정보주기관리 이론에 근거한 정보가치 기반문서 관리기법

임지훈**, 이칠기**, 이영중*

The Information value-based document management technique using the Information Lifecycle Management Theory

Ji-Hoon Im, Chil-Gee Lee, Young-Joong Lee

Abstract

Due to explosive expansion in R & D efforts for advancement of technological predominance by Enterprises, the volume of technical information rapidly increases and emphasize on the valuation of this information has grown ever increasingly important. Therefore the requirement for systematic management and safeguard and accumulation of these intellectual properties of the Enterprise is in very high demand. A lot of effort and research has been carried out and many on going studies in progress to try to derive the optimum solution on how to manage information retention policy, processes, execution method, and hardware to execute the information with and etc. The intent of this thesis is to recommend a way for the Enterprise on how to evaluate the valuation of the data and to suggest the method on how to manage these intellectual properties by way of using Information Lifecycle Management theory which manages data according to the business valuation of the data. The decision on valuation of data and retention cycle is based on analytic method of a nonparametric regression, experimentation was carried out by applying to Enterprise Document Management System to present the suitable retention cycle according to the valuation and variety of attribute of data.

Key Words: Information Lifecycle Management, Nonparameteric regression, Enterprise Document Management System(EDMS)

* 성균관 대학교
** 삼성전자

1. 서론

대부분의 기술 산업 분야에서 기술정보는 신제품 개발을 위한 다양한 연구 개발 정보, 제품의 성능 향상 및 원가 절감을 위한 제조 관련 기술의 개선 결과물, 제품 생산을 위한 개발규격, 설비 및 공정 표준, 특허 및 논문 등이며 이는 데이터, 메타데이터 및 다양한 형태의 문서로 기업의 산업 활동을 통해 끊임없이 생성되고 있다. 이에 기업의 중요한 기술 자산으로서 지적 가치의 축적과 보안에 대한 새로운 인식을 통해 기술정보에 대한 보다 체계적인 관리가 요구되어, 정보의 가치에 따른 효율적인 분류 모델과 그에 적합한 주기와 관리 프로세스를 구축하기 위한 많은 연구가 이루어지고 있다. 이러한 연구는 기술정보의 발생량이 증가하고 사용범위가 넓어짐에 따라 더욱 중요하게 인식되고 있다.

기업 내 기술정보의 증가는 네트워크의 급속한 확장과 인터넷 기술의 발전 등의 급속한 IT 기반의 변화와 기업 활동의 영역 확대 등 기업 환경의 변화에 기인한 것으로, 향후 이러한 증가폭은 더욱 커질 것으로 예상되고 있다. 이와 같은 변화는 기업으로 하여금 고성능 서버 및 대용량 스토리지 시스템의 구성을 유도하였고, 그 결과 많은 기업들이 누적되는 서버와 스토리지 장비로 인한 시스템의 복잡도와 유지비용의 증가, 시스템 효율 및 안정성의 저하와 같은 문제에 직면하고 있다.

이러한 시스템적 문제와 더불어 각 기업이 우려하고 있는 또 다른 상황은 물리적으로 증가하고 있는 천문학적인 정보량에도 불구하고, 이러한 기술정보가 적절하게 통제, 관리되고 있지 못함으로 인해 기술정보의 사용자로 하여금 오히려 사용의 효율을 감소시키는 정보 과부하 (information overload) [1] 의 상태에 이를 수 있다는 것이다.

이러한 문제의 해결을 위해 각 기업은 축적되어 있는 대량의 기술정보를 '어떠한 방법으로 관리하는 것이 가장 효율적이고 경제적인가' 하는 관점에서 바라보게 되었다. 그러나 대부분의 기업이 이러한 흐름을 이해하면서도 효율적인 정보관리 프레임워크를 쉽게 구성하지 못하는 이유는 정보 가치가 기업의 복잡한 비즈니스 환경과 밀접하게 연관되어 있어 정보 분류 체계 및 평가 기준을 마련하기가 어렵고, 경쟁력 있는 정보 생성을 위한 하부기반이 제대로 구축되어 있지 않기 때문이다.

이와 같은 근본적인 문제를 해결하기 위한 연구 가운데 하나가 정보에 대한 통제 및 관리를 강화할 수 있는 전반적인 정보관리 방법인 정보주기관리 (Information Lifecycle Management, ILM) 이다 [2]. 정보주기관리는 기업이 보유한 정보를 가치화하여 다양한 형태의 스토리지 미디어에 옮기고 보관하는 과정을 정책 기반의 자동화로 구현하고자 하는 것이다.

본 논문에서는 정보주기관리 이론에 따라 정보의 가치를 평가하는 과정을 통계학 분야에서 연구되고 있는 회귀분석기법 중 비모수 회귀분석 방법을 사용하여 제시하고 이 결과를 시뮬레이션을 통한 사례 검증으로 연구결과를 기술토록 한다.

2. 관련연구 및 이론

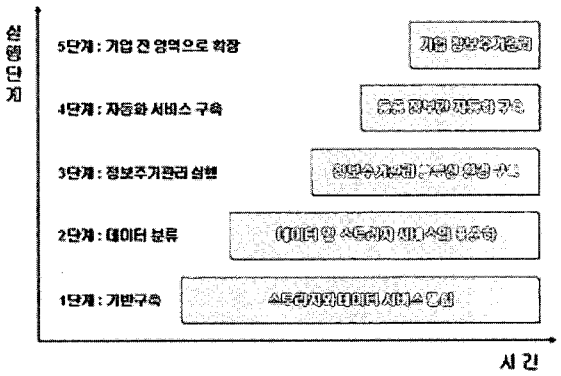
본 논문은 다음과 같은 두 가지 연구 분야, 즉 ILM 과 데이터마이닝에서 그 근원을 찾을 수 있다.

SNIA[1]의 데이터 관리 포럼 (Data Management Forum, DMF) 에 따르면 정보주기관리는 정보의 생성에서 소멸까지 가장 적절하고 효율적인 IT하부기반을 이용해 정보의 비즈니스 가치를 정렬하는 정책(Policy), 프로세스(Process), 실행(Practice), 도구(Tool) 를 의미한다 [2].

정보주기관리 환경을 구축하기 위해서는 먼

1 SNIA, the Storage Networking Industry Association, 는 스토리지 네트워킹 분야에서 보다 완벽하고 신뢰할 만한 솔루션을 제공하기 위해 스토리지(하드웨어 및 소프트웨어) 업체가 구성원으로 참여하고 있는 비영리 컨소시엄이다.

저 정보의 관리가 비즈니스의 주요 프로세스와 애플리케이션, 그리고 독창성과 밀접하게 연관되어져야 한다는 점(Business-centric)이다. 그리고 프로세스와 애플리케이션, 데이터 자원들을 포괄하는 전사적 차원의 정보 관리 정책을 기반으로 하여야하며(Policy-based), 모든 비즈니스 정보 자산들을 통합된 관점에서 관리할 수 있어야 하고(Centrally managed), 모든 종류의 플랫폼과 운영 체제 환경을 포괄해야 하며(Heterogeneous), 그리고 시간 흐름에 맞게 데이터의 가치를 비즈니스에 부합되도록 해야 하는(Aligned with the value of data) 것이다 [3]. 데이터 관리 포럼에서 효과적인 정보주기관리 환경 구축을 위해 제시하고 있는 전개 방법은 다음과 같이 다섯 단계로 구성되어 있다 [2].



<그림 1> 정보주기관리 구축 단계

<그림 1> 에서와 같이 정보주기관리를 위한 구축 단계는 1 단계에서 스토리지와 데이터 서비스의 통합을 통해 정보주기관리를 위한 기반을 마련하고 2 단계에서 데이터의 가치 평가를 위해 소규모 비즈니스 또는 기업 전체의 비즈니스를 반영하여 데이터를 분류한다. 3 단계는 가치 있는 정보와 어플리케이션을 효과적으로 지원하기 위한 환경을 구축하고 4 단계는 이에 대한 자동화 서비스를 구축하고 마지막으로 5 단계에서 기업의 전 영역으로 확산하는 과정으로 진행된다.

본 연구의 정보가치 분석에서 사용하고자 하는 회귀 (regression) 기법은 대표적인 통계 분석 기법으로 관측된 자료를 이용하여 독립변수와 종속변수간의 함수관계식을 찾아내고 변수간의 관계를 정확하게 수식적으로 밝혀내는 방법이다. 이는 모수회귀와 비모수회귀가 있으며, 비모수회귀방법은 사전에 회귀모형에 대한 일반적인 가정이나 제약 없이 시작되어지는 것으로 여러 분야의 자료에 응용이 가능하며 대표적으로 스플라인(spline) 방법과 커널(kernel) 방법이 있다. 본 연구에서는 평활회귀(smoothing regression) 의 방법을 통해 정보를 분석하고, 분석 결과를 바탕으로 정보의 가치를 평가하기 위해 선형 스플라인 회귀(linear spline regression) 를 사용하였다 [4].

3. 본론

3.1 정보가치 속성에 따른 평가

오랜 기간 다양한 채널을 통해 축적되어온 방대한 양의 기술정보를 일정한 기준을 정해 가치를 결정한다는 것은 쉬운 일이 아니다. 기술정보가 단순히 사실만을 기술한 데이터가 아닌 비즈니스를 내포하고 정보 간 상호 밀접한 연관을 가지고 관리되는 기업의 지적 자산이기 때문이다.

따라서 기업 내 기술정보의 효과적인 관리를 위해서는 비즈니스적인 요소를 반영한 다양한 시각을 통해 가치를 평가하기 위한 기준 정립이 선행되어야 하고, 이러한 평가 기준에 따라 정보의 관리 비용과 복잡성을 고려한 시스템적 기반이 뒷받침되어야 한다.

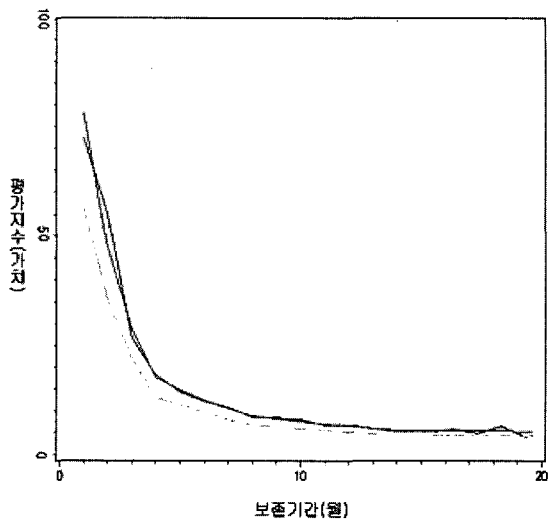
정보 가치에 영향을 미치는 속성은 기업이 가지는 비즈니스 환경에 따라 중요도 및 상호 연관성이 크게 달라지기 때문에 각 기업의 비즈니스 환경을 고려하여 적절한 항목을 도출하여야 한다. 일반적으로 정보의 가치에 영향을 주는 속성들은 다음과 같은 것들을 들 수 있다.

- **미리 정의된 가치** : 기업의 정보관리 정책, 비즈니스 민감도에 따라 이미 정의되어 있는 어플리케이션, 메타데이터, 데이터 등의 가치 등급
- **정보의 생명주기(lifecycle)와 관련된 속성** : 정보의 생성에 따라 결정된 보존 기간, 사용 범위 등의 정보의 lifecycle 과 관련된 속성
- **정보의 사용 빈도** : 정보 사용자에 의해 검색 또는 사용되는 횟수
- **사용자에 의해 요구되는 응답시간** : 정보의 성격에 따라 사용자에 의해 요구되어지는 정보 추출의 단계 및 시스템의 응답 시간
- **보존 기간** : 정보 생성 후 경과 시간
- **개정 횟수** : 정보의 개정 횟수와 원본 문서와의 연관성 정도
- **인용 빈도(Link)** : 다른 정보 내에 참조 또는 인용된 빈도

정보 가치의 평가는 가능한 많은 속성들을 고려하여 진행할수록 정확한 결과를 얻어낼 수 있지만, 정보의 구성이 단순하지 않고 정보간 상호 종속성을 이해하기가 쉽지 않아, 너무 많은 속성을 고려할 경우 자칫 분석 시간이 불필요하게 길어지고 자동화 구현이 어려워질 수 있다는 단점이 있다. 따라서 정보의 가치에 영향을 미치는 주된 속성을 적절하게 찾아내어 평가를 진행하는 것이 중요하다.

다음은 반도체 통합문서관리시스템(Enterprise Document Management System, EDMS) [5]의 일부 문서 데이터를 표본 데이터로 사용하여 보존 기간과 사용 빈도를 기준으로 개정 횟수와 참조 빈도 등의 다양한 속성을 더하여 실험한 예이다.

<그림 2> 의 그래프는 기술문서의 보존 기한과 사용빈도를 바탕으로 개정 횟수와 인용 가중치를 적용한 후 문서의 가치를 평가한 결과이다.



x: 보존기한, y: 사용빈도 고려
— x: 보존기한, y: 사용빈도, 개정횟수 고려
— x: 보존기한, y: 사용빈도, 참조빈도 고려

<그림 2> 속성에 따른 가치 평가

세 그래프는 모두 보존 기간이 증가함에 따라 전반적으로 가치가 감소하는 경향을 보이고 있지만 개정 횟수 및 인용 빈도를 분석 인자로 포함하였을 경우 보존 기한과 사용 빈도만을 고려했을 때와는 다른 감소 경향을 보임을 알 수 있다.

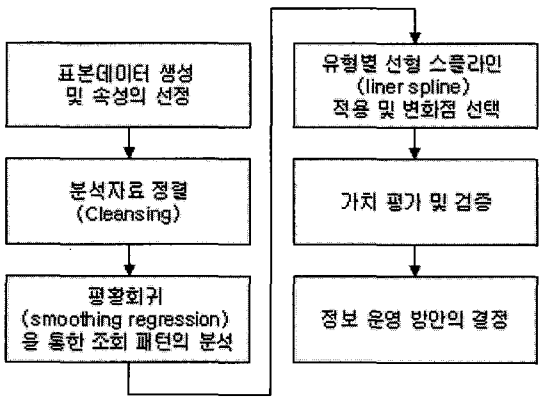
이는 가치를 평가하는 기준에 따라 결과가 달라질 수 있으며, 따라서 기업의 비즈니스 환경 또는 영향을 미칠 수 있는 요소들을 충분히 고려하지 않고 분석을 진행할 경우 정확한 연구결과를 얻어낼 수 없음을 단적으로 드러내고 있다.

이와 같은 과정을 통해 정보의 가치가 결정되면, 각각의 가치 등급에 따라 어떤 방법과 절차로 정보를 관리할 것인지를 결정한다. 이 단계에서는 가치화된 정보를 한정된 스토리지 자원에 최적으로 할당하고 운영하기 위한 환경을 구성하는 것이며, 이는 각 기업의 정보관리 구조 또는 시스템 인프라의 개선 및 재구성을 의미한다.

3.2 분석 및 구현절차

분석 절차는 반도체 통합문서관리시스템으로부터 표본 데이터를 추출하여 통계적 방법으로 데이터를 분석하고, 분석결과를 통해 정보의 가치를 결정한 후, 그 결과에 따른 가장 효율적인 정보 관리 방안을 찾는 방향으로 진행하였다. 관련 문서에 대한 자료 생성부터 최종평가를 거쳐 운영방안 결정에 이르기까지의 과정은 <그림 3> 과 같다.

이는 자료의 생성, 분석 및 구현 절차를 나타내는 것으로 먼저 반도체 통합문서관리 시스템으로부터 자료를 추출하여 표본 데이터를 생성한 후, 자료 정렬 (Cleansing) 단계를 통해 기준 정보 및 조회 정보가 잘못된 문서들을 찾아내어 잘못되거나 누락된 데이터를 삭제 또는 교정한다. 이 단계가 완료되면 평활회귀(smoothing regression) 기법을 이용하여 문서 분류별 보존 기한 및 사용 빈도를 바탕으로 정보 가치를 분석한 후, 보관 주기의 결정을 위해 각 유형별로 선형 스플라인 회귀(linear spline regression) 을 적용하여 분포를 적합(fitting) 시킨 후, 이를 바탕으로 이전 시점을 결정한다. 이 후 이 시점에 대한 검증 및 평가를 통해 이전 시점을 재조정하고 이에 적합한 운영방안의 결정이 이루어진다.



<그림 3> 분석 및 구현 절차

3.3 데이터 및 속성의 선정

실험을 위한 표본 데이터는 반도체 통합 문서관리시스템의 문서분류와 기술정보 구조를 바탕으로[5] 문서관리 정책과 비즈니스 환경을 고려하여 구성하였다. 이는 서비스 수준 정책이 다른 3개의 문서분류(분류1, 분류 2, 분류 3) 를 포함하고 있다.

문서분류 1 은 반도체 생산과 관련된 문서로 서비스가 빠르고 안정한 상태로 유지되어야 하며 장애 발생 시 2 시간 이내에 복구되어야 하는 중요 문서이다. 문서분류 2 는 연구, 개발 과정에서 발생하는 기술문서로 서비스 레벨은 중간 수준으로 유지되어야 하며 정책에 따라 24 시간 이내에 복구되어야 하며, 분류 3 은 타 시스템의 인터페이스를 통해 저장되는 문서이며 서비스 레벨은 낮으며, 48 시간 이내 복구되어야 한다.

반도체 통합문서관리시스템의 문서 가치에 영향을 미치는 주된 요소는 문서의 내용, 문서분류, 보존 기한, 사용 빈도, 개정 횟수 그리고, 문서 등급 등이 있다. 이 가운데 문서의 가치를 가장 잘 반영하고 있으며 분석을 위한 자료 수집도 용이한 생성일, 보존 기한 등의 문서 생명주기 관련 속성과 사용 빈도, 인용 빈도 등의 사용성과 관련된 속성을 선택하여 분석을 진행하였으며, 분석 과정에서 각 문서분류별로 요구되는 서비스 수준(상, 중, 하) 을 고려하여 문서의 가치를 도출하였다.

3.4 자료의 가공

선정된 표본 데이터 가운데 분석 작업을위해 필요한 정보인 문서명, 문서 생성일, 문서분류, 보안등급, 문서 적용상태, 조회 시간 등을 별도로 쿼리하여 새로운 테이블에 저장하였다.

테이블에 저장된 문서 데이터는 정렬과정을 통해 분석자료에 존재하는 결함을 조치하였다[8].

<표 1> 자료 결함 유형과 유형별 조치 방법

자료 결함 유형	조치방법
생성일 미존재	삭제
조회시간 불분명	삭제
잘못된 날짜 형식	수정
조회정보 중복	중복정보 삭제
문서 생성 전 조회 정보 존재	삭제

<표 1> 에서 분류된 것과 같이 생성일이 존재하지 않거나 조회일보다 생성일이 빠른 경우와 같이 원본 데이터를 추정할 수 없는 경우엔 분석 자료에서 제외하였으며, 날짜 형식이 잘못되었거나 조회정보가 중복된 경우[6]와 같이 수정이나 추가 조사를 통해 자료의 교정이 가능한 경우엔 형식에 맞게 날짜 형식을 교정하거나 중복된 데이터를 제거하는 방법으로 분석 자료를 교정하였다 [9].

이와 같은 과정을 통해 최종 분석 자료를 생성한 후 자료에 대한 분석은 4.3, 4.4 및 4.5와 같이 분석 방법 결정을 위한 조회 현황조사와 시점 선정을 위한 문서분류별 정보 가치 분석의 단계로 진행하였다.

3.5 조회 현황조사

조회 현황조사 과정은 조회패턴을 분석하기에 앞서 분석 방법을 결정짓기 위한 것이다. 각 문서분류별로 문서 생성 및 조회의 목적, 그리고 주된 사용자층이 다르기 때문에 문서 조회의 패턴 역시 다를 것으로 예상되었으나, 문서분류별 편차가 크지 않을 경우 분석을 위해 자료를 세분화할 필요가 없다고 판단하였다.

표본 데이터의 평균 조회율은 약 67.5% 이며, 각 문서분류별 조회율은 <표 2> 와 같다.

<표 2> 문서분류별 조회율

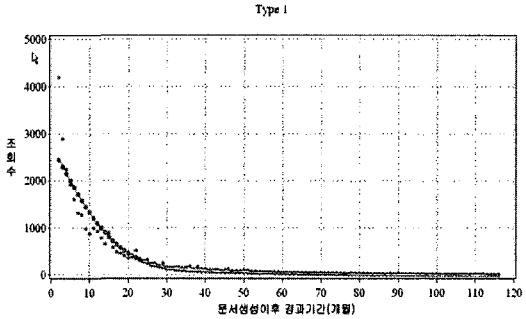
	조회율
문서분류 1	86.0 %
문서분류 2	70.6 %
문서분류 3	46.0 %

<표 2> 에서 보는 것과 같이 각 문서분류별 조회율은 46.0% ~ 100.0% 에 걸쳐 넓게 분포되어 있다. 이를 통해 문서 분류별로 조회율의 편차가 크다는 사실을 알 수 있었으며, 각 문서 분류에 따라 다른 분석 결과가 나올 것이 예상되어 각 문서분류별로 나누어 조회패턴의 분석을 진행하였다.

3.6 문서분류별 조회패턴 분석

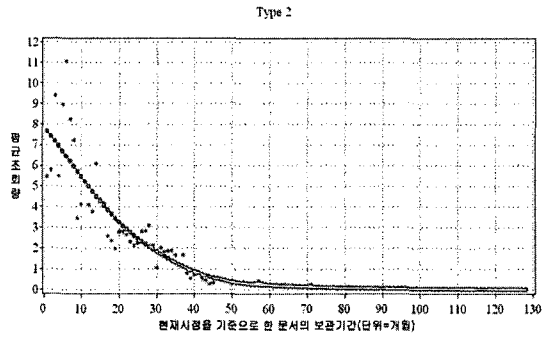
조회패턴의 분석은 두 가지 방법으로 진행하였다. 첫 번째는 생성 시점부터 현재까지의 조회 이력을 월별로 합산하여 분석하는 방법이고, 두 번째는 각 문서별 보관 기간에 따른 조회수의 평균값으로 분석하는 방법이다.

다음은 분류 1 에 대한 각각의 조회패턴 분석 결과이다.



X축 = 월 (month)
Y축 = 조회수 (sum of hits)

<그림 4> 문서분류1의 월 단위 조회현황



X 축 = 현재 시점을 기준으로 하는 문서의 보관기간
Y 축 = 평균조회량 (average of hits)

<그림 5> 분류 1의 문서나이에 따른 조회량

<그림 4> 는 문서분류 1의 월 단위 조회현황과 그에 따른 스플라인(spline) 회귀식을 그래프로 나타낸 것이고, <그림 5> 는 각 문서별 보관 기간에 따른 평균 조회수와 스플라인(spline) 회귀식을 그래프로 나타낸 것이다.
눈금이 * 표시된 선은 실제 조회수를 나타내는 것이며, • 로 표시된 선은 분산 시점을 도출하기 위한 스플라인(spline) 회귀식 적용 결과이다.

자료의 모형식은 조회 데이타가 비선형 형태를 띄고 있어, 비선형인 경우 사용하는 다항회귀식과 비모수회귀식 중 사전에 회귀함수의 모양에 제한을 두는 다항회귀식과는 달리 자료로부터 스스로 함수의 모양을 결정하여 자료를 더 많이 설명하는 스플라인 (spline) 회귀식을 선택하였다[4].

스플라인 회귀식의 비모수 회귀 방정식의 하나의 방법으로 스플라인 함수 s(t)를 이용하여 차수(r)는 3, 윈도우 크기(window size)(K)는 5로 추정하였다. 차수는 [4] 의 연구결과에 따라 3차 함수로 결정하였으며, 윈도우의 크기는 1 부터 10까지 각각의 크기에 대해 simulation을 수행한 후 이 중 분포의 설명력(R-square) 이 가장 높은 크기인 5로 선택하였다.

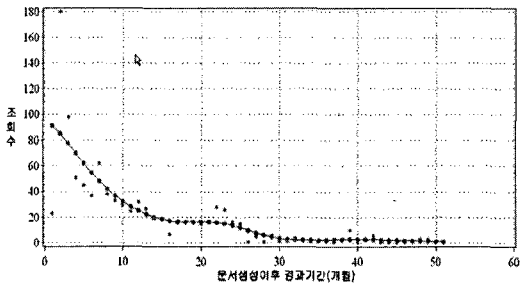
$$s(t) = \sum_{i=0}^{r-1} \theta_i t^i + \sum_{i=1}^{k} \delta_i (t-\theta)_+^{r-1}$$

여기서,

$\Theta_0,\ldots,\Theta_{r-1},\delta_1, \ldots,\delta_k$ 은 실계수이고,

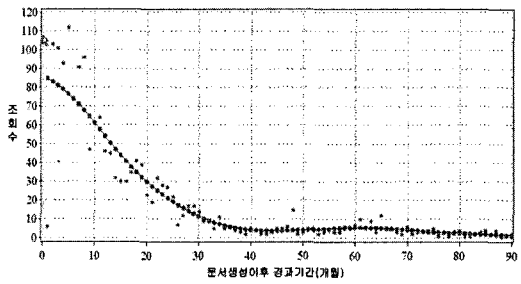
$$(t-\theta)_+ = \begin{cases} t-\theta & (x > \theta\text{일 경우}) \\ 0 & (x \le \theta\text{일 경우}) \end{cases}$$

이다.



<그림 6> 분류 2 의 월 단위 조회량

<그림 6> 은 조회정보를 월 단위로 합산하여 표시한 그래프로 시간의 경과에 따른 조회율의 분포와 문서 조회패턴의 변화추이를 판단하기가 용이하다.



<그림 7> 분류 3 의 월 단위 조회량

<그림 7> 의 그래프는 문서 생성시점부터 현재까지의 (본 논문에서는 2005 년 5월을 기준으로 함) 월 단위 기간을 (이하 문서나이) 바탕으로 평균 조회량을 나타낸 것이다.

평균 조회량(Ac)과 문서의 총 조회량(Tc) 및 문서나이(Ad) 의 관계는 다음과 같다.

$$Y = Ac = \frac{Tc}{Ad}$$

문서나이에 따른 조회패턴은 각 문서별 보관기간에 따른 참조 현황을 파악하는데 용이하지만, 시간의 경과에 따른 조회 분포를 반영하고 있지 않아 분석데이터로 사용하는데 적합하지 않다고 판단하였다.

또한 통계적으로 회귀모형의 유의성을 판단하는 설명력(R-square) 값도 전자의 모형식인 조회 수를 설명변수로 하였을 경우는0.90241, 문서나이를 설명변수(X축)로 하였을 경우는 0.8371로 전자의 모형식이 더 유의하다고 할 수 있다.

따라서 본 논문의 진행은 전자의 모형식을 통해 진행하되, 후자의 모형식은 문서의 보관주기가 결정된 후 보관주기를 적용하였을 경우의 문서나이별 조회 현황을 예측하기 위한 기초 데이터로 활용하도록 하였다.

3.5 의 결론에 따라 모든 분석은 문서분류별로 세분하여 진행하였다. 아래에서 보는 것과 같이 전반적으로 시간이 경과함에 따라 조회량이 감소하는 추세이며, 각각의 문서분류에 따른 다른 유형을 보이고 있음을 알 수 있다.

3.7 문서 가치의 평가

문서에 대한 가치는 이전 조회량을 기초로 조회량이 급속히 떨어지는 변화점을 추정하여 평가하였다.

변화점을 추정하는 방법은 조회량 데이터에 대한 선형 스플라인(spline) 회귀식을 이용하여 조회량의 기울기가 변화되는 시점을 회귀모형 선택 방법으로 반복 실행하였다. 스플라인(spline) 회귀방법은 전체 분포를 패턴이 유사한 구간으로 나누어 패턴의 형태를 추정하여, 기울기의 변화점을 자동적으로 찾을 수 있는 장점이 있어 이를 조회량의 변화점을 찾는 방법에 이용하였다. 회귀 모형 선택은 일반화 교차 타당성 (Generalized Cross Validation, GCV), 슈바르츠 베이즈 기준 (Schwarz's Bayesian Criterion, SBC) 통계량을 이용하여 최적모형 선택을 통해 변화점을 추정하였다.

본 논문에서 변화점을 검출하기 위하여 사용한 모형을 식으로 나타내면 다음 식과 같다.

$$y = \alpha_0 + \alpha_1 x + \beta_1 (x - t_1)_+ + \cdots + \beta_K (x - t_K)_+ + \varepsilon$$

여기서,

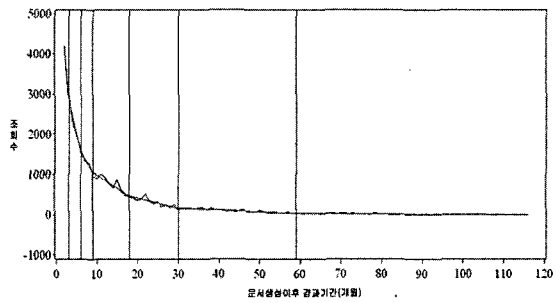
y = 조회수,

x = 월,

$$(x-t)_+ = \begin{cases} x-t & (x > t\text{일경우}) \\ 0 & (x \le t\text{일경우}) \end{cases}$$

$t_1,\ldots,t_k$ = 변화점 (월),

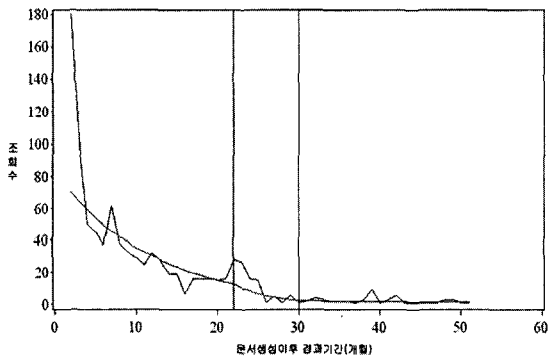
위의 모형에서 α_1 은 변화점 t_1 이전의 기울기를, β_1 는 변화점 t_1 이후의 기울기의 변화를 의미한다. 기울기가 동일한 각 구간에서는 직선의 형태를 갖고, 기울기가 변화는 시점(t)에서 연속이 되도록 추정한 결과는 다음과 같다.



<그림 8> 분류 1 의 추정 결과

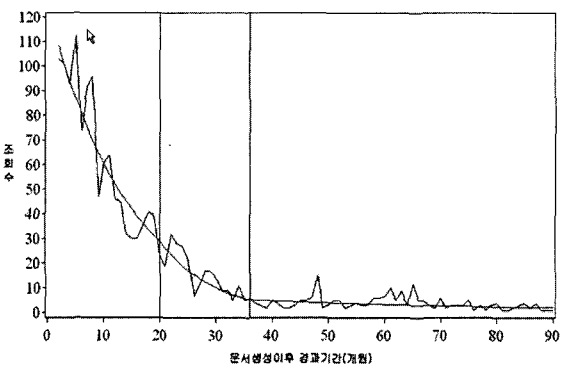
분류 1의 변화점은 3, 6, 9.5, 18, 30 및 59.5 개월의 지점에서 발생하였다. 문서의 중요도가 높은 점을 고려하여 사용빈도가 극히 낮은 문서만을 포함하는 59.5 개월의 변화점을 선택하였다.

분류 2의 변화점은 3, 15, 21 및 31 개월의 지점에서 발생하였다. 사용빈도가 전반적으로 낮은 수준에서 평준화된 31 개월의 지점을 선택하였다.



<그림 9> 분류 2의 추정 결과

분류 3의 변화점은 20 및 36 개월 지점에서 발생하였으며, 문서분류 3 은 중요도가 낮은 점을 고려하여 가장 좌측의 지점인 20 개월의 지점을 선택하였다.



<그림 10> 분류 3의 추정 결과

위의 결과를 적용하였을 경우 표준 데이터의 분류별 문서비율은 다음과 같다.

<표 3> 변화점에 따라 결정된 기간 및 문서비율(%)

구분	기준	문서비율
문서분류 1	30 개월 이전 문서	27%
	30 개월 이후 문서	73%
문서분류 2	21 개월 이전 문서	27%
	21 개월 이후 문서	73%
문서분류 3	20 개월 이전 문서	42%
	20 개월 이후 문서	58%

위와 같은 과정을 통해 결정된 시점은 문서의 가치를 판단할 척도로서 활용된다. 즉 제시된 기간 내의 문서는 활용도가 높은 문서이고 그 이후의 문서는 활용도가 상대적으로 낮은 문서임을 의미한다.

3.8 운영 방안의 결정

<표 3>의 각 문서분류별 기간에 따라 기간 내에 해당하는 문서는 고성능의 스토리지에 저장하여 신속한 서비스를 보장하고, 비즈니스 가치가 현저하게 낮아진 기간 외 데이터는 고성능 스토리지로부터 분리하여 별도의 저비용, 고용량의 스토리지 시스템에 저장함으로써 전체 스토리지 유지 비용을 줄이고 서비스 수준은 일정하게 유지할 수 있도록 구성한다.

즉, 가치가 높다고 판단된 문서는 소형 컴퓨터 시스템 인터페이스(Small Computer System Interface, SCSI) [10] 또는 파이버 채널(Fiber Channel) [11] 등을 통해 구성된 고성능의 스토리지(1차 스토리지)에 저장하고, 그렇지 않은 문서는 저렴한 직렬 에이티에이(Serial ATA) [12] 등의 기술을 이용한 저비용 고용량의 디스크 서브 스토리지(2차 스토리지)에

저장하는 것이다.

1차 스토리지의 전 문서를 대상으로 주기적으로 문서별 생명주기를 확인하고, 확인 결과를 바탕으로 스토리지간 문서의 이동을 관할하는 프로그램을 구현하여 모든과정을 자동화하여 구현한다.

3.9 평가의 검증

스토리지 장치를 포함한 데이터와 관련된 비용은 단순히 물리적인 저장 장치만으로 산출되는 것이 아니므로 이와 같은 절차에 따라 구성된 최종 결과가 얼마나 개선될지는 보다 다양한 각도에서 바라보아야 할 필요가 있다. 위 모든 과정에 대한 평가는 스토리지 관리 비용 측면, 서비스 측면으로 나누어 진행하였다.

3.9.1 모델의 구성 및 시뮬레이션의 진행

본 부분의 평가는 기존 모델과 신규 모델에 대한 시뮬레이션 모델을 구성한 후 평균 검색속도, 문서 및 데이터 로딩 시간, 장애 시 복구 예상 시간(1차 스토리지/2차 스토리지) 의 항목에 대해 시뮬레이션을 진행하였다.

3.9.2 기존 모델

1차 스토리지를 구성하여 발생하는 모든 문서를 가치와 무관하게 1차 스토리지 내에 저장한다.

3.9.3 신규 모델

스토리지를 4.6 에서 언급한 것과 같이 고성능의 1차 스토리지와 대용량의 2차 스토리지로 분리하여 구성한다. 1차 스토리지에는 4.5 의 문서 가치 평가 결과에 따라 가치가 높은 문서만을 저장하고 그 외의 문서는 2차 스토리지에 저장한다. 스토리지간 문서의 이동은 1개월 간격의 가치 평가에 의해 이루어진다.

3.9.4 모델의 분석

모델에 대한 시뮬레이션을 진행하기 위해 사용한 문서 발생 데이타는 2005년 5월까지의 월별 평균 발생 데이터이고, 문서 조회 데이터는 2003년 7월부터 2005년 8월까지 조회 이력을 사용하였다.

<표 4> 는 반도체 통합문서관리시스템의 성능을 가늠하는 세 가지 중요 성능지표에 대한 시뮬레이션 결과이다.

평균 검색속도는 문서검색 페이지에서 문서를 검색할 때 소요되는 시간을, 문서/데이터 로딩시간은 각 문서별로 문서 속성 및 첨부문서가 로딩되기까지의 시간을 의미한다. 데이터 복구 시간은 전 년도 총 장애 복구 시간 대비 백업/아카이빙 시스템 도입 후 예상되는 총 복구 예상 시간을 비교하여 나타낸 것이다.

<표 4> 서비스 부문 개선 평가

	개선도
평균 검색속도	약 38% 개선
문서/데이터 로딩시간	약 21% 개선
총 데이터 복구 시간	약 43% 개선

<표 4> 와 같이 문서 가치별 스토리지 분화를 통해 전체적인 스토리지 성능이 향상됨으로써 서비스 부문의 개선을 기대할 수 있게 되었다.

3.9.5 스토리지 관리 비용 측면

스토리지 관리 비용은 1차 및 2차 스토리지 및 백업/아카이빙 시스템의 도입, 하드웨어 및 소프트웨어 유지 보수, 시스템의 사용 등과 관련된 모든 비용을 포함하는 것이다. 신규 시뮬레이션 모델은 지난 10년간 스토리지 관리 비용을 기초로 연간 가중치를 적용하여 구성하였다. 시뮬레이션 결과에 따라 향후 10년간 약 28% 의 비용 절감 효과가 예상된다.

또한 스토리지의 통합으로 인한 서버 공간 50% 축소, 스토리지 운용을 위한 기타 비용 6% 감소, 백업(Backup) 을 위한 테이프 사용

량 절감 등으로 총 소유비용(Tocal cost of ownership, TCO) 절감 효과가 예상된다 [13].

4. 결론

본 논문에서는 지속적으로 증가하는 정보량으로 인해 발생하는 시스템의 복잡도 향상과 시스템 유지비용의 증가 등의 문제를 해결하기 위하여 정보주기관리 이론을 바탕으로 연구를 진행하였다.

문서 조회정보를 이용하여 조회 패턴을 분석하였으며, 스플라인 회귀(spline regression) 방법을 이용하여 적절한 보관 시점을 결정하고, 이를 검증하기 위한 시뮬레이션 실험을 진행하였다. 실험 결과 이러한 접근이 시스템 관리 비용 및 서비스 관점에서 개선 효과가 있음을 알 수 있었다.

본 과정을 통한 분석 및 구현은 다음과 같이 크게 두 가지 의의를 갖는다.

첫째, 각 산업의 특성, 업무 흐름, 비즈니스 및 관리 정책 등을 반영한 정보에 대한 관리 기준을 정하여 분석을 진행함으로써 각 기업의 비즈니스 환경에 보다 적합한 정보 관리 방법을 도출하는 절차를 마련하였다.

둘째, 분석 결과의 검증을 시뮬레이션을 통해 미리 구현하여 비교해봄으로써 시스템에 가장 적합한 구현 모델을 비용없이 미리 비교해 볼 수 있었으며, 향후 시스템 운영을 위해 지속적으로 시행해야 할 분석, 적용 및 평가 절차를 마련하였다.

참고문헌

[1] Butcher, H, "Information Overload in Management and Business", Information Overload, IEE Colloquium, 1995.

[2] Storage Networking Industry Association (SNIA), "Data Manangement Forum", www.snia.org/tech_activities/dmf, 2004.

[3] David Reiner, "Information Lifecycle Management: The EMC Perspective", International Conference on Data Engineering (ICDE), 2004.

[4] T.J.Hastie and R.j.Tibshirani, "Generalized Additive Models", chapman and hall, 1990.

[5] 장현성, 반도체 산업에서의 Enterprise Document Management Architecture 구현에 관한 연구, 한국경영과학회 추계학술대회 논문지, 2001.

[6] M. A. Hernandez and S. J. Stolfo, "Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem," Data Mining and Knowledge Discovery, vol. 2, pp. 9-37, 1998.

[7] M. Beigi, M. Devarakonda, R. Jain, M. Kaplan, D. Pease, J. Rubas, U. Sharma, A. Verma, "Policies for Distributed Systems and Networks, " Sixth IEEE International Workshop, pp. 139 148, June 2005.

[8] E. Rahm and H. H. Do, "Data Cleaning: Problems and Current Approaches," IEEE Bulletin of the Technical Committee in data Engineering, vol. 23, pp. 1-11, 2000.

[9] R. Kimball, "Dealing with Dirty Data," DBMS, vol. 9,pp. 55, 1996.

[10] ANSI. Fibre Channel Protocol(FCP), X3.269:1996. In 11 West 42nd Street, 13th floor, New York, NY 10036.

[11] ANSI. SCSI-3 Architecture Model (SAM), X3.270:1996. In 11 West 42nd Street, 13th floor, New York, NY 10036.

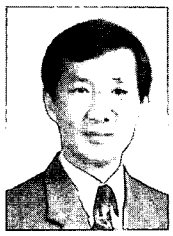
[12] R. Griswold, "Storage topologies," Computer Volume 35, Issue 12,56 63, 2002.

[13] Aziz, M.H., Ong Con Nie, Jesse Chan Mei Yam, Lee Chang Wei, "TCO Reduction," The 9th Asia-Pacific Conference on Volume 3,pp. 1147 1151, 2003.

주 작 성 자 : 임 지 훈
논문투고일 : 2005. 10. 06
논문심사일 : 2005. 10. 09(1차), 2005. 10. 10(2차),
2005. 10. 11(3차)
심사판정일 : 2005. 10. 11

저자소개

임지훈
1997년 전북대학교 전자공학과 졸업
1997년 ~ 현재 삼성전자 반도체사업부 선임 연구원
2004년 ~ 현재 성균관대학교 전자전기공학부 재학
관심분야: 컴퓨터 시뮬레이션, 공장자동화, 생산 정보시스템



이칠기
1980년 성균관대학교 전자공학과 졸업
1985년 Arizona State University 전기 및 컴퓨터 공학석사
1990년 University of Arizona 전기 및 컴퓨터 공학박사
1979년 ~ 1983년 한국 방송공사(KBS) 기술요원
1990년 ~ 1995년 삼성 정보통신 본부 컴퓨터 응용 개발실 수석 연구원
1995년 ~ 현재 성균관대학교 정보통신공학부 교수
관심분야: 컴퓨터 시뮬레이션, 객체지향모델링, 공장자동화, 전문가 시스템

이영중
1986년 인하대학교 기계공학과 졸업
1992년 Arizona State University 전기 공학석사
1996년 Arizona State University 산업 공학박사
1996년 ~ 현재 삼성전자 반도체사업부 수석 연구원
관심분야: 공장자동화, 지능시스템, 생산제어, 기업보안, ERP