# 2013 年全国大学生信息安全竞赛作品简介

作品名称：基于云计算的微博敏感信息挖掘系统

组　　长：吕若尘

组　　员：唐瞻立，李声龙，史乔茜

提交日期：2013-7-25

# 填写说明

1. 所有参赛项目必须为一个基本完整的设计。参赛作品简介旨在能够清晰准确地阐述（或图示）该参赛队的参赛项目（或方案）。

2. 参赛作品简介采用A4纸撰写。除标题外，所有内容必需为宋体、小四号字、1.5倍行距。

3. 参赛作品简介不超过6页A4纸。

4. 参赛作品简介模板里已经列的内容仅供参考，作者也可以多加内容。

# 一、摘要

伴随互联网的普及、网民社会责任意识的提高，网络舆情爆发出了不容忽视的巨大能量。在中国涉及网络舆情的媒体当中，微博是最有影响力的媒介之一，因此对微博的舆情进行有效监控刻不容缓。而现有的舆情监控系统虽然在一定程度上解决了微博舆情的发掘、追踪等问题，但普遍存在对微博敏感信息挖掘及时性差、准确率低、错误率高等几大问题。为解决现有系统存在的问题，我们团队开发了「基于云计算的微博敏感信息挖掘系统」，它是针对如今微博用户数量不断上涨以及微博舆情影响力与日俱增的现状应运而生的一款软件。该系统以突发事件网络舆情为研究背景，建立敏感事件话题语料库。并引入 PageRank 算法处理微博社交关系得到微博用户的影响力。通过运用自然语言分析等相关文本分析技术将高影响力微博用户发布的消息内容与语料库中关键词精确匹配，进而分析得到微博敏感人士，最终以响应式 Web 界面的形式将分析结果呈现给用户以便其及时查看、处理。 该系统解决了现有舆情监控系统中存在的问题，并首次在此类系统中用到基于人即以微博用户为监控对象的概念。系统利用 MongoDB 搭建了私用云平台，具有高效性、稳定性，将为微博平台提供最有效的敏感舆情监控保障。

# 二、相关工作

通过对国内外现有的这些舆情监测系统进行研究，我们发现它们主要有热点发掘、定向爬取、立场分析、话题追踪等功能，虽然在一定程度上解决了微博舆情的发掘、追踪等问题，但是仍存在以下三个主要问题：

1. 对敏感信息的挖掘不具及时性和准确性，甚至采用人工筛选的方式，导致成本增加、舆情挖掘效率低，并存在无用信息和遗漏等问题；

2. 对云平台的运用仅局限于节省硬盘空间以节约成本这一方面，而没有很好的利用云平台分配工作任务、缩短运算时间以提高系统效率；

3. 普遍采用基于文本的舆情监控方式，而将其工作方式用于微博这一以人（用户个体）为单位的平台，导致了监控对象的选取不具针对性。

基于对这三个主要问题的深入分析，我们团队针对它们解决现有系统存在的不足，开发了「基于云计算的微博敏感信息挖掘系统」。

# 三、本作品的研究内容

本系统分为表现层、服务层、业务逻辑层、数据访问层四层，系统分层架构图如图 1 所示：

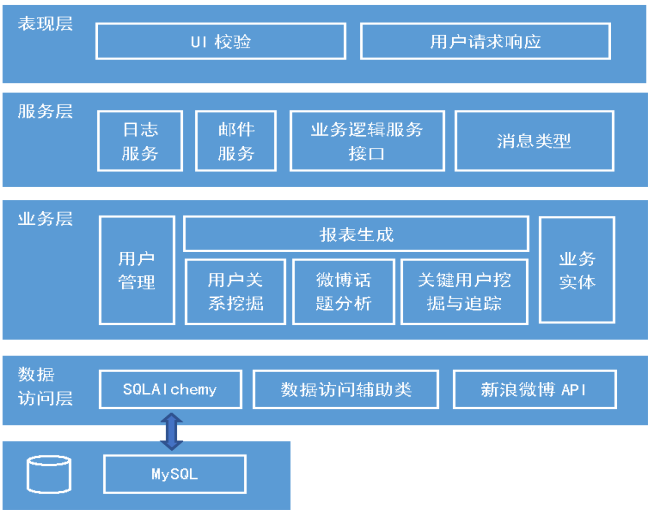


图 1 系统分层架构图

系统整体物理部署图如图 2 所示：

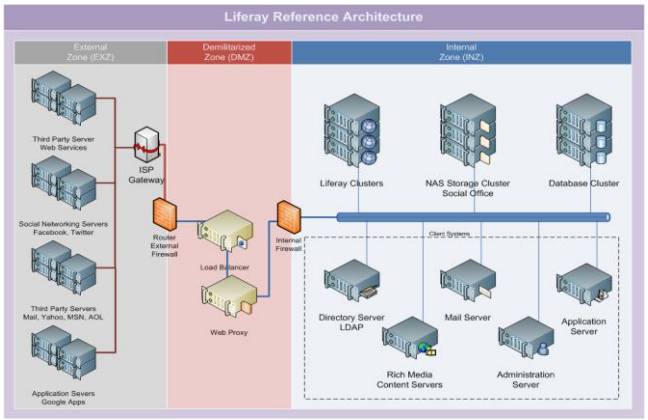


图 2 系统整体物理部署视图

针对现有系统存在的主要问题，我们团队从以下几方面研究开发了本系统：

1.系统以突发事件网络舆情为研究背景，通过处理境内及境外媒体网站上关于国内事件的报道，利用TF-IDF 文本聚类等相关技术手段提取每篇新闻的关键信息，获取国内敏感事件关键词列表，并加入事件发生时间、地点等相关信息，整理后建立「敏感事件话题语料库」。采用自然语言分析技术，将该语料库中的内容与用户微博内容进行精确匹对以挖掘出敏感的微博消息。

2. 我们利用 MongoDB 集群为系统构建一个单独的私有云平台，不仅继承了传统监控系统利用云平台节省硬盘空间这一优势。同时我们的云平台采用分布式的架构方式，合理利用了集群的运算优势，将计算分配给多个设备，使集群的运算时间大幅度减少，进一步提高了系统的工作效率，保证了系统对敏感舆情分析的及时性，弥补了现有系统热点挖掘时效性差这一不足。

3. PageRank 作为 Google 排名运算法则的一部分，是其用于标识网页的等级/重要性的一种方法。它根据网站的外部链接和内部链接的数量和质量来对网站进行排名计算。我们在系统中引入 PageRank 算法，用微博用户的关注动作代替 PageRank 算法中的链接动作，分析微博用户社交关系网络并计算用户影响力。最后，结合用户的影响力排名和其含敏感信息关键词的微博的发布量，进而分析得到微博敏感人士，提醒微博管理人员关注此类人群发出的微博。

# 四、 实验及结果

为了解基于云计算的微博敏感信息挖掘系统的质量如何，同时希望在本系统发布之前能将缺陷修复。我们对系统进行了全面测试。测试案例采用动态黑盒测试的方式，根据现有的需求规格说明书进行测试案例的设计和选择，仅适用于对系统的正确性测试。测试案例针对系统的核心功能进行测试，主要包括用户影响力计算，敏感用户挖掘与追踪以及敏感话题抓取和热门微博话题获取等。测试结果详见项目测试报告。

# 五、新点总结

1.基于云的 MongoDB 集群搭建技术。在现有微博舆情监控系统使用了云计算节省硬件资源这一优势的基础上，我们的系统更进一步，合理利用了云计算集群的运算优势，很大程度上提高了系统的运行效率。同时数据采用分布式的存储方式，也有效保证了系统数据的安全性，增强系统的稳定性。

2.基于 PageRank 的用户影响力排名算法。引入 PageRank 算法来分析处理用户间的社交关系，计算得到的 PageRank 值将作为微博用户影响力的重要衡量标准。

3.基于热点人物的敏感信息挖掘。我们团队开发的系统是第一款基于微博用户的微博舆情监控系统，区别于传统的微博舆情监控系统基于微博文本信息的方式。以用户之间的社交关系为系统的重要衡量依据，更适应微博这一应用场景。

4.基于境外报道的敏感事件话题语料库。系统通过文本聚类等相关技术处理境外媒体网站上关于国内事件的报道以建立「敏感事件话题语料库」，提高系统过滤微博敏感用户的准确性和效率。

5.基于响应式网页设计的数据图表展示。系统生成报告将以 Web 界面的形式呈现给管理人员，在以往系统的基础上，加入响应式的网页设计的新方式以确保管理人员能在任何网络环境下查看分析结果。

# 六、未来工作

本系统是契合现今网络舆情状况的一个尝试，目前支持在新浪微博平台上的使用。我们在系统中建立的一个与具体微博无关的抽象微博数据层，将方便团队在后期对系统进行改进以接入腾讯微博、网易微博等其他微博平台。

而随着网络的日益发展以及微博舆情的不断转变，市场对舆情监控系统的需求也会随之日益加剧。我们将对系统不断地进行改善优化并根据新的需求采取新技术，加入与之对应的新功能，使该系统能不断突破、不断发展。