특집 05

Social BigData 서비스 분석플랫폼 구축기술 소개

김은우 (EIC), 금득규 (한글과컴퓨터)

목 차 »

1. 서 론

기업의 데이터 증가와 데이터 속에서 가치를 찾으려는 노력은 십 수년 전부터 진행되어 왔다. 하지만 최근 스마트폰 등 모바일 환경으로의 변화와 SNS의 진화, 사물인터넷 환경의 성숙 등으로 데이터는 예상보다 급격히 증가하고 있으며, 이 속에서 가치 있는 정보를 찾아 활용하려는 움직임이 전세계적으로 진행되고 있다. 빅데이터 속에 기업이 원하는 가치가 담겨 있다는 것이다. 빅데이터 동향에 대하여 가트너는 2013년 10대 기술에 포함시킨 바 있다.

빅데이터기술은 무엇이고 어디부터 어디까지인가? 빅데이터시스템을 구축하기위해서는 무엇을 알아야 하고, 준비해야 하며, 표준은 무엇이고, 방법론은 무엇일까? 검색시스템을 위한 분산처리 환경과 NoSQL구조와 빅데이터분석플랫폼 구축을 위한 다양한 기술들이 융합하여 표현 가능한가? 현재 이것들에 대한 여러 연구가 진행되고 있으나 아직 부족한 부분이 많이 남아 있다. 우선, 빅데이터시스템을 구축하기 위해 빅데이터가 무엇인지, 어떤 핵심 원리를 포함하고 있는지 정의를 하면 다음과 같다.

빅데이터(big-data)란 우리가 생활하는 생활공간에서 생성되는 무작위의 데이터의 총 집합을 말하는 개념으로 단순히 사이즈가 큰 데이터만 의미하지 않는다. 물론 사이즈가 큰 데이터를 광의의 빅데이터로 보기도 하지만 그것 보다는 Relation Data Base(RDB)의 범주에 넣을 수 없는 비정형 데이터, 즉 일정한 포맷으로 규정 할 수 없는 데이터를 의미하는 것이 보다 광의의 의미로서의 빅데이터이다.

그렇다면 그런 빅데이터를 가지고 무엇을 할 수 있을까? 본 고에서는 그동안 필자가 연구한 내용과 실제 빅데이터시스템을 구축했던 경험을 통하여 특정 목적의 빅데이터 분석서비스시스템에 대하여 기본적인 구축기술에 대하여, 아키텍처 및 관련기술, 그리고 분석에 사용되는 R분석기법에 대한 예를 들고자 한다.

2. 분석서비스 플렛폼 구성 및 흐름 (flow)

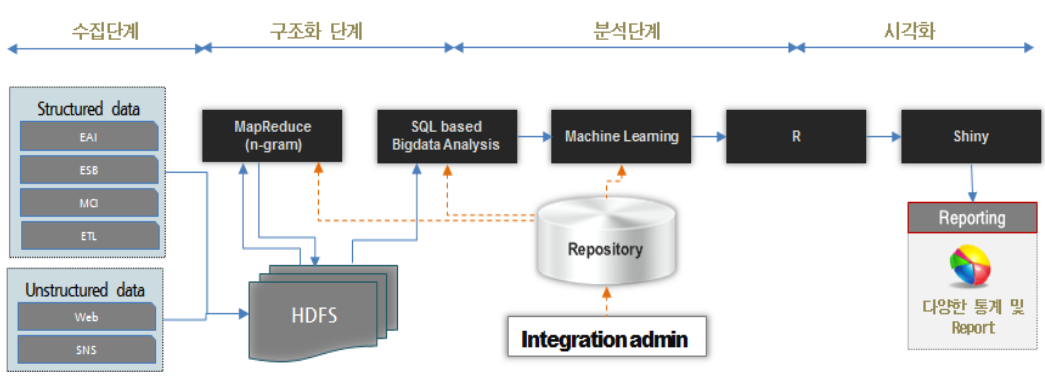
빅데이터의 성격상 데이터를 유용하게 사용하기 위해서는 기존 데이터 분석방법과는 조금 다른 패러다임으로 접근하지 않으면 유의미한 분석결과를 얻을 수 없다. 따라서 빅데이터 분석서비스 플랫폼이 구성되기 위해서는 데이터 수집단계와 수집된 데이터의 구조화 단계가 필수로 구성되어야 하는 것이 기존 정형데이터의 분석서비스와는 크게 다른 부분이다.

필자는 최근 'SocialFeel'이란 빅데이터분석서비스를 구축하였다. 이것은 소셜마이닝 알고리즘을 적용하였다. 목적은 웹/블로그 등에 게시된 수많은 글에서 특정 워드(word, 단어 또는 항목)를 이용하여 집단지성분석을 위한 것이었다.

이 시스템은 빅데이터 분석서비스 플렛폼으로서, 수집단계, 구조화 단계, 분석 단계, 시각화로 구성되어 있으며, 데이터 수집기능(연계, Crawling), 분산파일시스템 적재, n-gram, SQL기반 데이터 정렬, 분석알고리즘 및 시각화 기능을 가지고 있다. 지금부터 이 시스템에 대하여 자세히 설명하고자 한다.

BigData 분석서비스플렛폼은 다음과 같은 일련의 과정으로 수행된다.

1) EAI/ESB 등의 연계기술로 수집한 정형데이터와 Crawling을 통해 수집한 반/비정형 데이터를 분사파일시스템에 저장한다.
2) Integrator Admin에서 데이터 정제 및 분석서비스를 위한 설정을 Repository에 저장한다.
3) 구조화를 위해 분산파일 시스템에서 MapReduce를 진행 후 정제된 데이터를 분산파일시스템에 재저장 한다.
4) SQL기반 BigData 분석도구를 이용해 정제된 데이터 중 필요한 데이터를 추출하여 기계학습 도구로 데이터를 전달한다.
5) 기계학습 도구는 Repository에서 선택한 분석 알고리즘으로 데이터를 분석한다.
6) 분석된 결과는 R로 전송하여 시각화를 위한 작업을 진행한다.
7) 웹에 분석서비스 결과에 대한 화면을 표출하기 위해 R에서 Shiny[1] 서버로 결과를 전송하고, Shiny서버는 웹상에 결과를 표출한다.



(그림 1) 분석서비스 플렛폼 흐름도

1) Shiny? R 및 Java 프로그램을 활용하여 데이터를 분석한 결과를 웹으로 표현하기 위한 Web Application Framework

이제 전체 환경에 대한 대략적인 구성과 흐름을 보았다. 다음은 빅데이터시스템을 구성하는 논리적인 단계인 수집/저장, 구조화, 분석, 시각화로 나누어 세부적인 내용을 확인해 보자.

2.1 데이터 수집단계 (Data Gathering Phase)

데이터를 수집하기 위해서, 대상데이터의 형태에 따라 크롤링(Crawling), EAI/ESB 기술 등이 필요하다. 크롤러는 반/비정형 데이터(웹/문서 등)를, 정형데이터(주로 RDBMS/File)는 EAI/ESB 기술을 활용한다. 세부 프로세스는 다음과 같다.

비정형데이터 수집 (Crawling)

가) 블로그, 카페, 뉴스 등 Web페이지와 SNS에서 필요 데이터를 Crawling하여 수집한다.

나) 분산파일시스템(HDFS)에 메타데이터(Meta Data)로 저장한다.

정형데이터 수집 (EAI/ESB 또는 ETL)

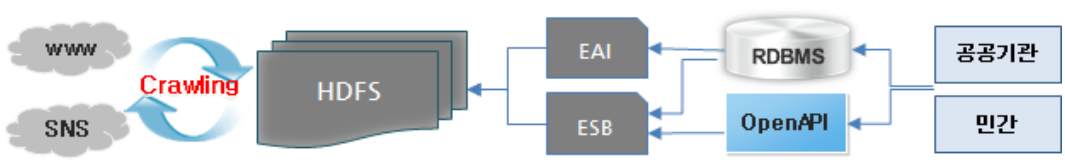
가) RDBMS에 적재된 정형데이터를 EAI/ESB(데이터연계기술)를 이용하여 분산파일시스 템에 적재한다.

나) OpenAPI를 이용하여 대상 데이터를 수집하여 분산파일시스템에 적재한다.

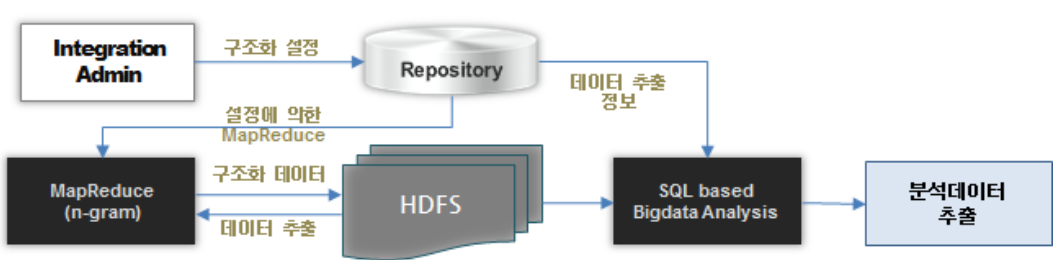
2.2 구조화 (Structuralization)

앞서 설명한 대로 일정한 포맷으로 정의 할 수 없는 빅데이터의 성격상 분석의 전처리 단계로 데이터를 구조화하는 단계가 반드시 필요하다. 이는 다양한 분석 알고리즘을 적용하기 위해서 데이터셋의 구조화 규칙을 설정한 뒤 설정된 규칙에 의거하여 빅데이터 저장소 (HDFS)에서 필요한 데이터를 추출하여 분석 데이터셋을 구성하는 절차로 진행되게 된다. 빅데이터 저장소에서 데이터를 추출하는 과정은 이미 알려진 대로 Map-Reduce를 사용하여 대용량의 데이터를 분산처리하여 시간을 줄이는 방법을 이용한다.

Integration Admin에서 구조화에 필요한 정보를 설정하여 Repository에 저장한다.



(그림 2) 데이터수집 아키텍처



(그림 3) 구조화 아키텍처

(그림 4) n-gram algorithm

분산파일시스템에서 데이터를 추출 후, 설정정보에 따라 n-gram[2] 알고리즘을 이용한 MapReduce 작업을 진행한다.

MapReduce 작업이 완료되어 생성된 구조화된 파일을 분산파일시스템에 재저장한다..

SQL기반 BigData 분석도구는 구조화된 파일에서 Integration Admin에 설정한 추출정보로 데이터를 정제하여 분석데이터를 추출한다.

n-gram은 별도의 의미 분석을 수행하지 않고도 단순한 계산 모델만으로 문장 표현들의 상대적인 사용 빈도를 효율적으로 표현할 수 있다는 장점으로 인하여 문자 인식 분야에서 인식기에 의하여 생성된 후보들의 리스트로부터 주위 문맥상 가장 자연스러운 표현을 도출하기 위하여 사용된다. 알고리즘은 도식화하면 다음과 같다.

※ n-gram 알고리즘

문자열을 입력한다.

빈칸, 마침표, 쉼표, 따옴표 등을 구분자로 모든 어절들을 추출한다.

불용어[3] 리스트를 이용하여 색인어로서 무의미한 어절들을 삭제한다.

비색인 분절을 삭제한다. 비색인 분절은 단일조사, 복합조사, 어미, 접미사 등이 결합된 다양한 형태의 음절 등을 포함한다.

나머지 색인 분절을 n-gram 들로 분할하여 색인어로 설정한다. n-gram 방법이란 인접한 N개의 음절을 말한다.

의미없는 n-gram의 생성으로 인해 질의에 부적합한 문서들이 검색될 가능성이 있으므로, 각각의 단어에 가중치를 부여한다.

2.3 데이터분석 (Data Analystic)

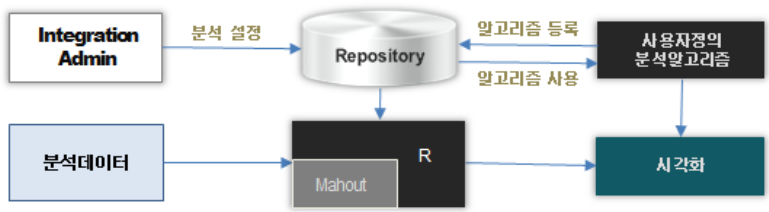
유의미한 데이터로 만들기 위해서 분석 목적에 맞는 알고리즘을 한 개 이상 선택하여 데이터 분석을 수행한다. 분석 알고리즘은 R, Mahout에서 제공하는 알고리즘과 사용자 정의 알고리즘 등으로 구성된다. 세부 프로세스는 다음과 같다

1) Integrator Admin에서 분석서비스에 대한 설정정보를 Rpeository에 저장한다.
2) Mahout에서 추천, 군집, 분류등을 이용하여 분석서비스를 수행한다
3) Mahout에서 분석서비스를 지원하지 않는 경우 R프로그래밍으로 분석서비스를 수행한다.
4) 확장성이 필요한 분석서비스의 경우 사용자정의 분석알고리즘을 Repository에 등록/사용한다
5) Mahout, R 또는 사용자 분석 알고리즘으로 분석서비스 할 데이터를 가치있는 데이터로 분석 후 시각화를 진행한다.

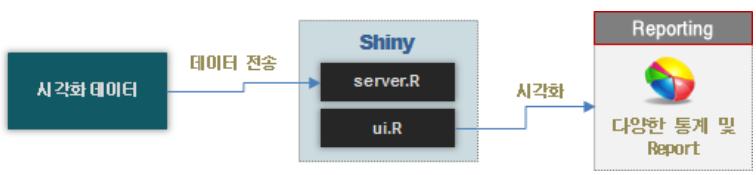
※ Mahout, R 및 사용자정의 알고리즘에 분석서비스를 수행 할 알고리즘이 존재하지 않으면 Repository에 해당 분석서비스 알고리즘을 신규 등록하고 추후 사용한다.(분석서비스 알고

2) n-gram이란? 1음절, 2음절, 3음절...n음절 등 음절단위의 색인어를 생성해 두고 검색어에 매칭시키는 방법이며 높은 재현율을 보장한다

3) '불용어(stopword)'란? 인터넷 검색 시 검색용어로 사용하지 않는 단어, 관사, 전치사, 조사, 접속사 등 검색색인 단어로 의미가 없는 단어를 말한다. 문장에서 내용을 나타내는데 큰 역할을 하지 않는 기능어이며, 검색엔진의 전처리 단계에서 의미를 가지는 어휘만을 추출하기 위해서 제외시키는 단어가 '불용어'에 해당한다.



(그림 5) 분석 아키텍처



(그림 6) 시각화

리즘 Repository화)

2.4 시각화 (Visuallization)

분석된 데이터를 도표나 그래프등으로 시각화하여 필요한 자료들을 효율적으로 찾을 수 있게 한다. 이러한 시각화를 위해서 플랫폼은 Shiny를 이용하여 표현한다. server.R은 서버로직이 ui.R은 웹화면에 표시되는 로직으로 구성되며 이를 도식화하면 다음과 같다.

1) 시각화 데이터를 서버로직을 수행하는 server.R로 전송한다
2) 통계 및 리포트등을 수행하는 ui.R을 이용해 시각화를 표현한다.
3) 웹브라우저를 통해 분석된 데이터를 확인한다.

3. 빅데이터 분석알고리즘

본 장에서는 2장에서 설명한 빅데이터서비스를 위해서 적용한 n-gram 알고리즘과 추가적으로 빅데이터시스템에서 자주 거론되는 마이닝기법(Ko)을 소개하고 어떻게 실무에 적용하는지 설명하도록 한다.

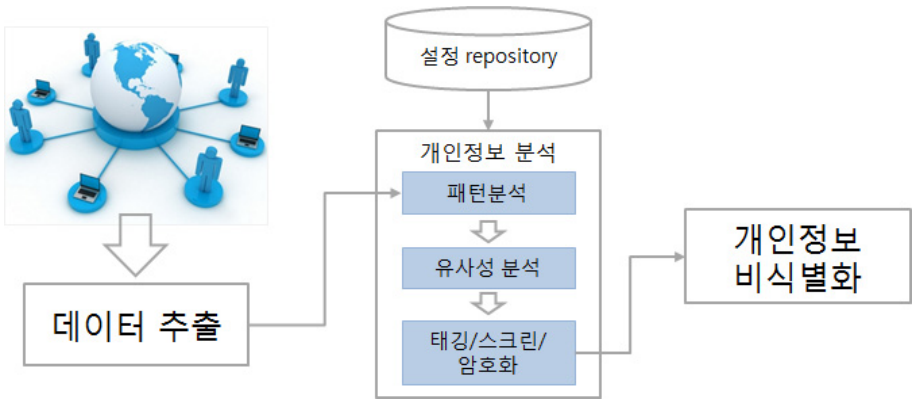
첫 번째, n-gram에 대하여 간단한 예를 들면, '잡학사전' 단어에 대한 2-gram은 '잡학', '학사', '사전' 이라는 3가지 하위 문자열들이 각각 빈도수 1로 생성하게 된다. n-gram은 서브스트링으로 나뉜 문자열을 통해 발생하는 빈도를 counting하여 데이터 분석에 사용하는 것이다

두 번째, 한글 빅데이터분석을 위해서 마이닝기법(R KoNLP패키지 사용)을 어떻게 적용하는지 실제 개발된 소프트웨어제품(개인정보 비식별화기능)[4]을 예를 들어 설명해 보겠다.

단, R패키지는 분석과정에 적용하며 이후 안정성 및 효율성이 검증되면 Mahout이나 Java환경으로 개발하게 된다.

빅데이터 분석 알고리즘에 대한 것은 현재 한글에 대하여 한글텍스트마이닝기법 분석패키지가 있으며, 그것을 이용하여 통계분석결과를 도

4) ㈜이아이컨설팅에서 개발(2013.11)한 개인정보비식별화 솔루션(PICleaner v1.0)으로 개인정보보호법 및 공공부문 대량데이터공개(정부3.0)에 따라 수집/공개되는 빅데이터(문서 및 텍스트)에서 개인정보항목을 인식/추출/필터하기 위해 한글텍스트마이닝 기법을 적용하여 제작된 소프트웨어.



출할 수 있는지에 대한 것을 다루고자 한다.

필자는 이 한글텍스트마이닝기법과 통계분석 알고리즘을 혼합하여 '개인정보비식별화' 소프트웨어를 제작하였다. 이것의 원리는 다음과 같다.

인터넷에서 수집할 수 있는 데이터는 웹 페이지 뿐만 아니라 요즘 중요성이 강조되고 있는 사회관계망서비스(SNS - 소셜) 데이터가 있다. 소셜데이터는 대부분 개인 정보가 포함되어 인터넷에 유포되고 있는 것이 보통이다. 이러한 데이터를 수집하여 활용함에 있어 가장 중요한 부분이 개인정보에 대하여 비식별화 과정이라는 정제과정을 거쳐 사용해야 한다는 점이다.

이러한 비식별화 과정은 위 그림에서 도식화한 대로 데이터 추출과정, 개인정보 분석과정, 최종적으로 개인정보 비식별화 과정으로 나뉜다.

1. 데이터 수집/추출과정

수집기(Crawler)를 이용하여 인터넷 상의 데이터를 수집하여 빅데이터 임시저장소에 저장한다.

2. 개인정보 분석 과정

임시저장소에 저장된 데이터는 설정 레포지터리에서 개인정보 처리 설정 값을 읽어와 개인정보 분석 과정을 거치게 된다. 설정데이터는 개인정보 종류 (이름, 전화번호, 이메일, 주소, 주민등록번호 등)와 비식별화 방법 (Tagging, Screening, Encryption 등) 등이 설정되어 있다.

3. 개인정보 비식별화 과정

설정된 값에 의거하여 저장된 데이터에서 패턴분석 및 유사성분석 과정을 거쳐 최종적으로 식별된 개인정보에 대하여 태깅(<Email>aaa@com.co.kr</Email> 형식), 스크리닝(a**@c**.**.kr), 암호화(tielkskfd482di!e=u) 등의 과정을 거쳐 개인정보에 대한 비식별화를 한다. 암호화 알고리즘에는 SHA, SEED, ARIA 등의 알고리즘을 사용한다.

4. 결 론

본고에서 빅데이터를 위한 모든 모듈을 설명하지는 않았다. 본고에서 다루지 않은 내용으로 보다 신뢰성 있는 빅데이터시스템으로 만들기 위해 고려해야 할 항목에 대하여 몇 가지 추가적으로 기술하여 보면 다음과 같다.

첫째 텍스트마이닝 기술과 인공지능 기술이 발전되어야 할 것이다.

- 텍스트마이닝은 마이닝마인즈 기술이 필요
- 대규모 연관분석

둘째 데이터분석이 선행되어 양질의 서비스를 도출하려는 노력이 필요할 것이다.

- 데이터분석의 한계성을 아는 것이 중요
- 고급SI시스템이라는 것을 인식

셋째 공공부문의 빅데이터시스템은 서비스핵심(=고차원적 데이터분석)은 미미한 반면, 분석을 제외한 빅데이터기술(=h/w 또는 Hadoop생태계)만 끼워 맞춰놓은 현상이다.

빅데이터분석 솔루션이 아무리 좋더라도 현재 100% 한글텍스트 마이닝을 지원할 수는 없다. 다만, 랭킹(n-gram), wordcounting 정도 가능하다.

향후 빅데이터기술의 핵심은 데이터마이닝기술에서 한걸음 진보한 마이닝마인즈가 될 가능성이 높다. 최근 CEO나 CIO는 빅데이터에 더 많은 관심을 가지며 빅데이터 기술을 바탕으로 빠른 의사 결정과 새로운 비즈니스 모델 개발에 초점을 맞추고 있다. 하지만 빅데이터 기술 특히 하둡 생태계는 범위가 넓어 하나의 솔루션이나 방법이 모든 것을 구사하기엔 무리가 있다. 또한, 관련된 한 가지 기술만을 섭렵하는 것도 쉽지 않아 보인다. 따라서 구축 후 유지운영상의 문제역시 고려하여 신속/정확한 선제 대응이 가능한 방향으로 구축하여야 할 것이다. 이를 위하여 꼭 필요한 기술만을 집적하여 효율적인 아키텍처(Simple Architecture)를 구성하는 것이 빅데이터시스템을 구축하기 위한 효과적인 방법이 될 것이다.

참 고 문 헌

[1] Hadoop (Original) ; Dec 2012.
[2] Hadoop & NoSQL ; 2013년 1월
[3] Mahout in Action ; Oct 2011.
[4] Programming Hive ; 2013년 4월
[5] BIG DATA BIG ANALYTICS ; Jan 2013.
[6] HeadFirst Data Analystic ; 2013년 4월
[7] HeadFirst Staticstics ; 2012년 4월
[8] R and DATA MINING ; Dec 2012.
[9] R과 함께하는 통계학 ; Jan 2010.
[10] R과 함께하는 분산분석 ; 2011년 3월
[11] R과 함께하는 상관 및 회귀분석 ; 2010년 2월
[12] R과 함께하는 판별분석과 로지스틱 회귀분석 ; 2013년 4월
[13] R을 이용한 데이터마이닝 ; 2011년 7월
[14] Machine Learning in Action ; 2013년 6월
[15] 집단지성 프로그래밍 ; 2010년 12월
[16] R Graphics ; 2013년 9월
[17] 한국정보화진흥원(NIA), 공공정보 개방·공유에 따른 개인정보보호지침, 2013년 9월.

저 자 약 력

김 은 우

이메일 : ew.kim@ei-consulting.kr

- 2014년 현재 ㈜이아이컨설팅 대표 (빅데이터 및 연계통합 전문 솔루션 및 컨설팅)
- 2014년 서울시청 'R을 이용한 빅데이터분석기법'강의
- 2013년 안전행정부 '빅데이터공통기반시범구축사업'
- 2013년 한국정보화진흥원(NIA) '빅데이터공유활용체계수립'
- 2013년 웹크롤러/개인정보비식별화 솔루션개발
- 2012년 KT 서비스플렛폼 통계분석시스템 컨설팅/구축
- 2010년 인천국제공항SAP-ERP(아우리)통합연계(EAI) 총괄
- 2007년 디지털예산회계(D-Brain)통합연계(EAI)구축 총괄
- 1992년 경남대학교 전산통계학과(학사)
- 관심분야 : 한글텍스트마이닝-R패키지개발, 인공지능, 마이닝마인즈, 감성ICT, 빅데이터서비스/플렛폼아키텍처

금 득 규

이메일 : dkkum@hancom.com

- 2012년~현재 한글과컴퓨터 미래전략실 수석
- 2009년~현재 동서울대학교 컴퓨터소프트웨어과 겸임교수
- 2013년 국립국어원 자문위원
- 2012년 숭실대학교 전산학과(박사)
- 2007년 한국 BPM 표준화분과위원회 위원
- 관심분야 : 서비스지향 아키텍처, 클라우드 컴퓨팅, 빅데이터 분석기술 등