ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202310.00905

Postprint: Innovative Practice of "Word Cloud" as a News Data Visualization Tool in New Media

Authors: Qin Yufang, Li Ruonan, Liu Yingxu

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

Abstract

Fueled by the vigorous development of Internet technology and new media services, innovative data journalism products are emerging in an endless stream, and data visualization of news has become a prevailing trend. Data visualization practices both domestically and internationally are being actively pursued. Among the numerous forms of news visualization, word clouds can provide a quick overview of news content and extract key information, feature diverse presentation forms, and are favored by editors, journalists, and readers. This paper takes the news data visualization tool "word cloud"as the entry point, and based on the scenario requirements for news reporting, implements a word cloud tool using news keyword extraction algorithms and data visualization technology, effectively improving the efficiency of Xinhua News Agency' s editors and journalists in creating word clouds.

Full Text

Innovative Practice of Word Cloud: A News Data Visualization Tool in New Media

Qin Yufang, Li Ruonan, Liu Yingxu
(Xinhua News Agency Communication Technology Bureau, Beijing 100803)

Abstract: With the vigorous development of Internet technology and new media services, innovative data journalism products continue to emerge, making data visualization an established trend. Data visualization practices are in full swing both domestically and internationally. Among the various forms of news visualization, word clouds enable rapid overview of news content and quick access to key information, offering diverse presentation formats that have gained popularity among editors, journalists, and readers. This paper focuses on the word cloud as a news data visualization tool, implementing a word cloud solution based on news keyword extraction algorithms and data visualization

technologies tailored to news reporting scenarios, which effectively improves the efficiency of word cloud production for Xinhua News Agency editors and journalists.

Keywords: Internet technology; word cloud; keyword extraction; data journalism; visualization; data
Classification Code: TN816
Document Code: A
Article ID: 1671-0134(2021)09-046-04
DOI: 10.19483/j.cnki.11-4653/n.2021.09.013
Citation Format: Qin Yufang, Li Ruonan, Liu Yingxu. Innovative Practice of Word Cloud: A News Data Visualization Tool in New Media[J]. China Media Technology, 2021(09).

1. Product Form

The word cloud tool is an efficient and user-friendly data visualization solution designed for editors and journalists, integrating text analysis and word cloud generation capabilities that enable users without data analysis or graphic design backgrounds to easily create exquisite visualizations. As shown in [Figure 1: see original paper], the interface is clean and straightforward, divided into three main areas: a menu bar, a left display zone, and a right configuration panel. The core functionalities are concentrated in the right configuration panel, which comprises two major sections: data editing and chart settings. The data editing area provides text analysis capabilities, allowing users to upload files and perform statistical analysis of keyword frequency and weight. The chart settings area offers word cloud styling options for customizing shapes, fonts, colors, and other visual parameters.

Word clouds, also known as text clouds, visually represent "keywords" that capture main ideas within a text. The primary workflow involves extracting key information from text through algorithmic or statistical methods and presenting it through diverse visual formats. This approach provides readers with a more intuitive understanding of textual content and has become one of the most commonly used news data visualization forms for editors and journalists. While numerous word cloud tools with various functions and forms exist in the market, their application in news reporting scenarios presents challenges: the generated effects often fail to meet professional requirements, and security and copyright issues may arise. To address these limitations, we developed an online word cloud tool with simple operation and clean styling, specifically tailored to Xinhua News Agency' s reporting characteristics and based on thorough research into editors' and journalists' needs. Throughout the development and iteration process, we conducted in-depth exploration and practice in product design, overall technical architecture, and news keyword extraction algorithms.

1.1 Product Features

1.1.1 Data Editing Users can input text through file uploads, direct text entry, or by providing valid URLs. After selecting extraction types such as frequency or weight, the tool automatically analyzes the text and generates a default word cloud visualization in the left display zone. The extracted keywords are displayed in a data table where users can modify, add, or delete entries online. The tool also supports exporting processed datasets for download.

The automatic keyword extraction supports both long-word and short-word frequency algorithms (counting keyword occurrences in text) as well as weight algorithms (calculated based on semantic relationships within text context). File upload supports multiple formats including txt, doc, docx, and pdf, in addition to direct text input or URL submission.

1.1.2 Chart Settings To accommodate different news scenarios, the tool supports personalized configuration of word cloud shape, color, font, and animation. Users can see real-time rendering effects in the left display zone after parameter adjustments, ensuring both usability and efficiency.

The shape library offers rich templates including emojis, geometric shapes, numbers, and sports-related forms. Users can also upload custom images to define word cloud outlines, effectively expanding application scenarios. For color themes, the tool provides default palettes such as serious, lively, solemn, soft, and gradient options, and supports extracting color pixels from images as text colors, allowing flexible customization based on preferences. The rich color themes enhance visualization effects, making word clouds "stand out" in news reports.

The tool also provides multiple free commercial fonts that editors can use without additional loading or installation, enabling real-time font preview. Personalized configuration options allow users to adjust font size, outline, and animation parameters to optimize final results, significantly improving user experience.

2. Overall Technical Architecture

During system architecture design, the project team conducted thorough preliminary research through multiple discussions with editors and journalists, completing overall product design by referencing various commercial solutions. Considering that B/S architecture offers flexibility and low maintenance costs—requiring only network access and a browser for anytime, anywhere word cloud querying, creation, and modification—the team leveraged existing frontend visualization technology expertise to finalize technical strategies. The architecture is detailed below from both browser and server perspectives.

2.1 Browser Side

The browser side employs industry-leading technologies including the React framework, Ant Design UI library, and G2 visualization engine to implement a clean, user-friendly, and robust interactive interface.

React is a JavaScript library for building user interfaces that focuses on lower-level implementation logic, facilitating flexible custom component construction. Its Virtual DOM design minimizes unnecessary operations on the actual DOM through diff algorithms, delivering superior performance that has made it widely adopted in large-scale system architectures.

Ant Design is a high-quality, ready-to-use React component library offering rich foundational components with clean and aesthetically pleasing visual styles. Covering most application development scenarios and integrated with React' s robust ecosystem, it forms a complete frontend solution for efficiently customizing user interfaces and managing frontend projects.

G2 is an open-source charting library from Ant Financial featuring high usability and extensibility. Based on data-driven, highly interactive visualization grammar, it can flexibly render various chart types to facilitate visual analysis. In this project, we utilize G2' s graphics grammar to optimize the D3-based word cloud layout algorithm at the 底层, dynamically rendering large numbers of text tags.

The word cloud layout algorithm works as follows: First, keyword configuration parameters are initialized and data is sorted, beginning layout with the highest-weight keyword. Each keyword comprises four vertices represented as a rectangular area. During placement, a collision detection algorithm checks for conflicts with previously positioned keywords. If a conflict is detected, the keyword is repositioned along an Archimedean spiral. If the keyword cannot be placed at any point along the spiral, the algorithm proceeds to the next keyword.

2.2 Server Side

The server side is primarily based on Node.js, MySQL, and Redis technologies, with business logic developed using the Express framework for rapid and convenient API service creation. MySQL and Redis employ clustered data storage to improve system reliability and performance.

Node.js is an event-driven, non-blocking I/O JavaScript model based on Chrome' s V8 engine, providing various rich JavaScript module libraries that greatly simplify web application development. Express is a flexible and convenient web development framework for the Node.js platform, offering powerful features for quickly creating web applications and HTTP utilities. Its core features include: (1) responding to HTTP requests through middleware configuration; (2) defining routing tables for different HTTP requests; and (3) dynamically rendering HTML pages by passing parameters to templates. Due to its extensive

functionality packages, Express is commonly used as a middleware service layer framework for handling business logic in large projects.

MySQL is a multi-threaded SQL database server capable of fast, efficient, and secure processing of large data volumes. Compared to databases like Oracle, MySQL is more concise and has gained widespread use in web applications due to its speed, robustness, and usability. Redis is a high-performance data structure server that can function as a database, cache, or message broker, supporting structures including strings, hashes, lists, sets, sorted sets, bitmaps, and hyperloglogs. Running in memory while supporting disk persistence, Redis serves as an excellent complement to relational databases.

In this project, MySQL stores template configurations and user records while handling complex statistical queries. Redis stores task IDs for large-scale text keyword extraction, ensuring interface stability and accessibility through asynchronous polling.

3. Keyword Extraction Algorithm

The word cloud tool employs an unsupervised method based on word graph models for text keyword extraction. Document keywords represent thematic and critical content, serving as an important means for people to quickly understand and grasp document topics. Keywords are widely used in news reports, scientific papers, and other fields to facilitate efficient document retrieval, management, and search.

Keywords must simultaneously possess readability, relevance, and coverage. Readability means keywords themselves should be meaningful words or phrases. Relevance requires keywords to be thematically related to the document. Coverage demands that keywords adequately represent the document' s themes without focusing solely on one aspect while ignoring others.

Text keyword extraction methods fall into three categories: supervised, semi-supervised, and unsupervised. Supervised methods transform keyword extraction into a classification problem for each word, requiring extensive data annotation. Semi-supervised methods use small training samples to build models, then iteratively refine them through human filtering of extracted keywords. Unsupervised methods require no manual annotation, using various techniques to identify important words as keywords. Due to high labor costs for supervised methods and the need for manual intervention in semi-supervised approaches, current text keyword extraction primarily employs more adaptable unsupervised methods.

Unsupervised methods mainly include statistical feature-based, word graph model-based, and latent topic model-based approaches. Statistical feature methods rank words or phrases based on quantitative metrics such as part-of-speech, frequency, TF-IDF, position information, mutual information, and word span. Word graph model methods construct directed graphs with

candidate keywords as vertices and co-occurrence relationships as edges, then select important vertices as keywords using specific algorithms. Latent topic model methods leverage topic distribution properties for keyword extraction. Word graph model methods are currently the most commonly used, and our approach optimizes this methodology.

3.1 News Keyword Extraction Algorithm Flow

Considering that actual application documents are news articles, extracted keywords must be meaningful words or phrases to effectively summarize content. We employ the following approach:

Preprocessing: Remove special characters unrelated to content and structure from news articles, then perform word segmentation, part-of-speech tagging, and named entity recognition.

Generating Candidate Keyword Sets: This includes three steps: 1. Key phrase generation: Using only word segmentation results fails to identify key phrases due to overly fine granularity. We employ dependency syntax-based key phrase generation. 2. Obtaining positive words from news articles based on rules and word lists. 3. Filtering appropriate keyword candidates based on rules and metrics.

Keyword Ranking and Selection: Two ranking methods are used: weight-based and frequency-based. After ranking, top-ranked words are selected as keywords based on the required number.

Algorithm Details: The word segmentation and POS tagging algorithm can annotate 59 parts of speech. Named entity recognition identifies person names, locations, organization names, conference names, and temporal words. Selected keywords primarily include various nouns, verbs, and entity words, while excluded words include: the verb "是" (to be), the verb "有" (to have), directional verbs, formal verbs, modal verbs, and numerals.

3.2 Using Dependency Syntax for Phrase Extraction

Dependency syntax relationships describe syntactic-level relationships between sentence components, representing dependency relations between words. These relationships indicate not only syntactic collocations but also certain semantic associations. To identify phrases with practical meaning, we compiled 15 types of syntactic dependency relationships to describe dependencies in news articles: Subject-Verb (SBV), Verb-Object (VOB), Indirect Object (IOB), Fronted Object (FOB), Double Object (DBL), Attribute-Head (ATT), Adverbial-Head (ADV), Verb-Complement (CMP), Coordinate (COO), Preposition-Object (POB), Left Adjunct (LAD), Right Adjunct (RAD), Independent Structure (IS), Punctuation (WP), and Head (HED).

To ensure syntactic dependency effectiveness for news data, we selected 1,672 sentences from news articles, annotated their dependency syntax, and added

them to the original training data to improve algorithm adaptability. After POS tagging and dependency parsing, analysis results are merged, primarily combining attribute-head structure phrases. To prevent over-merging, rules are applied: conjunctions, auxiliary words, and punctuation are excluded from phrase structures; subject-verb relationships cannot be merged; and some verb-complement structures can form phrases.

compares word segmentation and phrase extraction results, demonstrating that phrase extraction better captures meaningful units.

3.3 Setting Positive Words

The inclusion of positive words enhances keyword extraction effectiveness for political news. Positive words are those that summarize content and have positive connotations, providing guidance in certain categories (e.g., political news). Through user communication, we obtain positive words in two ways: (1) extracting text within quotation marks, parentheses, and book title marks, and (2) allowing users to define positive word dictionaries.

In certain categories, positive words play a crucial role. To increase their weight during weight-based ranking, we boost their initial weight to ensure higher probability of appearing in top-ranked positions. User feedback confirms that positive words improve keyword extraction effectiveness for political news.

3.4.1 Weight-Based Method

After obtaining the candidate word set, we use a TextRank-based approach for ranking. TextRank calculation depends only on co-occurrence between words or phrases. The process involves: (1) constructing a graph with words as vertices and building edges when two words co-occur within a certain window, (2) applying the PageRank algorithm or similar methods to obtain each vertex' s weight, and (3) ranking vertices by weight and selecting top words as keywords.

This method' s advantage lies in determining word weight based on associations with surrounding words. Compared to TF-IDF and similar algorithms, it judges word importance through inter-word connections. In our implementation, we use word TF as initial values, select a word window of 5, and incorporate the positive word mechanism to ensure their weight during initial weight assignment.

3.4.2 Frequency-Based Method

Word frequency refers to a word' s occurrence count in an article—higher frequency indicates greater importance. Based on user feedback indicating that weight values from ranking algorithms don't intuitively reflect vocabulary importance, we offer a frequency-based alternative after generating candidate keyword sets.

Frequency statistics include two approaches: (1) **Segmentation frequency**—counting occurrences after word segmentation, POS tagging, and named entity

recognition, which yields accurate counts of properly segmented words; and (2) **Occurrence frequency**—counting each word's appearances in text using string matching, which is more intuitive and consistent with common search methods. Users can freely choose between these two options.

Additionally, editors noted that algorithm-extracted words were too short to convey full meaning, and that some visualizations suit longer words while others suit shorter ones. To address this, we provide a short-word filtering method. Short-word length is set through system defaults or user specification—words equal to or shorter than this length are considered short words, while longer ones are long words. If a short word appears within a long word, the short word is removed from the candidate set. This preserves long-word effects while ensuring important short words aren' t filtered, an approach validated by user feedback.

4. Results and Future Improvements

Through preliminary research, product design, independent R&D, and algorithm optimization, the project team rapidly completed coding and efficient iterative development, achieving the construction and optimization of the entire word cloud tool. Using this tool, journalists and editors produced numerous outstanding works during major 2021 reporting events such as the Spring Festival and Two Sessions, affirming the tool' s value while providing valuable feedback and suggestions.

Future improvements will focus on optimizing user experience, providing more flexible customization settings, and offering more refined styling options. Through continuous refinement, the product will continue to evolve.

5. Conclusion

Through independent R&D targeting news scenarios, combined with business analysis and improvements, we have developed an online word cloud tool that meets news reporting requirements and effectively assists editors in rapidly creating data journalism. This foundation will support the exploration of additional news data visualization tools.

References: [1] Leland Wilkinson. The Grammar of Graphics[M]. Springer,
[2] Jonathan Feinberg. Beautiful Visualization[M]. O'Reilly Media, 2010: 37-58.
[3] 朴灵. 深入浅出 Node.js[M]. 北京：人民邮电出版社,
[4] Hasan K S, Ng V. Automatic Keyphrase Extraction: A Survey of the State of the Art[C]. Meeting of the Association for Computational Linguistics. 2011.
[5] Page, L & Brin, S & Motwani, R & Winograd, T. The PageRank citation ranking: Bringing order to the Web. Technical report.

Author Biographies:
Qin Yufang (1987-), female, from Jiaozuo, Henan, Engineer, research direction: data journalism;

ChinaRxiv 中X

Li Ruonan (1993-), female, from Helan, Ningxia, Engineer, research direction: data journalism;
Liu Yingxu (1988-), female, from Yulin, Shaanxi, Engineer, research direction: data journalism.

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

Source: ChinaXiv —Machine translation. Verify with original.