보이고 있다. 그 이유는 원시데이터인 fact 데이터의 경우 소스데이터베이스의 원시데이터 자체가 관련된 테이블이 극히 적고 통합 시에도 대부분 관계가 유지된다. 반면에 설정데이터의 경우엔 소스 데이터베이스부터 많은 관계가 설정되어 있고 통합 시 많은 오류를 보이고 있다. 실험 결과 현재 상용 cleansing 제품들에서는 통합 이전의 소스 데이터들 간의 관계를 고려하고 있지 않아 소스데이터들 간의 관계가 복잡하게 구성되어 있는 경우 많은 문제점을 야기시킬 수 있다.

6. 결론

본 논문에서는 이러한 요구에 따라서 대량의 교통데이터 웨어하우스를 구축 한 국내의 유사시스템의 사례를 조사 분석하고 효율적으로 교통데이터 웨어하우스를 구성하기 위한 물리적, 논리적 방안에 대해서 제안하였다. 본 논문에서 기술한 교통데이터 웨어하우스는 연구 목적으로 새로운 서비스를 발굴하거나 사용자로부터의 새로운 요구 사항을 도출하고 활용분야를 수립하는데 응용될 수 있다. 또한, 기존의 운영체 시스템에서 교통데이터 웨어하우스로 교통데이터를 통합하기 위한 추출, 변환, 적재 방법에 대해서 기술하였다. 효율적인 교통데이터의 품질관리를 위해서 데이터 값뿐만 아니라 데이터 구조 관점에서의 통합 데이터처리과정을 제안하고 데이터처리 과정을 통해서 만들어진 분석 데이터를 평가하기 위한 통합 데이터 평가 프로세스를 제시하였다. 제안된 교통데이터를 위한 통합 데이터 품질관리 프레임워크에서는 교통공학 관점과 데이터베이스 관점을 접목시킴으로써 대용량 교통데이터 웨어하우스에서 교통공학적인 측면과 데이터베이스 측면 모두의 품질관리를 쉽게 수행할 수 있다. 또한 제안된 교통데이터 웨어하우스의 통합 품질관리 프레임워크의 유용성을 보이기 위해서 일부 고속도로 데이터와 우회도로 데이터를 가지고 구현하여 실험 평가 하였다.

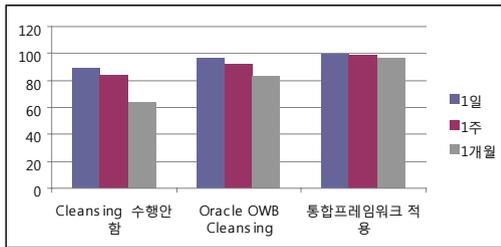


그림 7. 기간별 엔터티 품질평가 수행비교

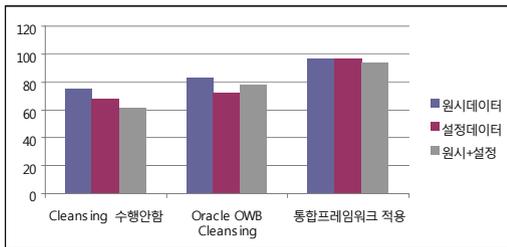


그림 8. 데이터별 엔터티 품질평가 수행비교

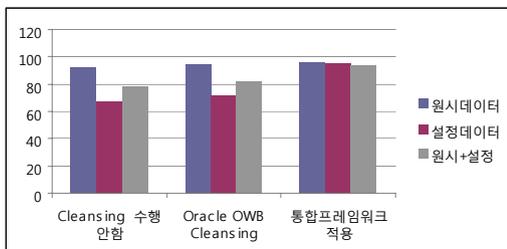


그림 9. 데이터 관계 품질평가 수행비교

참고 문헌

[1] Smith, B. and S. Babiceanu, "Investigation of Extraction, Transformation, and Loading Techniques for Traffic Data Warehouses," *Transportation Research Record*, Vol. 1879, pp.9-16, 2004.

[2] Smith, B. D. Lewis, R. Hammond. "Design of Archival Traffic Databases: Quantitative

Investigation into Application of Advanced Data Modeling Concepts,” *TransportationResearchRecord*, Vol.1836, pp.126-131,2003.

- [3] 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,” *TransportationResearchRecord*, Vol. 1867, pp. 116-126, 2004.
- [4] Wall, Z., and D. Dailey, “Algorithm for Detecting and Correcting Errors in Archived Traffic Data,” *TransportationResearchRecord*, Vol. 1855, pp. 183-190, 2003.
- [5] Turner, S., L. Albert, B. Gajewski, and W. Eisele, “Archived Intelligent Transportation System Data Quality: Preliminary Analyses of San Antonio TransGuide Data,” *TransportationResearchRecord*, Vol. 1719, pp. 77-84, 2000.
- [6] Zong, M., S. Sharma, and P. Lingras, “Genetically Designed Models for Accurate Imputation of Missing Traffic Counts,” *TransportationResearchRecord*, Vol. 1879, pp. 71-79, 2004.
- [7] Smith, B.L., Scherer, W. T. and Conklin, J. H., “Exploring Imputation Techniques for Missing Data in Transportation Management Systems,” *TransportationResearchRecord*, 2003.
- [8] 김태용, 강선희, 윤명수, “중앙정보센터 교통정보 가공처리 프로세스 및 시스템 개발,” 2003년도 추계 학술발표대회, 제 44회 학술발표대회, 대한교통학회, 2003, pp245.
- [9] 한 대철, 최숙양, 문학룡, 김성현, “ITS시스템의 결측보정에 관한 연구,” 한국 ITS 제 3회 정기총회 및 추계 학술대회, pp270~275, 2004.



황 재 일

1990년 단국대학교 전자공학과(공학사)
 1999년 단국대학교 산업대학원 정보처리학과(공학석사)
 2000년~단국대학교 대학원 컴퓨터 공학과 (박사과정)
 관심분야는 데이터베이스, 멀티미디어 데이터베이스, 대용량 분산 데이터베이스, 데이터 웨어하우스



박 승 용

1997년 단국대학교 컴퓨터공학과(공학사)
 1999년 단국대학교 컴퓨터공학과 (공학석사)
 2007년 단국대학교 컴퓨터공학과 (공학박사)
 2002년~2006년 : (주)엔다인 책임연구원
 2006년~현재 (주)내쇼널테스팅랩아시아 연구소장
 2006년~현재 (주)리얼구루즈 대표이사
 관심 분야는 데이터베이스, 데이터 모델링, 대용량 분산 데이터베이스, 멀티미디어 데이터베이스



나 연 목

1986년 서울대학교 컴퓨터공학과(공학사)
 1988년 서울대학교 대학원 컴퓨터공학과 (공학석사)
 1993년 서울대학교 대학원 컴퓨터공학과 (공학박사)
 1993년~현재 단국대학교 컴퓨터학부 교수
 1991년 IBM T. J. Watson 연구소 객원연구원
 2001년~2002년 University of California, Irvine 객원교수
 관심분야는 데이터베이스, 데이터 모델링, 객체지향 데이터베이스, 멀티미디어 데이터베이스, 멀티미디어 정보 검색, 시맨틱 웹, 이동객체 데이터베이스, 대용량 분산 데이터베이스