GitViz: An Interactive Visualization System for Analyzing Development Trends in the Open-Source Software Community

Chanhee Park Sungjun Do Eunjeong Lee
Hanna Jang Sungchan Jung Hyunwoo Han Kyungwon Lee
Ajou University

ABSTRACT

This study proposes a visualization that can assist computer scientists and data scientists to make decisions by exploring technology trends. While it is important for them to understand the technology trends in the rapidly changing computer science and data science fields, it takes considerable time and knowledge to acquire good information about these trends. Particularly, data/computer scientists with little experience in the field find it difficult to obtain information on such trends. Therefore, we propose a visualization system that can easily and quickly explore the technology trends in computer and data science. This study aims to identify the key technologies and developers in a specific field, and other technologies deeply related to specific technologies, and explore the changes in popularity of technologies, languages, and libraries over time. This study includes two case studies to obtain information using the proposed visualization. We demonstrate our system with GitHub repositories data.

Keywords: Trends analysis, decision making, Radviz.

Index Terms: H.5.2 User Interfaces: graphical user interfaces.

1 INTRODUCTION

Computer science and data science fields are advancing rapidly. Nw technologies are emerging every day with their trends changing, every year. Identifying the changing trends is extremely important for developers in these areas. Developers require information on recent technological issues in the development field, the most influential developers in each field, and how their keywords of interest have changed over time. However, it is not easy for ordinary developers, especially technicians studying the field for the first time, to quickly understand the trends.

Companies and developers ahead of the trends release new technologies and their developments to the public through open source. Therefore, we visualize open source projects to solve these problems related to acquiring information on trends. The largest open platform for open source in the world is the 'GitHub' [10]. In this study, we design a system for analyzing technology trends in the computer/data science field by using open-source projects shared through GitHub.

* Chanhee Park, Sungjun Do, Eunjeong Lee, Hanna Jang, Sungchan Jung, Hyunwoo Han, and Kyungwon Lee are with the Ajou University. E-mail: {ch13p, kand148, dnswjd2340, jhn9592, tmdcks6628, ainatsumi, kwlee} @ajou.ac.kr

2 RELATED WORKS

Decision making sometimes requires the integration of information. For example, a student wanting to get a job needs to possess the skillset and industry knowledge related to a specific job. However, it is sometimes a challenge to integrate and explore information because of the diverse nature of information and lack of prior knowledge. There are certain studies providing solutions that aid in decision making. Du, Fan, et al. propose a visualization system [1] that identifies processes, similar to the ones in hand, to guide during decision making. This system assists in decision making by showing people's decisions and results in comparable situations. Liu et al. present JobViz [2], a circular visualization that provides students with the information required to land a job. It explores the connection between student's majors and the job postings and shows the skills and industry knowledge required for a specific job.

Some studies suggest ways to effectively explore researchers and research topics. Computer/data scientists obtain information from open sources. This is similar to how researchers pull information from previous studies. Kurosawa et al. [5] propose the visualization system for co-authorship networks. This study suggests a creative way to show an individual researchers' past and current interests using a pie chart. Lee et al. [4] suggest a radial visualization method to find interdisciplinary research keywords. This visualization identifies interdisciplinary research topics by placing disciplines on a circular axis and keywords within the circle. Heimerl et al. introduce CiteRivers [7], a visual analytics approach useful for the analysis of scientific literature. This visualization shows subject quantities and citation rankings with actual data in a subview. It explores how the popular subject and citation information are altered over time. Yang et al. propose VISTopic [8], an interactive visualization to help users discover knowledge with topical information, temporal information and bibliographic information simultaneously. This visualizes the hierarchical topic model through Sunburst.

3 DESIGN GOALS

The Purpose of this system is to provide easy and quick solutions for these users to explore the technology trends in computer and data science. We have set three visualization goals to develop effective interactive visualizations.

G1: Identify core technologies and key developers in a specific field.
G2: Identify other technologies relevant to a particular technology.
G3: Explore changes in popularity over time of technologies, languages, and libraries that interest major developers.

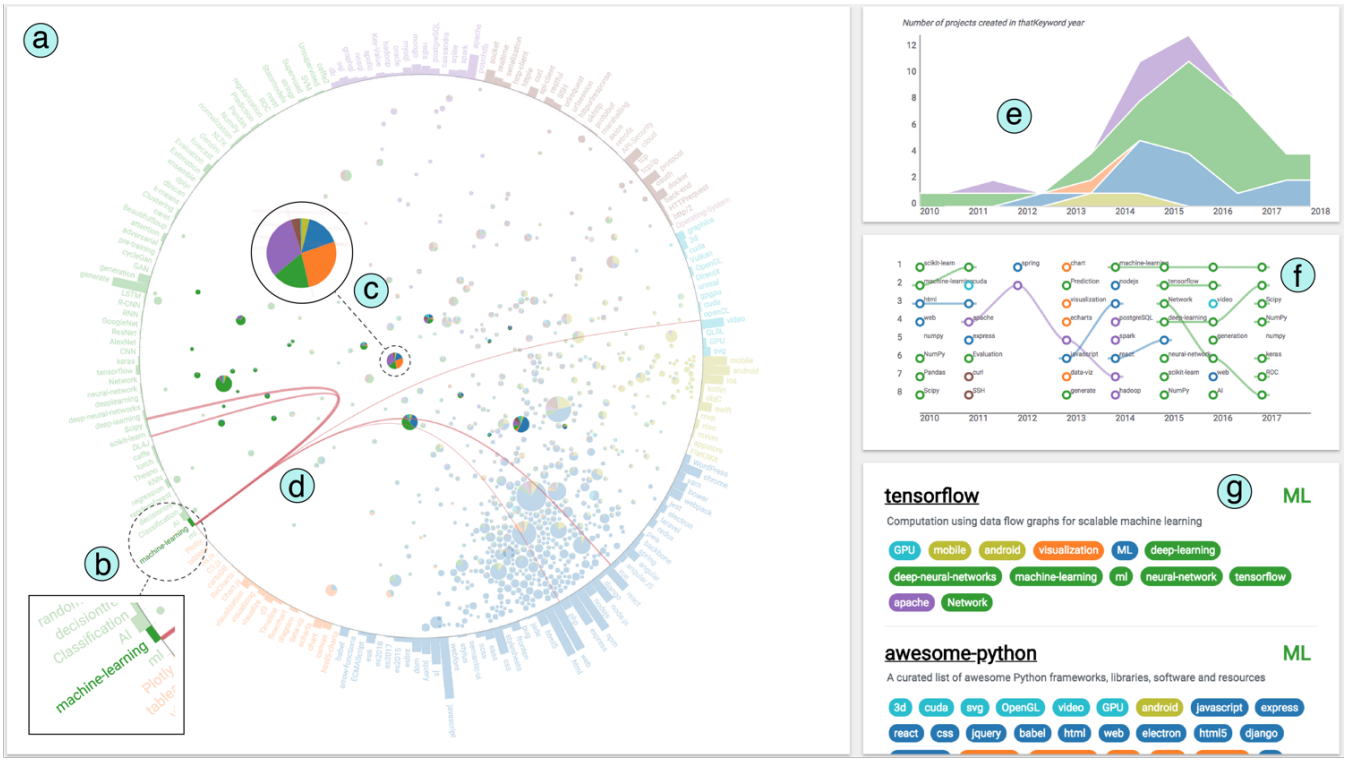

Figure 1: (a) In this visualization, Radviz uses the keyword as a dimension anchor and the GitHub user as an instance. (b) Keyword dimension anchors are displayed in different colors for each field and have a bar chart proportional to the number of related projects. (c) The colors of the user nodes are displayed in a pie chart according to the ratio of the fields. (d) If two keywords appear in the same repository, they are connected. (e) The cumulative area chart shows the change in the amount of generation of the repositories by year. (f) Changes in the keyword ranking of the repository are shown. (g) List of repositories related to the currently selected condition, in the order of popularity.

4 Visualization System

Our visualization shows the most popular 1000 repositories and their developers in GitHub. In this section, we describe the dataset and the five classifiers that are used to illustrate the visualization. Further, we propose a visualization system to analyze multiple classifiers.

4.1 Data

GitHub API is used for collecting data. In GitHub, a project is managed through a single repository. The user of the GitHub leaves a star to show interest in the repository or recommend a repository of similar themes. In other words, repositories with many stars are considered influential projects. This study uses 1000 repositories in GitHub that have the maximum number of stars and the details of the users who contributed to these repositories. We have collected the title, tag, description and readme.md of each repository, as well as information on users who contributed. If the repository is managed by a team or a company, we have collected that information as well. In this visualization system, we assume that a team or a company is a single user as long as it has one GitHub account.

The tags, descriptions, and readme.md of the repository often contain keywords that represent the fields and developer interests of the repository [9]. We use this data to identify related fields of GitHub users and repositories. We perform the following steps to find the computer/data science keywords used in GitHub. First, we count the number of occurrences of all words in the tag, description, and readme.md in the repository. After that, we pick out the words with high frequency of appearance. Finally, we use this as a keyword, if the word corresponds to a specific technology (e.g., TCP/IP), a programming language (e.g., C ++) or a library (e.g., tensorflow). Keywords are categorized into seven fields that include graphics, mobile, web, visualization, data analysis/machine learning, databases, and networks. These seven fields cover more than 95% of all the keywords.

4.2 Visualization and Interaction

Our visualization system provides four views in one screen. 1) Radviz view to find the key technology keywords and major developers. 2) A stacked area view showing the change in popularity of the development field over time. 3) Keyword ranking view that analyzes the popularity change of technology keywords over time. 4) A repository list view that shows the projects related to the selected keywords and developers. Users choose those keywords and developers from Radviz that they want to explore further. The other three views change as per the chosen keywords and developers. This system provides effective interaction and supports users to smartly explore technology trends.

4.2.1 Radviz

Radviz is a circular visualization based on the spring paradigm, which projects the data of N-dimensions into a plane [3]. In Radviz, the dimension of data, called the dimension anchor, is placed on a circle, in the form of an axis, and the dimension anchor is connected to the inner instance of the circle. The location of the instance is determined by the value of the connected dimension. Dimension anchors that are associated with an instance pull the instance. In this study, Radviz shows the relationship between technologies and GitHub users. Development-related keywords are used as dimension anchors and placed on a circle. Users are used as instances and placed inside the circle. The user instance is linked to its associated keyword dimension anchor thereby enabling

keywords to determine the location of the user instance. That is, the user instance is positioned around keywords associated with.

However, Radviz has some limitations [3]. If an instance is in the middle of a Radviz, it is difficult for Radviz to determine which one is associated with a particular dimension anchor. In addition, as the number of dimensions increases, it is difficult to locate the optimal dimension anchor and find meaningful dimensions. We have added three features to Radviz in order to overcome these limitations and solve problems such as 'find key developer' (G1) and 'analyze keyword relevance' (G2).

The first feature is color. Each field has a color. Different colors are assigned to each development field; For example, blue for the web and green for the data science. This makes easier to explore instances and dimensions [4]. A dimension anchor (keyword text) has its own field color. And an instance (user) are represented on a pie chart such that the percentage of each field in which a developer participates. It makes to identify how much the user has contributed to each field, even if the instance is in the middle of the radviz. In addition, the pie chart shows each developer's main field and other fields. It enables the following question: "What other areas of interest for developers interested in X?"

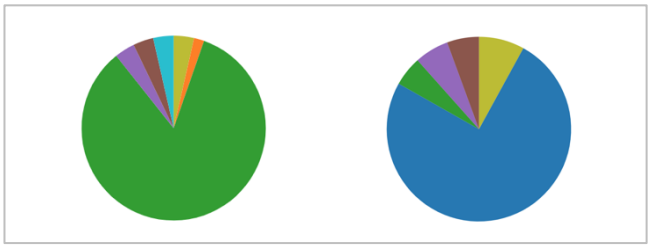

Figure 2: User node of tensorflow (left) and Facebook (right).

The second feature is size. The keyword dimension anchor is combined with a bar chart to show how many of the keywords came from each repository [2]. This shows the key technologies in the field. The user instance has a radius proportional to the user's popularity. Users with many stars show up large. It is useful to identify who is the most important GitHub user in each field.

The last feature is the connection. In this study, we have added connection lines between related dimension anchors and instances [2, 6]. This technique is mainly used to connect two of the nodes in a network visualization. We assume that if two keywords appear frequently in the same repository, they are related to each other. We link these two keyword dimension anchors with red lines. Likewise, when two users appear frequently in the same repository, we assume that they are related to each other. We link these two keyword dimension anchors with black lines. A black line represents collaboration between two developers. Both of these lines express the magnitude of association between the two connecting elements using the thickness of the line. This allows users to explore other keywords related to the ones they are interested in or look at the projects they collaborate with.

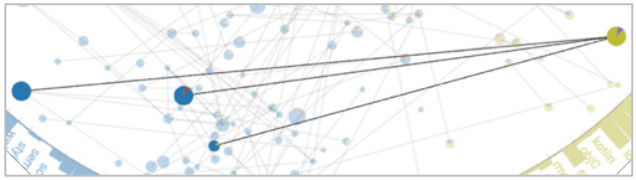

Figure 3: A Link between Web (blue nodes) and Mobile (a yellow node) developers.

For effective navigation, our system provides mouse-over interaction to user nodes. When the user moves the mouse over a node, it displays the name, popularity, and related keyword list, pertaining to the node. The system also provides selective interaction with nodes and keyword dimension anchors. The selected node and keyword dimension anchor can be explored using the three subviews.

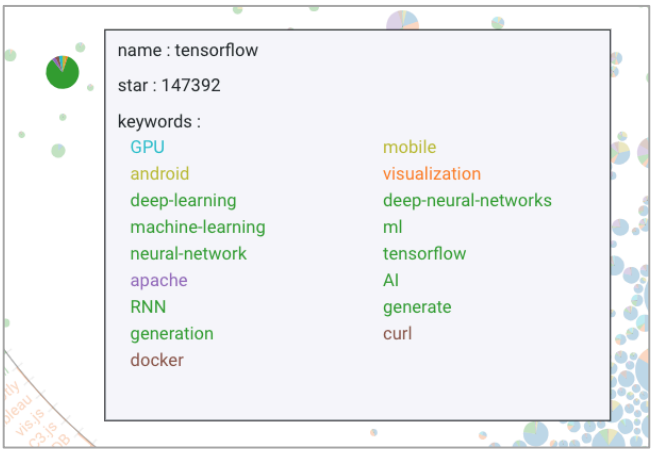

Figure 4: A Link between Web (blue nodes) and Mobile (a yellow node) developers.

4.2.2 Stacked Area Chart

The stacked area chart shows changes in creation amount of repositories. In the open source community, the popularity of the field is reflected in terms of the number of projects created. It shows the number of creation of all the repositories that are associated with the selected keywords and users. The horizontal axis represents time and the vertical axis represents the number of repositories created for the year. The color of the area is consistent with the field-specific color used in Radviz.

This view shows the change in popularity of each field at a glance. (G3) In particular, it answers specific questions such as "When did the selected developers start developing data science?" This chart is more effective when used with keyword ranking visualization.

4.2.3 Keyword Ranking Visualization

Keyword ranking visualization shows the change in the frequency of keyword appearances in all repositories related to selected keywords and users [7, 8]. Like the stacked area chart, the horizontal axis represents time and the vertical axis represents the ranking in the year. Colors are consistent with field-specific colors in Radviz. If a keyword is ranked for two consecutive years, it is linked by a line. Connecting it to the line in the next year helps in easily recognizing the ranking change.

This view shows the change in popular keywords by years. (G3) In particular, it answers specific questions such as " What are the most popular libraries in the selected technology?" and "What technologies are popular by selected GitHub developers and companies recently?".

4.2.4 Repository List

This visualization system primarily aims to support developers navigate development trends and pick an area of interest. The previous three views including Radviz, stacked area chart, and keyword ranking visualization, provide development trends. The repository list view furnishes information for deeper exploration and learning. This view shows the title, description, field, and related keywords of the repository along with GitHub links. By presenting the actual data that makes up the visualization, this view

not only allows the user to trust the visualization, but also enables deeper exploration.

5 Case Study

We illustrate two case studies to prove that visualization has achieved its goals completely. A user explores trends through the following process. First, the user selects a keyword or a node based on the field of interest in Radviz. The user then analyzes trend changes using stacked area charts and ranking visualization. There are three things the user can ascertain at this point of time. 1) Is this technology popular always? 2) Is this technology coming up recently? 3) Is this technology showing a recent decline? Finally, once the user selects a particular technology or project after this analysis, she or he navigates to the GitHub using the repository list view to explore in depth.

Figure 5 shows the popularity of the language and libraries in the Web front end. The keyword 'JavaScript' has been used consistently throughout all the years. This shows that 'JavaScript' is a key development language for the web front end. The popularity of 'jQuery' has shrunk since 2011 and has been pushed out of the rankings after 2014. In 2011, when 'jQuery' began to decline, 'react', the library developed by Facebook, first appeared with a keyword ranking '6' and has been in the first place since 2013, as a result of explosive growth. It also provides the techniques delivered by 'jQuery' in 'react', which seems to have partially replaced 'jQuery'.

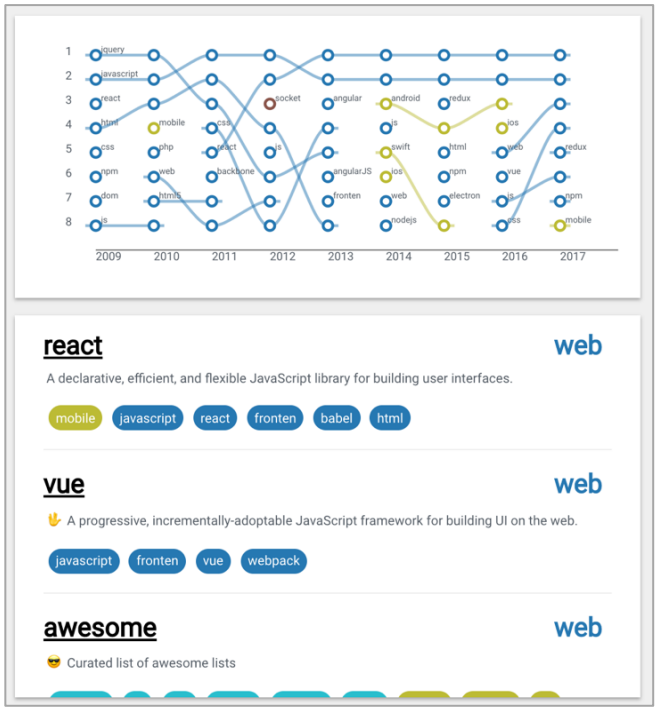

Figure 5: Keyword ranking and repository list generated by the popular user nodes in the Web front-end.

Figure 6 shows the yearly repository yields and keyword rankings for developers involved in machine learning. Cumulative area charts indicate that there were not many machine learning related projects before 2014. The number of projects related to machine learning increased during 2015 and 2016, when 'tensorflow' and 'deep learning' appeared for the first time. It appears that 'tensorflow' and 'deep learning' contributed to the growth of the machine learning field.

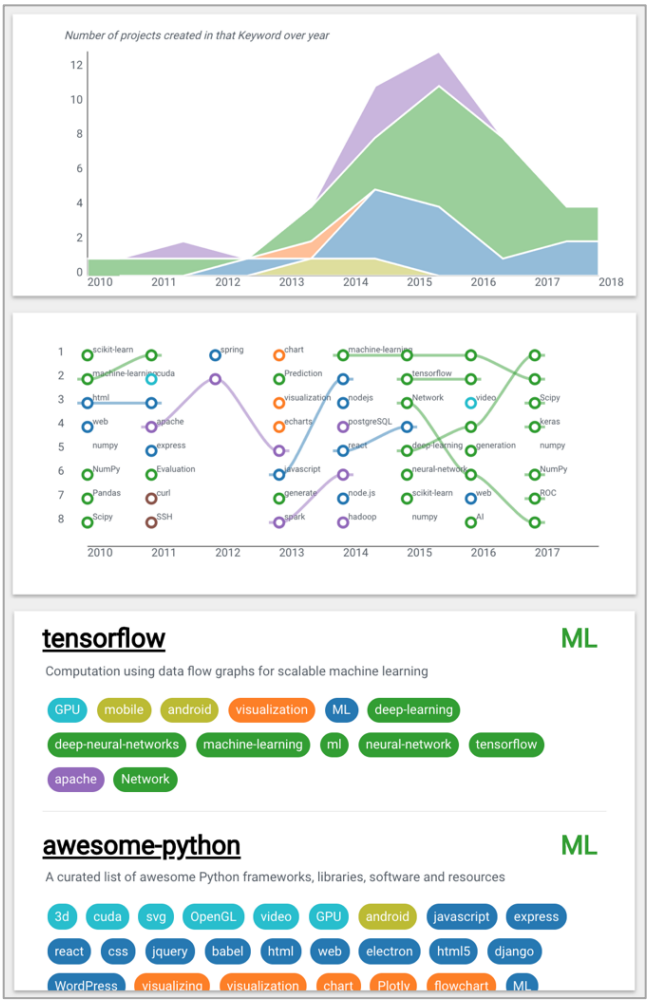

Figure 6: Stacked area charts and keyword ranking visualization generated by repositories associated with the keyword 'machine learning'.

6 Conclusion

This study focused on the technological trends of computer/data science using data visualization and case studies. Through case studies, we found that 'JavaScript' is steadily strong in the Web front-end area, and that interests in 'html' and 'jQuery' have declined owing to the explosive popularity of 'react' since 2012. The findings also demonstrate that the rapid growth of the machine learning field in 2014 is due to the emergence of 'deep-learning' and 'tensorflow' projects.

In this study, we limited the range of data in a repository to that created by the 1000 most popular users in GitHub. This was not sufficient to explore the patterns of repositories with keywords in different fields at the same time. Future studies will have ample data to improve on this aspect. We will also add interactions, such as allowing instances of to be viewed in repositories rather than users or adding custom keywords to allow for various searches. The future work will also include evaluations to verify that the proposed visualization is useful to practitioners in reality.

REFERENCES

[1] Du, Fan, et al. "Finding similar people to guide life choices: Challenge, design, and evaluation." *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017.

[2] Liu, Li, Deborah Silver, and Karen Bemis, "JobViz: Interactive Visualization of Majors & Jobs", IEEE VIS, 2016.

[3] Gee, Alexander G., Min Yu, and G. G. Grinstein, "Dynamic and interactive dimensional anchors for spring-based visualizations", Technical report, computer science, University of Massachussetts Lowell, 2005.

[4] Jihye Lee, Geon Hur, Y ongkyun Lee, Y erin Ga, LeeKyung Hong, Kyungwon Lee, "Research Trend Case Study: For Understanding Interdisciplinary Keywords in South Korea", IEEE VIS, 2015.

[5] Kurosawa, Takeshi, and Yasufumi Takama, "Co-authorship networks visualization system for supporting survey of researchers' future activities", Journal of Emerging Technologies in Web Intelligence 4.1, 2012.

[6] Liu, Shixia, et al, "Topicpanorama: A full picture of relevant topics", Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on. IEEE, 2014.

[7] Heimerl, Florian, et al, "CiteRivers: Visual analytics of citation patterns", IEEE transactions on visualization and computer graphics 22.1, 190-199, 2016.

[8] Yang, Y. I., Quanming Yao, and Huamin Qu, "VISTopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling", Visual Informatics 1.1, 40-47, 2017.

[9] Lee, R. K. W., & Lo, D. GitHub and Stack Overflow: Analyzing developer interests across multiple social collaborative platforms. In International Conference on Social Informatics (pp. 245-256). Springer, Cham, 2017.

[10] Chanhee Park, et al, "GitHub Viz: An Interactive Visualization to Acquire Knowledge from Authoritative Developers", InfoVis, IEEE VIS, 2018.