STACKOVERFLOW VISUAL ANALYTICS: AN INTERACTIVE VISUALIZATION AID FOR THE STACKOVERFLOW FORUM

Anvitha Rao
Master of Computer Science
Arizona State University
Tempe, USA
arao30@asu.edu

Bhavani Balasubramanian
Master of Computer Science
Arizona State University
Tempe, USA
bbalasu6@asu.edu

Ghanshyam Lele
Master of Computer Science
Arizona State University
Tempe, USA
gdlele@asu.edu

Arun Subramanian
Master of Computer Science
Arizona State University
Tempe, USA
asubra29@asu.edu

ABSTRACT

Stackoverflow is one Questions and Answers forum that does not need any introduction. Every software engineer, aspiring coder and students have used Stackoverflow on a daily basis. Stackoverflow hosts over a million questions in its database and has an average incoming rate of more than 3 questions every minute. The stackoverflow community appoints moderators to keep track of the validity of questions and answers and ensure that a maximum number of questions are answered. However, Stackoverflow can be used in much more efficient way than as just a queestion and answer forum. This paper researches an additional usage, using Stackoverflow for the purpose of teaching, by using several interactive visualizations, each associated with an appropriate idea on how to use that visualization. In the second part of the paper, we talk about the evaluating techniques that can be used to evaluate the effectiveness of the visualizations. We also go on to mention the future scope and improvements that can be performed on our visualization to improve its usability.

Keywords

Stackoverflow, Visualization, Data Crawling, Visual Analytics, Topic recommender, Concept Map, Heat Map, Bubble Chart, Text mining

1. INTRODUCTION

Stackoverflow is one of the biggest Question and Answer forum in the entire world for Computer and Software engineers hosting more than 15 million questions and 24 million answers. It has more than 8 million users, thereby making it the largest community in the entire world for Computer and Software Engineers. The primary purpose of the forum for people to ask questions about problems they face and users from all around the world provide answers and solution to it. Each question posted in Stackoverflow has a list of associated tags with it. Tags represent the most related technology and concept used in the question. Some of the most common tags would be "Java", "Python", "Stack", etc. There are more than 50,000 tags available in Stackoverflow which covers near about every topic and technology that is available. These tags are a great indicator of what the question basically and is used to filter questions out.

The biggest problem with Stackoverflow is that it is used only as a Question and Answer forum. There is so many more potential applications that can be developed from Stackoverflow but have not been developed yet. One such application that can be developed is basically identifying the most trending tags of all times indicating the technology that is used in the current trend. Mining for such information in the huge dataset available in Stackoverflow was the main Objective of this project.

In this research work, we have identifies three such visualizations, which we will discuss in detial throughout the course of the entire paper. With these visualizations, we hope to help students identify technologies that they can prepare themselves in so that they would be ready for the job market when they graduate with their degree.

2. MOTIVATION

The main motivation behind this paper is to help students identify technologies that they can master so that they would be equiped with a set of skills that will be in demand when they graduate. This ensures that these students have a better chance of getting a job. Though Stackoverflow is an excellent source of knowledge containing more than 15 million questions, the problem is that students sometimes get intimidated with the reponses available in Stackoverflow. Often, the replies and the answers provided by the community expects the person who asked the question to have some prior knowledge and are not willing to explain the underlying basic concepts. This might lead to studentss feeling discouraged.

The first problem with Stackoverflow, that we have discussed, is its inability to recommend most similar technologies for students. Because of Stackoverlow's huge database of questions, searching any bunch of tags will give us questions asked on those tags irrespective of where or not they are related. This might lead to students studying more than 2 different languages of the same application. For example, studying JAVA and NodeJS at the same time is not useful as

they both are back-end web technologies that can be used in the same project at the same time. Instead, studying Bootstrap with NodeJS will be a lot more useful for students. Though Stackoverflow has the content to make this relation, it does not do it.

The second problem with Stackoverlow is its permanent back-end database. Becuase, stackoverflow has been around for a long time and the system does not automatically delete questions over time, Stackoverflow will list out all questions on a given technology despite the fact that that partiicular technology could possibly be dead. For example, the number of Software Developers who use LISP and Fortran are extremely low in the real world. However, becuase Stackoverflow has a permanently persistant database, a student might still find hundreds of question on the same topic which might not be useful.

The third and final aspect of our visualization is aimed at helping students identify if a particular technology has been popular of the span of the last 10 years. This will help students understand the trend that is going across the years in the industry and help them in learning technologies that are more useful towards getting a job.

This paper aims in helping students identify a variety of uses that they can make out of Stackoverflow and understand the depth of meta knowledge that they can possibly gain by using this visualization. We hope to help students in their career by helping them make intelligent decisions on yhe kind of technology that they should lrarn in order to be successful.

3. VISULAIZATION DESIGN

This section of the paper talks about the actual visualization that we implemented for our project on visualizing data from Stackoverflow, the idea behind each visualization, the technique that was used to visualize the idea and also the key learning objective of the visualization. We also discuss on what other visualization technique could be used for the same idea. We also mention some design principles which made us choose a particular design technique over others.

3.1 Most related Technology

Most common problem students and young professionals face while shaping their career is what they should learn next. Market needs for technologies keep changing fast and students need to keep up with the trends. Instead starting to learn something new from scratch, it is easier for them to learn based on something they already know. The job market today requires professionals to not only know a particular programming language, but also the set of tools that go along that language. Professionals who have the knowledge of a particular tech stack or an ecosystem, are in demand. Thus, the aim of this chart is to provide the students an information about what they should learn next. For example, if you know Java, it makes more sense to learn Android development or Spring boot instead of learning Node.js, because java and node.js are independent and there is practically no use of learning both. Our visualization of a solution for this problem is described below.

3.1.1 Visualization Technique: Bubble Chart

In this visualizaion, students are allowed to choose form a set of tags that have been in the trending list for more than a decade. Some examples of such tags would be JAVA,

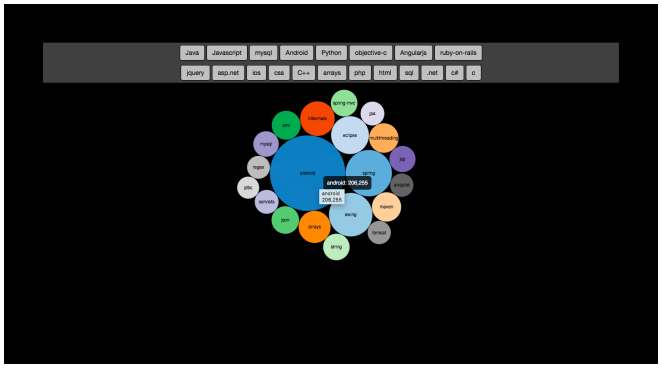

Figure 1: Most Related Technology

JavaScript, Python, MySQL, etc. These tags are presented as individual buttons to the users. When a student or an young professional clicks on one of these buttons, a bubble chart populated below giving them a fair idea of what set of technologies are most common with the technology clicked. For example, if a student clicks on button JAVA, it will form a bubble chart which will show that studying Spring Boot and Android makes a lot more sense than studying something like NodeJS or ExpressJS which will practically be useless because the two technologies are mutually exclusive.

For this visualization we used a bubble chart. Since we wanted to visualize a set of related technologies to a chosen technology, we used a bubble chart where each bubble represents one technology and itsâĂŹs area represents the number of questions asked on the technology. The human visual system naturally experiences a disk's size in terms of its area. And the area of a diskâĂŤunlike its diameter or circumferenceâĂŤis not proportional to its radius, but to the square of the radius[3]. Thus, we should change the area, not the radius of the bubble Bubble charts are intuitive as a discovery tool. Thus, a technology that is used in the industry quite a lot, will have a bigger bubble. Also, each bubble has a different color for aesthetics and for quicker differentiation. The bubble chart is interactive and it displays the number of questions asked for a particular technology when you hover over it. The technology that you want to find related tag against, can be chosen as a parameter via a button.

3.1.2 Reason for using Bubble Chart

We could have used a bar chart with technology on x axis and number of questions on y axis. But since the number of technologies is quite large, bar graph would not have been a god choice. Bubble chart looks aesthetically pleasing and good for large number of tags. We also could have used a scatter plot . But again, since we can change the size of bubbles according to the number of questions, it is more intuitive for the kind of application that is intended for this visualization.

3.1.3 Alternate Techniques

Another technique would have been a hierarchical tree based structure with weights for each edge connecting from the main technology chosen by the student or the young professional. We could depict a tree where each technology leads to it's own sub-branch thereby giving a more broader

overview of the visualization. The reason we did not do this was because it will appear a little bit more complicated and might lead to a misconception that learning one technology might lead to not learning another technology.

3.2 Technology Trends

While learning related technology is extremely important, it does not make a lot of sense to learn a technology that is no longer used in the field. This is quite a common problem that students face because they might not know what the industry actually requires at present. They would be excellent in their skills with a particular language but might struggle with learning a new language due to time crunch and hence might face difficulty in landing a job.

3.2.1 Visualization Technique: Heat Map

This visualization helps students and young professionals to assess their current skill sets. It can help them to check if they are market ready. It can help them to plan their career. This visualization can also be used to predict the future of a particular technology. If the current skill sets of an engineer are more likely to be less valued In the future, they can decide to learn and switch to a different technology.

We wanted to display the trending technologies over the years. We not only wanted to display how famous a particular technology is, but we also wanted to show how the usage has been developed over the years. Thus, this visualization is a combination of a timeline chart and a bar chart. We decided to use a colored heat map because humans can perceive more shades of color than they can of gray, and this would purportedly increase the amount of detail perceivable in the image [1]. Th X - axis of the graph represents each year from 2009 to 2018 and the Y - axis are all the different technologies. The heat map is interactive, and each cell of the heat map has different color shades. It shows the value of a tile when you hover over it. The darker the shades gets, higher is the value. The cluster heat map is an ingenious display that simultaneously reveals row and column hierarchical cluster structure in a data matrix. It consists of a rectangular tiling, with each tile shaded on a color scale to represent the value of the corresponding element of the data matrix. The rows (columns) of the tiling are ordered such that similar rows (columns) are near each other.[2]

For this visualization, we scrapped our data from Stackoverflow for the past 10 years (from 2009 to 2018). Based on the data that we were able to scrap, we figured out the most trending tags across the entire decade and chose the ones that occured in all years. This is to ensure that the heat map is filled with all technologies that always trended across the entire decade and not just a few years. This ensured that data was evenly distributed and not manipulated upon to give false positives.

3.2.2 Reason for using HeatMaps

We could have used a histogram based chart to display the frequency, or a bar chart with time on X - axis and frequency on Y - axis. But this would have had us select the technology as parameter which would mean seeing trend of only one technology at a time. Sine the purpose of thus visualization is to see the trends of all technologies together over time and compare them, we ended up going for heat map. It enables us to show three parameter, technology, time and frequency all at the same time

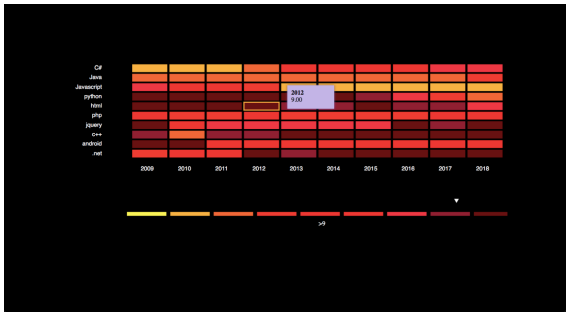

Figure 2: Technology Trends

3.2.3 Alternate Techniques

To gain even more insight, we could also use chained heatmaps where in we click on a technology of a particular year and another heatmap crops up with month-wise heatmap. We could also do the same for a year. We could also maybe create a histogram to depict the month-wise number of questions asked. However, that type of visualization might not be actually helpful for the student as the number of questions asked in every month does not have any relation to the increasing or decreasing trend of that particular language or technology.

3.3 Evergreen Technology

Evergreen Technology visualization deals with figuring out the actual impact that a technology has had in the industry over the past decade as a whole. This visualization was necessary to ensure that the previous visualization was not read wrong. Consider the following example, A particular technology 'A' had received around one 100,000 questions in the year 2009 and that dropped to 90,000 in the next year. This would have been shown in the heatmap as a gradual decrease. Now consider another technology 'B' which had only 10,000 questions in the year 2009 and that saw an increase in the next year to 25,000. This obviously means that the popularity of B has increased but that does not mean that the popularity of A has decreased. However, that particular information might not have been available in the heatmap leading to an incorrect assumption that the technology might not be used.

3.3.1 Visualization Technique: Word Cloud

The Third visualisation is a word cloud which assists users in identifying the top tags for a span of 10 years. This visualisation was based on tags and their frequencies for 10 years. Words are constructed based on frequencies of tags. This essentially meant that larger the count of the questions asked in a particular technology in the past 10 years, bigger it's size would be in the visualization. We constructed this visualization with the help of the data that we scrapped for the previous visualization.

The number of questions for each year of every tag was calculated and a sum of this gave the number of questions for that tag across a period of 10 years. This was repeadtedly done for multiple tags and then the top tags were fed to the word cloud which allotted the color and orientation for each word at random. The key factor was the size. Bigger size meant that the tag was more used. This helps us in understanding that though the trend shows that the questions

Figure 3: Evergreen Technology

had decreased over the year, it still is a majoe player in the tech industry and hence learning the concepts under those technologies will certainly help in the long run.

3.3.2 *Reason for using Word Cloud*

The primary reason behind using word cloud was its innovative design which could be understood by anyone. This means that a person who is looking at this visualization does not have to know about visualization techniques or the type of visualization used. Any personal with a reasonable amount of intelligence will be able to assume that bigger the font size of the word, more prominent it is.

Another advantage of using this visualization is that it does not have to be looked at too close. Unlike bar charts and pie charts, they do not reflect any mathematical number on them. This definitely helps get the point across to a wider audience.

3.3.3 *Alternate Techniques*

The same word cloud technique can be smoothened out to form an even visualization. We could also bring the concept of font-color into picture where in we could have two different aspects of the data being represented by the same visualization. For example, increasing trends across the years can be represented by one type of color while the decreasing trends can be depicted by another color. Size can depict the overall evergreeness of the tag or technology. Moreover, the current visualization populated the word cloud with random orientations. This suited our purpose behind the visualization but if needed, the direction could also be considered for a third dimension in visualization. An example of such a use would be incase we were segregating the tags into groups like Technologies which could include Java, JavaScript, HTML and other such languages and Techniques like Stack, Queue and other such concepts.

4. METHODOLOGY

The Stackoverflow data provided to us consisted of the following:

- Type:

  This dealt with the actual type of the post in Stackoverflow. There are three types of post: Question, Answer and Accepted-Answer. Questions relate to the actual question posted by an user. Answer refers to a post that was made in attempt to answer the question asked while the accepted answer refers to the Answer post that was accepted as the answer for question posted by the user. Only the user who posted the question has the ability to mark an answer as accepted.

- Title:

  Title is the actual title of the question. This will not include any code as such. It will be a brief statement as to what the crux of the problem is. The title is a very important part for search engines such as Google because they generally index pages with the title for the question.

- Content:

  This is the entire content of the post. This includes everything such as the text in it, code, hyperlinks etc inside the post. Essentially, everything that the user types out as a part of the question or answer is a part of this.

- Text:

  This is the String content of the post. This could be the data typed out in the answer or the question without including any hyperlinks or code. This is just a code-free version of the post. This generally contains all the explanation in the post which could be the description of the question or the answer.

- Code:

  This is a subset of Content where-in only the code part of the post is present. All other comments, texts, links, etc are striped out.

- User_Id:

  This is an unique identifier assigned by Stackoverflow to every user who creates an account with Stackoverflow. It is mandatory to create an account with Stackoverflow for creating, editting, deleting and voting on posts. This is used to identify who created what post in Stackoverflow.

- Time

  This is the time at which the post was created. This is usually the local time in which the question was created.

- Vote:

  This refers to the number of votes each post received. Each vote is an integer and depends entirely on the quality of the answer. It is generally expected that an accpeted answer has a positive vote associated to it while an incorrect or out-of-context answer is given a negative vote.

- Repuation: Reputation is an integer representation of an user. A high value for reputation generally means that the user is extremely good with his answers and that his answers are usually correct. New users start with 0 reputation and reputation is built based on their contribution to the forum.

- Tag: This is one of the most important data provided. Tags refer to the keywords, techniques or technologies that a given question is related to. Tags are used for

indexing by search engines and also for search inside Stackoverflow. Tags are the best way to identify the topics/techniques a given questions fall under. Our visualization makes extensive use of this tag attribute. Every question has multiple tags in case if the question spans over multiple technology or techniques or both.

4.1 Data Sources

Despite the fact that we had the entire list of questions on Java for the year 2014 posted on Stackoverflow, our data requirements exceeded what was given. Hence we crawled additional data from Stackoverflow. We used Python and BeautifulSoup for crawling and cleaning the crawled content. The actual data used for each of the visualizations is mentioned below.

4.1.1 Most Related Technology

For this visualization, we had to restructure our data as follows:

$$\{tag1Name,\ tag2Name,\ count\}.$$

tag1Name represents the tag that the user clicks from the list of available buttons. tag2Name refers to all tags that occured along with tag1Name in the question posted on Stackoverflow. The count is the number of times the two combination of tags appeared together. Since this relation is commutative, we could make the assumption that the number of times tag1Name and tag2Name appeared together is exactly equal to the number of time tag2Name and tag1Name appear togehter. However, while visualizing this content, it is not necessary that they will appear with the same size. This is simply because the bubble chart is relative and hence if tag2Name is the most common for tag1Name and tag3Name is the most common for tag2Name, then the size of tag2Name when tag1Name is selected will be bigger than the size of the bubble for tag1Name when tag2Name is selected.

4.1.2 Technology Trends

For the technology trends visualization, we scrapped the entire website for all questions between 2009 and 2018. Based on the results, we created the following structure for the data we crawled:

$$\{tagName,\ frequency,\ year\}.$$

The tagName refers to the technology or technique while the frequency refers to the number of times that particular tag has been used to tag a question in the year this question was posted. We collected this information for a lot of tags but decided on the top 10 based on how frequently they get featured across the years. This was to ensure that we do not choose a tag that was a featured tag in one year but was not featured in another.

After choosing the top tags, we made the heatmap based on the frequency of each tag across the years. We made increments of 10% for a change in color. This was to ensure that certain tags that might have had a low value with high increase rate do not remain in the same color.

4.1.3 Evergreen Technology

For the evergreen technology visualization, we used the data collected for the previous visualization. However, for this visualization, we added the frequency of the tags for each year. This resulted in one entry for one tag with the following data structure:

$$\{tagName,\ frequency\}.$$

where the tagName is the name of the tag and the frequency is the sum of frequencies across the 10 years. This ensured that irrespective of a decline or an incline in the usage of the technology, we get a normalized result based on the actual number of questions asked over the past 10 years.

5. ANALYSIS OF THE RESULT

After performing these visualization, we could gain several inferences from them. We could understand what was the most sought after technology based on the number of questions that users across the world were asking. This is a summary of the results individually:

5.1 Most Related Technology

As expected, we were able to find out that the expected tags for each technology were extremely related together. For example, clicking on JavaScript came up with JQuery, AJAX, HTML, CSS, AngularJS, Ajax, etc. while clicking on iOS came up with X-Code, Swift, Objective C, etc. This helps students and young professionals understand the grouping of technologies that form a stack.

We were also able to identify the correct number of questions that lead to the bubble being the size it was. This visualization hence proved to be a very useful one in identifying a set of corelated technology.

5.2 Technology Trends

The technology trends visualization was aimed at identifying which technologies were most commonly sought after based on the number of questions that was asked on that tag for that year. It was observed that Python had a very high density of questions for the year 2009. This could be coraborated with the fact that the Flask API for Python was released in the year 2009 and a lot of people had questions on Pythong and Flask. However, the questions on HTML had a constant decline because the concept of HTMl is getting replaced by very simple HTML code with a very heavy use of CSS and JavaScript to embed and format the looks. Introduction of Bootstrap and other such technologies has reduced the burden on the actual HTML to design pages and hence have seen a decline.

5.3 Evergreen Technology

This visualization considered the number of questions over the past 10 years and the size of the visualization was based on the frequency of the questions asked. While trend showed that the number of Java questions asked became constantly lower, Java was still one of the major contributors and hence it was written with a very high font size. The declining trend in Java could be due to the fact that a lot of universities and schools have taken Java as the primary language for programming replacing C and C++.

6. EVALUATION PLAN

Visualizations need to be evaluated because the current methods evaluate based on small data sets , with university students using small taks.[4] Visualizations are not only

needed to be evaluated based on Human Computer Interaction, but also how useful they are to the consumer. This is of paramount importance because of the simple fact that the number of people who use visualizations to communicate has increased a lot. Visualizations are now the primary medium of communication for the untrained eye. There are two classes of evaluation for data visualizaion. These are evaluating information visualizations qualitatively as well as quantitatively[4].

- Quantitative evaluation:

  Quantitative evaluations, most well known as laboratory experiments or studies, are those methodologies in which precision is relatively high and in which some declaration can be made about the possible generalization to a larger population. These declarations can include information about the characterization of this larger population and how likely it is that the generalization will hold. In quantitative evaluation, we intend to focus on hypothesis development identification of independent variables, control of independent variables, elimination of complexity and measurement of dependent variables[4].

- Qualitative evaluation:

  Qualitative inquiry works toward achieving a richer understanding by using a more holistic approach that considers the interplay among factors that influence visualizations, their development, and their use.[4] In qualitative evaluation, we intend to focus on in-situ observational studies, participatory observation, laboratory observational studies and contextual interview [4].

7. TECHNOLOGIES USED

This project saw the use of a variety of technology for the purpose of data visualization. Each of these technologies played a pivotal role in furnishing the end product that we were able to present. These technologies are listed below alogn with a short description of what they are and how they are used:

- Python

  Python is very well-known scripting language on which web crawling can be performed easily. In our project, we used Python 3.4 to hit the Stackoverflow URL and crawl the data that was required for the project.

- Beautiful Soup

  Beautiful Soup is an extremely well known library for URL parsing and data cleaning written in Python. We used Beautiful Soup to parse the data returned by our crawler and manipulate the data that was required and restructre the data returned to our needs.

- HTML

  The Hyper Text Markup Language is the backbone for all web pages and we used it to draw out SVGs and other graphs.

- Vanilla JavaScript

  Vanilla Javascript was used for the purpose of calculation on the page other than just drawing the charts.

- D3.JS

  D3.JS is the abbreviation of Data Driven Documents and is one of the best visualizing library available in the market. D3,JS offers powerful tools to create SVGs without actual knowledge of SVG themselves. D3 was written in JavaScript and acts as a wrapper class around SVG to create charts and diagrams with simplified code. D3 acts as a transalator between JavaScript and SVG and converts the user data into SVG charts.

- Bootstrap

  Bootstrap is a CSS based Framework that makes web pages responsive to the size of the screen that is being used to view the page. This ensures that a healthy user-interface is designed which adapts itself to the screen in which the student is viewing the page. This is of great asthetic importance in visualization.

- Heroku

  Heroku is a cloud platform created by Salesforce to helo small-scale companies with cloud platform services. This helps the users by removing importance from the infrastructure and ensures that the users can focus more on the application itself.

8. ACKNOWLEDGEMENT

We would like to express our sincere appreciation and gratitude to Dr. Sharon Hsiao, the course instructor, for her guidance in developing this data visualisation system. Her untiring guidance and support was cruicial for the success of this project. We would also like to thank Mr. Yihan Lu, the teaching assistant for his inputs and constructive feedbacks with regards to project implementation. His questions were thought provoking and at the same time created an environment for healthy discussions on design decisions. We would also like to thank all other teams who offered us healthy competition while at the same time being helpful with suggestions.

References

[1] *Bubble chart*. Apr. 2018. URL: `https://en.wikipedia.org/wiki/Bubble_chart`.

[2] Sheelagh Carpendale. "Information Visualization: Human-Centered Issues and Perspectives". In: Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. Chap. Evaluating Information Visualizations.

[3] *Heat map*. Apr. 2018. URL: `https://en.wikipedia.org/wiki/Heat_map`.

[4] Leland Wilkinson and Michael Friendly. "The History of the Cluster Heat Map". In: *The American Statistician* 63.2 (2009), pp. 179–184.