Analyzing and Comparing Different Users of CrossValidated

Kavita Rathore
1140 E Orange Street
+1(480)-416-1161
krathor1@asu.edu

Kshama Jain
1140 E Orange Street
+1 (480)-765-4885
kjain10@asu.edu

Nishi Shah
1140 East Orange Street
+1(669)264-8479
nkshah6@asu.edu

Prachi Arun More
1255 East University Drive
+1(248)607-4212
pamore@asu.edu

ABSTRACT

With rapid development of online learning, number of users on online discussion forums have increased by huge number. On various online discussion forums one can look into statistics of a particular user but cannot visualize statistics for more than one users simultaneously. In this paper, we have implemented a visualization in which two different users on CrossValidated forum can be compared based on their profile i.e. number of posts, score , reputation, votes and views. One can also compare text/words used by users in their posts. Text Modeling has been implemented to show comparative analysis of texts of two different users. This paper visualizes users specifically who have posted question/answer in R domain. This visualization also tries to find correlation between different features of user profile.

Keywords

Visualization; Cross Validated; R; Text Modeling; d3; Bubble Chart

1. INTRODUCTION

Figure 1: Visual Analytics[1]

Visual analytics is a way of analytical reasoning which is facilitated by interactive visualizations. It is an approach which combines visualization, human factors and data analysis. This helps a user to draw intelligence from unprocessed data. The concept of using pictures to understand data has been around since years, from basic maps in past centuries to the invent of interactive visual tools, like gephi in recent decades. Pictures and diagrams are much more easy to understand as compared to numerical data values as they give you a quick summary of the data as shown in figure 1. Data visualization is a quick and easy way to convey ideas. Moreover, data is increasing at an incredible rate and data collection and storage is also increasing at a same rate, but understanding the data, like drawing out patterns and trends remains the bottleneck in data -analysis because of the inability to analyze it. Visual Analytics methods allows user to break this bottleneck and get insight into the data to make well informed decisions in complex situations. For our project, we are working on visual analytics on the data of online learning forum called CrossValidated specific to R users. Online learning forums have attracted a lot of users in recent time. Analysing these huge amount of users of CrossValidated forum provides an opportunity to answer some research questions as discussed in further sections of the paper.

2. MOTIVATION

With large data there's potential for great opportunity. But many such opportunities don't turn into constructive solutions because of lack for good data analytics tool. One such potential research problem is the analysis of different users on online learning forums. Many online discussion forums such as quora allow a user to view his/her own statistics such as how as number of posts on a timeline, votes and reputations over a time and many more. Inspired by above design we decided to create a visual analytical tool which focuses on comparative results of two users. In this project, we provide the user the ability to view the number of posts, reputation and many more features and this can assist the user in evaluating the different users.

Being a user on an online forum, gives them privileges to view their own profile, unanswered questions and posts made by other users. But allowing the user to monitor his/her progress and understand how to improve oneself is not an option. Allowing users to compare their profile with other users, inspires them to improve their performance in the online forum by increasing their posts and score. It also informs them about the tag activity and gives them an insight about the forum and its users.

3. VISUAL DESIGN

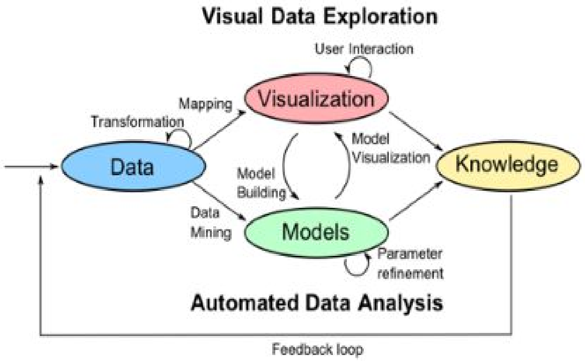


Figure 2. Data Analysis FlowChart[2]

Flow for visual analytics implementation is as shown in figure 2. Firstly, data is collected in .csv format to implement this visualization design. After which data is cleaned and converted to proper format. Henceforth data is analyzed to understand the trends in data. Later, visualization design is selected in which the processed data can be visualized properly. Lastly, we have implemented text analysis algorithm to compare text/words of two different users.

3.1 Data Collection and Cleaning

We started collecting users' data from Stack Exchange api [1] as shown in figure 3. But we were facing problem in collecting data for users in R specific domain. Collecting data for all users who have posted answer/question in topic 'R' was not directly possible. To collect data in R specific domain we had to manually check tags for each user and then combine all such users. Hence, we moved to search for other source for Cross Validated data collection. As a result, we have used Stack Exchange Data Explorer website to collect the data [2] as shown in figure 4. This website provides feature where one can compose SQL query to fetch data from Cross Validated database. Database schema for Cross Validated contains 27 tables. From these 27 tables we have used 4 tables which are Posts, Users, PostTags, Tags. Tags table gives us 'tagid' for tag 'R'. PostTags gives us all posts posted with tag as 'R'. Whereas, Posts table gives us all users who have posted question/answer in R domain and number of posts for such users. Posts table also provides post body for all such posts. Lastly, Users table gives us profile data such as reputation, upvotes, downvotes, score in R for all users who have posted in R domain. We grouped Posts data on basis of users so that we can get all the posts combined for each user. Result of SQL query comes in the table format. Following is one such SQL query:

select OwnerUserId,count(AcceptedAnswerId) as

Total_Accepted, sum(Score) as score,count(Id) as total_posts

from Posts where Id in (select PostId from PostTags where tagid = 41) and AnswerCount > 0 and OwnerUserId >0 group by OwnerUserId;

Stack Exchange API v2.2

This is the documentation for the v2.2 Stack Exchange API (with both authentication and write support). If you have additional questions, or believe you have encountered a bug, don't hesitate to post a question on Stack Apps.

If your application is in a runnable state (even a beta one), Stack Apps is also the place to list it.

General

Applications should be registered on Stack Apps to get a request key. Request keys grant more requests per day, and are necessary for using access_tokens created via authentication.

All API responses are JSON, we do support JSONP with the callback query parameter. Every response in the API is returned in a common "wrapper" object, for easier and more consistent parsing.

Additionally, all API responses are compressed. The Content-Encoding header is always set, but some proxies will strip this out. The proper way to decode API responses can be found here.

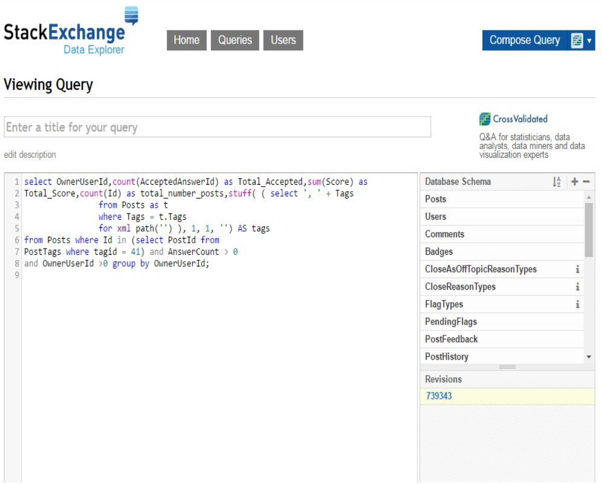
API usage is subject to a number of throttles. In general, applications should strive to make as few requests as possible to satisfy their function.

API responses are heavily cached. Polling for changes should be done sparingly in any case, and polling at a rate faster than once a minute (for semantically identical requests) is considered abusive.

Developers can trim API responses down to just the fields they are interested in using custom filters. Many types have fields that are not normally returned (question bodies, for example) that can likewise be requested via a custom filter.

Figure 3. Stack Exchange API[3]

Data obtained in .csv format is then processed and cleaned. Posts for which there are no answers are filtered and removed. Also, posts for which UserId is missing are also removed in order to get clean and noise less data. Posts body from all posts by each user are combined and cleaned for text analysis. Finally cleaned data has 5704 users specific to R domain.



Viewing Query

Enter a title for your query

edit description

```
select OwnerUserId,count(AcceptedAnswerId) as Total_Accepted,sum(Score) as
Total_Score,count(Id) as total_number_posts,stuff( ( select ', ' + Tags
                from Posts as t
                where Tags = t.Tags
                for xml path('') ), 1, 1, '') AS tags
from Posts where Id in (select PostId from
PostTags where tagid = 41) and AnswerCount > 0
and OwnerUserId >0 group by OwnerUserId;
```

Figure 4. Stack Exchange Data Explorer[4]

3.2 Visualization Design

Our first primary task was to design the visualization for the users who contributed for R language on Cross Validated website. First we decided to display the users on Radial Chart, where angle would be according to their number of posts and distance from center would be according to score of users in R domain.

But in our dataset there were so many users for particular post. So most of the datapoints of users were displaying along the same radial axis.. As our dataset did not go well with Radial Chart visualization design, we changed our visualization design and decided to go with bubble chart.

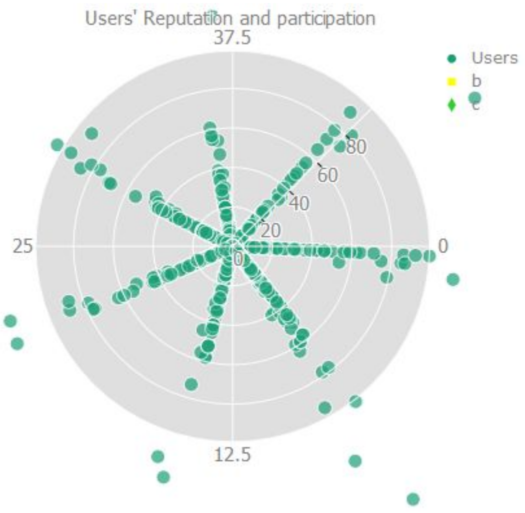


Figure 5. Radial Chart representing all users in R domain

We plotted bubble chart using d3 Force API [15]. The radius of circles denote the score of users in R domain and Color of data points denote total accepted answers. We have chosen gradient color. Lighter color represents less number of accepted answers than darker ones. Figure 6 shows our visual design of bubble chart.

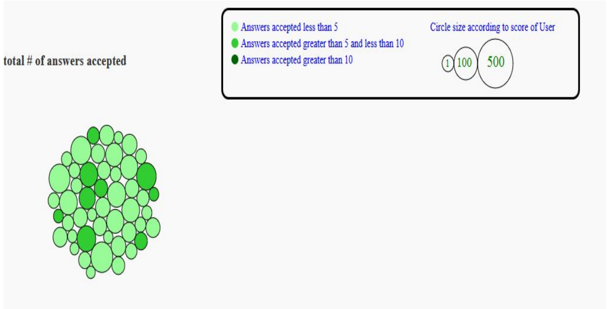

total # of answers accepted
● Answers accepted less than 5
● Answers accepted greater than 5 and less than 10
● Answers accepted greater than 10
Circle size according to score of User
1
100
500

Figure 6. Bubble Chart representing users in R domain

There were so many users in the dataset. So we decided to add slider for better visualization as shown in figure 7. Slider value represents number of posts given by users. By adjusting different values of slider, user can analyze users with particular number of posts.

Also, our visualization shows the number of users for different posts on a bar chart.

We also added the feature to combine and divide the users on the basis of accepted answers. So when the user will check on the check box, all the bubble points of the users will segregate on the basis of accepted answers. So the user can easily visualize that how many users are there for different range of number of accepted answers for particular number of posts which can be adjusted using slider. Figure 8 represents segregated users.

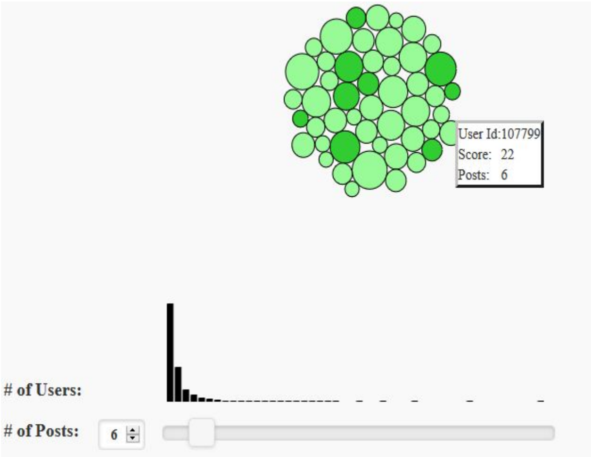


Figure 7. Slider to filter users depending on number of posts

It is easily visualized from figure 8 that there are large number of users who have accepted answers less than 5.

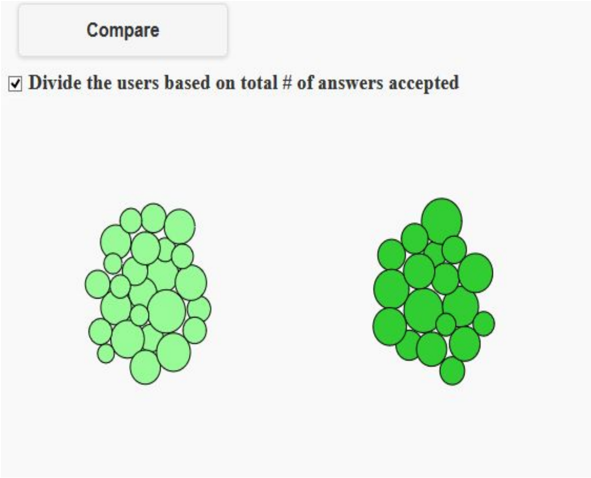


Figure 8. Checkbox used to divide users based on number of answers accepted

We have used following color scale to display the bubble points for different range of accepted answers.

i.e. Pale green 0< Number of Accepted_Answers <5

Lime green 5< Number of Accepted_Answers < 10

Dark green Number of Accepted_Answers > 10

We have added tooltip to display user_id, posts and user's score in R domain. For better analyzing user profiles, all attributes of the users must be displayed. It is not feasible to display all the information of a user on tooltip. Hence as shown in figure 9, we added horizontal bar chart to easily visualize all the attributes of users where color is used to distinguish the two users. In figure 9 comparison between profile information of two users with id 261 and 442 is displayed.

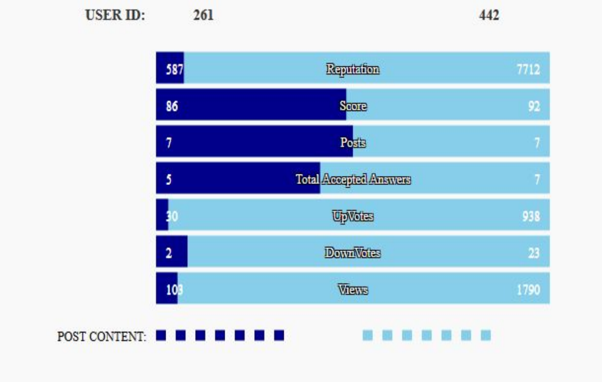
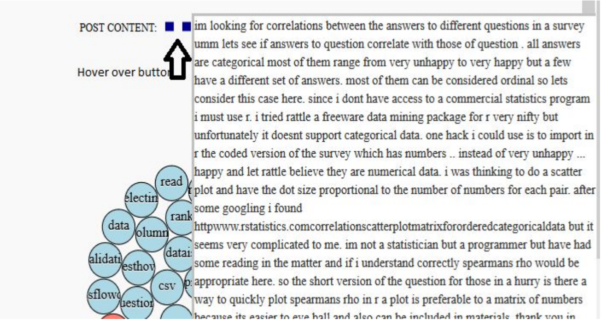


Figure 9. Horizontal bar chart to compare user profiles

Under this barchart we have also added boxes to display content of posts by the users. As shown in figure 10, when the user will hover over a box, the text for those particular posts will be displayed.



im looking for correlations between the answers to different questions in a survey umm lets see if answers to question correlate with those of question . all answers are categorical most of them range from very unhappy to very happy but a few have a different set of answers. most of them can be considered ordinal so lets consider this case here. since i dont have access to a commercial statistics program i must use r. i tried rattle a freeware data mining package for r very nifty but unfortunately it doesnt support categorical data. one hack i could use is to import in r the coded version of the survey which has numbers .. instead of very unhappy ... happy and let rattle believe they are numerical data. i was thinking to do a scatter plot and have the dot size proportional to the number of numbers for each pair. after some googling i found httpwww.rstatistics.comcorrelationscatterplotmatrixfororderedcategoricaldata but it seems very complicated to me. im not a statistician but a programmer but have had some reading in the matter and if i understand correctly spearmans rho would be appropriate here. so the short version of the question for those in a hurry is there a way to quickly plot spearmans rho in r a plot is preferable to a matrix of numbers because its easier to eye ball and also can be included in materials. thank you in

Figure 10. Each post of a user represented as a box. Tooltip shows the content of the post

We added extra feature for highlighting the bubbles using border line when they are clicked for comparison. So when any user will click on two bubbles and after that click on ‘COMPARE’ button, at that time Bar Chart to compare the user’s profile will be displayed.

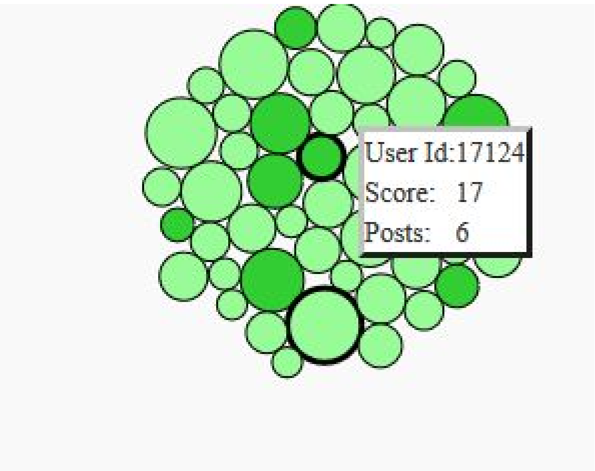


Figure 11. Selecting two users for comparison

After analyzing the user’s profile, we wanted to compare the text of posts for high scoring users and low scoring users in R domain. So we showed the words which the users have used in the form of bubble clouds.

Figure 12 shows bubble clouds for text content of the posts. Here red colored bubble indicates tag words and blue colored bubble indicates non tag words.

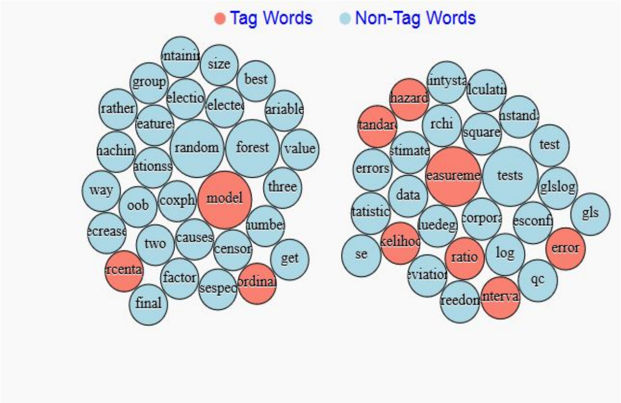


Figure 12. Text analysis of each user in form of bubble cloud

3.3 Text Analysis

For text analysis, we wanted to visualize the frequently used words by users who have low score in R with respect to users who have high score in R . For this we collected data using the following SQL query from the database.

```
SELECT OwnerUserId,
    STUFF(
       (SELECT DISTINCT ',' + Body
        FROM Posts
        WHERE OwnerUserId = a.OwnerUserId
        FOR XML PATH (''))
        , 1, 1, '') as Post_body
FROM Posts AS a where Id in (select PostId from PostTags
where tagid = 41) and OwnerUserId >0 and AnswerCount > 0
GROUP BY OwnerUserId
```

The data collected from above contained 5704 rows of R users on CrossValidated. But the results obtained were not accurate for text analysis.

Data contained a lot junk values which were required to be cleaned before text analysis. For data cleaning, we used R and Python scripts. In that, data gets loaded into a list where each list denotes a user. Now for each user in the list we have its user_id and posts. After extracting post of each user from the list, we performed various data cleaning tasks using libraries such as csv, sys, re (regular expression), nltk (natural language toolkit) numpy and stopwords.

Post of each user is first converted to lowercase. After that we removed various HTML tags, non-alphabetical characters, special character like #,$ using regular expressions. We also removed various single alphabets from the data except r as R is one of the tag word. After that we tokenized every post so as to treat every word of a sentence independently. Now we want to keep only important words and not the words like “the”,”is”,”a” ,”an” and many more. So for this we used stopwords library in Python, After removal of the stopwords from the posts, we performed data on the reduced data. Stemming changes the different words from same root to same word, for example after stemming walking and walked will be reduced to same word to walk. After this, we were left with only useful words in a post of every user. Now, we calculate word frequency in every post for various user. We also extracted a csv file which contained only tags in CrossValidated and added one more column in our data which contained words and its frequency which shows that whether a particular word is classified as a tag or not. Below is the Pseudo code for cleaning the data where file is a list containing all the instances user.

```
file<-read data  from csv
function clean_data:
        data<- get data from file
        convert data to lower case
        use regular expression to get only alphabetical characters
        tokenize the data
        remove punctuation marks
        remove all single alphabets except R
        remove stopwords
        perform stemming
        get count of each word now
        append “yes” or “no” if its a tag or not respectively
        get top 20 words for each user
        append in a file
for each instance of file:
        call clean_data
```

So our initial data for a user looked like this:

Table 1: data without cleaning

user_id	Post
2917	&lt;p&gt;A basic rule of thumb for when to neural networks is, if you# (as a human) can$ see what the right answer is quickly,...............

After cleaning the data, it contained four fields, word, frequency, user_id and yes/no for whether it is a tag or not

Table 2: data a of single user after cleaning it.

word	frequency	user_id	yes/no
regression	6	2917	no
R	3	2917	yes
neural	2	2917	yes
networks	1	2917	no
model	1	2917	no

After cleaning of data, we saved only top 20 most frequently occuring words of each user. In the end, we got 5704*20 = 114080 rows similar to as shown in Table 2. The final data was imported into a CSV file. After which top 20 frequents words are showed and compared for two selected users. Initially, we compared frequent words for selected two users by visualizing words in form of word cloud. However, comparing text for two different users on word cloud was not providing clear visualization. Hence, we moved to bubble cloud. Frequent words were plotted in form of Bubble Cloud. In each cloud a bubble is colored red when word corresponding to that bubble is one of the tags. And, bubble is colored blue when word corresponding to that bubble is not from the tags list. Each cloud shows top 20 frequent words for that particular user.

4. METHODOLOGY

This section outlines our research questions and the analysis methods.

4.1 Research Questions and Analysis

The focus of this visualization is to compare any two Cross Validated users based on their posts, scores, votes, views and number of answers accepted. In order to find the factors affecting the score of a user it is important to analyze and compare two users and find the differences and commonalities so as to better understand the progress of a user. It is only when we compare a user with other users, that we understand if the user is performing better or worse.

Our next research question is whether the use of tag words in post affects the score of a user? We wanted to find whether the high scoring users are using more tag words in their posts. The tag words are used in posts to categorize the post in a topic. By categorizing this way, when anyone searches for that topic, the post appears. This helps increase the views to that post and possibly lead to a vote. So, the use of tags words relevant to the post may help the user increase their views and score.

By performing text analysis, we analysed the post context of users. After depicting the results of text analysis on a bubble cloud, it was visible that users were not using tag words purposefully in their posts in order to receive higher views. There was no relation found between the use of tag words by users and their views or score.

Our final research question is to find a correlation between any two features of the user profile like score in a domain, views, votes, tags used, reputation, answers accepted and posts.

Analysing the users based on number of posts and answers accepted showed us a positive correlation between the number of answers accepted and the number of posts given by the user. So, a user can improve their score in R domain by increasing the posts made and improving the quality of the answers in order to increase its chances of being accepted by the asker. A good quality answer is one which is written-well, suggests a good practise and is helpful to the asker.

5. EVALUATION PLAN

Good Visualization is one that helps user understand and explore data without any extra efforts. Visualization should convey maximum information with minimal complexity.

For good visualization following criterias should be fulfilled. Main idea of the visualization should be clear and distinct. Visualization and dashboard should be simple and understandable. Including complex graphics and overwhelming the visualization with design decreases the quality of visualization. Most important features must be visualized at the first glance. Right visual should be used according to the purpose [25]. Colors given in the visualizations must have some meaning and relation with data. Visualization should be interactive and interesting. It should keep user engrossed and engaged. Scale and legend should be clearly stated so that viewer can understand the attributes of visualization easily. A good visualization should convey some interesting story in order to be meaningful.

In our visualization we have taken care that the dashboard doesn't become complex. It is simple enough that the viewer can have an easy walk through the visualization and dashboard. As, for good visualization proper design should be selected according to the purpose. In our visualization we have used bubble chart to analyze user and scale to divide users across number of posts as there were huge number of users to analyze together. We have also used horizontal chart to compare two different users which makes it easy for viewer to get clear idea of both the user's profile. For text comparison we have used bubble cloud instead of word cloud as effective analysis of result was not possible with word cloud. For effective use of color we have used shades of green to analyse number of accepted answers of the user. If bubble of the user is colored with darker shade of green, it implies that the user's profile is of expert level as his/her more number of answers are accepted. Whereas, if a user's bubble is colored with lighter shade of green it depicts that his/her profile is still at novice level. Scale and legend in this visualization is clearly mentioned at the upper right corner of visualization. Also, after the bubble clouds are shown for text comparison between two different users, legend for that bubble cloud can also be seen clearly. To clear convey every information we have also included the feature of tooltip. If a word is not visible in the bubble cloud because of if larger length, one can clearly view the word in tooltip. Moreover, this visualization is interactive enough for a viewer to play around the visualization. It can keep viewer interested and to explore more.

We have considered all of these criterias and included them in our interactive visualization to match with standard criterias which are required for good visualization.

6. CONCLUSION AND DISCUSSION

While analysing users using our interactive visualization, we came across about some of the following interesting facts.

There are significant amount of users with number of posts less than 5 in R domain. Number of skilled, high scoring users in R are only 3% of all the users. Users with same number of posts differ in their score because of their number of accepted answers.

Even if some of the users have posted very less number of posts, their score in R domain is very much high because of large number of view counts and upvotes, which supports the fact that the quality of the post and the views it receives play an major role in developing the user's score in R domain. For example. the user with id '890' has only 1 post. But his score in R domain is 296.

We found that there is a positive correlation between the amount of answers accepted and the number of posts given by the user. As discussed in methodology, we analysed different word clouds to validate our assumption of high scoring users might use high frequency of tag names in their posts. But we could not find any such scenarios from word clouds of posts of high scoring users.

By implementing this comparative interactive visualizations, we could easily visualize profiles of different users and compare different attributes like score in R domain, total number of posts, total accepted answers and upvotes achieved by the users. So that we could analyze different users easily and visualize how better they are in different domains from others.

7. FUTURE WORK

This project could also be further implemented by ranking the users based on the Elo Ranking System. The Elo Ranking system can be modified for multi-players. By ranking the users, we can determine their skill based on a rank value. Another addition to this visualization is to create a heat map for categorizing users based on their reputation.

This visualization can be extended to become a visual recommender allowing each user to understand where they stand when compared to other users. Also, a recommendation can be provided to the user which contain posts the user can possibly answer based on their expertise to increase their score in a domain. Featured posts can be recommended to the user giving them a chance to increase their reputation by answering questions which are high in demand .

Finally, the size and the growth trends of each tag can be compared to understand the most demanding and trending domains in Cross Validated.

8. REFERENCES

[1] http://atheonanalytics.com/visual-analytics/

[2] http://www.visual-analytics.eu/faq/

[3] http://api.stackexchange.com/docs

[4] http://data.stackexchange.com/stats/queries

[5] https://bost.ocks.org

[6] http://bl.ocks.org/mbostock

[7] http://www.lifewithalacrity.com/

[8] http://stackrating.com/stats

[9] http://omnipotent.net/jquery.sparkline/#s-docs

[10] https://www.youtube.com/watch?v=FUJjNG4zkWY

[11] http://jstricks.com/animated-circular-statistics-jquery/

[12] http://stackoverflow.com/questions/21945339/rating-system-for-multiple-competitors

[13] https://codepen.io/

[14] http://www.niemanlab.org/2011/10/word-clouds-considered-harmful/

[15] http://betterevaluation.org/en/evaluation-options/wordcloud

[16] https://www.surveygizmo.com/survey-blog/what-you-need-to+-know-when-using-word-clouds-to-present-your-qualitative-data/

[17] https://github.com/d3/d3-force#forces

[18] http://valt.cs.tufts.edu/pdf/griffin2011analytical.pdf

[19] http://www.nytimes.com/interactive/2012/09/06/us/politics/convention-word-counts.html?_r=1&#!

[20] http://alignedleft.com/tutorials/d3/making-a-bar-chart

[21] http://meta.stackexchange.com/questions/255830/has-anyone-tried-to-estimate-stack-overflow-users-skill-by-analyzing-the-data-du

[22] http://blog.blprnt.com/about

[23] http://vis.cs.ucdavis.edu/papers/wu_semantic-preserving-word-cloud.pdf

[24] http://geoffreyrockwell.com/publications/WhatIsTAnalysis.pdf

[25] https://www.gooddata.com/blog/5-data-visualization-best-practices