A Preliminary Study on the Application of the Internet News Publishing Public Opinion Monitoring System in Liaoning Province's Integrated Internet Content Supervision Platform (Postprint)

Authors: Liu Ning

Date: 2023-10-08T00:00:00+00:00

Abstract

This paper presents a technical analysis of data collection technologies for internet news publishing public opinion monitoring systems.

Full Text

A Preliminary Study on the Application of the Internet News Publishing Public Opinion Monitoring System in Liaoning Province's Integrated Internet Content Supervision Platform

Abstract: This paper elaborates on the data collection technology analysis of the internet news publishing public opinion monitoring system.

Keywords: system; public opinion topic-focused crawler; text and sentiment analysis technology

Author: Liu Ning

1. Overview

The integrated internet content supervision platform is built upon unified modules for data collection, data analysis, and statistical reporting. It forms an extensible, integrated intelligent monitoring platform for different monitoring

domains, including mobile APP audio-visual programs, internet news publishing public opinion, and online illegal publications. The platform comprises three main components: "Internet News Publishing Public Opinion Monitoring," "Mobile APP Audio-Visual Program Monitoring," and "Network Illegal Publications Monitoring."

This system can comprehensively monitor real-time public opinion hotspots related to news publishing across the internet, audio-visual programs published in mobile APPs, and the dissemination of various online publications (such as online literature, web comics, and online games). It enables timely detection of illegal audio-visual programs, harmful online publications, and negative public opinion disseminated online, while providing download and evidence collection capabilities. The system delivers comprehensive, complete, and detailed daily monitoring data and information to regulatory departments.

The following sections will focus specifically on introducing the data collection aspects of the internet news publishing public opinion monitoring system within this integrated platform.

2.1 Data Processing in the Public Opinion Monitoring System

The data processing workflow begins with web spiders crawling data from the internet. As data is crawled, information is simultaneously transmitted to the application server, which delegates tasks to intelligent agents for processing. These agents represent the core functional subsystem, conducting comprehensive analysis and filtering of all crawled network data to identify regulated illegal information for further processing by other subsystems. The intelligent agents can continuously self-learn to improve their knowledge systems and enhance their intelligence.

Users simply need to configure the homepage addresses of target websites, and the spider program will download corresponding pages according to the specified sites and transmit them to backend processing programs. The crawled data is packaged into temporary data packets, with new tasks sent to application servers. The server selects an idle intelligent agent to assign the task for analysis. Upon completion, the agent returns feedback to the application server, which may then assign the task to an idle storage processor.

The storage component performs several key functions: it stores suspected case data and all temporary data into case databases and master databases; parses IP addresses from discovered case URLs; highlights text content according to rule numbers for classified cases; transfers cases from temporary databases to master case databases; statistics on case types and quantities discovered in temporary tables trigger server alerts; and normal information is transferred to the master Total database.

When suspected illegal public opinion information is identified, the storage com-

ponent imports it into the database while also transferring normal information to the master database. If violations are detected, notifications are sent to users responsible for monitoring the relevant topic based on violation type. Client users can then perform auditing, feedback, confirmation, and printing functions on cases. The entire system data processing workflow is illustrated in the figure below.

2.2 Text and Sentiment Analysis Technology

Public opinion event information disseminated through various internet information systems expresses not only objective facts but also user viewpoints and emotions, such as support, opposition, or neutral attitudes toward events. These emotional attitudes are primarily expressed through text published by ordinary internet users, containing diverse viewpoints and positions on social phenomena. As individuals and organizations increasingly utilize online sentiment information for decision-making, sentiment analysis technology has emerged.

Sentiment analysis technology plays a crucial role in describing and predicting the development trends of online public opinion events. However, due to the diversity of online public opinion information and the particularities of Chinese text processing, Chinese sentiment analysis for online public opinion events faces several challenges.

First, emotional judgment of online public opinion events is highly subjective. Different people, constrained by their backgrounds and cognitive levels, do not consistently judge the same information, making unified standards difficult to establish and machine-based sentiment determination particularly challenging. Second, online information carriers are diverse with non-uniform data formats and types. Public opinion events may be expressed through long texts such as news articles and blogs, or through short texts on forums and microblogs. Written language mixes with colloquial expressions, while new internet vocabulary and variant terms proliferate, significantly increasing the difficulty of sentiment analysis.

Third, relevant corpora for online public opinion events are difficult to obtain. Although current construction of Chinese and English corpora related to online public opinion events remains incomplete, the primary technologies for sentiment analysis require substantial corpus support. Fourth, Chinese sentiment analysis is particularly challenging. While considerable research has been conducted on English sentiment analysis, Chinese presents unique difficulties. Its accuracy is directly correlated with the precision of Chinese word segmentation, named entity recognition, syntactic parsing, and other tools, whose accuracy substantially impacts Chinese sentiment identification.

The text sentiment analysis process proceeds as follows: First, input text undergoes preprocessing, where it is segmented into sentences and then into words. Second, word-level sentiment analysis determines the emotional orientation of individual words within each sentence. Third, sentence-level sentiment analysis

applies these word sentiments to determine the overall emotional orientation of each sentence. Finally, the importance of each sentence within the document is calculated and combined with sentence-level sentiments to produce the final positive or negative classification of the document. The figure below illustrates the text sentiment analysis workflow.

The final internet news publishing public opinion analysis system can actively discover domestic and international hotspots, harmful information, and actionable intelligence related to national and "concerning-us" (news publishing-related) interests, while tracing their dissemination paths. It supports the detection and discovery of hotspots concerning specific social groups related to business operations. The system characterizes network hotspots through hotspot cloud visualization and multi-hotspot analysis indices.

Through thematic analysis, the system enables operators to perform monitoring tasks driven by specific topics or events, providing one-stop services from active data collection and analysis to statistical reporting and brief generation. It supports analysis of event trends, current influence, evolutionary stages, information traceability, social network dissemination, online promoter identification, regional distribution of netizens, netizen sentiment analysis, and viewpoint extraction, with automatic brief generation capabilities.

Note: Figure translations are in progress. See original paper for figures.