Mining the Trend of the Most Frequently Used Programming Language and Tools of the Last Decade

Ning Chih Chang
College of Engineering and Applied Science
University of Colorado Boulder
Boulder, CO, USA
nich1985@colorado.edu

1 Motivation

As a computer science student, knowing the programming languages and development tools that have been frequently used in industry for the past years is important because it would help me understand what skillsets I would need to gain to prepare myself for the workforce. I am interested to know the trend of the most used programming languages and tools in the past 10 years. Is there a particular collection of programming language and tools that have been frequently used by a certain demographic? Knowing the trend of programming languages and tools would not only benefit students and professionals to keep track of the latest technology for career development, but also would create opportunities for businesses and educational institutions in the software and programming field.

2 Literature Survey

There are some well-established indicators for the most popular languages on the web that are free to the public. The TIOBE index gives the most popular 100 programming languages which is updated monthly and is based on the number of hits on various websites [9, 14]. The IEEE Spectrum provides a similar indicator that is generated from the popularity among the IEEE members and developers, the employer demands and the current trend [2, 3]. However, both indicators currently do not cover programming tools such as development environment or frameworks.

Studies have been done in finding trends in programming tools using Stack Overflow posts. Approximately 11 million user question and answer posts of Stack Overflow have been used to find the most popular languages, tools and topic trends between 2014 and 2015 by topic modeling [10]. Trends of NoSQL database from 2008 to 2017 have been investigated by using the normal interest score of Stack Overflow posts [6]. Moreover, posts on Stack Overflow that are related to C, Java and Python over a 14-year period have been studied and interestingly the user's country and reputation are included in trend analysis [11].

Surveys have also been used to find trends. The Stack Overflow Developer Survey has been used in investigating gender insights [12] and finding the most used programming languages and tools for 23 different IT roles [5]. Interestingly, a study that collected data from surveys and multiple sources from schools and industries in 1993, 1998 and 2003 formed a regression model that illustrated the trend [4].

I would like to explore the trends of the most frequently used programming language and tools with demographic data to see if there are any interesting patterns. The Apriori algorithm can be used to mine frequent itemsets [7] and has been applied to find trends in major selection among university students [1] and frequent patterns in human drug addiction behavior [8]. This would be an exciting tool to use in this study to find hidden patterns.

3 Data Set

The data set of this study is retrieved from Stack Overflow Annual Developer Survey webpage [13] and surveys from 2015 to 2024 are included, which is 10 sets in total.

The number of responses for each survey from 2015 to 2024 are 26086, 56030, 51392, 98855, 88863, 64461, 83439, 73268, 89184 and 65437, respectively. The number of attributes for each survey from 2015 to 2024 are 45, 65, 153, 128, 84, 60, 40, 70, 68 and 87, respectively.

The attributes among these surveys are not exactly the same, but there are nine shared concepts, which are occupation/devtype(nominal), country(nominal), years of coding/programming experience(ordinal/ratio), compensation/salary(ordinal/ratio), education(nominal), languages worked in the past year(nominal), languages want to work in the next year(nominal), tools worked in the past year(nominal) and tools want to work in the next year(nominal).

4 Proposed Work

My research plan is similar to the study that analyzed the most used languages and technologies to the 2020 Stack Overflow Developer Survey [5], yet it will be different because this study includes data across 10 years as well as exploring frequent patterns.

Since we are combining 10 sets of surveys, we need to make sure that only the shared attributes are included, and they should be consistent among surveys. Data preprocessing is crucial and going to be challenging in our case.

4.1 Data Cleaning

Missing data has been found in the occupation/devtype attribute and may appear in other attributes as well. If the total number of missing data in occupation/devtype is relatively small, then dropping the missing data would be considered. If not, then we may need to look into other options (e.g., replace with "NA") to manage missing data.

There are some "NA" values in the compensation/salary attribute. It would be logical if the compensation is not applicable if the respondent is a student, however there may be cases where the respondent does not want to reveal this information. However, we cannot assume that the "NA" response equal to $0 compensation for students. Therefore, if the total number of "NA" values is significant, then we need to take caution when explaining the results.

4.2 Data Transformation

Data transformation needs to be performed on the shared attributes before data integration. Although all the surveys have the compensation/salary and the years of coding/ programming experience attributes, some of the surveys ask for numeric response yet some provide a selection of ranges (i.e., ordinal) for respondents. To have a consistent attribute type for the integrated data set, there will be data discretization performed.

Moreover, even if the share attribute is ordinal, the selections may not be in the same size. For example, the answer selections of the years of coding/programming experience attribute are "less than 1 year", "1-2 years", "2-5 years", "6-10 years" and "11+ years" in the 2015's survey. However, in the 2018's survey, the available answers are "0-2 years", "3-5 years", "6-8 years" and continue every two years until "30+ years". Therefore, data transformation is needed to ensure the interval labels of the same attribute are consistent among surveys.

There are multiple attributes (e.g., framework and development environment) in each survey that can be joined into a new attribute (e.g., programming tools). Therefore, there will be new attribute creation and in general, the final attributes will represent the nine shared concepts mentioned above.

4.3 Data Integration

After ensuring that the shared attributes are consistent across 10 surveys, the shared attributes will be combined into one large data set and the rest of the attributes will be discarded.

4.4 Derive The Most Used Language

To see the trend of the most used language in the past 10 years, we can retrieve the programming language with the highest frequency by year from the data set. Furthermore, we can dig deeper by looking at different occupation/devtype, years of coding/programming, compensation, education and country.

4.5 Derive The Most Used Tools

Programming tools with the highest frequency can be calculated by year to observe the trend of the most used tools in the past 10 years. Additionally, we can examine this with different occupation/devtype, years of coding/programming, compensation, education and country.

4.6 Derive Frequent Patterns with Apriori

To find the frequent patterns, each respondent's language, tools, occupation/devtype, years of coding/programming, compensation, education and country data will be combined into an item. The frequent itemsets can be found after running all items with the Apriori algorithm.

5 Evaluation Methods

There will be limited validation methods for this study because the aim of this study is not establishing a prediction model, which means it would not be logical to reserve partial data for validation. The frequent patterns that we will find using Apriori may overlap with the most frequent languages and tools using descriptive statistics, which may serve some degree of validation. Also, frequency histograms can be generated to check the accuracy of the result.

6 Tools

The Python language and Pandas, NumPy and Matplotlib libraries will be used to manage and perform statistic analysis on large datasets as well as to provide visualization of results.

7 Milestones

Data collection and search for common attributes have been done. I plan to finish data cleaning, data transformation and data integration before October 28th, to obtain the most used language and tools by November 4th, and to complete the frequent patterns by November 11th.

REFERENCES

[1] Almahdi Alshareef, Salem Ahmida, Azuraliza Abu Bakar, Abdul Razak Hamdan, and Mohammed Alweshah. 2015. Mining survey data on university students to determine trends in the selection of majors. In *2015 Science and Information Conference (SAI).* 586-590. https://doi.org/10.1109/SAI.2015.7237202

[2] Stephen Cass. 2024. The Top Programming Languages 2024: Typescript and Rust are among the rising stars. (August 2024). Retrieved October 14, 2024 from https://spectrum.ieee.org/top-programming-languages-2024

[3] Stephen Cass. 2024. The Top Programming Languages Methodology 2024: How we construct the rankings. (August 2024). Retrieved October 14, 2024 from https://spectrum.ieee.org/top-programming-languages-methodology-2024

[4] Yaofei Chen, Rose Dios, Ali Mili, Lan Wu and Kefei Wang. 2005. An empirical study of programming language trends. *IEEE Software* 22, 3 (June 2005), 72-79. https://doi.org/10.1109/MS.2005.55

[5] Oluwaseun Alexander Dada, George Obaido, Ismaila Temitayo Sanusi, Kehinde Aruleba, and Abdullahi Abubakar Yunusa. 2023. Hidden gold for IT professionals, educators, and students: Insights from stack overflow survey. *IEEE Transactions on Computational Social Systems* 10, 2 (April 2023), 795-806. https://doi.org/10.1109/TCSS.2022.3151130

[6] Harshit Gujral, Abhinav Sharma and Parmeet Kaur. 2018. Empirical investigation of trends in NoSQL-based big-data solutions in the last decade. In *2018 Eleventh International Conference on Contemporary Computing (IC3)*. 1-3. https://doi.org/10.1109/IC3.2018.8530582

[7] Jiawei Han, Micheline Kamber, and Jian Pei. 2012. *Data Mining: Concepts and Techniques* (3rd. ed.). Morgan Kaufmann Publishers, Waltham, MA.

[8] Md. Mehedi Hassan, Sadika Zaman, Swarnali Mollick, Md. Mahedi Hassan, M. Raihan, Chetna Kaushal, and Rajat Bhardwaj. 2023. An efficient Apriori algorithm for frequent pattern in human intoxication data. *Innov. Syst. Softw. Eng.* 19, 1 (March 2023), 61-69. https://doi.org/10.1007/s11334-022-00523-w

[9] Paul Jansen. 2024. TIOBE index for October 2024. (Oct. 2024). Retrieved October 14, 2024 from https://www.tiobe.com/tiobe-index/

[10] Vishal Johri and Srividya Bansal. 2018. Identifying trends in technologies and programming languages using topic modeling. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*. 391-396. https://doi.org/10.1109/ICSC.2018.00078

[11] Yusuf Ade Putra Perdana and Yusuf Sulistyo Nugroho. 2024. An empirical study of the popularity of three programming languages on stack overflow: Trends, solutions, and users. *AIP Conf. Proc.* 2926, 1 (Jan. 2024), 020092-1-020092-8. https://doi.org/10.1063/5.0182877

[12] Karina Kohl Silveira, Soraia Musse, Isabel H. Manssour, Renata Vieira, and Rafael Prikladnicki. 2019. Confidence in programming skills: Gender insights from StackOverflow developers survey. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*. 234-235. https://doi.org/10.1109/ICSE-Companion.2019.00091

[13] Stack Overflow. 2024. Stack Overflow Annual Developer Survey. Retrieved October 14, 2024 from https://survey.stackoverflow.co/

[14] TIOBE Software BV. 2024. TIOBE Programming Community Index Definition. Retrieved October 14, 2024 from https://www.tiobe.com/tiobe-index/programminglanguages_definition/