Post-print: Practical Exploration of Big Data Analysis Applications in Policy Interpretation

Authors: Tan Xin

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

Abstract

The acquisition and research of information regarding the latest policies and their interpretations hold significant value for various sectors, including media, enterprises, and research institutions. By leveraging web data scraping and intelligent text mining technologies to monitor and collect data from first-hand policy release sources and interpretation articles, and subsequently performing data processing and intelligent mining analysis to develop policy interpretation application data products, the efficiency of information acquisition and policy research can be substantially enhanced. This paper presents, from a practical perspective, solutions to key challenges and application functionalities within policy interpretation applications.

Full Text

Preamble

Theoretical Research · Media Theory: Practical Exploration of Big Data Analysis Applications for Policy Interpretation
ChinaXiv Cooperative Journal

Abstract: Accessing and researching the latest policy releases and policy interpretation information holds significant value for industries including media, enterprises, and research institutions. By employing web data scraping and intelligent text mining technologies to monitor and collect first-hand policy release sources and interpretation article data sources, and conducting data processing and intelligent mining analysis, the resulting data products for policy interpretation applications can greatly enhance the efficiency of information acquisition and policy research. This paper introduces solutions to key problems and application functions in policy interpretation practice from a practical perspective.

Keywords: policy interpretation; big data analysis; text mining; data product; value-added services

CLC Number: TP393
Document Code: A
Article ID: 1671-0134(2019)03-022-02
DOI: 10.19483/j.cnki.11-4653/n.2019.03.001
Author: Tan Xin

Each year, government departments at all levels release a large number of policies covering various aspects of social development and people' s livelihood, making each new policy release a focus of public attention. Questions such as what is the purpose of the policy release, how to interpret the policy in detail, and which industries and sectors are affected after the policy release have become focal points for all sectors of society, including the media, business, academia, and the general public. Therefore, the rapid and comprehensive collection and analysis of policy and interpretation data is of practical significance.

This paper describes how automated web information scraping technology, big data technology, and intelligent text mining technology can be used to quickly and efficiently aggregate the latest policy and policy interpretation information. Based on this foundation, the paper further discusses processing techniques including extraction of key policy element information, multi-dimensional data classification, and data association, thereby forming a comprehensive data product for policy interpretation that provides multi-scenario, multi-role information services for media practitioners, industry researchers, and the business community, ultimately improving the efficiency of information acquisition and data research.

1. Challenges in Policy Information Acquisition

Industries have high demands and requirements for accessing the latest policy information. For the media industry, obtaining policy release information in a timely manner is essential for writing promotional reports or interpretive articles on new policies, where the timeliness of publication directly reflects and influences media reach and impact. For enterprises, timely research on how new policies affect corporate development, R&D investment, market changes, and decision-making analysis is crucial for formulating internal strategies, as many policy releases have a critical impact on business development directions. However, challenges such as the wide range of policy information sources, non-concentrated release times, and the need for information integration constrain policy information acquisition.

1.1 Wide Range of Policy Data Sources

Policy releases originate from government departments at all levels, making source diversity a major difficulty for individuals seeking policy information. First, authoritative policy sources include government official websites, official news apps, official WeChat accounts, and officially verified Weibo accounts, representing multiple types of release channels. Second, government departments

are numerous across different administrative levels and departmental categories, requiring attention to multiple official information sources even when focusing on a single policy domain. Third, regarding post-release policy interpretation articles, besides official interpretations on government websites, articles written by government officials, domain experts, researchers from institutions, and media professionals also hold significant reading value. However, these sources are even more widespread, potentially appearing on news portals, news apps, newspapers or digital editions, institutional WeChat or Weibo accounts, and personal blogs, WeChat accounts, or Weibo accounts of domain experts. In summary, quickly browsing through various policy interpretation perspectives from all parties presents considerable difficulty.

Only a small portion of policies released each year are published at fixed times; the majority are introduced in real-time according to social development needs, making advance preparation for policy information acquisition impossible.

In conclusion, in today' s era of information overload, the key issues that policy interpretation applications must address include how to quickly and efficiently obtain policy and interpretation information, how to precisely acquire data needed by researchers across industries, and how to leverage artificial intelligence and machine analysis capabilities to aggregate and analyze information for researchers.

2. Application Practice of Big Data Analysis for Policy Interpretation

Policy interpretation applications leverage big data and artificial intelligence technologies to achieve automated information collection, multi-dimensional automatic indexing, text mining, and association analysis, thereby enhancing the utilization efficiency of policy and interpretation information through visual presentation.

2.1 Automated Monitoring and Collection

By employing mature automated web information scraping software, the application implements real-time monitoring of target websites that serve as policy and interpretation release sources, promptly collecting the latest web pages to local storage for content analysis and information filtering, thereby completing local storage of policy interpretation information.

During the data collection process, the application not only transforms unstructured web data into semi-structured data but also automatically extracts policy-related data including policy names, release dates, policy text content, issuing unit names, publishing website names, channel names, and article URLs. Subsequent text mining and processing build a policy metadata database that provides foundational data services for policy interpretation applications.

The collection focuses primarily on first-hand release data sources rather than

second-hand reposted data to ensure timeliness, accuracy, and reliability of information acquisition.

2.2 Processing of Policy and Interpretation Information

Data collection technology enables real-time monitoring and collection of the latest online data, freeing individuals from browsing and searching time. Text mining technology provides a series of data processing capabilities including automated classification, automatic clustering, intelligent information extraction, data association analysis, and automatic data indexing, solving the problem of policy data silos and making policy data applications more effective. [Figure 1: see original paper] illustrates the processing workflow for policy and interpretation data.

In front-end application functions, these policy tags are utilized to display policy lists through refined navigation. Combined search functions enable users to search for policies and interpretation articles, allowing personalized retrieval results through custom keywords for rapid and comprehensive information understanding. Full-text search functionality is provided for titles, bodies, and themes of policies and interpretation articles. Filtering capabilities are available for fields such as issuing unit name, document type, industry domain, geographical region, and release year. Multi-dimensional classification tags also enable more flexible page filtering and provide basic options for personalized subscriptions.

2.3 Multi-dimensional Classification and Indexing

Employing automatic classification and rule-based classification technologies, the application performs multi-dimensional classification and indexing of policies to help users with different identities and needs quickly and effectively locate required categories and corresponding policy information across various scenarios.

Classification tags include industry domain, geographical region, issuing unit, subject matter, document type, and release year. Table 1 provides examples of classification categories.

Table 1: Examples of Policy Classification Names and Values - Industry Domain: Finance, Manufacturing, Internet···- Geographical Region: Beijing, Shandong, Hubei···- Issuing Unit: State Council, National Development and Reform Commission, Ministry of Industry and Information Technology···- Subject Matter: Financial Regulation, Personal Income Tax, E-commerce···- Document Type: Announcement, Notice, Proposal···- Release Year: 2019, 2018, 2017···

2.4 Policy Text Mining

Conducting data mining and key information extraction on policy texts forms the foundation for policy indexing and retrieval, information association analy-

sis, and multi-dimensional classification indexing. Employing natural language processing technologies such as automatic text segmentation and part-of-speech tagging, and combining rule-based and statistical approaches, the application performs Chinese word segmentation and policy information extraction, including structured extraction of policy theme keywords, related persons, institutions, and regional names, thereby completing keyword and entity indexing of policies.

In policy information display functions, multi-dimensional intelligent analysis and association help users quickly discover key information within a policy and related articles. Policy theme words, policy entity mining results (related persons, institutions, regions), and temporal trends and volumes of policy interpretation articles are presented in chart form. Related policies, interpretation articles, and media reports are displayed as article title lists. [Figure 2: see original paper] shows a schematic diagram of the display results.

Rapid mining and comparative display of multiple perspectives enable users to grasp policy content more comprehensively. Using semantic analysis technology, articles with correlation exceeding a certain threshold are associated to achieve deep mining of complex semantic relationships, thereby completing the association between policies and official interpretation articles, media interpretation articles, related reports, and previous years' policies.

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