Study on the Frequency of Computer Language Use Based on Big Data Analysis

Boyang Liu[1,a,*]

[1]Operation Department, ONUS Global Fulfilment Solutions, Richmond V6W 1G3, British Columbia, Canada
[a]p7908686@gmail.com
*Corresponding author

Abstract: In today's digital age, computer language plays a vital role as the cornerstone of information technology. With the rapid iteration and update of technology, the types of computer languages and application scenarios have become increasingly diversified. Different programming languages, with their unique characteristics, serve a wide range of applications from low-level system development to advanced data analysis. Along the way, developers are constantly creating new programming languages to meet diverse needs. Some languages focus on specific domains, while others are widely used in multiple domains. Based on IEEESpectrum's latest ranking, this article will provide an in-depth analysis of the current usage of major programming languages in areas such as data science and cloud development. At the same time, through big data analysis, the use frequency of computer language and its changing trend are analyzed comprehensively. We will look at multiple dimensions, including global language usage, industry and domain differences, and regional language usage characteristics. Through this analysis, we hope to provide valuable references for developers, business decision makers, and academic researchers to help them make more informed decisions when choosing programming languages and developing technology strategies.

Keywords: Computer language, Frequency of use, Big data, Digitization

1. Introduction

With the rapid development of information technology, the importance of computer language is increasing day by day. Different computer languages are used at different frequencies in different domains and application scenarios, and these changes in frequency reflect technology trends, industry needs, and the dynamics of the developer community. Different programming languages, with their unique characteristics, serve a wide range of applications from low-level system development to advanced data analysis. The popularity of computer languages is not static, with the emergence of new technologies and changes in industry needs, the popularity of programming languages is also constantly changing. Python has quickly risen to become one of the most popular programming languages in the world due to its powerful capabilities in data science and artificial intelligence. Traditional programming languages such as Java and C++ still play an important role in enterprise applications, but their relative popularity has shown a decline. Such changes are not only the evolution of the technology itself, but also reflect the dynamic changes of the entire technology ecosystem. Analyzing and understanding the frequency of use of these languages is not only important for technical practitioners, but also for business decision making. The rise of big data technology provides a new method and perspective for analyzing the frequency of use of computer language. Through the analysis of massive data, it can reveal the global use trend of different computer languages, analyze their application in different industries and different fields, and predict the future direction of technology development. Compared with the traditional market research and expert evaluation, this analysis method is more objective, accurate, and has strong timeliness. With the continuous evolution of computer language, the study of its usage frequency not only helps us to understand the current status of technology ecology, but also provides an important basis for predicting the future technology trend [1].

2. Study on the Frequency of Computer Language Use in the Context of Big Data

2.1. Data Source

To study the frequency of computer language use in the context of big data, it is first necessary to identify and collect high-quality data sources to ensure the comprehensiveness and reliability of the analysis. The frequency of use of programming languages usually depends on the joint analysis of multiple data sources, which can comprehensively cover the use of programming languages in actual development, learning, discussion and other aspects. As the world's largest code hosting platform, the number of projects and developer activity on GitHub can directly reflect the actual use of a certain programming language. By analyzing the number of repositories, commits, and developer activity of different languages on GitHub, we can get a first-hand look at how frequently the language is being used. Stack Overflow is the world's most influential programming Q&A community where developers share problems and solutions. By analyzing the number of questions related to different programming languages, the frequency of tag usage, the speed of response to questions, and the resolution rate, we can indirectly reflect the popularity and breadth of use of the language in the developer community. TIOBE Index and PYPL (PopularitY of Programming Language) index rank the popularity of programming languages worldwide based on Internet search engine data. We also analyze relevant academic papers and technical literature, and the application of certain programming languages in specific fields can be measured by the number of academic papers cited and technical reports published [2].

2.2. Data Processing

After data collection, one is data cleaning. Remove noise and outliers from the data to ensure that only valid data is retained. In GitHub data, you need to remove empty items, duplicate items, and test code bases. In Stack Overflow data, you need to filter out duplicate questions, invalid responses, and language-neutral discussions. The second is standardized treatment. Since the data comes from different platforms, there may be significant differences in the size and presentation of the data from each platform, so it is necessary to standardize the data. The number of code commits on GitHub and the number of questions on Stack Overflow are normalized to ensure that these data are comparable when analyzed. The third is time series analysis. In order to identify the time trend of programming language usage frequency, time series analysis must be performed. This involves grouping data into time dimensions, analyzing trends across programming languages over time, and identifying patterns of growth, leveling off, or decline. Fourth, cluster analysis. When dealing with large amounts of data, cluster analysis is an effective means to help identify groups of programming languages with similar characteristics. Cluster languages by application area, developer community size, or market share to discover language groups with similar characteristics and their positioning in the market [3].

3. Analysis of Geographical Distribution of Programming Language Usage Frequency

The frequency of programming language usage is not only influenced by technology trends and industry needs, but also closely related to geographic regions. Factors such as the level of economic development, technological industry structure, education system and language culture in different regions have an important impact on the popularity of programming languages. By analyzing the geographic distribution of programming languages on a global scale, big data can reveal regional preferences for specific programming languages, which has important implications for understanding the global technology ecosystem.

3.1. Data Sources and Analysis Methods

For the geographic distribution analysis, this article relies on GitHub open source project data, Stack Overflow Q&A data, Google Trends search data, and data from the popularity rankings of programming languages such as TIOBE and PYPL. This data covers the behavior patterns and interest tendencies of developers around the world, which can effectively reflect the use of programming languages in different regions. In order to carry out the geographical distribution analysis, it is necessary to mark the geographical location of the data, and determine the geographical attribution of the data combined with the registration place and IP address of the developer. Through data cluster analysis and thermal map visualization, the frequency of programming languages used in various

regions of the world is shown [4].

3.2. Analysis of Programming Language Usage Preferences by Region

In technologically developed regions such as North America, Europe and Asia, the use of programming languages has a strong diversity. These regions have developed technology industries and active developer communities, and mainstream programming languages such as Python, Java, and JavaScript are widely used. In the United States, for example, Python is one of the most popular languages due to its extensive use in data science and artificial intelligence, while Java is prominent in enterprise-level application development.

In Europe, especially Western European countries, JavaScript and TypeScript are popular due to their importance in Web development. Rust is gaining traction among developers in countries such as Germany, mainly because of its advantages in systems programming and high performance computing.

Programming language usage in Asia shows some regional differences. China and India, as countries with the largest number of developers in the world, have a wide range of applications for Python and Java, with Python especially popular in the field of data analysis and machine learning. In Japan, Ruby still holds a market share due to its efficiency and flexibility in Web development. The Korean developer community has shown a preference for C++ and C#, mainly in game development and embedded systems.

In emerging markets such as South America and Africa, Java and PHP are used more frequently. Due to the relative lack of IT infrastructure and educational resources in these regions, developers tend to choose programming languages with a flatter learning curve and strong community support. Java, as a language with a wide range of application areas, has been favored by many developers, while PHP is widely used by small and medium-sized enterprises because of its simplicity and cost effectiveness in Web development [5].

3.3. Research on the Correlation between Languages and Regions

The popularity of certain programming languages is closely related to the structure of the technology industry in a region. For example, technology companies in North America and Europe tend to adopt JavaScript and Python because the Internet and technology companies in these regions are more developed and the related technology ecosystems are more mature. In contrast, in some manufacturing-dependent countries, such as Japan and South Korea, C/C++ and Java are used more frequently, which is closely related to the application of these languages in the field of embedded systems and industrial control [6].

The geographic distribution of programming languages reflects the diversity and complexity of the global technology ecology. Understanding programming language usage preferences in different regions not only helps developers and businesses make more accurate technical decisions, but also provides valuable reference data for educational institutions and policy makers. In the future, with the globalization of technology and the rise of emerging markets, the geographical distribution pattern of programming languages will continue to evolve.

4. Analysis of Factors Affecting the Popularity of Programming Languages

4.1. Developer Community and Support

The size and activity of the developer community is an important factor in the popularity of a programming language. A large and active developer community usually means rich online resources, frequent discussions, quick technical support, and a large number of open source projects, all of which help to reduce the cost of learning and difficulty of development for developers. The rise of Python, for example, is largely due to its large developer community, which not only has a large number of open source libraries and frameworks, such as TensorFlow, Django, etc., but also many high-quality tutorials and documentation to help beginners get started quickly. The community-driven rate of language renewal and evolution also plays a key role in the language's vitality. Active communities tend to be quick to respond to developer needs and iterate and optimize the language. For example, JavaScript has experienced significant growth over the past few years, thanks to the promotion of Node.js and a large number of community-contributed libraries such as React and Vue.js, making its position in front-end

development more secure [7].

4.2. Ecosystems and Tool Chains

The degree of perfection of a language ecosystem and tool chain directly affects the efficiency and scope of application of the language in practical development. A well-developed ecosystem typically includes rich libraries, frameworks, development tools, integrated development environments (ides), etc., all of which can significantly increase developer productivity. One of the reasons why Java dominates enterprise development is its mature ecosystem, which covers the complete toolchain from the Spring framework to Maven build tools, from Jenkins continuous integration to Docker containerization deployment, greatly facilitating all aspects of development and operations. Python's popularity in data science is also due to its powerful data processing and analysis libraries such as Pandas, NumPy, SciPy, as well as machine learning libraries Scikit-learn and deep learning frameworks such as TensorFlow and PyTorch. The richness of these toolchains makes Python the language of choice for data scientists and AI researchers.

4.3. Language Learning Curve and Development Efficiency

Learning curve and development efficiency are important considerations when choosing a programming language. The learning curve refers to the time and effort it takes for a developer to go from beginner to proficient in a language. The smoother the learning curve, the easier it is for beginners to accept and master a language. Python's simple syntax and high readability make it the language of choice to get started in programming, especially among developers with non-computer backgrounds. In contrast, C++, while powerful, has a steep learning curve due to its complex syntax and memory management mechanisms, which limits its widespread use in some application scenarios. Development efficiency is also an important criterion for evaluating the popularity of a language. Development efficiency includes not only coding time, but also debugging, testing, and maintenance. Frameworks like Ruby on Rails enhance Ruby's appeal by significantly improving development efficiency through the "convention over configuration" design philosophy [8].

The popularity of programming languages is influenced by many factors, forming a complex dynamic system. The support of the developer community, the integrity of the toolchain and ecosystem, and the learning curve and development efficiency of the language all influence the frequency and popularity of programming languages to varying degrees. In the future, with the continuous progress of technology and the diversification of application scenarios, the importance of these factors will also change, which will affect the popularity of programming languages.

5. Forecast and Summary of Future Trends

With the continuous progress of science and technology and the wide application of information technology, the ecosystem of programming languages is constantly changing. The frequency with which a programming language is used is influenced by many factors. Artificial intelligence (AI) and machine learning (ML) are one of them. As the most popular technology field in recent years. These areas have distinct requirements for programming languages that can support large amounts of data processing, high-performance computing, and libraries and frameworks that are easy to integrate. Python has become the language of choice for data science and AI because of its rich libraries (such as TensorFlow, PyTorch) and easy to learn syntax. As the requirements of AI and ML become more complex, other programming languages will also emerge in specific areas. Due to its excellent numerical computing capabilities and high performance characteristics, Julia is expected to occupy a place in future AI applications.

With the diversified development of technology, the coexistence of multiple languages has become an irreversible trend. Developers are no longer limited to using a single programming language, but choose the most appropriate combination of languages according to the needs of the project. Front-end development typically uses JavaScript and its frameworks (such as React, Vue), while the back-end uses Python, Java, or Go for business logic and data processing. The rise of cross-platform development tools, such as Flutter and React Native, has also enabled developers to write cross-platform applications in one language, further driving the trend of multilingual coexistence. The popularity of a programming language also depends on how well its toolchain and ecosystem are developed. A robust ecosystem includes not only the language itself, but also support for compilers,

debugging tools, integrated development environments (ides), libraries, and frameworks. In the future, languages with well-established ecosystems will become more popular.

We believe that developers should remain sensitive to emerging programming languages and continue to learn to adapt to rapid changes in technology. For enterprises, it is important to be flexible in choosing languages based on project needs and to invest in improving the multilingual development capabilities of teams. Educational institutions should adjust the curriculum to emphasize cross-language programming ability and problem solving ability in multi-language environment, and cultivate versatile talents to meet the needs of future technology. The future of programming languages will show a trend of diversification and specialization. With a deep understanding of these trends, developers, businesses, and educational institutions can all be well-positioned to ride the next wave of technology.

References

[1] Xia Gao. Research on the frequency of computer language use based on big Data analysis. Information and Communication, 2022 (035-003).

[2] Qinghai Bai. Based on analysis of large data computer language use frequency research. Journal of Inner Mongolia national university: natural science edition, 2019, 34 (6): 4. DOI: CNKI: SUN: NMMS. 0.2019-06-007.

[3] Huang L, Zhang Y. Research on Application of Computer Artificial Intelligence Technology in Sports Economic Big Data Analysis Highlights in Science, Engineering and Technology, 2022. The DOI: 10.54097 / hset v16i. 2593.

[4] Xiaoyan Liu. Research on night image enhancement method based on Python. Natural Science, 2019, 037 (001): 85-89.

[5] Yuanyuan Wang, Bu Fanliang. Simple Application of digital image processing based on Python. Electronic Technology and Software Engineering, 2022 (003): 56-68.

[6] Jiming Du. Analysis of Web development application of Python language. Digital Communication World, 2022 (001): 89-97.

[7] Jianing Li. Research on the development and application of computer programming language. Digital Design (I), 2021, 010 (003): 2-3.

[8] Zhiyun Zhang. Research on the application of Java programming language in Computer Software development. Information Recording Materials, 2023, 24 (4): 110-112.