자료의 질 향상을 위한 데이타베이스의 최적감사시점+

김기수*

Optimal Database Audit Timing for Data Quality Enhancement

Kisu Kim

〈요 약〉

정보시스템이 효과적이기 위해서는 정보가 도출되는 자료의 무결성이 우선 전제되어야 한다. 특히 오늘날과 같이 사회가 다양한 활동들을 지원하기 위해 컴퓨터를 이용한 정보시스템에 점점 더 의존해감에 따라 정보시스템에서 사용되는 자료의 질을 적절한 수준으로 유지 및 관리해야 할 필요성이 더욱 절실히 대두되게 되었다. 그럼에도 불구하고 여전히 관리자들은 효과적인 의사결정 및 활동을 위해 필요한 최신의 정확한 자료들을 제공 받지 못하고 있으며 [Nesbit 1985], 정보시스템이 기대 이하의 성능을 나타내는 가장 단순하고 일반적인 원인은 정보시스템에 입력된 자료가 부정확하거나 불완전하기 때문인 것으로 나타나고 있다 [Ballou and Pazer 1989]. 낮은 질의 자료는 즉각적인 경제적 손실뿐만 아니라 보다 많은 간접적이고 경제적으로 측정하기 어려운 손실들을 초래한다. 그리고 아무리 잘 관리되는 시스템에도 시간이 흐름에 따라 여러 가지 원인에 의해 저장된 자료에 오류가 발생하게 된다. 자료의 질을 적절한 수준으로 유지하기 위해서는 이와 같은 오류는 주기적으로 발견 및 수정되어야 한다. 이와 같은 작업을 데이타베이스 감사라고 한다. 본 논문에서는 데이타베이스에 저장된 자료의 질을 주기적으로 향상시키기 위한 최적 데이타베이스 감사시점을 일반적인 비용모형을 통해 결정하는 과정을 제시하고, 그와 관련된 사항들에 대해 논의하였다. 데이타베이스는 오류 발생률도 다르고, 오류의 결과도 상당히 다른 여러 개의 자료군들로 구성되어 있다고 가정하였다. 그리고 각 자료군에서의 오류 누적과정은 확정적이 아닌 확률적인 과정으로 모형화하고, 단순한 오류의 발생뿐만 아니라 오류의 크기도 확률적으로 변하는 상황을 모형에 반영하여 보다 현실성있게 모형화하였다.

+ 이 논문은 1994년도 한국학술진흥재단의 공모과제 연구비에 의하여 연구되었음

* 영남대학교 상업교육과

1. 서 론

과거 몇십년 동안 경영자들은 가장 최근의 정보기술들을 수용하려는 많은 노력을 기울여왔다. 그 결과 광범위한 요구들을 충족시키기 위해 고안된 다양한 정보시스템들이 지속적으로 쏟아져 나오고 있다. 그러나 이와 같은 시스템들의 두드러진 특징들 가운데 하나는 사용자의 기대와 실지로 실현되는 정보기술의 성능사이에 괴리가 자주 발생한다는 사실이다. 이와 같은 상황이 발생하는 중요한 요인 가운데 하나가 자료의 질 문제이다 [Ballou and Pazer 1985]. U. S. National Bureau of Standards [1979]에 따르면 응용시스템이 바람직하지 못한 성능을 나타내는 가장 단순하고 일반적인 원인은 잘못된 입력자료에 있다고 한다.

정보시스템의 평가는 일반적으로 의사결정과정에서 정보에 의존해야 하는 주로 의사결정자인 사용자에게 제공되어지는 정보의 효익에 초점을 맞춘다 [Ahituv 1980]. 정보시스템은 필요한 자료가 처리 및 사용자 요구와 관련해서 가지고 있는 무결성(integrity)의 수준 이상으로 사용자들에게 효과적일 수는 없다. 그리고 정보가 효과적이기 위해서는 정보가 도출되는 자료의 적절성(validity)이 우선 전제되어야 한다.

정보시스템 문헌들에서 자료는 사실에 대한 "객관적(objective)"인 묘사(descriptions)인 반면 정보는 특정 의사결정자의 관점에서 자료의 "주관적(subjective)"인 해석이라고 정의하면서 자료와 정보를 구분하고 있다 [Agmon and Ahituv 1987]. 자료가 정보가 되기 위해서는 몇 가지 조건이 충족되어야 한다 [Ahituv et al. 1981, Davis and Olson 1985]: 자료는 의미 있고 이해가능해야 한다; 자료는 의사결정 문제에 적절하게 관련이 있어야 한다; 자료는 의사결정자에 의해 자기 목표에 기여할 것으로 인식되어야 한다; 자료는 이용(접근) 가능해야 한다. 이와 같은 모든 조건들이 만족되었다 하더라도 여전히 자료의 신뢰성(질)이라는 한가지 문제가 남는다. 사실 자료의 질에 대한 평가는 앞의 어떤 조건에 대한 검토보다도 우선되어야 한다.

다양한 활동들을 지원하기 위해 오늘날 사회가 컴퓨터를 이용한 정보시스템에 점점 더 의존해감에 따라 오랫동안 인식되어왔던 적절한 자료의 질에 대한 중요성이 더욱 절실하게 되었다. 더욱이 자료자원 및 그 관리의 분산 (예를 들면, 분산 컴퓨팅/분산 데이타베이스시스템)과 같은 정보시스템의 최근 추세 때문에 자료의 무결성을 적절한 수준으로 유지하기가 더욱 어렵게 되었다. 따라서 자료의 무결성을 확보하여야 할 필요성은 오늘날 조직에서 더욱 커지고 있으며 저장된 자료가 완전하고(complete), 정확하고(accurate), 최신(up-to-date)이 되게 할 필요성이 절실히 나타나게 되었다 [Ballou and Tayi 1989].

그럼에도 불구하고 경영자들은 여전히 효과적인 활동 및 의사결정에 필요한 정확한 자료들을 가지고 있지 못한 것으로 나타나고 있다 [Nesbit 1985]. 또한 많은 경영자들은 그들이 사용하고 있는 자료의 질에 대해 잘 모르고 있으며 정보기술이 자료를 완벽하게 해준다고 생각하고, 자료의 질이 낮은 것이 예외적인 것이 아니라 일반적인 현상임에도 불구하고 대부분 자료의 질 문제에 무관심한 것으로 나타났다 [Nesbit 1985]. 질이 낮은 자료는 즉각적인 많은 경제적 손실뿐만 아니라 보다 간접적이고 경제적으로 측정하기 어려운 비용도 수반한다. 예를 들면, 어떤 은행에서 고객의

주민등록번호를 정확하게 가지고 있지 않다면 고객은 자기 돈을 관리할 수 있는 그 은행의 능력에 대해 심각하게 의심을 할 것이다. 한 부서(주문입력 부서)로부터 나온 자료가 다른 부서(고객 지불청구부서)에서 사용될 때 그 자료에 오류가 있으면 상호 불신은 증가하게 된다. 최근에 발표된 자료의 오류로 인한 경제적 손실의 예로는 부정확한 신용보고서, 부정확한 지방세 납부로부터 소비자들을 보호하기 위한 소송과 부정확한 제품레이블에 기인한 부분환불(rebates) 등이 있다 [Redman 1995]. 커뮤니케이션, 금융서비스, 제조, 건강진료 등 어떤 산업조직도 이와 같은 비용으로부터 자유롭지 못하다. 정부조직도 마찬가지이다.

뿐만 아니라 재무 및 다른 관리시스템들에서의 낮은 자료의 질은 전략을 효과적으로 수행할 수 없게 하고, 또한 부정확한 자료는 JIT제조 및 리엔지니어링도 불가능하게 한다.

그리고 아무리 잘 유지·관리되는 시스템에서도 여러 가지 원인에 의하여 시간이 지나면 저장된 자료에 결함이 생기게 된다. 따라서 자료의 질 향상을 위해서는 저장된 자료들에 대해 주기적으로 오류를 발견하고 그것을 수정하는 감사(audit)가 이루어져야 한다.

본 논문에서는 저장된 자료의 부적절한 무결성(질) 문제와 그것의 향상 방법을 다루기 위해 시간이 흐름에 따라 서로 다른 오류 발생률을 가지고 오류로부터 발생하는 결과도 서로 다른 복수의 자료군들(data sets)로 이루어진 데이타베이스의 자료들에 대한 최적감사시점(optimal audit timing)을 결정하는 방법을 제시하고 관련사항들에 대해 논의하려고 한다.

2. 연구배경

자료에 내재해 있는 오류의 정도 및 영향에 대한 인식이 증가하고 있다. Computerworld와 Information Week가 최근 대기업과 독자들을 조사한 결과 대부분의 조직이 자료의 질 문제로 고민하고 있다는 결론을 내렸다 [Knight 1992, Liepens 1989]. 낮은 질의 자료가 비용을 수반한다는 데는 일반적으로 공감하고 있으나 자료의 질 수준 또는 자료의 신뢰도(data reliability)에 대한 측정이나 실증적 통계분석을 위한 체계적인 연구는 상대적으로 적은 편이다. 그리고 이들 연구들은 자료의 질 가운데 정확성(accuracy) 차원만 반영하고 적합성(relevancy), 일관성(consistency) 등과 같은 자료 질의 다른 문제들은 반영하지 않고 있다.

Johnson et al. [1981]은 미수금계정과 재고계정 감사에서의 오류율을 연구했다. 그들은 오류율과 오류의 크기는 계정의 종류에 따라 다르고, 오류율은 거래(transaction)의 수가 증가함에 따라 증가하고, 소규모 기업뿐만 아니라 대기업들도 상당한 오류율을 가지고 있다는 사실을 발견했다. 상대적으로 오류가 없을 것이라고 일반적으로 생각되는 금융시스템에서도 조사된 시스템들 가운데 사분의 일이 9퍼센트를 넘는 오류율을 가진 것으로 나타났다. Morey [1982]는 군대 인사시스템에 저장된 레코드들의 오류율이 10퍼센트가 넘는다는 결론을 내렸다. 많은 인사시스템들이 이 정도의 오류율을 가지고 있을 것이라고 생각해도 무리가 없을 것이다. Laudon [1986]은 형사재판 활동을 지원하는 정보시스템에 저장된 자료의 질에 대한 방대한 연구를 했다. 조사된 두 시스템 가운데 전산화된 시스템에서는 완전히 정확한 레

코드는 절반도 되지 않고, 비전산시스템에서는 사분의 일 정도만 완전히 만족할 만한 수준임을 발견했다. Laudon은 이와 같이 낮은 자료의 무결성을 나타내는 것은 형사재판 정보시스템의 조직간 특성 때문이라고 지적했다. Agmon과 Ahituv [1987]는 품질관리분야에서 사용되는 신뢰도라는 개념을 정보시스템에 적용했다. 이들은 자료의 신뢰도에 대한 세가지 측도를 개발했다. 즉 다양한 자료항목들에 "공통으로 받아들여지는" 특성을 반영하는 내부신뢰도(internal reliability), 자료가 사용자의 요구들에 일치하는 정도를 나타내는 상대적 신뢰도(relative reliability), 그리고 자료항목들의 실제사항과의 유사성 정도를 결정하는 절대적 신뢰도(absolute reliability)가 그것이다. 이들 세가지 측도 사이의 관계에 대해서 논의하고 실증적 연구결과를 제시했다.

지난 수년동안 자료의 질 및 정확성 차원에 대한 중요성이 인식되면서 자료의 무결성을 확보하기 위한 효과적인 절차 및 통제들이 개발되었고, 그중 일부는 성공적으로 실행되었다. 정보시스템에서 자료 질의 영향을 평가하는 일반모형의 개발을 위한 상당한 연구가 또한 있었다. 이와 같은 연구는 주로 회계문헌들에 나타났으며 회계시스템에 들어가기 전에 자료에 있는 오류를 발견해서 수정하기 위한 효과적인 내부통제시스템의 개발에 초점을 맞추었다. Cushing [1974]에 의해 신뢰도모형이 제안되었고 Bodner [1975], Stratton [1981], Hamlen [1980] 등에 의해 확장되었다. 예를 들면 Halmen [1980]은 Cushing [1974]에 의해 원래 개발된 모형의 틀을 확장해서 구체적인 오류수준을 확보하는 통제시스템을 설계하는데 이용될 수 있는 모형을 확률적 제약조건을 가진 혼합정수계획(chance-constrained mixed integer program)으로 개발했다. 분산시스템 환경에서의 감사에 관련된 문제들은 Hansen [1983]에 의해 논의되었다. 대부분의 이와 같은 연구들은 컴퓨터시스템에 잘못된 자료들이 저장되는 것을 방지하는 절차에 주로 초점이 맞추어져 있다.

시간이 흐름에 따라 회계자료에 누적되는 오류의 영향과 감사 및 수정에 관한 연구는 그렇게 많지 않으며 조직의 내부통제시스템의 일부로 최적감사시점이란 제목으로 이루어졌다. Hughes [1977]는 내부감사의 최적시점을 결정하는데 적절한 일반적인 의사결정모형을 구축했다. 그는 이 문제를 동적계획 문제에 일반적인 후진귀납(backward induction)방법을 사용해서 풀었다. Boritz와 Broca [1986]는 확정적(deterministic) 모형을 이용해서 감사사이의 최적고정시간간격(optimal fixed time interval)을 결정했다. Morey와 Dittman [1986]은 감사사이에 한 계정에 허용가능한 최대누적화폐차액(maximum accumulated dollar discrepancy)과 그 한계치를 넘지 않을 신뢰수준을 목표로 필요한 감사시점을 결정하는 모형과 해법을 제시했다. 그러나 이들은 모두 거래처리에 관련된 단일 회계계정에 대한 감사만 고려하고 보다 포괄적인 데이타베이스의 자료들에 대한 감사는 고려하지 않고 있다. Morey [1982]는 자료 갱신 및 수정 절차들의 저장된 자료자원의 오류율에 대한 영향을 평가하는 한 방법을 제시했다. Ballou와 Pazer [1985]은 최종사용자를 위해 다양한 처리과정을 거치는 여러 자료집합들의 각각에 있는 오류의 결과를 분석했다. 복수응용환경(multi-application environment)에서는 이와 같은 처리과정들이 또한 잠정적으로 오류를 발생시킬 수 있다. 이들의

모형은 다수의 중간 및 최종출력에 내재해 있을 수 있는 오류들을 입력 및 처리과정의 오류함수로 표현한다. 이 오류에 대한 표현을 이용해서 선택된 출력에 대해서 서로 다른 질 통제절차들(alternative quality control procedures)이 가지는 영향을 분석할 수 있다. 시스템범주의 비전산 및 전산 통제와 절차들의 역할을 포함해서 자료질의 다양한 면들은 Baily [1983]에 의해 논의되었다.

Janson [1988]은 최종사용자 컴퓨팅에서의 자료의 질 문제를 다루었다. 낮은 자료의 질이 최종사용자 컴퓨팅에서 주요한 위험요소이고 공공조직에서 컴퓨터의 효과적인 사용에 중요한 방해요소가 되는 것으로 밝혀졌다. 예로서 그는 1970년대 미국에서 에너지 절약을 위해 추진된 에너지정보시스템이 실패한 주요 원인 가운데 하나가 낮은 자료의 질 때문이었다는 사실을 든다. 그는 최종사용자 컴퓨팅에서 자료의 질을 향상하기 위해 실험통계기법들(exploratory statistical techniques)을 제시하고 앞의 예를 이용해서 자료획득의 전 단계에서 자료의 질을 상당히 향상시킬 수 있다는 것을 보였다. 또한 이 기법들은 이미 수집된 자료의 질 통제와 검증(validation)에 특히 중요하다는 것도 보였다. Ballou와 Tayi [1989]는 자료의 질 향상을 위해 다수의 서로 다른 자료군들에 한정된 자원을 최적으로 할당하는 방법을 제시했다. 보다 최근에는 Liepens [1989], Fox, et al. [1994], Knight [1992], Redman [1995] 등이 자료의 질 문제와 현대 정보시스템에서 정확한 자료의 중요성에 대해 논의했다. 특히 Redman은 AT & T사에서 자료의 낮은 질을 인식하고 그 질을 향상시키위해 사용하는 한 과정을 설명하고 있다. 그는 자료의 질 문제를 확인하고, 자료를 중요한 자산으로 취급하고, 품질시스템(quality system)을 자료를 만들어내는 과정(process)에 적용하는 3단계 방법을 제안했다.

실증적 연구결과 자료의 무결성 정도는 자료군마다 매우 다르다는 사실을 입증하고 있다. 이것은 여러 가지 요인들에 의해 설명될 수 있다. 아마 가장 중요한 요인은 시스템마다 자료의 질에 대한 요구가 다르기 때문일 것이다. 특정 자료를 사용해서 이루어지는 의사결정의 비용과 효익 그리고 의사결정의 저장된 자료의 질에 대한 의존성이 자료의 질에 대한 요구를 결정하는데 주요 고려사항이 된다. 한 예로 의사결정에 사용되는 자료의 통합 정도를 들 수 있다. 만일 자료에 어떤 체계적인 편차(systematic biases)가 없다면 자료의 통합은 상대적인 오류를 줄이고, 따라서 높은 자료의 무결성에 대한 요구를 불필요하게 하는 상황이 될 수 있다. 한편 개인의 의료기록에 기초한 적절한 치료에 관한 의사결정은 가능한 한 오류가 없는 자료를 필요로 할 것이다.

자료의 질 수준에 영향을 미치는 또 하나의 요인은 자료의 소유(ownership)이다. 만일 어떤 자료를 필요로 하는 사람이 또한 그 자료를 수집하고, 저장하고, 유지하는 책임도 가지고 있다면, 높은 수준의 자료의 무결성을 확보하기 위한 내적 동기가 있게 된다. 그러나 최종사용자 컴퓨팅에서도 자료의 질이 낮은 것이 일반적이고 자료의 질이 최종사용자정보시스템의 성패에 중요한 요인이 된다고 한다 [Janson 1988]. 데이타베이스 환경에서는 자료의 저질화 가능성에 대해 두 가지 상충되는 면이 있다. 한편으로는 데이타베이스관리시스템(DBMS)과 자료사전(data dictionary)이 다양한 무결성 조건들을 구현할 수 있고, 데이타베이스 관리자(database administrator)의 책

임이 자료의 질에 대한 규제들의 수립과 준수에 도움을 줄 수 있다. 그러나 데이타베이스 시스템에서는 자료 값에 대해 하나의 원천에만 더욱 의존하고, 따라서 그것이 부정확하면 그 값을 사용하는 모든 응용에 나쁜 영향을 미치게된다 [Brodie 1980, Davis and Olson 1985]. 더욱이 비전산(manual) 혹은 특수목적(dedicated) 시스템에서는 자료를 자동적으로 모니터 하는 사람들이 있으나 데이타베이스 환경에서는 그렇게 할 수 없다. 특히 데이타베이스에서의 자료의 질에 관한 다른 고려사항들은 Brodie [1980]에 의해 검토되었다.

자료의 질에 영향을 미치는 또다른 요인으로는 오류에 의해서 누가 어느 정도의 영향을 받느냐는 것이다. Laudon [1986]에 의해 검토된 형사재판시스템의 경우와 같이 오류의 결과가 시스템의 외부에 있는 사람들에게 주로 돌아간다면 자료의 무결성을 유지할 동기가 줄어들게 될 것이다. 또한 저장된 자료의 장기적 질은 자료가 사용되고 검토되는 빈도에 많은 영향을 받게 된다. 예약시스템에 있는 오류는 상대적으로 빨리 발견되는 반면 신용평가(등급) 파일에 있는 오류는 그 심각성 정도에 따라 결코 발견되지 않을 수도 있다.

3. 문제의 정의 및 논의

정보시스템의 목적은 일반적으로 의사결정과정에서 사용자에게 필요한 양질의 정보를 제공하는데 있다. 정보의 질은 그것이 어떻게 인간 행동에 동기를 부여하고 효과적인 의사결정에 기여하느냐에 의해 결정된다. 의사결정자의 관점에서 정보의 질은 정보의 효용성(utility of information), 정보만족(information satisfaction), 정확성(accuracy) 등을 포함한다 [Davis and Olson 1985]. Andrus [1971]는 정보는 정확성 외에 정보의 사용을 촉진하거나 저해할 수 있는 효용성으로 평가될 수 있다고 했다. 그는 효용을 형태효용(form utility), 시간효용(time utility), 장소효용(place utility, physical accessibility), 소유효용(possession utility, organizational location)으로 구분했다. 정보의 형태가 의사결정자의 요구와 보다 밀접하게 일치할수록 정보의 가치는 증가한다. 정보는 필요할 때 이용할 수 있다면 더 많은 가치를 가진다. 정보에 쉽게 접근하거나 정보가 쉽게 전달된다면 정보는 더 많은 가치를 가진다. 온라인 시스템은 시간효용과 장소효용을 모두 최대화 해준다. 정보의 소유자는 다른 사람들에게로의 정보의 확산을 통제함으로써 정보의 가치에 큰 영향을 미친다. 정보만족은 의사결정자가 공식적인 정보시스템으로부터 나온 출력에 만족하는 정도이다. 만일 의사결정자가 공식적인 정보시스템이 필요한 정보를 제공해줄 수 있어야 한다고 생각한다면 그는 우선 그기에 필요한 정보에 대해 질의할 것이다. 거기서 필요한 정보를 즉시 얻을 수 있으면 그 정보시스템에 대한 만족은 강화될 것이다. 그러나 그렇지 않으면 공식적인 정보시스템에 대한 좌절이나 불만족이 강화될 것이다. 정보만족은 공식적인 정보시스템들을 평가하는 하나의 기준으로 사용될 수 있다. 정확성은 정보에 내재된 오류의 정도를 나타낸다. 본 논문에서는 정보의 정확성에 초점을 맞춘 정보의 질을 논의의 대상으로 하려고 한다. 왜냐하면 위에 언급한 정보의 질에 대한 세 가지 측면들 가운데서 정확성이 가장 기본적이고, 객관적 평가가 가능하고, 다른 것들 보다 우선적으로

고려되어야 한다고 생각되기 때문이다. 그리고, 물론 자료에 있는 오류는 처리과정에 의해 그 오류가 증폭되기도 하고, 축소되기도 하고, 그대로 남아있기도 하지만, 정보의 정확성은 그 정보가 도출되는 자료의 정확성에 크게 달려있다.

오늘날 어떤 조직에 내재하고 있거나 조직에 의해 생성되는 자료는 관리되어지고, 합법적으로 필요로 하는 모든 사람이 사용할 수 있는 조직의 자원으로 간주되고 있다. 모든 형태의 저장된 자료라는 의미로서의 정보자원은 다른 자원과 마찬가지로 효과적으로 관리되어야 한다. 이와 같은 정보자원은 오늘날 일반적으로 데이타베이스의 형태로 관리된다. 그러나 여러 가지 원인에 의해 데이타베이스에 저장된 자료들에 오류가 발생하게 된다. 예를 들면 부정확한 자료의 측정 및 수집방법, 정확한 처리절차들을 따르지 못함, 자료의 손실이나 비처리, 자료를 잘못 기록하거나 수정함, 잘못된 매스터파일 혹은 다른 매스터파일의 사용, 컴퓨터프로그램 오류와 같은 처리절차의 오류, 고의적인 오류 등의 결과이다. 여기서 오류는 무작위 오류, 편차, 불완전한 자료, 자료의 불일치 등 자료의 무결성을 저해하는 모든 형태의 오류를 포함한다. 대부분의 정보시스템들에서는 정보를 받는 사람은 그 정보의 질에 영향을 미치는 오류에 대해서 모르고 있다. 보고서 형태로 제공되는 정보는 그것이 작성되는 과정과 보고서에 포함된 자료의 정확성에 의해 영향을 받기 때문에 정보의 사용자는 그 정보의 정확성을 확신할 수가 없다. 예를 들어 재고보고서에 347개의 부품이 현재 재고로 남아있다고 하자. 그러나 이 숫자는 지속적인 장부가치로서의 재고에 기초한 것일 가능성이 있다. 여기에는 재고 입·출고 등을 기록하는데 다양한 오류가 포함될 수 있기 때문에 많은 경우에 적은 오류 그리고 몇몇 경우에는 큰 오류가 있을 수 있다. 그렇기 때문에 장부가치의 재고를 수정하기 위해 주기적인 재고실사가 이루어진다. 오류를 발견하기 위한 내부통제, 내부 및 외부 감사, 자료에 "신뢰한계"를 추가, 사용자들이 가능한 오류들을 평가할 수 있도록 측정 및 처리 절차에 대한 사용자 교육 등을 통해 자료의 오류 문제들을 어느 정도 극복할 수 있다.

구축된 데이타베이스를 평가하는데 고려되어야 할 중요한 한가지 요소는 자료의 무결성(정확성, 완전성 등)이 유지될 수 있는 확률이다. 데이타베이스의 평가는 또한 자료 질의 저질화로부터의 위험도 고려해야 한다. 자료의 질을 유지하는 조직의 능력은 조직적 요인과 자료 요인 모두에 달려있다 [Davis & Olson 1985]. 첫째는 오류-효과 사이클의 길이이다. 만약 오류들이 즉각적인 효과를 가지고 있으면 오류의 효과가 장기적일 때보다 조직의 자원이 보다 쉽게 사용될 것이다. 그러므로 계산청구서에 있는 오류들이 고용자 정보파일에 있는 오류 보다 관심을 더 얻을 수 있다. 둘째는 측정의 규칙성이다. 자주 그리고 규칙적인 간격으로 이루어지는 자료의 수집은 조직의 절차들을 통해 규정화 될 가능성이 높다. 가끔 임시변통으로 이루어지는 자료의 수집은 잊어버릴 수도 있고 잘 이루어지지 않는다. 지출보고의 일부로서 영업직원으로 부터 오는 경쟁사 정보(competitor intelligence)에 대한 규칙적인 주별보고는 가끔씩 이루어지는 정보보고 보다 높은 무결성을 가질 가능성이 높다. 셋째는 사용자와 자료제공자 사이의 연결성이다. 자료가 제공하는 기능이 그 사용자와 아무런 조직적 연결성을 가지고 있지 않을 때 높은 질의 자료를 유지하기가 더 어렵게 된다. 만일 자료를 외부원천으로부터 얻는다면 그 조직은 자

신의 질통제표준(quality control standards)을 적용할 수 없게 된다. 넷째는 자료제공자의 자료기강(data discipline)이다. 완전하고 정확한 자료를 달성하려는 연습(훈련)이 따르는 어떤 기능의 자료기강의 정도는 그 기능의 배경, 훈련, 문화의 결과이다. 예를 들면 회계기능은 마케팅 보다 일반적으로 높은 자료기강을 가지고 있다. 다섯째는 검증(verification)의 용이성이다. 어떤 자료항목은 다른 저장된 자료나 물리적 증거와 비교함으로써 쉽게 확인될 수 있다. 미수금계정의 차변은 현금계정의 대변에 의해 (기타 정정과 함께) 쉽게 검증될 수 있다.

이와 같은 요인들에 의해 데이타베이스에 들어있는 다양한 자료군들은 서로 다른 오류율을 가지고 오류의 결과도 상당히 다를 수 있다. 어떤 자료군은 적은 오류율을 가지지만 각 오류가 심각한 결과를 가져올 수 있는 반면 다른 자료군은 높은 오류율을 나타내지만 각 오류의 영향은 크지 않을 수 있다. 예를 들면 환자들의 의료데이타베이스에서 각 환자의 특정 약에 대한 엘러지나 각종 의료검사결과들은 환자의 주소나 보험상태 보다 오류율은 낮지만 오류가 있을 때는 치명적인 결과를 초래할 수 있다.

지금까지의 대부분의 연구들은 컴퓨터시스템에 오류가 있는 자료가 저장되는 것을 방지하는 절차들(주로 data validation방법들)을 제시하고 있다. 그러나 저장된 자료의 무결성(질)을 만족할 만한 수준으로 유지하는 것은 계속해서 비용이 수반되는 지속적인 작업이다. 또한 아무리 잘 관리되는 시스템에서도 시간이 흐름에 따라 저장된 자료에 오류가 생기고 자료의 질은 떨어지게 된다. 따라서 데이타베이스에 저장된 자료들은 주기적인 감사를 통하여 오류를 발견하고 수정함으로써 적절한 수준으로 자료의 질이 유지되도록 해야한다.

이와 같은 관점에서 본 논문은 시간이 흐름에 따라 확률적으로 자료의 질이 변하는 상황에서 자료의 질 저하로 발생하는 비용과 자료감사에 수반되는 비용을 고려한 최적감사시점을 결정하는 모형 및 방법을 제시하고자 한다. 앞에서 언급되었듯이 데이타베이스는 조직적 요인과 자료요인에 따라 오류 발생율도 다르고 각 오류로부터 발생하는 비용도 서로 다른 논리적으로 관련된 자료군들로 구성되어 있고 이들은 동시에 감사된다고 가정한다. 여러 자료군들을 동시에 감사하면 자료군들 사이에 있을 수 있는 상호의존성(interdependencies)을 이용하여 자료군별로 각각 감사할 때보다 감사비용을 줄일 수 있다. 예를 들면 한 자료군의 감사로부터 나온 정보가 다른 자료군들의 오류에 대한 정보를 제공해 줄 수 있기 때문이다. 물론 감사기술 혹은 특성상 데이타베이스에 저장된 자료군들을 모두 동시에 감사할 수 없는 경우도 있을 수 있고, 극단적으로는 각 자료군별로 독자적인 감사가 이루어져야 할 수도 있다. 이 경우는 본 논문에서 사용된 모형의 특수한 경우로 취급해서 동시에 독자적으로 감사가 이루어질 수 있는 자료군별 최적감사시점을 각각 구할 수 있을 것이다.

그리고 자료의 오류로 인한 비용은 자료군별로 다를 뿐만 아니라 같은 자료군 내에서도 각 오류의 정도에 따라 비용이 다르게 나타날 수 있다고 가정한다. 또한 각 자료군의 오류누적과정(error accumulation process)은 시간이 흐름에 따라 단순히 증가만 하는 것이 아니라 증가도 할 수 있고 감소도 할 수 있다고 가정한다. 실지로 거래처리과정에서 발견되는 오류는 즉각 수정되

고 발견되지 않는 오류만 누적되기 때문에 누적오류는 줄어들 수도 있다. 이와 같이 본 논문에서는 가능한 한 여러 가지 가정들을 일반화해서 실지상황에 보다 잘 적용될 수 있도록 하였다.

3.1 변수들의 설명

모형의 개발 및 최적감사시점의 표현을 위해 다음과 같은 기호(변수)를 사용한다.

n : 데이타베이스에 저장된 자료군들, S_i, i = 1, 2, . . ., n, 의 수

N_i : 자료군 i에 들어있는 자료단위(data units)의 수

A_{ik} : 자료군 i에서의 k번째 오류의 크기

$L_i(x)$: 자료군 i에서 각 오류의 크기(A_{ik})가 x일 때의 단위시간당 비용

λ_i : 자료군 i의 정확한 각 자료항목의 단위시간당 오류발생율

μ_i : 자료군 i의 오류인 각 자료항목의 단위시간당 오류발견 및 수정율

F : 고정 감사비용

c_i : 자료군 i에서 감사시 발견된 각 오류를 수정하는 비용

d_i : 자료군 i의 자료항목당 변동 감사비용(오류발견비용)

$f_i(x)$: 자료군 i의 각 오류의 크기(A_{ik})에 대한 확률밀도함수

자료군의 수(n)는 자료관리자가 무결성을 유지하기 원하는 데이타베이스에 저장된 레코드, 필드 혹은 필드의 집합들에 달려있다. 또한 동시에 같이 감사될 수 있는 자료의 범위에 따라서도 달라질 수 있다. 자료군 i는 어떤 파일에 들어있는 모든 레코드, 그 파일에 들어있는 레코드들 중에서 모든 활동적인 레코드, 혹은 그 파일의 모든 또는 일부 레코드로부터 나온 필드들의 부분집합이 될 수 있다. 다른 경우들도 있을 수 있다. 예를 들면, 한 자료군은 어떤 파일의 각 레코드로부터 나온 필드들의 부분집합으로 구성될 수 있는 반면 다른 자료군은 같은 파일로부터 나온 필드들의 또 다른(disjoint) 부분집합이 될 수도 있다. 이와 같은 현상은 필드들의 두 부분집합이 상당히 다른 오류발생율과 오류발견 및 수정율 그리고 서로 다른 오류비용 및 감사비용을 가지고 있을 때 가능하다. 만일 각각 같은 형태의 여러 파일들이 동일한 오류발생율과 오류발견 및 수정율 그리고 동일한 오류비용 및 감사비용을 가지고 있다면 이와 같은 여러 파일들을 합쳐서 한 자료군으로 형성할 수도 있다.

N_i의 값은 자료군 i에 있는 자료항목(data items)의 수를 나타낸다. (하나의 자료항목은 구체적인 자료군의 구성단위이다. 이것은 자료파일로부터 나온 필드들 가운데 일부일 수 있다.) 앞에서도 언급했듯이 이 자료항목의 수는 어떤 파일에 있는 레코드의 수와 같을 수도 있고 다를 수도 있다. 많은 데이타베이스의 경우 N_i의 값은 아주 큰 수가 될 것이다.

$L_i(x)$는 자료군 i에 있는 각 오류의 크기(정도)가 x일때의 단위시간당 오류비용을 나타낸다. 이것은 자료군 i의 각 자료항목에 있는 오류나 결함이 가지는 조직에 대한 악영향의 정도를 오류의 크기에 대한 함수로 표현한 것이다. 물론 각 오류는 조직에 독특하고 다른 오류와 구별되는 영향을 미칠 수 있지만, 여기서는 같은 자료군 내에서 같은 크기의 오류는 동일한 비용(평균비용)을 가지는 것으로 가정한다. 따라서 하나의 자료군에서

발생되는 오류들의 영향이 상대적으로 동일하도록 자료군을 구성하는 것이 중요하다. 그러나 다른 자료군들의 오류비용은 서로 상당히 다를 수 있다. 어떤 자료군의 오류는 그 크기가 상당히 크지만 그 영향은 그렇게 크지 않을 수 있으며 반면 어떤 자료군의 오류는 조그만 오류에도 상당히 민감한 영향을 미칠 수 있다.

자료군 i의 단위시간당 오류발생율 λ_i는 한 자료항목이 단위시간 내에 오류가 될 가능성(확률)의 추정값이다. 마찬가지로 μ_i는 자료군 i의 한 자료항목이 단위시간내에 오류상태에서 수정되어 정확하게 될 가능성에 대한 추정값이다. 본 모형에서는 거래처리과정에서도 데이타베이스에 있는 오류가 발견될 수 있고 이 때 발견된 오류는 즉시 수정될 수 있다고 가정한다. 따라서 거래처리과정에서 발견되지 않은 오류들만 데이타베이스에 누적되고, 이 누적된 오류의 수는 시간이 흐름에 따라 증감한다. 이 추정값들은 특정 자료군의 각 자료항목이 정확한(오류의) 상태에서 오류의(정확한) 상태로 변하는데 걸린 평균시간을 각각 측정하여 그 역수를 취함으로써 얻을 수 있다. 여기서 오류라는 것은 각 자료군에서 발견되는 모든 형태의 오류를 포함한다. 이 값들은 단위시간을 무엇으로 (예를 들면, 시간, 일, 주, 달, 년 등) 잡느냐에 따라 상당히 다르게 나타날 수 있다.

$f_i(x)$는 자료군 i의 각 자료항목에 발생하는 오류의 크기에 대한 확률밀도함수이다. 동일한 자료군 내에서 발생하는 오류들의 크기는 서로 다를 수 있지만 동일한 확률밀도함수를 가진다고 가정한다. 이 값들은 자료군별로 표본추출된 오류들의 크기를 분석함으로써 얻을 수 있다. 자료군마다 오류의 크기가 상당히 다를 수 있다. 예를 들면, 고가품의 경우 재고수량의 오류의 크기는 그렇게 크지 않아도 재고금액에서의 오류는 상대적으로 훨씬 클 수 있다.

C_i는 감사과정에서 오류로 밝혀진 자료군 i의 각 자료항목을 수정하는데 드는 비용이다. d_i는 자료군 i의 각 자료항목을 감사하는데 드는 비용이다. 실지로 데이타베이스 감사는 각각 비용을 수반하는 오류를 발견하고 수정하는 두단계 과정의 작업이다. F는 감사를 실시할 때 발생하는 준비비용 등의 고정비용이다.

3.2 모형의 개발 및 설명

우선 각 자료군별 오류누적과정을 birth and death 과정으로 모형화한다. 작은 시간구간 δt를 생각하자. $X_i(t)$를 t시점에서 자료군 i의 저질화 정도를 나타내는 오류인 자료항목의 수라고 하고, 정확한 각 자료항목이 시간 t와 $t+\delta t$사이에 오류가 될 확률을 $\lambda_i\delta t + o(\delta t)$라고 하자. 여기서 $o(\delta t)$는 δt보다 빨리 0에 수렴하는 δt의 함수이다 ([Ross 1981], pp.171). 또한 둘 이상의 정확한 자료항목이 t와 $\mu_i\delta t + o(\delta t)$사이에 오류가 될 확률은 $o(\delta t)$이라고 가정한다. 같은 방법으로 $\mu_i\delta t + o(\delta t)$를 자료군 i의 오류인 각 자료항목이 시간 t와 $t + \delta t$사이에 수정되어 정확한 자료항목으로 될 확률이고, 같은 시간사이에 둘 이상의 오류인 자료항목이 수정되어 정확한 자료항목으로 될 확률은 $o(\delta t)$라고 하자. 그리고 오류발생과 오류발견 및 수정은 각각(상호) 독립적으로 일어난다고 가정하자. 그러면 자료군 i에 k개의 오류인 자료항목이 있을 때 정확한 자료항목의 수는 $(N_i - k)$이므로 자료군 i에서 시간 t와 $t + \delta t$사이에 정확한 한 자료항목에 오류가 발생할 확률은 $\lambda_i(N_i - k)\delta t + o(\delta t)$이고, 같은 시간사

이에 오류인 한 자료항목이 수정되어 정확한 자료항목으로 될 확률은 $\mu_t k\delta t + o(\delta t)$이다. 따라서 $k = 0, 1, \ldots, N_i$에 대해 $\lambda_k = \lambda_t(N_i - k)$이고 $\mu_k = \mu_t k$인 birth and death 과정이 된다.

만일 $P_{ki}(t)$를 자료군 i에서 시간 t에 오류가 있는 자료항목의 수가 k일 확률 (즉 $P_{ki}(t) = P(X_i(t) = k)$)이라고 하면 다음과 같은 $P_{ki}(t)$에 대한 표현을 얻을 수 있다.

$$\begin{aligned}
P_{0i}(t+\delta t) &= (1 - \lambda_t N_i \delta t)P_{0i}(t) \\
&\quad + \mu_t \delta t P_{1i}(t) + o(\delta t), \\
P_{ki}(t+\delta t) &= (N_i - k + 1)\lambda_t \delta t P_{(k-1),i}(t) \\
&\quad + (k+1)\mu_t \delta t P_{(k+1),i}(t) \\
&\quad + \{1 - (N_i - k)\lambda_t \delta t\} \\
&\quad \{1 - k\mu_t \delta t\} P_{ki}(t) \\
&\quad + o(\delta t), \\
&\quad \text{for } k = 1, 2, \ldots, N_i - 1 \\
P_{Ni}(t+\delta t) &= (1 - N_i \mu_t \delta t)P_{Ni}(t) \\
&\quad + \lambda_t \delta t P_{Ni-1}(t) + o(\delta t). \quad (1)
\end{aligned}$$

이 상황을 정확한 자료항목들로 구성된 모집단(population)과 오류인 자료항목들로 구성된 모집단 사이를 이동하는 과정으로 생각할 수 있다. 자료군 i의 오류인 자료항목의 수가 k일 때 정확한 자료항목들이 오류인 자료항목으로 이동하는 순간율(instantaneous rate)은 $(N_i - k)\lambda_t$ 이고 반대방향으로 이동의 순간율은 $k\mu_t$ 이다. 여기서 λ_t는 자료군 i에 대한 자료 입력과정에서의 통제(오류 방지노력) 정도를 나타내고, μ_t는 자료군 i에 대한 오류발견 및 수정 노력의 정도를 나타낸다고 볼 수 있다.

$P_{ki}(t)$항을 등호의 좌변으로 이항하고 등호의 양변을 δt로 나눈후 $\delta t \to 0$의 극한을 취하면 위의 등식들은 다음과 같은 미분차분방정식 (differential difference equations)으로 표현된다.

$$\begin{aligned}
P_{0i}'(t) &= -\lambda_t N_i P_{0i}(t) + \mu_t P_{1i}(t) \\
P_{ki}'(t) &= (N_i - k + 1)\lambda_t P_{k-1,i}(t) \\
&\quad + (k+1)\mu_t \delta t P_{k+1,i}(t) \\
&\quad - \{(N_i - k)\lambda_t + k\mu_t\} P_{ki}(t), \\
&\quad \text{for } k = 1, 2, \ldots, N_i - 1 \\
P_{Ni}'(t) &= -N_i \mu_t P_{Ni}(t) + \lambda_t P_{Ni-1}(t).
\end{aligned} \quad (2)$$

이 미분차분등식들이 각 자료군에서 오류누적과정의 확률적인 활동(behavior)을 정의한다.

다음과 같은 발생함수(generating function)을 정의하자:

$$\Psi_{N_i}(z, t) = \sum_{k=0}^{N_i} z^k P_{ki}(t). \quad (3)$$

그리고 등식들 (2)의 양변에 Z^n을 곱해서 모두 더하면 다음과 같은 등식을 얻는다:

$$\begin{aligned}
\sum_{k=0}^{N_i} z^k P_{ki}'(t) &= -\lambda_i \sum_{k=0}^{N_i - 1} (N_i - k) z^k P_{ki}(t) \\
&\quad - \mu_i \sum_{k=1}^{N_i} k z^k P_{ki}(t) \\
&\quad + \lambda_i \sum_{k=1}^{N_i} \{N_i - (k-1)\} z^k P_{(k-1),i}(t) \\
&\quad + \mu_i \sum_{k=0}^{N_i - 1} (k+1) z^k P_{(k+1),i}(t).
\end{aligned} \quad (4)$$

$\frac{\partial \Psi_{N_i}(z, t)}{\partial t} = \sum_{k=0}^{N_i} z^k P_{ki}'(t)$ 이고

$z\frac{\partial \Psi_{N_i}(z, t)}{\partial z} = \sum_{k=1}^{N_i} k z^k P_{ki}'(t)$ 이라는

사실을 인식하면 다음과 같은 $\Psi_{N_i}(z,t)$에 대한 편미분 방정식을 얻는다.

$$\frac{\partial \Psi_{N_i}(z,t)}{\partial t} = -\lambda_i N_i (1-z)\Psi_{N_i}(z,t) + (\lambda_i - \mu_i) z(1-z)\frac{\partial \Psi_{N_i}(z,t)}{\partial z} \quad (5)$$

위의 식 (5)로 부터 시간 t에 자료군 i에 오류 자료항목의 수가 k일 확률, $P_{ki}(t) = P(X_i(t) = k)$, 을 다음과 같이 구할 수 있다:

$$P_{ki}(t) = \binom{N_i}{k} B_i(t)^k (1 - B_i(t))^{N_i - k},$$

여기서 $B_i(t) = \frac{\lambda_i}{\lambda_i + \mu_i}(1 - e^{-(\lambda_i + \mu_i)t})$

이다. (6)

이것은 이항분포의 형태이고, 또한 (2)의 미분 방정식들을 만족시키는 해이다.

본 연구에서는 또한 각 자료의 오류비용은 오류의 크기(정도)에 따라 다르다고 가정한다. 즉 각 오류의 크기가 오류비용에 영향을 미친다는 것이다. 자료군 i의 각 오류의 크기 A_{ki}는 확률분포를 하며, 만일 그것이 연속적이면 확률밀도함수 $f_i(x)$를 가진다. 물론 A_{ki}가 이산적이면 $f_i(x)$를 이산확률함수(probability mass function)로 보면 된다. 한 자료군의 각 자료항목이 boolean형태의 값이라면, 자료항목이 오류일 때 $A_{ki} = 1$로 그리고 자료항목이 오류가 아닐 때 $A_{ki} = 0$으로 표현할 수 있다. A_{ki}는 k=1,2, ..., N_i는 *iid*이고, 자료군 i에 누적된 오류인 자료항목의 수 $X_i(t)$와 각 A_{ki}는 서로 독립이라고 가정한다. 그리고 $A_{ki}=x$일때 오류비용 $L_i(x)$가 가 발생하며, 자료군의 특성에 따라 여러 가지 손실함수(loss function) 형태를 취할 수 있다. 일반적으로 오류는 그 크기에 정확히 비례해서 비용이 수반되는 것이 아니다. 많은 경우 작은 크기의 오류에 따른 비용은 사소해서 무시되거나 낮은 비용을 수반하고, 어떤 수준 이상 크기의 오류는 시스템에 심각한 영향을 미치고 따라서 많은 비용을 유발한다. 예를 들면, 자료군 i의 각 오류의 크기가 x일 때 다음과 같은 형태를 취할 수 있다.

$$L_i(x) = \begin{cases} a_1 + a_2|x|, & |x| > M_i \geq 0 \\ b_1 + b_2|x|, & 0 < |x| \leq M_i \end{cases} \quad (7)$$

여기서 a_1, a_2, b_1, b_2는 양의 상수이고 M_i는 자료군 i의 오류비용이 갑자기 상승하는 오류크기의 한계점(threshold)이다. 이외 오류비용함수는 이차다항식(quadratic polynomial)이 될 수도 있고, 과대기록(overstatement)과 과소기록(understatement)이 서로 다른 영향을 미칠 수도 있다.

자료의 질 향상을 위한 데이타베이스의 감사는 주기적으로 이루어지며, 그때마다 고정비용 F와 자료군 i의 각 자료항목당 평균오류발견비용 d_i와 오류인 각 자료항목당 평균오류수정비용 C_i가 발생한다. 오류발견비용은 각 자료군의 모든 자료항목에 적용되는 반면 오류수정비용은 오류인 자료항목에만 적용된다. 그리고 데이타베이스 감사가 이루어지면 모든 오류자료들은 수정된다고 즉 완전감사(perfect audit)를 가정한다. 실제

로 완전감사가 이루어지지 않는다 하더라도 보다 현실적으로 데이타베이스의 감사 시에 발견되지 않거나, 발견되어도 수정되지 않은 오류 자료항목은 오류로 취급하지 않는다고 가정하는 것과 같다. 또한 오류수정에 소요되는 비용이 너무 커서 오류비용을 초과할 경우도 그 자료항목은 그대로 두고 오류로 취급하지 않는다. 즉 데이타베이스의 감사에서도 수정되지 않는 자료항목은 그냥 오류가 없는 자료항목이라고 가정한다는 것이다.

4. 최적감사시점의 유도

본 절에서는 최적 데이타베이스 감사시점을 결정하기 위해서 단위시간당 기대총비용(즉, 오류로 인한 기대비용과 기대감사비용의 합)을 시간의 함수로 공식화하고, 이것을 최소화하는 감사시점을 구하고자 한다. 오류로 인한 비용의 계산에 (6)의 확률이 사용된다. 우선 시간 t에서의 자료군 i의 오류로 인한 기대비용은 조건부에 의한 기대치의 계산 방법에 의해 (예를 들면, [Ross 1981] pp.81-85 참조) 다음과 같이 표현될 수 있다:

$$E\left[\sum_{k=1}^{X_i(t)} L_i(A_{ki})\right] = E[X_i(t)] \cdot E[L_i(A_{1i})]$$

여기서 A_{1i}는 자료군 i의 첫번째 오류의 크기를 나타내고, 자료군 i의 각 오류의 크기 A_{ki}, $k=1,2 \ldots$, N_i는 iid이고, $X_i(t)$와 각 A_{ki}는 서로 독립이라고 가정했다. (6)으로부터 $E[X_i(t)] = N_iB_i(t)$라는 사실을 알 수 있다.

$E[L_i(A_{1i})] = \int_{-\infty}^{\infty} L_i(x) f_i(x)\,dx$ 이고

$L_i(x)$는 자료군에 따라 상당히 다른 형태로 나타날 수 있다. 시간 t에 데이타베이스 감사가 이루어질 때 발생하는 기대비용은 다음과 같이 표현될 수 있다:

$$F + \sum_{i=1}^{n} d_iN_i + \sum_{i=1}^{n} c_iE[X_i(t)] = F + \sum_{i=1}^{n} d_iN_i + \sum_{i=1}^{n} c_iN_iB_i(t)$$

따라서 만약 T를 본 모형의 의사결정변수인 데이타베이스 감사가 이루어지기 전까지의 시간이라고 하면 시간 T까지의 단위시간당 기대총비용은 다음과 같이 표현될 수 있다.

$$C(T) =$$

$$\frac{1}{T}\left\{\int_0^T \sum_{i=1}^{n} \left(\int_{-\infty}^{\infty} L_i(x) f_i(x)dx\right) N_iB_i(t)dt + F + \sum_{i=1}^{n} d_in_i + \sum_{i=1}^{n} c_iN_iB_i(T)\right\}$$

$$= \sum_{i=1}^{n} \frac{EL_iN_i\lambda_i}{\lambda_i+\mu_i} + \frac{1}{T}\left\{\sum_{i=1}^{n} \frac{N_i\lambda_i}{(\lambda_i+\mu_i)^2} (EL_i - c_i(\lambda_i+\mu_i))(e^{-(\lambda_i+\mu_i)T}-1) + F + \sum_{i=1}^{n} d_iN_i\right\} \quad (8)$$

여기서 편의상 EL_i는 $E[L_i(A_{1i})] = \int_{-\infty}^{\infty} L_i(x) f_i(x)\,dx$ 를 나타낸다.

최적감사시점은 단위시간당 총비용(식 (7))을 최소화하는 T이다. 이것을 찾기 위해 일반적인 방법으로 $C(T)$를 T에 대해 일차미분하면 다음과 같다.

$$\frac{dC(T)}{dT} = C'(T) =$$
$$-\frac{1}{T^2}\left\{\sum_{i=1}^{n}\frac{N_i\lambda_i}{(\lambda_i+\mu_i)^2}\right.$$
$$(EL_i - c_i(\lambda_i+\mu_i))(e^{-(\lambda_i+\mu_i)T}-1)$$
$$\left. + F + \sum_{i=1}^{n} d_iN_i\right\}$$
$$-\frac{1}{T}\left\{\sum_{i=1}^{n}\frac{N_i\lambda_i}{(\lambda_i+\mu_i)}(EL_i - c_i(\lambda_i+\mu_i))\right.$$
$$\left.(e^{-(\lambda_i+\mu_i)T})\right\} \qquad (9)$$

이 일차미분(식 (9))을 0으로 두면, 최적감사시점 T^*는 다음을 만족시킨다는 것을 알 수 있다.

$$-\frac{1}{T^2}\left\{\sum_{i=1}^{n}\frac{N_i\lambda_i}{(\lambda_i+\mu_i)^2}(EL_i - c_i(\lambda_i+\mu_i))\right.$$
$$\left.(e^{-(\lambda_i+\mu_i)T}-1) + F + \sum_{i=1}^{n} d_iN_i\right\}$$
$$-\frac{1}{T}\left\{\sum_{i=1}^{n}\frac{N_i\lambda_i}{(\lambda_i+\mu_i)}(EL_i - c_i(\lambda_i+\mu_i))\right.$$
$$\left.(e^{-(\lambda_i+\mu_i)T})\right\} = 0 \qquad (10)$$

그리고 $\frac{d^2C(T)}{dT^2} = C''(T) > 0$ 이 되기 위한 충분조건은 다음과 같다.

$$\sum_{i=1}^{n}\frac{N_i\lambda_i}{(\lambda_i+\mu_i)^2}(EL_i - c_i(\lambda_i+\mu_i))$$
$$> F + \sum_{i=1}^{n} d_iN_i > 0. \qquad (11)$$

따라서 조건 (10)을 만족시키면 방정식 (9)의 해가 $C(T)$를 극소화하는 최적감사시점(T^*)이 된다. 조건 (10)은 자료항목당 평균오류비용과 오류수정비용의 차이의 합이 고정감사비용보다 크다는 것을 의미한다. 이 조건은 당연히 만족되어야할 조건이다. 왜냐하면 만일 그렇지 않으면 감사를 하지 않고 그냥 두는 것이 장기적으로 비용면에서 유리하게 되기 때문이다. 이것은 또한 시간 T에 데이타베이스의 자료를 감사할 때 발생하는 단위시간당 총비용(식 (7))과 자료감사를 하지 않고 그냥 두었을 때의 단위시간당 평균오류비용을 비교함으로써도 알 수 있다. 우선 식 (6)으로부터 $t=\infty$때 $X_i(t)$의 한계(limiting)확률들은 다음과 같이 이항분포로 나타난다는 것을 알 수 있다.

$$P_{ki} = \binom{N_i}{k}\left(\frac{\lambda_i}{\lambda_i+\mu_i}\right)^k\left(\frac{\mu_i}{\lambda_i+\mu_i}\right)^{N_i-k} \qquad (12)$$

이 확률을 이용해서 감사를 하지 않고 그냥 두었을 때의 장기적인 단위시간당 평균오류비용을 구하면 $\sum_{i=1}^{n}\frac{EL_iN_i\lambda_i}{\lambda_i+\mu_i}$ 이 된다.

따라서 자료감사를 함으로써 비용절감을 얻을려면 두 비용의 차이인

$$\frac{1}{T}\left\{\sum_{i=1}^{n}\frac{N_i\lambda_i}{(\lambda_i+\mu_i)^2}(EL_i - c_i(\lambda_i+\mu_i))\right.$$
$$\left.(e^{-(\lambda_i+\mu_i)T}-1) + F + \sum_{i=1}^{n} d_iN_i\right\}$$ 이 0보다

작아야 한다. 이 조건으로부터도 조건 (10)이 나

오게 된다. 조건 (10)을 만족하기 위한 하나의 필요조건은 $-1 < (e^{-(\lambda_i+\mu_i)T} - 1) < 0$이므로 각 자료군에 대해 $EL_i - c_i(\lambda_i+\mu_i) > 0$ 이어야 한다는 것이다. $\lambda_i+\mu_i$는 단위시간을 조정함으로써 항상 1보다 크게 할 수 있으므로 이것은 각 자료항목의 오류비용이 오류수정비용 보다 크야한다는 것을 의미한다. 앞 절에서 언급된 어떤 자료항목의 오류수정비용이 오류비용보다 클 경우에 그 자료항목은 그대로 두고 오류가 아닌 것으로 간주한다는 가정을 뒷받침 해주는 결과이다.

식 (9)로 부터 최적감사시점 T^*를 명료하게 표현할 수는 없다. 그러나 수치적으로 충분한 근사치는 쉽게 구할 수 있다. 다음 절에서는 하나의 가상적인 상황에서 수치적인 결과를 보여줄 것이다.

5. 수치예

본 모형의 적용과 최적감사시점을 설명하기 위해 다음과 같은 간단한 가상적인 상황을 고려하자. 여기서 단위시간은 1 작업일(one working day)이고, 앞 절의 조건들을 만족시키도록 변수들의 값이 다음과 같이 주어졌다고 가정한다:

(1) 데이타베이스에 저장된 자료군들의 수는 3 즉 $n = 3$이다.

(2) 각 자료군에 들어있는 자료항목(data units)의 수는 각각 $N_1 = 100$, $N_2 = 200$, $N_3 = 150$이다.

(3) 각 자료군의 정확한 각 자료항목의 단위시간당 오류발생율은 각각 $\lambda_1 = 3$, $\lambda_2 = 4$, $\lambda_3 = 1$이다.

(4) 각 자료군의 오류인 각 자료항목의 단위시간당 오류발견 및 수정율은 각각 $\mu_1 = 1$, $\mu_2 = 1$, $\mu_3 = 0.5$이다.

(5) 각 자료군의 오류 자료항목당 오류의 크기(A_{ik})는 각각 평균 0이고 표준편차 1인 정규분포, 평균 10인 지수분포, 그리고 오류율이 0.2인 Bernoulli분포(오류이면 1, 그렇지 않으면 0)를 가진다.

(6) 각 자료군의 오류비용함수($L_i(x)$)는 식 (7)과 같은 형태이고, 자료군 1은 ($a_1 = \$0$, $a_2 = \$5$, $b_1 = \$25$, $b_2 = \$20$), 자료군 2는 ($a_1 = \$1$, $a_2 = \$0.5$, $b_1 = \$4$, $b_2 = \$2$), 자료군 3은 오류일 때 \$60의 오류비용을 가진다. 그리고 오류 크기의 한계점은 각각 $M_1 = 1$(평균 + 1 표준편차), $M_2 = 10$, 자료군 3은 boolean형태이므로 $M_3 = 1$이다.

(7) 각 자료군에서 감사시 발견된 각 오류를 수정하는 비용은 각각 $c_1 = \$0.4$, $c_2 = \$0.3$, $c_3 = \$0.5$이다.

(8) 각 자료항목의 자료항목당 오류발견비용은 각각 $d_1 = \$0.1$, $d_2 = \$0.5$, $d_3 = \$1$이다.

(9) 고정 감사비용은 \$1000 즉 $F = \$1000$이다.

우선 각 자료군의 오류 자료항목당 평균오류비용(EL_i)를 다음과 같이 계산할 수 있다:

$$EL_1 = \int_{-1}^{1}(0+5x)f_1(x)\,dx + 2\int_{1}^{\infty}(25+20x)f_1(x)\,dx \approx \$18,$$

$$EL_2 = \int_0^{10} (1+0.5x) f_2(x)\,dx$$

$$+ \int_{10}^{\infty} (4+2x) f_1(x)\,dx \approx \$14,$$

$$EL_3 = 60 \cdot (0.2) = \$12.$$

그러면 모든 $T > 0$에 대해서

$\frac{d^2C(T)}{dT^2} = C''(T) > 0$ 이 되고, 식 (9)를 수치적으로 풀면 $T^* = 40$이 되어 40일에 한 번정도씩 데이타베이스를 감사하여 모든 자료 오류들을 수정하는 것이 최적이라는 결과가 된다. 이 때의 단위시간당 평균총비용 즉 $C(T^*)$는 $4781이 되고, 데이타베이스 감사없이 그냥 두었을 때의 단위시간당 오류비용 $4915보다 $134이 적다. 식 (9)의 수치적인 풀이를 위해서는 Pascal과 같은 프로그래밍언어로 간단히 프로그램 하거나 MathCAD [4]와 같은 계산수학 소프트웨어 패키지를 사용할 수도 있다.

6. 결론

다양한 활동들을 지원하기 위해 오늘날 사회가 컴퓨터를 이용한 정보시스템에 점점 더 의존해 감에 따라 정보시스템에서 사용되는 자료에 내재해 있는 오류의 정도 및 영향에 대한 인식이 증대되고 있다. 더욱이 자료자원 및 그 관리의 분산(예를 들면, 분산 컴퓨팅/분산 데이타베이스시스템)과 같은 정보시스템의 최근 추세 때문에 자료의 무결성을 적절한 수준으로 유지하기가 더욱 어렵게 되었다. 따라서 자료의 무결성 확보의 필요성은 더욱 커지고 있다.

자료 질의 중요성은 오랫동안 인식되어 왔다. 그 결과 지난 수년동안 자료의 무결성을 확보하기 위한 효과적인 절차 및 방법들이 연구되었다. 대부분의 이들 연구들은 컴퓨터시스템에 오류가 있는 자료가 저장되는 것을 방지하는 절차들을 주로 제시하고 있다. 그러나 저장된 자료의 무결성(질)을 만족할만한 수준으로 유지하는 것은 계속해서 비용이 수반되는 지속적인 작업이다. 또한 아무리 잘 관리되는 시스템에서도 시간이 흐름에 따라 저장된 자료에 오류가 생기고 자료의 질은 떨어지게 된다. 따라서 데이타베이스에 저장된 자료들은 주기적인 감사를 통하여 오류를 발견하고 수정함으로써 적절한 수준으로 자료의 질이 유지되도록 해야한다.

이와 같은 관점에서 본 논문은 데이타베이스의 저질화과정을 시간이 흐름에 따라 확률적으로 변하는 상당히 일반적인 확률모형으로 모형화하고, 자료의 질 저하로 발생하는 비용과 자료감사 및 수정에 수반되는 비용을 고려한 최적감사시점을 결정하는 모형 및 방법을 제시하고 관련사항들에 대해 논의하였다. 본 연구에서 데이타베이스는 조직적 요인과 자료요인에 따라 오류 발생률도 다르고 각 오류로부터 발생하는 비용도 서로 다른 논리적으로 관련된 자료군들로 구성되어 있고 이들은 동시에 감사된다고 가정하였다.

본 논문에서 개발된 모형과 감사시점은 여기에 사용된 여러 가지 변수들(parameters)에 대한 정보를 얻을 수 있고 그 변수 값들은 상당히 안정적이라는 가정을 하고 있다. 따라서 사용된 변수값들의 추정치에 상당히 의존적이라고 할 수 있다. 실지로 오류 발생률과 거래처리과정에서의 오류 발견 및 수정율, 그리고 오류의 크기에 대한 확률분포 등은 각 자료군별로 무작위로 추출된 자료항목들에 대한 관찰치들을 통계 처리함으로써

추정할 수 있다. 이렇게 되면 주기적인 변수 값들의 추정작업과 그기에 따른 데이타베이스 감사시점(빈도)의 갱신(update)으로 변수 값들의 변화가 반영되어야 한다. 오류비용의 추정은 좀더 힘지만 그렇게 어렵지는 않다. 데이타베이스의 사용자들의 경험으로부터 자료군별 오류의 상대적 영향과 감사비용과의 상대적 비교를 통해서 본 모형에서 사용될 수 있는 추정치를 구할 수 있다. 출력되는 정보에 대한 자료 질의 영향을 평가하는 분석적 모형을 (예를 들면, [Ballou and Pazer 1985]) 이용할 수도 있다. 자료의 결함이 조직에 미치는 악영향을 계량화하는 연구는 별로 이루어져 있지 않으나 보다 나은 추정치를 얻기 위해서는 이와 같은 정보가 필수적이다. 앞으로 여기에 대한 연구가 많아질 것으로 기대되며 그렇게 되면 보다 나은 오류 비용에 대한 추정치를 얻을 수 있을 것이다. 데이타베이스 감사비용은 매번 감사가 수행될 때마다 그 추정치를 얻을 수 있다.

이와 같은 추정치들을 구하는 과정에는 자료의 사용자와 수집자 뿐만 아니라 데이타베이스 관리인력도 함께 참여하게 된다. 이와 같은 활동은 자료의 무결성 확보의 구체적인 문제들에 대한 지식을 가진 사람들이 참여하게 되어 그 자체로서도 조직에 도움이 된다. 이와 같은 활동을 통해서 조직의 구성원들이 자료 질의 유지 및 관리 문제를 보다 더 인식하게 되고 이 문제에 더 민감하게 되는 것이 본 연구 결과의 실행과정에서 얻을 수 있는 중요한 부수적 이익의 하나가 될 수 있다.

물론 본 연구에서 제시하는 방법이 가장 효율적이라고 할 수는 없고, 분명 데이타베이스 관리자에게 몇 가지 변수들을 추정해야 하는 짐을 지우는 것은 사실이지만, 본 논문은 데이타베이스에 저장된 자료 질의 지속적인 관리문제와 데이타베이스의 저질화 과정, 오류 비용, 감사 비용 사이의 관계를 제시하고, 다양한 오류 예방노력의 비용효과와 감사시점에 미치는 영향 등을 평가하는 데도 강력한 의사결정지원을 해줄 수 있다. 예를 들면, 오류 예방노력의 효과는 오류 발생률(λ_l)과 거래처리과정에서의 오류 발견 및 수정율(μ_l)로 나타나며, 오류 예방노력들은 서로 다른 비용을 수반한다. 따라서 각 오류 예방노력별로 최적감사시점(T^*)과 그때의 단위시간당 총평균비용($C(T^*)$)을 계산하고, 각 오류 예방노력의 단위시간당 평균비용과 $C(T^*)$의 합을 최소화하는 오류예방노력이 가장 경제적인 오류 예방노력이 된다는 사실도 알 수 있다. 또한 본 논문에서는 오류 비용함수, 오류 크기의 확률분포 등을 특정한 형태로 고정하지 않고 일반화하고, 오류 누적과정을 단순증가가 아닌 증감할 수 있도록 모형화 함으로써 보다 여러 가지 구체적인 현실상황에 적용될 수 있도록 하였다.

본 연구와 관련된 추후 연구과제로는 각 자료군의 자료항목 수가 고정되어 있지 않고 증가하는 경우와 완전 감사가 아닌 서로 다른 비용을 수반하고 감사의 효과도 서로 다른 다양한 감사방법들을 모형에 포함시켜 비용효과를 비교해 보는 분석이 이루어질 수 있다. 또한 다양한 실제 데이타베이스들의 오류 특성, 오류 예방노력, 감사절차, 오류의 영향 등에 대한 실증적 연구도 있어야 할 것으로 생각된다.

〈참고문헌〉

Agmon, N. and N. Ahituv, "Assessing Data Reliability in an Information System", *Journal of Management Information Systems*, Vol.4, No.2 (1987), pp.34-44.

Ahituv, N., "A Systematic Approach Toward Assing the Value or an Information System", *MIS Quarterly*, Vol.4, No.4 (1980), pp.61-75.

Ahituv, N., M. C. Munro, and Y. Wand, "The Value of Information in Information Analysis", *Information and Management*, Vol.4, No.3 (1981), pp.143-150.

Anderson, R. B., *The Student Edition of MathCAD*, Verson 2.0, Addison-Wesley, Reading, M. A., 1989.

Andrus, R. R., "Approaches to Information Evaluation", *MSU Business Topics*, Summer (1971), pp.40-46.

Baily, R. W., *Human Error in Computer Systems*, Prentice-Hall, Englewood Cliffs, N. J., 1983.

Ballou, D. P. and H. L. Pazer, "Modeling Data and Process Quality in Multi-input, Multi-output Information Systems", *Management Science*, Vol.31, No.2 (1985), pp.150-162.

Ballou, D. P. and G. K. Tayi, "Methodology for Allocating Resources for Data Quality Enhancement", *Communications of the ACM*, Vol.32, No.3 (1989), pp.320-329.

Bodner, G., "Reliability Modeling of Internal Control Systems", *Accounting Review*, Vol.50, No.4 (1975), pp.747-757.

Boritz, J. E. and D. S. Broca, "Scheduling Internal Audit Activities", Auditing: *A Journal of Practice & Theory*, Vol.6, No.1 (1986), pp.1-19.

Brodie, M. L., "Data Quality in Information Systems", *Information and Management*, Vol.3, No.3 (1980), pp.245-258.

Cushing, B. E., "A Mathematical Approach to the Analysis and Design of Internal Control Systems", *Accounting Review*, Vol.49, No.1 (1974), pp.24-41.

Cushing, B. E. and M. B. Romney, *Accounting Information Systems and Business Organizations*, 4th Ed., Addison-Wesley, Reading, M.A.,1987.

Davis, G. B. and M. H. Olson, *Management Information Systems: Conceptual Foundation, Structure and Development*, McGraw-Hill, New York, 1985.

Feller, W., *An Introduction to Probability Theory and its Applications*, Volume 1, 3rd Ed., John Wiley & Sons, New York, 1970.

Fox, C. J., Levitin, A. V., and Redman, T. C., "The Notion of Data and Its Quality Demensions", *Information Processing and Management*, Vol.30 (1994), pp.9-19.

Ham, J., D. Losell, and W. Smieliauskas, "An Empirical Study of Error Characteristics in Accounting Populations",

Accounting Review, Vol.60, No.3 (1985), pp.387-406.

Hamlen, S. S., "A Chance Constrained Mixed Integer Programming Model for Internal Control Design", *Accounting Review*, Vol.55, No.4 (1980), pp.578-593.

Hansen, J. V., "Audit Considerations in Distributed Processing Systems", *Communications of the ACM*, Vol.26, No.8 (1983), pp.562-568.

Hughes, J. S., "Optimal Internal Audit Timing", *The Accounting Review*, Vol.12,No.1 (1977), pp.56-68.

Janson, M., "Data Quality: The Achilles Heel of End-user Computing", *OMEGA International Journal of Management Science*, Vol.16, No.5 (1988), PP.491-502.

Johnson, J. R., R. A. Leitch, and J. Neter, "Characteristics of Errors in Accounts Receivables and Inventory Audits", *Accounting Review*, Vol.56, No.2 (1981), pp.270-293.

Knight, B., "The Data Pollution Problem", *Computerworld*, Vol.28, September (1992).

Kreutzfeldt, R. W. and W. A. Wallace, "Error Characteristics in Audit Populations: Their Profile and Relationship to Environmental Factors", *Auditing: A Journal of Practice & Theory*, Vol.5, No.1 (1986), pp.20-43.

Laudon, K. C., "Data Quality and Due Process in Large Interorganizational Record Systems", *Communications of the ACM*, Vol.29, No.1 (1986), pp.4-18.

Liepens, G. E., "Sound Data Are a Sound Investment", *Quality Progress*, September 1989, pp.61-64.

Morey, R. C., "Estimating and Improving the Quality of Information in a MIS", *Communications of the ACM*, Vol.25, No.5 (1982), pp.337-342.

Morey, R. C. and D. A. Dittman, "Optimal Timing of Account Audits in Internal Control", *Management Science*, Vol.32, No.3 (1986), pp.272-282.

Nesbit, I. S., "On Thin Ice: Micros and Data Integrity", *Datamation*, Vol.31, No.21 (1985), pp.80-85.

Redman, T. C., "Improve Data Quality for Competitive Advantage", *Sloan Management Review*, Winter 1995, pp.99-107.

Ross, S. M., *Introduction to Probability Models*, Academic Press, New York, 1981.

Stratton, W. O., "The Reliability Approach to Internal Control Evaluation", *Decision Science*, Vol.12 (1981), 51-67.

Taylor, H. M. and Karlin, S., *An Introduction to Stochastic Modeling*, Academic Press, Orlando, FL, 1984.

U. S. National Bureau of Standards, *Guideline for Automatic Data Processing Risk Analysis*, FIPS Publication 65, U. S. Dept. of Commerce, Springfield, V. A., 1979.