Research on Visual Analysis of Popular Science Content Dissemination Hotspots

Xing Xu, Jiya Jiang(✉), Yonglian Wei, and Wei He

Science Communication Center,
Beijing Academy of Science and Technology, Beijing 100089, China
{xuxing,jiangjiya}@bjast.ac.cn

Abstract. In this paper, crawler technology is used to obtain data sets, and after automatic classification, visualization technology is used to analyze and study popular science content. The crawler technology uses the Python web crawler library, the automatic classification algorithm is based on Poisson distribution of Bayesian algorithm, and the visualization is realized through tableau. This paper observes and analyzes the data sets within the selected range from a new perspective, text clustering and research the digital dissemination of popular science content by hot spot analysis of the distribution of popular science news, so as to provide decision-making service reference for science popularization workers.

Keywords: Web crawler · Bayesian algorithm · Hot Spot Analysis · scientific popularization

1 Introduction

Xi Jinping delivered an important speech at the National Science and Technology Innovation Conference, the Academician Conference of the Chinese Academy of Sciences and the Ninth National Congress of the Chinese Association for Science and Technology. It systematically and comprehensively explains the "Two-Wings Theory", and points out that scientific and technological innovation and scientific popularization are the two wings to achieve innovation and development, and scientific popularization should be placed in the same important position as scientific and technological innovation. In September 2022, the General Office of the CPC Central Committee and the General Office of the State Council issued the Opinions on Further Strengthening the Popularization of Science and Technology in the New Era, which proposed the digital dissemination of popular science content. In order to strengthen the analysis of popular science content dissemination hotspots, this paper uses data visualization [1]for analysis and research.

Visualization technology is a very effective means of data understanding, which was first applied in the field of scientific and engineering computing, and has now developed into a very hot research field.In short, visualization is to express complex information or laws in the form of graphical symbols, so that people can quickly obtain the key information contained in the data [2]. Visual analysis technology is the integration of

C. F. Peng et al. (Eds.): EIMT 2023, AHSSEH 8, pp. 609–614, 2023.
https://doi.org/10.2991/978-94-6463-192-0_79

visualization, data mining, natural language processing and other analysis technologies, it is a process of mining valuable information from a large number of complex data, analyzing and reasoning, and understanding the internal structure and correlation of data [3]. This paper applies visual analysis technology to the analysis of digital dissemination and distribution of popular science content, and studies the digital communication of popular science content by observing and analyzing the data sets within the selected range from a new perspective, text clustering, and hot spot analysis of popular science news distribution, so as to provide decision-making service reference for popular science workers.

2 Dataset Acquisition Based on Crawler Technology

Web crawler also referred to as Internet robots, and it can get data and information in the internet automatically through its own rules. Compared with the traditional manual data collection, Web crawler can avoid the low efficiency, it can avoid high cost and cumbersome problems[4]. There are many programming languages for crawlers, Python is one of the mainstreams, and the Python toolbox covers the common library and external tools for data crawlers, data analysis, deep learning and so on. It provide a large number of libraries for realizing web crawler [5, 6]. Batch data processing can be applied to many application scenarios in the era of big data[7–12]. In this paper, firstly, we use Python's crawler framework crawley for Crawling. Secondly, we use the requests to implement the HTTP request operation. Thirdly, we use beautifulsoup to extract information from the web page. At last, we use PymysqlPython interact with the database, stores the parsed valid data into database MySQL. After this steps, we complete data acquisition.

3 Data Clustering Based on Bayesian Algorithm and Poisson Distribution

Data cluster based on Bayesian algorithm and Poisson distribution.

Dividing the training materials, removes the word frequency statistics in the training set classification. After the training is completed, you can automatically classify the article of the new storage, and you can keep the classification accuracy. This paper uses a Bayesian classification algorithm based on Poisson distribution.

The classification task is to find a class that makes a given instance with the maximum probability. Consider an example $\boldsymbol{B}$=$\lfloor B_1, B_2, \cdots, B_n \rfloor$ with N attributes, this example is a probability of a class A_i is $P(A = A_i|B_1 = b_1 \Lambda \cdots B_n = b_n)$.

According to Bayes formula, it's

$$P(A_i|\boldsymbol{B}) = \frac{P(B_1 = b_1 \Lambda \cdots B_n = b_n|A = A_i) * P(A = A_i)}{P(B_1 = b_1 \Lambda \cdots B_n = b_n)} = \frac{P(\boldsymbol{B}|A_i) * P(A_i)}{P(\boldsymbol{B})}$$

Among them, $P(\boldsymbol{B})$ is independent of A_i, it can be seen as a constant, which can be obtained from the training data. In order to find the probability $P(A_i|\boldsymbol{B})$, the $P(\boldsymbol{B}|A_i)$ is first requested.

In fact, it doesn't know the exact distribution of the instance, and it is necessary to construct the density function approximation condition probability $P(\boldsymbol{B}|A_i)$. This paper uses Poisson distribution approaching it.

Set one-dimensional random variable X obedience to Poisson distribution of paramete λ

$$P(X = k) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}, \lambda > 0, \mathrm{k} = 0,1,2,\cdots,$$

Assuming that class property is distributed from Poisson distribution, you can get a Bayesian classifier. The corresponding classification criterion is: $P(\boldsymbol{B}|A_i) * P(A_i) > P\left(\boldsymbol{B}|A_j\right) * P\left(A_j\right) P(\boldsymbol{B}|A_i) * P(A_i) > P\left(\boldsymbol{B}|A_j\right) * P\left(A_j\right)$.

This paper trains the information on the Internet based on Poisson distribution of Bayesian algorithm, and then continuous learning in applications, and has high classification accuracy.

4 Data Visualization Technology Based on Image Processing

This paper uses data visualization technology for image processing, data visualization analysis technology is a branch of data visualization, which is based on the image form planning and analysis process, and can strengthen the ability of data mining analysis and display effect. For popular science data, the same data is presented in different image forms around the same data. we interpret data from different perspectives in different communication platforms. In all, we should let the key information be displayed in the most suitable form and let the audience accept it more effectively.

The image processing can be applied to three stages of the information analysis process. Firstly, complete data mining and data cleaning.Secondly, use various professional tools to express the science popularization data in a static, dynamic and product-based way. Thirdly, analyze and implement to verify its communication effect, as shown in Fig. 1.

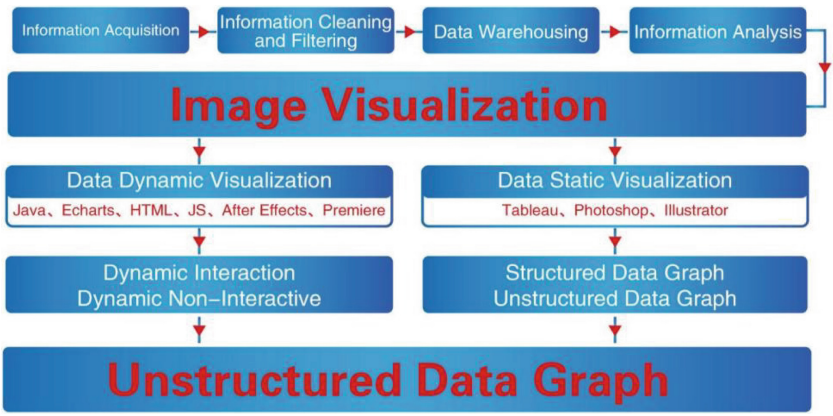

Fig. 1. Technology roadmap

5 Example Analysis

Grab the key words of the name of the author's unit Beijing Academy of Science and Technology and popular science on the Internet through crawler technology, capture the dynamics of science popularization related activities and articles in 2022, A total of 99417 articles in the media types which were initially screened manually, and 39781 articles were de-duplicated, the distribution of hotspots can be seen from the following figure by clustering with Bayesian algorithm, the analysis results is shown in Fig. 2.

Through statistical analysis, the number of reprints can be calculated, and the number of reprints can be used to mine hot news. For example 'The maximum full moon of the year! The super moon and other wonderful celestial phenomena will be staged in July' has been forwarded 469 times by various media such as People's Daily, Sina, Tencent and WeChat, the following figure shows the media reprinting situation, reprinting trend and reprinting peak of the article, the analysis results is shown in Fig. 3.

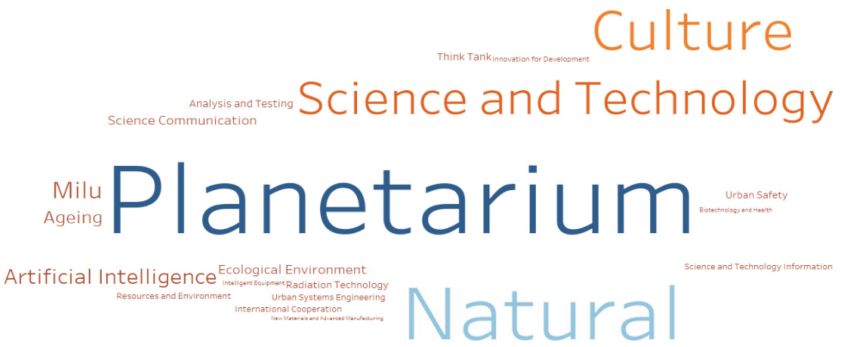

Fig. 2. Distribution of hotspots

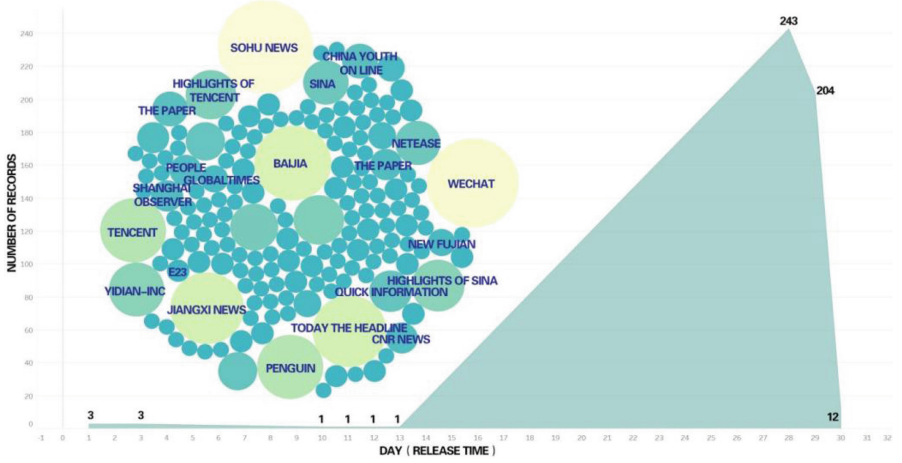

Fig. 3. Media reprinting situation and trend

6 Conclusion

According to the data analysis results, in 2022, Beijing Academy of Science and Technology reported a large number of popular science content, however, in terms of media types, number of reports and reprints, 92.95% of popular science content is concentrated in popular science units and bases, in the future, we should further give full play to the advantages of combining scientific and technological innovation with scientific popularization, carry out popularization of science and technology resources and strengthen the dissemination of scientific knowledge in scientific research institutions.

In the way of spreading popular science content, we should create and distribute adaptive content according to the characteristics of websites and We-media platforms, and form a diversified and differentiated media matrix, to optimize the integration and linkage effect of multiple platforms, expand the breadth, strength and depth of scientific communication. For example, We-Media for Science Popularization can use short video platforms such as Tiktok and Kwai to produce interesting short videos for science popularization, and present abstract knowledge concretely; At the same time, we can also use WeChat Channels, China Media Group Mobile and other platforms to broadcast live, and realize real-time interaction of scientific communication. In addition, While realizing the full coverage of the audience's information needs, we should also avoid the homogenization of information, so as to produce more hot information.

References

1. CHEN W, SHEN Z Q, TAO Y B. Data visualization[M]. 2nd edition. Beijing: Publishing House of Electronics Industry, 2019: 2-5.
2. Zeng You. The Concept Study of Data Visualization Under the Ackground of Big Data[D]. Hangzhou: Zhejiang University, 2014.
3. Keim D A, Mansmann F, Schneidewind J, et al. Challenges in Visual Data Analy-sis[C]//Tenth International Conference on Information Visualization(IV'06). London: IEEE, 2006: 9–16.
4. TANG S, CHEN Z Q. Python Web Crawler from Introduction to Practice[M]. Beijing: Chi-na Machine Press, 2017.
5. HU S T. Python Web Crawler Practice [M]. Beijing: Tsinghua University Press, 2017.
6. TOMORROW'S TECHNOLOGY, Python goes from beginner to proficient[M]. Beijing: Tsinghua University Press, 2018.
7. Tsourakakis CE. Fast counting of triangles in large real networks without counting: Algorithms and Laws, 2008. 608 617. doi: https://doi.org/10.1109/ICDM.2008.72.
8. Chen Y, Alspaugh S, Katz R. Interactive analytical processing in big data systems: A cross-industry study of MapReduce workloads. Proc. of the VLDB Endowment, 2012,5(12):180213. doi: https://doi.org/10.14778/2367502.2367519.
9. Stupar A, Michel S, Schenkel R. RankReduce-Processing k-nearest neighbor queries on top of MapReduce. Large-Scale Distributed Systems for Information Retrieval, 2010. 13 18.
10. Zhou MQ, Zhang R, Xie W, Qian WN, Zhou AY. Security and privacy in cloud computing: A survey. IEEE, 2010. 105 112. doi: https://doi.org/10.1109/SKG.2010.19.
11. Feblowitz J. Analytics in oil and gas: The big deal about big data. In: Proc. of the SPE Digital Energy Conf. 2013. doi: https://doi.org/10.2118/163717-MS.
12. Yu H, Wang D. Research and implementation of massive health care data management and analysis based on hadoop. IEEE, 2012. 514 517. doi: https://doi.org/10.1109/ICCIS.2012.225.