보건복지 개인정보통합관제시스템의 빅데이터 도입방안을 위한 벤치마크테스트

정채용(보건복지부), 이야리(한국보건사회연구원), 한경석(숭실대학교)

차 례

■ keyword : | Big Data | Personal Information Protection | Monitoring System | Bench Mark Test

1. 서론

빅데이터 시대에서의 데이터를 이용한 정보의 가치창출은 무한생산이라고도 할 수 있다. 세계 각국의 정부와 기업들은 빅데이터가 향후 국가와 기업의 성패를 가름할 새로운 경제적 가치의 원천이 될 것으로 기대하며 다양한 부문에서 빅데이터의 적극적인 활용을 시도하고 있다[1].

빅데이터에 대한 정의를 살펴보면 매킨지(McKinsey)(2011)의 '일반적인 데이터베이스 소프트웨어가 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터', IDC(2011)에 의하면 '다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고 데이터의 초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처', 위키피디아(Wikipedia)에서는 '기존 데이터베이스 관리도구의 데이터 수집·저장·관리·분석의 역량을 넘어서는 대량의 정형 또는 비정형 데이터 세트 및 이러한 데이터로부터 가치를 추출하고 결과를 분석하는 기술', 가트너(Gartner)는 '데이터 양, 데이터 속도, 데이터 다양성의 3V(Velocity, Volume, Variety)에 복잡성(Complexity) 추가', IBM에서는 '3V(Velocity, Volume, Variety)에 데이터의 진실성(Veracity) 추가' 등으로 최근에는 빅데이터를 분석한 결과를 토대로 예측 가능한 미래에 대한 트렌드 분석까지 범위가 확대되고 있다[2][4].

기존 방식과 마찬가지로 빅데이터 환경에서도 데이터 수집, 축적, 다양한 가공이 일어나게 되며 개인 신상에 관한 데이터 역시 수집, 축적, 가공이 동반될 수밖에 없다. 동전의 양면과 같이 빅데이터의 순기능과 함께 프라이버시 위협과 같은 역기능에 대한 우려도 함께 안아야만 하는 현실적 부담감이 크게 작용한다[3]. 스마트 모바일 리터러시, 언어폭력과 스토킹, 불건전 정보의 유통 등 사회문화적 측면에서의 역기능과 바이러스 및 악성코드, SNS 피싱, 스팸, 개인정보 또는 기밀정보의 유출, 프라이버시 위협 등 기술적 측면에서의 역기능이 발생하고 있다. 이러한 위협은 개인정보 침해와 기업의 기밀정보 유출로 인한 피해뿐만 아니라, 국가 주요 인프라 시설에까지 확대되어 커다란 위협의 가능성을 일으키고 있다[5].

보건복지부와 소속·산하기관(이하 "보건복지분야"라 한다)에서 보유하고 있는 개인정보는 의료정보, 건강정보, 연금정보, 사회복지정보 등 민감한 정보가 대부분이며, 우리나라 정부기관 보유 전체 개인정보의 약 82%에 해당하는 개인정보를 보유하고 있다(2014년 1월 기준, 안전행정부 개인정보파일 등록현황). 따라서 2009년부터 보건복지부에서는 내부 사용자에 대한 개인정보 오남용 및 유출방지를 위한 관리체계를 갖추고 상시 모니터링할 수 있는 개인정보통합관제시스템을 구축하여 운영하고 있다. 그러나 보건복지 개인정보통합관제시스템(이하 '관제시스템')은 정형화된 데이터 분석을 위한 한정된 구조로 인해 다변화하는 IT업무에 대한 능동적 대처를 마련할 필요가 있다. 이에 본 연구에서는 빅데이터 관련한 기계 검색 엔진들 중 빅데이터 기계검색엔진(빅데이터 기계검색엔진)을 이용하여 실시간, 시계열 분석이 가능한 차세대 데이터 분석 방법 도입을 위한 벤치마크테스트를 수행하여 결과 및 시사점을 도출하였다.

2. 관련 연구

2.1 시스템 로그 정의

로그는 시스템의 모든 기록을 담고 있는 데이터라고 할 수 있다. 이 기록에는 시스템에 관한 성능, 오류, 경고 및 운영 정보 등과 같은 정보가 포함되며, 숫자와 기호 등으로 구성된 텍스트 파일형태이다. 이러한 로그는 전문적으로 분석한 후에만 활용할 수 있다. 웹 서버의 경우 많게는 하루에 수백 메가에서 기가 단위의 로그가 쌓이기도 하며, 이러한 방대한 양의 로그는 원하는 정보로 변환하는 일이 필요하다. 따라서 로그 데이터는 분석 행위를 필요로 한다[6]. 로그 데이터 분석을 통해 얻을 수 있는 정보로는 시스템 외부로부터의 침입 감지 및 추적, 성능관리, 장애 원인 분석, 취약점 분석, 마케팅·영업의 전략으로까지 활용될 수 있다. 즉, 시스템에서 발생하는 대부분의 문제에 대한 단서로 제공될 수 있으며, 시스템에서 발생한 보안 결함에 대한 검색이 가능하고 장애 발생 시 복구에 필요한 정보로 활용되는 등 시스템에 내포된 잠재적인 문제를 예측할 수 있다[7]. 최근 이슈화되고 있는 개인정보 침해사고 시 근거 자료로 활용될 수 있도록 각종 법규 및 지침에서 관리가 의무화되었다.

로그분석을 위해서 로그 설정방법, 파일의 저장위치, 로그에서 나타내는 정보 등이 필요하다. 또한 로그를 관리하기 위해서는 로그의 정확성, 최신성, 무결성을 확보해야 하고 이를 위해 적합한 시스템 환경을 갖추어야 한다.

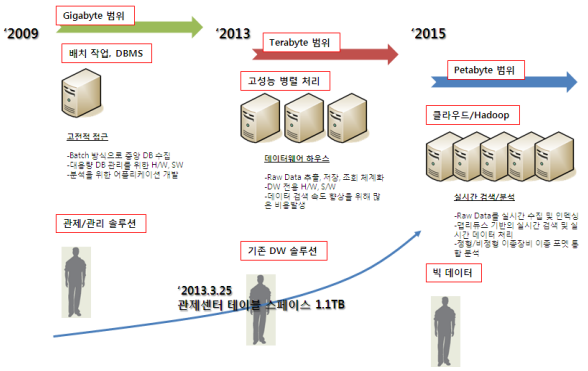
2.2 관제시스템 개요

보건복지 개인정보통합관제시스템은 보건복지부 및 소속·산하기관에서 사용하는 업무시스템, USB보안솔루션, DBMS, 메일서버, DRM솔루션, DATA WAREHOUSE 등 개인정보를 보유하고 있는 다양한 시스템을 대상으로 개인정보취급에 관한 로그를 수집하여 개인정보의 생명주기(수집, 저장, 이용, 제공, 파기)에 따른 개인정보의 체계적 관리, 오남용 및 대량노출을 예방을 위한 매일 상시적으로 관제할 수 있는 개인정보 통합관제 시스템이다. 관제시스템의 주요기능은 추출로그를 바탕으로 직원별·기관별 개인정보 오남용 분석기능과 추출된 로그를 바탕으로 오남용 의심이 되는 로그에 대하여 소명처리 기능 및 추출조건별 보고 기능 등이 있다[8].

3. 빅데이터 기계검색엔진 벤치마크테스트를 위한 요구사항 및 환경 분석

3.1 대용량 업무로그 데이터 활용을 위한 요구사항

현재 관제시스템에서는 각 대상 기관 업무로그가 즉각적으로 분석서버로의 적재가 이루어지지 않아 로그분석까지의 프로세스 단계 중 로그의 적재시간에 대부분의 시간이 소요되어, 분석이 지연되고 있어서 그림1의 분석과 같이 관제시스템에서 Terabyte급 데이터는 실시간 운영 시스템에서 데이터 추출 및 검색에 관한 심각한 문제를 발생시킬 수 있다.



▶▶ 그림 1. 데이터 증가에 따른 업무로그 데이터 관리의 예측상황

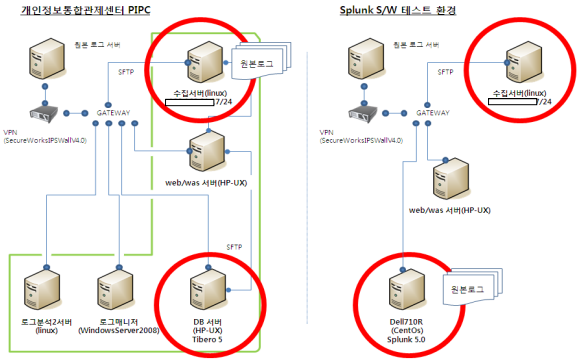
관제센터에서 관리하고 있는 업무로그 용량은 최근 Gigabyte 수준의 대용량 데이터를 뛰어 넘어 Terabyte('13.12월 기준 1.1TB 테이블스페이스)로써 담당자별로 과거 업무 로그에 대한 시계열 패턴 분석 과정이 복잡하고 어려워 정형화된 분석 패턴에 대한 구축이 필요한 상황이다. 또한 관제시스템의 정형화된 구조로는 변화되는 업무 환경에 대응할 수 있는 분석 업무나 비정형화된 데이터에 대해 빠르게 도입하여 적용하기가 쉽지 않다. 따라서 빅데이터 개념에서의 관제시스템의 업무로그 데이터를 빅데이터(Big Data)로 정의하여 빅데이터 시각으로 관제시스템에서의 기계 검색엔진을 활용하기 위한 벤치마크테스트(이하 'BMT')를 수행하였다.

3.2 BMT 환경 및 대상

본 BMT에서는 테스트를 수행하기 전에 기계 검색 엔진인 빅데이터 기계검색엔진 데모 및 세미나를 진행하였고 테스트 배경에 따른 분석과 시연을 통해 관제시스템에 대한 적합성 여부에 대해 논의하였다. 그 결과로써 하드웨어 및 운영체제, 테스트 데이터의 구성(3개 기관, 1개월 데이터 100GB를 이용한 테스트)을 위한 방안으로 표1과 같이 구성하였고 그림2의 환경 하에서 테스트하였다.

표 1. BMT 환경

연번	비교 구분		세부 내용	
1	Brand 분류		HP-UX	LINUX(Open Source)
2	장비	Brand	HP RX6600	DELL 710 R
		CPUs	Quad-Core*2	Quad-Core
		disk	300GB*8ea(SAS)	300GB*4ea(SAS)
3	운영체제(OS)		HP-UX 11.31 U ia64	CentOS(Linux) 6 x86_64
4	테스트 데이터		3개 기관의 1개월 원본로그	
5	원본 로그 크기		100GB	



▶▶ 그림 2. 관제시스템 프로세스 구성 및 빅데이터 기계검색엔진 프로세스 테스트 환경

BMT의 대상으로는 수집·적재·분석·리포트 기능에 대해 표2.와 같이 테스트하였으며, 기간은 총 4주에 걸쳐서 수행하였다.

표 2. 벤치마크 테스트 일정

연번	일정내용	일정	비고사항
1	데이터 마이닝 및 시뮬레이션	1주차	기관별 테스트 적합성 검토
2	샘플 업무로그 데이터 적용	2주차	기관별 테스트 데이터 확보
3	테스트 및 적용 가능성 분석	3주차	테스트 및 시뮬레이션
4	BMT 결과보고서 작성	4주차	-

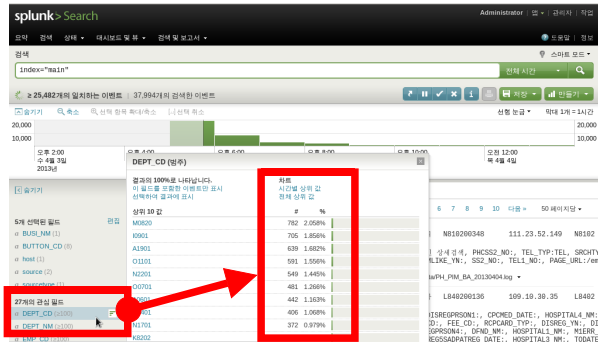
첫 번째 수집 기능 BMT에서는, 관제시스템의 각 기관에서 생성하는 데이터를 일별 Batch 업무로 수집하고 수집된 업무를 재적재하는 과정으로 수집 자료와 적재 자료의 이중 적재로 각 기관별 대량의 업무로그는 수집량이 늘어날수록 전송을 위한 시간에서 크게 낭비되고 있다. 이에 따라 빅데이터 기계검색엔진을 이용한 디렉터리(Directory) 수집 테스트를 수행하였다. 두 번째 적재기능 BMT에서는, 수집로그를 DBMS에 적재하는 과정에서 현재 관제시스템의 주된 단점으로 분석시간보다 더 많은 적재시간이 소요되고 있어서 빅데이터 기계검색엔진을 이용한 실시간 적재 테스트를 수행하였다. 세 번째 분석·리포트 기능 BMT에서는 빅데이터 기계검색엔진의 주요 기능 중 실시간 적재 및 실시간 분석 기능을 관제 업무 측면에서 테스트와 빅데이터 기계검색엔진을 이용한 실시간·시계열 분석 테스트, 빅데이터 기계검색엔진을 이용한 분석 결과 보고서 테스트, 다양한 형태로 오남용 추출 결과에 대한 사유 근거를 제시 가능성에 대해 테스트하였다.

3.3 BMT 내용

빅데이터 기계검색엔진은 그림3과 같이 인덱싱 과정이 적재 및 분석 과정에서 중요한 프로세스이다.

▶▶ 그림 3. 빅데이터 기계검색엔진 구성도

다양한 포맷의 RAW데이터를 추가적인 데이터베이스화 및 별도의 구문분석 규칙(Parse Rule)없이 본래의 데이터 포맷을 그대로 저장하고 자동 인덱싱(INDEXING)할 수 있는 빅데이터·스마트 데이터 분석도구를 활용해서 그림4와 같이 정형 및 비정형 데이터에 대한 적재와 인덱싱을 한다.



▶▶ 그림 4. 빅데이터·스마트 데이터 분석도구

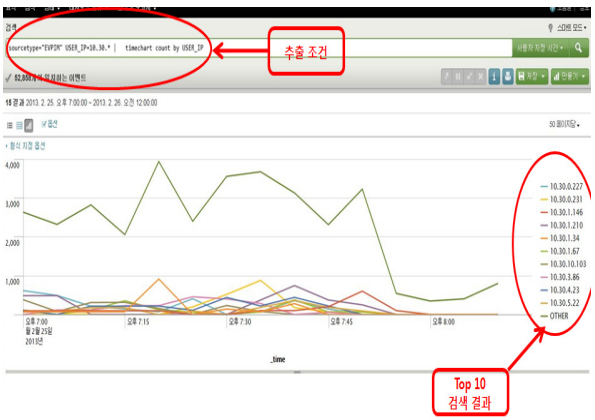
분석도구를 활용한 인덱스 특징으로는 표3.과 같이 스

키마가 별도로 요구되지 않고 쿼리와 검색이 상황에 맞게 동적으로 구성되는 점 등이 있다.

표 3. 빅데이터 기계검색엔진의 인덱스 특징

연번	구조	데이터
1	Schema 가 요구되지 않음	여러 이종의 데이터 수용 모든 종류의 Raw데이터 수용
2	데이터의 속성이 검색과 함께 정의됨	지속적인 변경을 수용
3	Queries, 검색은 상황에 맞게 동적 구성	Conversion 이나 데이터 규격에 따른 제약 조건이 없음.

수집된 다양한 데이터는 자동으로 여러 패턴이 인식되고 10:1 압축 포맷으로 저장되며 빅데이터 검색 엔진을 통해 결과를 분석하고, 알람 생성, 리포팅, 타 시스템으로 서비스 결과를 제공하게 된다(그림 4 참조). 빅데이터 기계검색엔진의 통합 인덱싱 데이터에서 장애, 변경, 성능, 메시지 등의 다양한 데이터를 실시간으로 처리하고 타 시스템의 API 혹은 Script와 연계하여 다양한 용도의 시스템과 연동 기능을 제공(2Tier 및 3Tier)한다. 지정된 적재 디렉터리에 파일 전송을 완료한 이후에는 빅데이터 기계검색엔진 에이전트(Agent)가 자동 인식하여 인덱서(Indexer) 서버로 데이터를 적재한다. 이러한 적재 과정에서도 적재된 데이터의 실시간 조회가 가능하다. 즉 관제시스템의 DBMS와 같은 상용 RDBMS는 Commit 완료 후 적재가 완료되지만 빅데이터에서는 각 적재 업무로그는 실시간으로 적재되며 시계열 데이터 분석을 수행할 수 있다. 이를 위해 그림5와 같이 검색창에 추출 조건을 입력(추출 조건 : 조회한 PC IP와 조회 및 Top 10 추출)하였고 Drill-Down 방식으로 시간대별, 분, 초 단위까지의 세부 검색을 수행하였다.



▶▶ 그림 5. 빅데이터 기계검색엔진의 실시간 분석 서비스

인덱싱 과정에서의 적재되는 데이터에 대한 실시간 조회가 가능하였고 관제시스템의 업무로그는 시간 개념적인 로그로써 시계열 분석이 가능하였다. 결과적으로 표4.와 같이 기존 DBMS에서는 업무로그 적재 이후 필요에 따라 사용 가능하도록 인덱싱되나 빅데이터 기계검색엔진은 적재 시 실시간 인덱싱을 수행하였다.

표 4. 관제센터 업무로그 적재 분석 방법과 빅데이터 기계검색엔진의 적재 분석

연번	구분	관제시스템 DBMS 분석서버	기계검색엔진 분석 시스템
1	실시간 조회	DB 서버 1차 적재, ETL 작업 시간 동안 실시간 조회 불가능	Indexer 서버에 데이터 적재, Indexing 과정에서 분석 담당자는 기계검색엔진 Header를 통해 시계열 분석 및 실시간 조회 가능
2	시계열 분석	사용불가	기관별 1일 로그 데이터 적재 시 1시간 단위의 조회량을 실시간으로 시계열 분석 처리(의심 로그에 대한 추이 분석 가능)
3	실시간 로그인덱싱	신규 추출 정책에 대한 검증 방식이 쿼리 방식으로 진행 쿼리 방식으로 데이터 검증 과정 시 많이 시간 소요 신규 쿼리 작성에 많은 시간 낭비	신규 추출 조건에 대한 검증이 시각적이며, 오남용 확인 절차 간소화
4	통계	별도 통계 구축	시계열 분석을 통한 리포트 및 통계 보고서 가능

3.4 BMT 결과

빅데이터 기계검색엔진을 이용한 관제시스템의 업무로그 데이터 활용성에 대한 분석 결과 장점과 단점을 도출하였으며 대용량 및 실시간 처리를 위한 기술 환경 도입 측면에서는 긍정적 효과를 가졌으며 비용 측면에서는 재검토를 고려할 필요가 있었다. 빅데이터 도입 시 주요 장점으로는 표5.와 같이 정형·비정형 데이터를 모두 사용할 수 있다는 점과 분석 데이터에 대해 수집 즉시 인덱싱하여 즉시 검색할 수 있고 수집 이벤트 발생 시 즉각적인 리포트 행위가 가능하다는 점이다. 그러나 주요 단점으로는 오픈소스로 인해 국내 개발 자원(개발자, 개발사)의 부재로 지원 환경이 부족하고 성능에 비해 낮은 보안성을 나타내고 있으며 원본데이터에 비례하여 별도 공간이 필요해서 비용증가를 고려해야 하는 점이다.

표 5. BMT 결과(관제센터 현황 대비 빅데이터의 장점)

내용	세부 내용	
	DBMS	빅데이터
실시간 적재성	△	○
실시간 분석 및 테스트 내역 조회	×	○
시계열 분석에 따른 분석 적중률	×	○
다양한 분석 쿼리에 대한 도표화	×	○
비정형 분석지원	×	○
다양한 형태 로그 취합 분석 가능(확장성)	△	○
분석 업무 패턴 표준화	○	○
인력이동에 따른 관제 질 저하 방지	△	○

4. 결론

본 연구에서는 다변화하는 IT환경에 대응하기 위한 관제시스템의 빅데이터 기계검색엔진 도입방안을 검토하기 위해 약 4주간 일정으로 BMT를 수행하였다. BMT 수행의 목적은 새로운 데이터 분석 방식을 도입함으로써 대상기관의 로그파일 적재시간을 단축시키고 시계열 분석을 통한 분석의 정확도를 향상시키는 것이다. 또한 분석 패턴의 표준화를 정립하고 데이터 추출 방식 및 로그 수집 대상의 확대와 같은 신규 정책 수립 시 효용성을 테스트하기 위함이었다. 기존 DBMS의 테이블스페이스는 GB(Gigabyte)를 넘어 TB(Terabyte) 수준으로써 계속적으로 증가하는 추세이며 PB(Petabyte)까지의 데이터로 증가할 경우에는 기술적인 운영 및 관리가 어려운 수준이 될 것으로 예상되는 바, 현재 관제시스템은 GB 수준의 기술로 운영 중이며 대용량 처리 및 운영을 위해서는 파티션구조로 관리하고 있어서 대용량처리는 가능하나 시계열 분석 방법은 어려운 상황이다. 따라서 빅데이터 기계검색엔진 BMT를 통한 기능 테스트를 수행하여 실시간, 시계열 분석 방법에 대해 검토하였고 그밖에도 비정형 데이터에 대한 로그 취합 및 분석 가능성과 인력이동에 따른 관제업무의 질적 저하 방지와 같은 장점을 취할 수 있음을 알게 되었다. 한편 초기 도입을 위한 높은 비용, 빅데이터 솔루션에 대한 사용 경험 부재, 낮은 보안성 등의 단점이 존재하므로 향후 계획 수립 시 신중하게 고려해야 할 필요가 있다.

참고문헌

[1] 이각범, 빅데이터를 활용한 스마트 정부 구현(안), 국가정보화전략위원회 보고서, 2011년 11월.

[2] 김동한, 빅데이터 환경에서 지능형 로그 관리 플랫폼으로 진화하는 보안 정보/이벤트 관리(SIEM) 동향, 정보통신산업진흥원 포커스, 2013년 8월.

[3] 이가원, 김용현, 정병호, 허의남, 클라우드 기반 빅데이터 플랫폼 요구사항 및 기능 분석, 한국정보과학회 제40회 정기총회 및 추계학술발표회, 2013년 11월.

[4] 김동한, 빅데이터의 핵심 플랫폼, 기업용 하둡 동향, 정보통신산업진흥원 IT 기획시리즈, 2013년 2월

[5] 윤재석, 2011 경제발전경험모듈화사업: 한국의 정보보호 활동과 시사점, 한국방송통신위원회, 한국인터넷진흥원 연구보고서, 2012년 5월

[6] 김석기, 안정용, 한정수, 웹 로그 데이터 분석 방법에 관한 연구, 응용통계연구 제14권2호, 2001년.

[7] 유기순, 임설화, 김학범, 통합로그관리시스템의 기술 동향과 발전 방향, 정보보호학회지 제23권 제6호, 2013년 12월.

[8] 보건복지부 정보화담당관, 보건복지 개인정보통합관제시스템 관리·운영가이드라인, 2012년 3월.

저자소개

● 정 채 용(Chae-Yong Jung) 정회원

- 1989년 2월 : 숭실대학교 전자계산학과(공학사)
- 2003년 2월 : 숭실대학교 정보과학대학원(공학석사)
- 2013년 3월 ~ 현재 : 숭실대학교 IT정책학과 박사과정 재학
- 1989년 : 총무처 정부전자계산소(GCC)
- 2005년 : 여성가족부 정보화전략팀장
- 2009년 : 국립의료원 홍보전산팀장
- 현재 : 보건복지부 정보화담당관

<관심분야> : 개인정보보호, 정보통신보안, IT정책관리 등

이 야 리(Ya-Ri Lee) 정회원

- 1990년 2월 : 고려대학교 전자전산공학과(공학사)
- 1999년 2월 : 동국대학교 교육대학원(컴퓨터교육학석사)
- 2002년 8월 : 동국대학교 컴퓨터공학과(공학박사)
- 2004년 3월 ~ 2012년 5월 : 동국대학교 등 강의
- 2012년 6월 ~ 현재 : 한국보건사회연구원 개인정보통합관제센터 팀장

<관심분야> : 개인정보보호, IT융합, 클라우드컴퓨팅, 모바일프로그래밍, 형식언어

한 경 석((Kyeong-Seok Han)

- 1984년 2월 : 서울대학원 경영대학원(석사)
- 1989년 : 미국퍼듀대학 MIS(박사)
  미국 휴스턴대학교 조교수 역임
- 현재: 숭실대학교 경영학부 경영정보시스템 교수

<관심분야> : e-Business, ERP, AIS, 중소기업정보화