

Intelligent text big data analysis display platform

Yingzi Yuan
Hongli Zeng
Spring Meng

Dec 2022 – Feb 2023
INSIS, Beijing

Collaboration with Gilight Education Technology Co., Ltd



Project Background

With the **continuous development of campus informatization**, various campus information systems (such as portal websites, team sites, personal homepages, and campus forums) have become **increasingly complex**.

These systems generate **massive amounts of data**, but there is a **lack** of effective integration, management, and analysis.



Current Problems:

- **Difficult information retrieval** due to scattered data sources and lack of a unified search entry.
- **Lack of effective supervision**, making it difficult to monitor content quality and detect sensitive or inappropriate content.
- **Potential security risks**, including personal information leaks and harmful content.
- **Limited support for decision-making**, as there are no effective tools to extract insights from text data.

System Architecture



Portal websites,
team sites,
personal homepages,
campus forums
...



Exact search
Fuzzy search
Multi-word search
Statistical analysis
Hot topic discovery
Similarity calculation
Update frequency monitoring



User management
Keyword search
Statistics
Visual reports

Data Sources

Data Storage & Processing

Key Functions

User Interface

Security Measures

Data masking, encryption,
and authorization to ensure
data privacy and security.

Operations Assurance

Monitoring, process management,
operational dashboards, and reporting tools
to ensure smooth system operation.

Problems to Solve

**Campus Text Data
Collection &
Management**

**Campus Text Data
Retrieval**

**Campus Text Data
Analysis**

Overall Design Concept

Data Collection:

Real-time crawling of all campus information release pages

Data Management:

Text data is pre-processed using word segmentation, stored using the Hadoop ecosystem and MongoDB, and indexed using Elasticsearch.

User Interface:

Based on Java and VUE framework, providing users with functions such as configuration, login, search, data visualization, and export.

Key Text Analysis Features

Hot Topic Discovery:

Identify trending topics over time and generate ranking lists to help staff follow the latest developments.

Monitoring and Alerts:

Detect sensitive words, inappropriate remarks, and visualize results for administrators to take immediate action.

Content Quality Monitoring:

Analyze publication frequency and readership to assess content quality. Missing content is flagged for review.

Relationship Analysis:

Explore hidden correlations between different groups and content types to improve content recommendation accuracy.

Home page overview

Through **comprehensive event analysis across the network**, **heat index tracking**, and **comment analysis**, the platform enables **real-time monitoring of published content across all sites and pages**, while also **tracking and counting the number of detected anomalies** to meet the **monitoring requirements of customers**.



Search & Query

Through **keyword extraction**, related **word analysis**, and **precise query capabilities**, the platform enables **multi-functional monitoring of website text, images, and videos**, achieving **accurate identification of the target information**.

✔ Supports **exact search** and **fuzzy search**

✔ Supports **single keyword search** and **multi-keyword combined search**



Anomaly Alerts

Based on **pre-configured alert conditions set by the user**, the platform can **intelligently identify abnormal information**. When **content matching the alert criteria** is detected, the system will **automatically trigger an alert**, allowing users to take immediate action.

Key monitored anomalies:

- Sensitive words
- Broken links
- Empty content
- Zombie websites (inactive sites)

The screenshot displays the 'Text Big Data Platform' (文本大数据平台) interface. The left sidebar contains navigation options: 首页 (Home), 异常报告 (Anomaly Report), 检索查询 (Search), 任务列表 (Task List), 统计分析 (Statistical Analysis), and 设置 (Settings). The main content area is divided into two panels. The left panel, titled '敏感词 (310)' and '异常链接 (130)', shows a table of monitored items. The right panel, titled '敏感词 (310)' and '异常链接 (130)', displays a list of alerts with details and a '取消预警' (Cancel Alert) button.

#	类型	站点	URL
10	死链接	北京交通大学环境学院	https://env.bjtu.edu.cn/
11	死链接	计算机与信息技术学院	http://scit.bjtu.edu.cn/
12	内容为空	北京交通大学国有资产管理处	http://gzc.bjtu.edu.cn/
13	内容为空	土木工程实验教学示范中心	http://tcec.bjtu.edu.cn/
14	死链接	北京交通大学招生资讯网	https://zsww.bjtu.edu.cn/
15	死链接	北京交通大学招生资讯网	https://zsww.bjtu.edu.cn/
16	内容为空	北京交通大学信息化办公室	http://ic.bjtu.edu.cn/
17	内容为空	北京交通大学信息化办公室	http://ic.bjtu.edu.cn/
18	内容为空	北京交通大学信息化办公室	http://ic.bjtu.edu.cn/

敏感词 (310) 异常链接 (130)

马克思主义学院韩振峰教授作党的二十大精神专题辅导报告
发布日期: 2022-11-19
为进一步学习宣传党的二十大精神, 深刻领会新时代新征程中国共产党的使命任务...
时间: 2022-11-19 11:46:08 浏览量: 0
URL: http://mkszyxy.bjtu.edu.cn/zyim/xzhd/187979.htm

取消预警

马克思主义学院等单位举办第十四届青年教师教学基本功大赛初赛
为促进青年教师教学技能水平的提高, 推动教师教学质量建设, 2022年10月27...
时间: 2022-10-30 11:46:08 浏览量: 0
URL: http://mkszyxy.bjtu.edu.cn/zyim/xyxw/187457.htm

取消预警

威海国际学院推荐2023届优秀应届本科毕业生免试攻读研究生(普通)名单公示
威海国际学院推荐2023届优秀应届本科毕业生免试攻读研究生(普通)名单公示 时间: 2022-09-19 经学生本人申请, 威海国际学院推免工作办公室审核报名资格及申请材料, 并对报名...
时间: 2022-09-19 16:45:35 浏览量: 0
URL: http://wh.bjtu.edu.cn/tzgg/announcement/186698.htm

取消预警

教育部教师队伍建设改革近期工作举措
教育部教师队伍建设改革近期工作举措 日期: 2022-09-06 月6日, 在教育部新闻发布会上, 教育

Rectification Tracking

Based on **pre-configured alert conditions set by the user**, the platform can **intelligently identify abnormal information**.

When **content matching the alert criteria** is detected, the system will **automatically trigger an alert**, allowing users to take immediate action.

✔ **Track rectification progress for each issue, ensuring timely resolution and content quality improvement.**

Status categories:


- **Pending rectification**
- **Completed rectification**



Statistical Analysis

Key Analysis Functions:

- Comprehensive site statistics
- Trending content identification
- Popular keyword tracking



The screenshot displays the 'Text Big Data Platform' (文本大数据平台) interface. The left sidebar contains navigation options: Home (首页), Abnormal Warning (异常预警), Search (检索查询), Task List (任务列表), Statistical Analysis (统计分析), and Settings (设置). The 'Statistical Analysis' option is currently selected. The main content area shows a table of site statistics with columns for No., Site (站点), Content Update (内容更新), Sensitive Word Alert (敏感词预警), Abnormal Link (异常链接), Content Measurement (内容测量), and Data Update Time (数据更新时间). The table lists 12 sites, including various departments and centers of Beijing Jiaotong University.

No.	站点	内容更新	敏感词预警	异常链接	内容测量	数据更新时间
1	马克思主义学院	23	5	0	0	2022-11-19 11:46:08
2	北京交通大学人事处网站	23	0	0	0	2022-10-24 16:36:58
3	北京交通大学人文社会科学处	21	0	0	1196	2022-11-08 17:09:08
4	北京交通大学后勤产业集团	17	0	0	0	2022-11-14 14:49:09
5	北京交通大学工程硕士中心	17	0	0	15062	2022-11-09 11:41:08
6	北京交通大学法学院	17	0	0	0	2022-11-07 17:38:08
7	詹天佑学院	16	0	0	0	2022-11-04 23:06:08
8	土木工程实验教学示范中心	15	0	0	0	2022-11-01 14:38:08
9	北京交通大学电气学院	15	0	0	0	2022-10-18 16:30:49
10	北京交通大学保卫处	14	0	0	0	2022-11-01 11:07:08
11	北京交通大学招生资讯网	14	0	0	0	2022-11-18 16:00:09
12	北京交通大学国际教育学院	14	0	0	0	2022-11-12 17:58:08

Summary

This platform provides comprehensive **campus text data management**, including:

- Real-time data collection.
- High-performance search.
- Advanced text analysis and monitoring.
- Visual reporting and decision support.

It helps the university:

- Monitor **public opinion** and campus sentiment.
- Support **decision-making** with data insights.
- **Enhance data security** and content management.



Thank You

Yingzi Yuan
Hongli Zeng
Spring Meng

Dec 2022 – Feb 2023
INSIS, Beijing

Collaboration with Gilight Education Technology Co., Ltd

