Postprint: News Republication and Citation Analysis for Internet Data

Authors: Chen Xinyi, Chen Jun, Wang Yi

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

Abstract

The development of Internet, big data, and new media technologies has brought revolutionary changes to media communication channels and content forms. Analyzing the adoption and dissemination of news across different media channels constitutes a crucial component in building big data-driven editorial and dissemination decision-making, and holds significant importance for enhancing the domestic and international communication capabilities of news agencies. However, due to issues such as non-standard data formats on the Internet and new media, and the lack of source attribution in reprinting and citation, analyzing news reprinting and citation in new media presents considerable challenges. This paper collects news data from multiple sources, including websites, electronic newspapers, WeChat official accounts, and mobile clients, covering over 5,000 Chinese and English media outlets and more than 400,000 new media accounts worldwide. By leveraging intelligent information matching technology to track the adoption of news across all media platforms, we have constructed a news reprinting and citation analysis system, which lays the foundation for further analyzing media communication pathways, grasping the patterns of domestic and international media dissemination, and enhancing the influence of public opinion both at home and abroad. This article introduces the working principles and construction significance of news reprinting and citation analysis, conducts in-depth research on key technical implementations, and proposes future development recommendations for news reprinting and citation analysis based on this foundation.

Full Text

Analysis of News Reprint and Citation for Internet Data

Abstract: The development of internet, big data, and new media technologies has brought revolutionary changes to media communication channels and

content formats. Analyzing how news is adopted and disseminated across different media channels constitutes a critical component of building big data-driven editorial and dissemination decision-making systems, and holds significant importance for enhancing the domestic and international communication capabilities of news agencies. However, due to issues such as non-standardized data formats in internet and new media, and the frequent omission of attribution in reprints and citations, analyzing news reprint and citation patterns in new media environments presents substantial challenges. This paper collects news data from multiple sources—including websites, electronic newspapers, WeChat official accounts, and mobile clients—covering over 5,000 Chinese and English-language media outlets and more than 400,000 new media accounts worldwide. By leveraging intelligent information comparison technology to track news adoption across all media platforms, we have constructed a news reprint and citation analysis system that lays the foundation for further analysis of media dissemination paths, understanding of domestic and foreign media communication patterns, and enhancement of public opinion influence. This paper introduces the working principles and significance of news reprint and citation analysis, conducts in-depth research on key technical implementations, and proposes future development recommendations for news reprint and citation analysis.

Keywords: News reprint and citation; Large-scale text similarity; Hadoop Spark

Classification Code: TP392
Document Code: A
Article ID: 1671-0134(2017)11-089-03
DOI: 10.19483/j.cnki.11-4653/n.2017.11.029

1. Concept of News Reprint and Citation Analysis

News reprint and citation analysis aims to identify, within massive real-time internet datasets, media outlets that have reprinted or cited a particular original news article through a series of technical methods.

Reprint refers to the practice where newspapers, websites, or other media publish news articles previously released by other media. In domestic reporting, full-text reprinting of news by media outlets is relatively common. **Citation** refers to the partial use of statements or information from news articles previously published by other media. In international reporting, overseas media—particularly major international outlets—typically cite a paragraph or sentence from a news article, or paraphrase the original information. In news reporting, citations serve two primary purposes: either to quote facts before delving deeper into coverage, or to quote viewpoints for elaborating concordant or opposing perspectives.

Explicit reprint and citation occurs when media outlets clearly attribute

the source, either by retaining the original dateline in reprints or by specifying "according to a report from X media outlet" when citing. **Implicit reprint and citation** occurs when media outlets reprint or cite content without attribution, a practice known as implicit reprinting or citation. Compared to explicit cases, implicit reprint and citation present significantly greater identification challenges. As internet technology evolves and various new media platforms continuously emerge, they expand communication boundaries while simultaneously exhibiting irregularities in attribution practices.

2. Significance of News Reprint and Citation Analysis

By analyzing reprint and citation patterns of news across Chinese and English-language websites, electronic newspapers, WeChat, and mobile clients—marking cited paragraphs and sentences, and identifying adopting media, adoption time, and placement information—we can promptly track and analyze news adoption across all media platforms. This enables statistical evaluation of editorial staff performance and analysis of article dissemination effectiveness, providing data support for guiding efforts to enhance news communication impact.

3. Working Principle of News Reprint and Citation Analysis

This paper proposes a news reprint and citation analysis technique based on textual semantic comparison, comprising four main steps: news feature extraction, similar news clustering, news reprint and citation relationship determination, and result verification.

News Feature Extraction: This step employs web information extraction technology to extract features from internet news data. For each article, the publication time is extracted by analyzing webpage structure using a hybrid machine learning and rule-based algorithm.

Similar News Clustering: A similarity cluster partitioning algorithm groups collected internet news data by semantic similarity. News articles within each cluster are semantically similar and may have implicit reprint relationships.

News Reprint and Citation Relationship Determination: Based on similarity scores among news articles within clusters and publication timestamps, empirical thresholds are applied to analyze and determine reprint and citation relationships.

Result Verification: The determined results undergo secondary verification.

4. Technical Architecture of News Reprint and Citation Analysis

The overall system data processing architecture is illustrated in Figure 1 [Figure 1: see original paper]. The primary architectural design concepts and data

processing procedures consist of the following components:

Data Ingestion Layer: Internet news data collected through large-scale web crawling and third-party sources first undergoes deduplication using Redis, followed by preprocessing and ETL to produce structured, normalized data.

Task Scheduling Layer: Built upon the Kafka distributed message queue, this layer enables data access and buffering. Data in Kafka queues is processed using both real-time Spark Streaming for stream computing and offline large-scale MapReduce computing frameworks for news reprint and citation analysis.

Data Storage Layer: For massive news datasets, distributed storage enables efficient business logic operations, scalable storage deployment strategies, and highly available redundant storage. MySQL serves as the foundational storage database for reprint and citation statistics, responsible for data model definition and accumulation but not for complex query services. ElasticSearch functions as a mirror of core MySQL business tables, enabling data synchronization, multi-table joins, and data redundancy to enhance query performance. Additionally, it acts as the real-time server for data service operations, providing online query capabilities. Hive serves as the offline server for data services, offering large-scale offline data query and analysis services. FastDFS functions as a distributed file storage system for managing images, PDFs, and Excel reports.

Integration Service Layer: Addressing business requirements, this layer leverages service bus technology to publish underlying data through flexible query and extraction logic to upper-layer service interfaces, implementing universal service interfaces. Based on Zookeeper and Dubbo, the service bus enables unified coordination, scheduling, and configuration management.

5. Key Technology Implementation

5.1 Web Information Extraction

Traditional methods for parsing content from webpage source code typically employ recursive parsing of sub-tags to extract tag content individually. However, in practical applications, this approach exhibits excessive complexity and resource consumption when parsing complex webpage source code. To address this issue, this paper designs a webpage content parsing algorithm that combines XPATH technology with recursive parsing of webpage structure trees to extract content. XPath (XML Path Language) is a language used to locate specific portions of XML documents, providing node-finding capabilities within data structure trees.

Main body content typically resides within specific HTML tags or their sub-tags. This algorithm first uses XPATH to identify the main content blocks of a webpage. For each content block, a webpage structure tree is constructed, and all tags are traversed recursively on this tree. During traversal, the full tag path is recorded to avoid duplicate parsing. This process captures webpage information including title, body text, links, source, and publication time.

5.2 Text Similarity Comparison

Using text similarity comparison algorithms, articles are partitioned into different similarity clusters. This paper employs the classic Vector Space Model (VSM) and Bag of Words (BOW) as document representation models. The fundamental concept involves dividing documents into feature terms and quantifying feature term weights to represent documents as mathematical vectors, enabling inter-document similarity calculations. The TF-IDF algorithm is used for feature term weighting. The text similarity calculation process is illustrated in Figure 2 [Figure 2: see original paper].

In text similarity algorithms, similarity measurement methods play a crucial role. Common methods include Pearson Correlation Coefficient (PCC), Cosine Similarity, and Euclidean Similarity. Comparative analysis reveals that Pearson Correlation Coefficient is most suitable for this algorithm. PCC measures linear correlation between two vectors, calculated as follows:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

5.3 News Reprint and Citation Relationship Construction

Based on similarity clusters and leveraging source information and publication timestamps, reprint and citation networks are constructed. This paper utilizes graph databases to build and store these networks, supporting dynamic updates and multi-level reprint and citation relationship queries. Finally, using network path tracking technology, the reprint and citation path of each news article can be traced to locate and track reprint and citation patterns.

6. Specific Technical Implementation of Similar Text Clustering

Two similar text clustering systems have been developed for different application scenarios: a Hadoop platform-based batch processing system and a distributed in-memory real-time computing system.

6.1 Hadoop Platform-Based Similar Text Clustering

Technologies represented by the Hadoop Distributed File System (HDFS) and MapReduce computing model have achieved significant success in big data batch processing. Additionally, Hadoop' s mature ecosystem provides rich component support, and this paper employs algorithms from the Mahout data mining toolkit, substantially simplifying processing complexity.

6.2 Distributed In-Memory Real-Time Similar Text Clustering

The distributed in-memory real-time similar text clustering system targets scenarios with high real-time requirements. This system achieves sub-second response times for data processing. The processing framework is illustrated in Figure 3 [Figure 3: see original paper].

7. Integration and Testing Results

After multiple rounds of testing and algorithm optimization, the current system achieves over 95% accuracy for Chinese text news reprint and citation analysis, and over 90% accuracy for English text analysis.

8. Application Prospects for Internet Big Data-Based News Reprint and Citation Analysis

Dissemination Path Analysis: Combining similarity text clustering with analysis of complete news dissemination paths to identify key media or new media accounts in the propagation chain.

Thematic Reporting Analysis: Conducting reprint and citation analysis on a set of news articles within thematic coverage, analyzing and summarizing dissemination patterns in conjunction with timing, geography, and event development processes.

Public Opinion Guidance Analysis: Within coverage of a news event, analyzing news reports before and after a particular article to study its public opinion guidance role and achieved effects.

In April 2017, the system entered trial operation, providing editorial staff company-wide with real-time queries on article adoption across all media platforms. It provides foundational data for newsroom business statistics and performance evaluation at both headquarters and branch offices, and supports large-screen displays of company-wide reporting, adoption, and engagement metrics, achieving promising initial results. As applications deepen, both editorial and statistical staff have proposed new requirements. The system will continue to address challenging topics including adoption analysis for multimedia content (images and video) and analysis for minor-language articles.

References

[1] Holden Karau, et al. *Spark Fast Data Analysis* [J]. Beijing: Posts & Telecom Press, 2015(10): 161-185.
[2] Sean Owen, et al. *Mahout in Action* [J]. Beijing: Posts & Telecom Press, 2014(3): 40-47.
[3] Tom White. *Hadoop: The Definitive Guide* [J]. Beijing: Tsinghua University Press, 2011(7): 160-174.

(Author Affiliation: Xinhua News Agency Communication Technology Bureau)

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