Visual Analysis of Stackoverflow

Finding popularity, correlation and knowledge index of tags

Nishant Bansal
CIDSE, Arizona State University
Tempe, USA
nbansal7@asu.edu

Prerna Satija
CIDSE, Arizona State University
Tempe, USA
psatija1@asu.edu

Nitesh Kedia
CIDSE, Arizona State University
Tempe, USA
nkedia1@asu.edu

Abstract— **In this paper, we explore the prominence and association of different programming languages used on a popular Community based Question & Answer (CQA) website, called Stackoverflow.com. CQA websites are a great platform to post technical queries related to a language and obtain solutions in real time from users of any part of the world. We analyzed the different programming languages to find some useful features namely language popularity, correlation between languages, and knowledge index across different countries. These features were then used to create a visual interaction in order to make it easier to examine the common trends among languages and compare them. The interactive visualizations enable the user to select a particular language and study its correlation with other languages and geographical knowledge index in a particular quarter. The analysis was done on the data obtained from www.stackoverflow.com for a period of 4 years.**

Keywords—popularity; correlation; Knowledge Index; stackoverflow; visual analytics; tags.

I. Introduction

CQA sites are question-answer communities, beneficial to users belonging to any geographical location. One such platform is provided by Stackoverflow for people interested in Computer Science, software and technology. When one starts to learn a new programming language or works on a project to develop new software, it is very useful to know a few things about that language/ software.

We analyzed the Stackoverflow data to find some useful features particularly language popularity, correlation between languages, and knowledge index of a language across the world. These features are useful as how popular a language is implies how much user support would be available. Also if you are using a language you might want to know other languages used along with the chosen language. If a user wants to find out the most correlated skill to Android, he can get an idea that Java is the most popular choice of language used with Android. Knowing which country has more knowledge for a particular language can help you hire the right people. To make it easier and more efficient for the user to use these features, we developed visual interactions to provide the analysis.

Data visualization is a powerful medium of expressing an information. It involves representing information through visual tools like bar charts, bubble charts and maps. For our study, we used a word cloud to show popularity, a doughnut chart to depict the correlation between languages, and a Choropleth map to plot the knowledge index. In this paper, languages and tags will be used interchangeably.

II. Motivation

The main purpose of the Visual Analytics study of Stackoverflow data is to visualize the growth and decline of programming languages over time. Correlation of tags analysis is useful for some one starting a project in a language of his interest. Or even for some one preparing for a Technical Interview as an interviewer testing one's Object Oriented Programming (OOP) skills is highly likely to ask questions related to Java, C++, classes etc. Also our knowledge index analysis helps to assess useful information like the users of which countries are more proficient in their knowledge of a particular technical skill. This analysis is helpful for both job seekers who can better prepare themselves by becoming adept at the more popular tags used in a country and recruiters of software companies who are looking for some specific skills.

III. Data

A. Data Source and other details

It is the Stackoverflow dataset available at Stackexchange.com. We downloaded the dataset from Kaggle.com. The url is *www.kaggle.com*. It is the data for 4 years spanning from July 31, 2008 to July 31, 2012. It consists of many xml files but for our study we primarily used posts.xml and users.xml. Each row of posts.xml is a record of anything that was posted on Stackoverflow. Total number of posts is a little more than 10 million. The structure of a row in posts.xml is as follows.

<Id, PostTypeId, ParentId, AcceptedAnswerId, CreationDate, Score, ViewCount, Body, OwnerUserId, LastEditorUserId, LastEditorDisplayName, LastEditDate, LastActivityDate, CommunityOwnedDate, ClosedDate, Title, Tags, AnswerCount, CommentCount, FavoriteCount>

Similarly, users.xml consists of information of all registered users of Stackoverflow with the structure given below. It has approxnoimately 1.2 million users.

<Id, Reputation, CreationDate, DisplayName, EmailHash, LastAccessDate, WebsiteUrl, Location, Age, AboutMe, Views, UpVotes, DownVotes>

B. *Preprocessing*

For our visual analysis, we did not use all of the features mentioned in the posts.xml and users.xml. Some of the features in the dataset were used to create new features. To observe the change over time we divided the data into 16 progressive quarters. From users data we were interested only in the location information of the user. The locations available in the data were unstructured addresses of the user as provided by the user. We were more interested in the country wise location of all the users. We used Google Geocodes API [1] to convert these unstructured addresses to get country names and codes. The users with empty or invalid locations were removed and we were left with around 0.2 million users with country-wise location.

From posts data we concentrated mainly on tags. We observed that the posts of type 'answer' did not have tags in it so we used '*ParentId*' to get the tags of their parent (tags corresponding to the question post). The '*CreationDate*' was used to split the data into 16 quarters. '*OwnerUserId*' was used to obtain the country location of the post by mapping it with the user.

For backend data preprocessing we have extensively used Java and Matlab. For all the visualization we have created json files to store the processed data. After data pre-processing, we were able to extract useful information and compress the initial data-set of 40gb to 15mb.

IV. Analysis

Data analysis is the process of cleaning and preprocessing data with the purpose of extracting useful information from it. Stackoverflow data is an interesting dataset to analyze. We explored the stackoverflow dataset to find the most popular tags and the change in their popularity with each progressive quarter. We also determined the most corelated tags to a particular tag. Third, we analyzed how knowledge of a language is spread across different countries of the world. All the analysis is done over progressive quarters where, progressive quarter means the cumulative data from start date of the dataset till the end of that quarter.

A. *Popularity*

The popularity of a tag is directly related to the cumulative number of posts of a tag on stackoverflow.com till a particular quarter.

$$\text{Popularity} = \frac{\text{Number of posts of a tag}}{\text{Number of all posts}}$$

The higher this ratio, the more popular that tag is. The popularity of a tag is useful to know which languages are most commonly used in real world software development or in academic projects as the users of Stackoverflow.com are mostly students or passionate software developers. It is also useful to know the common trend among the skills being looked for by hiring companies. After analysis, we have stored the processed data in "*popularity.json*" with the format shown below.

{
"Q_n": [
{"text": "T_m", "size": popularity},
...
]
...
}
where, Q_n is the quarter,
T_m is the m^{th} tag, and
popularity is the normalized popularity of T_m

B. *Correlation*

Correlation of tags is referred to as the association between tags. To compute the correlation of tags, we used a Tag × Tag correlation matrix. The n_{12} entry in this matrix (Table1) denotes the number of posts where tag t_1 and t_2 co-occur. If the number of these posts is high it means high correlation between tags t_1 and t_2. For instance, the most related tags for HTML are Javascript and CSS. This information is useful to know what are the most common languages used with HTML if someone wants to do frontend development. Also the correlation analysis has a lot of usability from the point of view of companies who want to start a new project or venture into a new domain. Since these correlations can change as technology evolves and more new languages develop and emerge, we have analyzed the correlations over time.

Tags	**t_1**	**t_2**	**t_3**	**t_4**
t_1	n_{11}	n_{12}	n_{13}	n_{14}
t_2	n_{21}	n_{22}	n_{23}	n_{24}
t_3	n_{31}	n_{32}	n_{33}	n_{34}
t_4	n_{41}	n_{42}	n_{43}	n_{44}

Table1. Co-occurrence Matrix

We have processed the data to compute the correlation between tags and then we have written this data in "*correlation.js*" in the following format.

[
{
"quarter": N,
"tag": "T_a",
label: "T_b",
value: correlation value
},
...
]
where, quarter is the key for N^{th} quarter,

tag is the key for the selected tag T_a,
label is the key for the correlated tag T_b, and
value contains the correlation value.

C. Knowledge Index

The Knowledge Index Analysis is the geographical distribution of cognizance of a tag over different quarters of time. For instance, if you want to know which regions of the world are rich in knowledge of Ruby on Rails, you can look up at this analysis. As with popularity and correlation discussed in the previous sections, the knowledge quotient or prominence of a language in a region also changes or grows with time. In some cases, it may even remain the same. To examine such trends, we need an analysis over both geographical location and time. The Knowledge Index of a tag t_1 in a country c_1 depends on two factors Global Knowledge Index (GKI) and Local Knowledge Index (LKI) as shown in the formula used below:

$$KI = b^{*}\ LKI * GKI$$

Where,

$$GKI = \frac{\text{Number of answers of } t_1 \text{ by country } c_1}{\text{Total number of answers of } t_1}$$

$$LKI = \frac{\text{Number of users of } c_1 \text{ answering for } t_1}{\text{Total number of users of } c_1}$$

And, b is some constant. For our calculations, we used b =100.

GKI is a global measure because it is the ratio of the number of answers for a tag by a country to the total number of answers by the world. It signifies the participation of that country for that particular tag. Higher GKI means the contribution of that country is more and thus its knowledge index is high but if the population of that country is high the knowledge index will still be high. Thus to remove this bias, we used a local knowledge index parameter which is the ratio of number of users of a country who have answered posts for a tag to the total number of users of that country on stackoverflow.com. By doing so, we get a fair knowledge index irrespective of the user base size of a country.

We have computed the knowledge index for the data and created a set of json files for each quarter. The format of these json files is as follows.

```
{
        "Ta": {
                "Cm": knowledge index,
                ...
                "Cp": knowledge index
        }
        ...
}
```

where, T_a is the selected tag,
C_m and C_p are country codes with knowledge index

V. Visualization

Data visualization is a very influential technique to represent any data visually. It is important to convey a story effectively and so the choice of visualization tools matters. A good visualization not only enables its target audience to understand the analysis through visual effetcs but also gives them the flexibility to explore data according to their needs through visual interactions. In this paper, we have shown three visualizations for the three types of analysis: popularity, correlation and knowledge index. All these visualizations are for the top 400 popular tags.

We have strictly followed the Visualization Mantra of “Overview first, zoom and filter, then details-on-demand”. First, we have provided a generalized context for understanding the dataset. Our index page is the general word cloud for all tags. On selection of a tag, we delve into the specifics of that tag, thus following the zoom and filter strategy. If a user wants to see more details, they can futher explore the correlation modal or knowledge index modal for any tag and quarter of his choice and a customized visualization for the user can be generated. Thereby enabling us to implement the “details on demand” feature.

A. Popularity Visualization

To visualize the popular tags used on stackoverflow.com, we have used a word cloud. This visual design is useful for quickly perceiving the most popular terms where the size of a tag depicts its popularity. Bigger size of a word denotes high popularity of that tag. Tags which are smaller in size are the less popular tags meaning that they have less number of posts on stackoverflow. For instance in Fig1., tags like “java”, “php”, “c#” are more popular tags as compared to “ruby on rails” and “objective-c”.

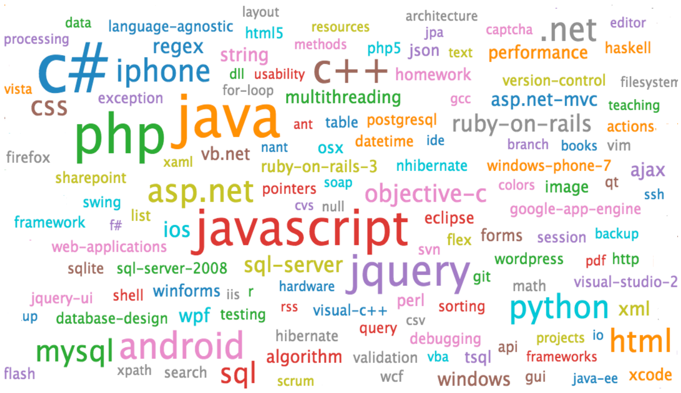


Fig 1. Word cloud for popularity

This visualization can be dragged and any tag can be selected by clicking on it. On selecting a particular tag, more visualizations can be seen. Also we have used a time slider which shows the change in the tag popularity with each progressive quarter.

A.1 Choice of Visualization Tool

We have chosen word cloud for showing popularity of tags because the number of tags is very high. It takes less space and is easier to interpret. Using different colors for different tags makes it look good and increases the readability of this visualization. Initially we used a bubble chart where size of the bubble denoted popularity of the tag. But since the number of tags was huge, it could not be adjusted on the window frame. Also the text on smaller bubbles was very hard to read.

A.2 Implementation

We have used d3.js to implement this visualization and we have scaled the tag font size based on popularity. We have also implemented the on-click functionality to open a modal. Also a time slider is implemented using jQuery which updates the word cloud and shows the transitions over each quarter.

B. Correlation Visualization

Correlation between tags is an important visualization for anyone who wants to find the most related tags to a development tool/language they have interest in. For this visualization we have used a doughnut chart. The doughnut (or donut) chart is divided into ten slices for the top ten co-related tags. The region covered by each slice is directly related to the correlation percentage. The bigger slice with more area on the donut chart represents a more corelated tag compared to the smaller slice. We have also included percentages in each slice to view these correlations. The visualization shown in Fig 2., depicts the top ten correlated tags to HTML. One can easily get an idea from this visualization that CSS and Javascript are the most commonly used languages with HTML as they have higher percentages in the chart. Similarily the donut chart for Java shows Eclipse as one of the top correlated tags thus giving an idea that Eclipse is the most common choice of IDE used to develop and run java programs.

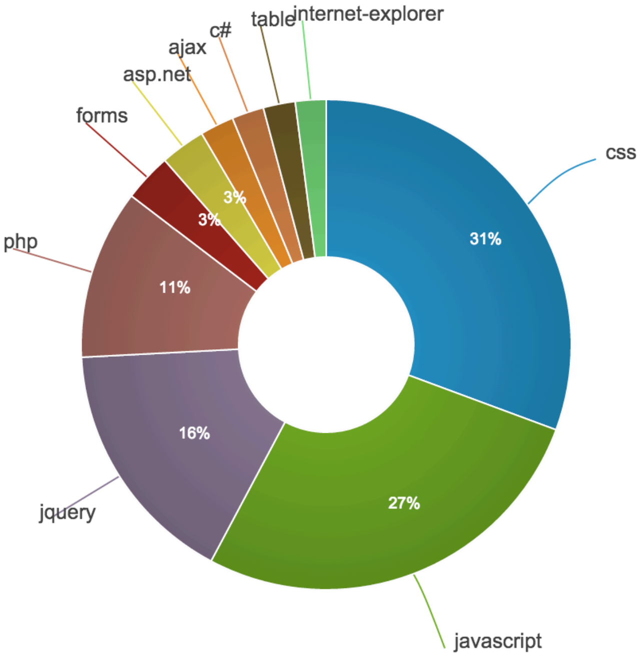


Fig 2. Donut chart for most correlated tags to HTML tag

B.1 Choice of Visualization Tool

We have chosen donut chart as a visualization tool to depict the most correlated tags to a selected tag. Initially we thought of using a Pack Layout but the problem we faced was that the smaller bubbles reduced the readability of the tags. So we chose donut chart as it clearly indicates the top ten tags along with their percentage of corrletaion. Giving unique colors to each tag in the chart makes it look visually more appealing.

B.2 Implementation

We have used d3pie library of d3 to create the donut chart. On clicking a slice in the donut chart, the onclick event gets called which updates the chart and shows the correlation of the clicked tag. For tags with correlation percentage less than 3% we have not shown the percentage value because the visualization gets cluttered. We have created a time slider using jQuery which updates the correlation of tags over progressive quarter.

C. Knowledge Index Visualization

The knowledge index analysis is an analysis over geographical locations. So we have used World Map to visualize this story. The World map depicts different regions with different shades of a color to show the knowledge index. The countries with brighter shades have higher value of knowledge index compared to countries depicted with lighter shades.

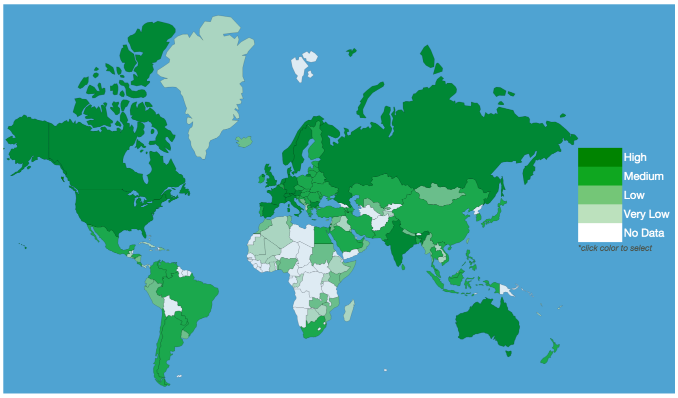


Fig 3. World Map for Knowledge Index of Java

The visualization in figure 3 shows the world map for knowledge index of Java. We can observe that countries like United States of America, India, Russia etc. have brighter shade of green which implies high knowledge index whereas countries like Algeria, Zimbabwe etc. have very light shade of green representing low knowledge index. For countries like Niger, Angola, Afghanistan etc. for which we don't have enough data to compute the knowledge index, we have used white color.

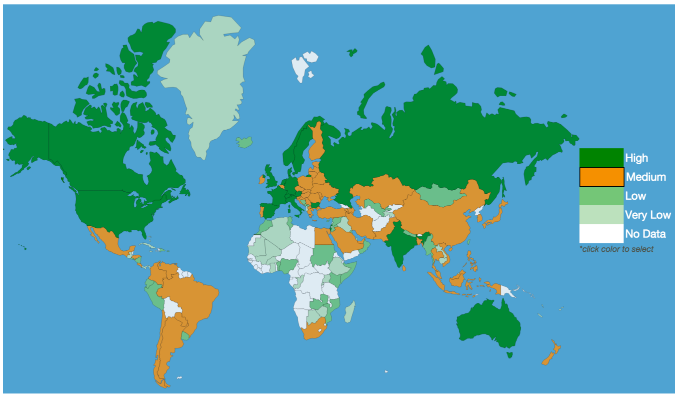


Fig 4. Highlighted countires with “medium” knowledge index

We have included a clickable legend for more effective visual analysis. As shown in figure 4, when you click on a color in the legend, it highlights all the countires corresponding to that knowledge index. So China, Finland, Brazil etc. got highlighted with orange color when “medium” index legend is selected.

C.1 Choice of Visualization Tool

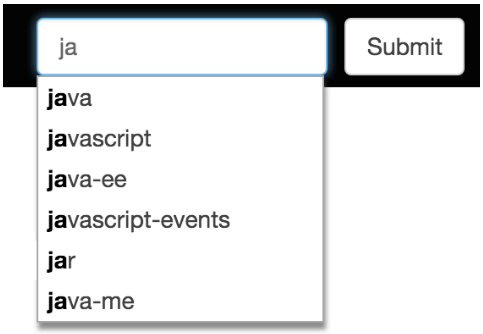
We have used world map because knowledge index is a geographical property. Particularly, we have chosen a choropleth map because it shows the level of variablity across the world. We used shades of green because it symbolizes growth and the power to expand and knowledge index also depicts the growth of knowledge of a tag. Green also symbolizes peace and harmony and knowledge connects the world thereby bringing in peace. The background color is blue to symbolize water in those regions. We used a shade of orange to highlight the corresponding countries when a legend color is selected because it is a hot color and will overshadow the green color, thus prefectly highlighting the countries. We considered red color to highlight but it is too hot for the human eyes and is an aggressive color. On hovering over a country, the color of the country changes to yellow to temporarily highlight it and displays the country name.

C.2 Implementation

We have used Ammaps to create the world map for knowledge index. We implemented the hover such that when a country is hovered it gets highlighted and the name of country is shown in the balloon text. Also, we have created a table to design the clickable legend. We have implemented on click events for all colors in the legend to update the map. If a selected color is clicked, then it gets unselected to remove the highlight. We have not included the zoom functionality because zoom is used to drill down the map to get more information but in our case all the relevant information is clearly shown on the full world map. To show the change in knoeledge index over time we have developed a time slider.

D. Other Interactions

We have designed a search feature with auto-complete functionality to easily search for tags. This is provided in the header of the main webpage so if a user is looking for analysis of a particular tag then he/she can directly search instead of finding it in the word cloud. Auto-complete functionality is also included to make it easier for the user. It is implemented using jQuery. Figure 5 shows a screenshot of the auto-complete functionality.



ja

Submit

java
javascript
java-ee
javascript-events
jar
java-me

Fig 5. Search and Auto-complete functionality

In the time slider implemented to show the change in all the three visualization, we also included play, pause and stop functionality so that the user can see the change in the form of an animation. The slider also includes drag and click option to jump between quarters so that if the user is only interested for a particular quarter then he/she can easily switch to that quarter. The time slider is shared among the three main visualizations so that if you switch between visualization then it should show the analysis for the quarter last selected.

The header of the main webpage also includes links to documentation and a dropdown to see the team members who worked on this project. Clicking on the name of a team member will take you to the linkedIn page of that member.

VI. Evaluation

We have evaluated our visualization design for Popularity index using TIOBE index. We have found a similar pattern for technologies that have a sharp rise and decline in popularity in our dataset. According to our data analysis, popularity of jQuery, Android, Javascript has increased drastically between Jul 2008 - Jul 2012 that has been validated using TIOBE index. Similarly, popularity of C#, asp.net has decreased over time which has also been validated.

We have evaluated correlation between different technologies from the popularity point of view. For example Animation used to have strong correlation with win-forms and .net in 2008 but with the decline in their popularity over time and simultaneous rise in popularity of android, javascript etc, correlation of animation with these technologies has increased drastically.

Moreover, technologies that are strongly dependent on each other like .net and C# maintain their correlation over time. Similarly with the rise in popularity of jQuery, correlation between jQuery and javascript has increased because of strong mutual dependency on each other. We also evaluate the correlation result from our data analysis with the submissions made on Kaggle contest with a good accuracy.

VII. Related Work

Most of our design and analysis is inspired from the work done by [2] Andrea Cuttone, Dept. of Applied Mathematics and Computer Science, Technical University of Denmark. His visualization is focused on analyzing time trends, correlations, location trends, and recommendation and twitter network of stack overflow dataset. Although his work also deals with user network, for the purpose of this study we focus on tags [3]. We have also taken motivation from different visualization designs on Kaggle contest for predicting closed questions for the same stack overflow dataset, which is used in this visualization work. Here, the popularity of a language and its correlation with each other has been visualized using a scatter plot and we have divided that into two different visualizations to make it more interactive and informative.

VIII. Conclusion

With the growth of technology and Internet, use of Internet for study of a language, sharing knowledge on online communities has increased. Since the number of Internet users and the awareness about different programming languages has increased; the amount of data on online platforms like StackOverflow has also increased. New languages have emerged leading to more curiosity among Computer Science students and Software Developers. Thus data analysis has become necessary to scrutinize trends among different programming languages to see how each language is growing or declining with time. It is also important to find how correlated a language is to other languages. The geographical knowledge index is essential to analyze the proficiency of a tag in different countries.

Today, data visualization is used as the most impactful medium to show a story. Visualization tool selection is important in this as a good tool adds enhanceability and flexibility to the design. In our study we have chosen visualization tools that are most relevant to depict our analysis. For showing popularity of tags we have used a word cloud which clearly shows the direct relation between tag font size and popularity. Also for visualizing the correlation between tags, we have used a donut chart which clearly communicates the link between correlation and slice size of each tag in the donut chart. For our third analysis, we have used the world map to exhibit the spread of knowledge of each tag across different countries. We have used a time slider in all these visualizations, which can be dragged and a particular quarter can be selected by the user. Our visual analytics for StackOverflow is useful to people from various backgrounds as technology is used everywhere be it healthcare, banking or social media.

IX. Future Work

We plan to extend this visualization work to provide more granular details to the audience. We plan to change the quarters to monthly or weekly data to visualize the change in pattern for a shorter span of time. Apart from this we plan to accommodate user network and user reputation to build a recommendation system. This recommendation system would be able to post alerts to users who can give answers to a question being posted in real time. This could also be used to build a social network between users who match each other’s technologies and expertise level.

X. Resources Used

- d3.js
- bootstrap.js
- d3pie.js
- jsonsql.js
- lodash.js
- JSLINQ.js
- defiant.js
- jquery.mockjax.js
- jquery.autocomplete.js
- ammaps.js
- Java
- Matlab

Acknowledgment

We would like to thank Dr. Sharon Hsiao for helping us learn about data visualization and sharing the visualization mantra which enabled us to analyze the data and deliver an effective visualization.

References

[1] http://developers.google.com/maps/documentation/geocoding/

[2] http://code-trends.appspot.com/loctrend

[3] www.kaggle.com/c/predict-closed-questions-on-stack-overflow/prospector#188

[4] Heer, Jeffrey, Michael Bostock, and Vadim Ogievetsky. "A tour through the visualization zoo." *Commun. ACM* 53.6 (2010): 59-67.

[5] Sun, Lingling, and Julita Vassileva. "Social visualization encouraging participation in online communities." *Groupware: Design, implementation, and use*. Springer Berlin Heidelberg, 2006. 349-363.

[6] Cleveland, William S., and Robert McGill. "Graphical perception: Theory, experimentation, and application to the development of graphical methods."*Journal of the American statistical association* 79.387 (1984): 531-554.

[7] http://en.wikipedia.org/wiki/Tag_cloud

[8] http://d3pie.org/

[9] http://docs.amcharts.com/javascriptmaps/AmMap

[10] http://en.wikipedia.org/wiki/Choropleth_map

[11] Schenk, Dennis, and Mircea Lungu. "Geo-locating the knowledge transfer in StackOverflow." *Proceedings of the 2013 International Workshop on Social Software Engineering*. ACM, 2013.

[12] Agrawala, Maneesh, Wilmot Li, and Floraine Berthouzoz. "Design principles for visual communication." *Communications of the ACM* 54.4 (2011): 60-69.

[13] Thomas, James J., and Kristin A. Cook. "A visual analytics agenda." *Computer Graphics and Applications, IEEE* 26.1 (2006): 10-13.

[14] Andrienko, Gennady, et al. "Space, time and visual analytics." *International Journal of Geographical Information Science* 24.10 (2010): 1577-1600.