==========

Part1-啄木鸟信息挖掘分析系统（SNAS） 手册

牛力强 2015.3.19

1. 项目描述、成员

互联网时代，各种形式的网络平台（如微博、博客、报刊、新闻门户等）的急速发展带来的日益剧增的文本数据。啄木鸟信息挖掘分析系统旨在利用计算机相关技术挖掘多种来源的用户数据（主要为文本数据），进行大规模数据的分布式存储，以及对数据的深层次分析挖掘有价值的信息来为不同的用户提供服务。

URL：121.40.193.235:2014/yun-snas-nju

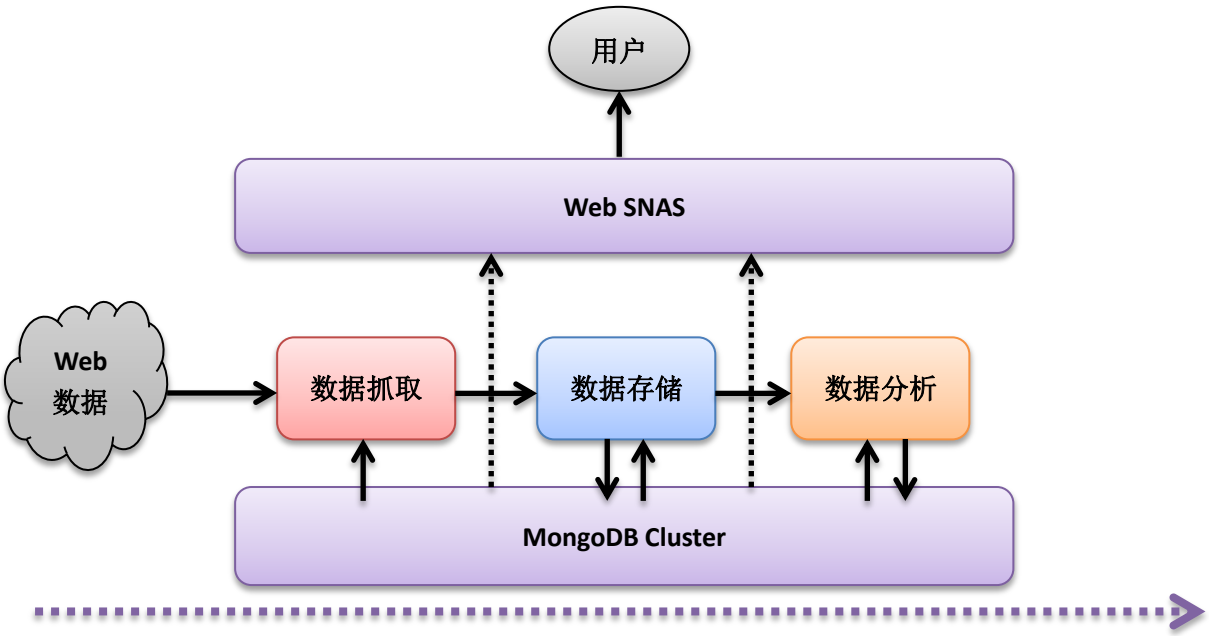
成员：戴新宇、邹远航、牛力强、程川、黄家君、郁振庭、尚迪、汤莲瑞

时间：2014.3 – 2014.11

南京大学自然语言处理研究室 nlp.nju.edu.cn

2. SNAS 系统架构

主要模块有数据抓取、数据存储、数据分析、web SNAS，如下图所示：



3. 数据抓取模块

3.1. 新浪微博

程川、郁振庭

模拟登录抓取指定关键字、指定用户的微博文本数据。
具体见 readmine 网站程川撰写文档。

3.2. 论坛、博客、媒体报刊

黄家君、汤莲瑞

模拟登录抓取指定关键字的相关文本数据。
具体见 readmine 网站黄家君撰写文档。

4. 数据存储模块

邹远航、牛力强

采用 NOSQL MongoDB Cluster。
具体见本文档第二部分“阿里云服务器 MongoDB 集群配置-主从复制”。

5. 数据分析模块

尚迪

对抓取的文本数据进行简单的统计、利用自然语言处理（NLP）以及机器学习（ML）技术进行文本情感分析、分类等。

6. web SNAS

牛力强

后端：Java wicket 框架
前端：Bootstrap 框架、CSS、JavaScript 等
具体看代码。

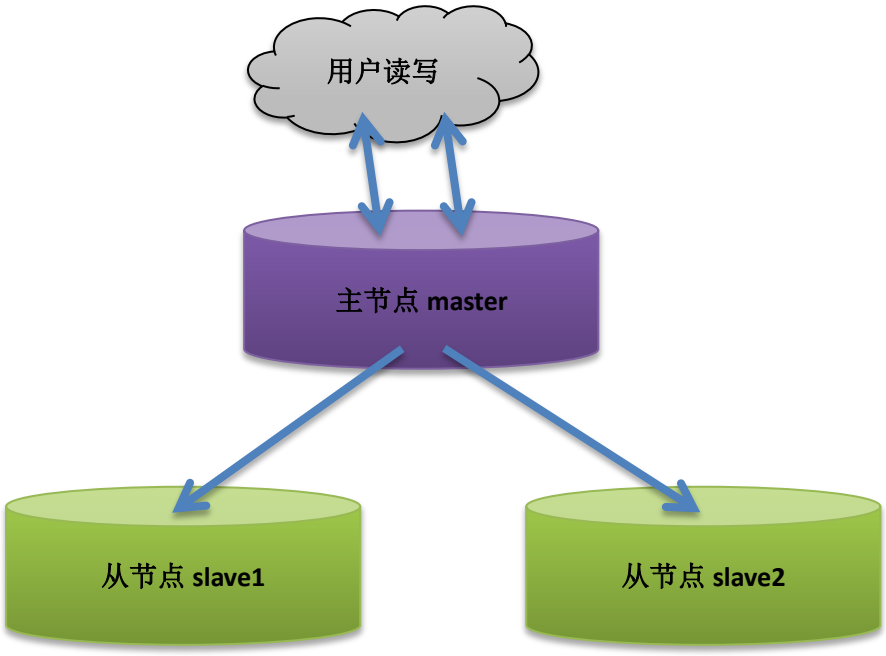
==========

Part2-阿里云服务器 MongoDB 集群配置-主从复制

2014 年 9 月 **22** 日 牛力强

1. 主从复制

主从复制是 MongoDB 最常用也是最简单的复制操作，常用于数据备份和故障修复等。其集群结构如下图所示：



说明：

1.MongoDB 集群包含有一个主节点 master 以及多个从节点 slave。

2.写入数据只能通过主节点，从节点负责从主节点拷贝数据用于备份。待主节点出现故障时从节点可以改为主节点（从节点以 master 形式重新启动即可）。

3.启动 MongoDB 服务采用 mongod（windows 下用 sc.exe 可以设置其为 windows service，随开机自动启动，服务启动 command 是 net start MongoDB，关闭服务 command 是 net stop MongoDB）（ubuntu 下启动命令为 sudo service mongod start，关闭为 sudo service mongod stop，重启为 sudo service mongod restart）。

4. 启动 MongoDB Shell 采用 mongo，格式：mongo ip:port（如 mongo 121.40.193.156:27017），即可连接到指定 ip 和 port 的 MongoDB 服务器进行 shell 操作。

2. MongoDB 集群配置

阿里云服务器资源限制，选取其中 3 台服务器作为 MongoDB 节点。具体如下：

MongoDB 节点	IP:PORT	磁盘 内存 OS	ssh 账号密码
主节点	~~121.40.193.156:27017(公网 IP)~~ 10.168.71.170:27017(内网 IP)	Disk D 100GB,RAM 1GB,OS Ubuntu 12.04 64	root 0870fb4e
从节点 1	~~121.40.193.23:27017(公网 IP)~~ 10.168.69.126:27017(内网 IP)	Disk D 100GB,RAM 1GB,OS Ubuntu 12.04 64	root 3e1b10e2
从节点 2	~~121.40.193.226:27017(公网 IP)~~ 10.168.70.15:27017(内网 IP)	Disk D 100GB,RAM 1GB,OS Ubuntu 12.04 64	root b9a8ee14

对于节点的文件目录说明：

MongoDB 配置文件默认：/etc/mongod.conf

数据存放路径：/var/lib/mongodb

日志：/var/log/mongodb/mongod.log

3. ubuntu 安装 mongoDB

http://docs.mongodb.org/manual/tutorial/install-mongodb-on-ubuntu/

4. 主从复制集群配置

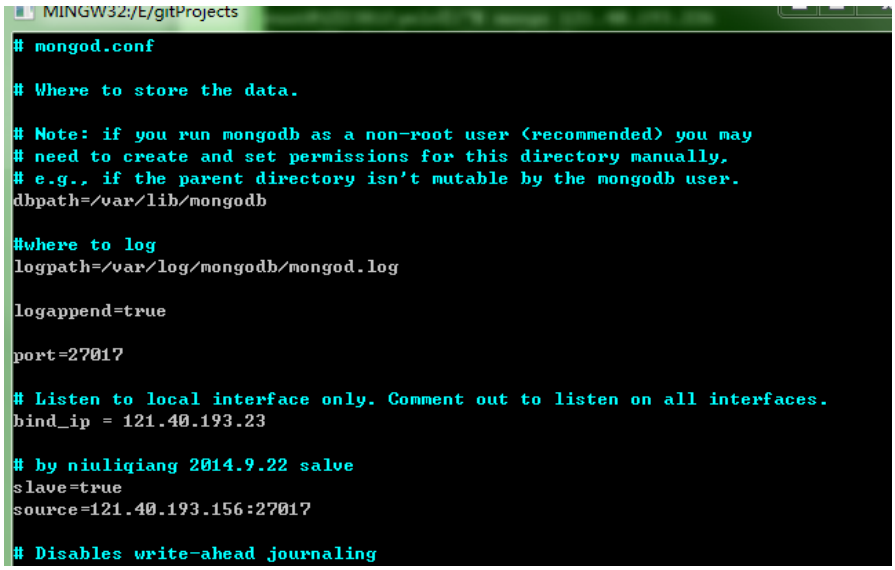
配置主节点 mongod.conf

MINGW32:/E/gitProjects

```
# mongod.conf

dbpath=/var/lib/mongodb

logpath=/var/log/mongodb/mongod.log

logappend=true

port = 27017

bind_ip = 121.40.193.156

# niuliqiang master node!
master=true
```

启动主节点：sudo service mongod start

配置从节点 mongod.conf



启动从节点：sudo service mongod start

==========

==========