비즈니스 인텔리전스 환경에서 변환 관리를 이용한 데이터 품질 향상에 대한 연구

A Study on Data Quality Management in Business Intelligence Environments

이 춘 열 (Choon Yeul Lee) 국민대학교 비즈니스IT전문대학원

요 약

비즈니스 인텔리전스를 위한 통합 정보시스템의 운영을 위하여서는 무엇보다도 기업 내부와 외부에서 발생한 자료들을 상호 연계하여 통합 관리하여야 한다. 데이터의 통합관리를 위하여서는 기존의 데이터와 데이터들 사이의 일대일 매핑이 아니라 데이터의 생성부터 통합 저장까지의 변환 과정을 총괄적으로 표현하고 관리하여야 한다. 본 연구는 정보구조그래프를 확장함으로써 데이터의 변환 구조들 뿐만이 아니라 세부 처리 단계들까지 통합 관리할 수 있는 방안을 제시하며, 이를 이용하여 비즈니스 인텔리전스와 같은 통합환경에서 데이터베이스의 품질 향상을 위한 활용방안을 제시한다.

키워드 : 메타데이터, 데이터관리, 데이터 품질, 비즈니스 인텔리전스, 데이터 처리 프로세스

Ⅰ. 서 론

최근의 기업정보시스템은 개별 시스템 구현보다는 이들 사이의 정보 교환 및 공유를 강조하고 있으며, 단순 데이터 처리보다는 정교한 정보와 지식의 도출이 주요 관심사로 대두하고 있다. 이러한 필요에 따라 정보 시스템의 형태 또한 통합화의 경향을 보이고 있으며, 전사적 시스템 통합(EAI: Enterprise Application Integration), 정보기술아키텍쳐(ITA: Information Technology Architecture) 등에 대한 관심이 증대하고 있다. 또한 데이터웨어하우징과 데이터 마이닝 등과 같은 고도화된 데이터 통합 관리 기법들이 도입되어 이용되고 있다.

정보시스템 고도화에 대한 요구는 기업 환경의 변화, e-비즈니스의 보편화 및 경제활동 범위의 확대 등에 기인한다고 볼 수 있다. 즉, 종래와 같이 거래가 단속적으로 이루어지며, 해당 거래가 다른 기업 활동에 미치는 영향 또한 시간적 격차를 두고 처리되는 환경으로부터 단일 거래들이 서로 연계되어 자동 처리되며, 처리 결과 또한 실시간으로 전달되는 환경이 되었다. 이러한 추세는 앞으로 경제 환경의 디지털화와 더불어 더욱 가속되리라 예상된다.

정보 시스템의 고도화를 위하여서는 응용 시스템의 실시간 연동 및 통합이 이루어져야 하는데, 이를 위하여서는 무엇보다도 기업 내부 및 외부에서 발생한 자료들을 상호 연계하여 통합 관리하여야 한다. 즉 거래 자료들을 상호 연계함으로써 하나의 응용 시스템에서 처리한 데이

터가 다른 시스템으로 자동으로 전달되어 처리되어야 하며, 이러한 데이터들이 전사적으로 통합 관리되어야 한다. 또한 데이터들은 상호 형태에 맞게 정제 및 가공이 이루어져야 한다. 이러한 데이터의 정제 가공 도구들을 ETT(Extraction, Transformation and Transition) 솔루션들이라고 하며, 거의 모든 데이터베이스관리시스템 공급 업체들이나 독립 데이터관리도구 공급 업체들이 데이터웨어하우스나 데이터마트 구축 도구의 일부로서 제공하고 있다.

이상에서 언급한 응용 시스템 통합이나 통합 데이터관리 솔루션들은 개별 시스템들을 결합하여 기업의 정보화 효과를 제고한다. 따라서 이들은 기업의 기간업무 시스템이나 데이터베이스들을 기반으로 하는 2차 솔루션들이라고 할 수 있다. 이들 2차 솔루션들이 효과적으로 활용되기 위하여서는 일차적인 업무처리시스템들이 구축되어 있어야 하며, 이를 통하여 2차 솔루션들이 필요로 하는 정보가 효과적으로 제공되어야 한다. 그러나 기업 안에는 다양한 응용 시스템들이 존재하며, 이들이 처리하는 데이터들은 각각의 데이터베이스나 파일에 다양한 형태로 저장되어 관리되고 있다. 따라서 무슨 데이터들이 어떠한 형태로 관리되고 있는가를 파악하기 위하여서는 데이터베이스의 자료 사전(data dictionary)들을 참조하여야 한다. 이러한 틀 안에서 요즘 제공되고 있는 EAI 솔루션이나 ETT 솔루션들은 거의 모두 원시 데이터를 지정한 후 이를 목표 데이터로 변환 시키는 점대점 매핑(point-to-point mapping) 형태를 취하고 있다.

현재의 솔루션들을 이용하여 데이터들을 통합 활용하기 위하여서는 기업 안에 어떠한 데이터들이 존재하는가를 개별 자료 사전을 탐색하여 사용자들이 스스로 파악하여야 한다. 그러나 정보화의 범위가 확대됨에 따라 비즈니스 인텔리전스와 같은 통합 환경에서는 무슨 데이터들이 존재하며, 이들이 어떠한 상호 연관성을 가지는 가를 일반 사용자들이 모두 파악하는 것은 거의 불가능하다.

데이터들 사이의 연관성을 파악하기 위하여서는 데이터의 외형적 특성을 나타내는 메타 데이터들뿐만이 아니라 데이터들의 변환 과정에 대한 메타 데이터도 같이 관리하여야 한다. 이러한 데이터의 변환 과정에 대한 정보를 체계적으로 관리함으로써 사용자들은 데이터의 의미와 이들 사이의 연관성들은 파악할 수 있다. 변환 과정에 대한 메타 데이터들은 특히 여러 출처로부터 자료들을 통합하여 활용하는 비즈니스 인텔리전스와 같은 환경에서 사용자들이 원시 데이터의 의미나 용도를 제대로 파악하기 위한 매우 중요한 정보를 제공한다.

이러한 필요성에 따라 본 연구는 데이터의 변환 과정을 체계적으로 표현할 수 있는 방안을 제시한다. 그리고 이러한 관리를 통하여 양질의 데이터들로 구성된 비즈니스 인텔리전스 환경을 구축하기 위한 방안을 살펴본다. 이하 제Ⅱ장에서는 데이터 품질에 대한 연구들을 데이터 변환 관리 관점에서 살펴보며, 제Ⅲ장에서는 데이터 변환 과정을 체계적으로 관리하기 위한 방안을 정보구조그래프를 근간으로 제시하며, 제Ⅳ장에서는 이렇게 표현된 데이터의 변환 과정을 이용한 데이터 품질 향상 방안을 살펴본다. 마지막으로 제Ⅴ장에서는 본 연구의 의의와 시사점을 결론으로 제시한다.

Ⅱ. 관련 연구

비즈니스 인텔리전스는 다양한 출처로부터 산출되는 데이터들을 통합하여 의사결정자들에게 유용한 정보를 제공한다. 이러한 비즈니스 인텔리전스의 효율성을 결정하는 중요한 요인들 중의 하나가 데이터의 품질이다. 데이터 품질이 비즈니스 인텔리전스 환경에서 더욱 강조되는 이유는 비즈니스 인텔리전스가 통합 데이터베이스를 근간으로 하며, 이를 통하여 많은 사용자들이 정보를 공유하게 때문이라고 할 수

있다. 이러한 통합 데이터베이스 시스템에서는 단독 시스템과 비교하여 사용자들이 데이터 오류를 식별할 수 있는 능력이 떨어진다. 따라서 정보 오류가 발생할 경우 여과없이 많은 사용자들에게 영향을 미치게 된다. 그리고 응용 시스템들이 상호 연관되어 있음으로서 이러한 데이터 오류들이 쉽게 확산될 수 있다.

통합 데이터베이스의 보편화와 이에 따른 정보 공유의 확산으로 데이터 품질에 대한 연구가 활발하게 이루어지고 있으며, 특히 데이터 처리 과정에 대한 연구가 데이터 품질 연구의 주요 부분으로 대두하고 있다. 이는 양질의 데이터를 확보하기 위하여서는 궁극적으로 데이터 처리 과정을 개선하여야 하기 때문이다.

데이터 품질에 대한 연구로서 Wang(1998)은 데이터 품질을 내재적 품질, 접근적 품질, 상황적 품질 및 표현적 품질로 분류하고 있다. 비슷한 구분으로 한국데이터베이스진흥센타(2002)는 데이터 품질을 데이터 자체에 대한 품질과 데이터 서비스에 대한 품질로 구분하며, 데이터 자체에 대한 품질로는 정확성, 완전성, 최신성, 포괄성 및 활용성을, 데이터 서비스에 대한 품질로는 검색성, 편의성, 지원성 및 시스템 성능을 제시하였다.

데이터 품질관리에 대한 연구로서 통합데이터품질관리(TQdM: Total Quality data Management)는 데이터의 출처로부터 사용자에 이르기까지의 전 과정을 포함한다. 즉 데이터 자체에 대한 관리와 더불어 데이터의 품질을 결정하는 프로세스에 대한 관리를 포함한다(English, 1999).

또한 데이터 품질관리에 대한 기초연구로서 데이터 처리 과정을 효과적으로 관리하기 위한 기법들이 제시되었다. Redman(1996)은 양질의 데이터를 서비스하기 위한 노력으로서 정보처리기능모형(Information Processing Model)을 제안하였으며, Wang(1998)은 정보제조시스템(Information Manufacturing System) 등을 제안하였다. Lee(2004)는 데이터의 생성 구조를 표현하기 위한 방안으로서 정보구조그래프(Information Structure Graph)를 제안하였다. 이들 모두 데이터 처리 프로세스를 데이터 관리의 관점에서 단순화하여 표현한 모형들이다.

데이터 변환 프로세스에 대한 관심의 증가는 업무 프로세스와는 다른 데이터 변환 프로세스를 표현하고자 하는 노력의 일환이라고 볼 수 있다. 데이터 흐름도(DFD: Data Flow Diagram)나 워크플로우 모형(WFM: WorkFlow Model)은 데이터 처리 프로세스나 업무 흐름을 표현한 모형들이다. 따라서 이들은 업무처리과정을 프로세스라고 하는 개념으로 표현하며, 해당 프로세스의 세부 내역은 자연언어로 표현한다. 그러나 데이터의 변환 과정은 매우 한정된 기능으로 분류할 수 있으며, 이러한 매우 한정된 기능들을 활용함으로써 보다 효과적인 데이터의 변환 과정을 표현할 수 있다. 예를 들면, 정보처리기능 모형은 데이터의 처리를 비교(Associate), 분류(Filter), 시작(Prompt), 대기(Queue), 검증(Regulate) 및 이동(Transit)의 기본 함수로 표현하며(Redman, 1996), 정보제조시스템은 데이터의 처리과정을 데이터(data unit), 공급자(vendor blocks), 프로세스블럭(process blocks), 품질블럭(quality blocks) 및 사용자(consumer blocks)로 모형화하여 공급자로부터 제공되는 데이터가 무슨 프로세스블럭과 품질블럭을 거쳐 사용자에 서비스되는 가를 나타낸다(Wang, 1998). 정보구조그래프는 데이터의 변환 과정을 데이터 갱신(value update), 개체 생성(object creation) 및 데이터 결합(aggregation)의 3가지 유형으로 구분하고 있다(Lee, 2004).

이러한 데이터 품질관리에 대한 관심의 증가와 더불어 데이터 품질 평가에서도 데이터 관리 프로세스를 포함하고자 하는, 즉 데이터 관리 프로세스에 대한 평가 척도를 데이터의 품질 평가에 추가하고자 하는 노력들이 경주되고 있다. 한국데이터베이스진흥센타는 기존의 데이터 자체에 대한 평가 척도나 데이터 서비스에 대한

평가 척도에 추가하여 데이터 관리 프로세스와 시스템을 평가하는 척도를 포함시키는 데이터 품질 평가 모형을 제시하고 있다(DPC, 2003).

Ⅲ. 확장 정보구조그래프를 이용한 데이터 변환 프로세스와 데이터베이스 스키마의 통합

양질의 데이터를 산출하기 위하여서는 무엇보다도 데이터베이스에 포함된 데이터 항목별로 이들이 어떠한 변환 과정을 거쳐 생성되고 갱신되는가를 관리하여야 한다. 즉 데이터 관리를 효과적으로 실행하기 위하여서는 데이터베이스에 포함된 데이터 항목들에 대하여 데이터의 변환 프로세스를 관리하여야 한다. 본 연구에서는 이러한 데이터의 변환 프로세스를 확장 정보구조그래프로 표현한다.

3.1 데이터 통합품질관리 아키텍쳐

데이터 품질 관리를 통합품질관리(TQM: Total Quality Management)와 연관하여 설명하면 통합품질관리를 위한 생산정보는 크게 다음과 같이 구성된다(Sartori, 1988).

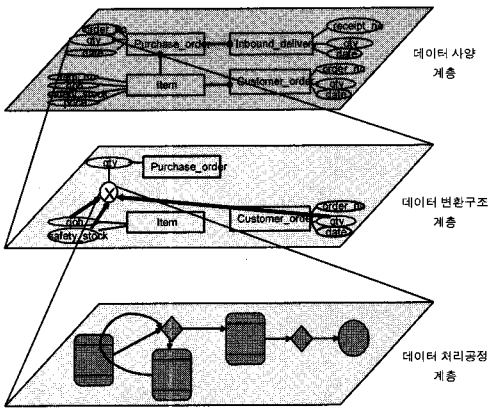
- 제품의 사양을 나타내는 품목 정보
- 제품의 구성을 나타내는 부품구성표
- 제품의 생산 공정을 나타내는 공정 정보

즉 제품을 생산하기 위하여서는 무슨 제품을 만들 것인가를 알아야 하는데 이를 나타내는 것이 품목 정보이다. 그리고 이 제품을 만들기 위하여서는 무슨 원재료나 부품이 필요한가를 파악하여야 하는데 이를 나타내는 것이 부품구성표이다. 마지막으로 이들 부품들을 어떠한 공정에 따라 생산하는가를 알아야 하며, 이를 나타내는 것이 공정 정보이다.

제품 생산의 통합품질관리를 참조로 할 때, 데이터의 경우에도 통합품질관리를 위하여서는 다음의 정보들이 필요하다,

- 무슨 데이터를 생성할 것인가를 나타내는 데이터 사양 정보
- 데이터를 만들기 위하여 무슨 원시자료나 기초 자료들이 활용되는 가를 나타내는 데이터의 변환 구조
- 이들 데이터들이 어떠한 과정에 따라 생성 또는 변환되는가를 나타내는 데이터 처리 공정

데이터에 대한 이러한 통합품질관리 정보들을 표시한 것이 <그림 1>의 데이터 통합품질관리 아키텍쳐이다. 데이터 통합품질관리 아키텍쳐는 3계층으로 구성된다. 첫 번째 계층은 데이터 사양을 나타내며, 두 번째 계층은 데이터베이스 스키마에 포함된 데이터 항목들의 변환구조를 나타내며, 세 번째 계층은 이들 변환구조에 대하여 보다 상세한 처리공정을 나타낸다.



〈그림 1〉 데이터 통합품질관리 아키텍쳐

데이터 품질관리에 대한 기존 연구들을 <그림 1>의 아키텍쳐를 참조로 살펴보면, 정보처리 기능모형은 데이터의 상세 변환과정, 즉 처리 공정을 주로 나타낸다. 그리고 정보제조시스템은 데이터 사양을 기반으로 데이터 처리공정을 모델링 한다. 정보구조그래프는 데이터 사양을 기반으로 데이터의 변환구조를 모델링 한 것이다.

이러한 데이터 품질관리 모형들을 통합품질관리 프레임워크에 맞추어 살펴볼 때, 이들 모형들은 <표 1>에 예시된 바와 같이 데이터의 사양을 나타내는 데이터베이스 스키마와 상호 유기적인 연관성을 가지는 것을 알 수 있다. 이는 데이터 품질관리의 궁극적 목적이 데이터베이스에 저장된 데이터의 품질 향상이라는 점을 고려할 때 당연한 현상이라고 할 수 있다.

〈표 1〉 데이터 통합품질관리 아키텍처와 기존 연구 모형들

계층		관련 모형
데이터 통합 품질관리 아키텍처	데이터 사양	데이터베이스 스키마
	데이터 변환구조	정보구조그래프
	데이터 처리공정	정보처리기능모형, 정보제조시스템

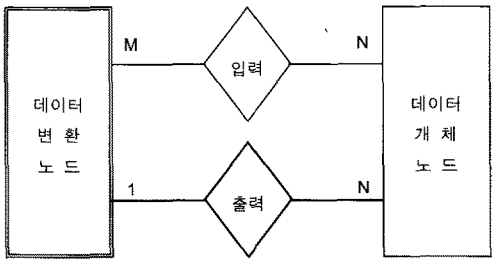
이러한 데이터 품질관리 모형들 중에서 정보구조그래프는 데이터의 변환구조를 모델링하며, 따라서 데이터 명세 및 처리공정을 나타내는 모형들과 연계가 용이하다. 이러한 특성으로 본 연구에서는 <표 2>에 제시된 바와 같이 정보구조그래프를 근간으로 이를 확장하여 데이터 통합품질관리를 위한 모형을 제시한다.

〈표 2〉 데이터 통합품질관리 아키텍처와 확장 정보구조그래프

계층		관련 모형
데이터 통합 품질관리 아키텍처	데이터 명세	데이터베이스 스키마
	데이터 변환구조	**확장 정보구조그래프**
	데이터 처리과정	

3.2 확장 정보구조그래프

정보구조그래프에서 노드(node)와 아크(arc)로 구성된 방향성 그래프 (directed graph)인데, 노드는 테이블 또는 컬럼과 같은 데이터 개체를 나타내며, 아크는 이들 노드를 연결한다. 즉 정보구조그래프는, <그림 1>의 데이터 변환구조 계층에 표시된 바와 같이, 무슨 데이터 개체들이 결합하여 무슨 데이터 개체가 만들어지는가를 나타낸다. 이러한 정보구조그래프를 관계형 데이터베이스의 프레임워크하에서 개체관계모형으로 표현하면 <그림 2>와 같이 모델링된다. <그림 2>는 데이터 개체를 나타내는 노드와 데이터 변환 프로세스를 나타내는 노드를 구분하여 표현하고 있다. 그리고 이들을 연결하는 아크는 관계로 표시된다. 즉 데이터 변환 프로세스와 입력 데이터 개체들을 연결하는 아크들은 입력이라는 관계로 표시되며, 데이터 변환 프로세스와 출력 데이터 개체를 연결하는 아크는 출력이라는 관계로 표현된다.



〈그림 2〉 정보구조그래프의 개체관계도

이론적으로 데이터 변환 노드는 데이터 개체 노드와 독립된 노드가 아니라 데이터 개체 노드의 생성을 표현하는 함수이다. 즉, 데이터 개체 노드 v의 생성 함수 $\mu(v)$는 데이터 개체 노드 v의 생성 유형[1]을 나타내는 함수이며, $\mu(v) = (\oplus, v_1, \cdots, v_n)$ 는 입력 데이터 개체 $v_1, \cdots, v_n$으로부터 생성유형 $\oplus$[2]에 의하여 출력 데이터 개체 v를 생성

1) 정보구조그래프는 모든 데이터항목들을 원시 데이터와 생성 데이터로 구분하며, 생성 데이터들은 다시 데이터 값 갱신, 데이터 결합, 데이터 생성의 3가지 유형의 프로세스에 의하여 만들어지는 것으로 분류하고 있다(Lee, 2004).

2) $\oplus$는 정보구조그래프에서 데이터의 생성 유형 중에서 데이터결합을 나타내는 부호이다.

하는 관계를 함수로 표시한 것이다(Lee, 2004). 이러한 생성함수를 정보구조그래프에서 가시적으로 표시한 것이 데이터 변환 노드라고 할 수 있다.

데이터 변환 노드는 출력 데이터 개체와 일 대 일로 대응하는 데이터 변환 프로세스를 나타낸다. 그러나 현실적으로 이러한 변환 프로세스들은 업무 프로세스나 컴퓨터 프로그램 등으로 구현되며, 따라서 1개의 프로세스에서 여러 데이터 개체들이 만들어질 수 있다. 이러한 가능성을 포함시키기 위하여 본 연구는 <그림 2>의 개체관계도와 <표 4>의 정보구조그래프 테이블에서 데이터 변환 프로세스와 출력 데이터 개체를 일 대 다수로 모델링하였다.

데이터베이스를 구성하는 테이블과 컬럼들이 <표 3>과 같이 표현된다고 가정하면, 정보구조그래프를 구성하는 데이터 개체들과 이들 사이의 변환구조는 <표 4>와 같이 표현된다[3].

〈표 3〉 자료사전 테이블

TBL (table_name[4], creator, table_type, …) COL (table_name, col_no, col_name, col_type, width, nulls, …)

〈표 4〉 정보구조그래프 테이블

ISG_data_node (data_object, data_object_type[5], …) ISG_process_node (output_data_object, process_id, process_name, data_creation_type, …) ISG_input_arc (output_data_object, input_data_object, seq_no, …)

3) <표 4>에서 산출을 나타내는 아크는 별도의 테이블로 표현되지 않으며, 데이터 변환 노드를 나타내는 테이블인 ISG_process_node에 같이 표현된다.

4) Table_name은 기저테이블 뿐만이 아니라 뷰도 포함하는 것으로 가정한다.

5) Data_object_type 은 해당 data_object가 기저 테이블인지, 가상테이블인 뷰인지, 또는 컬럼인지를 구분한다. 그리고 보고서와 같은 정보 산출물들은 모두 뷰로 표시된다고 가정한다.

확장 정보구조그래프는 이러한 정보구조그래프를 데이터 변환구조뿐만이 아니라 처리공정도 표현하도록 확장한 것이다. 여기서 데이터 처리공정은 정보구조그래프에서 정의된 데이터 변환 프로세스를 구성하는 세부 처리 단계들을 나타낸다. 이들 세부 단계들은, 실제로 데이터들이 만들어지고 저장되는 단계를 나타낸다. 예를 들면 현재고량이 안전재고량보다 낮게 되었을 경우 고객주문량을 고려하여 구매주문을 발주한다고 가정할 경우, 이러한 세부 작업 단계들을 데이터 처리공정을 나타내는 대표적인 모형인 정보처리기능모형(Redman, 1996)을 이용하여 작성하면 <그림 3>과 같다.

<그림 3>의 세부 작업단계들을 관계형 테이블로 나타내면 <표 5>의 ISG_sub_processes로 표현된다. 이들 세부 작업단계들의 입력 데이터개체와 출력 데이터개체들은 별도의 테이블인 ISG_sub_ process_IIP와 ISG_sub_process_OIP에 표현된다. <표 5>의 ISG_sub_processes 테이블이 나타내는 데이터 변환의 세부 처리단계들은 최종 산출물인 출력 데이터 개체를 산출하기 위한 중간단계들이다. 따라서 이들은 데이터 변환 프로세스를 매개로하여 <표 4>의 ISG_process_node와 연계된다. 이와 같이, 정보구조그래프는 데이터 통합품질관리 아키텍쳐에서 처리공정을 나타내는 세부 작업 단계들을 포함하도록 쉽게 확장된다.

확장 정보구조그래프는 데이터베이스 스키마에 포함된 모든 데이터 항목들에 대하여 이들을 생성하는 변환구조와 세부 처리공정을 같이 포함한다. 그리고 이들을 데이터 사양을 나타내는 데이터베이스 스키마와 동일한 형태로 표현한다. 이러한 확장 정보구조그래프를 이용함으로써 데이터베이스를 구성하는 모든 데이터 개체들에 대하여 이들의 변환구조뿐만이 아니라 세부 처리공정도 관리할 수 있게 된다.

〈표 5〉 데이터 변환 프로세스의 세부 처리공정 테이블

ISG_sub_processes (process_id, process_name, seq_no, process_type, process_description, duration, error_rate, ···)
ISG_sub_process_IIP (process_id, seq_no, data_object, data_object_type, ···)
ISG_sub_process_OIP (process_id, seq_no, data_object, data_object_type, ···)

단 계	1	2	3	4	5	6	7	8
함 수	시작	변환	비교	대기	검증	대기	검증	저장
설 명	현재고량이 안전재고량 보다 낮음	주문량계산	공급업체 선정	공급업체 확인대기	공급업체로 부터주문 확정	결재담당자 대기	결재	주문레코드 저장
임력데이터	주문필요 품목목록	현재고량, 고객주문량	공급업체 목록					주문레코드
출력데이터		주문량	공급업체		주문량			주문레코드
수행주체	데이터 담당자	데이터 담당자		공급업체			결재담당자	데이터 담당자

〈그림 3〉 구매 주문량 계산 프로세스의 세부 처리공정

Ⅳ. 확장 정보구조그래프를 이용한 데이터 품질 향상

확장 정보구조그래프는 데이터베이스에 저장된 모든 데이터 항목들에 대하여 이들의 변환구조와 세부 처리공정을 나타낸다. 이들 데이터의 세부 처리공정은 정보구조그래프에 포함된 데이터 변환 노드를 매개로 하여 데이터 개체 노드들과 연결된다. 이와 같이 확장 정보구조그래프는 데이터 변환 노드를 매개로 하여 특정 데이터 항목들을 위한 세부 변환 단계를 추적할 수 있으며, 역으로 이들 데이터 변환 단계들의 영향을 받는 데이터 항목들을 추출할 수 있다. 그리고 이들 세부 처리공정의 작업 시간이나 대기 시간, 오류 발생율들을 취합함으로써 특정 데이터 개체를 생성하기 위한 소요시간, 오류 발생율들을 예측하고 관리할 수 있다. 이러한 정보구조그래프를 이용함으로써, 기업은 여러 정보 원천들로부터 만들어지는 데이터들이 다양한 데이터베이스들에 어떻게 저장되고, 상호 연계 활용되는가를 쉽게 파악할 수 있다.

예를 들면, 이후에 표시되는 <그림 5>는 고객 데이터베이스의 고객 레코드로부터 고객 데이터웨어하우스의 고객별 주문량이 산출됨을 나타낸다. 이와 같이 데이터베이스들 사이의 변환 구조를 파악함으로써 전사적으로 보유하고 있는 데이터 항목들과 이들 사이의 연관 관계를 파악할 수 있으며, 이를 이용하여 양질의 비즈니스 인텔리전스 환경을 구축할 수 있다.

4.1 데이터 생성 프로세스의 품질 척도 관리

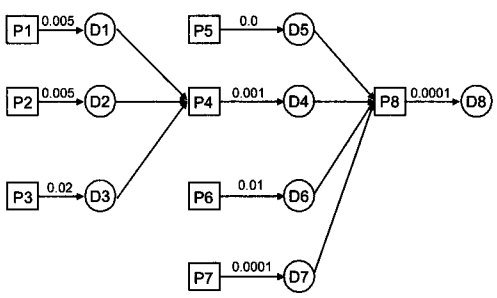
데이터 생성의 주요 품질 척도로는 데이터 생성에 소요되는 시간과 오류 발생율들을 고려할 수 있다. 물론 데이터의 품질 척도에 관하여서는 여러 연구들이 제시되고 있으며, 정확성이나 적시성 이외에도 완전성, 적정성, 이해가능성, 접근성 등의 여러 지표들이 제시되었다(Wang, 1998; DPC, 2003). 본 연구에서는 데이터의 변환 단계와 연관하여 가장 연관성이 높은 정확성과 적시성을 중심으로 품질척도 관리 방안을 제시한다.

데이터 개체 D_k를 생성하는 프로세스를 P_i라고 가정하고, P_i를 구성하는 세부 처리단계들을 P_{ij}라고 가정하자(프로세스 P_i는 ISG_process_node 테이블에 저장되며, 세부 처리단계 P_{ij}는 ISG_sub_processes 테이블에 저장된다. ISG_sub_processes 테이블에 저장된 process_id가 P_i이고 seq_no가 j인 레코드가 P_i의 j번째 세부 단계인 P_{ij}를 나타낸다). 그러면 세부 처리단계별 소요시간과 오류발생율이 duration(P_{ij}), error(P_{ij})로 측정되었다고 가정할 경우, 각 프로세스의 소요시간과 오류발생율은 다음과 같이 도출된다.

$$\text{duration}(P_i), = \sum_{j=1}^{n} \text{duration}(P_{ij}) \qquad (1)$$

$$\text{error}(P_i) = 1 - \prod_{j=1}^{n} (1-\text{error}(P_{ij})) \qquad (2)$$

이들 프로세스들은 정보구조그래프에서 데이터 변환 노드로 표현되며, 데이터 개체 노드들과 연결되어 그래프를 구성한다. 이러한 정보구조그래프를 표시하면 <그림 4>와 같다. 정보구조그래프에서 모든 데이터 노드들은 이에 대응하는 프로세스 노드를 가진다.



〈그림 4〉 정보구조그래프에서 품질척도의 표현 예

세부 처리단계들로부터 계산된 품질척도들은 데이터 변환 프로세스에 할당된다. 이러한 품질척도를 정보구조그래프에 나타내기 위하여서는 데이터 변환 노드에 대하여 표시하거나 또는 데이터 변환 노드와 데이터 개체 노드를 연결하는 아크에 대하여 표시할 수 있다. 이렇게 하면 모든 데이터개체들에 대하여 품질 척도들을 정보구조그래프에 표시할 수 있으며, 이를 이용하여 다음에 예시된 여러 가지 품질관리 기법을 적용할 수 있다.

- 주 경로를 식별함으로써 데이터의 품질에 영향을 미치는 입력 데이터 개체들을 찾아낼 수 있다. 예를 들면 <그림 4>에서 D8 구매주문의 정확성에 가장 영향을 많이 미치는 것은 D6 고객주문량임을 쉽게 알 수 있다.
- 산출된 데이터의 정확성을 모든 프로세스의 오류발생율로부터 누적적으로 계산할 수 있다. 즉 경로를 구성하는 프로세스 노드들의 정확성은 식 (2)와 같이 계산되며, 이들 노드들을 포함하는 경로의 오류발생율은 네트워크에 대한 기법들을 적용하여 산출할 수 있다. 예를 들면 데이터의 정확성을 계산하기 위하여서는 부품의 신뢰성으로부터 시스템의 신뢰성을 계산하는 신뢰성 예측 기법들을 원용할 수 있다.

  예를 들면 <그림 4>에서 데이터 개체 D8을 생성하는 경로는 4개가 있다. 이와 같이 네트워크를 구성하는 경로가 $R_1, \cdots, R_m$일 경우, 각 경로의 고장률은 앞에서 제시되었던 식 (2)의 고장률 계산식으로부터 계산되며, 전체 시스템의 고장률은 이들 경로의 고장률의 합으로 계산된다. 따라서 전체 경로의 고장률은 다음과 같이 계산된다.

$$\text{error}(R) = \sum_{j=1}^{m} \text{error}(R_i)$$

$$= \sum_{j=1}^{m} (1 - \prod_{j=1}^{m} (1-\text{error}(P_{ij})))$$

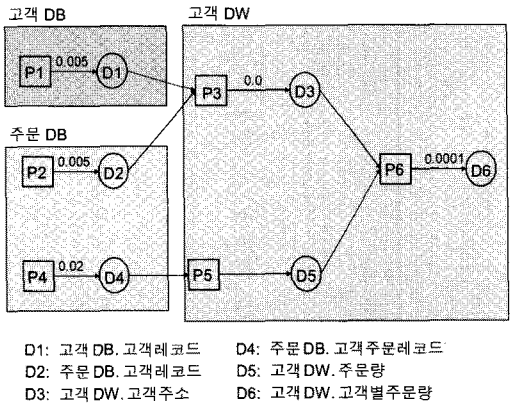
이는 각 경로가 서로 독립적이라고 가정할 경우의 고장율 계산식이다. 이와 달리 각 경

로들이 상호 보완적인 경우에는 병렬계 시스템의 고장률 예측 기법이나 몬테칼로 시뮬레이션(Monte Carlo Simulation)과 같은 보다 고도화된 기법을 사용할 수 있다(강영식 등, 2002).

확장 정보구조그래프는 단일 데이터베이스의 품질관리뿐만이 아니라 다양한 환경에서 데이터 품질관리를 위하여 활용될 수 있다. 특히, 여러 원천 데이터베이스로부터 데이터들을 통합 활용하는 비즈니스 인텔리전스 환경에서 보다 효과적으로 활용된다. 비즈니스 인텔리전스에서 데이터의 품질은 주요한 이슈로 대두되고 있으며, 이러한 데이터들의 품질을 향상시키는 것이 비즈니스 인텔리전스의 주요 성공 요인으로 평가되고 있다(English, 1999).

예를 들면, 고객 데이터베이스로부터 고객 신상정보들을 추출하고, 주문 데이터베이스로부터 고객의 주문 이력을 추출하여 고객 데이터웨어하우스에 저장할 경우, 데이터웨어하우스에 저장된 고객별 주문량의 품질은 이들 원천 데이터베이스의 품질과 변환 프로세스의 품질에 의하여 결정된다. 그리고 원천 데이타베이스의 품질은 이들 원천 데이터들이 만들어지는 프로세스에 의하여 결정된다. 즉 데이터웨어하우스에 저장된 고객별 주문량의 품질은 고객 데이터와 고객 주문이력과 같은 원천 데이터와 고객별 주문량을 생성하는 프로세스의 품질에 의하여 결정된다.

고객별 주문량의 품질을 관리하기 위하여서는 원천 데이터베이스에 포함된 데이터 개체들의 변환구조를 나타내는 정보구조그래프와 고객별 주문량의 변환구조를 나타내는 정보구조그래프를 변환 프로세스를 매개로 연결함으로써 쉽게 관리할 수 있다. 즉 <그림 5>에 표시된 바와 같이 고객 데이터웨어하우스에 저장된 고객별 주문량의 정보구조그래프는 고객 데이터베이스에 저장된 고객레코드와 주문 데이터베이스에 저장된 고객레코드 및 고객주문레코드와 이들을 처리하는 변환 프로세스들로 구성된다. 이와 같이 개별 데이터베이스의 정보구조그래프들을 연결함으로써 통합 데이터베이스의 정보구조그래프를 쉽게 구성할 수 있다. 그리고 이러한 성질을 이용함으로써 확장 정보구조그래프는 여러 데이터베이스들 사이의 데이터 변환 관리를 위하여 효과적으로 이용된다.



〈그림 4〉 데이터베이스간 정보구조그래프

4.2 데이터 품질의 영향 요인 분석

확장 정보구조그래프는 데이터베이스에 저장된 모든 데이터 개체들에 대하여 변환 프로세스와 이에 대한 품질 척도들을 관리한다. 이러한 데이터 품질 관리의 가장 큰 장점은 데이터베이스로부터 산출될 수 있는 모든 정보들에 대하여 품질을 예측할 수 있으며, 이들 정보의 품질에 영향을 미치는 데이터 개체들을 파악할 수 있다는 점이다.

[명제 1] 데이터베이스에 저장되거나 이로부터 도출되는 모든 정보들은 정보구조그래프의 데이터 노드로 표시된다.

(증명) 관계형 데이터 모형에 의하면 데이터베이스로부터 추출될 수 있는 모든 데이터 항목들은 관계형 테이블의 대수식으로 표현된다. 이들을 가상 테이블, 즉 뷰로 구성할 경우, 이들은 정보구조그래프의 데이터 개체 노드이며, 이들

을 산출하는 관계형 대수식은 데이터 개체의 변환 프로세스이다. 따라서 데이터베이스로부터 추출될 수 있는 모든 정보들은 데이터 개체 노드로 표시되며, 이들을 산출하는 프로세스는 데이터 변환 노드로 표시된다.

정보 시스템에 저장되어 있거나 산출될 수 있는 모든 데이터개체들은 정보구조그래프의 데이터 개체 노드로 표시함으로써 특정 데이터로부터 파생되는 모든 정보 항목들을 유추할 수 있으며, 특정 정보를 산출하기 위하여 필요한 모든 원시 데이터들을 파악할 수 있다. 즉 정보구조그래프는 데이터를 산출하기 위하여 필요한 입력 데이터 개체들을 나타내며, 이들을 전방 또는 후방으로 전개함으로써 데이터 개체들 사이의 연관성을 파악할 수 있다.

[명제 2] 정보 산출을 위하여 필요한 모든 원시 데이터들은 해당 데이터 개체 노드를 뿌리(root)로 하는 서브 그래프로 표시되며, 특정 데이터로부터 산출되는 모든 정보들은 해당 데이터 개체 노드를 시점으로 하는 경로로 식별된다.

(증명) 정보구조그래프는 방향성 그래프이다. 즉 특정 데이터 개체를 산출하기 위한 데이터 개체 노드들과 데이터 변환 노드들은 경로를 구성한다. 따라서 해당 데이터 개체를 생성하기 위하여 필요한 모든 원시데이터들은 이들 경로들로 구성된 서브 그래프로 표시된다.

특정 데이터로부터 산출되는 정보들 또한 이를 위한 데이터 변환 노드들과 같이 경로를 구성한다. 따라서 특정 데이터 개체를 시점으로 하는 경로에 포함된 모든 데이터 개체 노드들은 이로부터 생성되는 데이터 개체들을 나타낸다.

정보구조그래프의 서브 그래프를 구성하는 데이터 개체 노드와 데이터 변환 노드들을 추출함으로써 정보 산출을 위하여 필요한 모든 원시 데이터들과 이에 영향을 미치는 변환 프로세스들을 파악할 수 있다. 예를 들면, <그림 4>에서 구매주문에 영향을 미치는 데이터 변환 프로세스들은 다음과 같이 검색된다.

```
SELECT      output_data_object,
            input_data_object
FROM        ISG_input_arc
CONNECT BY  PRIOR input_data_object =
            output_data_object
START WITH  input_data_object = '구매주문'
```

위의 질의는 구매주문을 생성하기 위하여 사용되는 모든 입력 데이터 개체들을 검색하며, 검색 결과는 <표 6>과 같다. 이로부터 구매주문에 영향을 미치는 원인들을 식별할 수 있다.

〈표 6〉 데이터 품질 메타 데이터의 검색 결과

output_data_object	input_data_object
구매주문	안전재고량
구매주문	현재고량
구매주문	고객주문량
구매주문	공급자목록
현재고량	입고량
현재고량	출고량
현재고량	이월재고량

데이터의 품질에 영향을 미치는 원인의 파악과 더불어 이들 정보를 이용하여 역으로 이들 원시 데이터 개체들로부터 생성되는 모든 데이터 개체들을 검색할 수 있다. 예를 들면 입고량으로부터 산출되는 데이터 개체들은 다음과 같이 검색된다.

```
SELECT      input_data_object,
            output_data_object
FROM        ISG_input_arc
CONNECT BY  PRIOR output_data_object =
            input_data_object
START WITH  output_data_object = '입고량'
```

검색 결과는 <표 7>과 같이 나타나며, 이를 통하여 입고량으로부터 파생되는 모든 데이터 개체들을 식별할 수 있으며, 이들 데이터들의 품질에 입고량이 영향을 미침을 파악할 수 있다.

〈표 7〉 데이터 품질 메타 데이터의 검색 결과

input_data_object	output_data_object
입 고 량	현재고량
현재고량	구매주문

V. 결 론

지식 정보를 축적하기 위하여서는 양질의 데이터를 장기적으로 축적 관리할 수 있는 체계가 구축되어야 한다. 이러한 데이터 관리 체계의 구축에서 가장 기반을 이루는 것이 데이터의 생성 과정이다.

본 연구는 비즈니스 인텔리전스를 위한 통합 데이터베이스 환경에서 품질에 영향을 미치는 모든 정보들을 메타 데이터로 정의하고, 이들을 표현하고 관리하기 위한 방안을 정보구조그래프를 근간으로 하여 제시하였다. 확장 정보구조그래프는 데이터베이스를 구성하는 테이블과 속성들 사이의 변환 프로세스를 표현하는 그래프로서 이를 이용할 경우, 데이터베이스 스키마에 대한 메타 데이터와 변환구조에 대한 메타 데이터 및 세부 처리공정들을 효과적으로 표현하고 관리할 수 있다.

이와 같이 데이터의 변환 프로세스를 체계적으로 관리함으로써, 데이터의 외형적 특성뿐만이 아니라 데이터들 사이의 변환 구조와 연관성도 같이 관리할 수 있다. 이러한 데이터들 사이의 연관성과 변환 구조에 대한 메타 데이터들은 여러 출처로부터 자료들을 통합 활용하는 비즈니스 인텔리전스 환경에서 사용자들이 데이터의 의미나 용도를 제대로 파악할 수 있도록 하며, 양질의 데이터를 공유할 수 있도록 한다.

참 고 문 헌

강영식, 백종배, 이근오, 신뢰성공학, 동화기술, 2002.

DPC, DB 품질평가모델 연구보고서, 한국데이터베이스진흥센터, Dec. 2002.

DPC, DB 품질평가모델 확장 개발 연구보고서, 한국데이터베이스진흥센터, July 2003.

English, L.P., Improving Data Warehouse and Business Information Quality, John Wiley & Sons, Inc., 1999.

Lee, C. Y., "A Knowledge Mmanagement Scheme for Meta-Data: An Information Structure Graph", *Decision Support Systems*, Vol. 36, No. 4, March 2004, pp.341-354.

Redman, T. C., *Data Quality for the Information Age*, Artech House, 1996.

Sartori, L. G., *Manufacturing Information Systems*, Addison-Wesley, 1988.

Wang, R. Y., "A Product Perspective on Total Data Quality Management", *Communications of the ACM*, Vol. 41, No. 2, Feb. 1998, pp.58-65.

Information System Review
Volume 6 Number 2
December 2004

A Study on Data Quality Management in Business Intelligence Environments

Choon Yeul Lee*

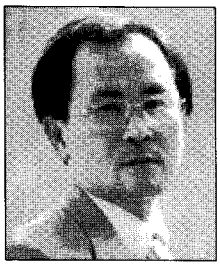
Abstract

Business intelligence assumes an integrated and inter-connected information resources. To manage an integrated database, we need to trace data transformation processes from its outset. For this purpose, this study proposes an extended Information Structure Graph that models data transformation steps in addition to data transformation structures. Using the graph, we can identify relationship among data entities and assign data quality measures to each nodes or arcs of a graph, thus eases management of data and enhanceing their quality.

Keywords: Metadata, Data Processing, Data Quality

* Data & Knowledge Engineering Department, Graduate School of Business IT, Kookmin University

◐ 저 자 소 개 ◑



이 춘 열 (cylee@kookmin.ac.kr)

현재 국민대학교 비즈니스IT전문대학원에 재직중이다. 서울대학교 산업공학과 학사, 서울대학교 대학원 경영학과 석사, 미국 University of Michigan에서 박사학위를 취득하였고(Computer and Information Systems 전공), 한국통신 소프트웨어연구소에 근무하였다. 주요관심분야는 데이터관리, 데이터베이스 모델링, 비즈니스 모델링, 메타데이터 관리, 데이터웨어하우징 등이다.

논문접수일 : 2004년 5월 17일 **게재확정일 : 2004년 11월 11일**
1차 수정일 : 2004년 6월 30일