“xxx”项目概要设计说明书

北京邮电大学 Pris Lab

一、概述：

项目的目标、用户等

二、项目需求：

1. 通过界面进行敏感话题关键字设置。利用关键字布控和语义分析，识别敏感话题。
2. 根据新闻出处权威度、评论数量、发言时间密集程度等参数，识别出给定时间段内的热门话题。
3. 对热点、敏感话题发展趋势进行图形化显示。
4. 对指定网站采集电信行业（包括中国电信）信息，根据信息进行分类。统计各运营行商相关的新闻数量等，新闻类别有：产品发布，人事变动等。界面中通过报表展示。
5. 。。。。。。
6.

三、系统功能：

1、
2、
3、
4、

四、平台：

1、开发环境

2、运行环境

五、系统逻辑结构图：

整个系统包含六部分：

1， 爬虫进程

该进程负责爬取。。。。。

2， 预处理进程

通过 socket 接收爬虫爬取的网页，对帖子正文和标题进行分词。结果全部写入到数据库中。

3， 。。。。。

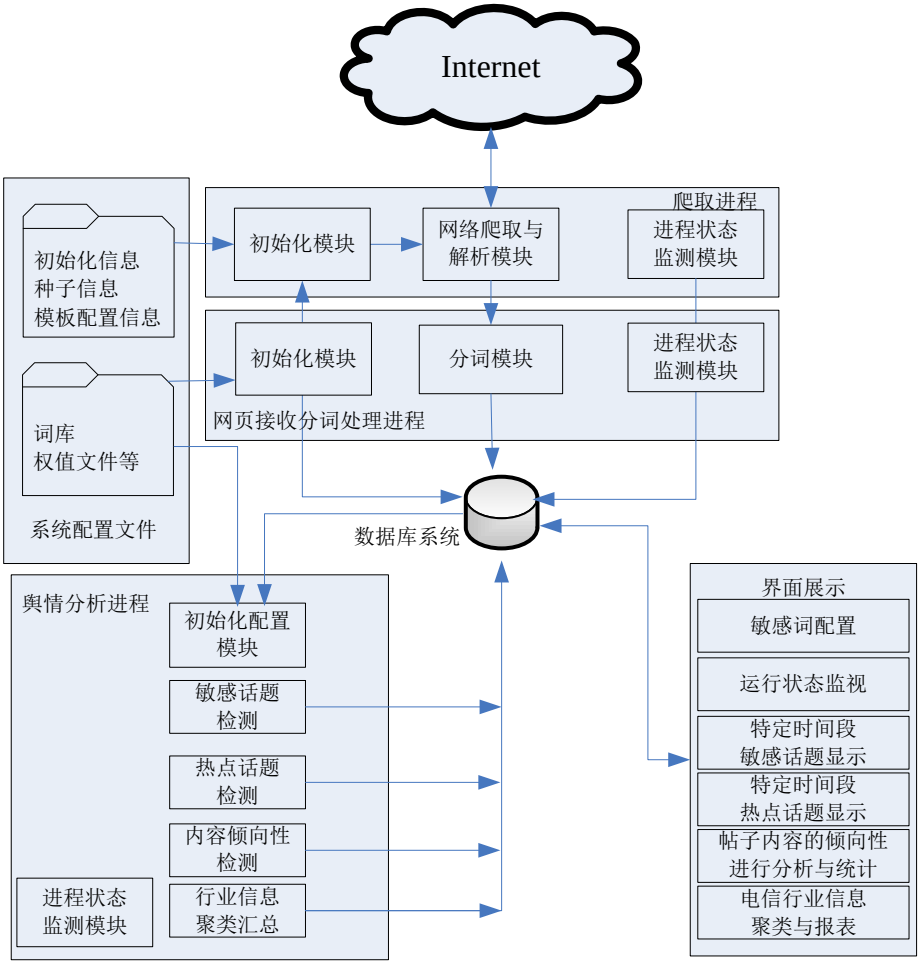
4， 界面展示模块

界面展示模块主要分为两大类，一类是配置界面，配置系统相关信息，如敏感词配置、网站权威度配置等。另一类是展示界面，主要展示系统运行状态、各分析模块分析结果等。

5， 系统配置文件和词库

系统运行需要的配置文件、词典等数据文件。

结构图如下图所示。



六、模块划分和接口设计：

1、xx 模块
功能。。。。。。
与模块 2 的接口
与模块 3 的接口

七、数据库设计

7.1.1 主贴信息表(postsinfo)

字段名	类型	注释
PID	int	文档 ID（auto_increment）
MD5	varchar(33)	根据 URL 获取 MD5
PostUrl	varchar(1000)	文档 URL
PostType	unsigned bigint	URL 的种类，1 表示新闻类（由敏感词处理模块和热点话题检测模块处理），2 表示电信行业信息相关（由行业信息处理模块和倾向性处理模块处理）
Title	varchar(200)	文档标题
Author	varchar(80)	文档作者
BoardCName	varchar(24)	版块中文名称
BoardEName	varchar(24)	版块英文名称
SiteName	varchar(24)	站点名称(例如北邮人,新浪,搜狐)
PublishTime	unsigned int	文档发表时间
ReplyTime	unsigned int	最后有人回复本文档的时间
CollectTime	unsigned int	爬虫爬取文档的时间
ClickNum	unsigned int	帖子点击数(未爬取,置为 0)
ReplyNum	unsigned int	帖子回复数
SeedID	unsigned int	种子 ID
PostContent	varchar(2000)	帖子正文
ReplyContent	varchar(2000)	回帖正文，目前为空
SegTitle	varchar(400)	标题分词结果
SegContent	varchar(4000)	正文分词结果
FreqTitle	varchar(400)	标题词频统计
FreqContent	varchar(4000)	正文词频统计
ProcessFlag	int	标记笨文档是否被处理，0 表示未处理，1 表示已处理；
主键	PID	
注释	没有成功抽取出的元数据比如点击数、回复数等都取值为 0；	

	分词结果举例：如原文档为字符串“我是中国人”，则分词后的结果为字符串“我 是 中国人”；

7.2.1 敏感帖子信息表（sen_topic_info）

字段名	类型	注释
pid	int(11)	帖子 id(外键)
sen_degree	int（11）	敏感度

1、 当某个帖子敏感度达到某个阈值被定义为敏感帖子时在该表格中插入一条记录。

2、 其中敏感文档 ID 字段“id”通过外键 PID 外链到表” postsinfo”。其中发布时间以绝对秒形式存储。

八、进度安排：

	Xx 模块	Y 模块
9.1-9.7	完 成 xxx 任 务/功能	
9.8-9.14		