**2018 Database System Project #3**

Text mining with MongoDB

컴퓨터공학과

박석 교수님

학번 : 20141556

이름 : 이성문

2018.6.7

**<목차>**

1. 프로젝트 목표
2. MongoDB 질의문 및 결과
3. RDB vs NoSQL DB에 대한 비교

# 1. 프로젝트 목표

\* 프로젝트 목표 : 본 프로젝트에서는 텍스트 마이닝 기법 중 하나인 Apriori Algorithm을 이용하여 제공된 뉴스 기사들을 분석하고 뉴스 기사에서 주로 쓰이는 단어들을 알아내는 프로그램을 작성한다. 또한 뉴스 기사 분석을 위해 비정형 데이터를 다루기 쉬운 NoSQL 기반 데이터베이스인 MongoDB를 사용함으로써 NoSQL 데이터베이스의 사용법을 익힐 뿐 아니라 관계형 데이터 베이스와 NoSQL 데이터베이스간의 차이점을 인식하는 것을 목적으로 한다. 텍스트 마이닝이란 비정형 데이터 마이닝의 유형 중 하나로 자연어 처리 기술과 문서 처리 기술을 적용하여 유용한 정보를 추출, 가공하는 것을 목적으로 하는 기술이다. 실생활에서 만들어 지는 대부분의 문서는 형태로 보관되며 제목, 저자, 출판 날짜 등과 같은 구조적인 특징들과 문서의 요약, 내용과 같은 크기가 일정하지 않은 비 구조적 요소들을 포함하기에 반 구조적 데이터로 분류된다. 응용 분야로는 Risk management, Knowledge management, Cybercrime prevention, Customer care service, Business intelligence, Spam filtering 등이 있다.

Apriori 알고리즘은 연관 규칙을 찾아 주는 알고리즘 중에서 가장 먼저 개발됐고 또 가장 많이 쓰이는 알**고리즘이다. 연관 규칙 분석이란 어떤 두 아이템 집합이 번번히 발생하는가를 알려** 주는 일련의 규칙들을 생성하는 알고리즘이다.

\* 사용환경은 다음과 같다.

서버: Host - dbpurple.sogang.ac.kr / Port - 22

운영체제: Ubuntu 14.04.5 LTS

데이터베이스: Mongodb 3.0.14

사용언어: python 2.7.6

라이브러리: pymongo, MeCab

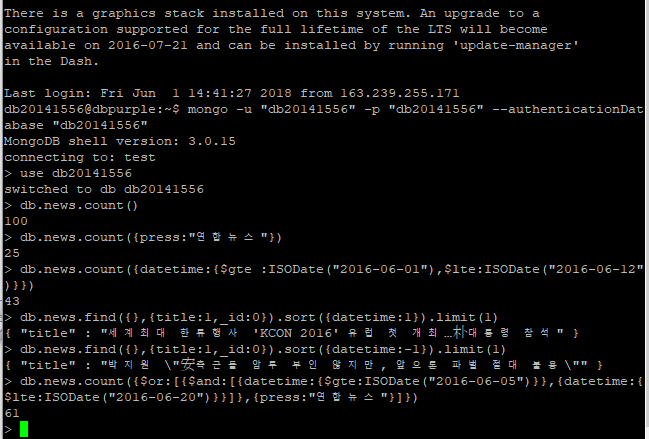
서버계정: db20141556

서버비번: db20141556

데이터베이스 계정: db20141556

데이터베이스 비번: db20141556

# 2. MongoDB 질의문 및 결과



<MongoDB 질의문 수행 장면 캡쳐>

1. **전체 기사의 개수를 구하시오**.

질의문 : db.news.count()

결과 : 100

1. **‘연합뉴스’사에 실린 기사들의 개수를 구하시오.**

질의문 : db.news.count({press:”연합뉴스”})

결과 : 25

1. **2016년 6월 1일부터 2016년 6월 12일 까지 실린 기사 개수를 구하시오.**

질의문 : db.news.count({datetime:{$gte :ISODate(“2016-06-01”),$lte:ISODate(“2016-06-12”)}})

결과 : 43

1. **전체 기사를 날짜 별로 정렬해서 가장 빠른 날짜와 늦은 날짜의 기사 제목을 출력하시오.**

->가장 빠른 날짜

질의문 : db.news.find({},{title:1,\_id:0}).sort({datetime:1}).limit(1)

결과 : { "title" : "세계최대 한류행사 'KCON 2016' 유럽 첫 개최…朴대통령 참석" }

->가장 늦은 날짜

질의문 : db.news.find({},{title:1,\_id:0}).sort({datetime:-1}).limit(1)

결과 : { "title" : "박지원 \"安측근들 암투 부인 않지만, 앞으론 파벌 절대 불용\"" }

1. **‘연합뉴스’ 사에 실린 기사 혹은 2016년 6월 5일부터 2016년 6월 20일까지 7월 한 달 동안 실린 기사의 개수를 구하시오.**

질의문 : db.news.count({$or:[{$and:[{datetime:{$gte:ISODate(“2016-06-05”)}},{datetime:{$lte:ISODate(“2016-06-20”)}}]},{press:”연합뉴스”}]})

결과 : 61

# RDB vs NoSQL DB에 대한 비교

RDB는 Relational Database로 관계형 데이터베이스이다. RDB는 Key(키)와 Value(값)들의 관계를 테이블화 시킨 간단한 원칙의 전산정보 데이터베이스이다. 즉, 모든 데이터를 Column(컬럼)과 Row(로우)를 이루는 하나 이상의 테이블(관계)로 정리하며 Primary Key(고유키)가 각 Row를 식별한다. Row는 레코드나 튜플이라고 한다. 일반적으로 각 테이블(관계)는 하나의 Entity Type)을 대표한다. Row는 그 Entity Type의 인스턴스를 나타내며 그 인스턴스의 속성이 되는 값들을 나타낸다. 테이블의 각 Row에는 각자의 Primary Key(고유키)가 있다. 한 테이블 안의 Row는 다른 테이블의 Row로 연결이 가능한데 이는 연결된 Row의 Primary Key(고유키)를 위한 Column을 추가함으로써 이루어진다. RDB에서 관계(Relation)에 대해서 살펴보자. 관계는 테이블 간에 둘다 존재한다. 이때 관계들은 일대일 , 일대다 , 다대다 이렇게 세 가지 종류가 있다. 대부분의 RDB들은 각 Row의 각 Column이 하나의 Value만을 가질 수 있도록 설계되어있다. 즉 각 Value는 원자적이다. 1970년대부터 관계형 데이터베이스의 발전은 시작되었고 현재 데이터베이스 시장을 지배하는 것으로 관련 제품으로 ORACLE, MS-SQL, MY-SQL, SYSBASE 등이 있다.

NoSQL DB를 설명하기 앞서 NoSQL이라는 단어에 대해서 살펴보니 SQL이 없다는 의미이라 생각하기 쉽지만 Not Only SQL의미로 SQL이 없으면서도 SQL뿐만 아니라 다른거도 있는 비관계형 데이터베이스를 의미한다. NoSQL DB은 기존의 RDB보다 덜 제한적인 일관성 모델을 이용하는 데이터의 저장 및 검색을 위한 구조를 제공한다. NoSQL DB의 동기에는 디자인의 단순화 , 수평적 확장성 , 세세한 통제를 포함한다. NoSQL DB는 단순 검색 및 추가 작업을 위한 매우 최적화된 Key값의 저장 공간으로 레이턴시와 스루풋과 관련하여 상당한 성능을 내는데 목적이 있다. NoSQL DB는 빅데이터와 실시간 웹 어플리케이션에서 상업적으로 널리 쓰이고 있다. 불과 2~3년 사이에 NoSQL 기술이 세계 DBMS 사용량 순위에 올랐으며 특히 NoSQL DB의 대표적인 MongoDB는 ORACLE, MYSQL,MSSQL에 이어 순위를 차지하며 새로운 데이터 기술의 등장을 알리고 있다.

 RDB와 NoSQL DB의 차이를 보겠다. 첫째로 RDB는 최초 테이블 생성 시 데이터간의 관계를 정의하지만 NoSQL DB는 최초 테이블 생성 시 데이터간의 관계를 정의하지 않는다. 따라서 RDB보다 NoSQL DB가 빠른 속도의 설계가 가능하고 빠른 속도를 낸다. 즉 RDB는 Record/Tuple-Table단위로 데이터를 다루고 NoSQL DB는 Document-Collection단위로 데이터를 다룬다. 여기서 Collection이란 문서의 모음으로 RDB에서의 Table이라 볼 수 있다. 둘째로 RDB는 복잡도와 용량의 한계가 있는데 이에 대해서 NoSQL DB는 극복하여 페타바이트 그 이상 급의 대용량 데이터를 저장하고 처리할 수 있다. 셋째로 NoSQL DB는 Scale Out을 이용해 노드 확장을 지원하므로 Scale Up방식의 RDB와 다르게 여러 일반적인 서버를 연결하여 데이터를 분산하고 저장한다. 이것을 통해서 NoSQL DB는 안전성이 보장이 된다. 더불어 NoSQL DB는 클라우드 기반 시스템에 적합하다. 넷째로 RBD는 정형데이터만 다루지만 NoSQL DB는 비정형데이터를 다룰 수 있어 현실 세계와 더욱더 관련이 깊어 현실세계의 데이터를 반영할 수 있다. 끝으로 RDB는 Fixed Schema이고 NoSQL DB는 Dynamic Schema이다. Fixed Schema는 고정된 유형과 고정된 속성에 의해 관리가 된다. Record는 테이블에 저장되며 각 테이블을 Relation의 정보로 Join , Union 등 많은 연산을 통해 관리한다. 그리고 트랜잭션으로 데이터를 관리하기 때문에 안전하게 데이터를 보관 및 저장할 수 있다. 그러나 Schema변경이 어려워 느린 속도를 가지며 확장성에 제한이 있는 문제가 있다. Dynamic Schema란 Mongo DB에 사용되는 스키마 형식이며 하나의 Collection문서가 모든 구조를 수용할 수 있다. 단 저장은 가능하나 일반적으로 데이터형태에 따라 Collection을 분리한다. Collection별로 데이터를 조회하는 것이 한 Collection에서 특정 데이터를 조회하는 것보다 훨씬 빠르며 인덱싱을 하게 되면 Document는 특정구조를 가져야 한다. 이때 인덱스는 Collection별로 정의한다. 단점으로는 Collection내에 다른 형태의 문서를 저장한다면 개발자가 관리하기 힘들다.

이번 프로젝트에 대해서는 100개의 기사를 기반으로 데이터베이스를 구축하고 데이터를 처리하는 프로젝트이다. 처리해야할 데이터가 100개의 기사이므로 수많은 비정형데이터의 집합으로 볼 수 있다. 이러한 대용량인 비정형데이터를 처리하는데 RDB보다 NoSQL DB가 더욱 적합하다. 수많은 트랜젝션을 바탕으로 priori 알고리즘을 적용해야하기에 계속해서 다른 트랜잭션을 탐색해 원하는 값을 비교하고 계산해야한다. 그러므로 적합한 것은 NoSQL DB가 적합하다.