User Roles in Enterprise Collaboration Systems

Rahul Bhanushali

University of Koblenz-Landau

rbhanushali@uni-koblenz.de

Abstract

Enterprise collaboration is a type of system where employees/people can communicate virtually to collaborate on corporate projects or for social networking. Such a collaboration platform enables people to share information from remote geographical locations using various software technologies. In such platforms, people perform different types of tasks differently. Hence, it is extremely important to introduce different roles which users perform while using Enterprise Collaboration Systems (ECS) such as Computer Supported Cooperated Work (CSCW), Social media platforms, Project Management platforms, etc. to improve virtual collaboration among people. In this paper, we use a Clustering algorithm to determine various user roles based on different user characteristics such as usage pattern, frequency of use, type and number of connections with other users, etc. We successfully represent results from the clustering algorithm to determine various user roles based on different characteristics.

Keywords—User Roles, Clustering, Collaborative systems, Stack Overflow

1. Introduction

Enterprise Collaboration Systems (ECS) support employees in all areas of their joint work and are an important enabler of the modern digital workplace (Schubert and Glitsch, 2016). Over the last decade, we have witnessed the emergence of a new type of collaboration software, the so-called "Enterprise Social Software" (ESS). The use of social media in private life has changed the way people communicate and exchange information (Schwade and Schubert, 2018). Petra Schubert and colleagues (Schubert and Glitsch, 2016) convey that Enterprise Social Systems (ESS) will soon become a necessary component of the basic IT infrastructure, especially in innovative and service-oriented companies. The growth of social media usage opens up new opportunities for analyzing several aspects of, and patterns in communication. For example, social media data can be analyzed to gain insights into issues, trends, influential actors, and other kinds of information (Stieglitz *et al.*, 2018). To perform various collaboration activities successfully, determining and knowing the roles and responsibilities is most important. An unclear role specification may create dysfunctional ambiguity and conflict in an organization(Zhu, 2006).

Integrated collaboration systems provided by multiple large software vendors such as IBM, Microsoft, etc. combine collaboration and social features in one platform and establish uniform access for employees (single-sign-on, uniform user interface). Integrated collaboration systems also provide activity logs that allow us to analyze multiple forms of collaborative work in the digital workplace in a new way, which presents a great opportunity for researchers to explore and better understand the collaboration activities that are going on in companies (Schwade and Schubert, 2017).

In this literature review, we try to define user roles by forming clusters of users performing similar activities or following similar patterns. The goal of clustering is to separate a finite, unlabeled dataset into finite discrete data and provide hidden data structures (RUI XU, DONALD C. and WUNSCH). The authors of this (RUI XU, DONALD C. and WUNSCH, no date) also define Cluster in simple terms as a set of entities that are alike and entities from different clusters which are not alike. Popular notions of clusters include groups with small distances between cluster members, dense areas of the data space, intervals, or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure(Wikipedia).
In this literature review, we are considering social media as a platform for our Cluster analysis to understand the behavior of users on social media based on different characteristics and define user roles accordingly.

Therefore, we define the following research questions:

1. What are the important dimensions of users to be considered while applying Social Media analytics?
2. What are the possible roles of users based on different dimensions of users on social media?

The growth of social media usage opens up new opportunities for analyzing several aspects of, and patterns in communication. In the field of Information Systems (IS), social media data is used to study questions such as the influence of network position on information diffusion (Stieglitz *et al.*, 2018).

Answers to the above research questions using multiple clustering algorithms (Borkar, Patel and Moparekar, 2019) will help us improve the communication between different users to share the data, career interests, ideas, and collaboration. Such cluster analysis will help developers and creators to detect communities on social media platforms based on similar interests and provide a better user experience and provide useful recommendations to communities.

This paper is structured as follows: First, we introduce terminology and related work on cluster analysis using multiple algorithms (section 2). Section 3 describes the overview of the research design followed for this literature review. We then implement different clustering algorithms (Borkar, Patel and Moparekar, 2019) on the social media dataset and perform cluster analysis. We conclude with an analysis of the results and comparisons of results from different algorithms.

2. Terminology & Related Work

ECS are knowledge systems that use several information technologies which provide good communication, coordination, and collaboration in organizations. The collaborative software is generally named groupware and it can be categorized as enterprise communication, collaborative, enterprise conferencing, and work management tools (Prakash *et al.*, 2020). IT-based collaboration is a tool that creates a workflow of information to a specific team and representative and allows exchanging ideas.

Types of ECS are "socially-enabled", providing social media functionality such as recommend, like, follow, tag, or rate, which are used on content items such as social profiles, microblogs, wiki pages, blog posts, files, or tasks (Schwade and Schubert, 2019). Term social media is a (public) platform for social interaction and information exchange. Such platforms are characterized by their openness (any interested person can register and use the platform) and by their ownership (they are usually provided by a company that owns the platform and, in most cases, using terms and conditions, also the user-generated content). In social media, people gather voluntarily in their free time to chat, exchange ideas, play together, and, most importantly, share information (photos, films, files) (Schwade and Schubert, 2017).

As a common word. 'role' is easily understood. However, as a concept, there is no fundamental discussion and clear definition of roles. There are many different role concepts applied in different systems (Zhu, 2006). Different researchers defined the term 'role' in different ways and corresponding to different systems. Zhu (Zhu, 2006) described the term *role* in collaborative systems. He describes a role as a category of users within the user population of a given application and all users in a certain role inherit a set of access control rights to objects within the applications. Based on labels of user roles, the system designer's user case-like structures define the different working processes for different roles (Zhu, 2006). With this view, once a collaborative system with roles is completed, it is very difficult to adjust the roles in the system, although the requirement of tuning roles is very common in collaboration.

In academic literature, the terminology in the area of Cluster Analysis is described. Cluster analysis is a generic name for a variety of mathematical methods, numbering in the hundreds, that can be used to find out which objects in a set are similar(Charles Rosemburg, 2004). Clustering is considered a well-known unsupervised learning technique that deals with unstructured data. Given unlabeled data set, it aims at categorizing the data objects into different groups (also known as clusters) based on some similarity measures like the distance between data points, characteristics that describe objects, etc. (Maniriho and Effendi, 2018).

3. Research Design

Understanding Social media users and grouping them according to their roles is important to understand the

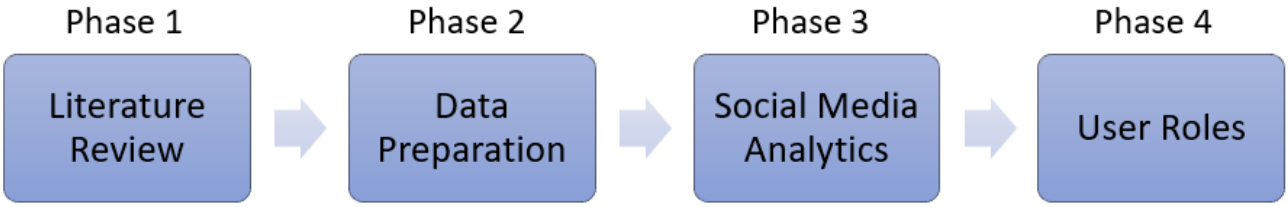

Figure 1: Research Design to define User Roles

current trends, for better recommendations, to analyze the key features of the platform, and improve them if necessary. Clustering algorithms such as Agglomerative Clustering, K-means clustering, and hierarchal clustering can help us analyze the user behavior in social media and to group them. User roles and behaviors differ in different collaboration platforms. Hence, to prepare a foundation for our analysis in social media using clustering algorithms, we organized our research design in four phases.

Figure 1 describes the flow of the research process which to define the user roles in the Collaborative system. In the first phase, we did a review of the literature study in a similar field. After the first phase, we moved to the second phase: the Data Preparation part in which data collection, data preprocessing, and data understanding were performed. After data was cleaned and preprocessed, we performed Social Media analytics in the third phase to get a better understanding of user behavior and important factors in determining user roles. In the last phase, we identified user roles based on information gathered from phase 3. Implementation of each phase in detail is mentioned in the next sections.

4. Detailed steps to determine User Roles

4.1 Literature Review

The first is the Literature review phase. For this phase, we started searching the commonly used keywords ‘User Roles in ECS’, ‘Cluster Analysis’, etc. in online databases such as Google Scholar, Mendeley, IEEEE Explore, etc. Later we also searched for related keywords such as ‘Community Detection’ and ‘Social Media Analytics’ to focus more on the core idea of the literature topic. In this phase, we reviewed a few literature reviews to establish a foundation for our analysis.

Brandtzæg in his research (Brandtzæg, 2010) defined Media behavior as the totality of human behavior about new media use, including both differentiated levels of participation (frequency of use) and content/activity preferences in media usage (forms of use). He categorized users into distinct user types that describe the various ways in which individuals use different media, reflecting a varying amount of activity/content preferences, frequency of use, and variety of use.

Brandtzæg (Brandtzæg, 2010) provided an overview of 22 media user typologies from the year 2000. One of the aims of this article is to detect the changing patterns of media usage over time; however, it is difficult to determine whether the characteristics of the typologies have changed over time (Brandtzæg, 2010). User roles will provide a more precise approach for the collaborative community to understand and identify users and to measure the heterogeneity of media behavior. A user typology describing the construction of a sub-group or user type, based on user activity; preferences/content selection; and the frequency and variety of use will not only contribute to a clearer view of diverse media behavior but will also indicate how people differ in their digital competence and how this might develop over time. Brandtzæg (Brandtzæg, 2010) defines eight types of users based on four dimensions for social media platform usage. Media use of sporadic is characterized by low frequency and variety of use. According to his User topology, media usage of lurkers is characterized by a medium frequency of use and a low variety of use. Entertainment users or socializers use a media platform for entertainment and connecting with other people. debaters participate in discussions and instrumental users use a media platform as a tool for a special purpose. Finally, Brandtzæg (Brandtzæg, 2010) defines advanced users as the users who are most active and use the most features of a media platform (Schwade and Schubert, 2019).

B. Armor and his colleagues (Amor *et al.*, 2016) stated that community detection is the graph-theoretical problem of identifying meaningful subgroups within a network. Many researchers performed a large number of algorithms to perform community detection. The variety of community detection methods reflects the fact that there cannot be a universal definition of what constitutes a 'good' partition of the network. However, most methods follow heuristics based on structural and combinatorial features of the network: typically, a subset of nodes is thought of as a good community if the connections between the nodes within the subset are denser than the connections with nodes outside of the subset (Amor *et al.*, 2016). B. Armor and his colleagues (Amor *et al.*, 2016) computed Bridgeness to identify the users important for information flow between two communities, we compute the shortest paths for all pairs of nodes (i, j) where $j \in C1$, $i \in C2$ and identify the between-community edges which feature in these shortest paths most often.

The user roles proposed by (Muller *et al.*, 2009) include lurkers, contributors, and uploaders. Lurkers only consume content. In contrast to this, contributors "do not upload files, but they do create metadata about files through actions such as commenting, sharing to specific other users, adding files to named collections of files, and adding tags to files such as downloading files". Thus, contributors rather create metadata about files. Finally, uploaders "create files in the service through upload operations" and thus create primary content (Schwade and Schubert, 2019).

As a result of this first phase (literature review), we identified major factors to define user roles in a Collaborative system as shown below in Table 1:

Factors/Dimension	Definition	Ref
Frequency of Use	How often do users use a platform	(Schwade and Schubert, 2019)
Content Preferences	Chosen content type	(Brandtzæg, 2010; Schwade and Schubert, 2019)
Platform Preferences	Chosen platform type	(Brandtzæg, 2010; Schwade and Schubert, 2019)
Betweenness	How influential is the user	(Amor *et al.*, 2016)
Variety of Users	Different types of Users	(Schwade and Schubert, 2019)

Table 1

Factor *Frequency of use* refers to how often and how long users use a platform. This is considered to be one of the most important dimensions to define user roles (Schwade and Schubert, 2019). On the other hand, dimension *Content preference* describes the type of content users prefer and the dimension platform preference describes the platform, which is preferred by users (Brandtzæg, 2010; Schwade and Schubert, 2019). Factor *Betweenness* refers to the number of in-degrees and out-degrees of a user in the social network of that particular platform. This dimension helps us to determine which group of users are influential and which are not in the social network. Community detection methods are used to compute *Betweenness* (Brandtzæg, 2010). Dimension *Variety of Use* states different types of users and their purpose of using the platform (Schwade and Schubert, 2019).

The results and discussions from literature reviews of (Muller *et al.*, 2009; Brandtzæg, 2010; Amor *et al.*, 2016; Schwade and Schubert, 2019) will serve as a starting for our analysis of stack overflow users and cluster them into roles based on type and frequency of usage.

4.2 Data Collection, Data Preprocessing and Data Understanding

4.2.1 Data Collection

The literature review phase provided the foundation for the second phase Data Preparation for gathering and preparing data of social media for analytics. In this phase, we gathered data of the Collaborative platform: Stack Overflow. Stack Overflow is a question-and-answer website for professional and enthusiast programmers (Wikipedia). Using this platform, users can post questions and answers on particular topics and thus can collaborate by sharing ideas and answers.
To understand more about the community and get answers to the questions such as 'How and when are users from all over the world are using Stack Overflow?', Are there any noticeable programmer communities on Stack Overflow?'. Data was gathered in a Stack Overflow Annual Developer Survey of 2020 to examine all aspects of the developer experience from career satisfaction and job search to education and opinions on open-source software. They ask questions like: Do you code as a hobby, are you currently enrolled in a formal degree-granting college or university program, and many other questions. The respondents were recruited through onsite messaging, blog posts, email lists, meta posts, banner ads, and social media posts. Highly engaged users on stack overflow were more likely to notice the links for the survey and click to begin it. This dataset is freely available on the Kaggle platform. Also, datasets about users, posts, comments, tags, votes, and badges from 2008 to 2020 are hosted on Google BigQuery.

As we know that, Stack overflow is a question-answer website for professionals to collaborate on different questions and problems and share ideas. Data about questions and answers play an important role while performing analysis. Questions from 2016 to 2019, with information like question id, answer count, creation time, score(upvote), owner user id, etc. and answers from 2016 to 2019, with information like answer id, creation time, parent(question) id, score(upvote), owner user id, etc. are hosted on Google BigQuery.

4.2.2 Data Preprocessing

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. It is not always that we come across clean and formatted data. While performing any operation with data, it is mandatory to clean it and put it in a formatted way.

A real-world data generally contains noises, missing values, and maybe in an unusable format that cannot be directly used for applying machine learning models (such as clustering algorithms). Data preprocessing is a required task for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model(Javapoint).

Salvador and colleagues in a research paper (García *et al.*, 2016) stated the importance of Data preprocessing and different efficient ways to implement data preprocessing. According to Salvador and colleagues, a set of techniques used before the application of a data mining method is named data preprocessing for data mining and it is known to be one of the most meaningful issues within the famous knowledge discovery from data process (García *et al.*, 2016). In his research paper, multiple steps of data preprocessing are described as shown below in Figure 1.

Missing values treatment is difficult. Inappropriately handling the missing values will easily lead to poor knowledge extracted and also wrong conclusions (García *et al.*, 2016). Missing values have been reported to cause loss of efficiency in the knowledge extraction process, strong biases if the missingness introduction mechanism is mishandled, and severe complications in data handling. For our case study, we handled missing values by removing them from the dataset as we need all answers from the survey for better analysis. Please note here, that we have not removed the complete record having missing data but only the missing value in that particular attribute.

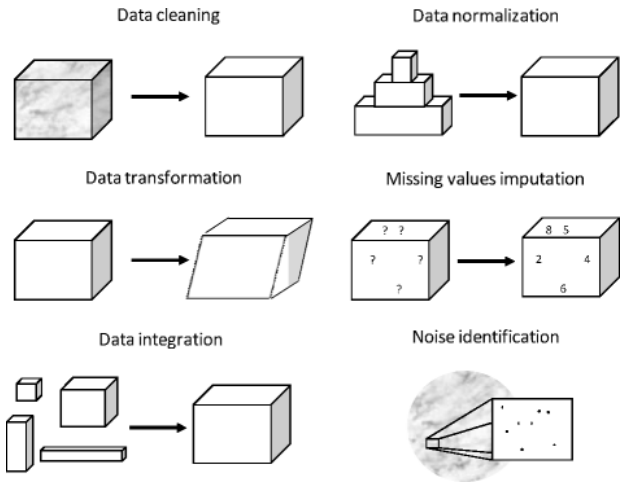

Figure 1

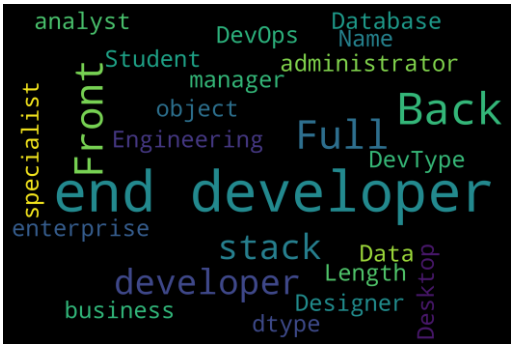

Figure 2

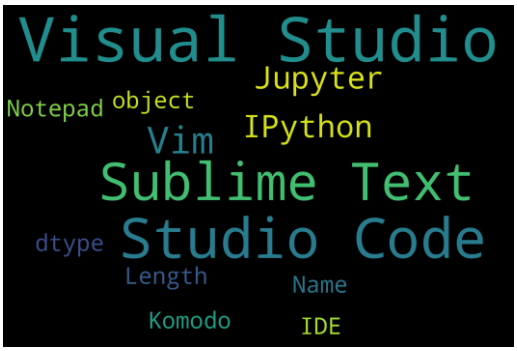

Figure 3

4.3 Social Media Analytics

After the data preprocessing phase, we performed social media analytics on the dataset to get insights about the types of users, any trend patterns, cross-relationship between different features, etc. Social Media Analytics is an emerging interdisciplinary research field that aims on combining, extending, and adapting methods for the analysis of social media data (Stieglitz *et al.*, 2014). There are a large number of different social media applications or platforms which in general can be categorized as weblogs, microblogs, social network sites, location-based social networks, discussion forums, wikis, podcast networks, picture, and video sharing platforms, ratings, and reviews communities, social bookmarking sites, and avatar-based virtual reality spaces. In a broader sense, social media refers to "a conversational, distributed mode of content generation, dissemination, and communication among communities. Stephan and Christoph have described different methods of social media analytics (García *et al.*, 2016). a. In the context of SMA, three main analysis methods almost always find their application: (1) text analysis/mining, (2) social network analysis, and (3) trend analysis.

4.3.1 Text analysis/mining

Text analysis/mining is a research technique within the field of content analysis that supports researchers in making replicable and valid inferences from texts to the contexts of their use (Stieglitz *et al.*, 2014). Automated quantitative methods of text analysis are required because of the massively growing amount of social media data. Based on these methods a broad variety of questions can be answered, among which are the classification of texts (i.e., most frequent words) and the identification and modeling of recurring topics. One important subfield of text analysis/mining is sentiment analysis or opinion mining, which has emerged as a distinct method to study people's opinions in terms of views, attitudes, appraisals, and emotions towards entities, individuals, issues, events, topics, and their attributes in a more thorough way (Stieglitz *et al.*, 2014).

Stack overflow being a question-answer website for the collaboration of enthusiasts and professional programmers, it is important to identify the different frequencies and types of developers using the platform. To achieve that we implemented the Word Cloud method as shown in above fig. 2 and fig. 3 on responses to survey

questions- Developer Type (DevType) and integrated development environment (IDE) they work upon. As a result, we found that majority of Frontend and Backend developers are using the platform followed by administrators and specialists. We can also see that Designers and enterprise developers are less in numbers in using the platform. From fig. 3 we can determine that majority of users use Visual Studio and Sublime text as IDE followed by Python. These results can be helpful to detect communities based on the type of Developers and IDE preferences in the Stack overflow.

4.3.2 Social network analysis (SNA)

The second main analysis method is social network analysis (SNA), which studies the relationships between persons, organizations, interest groups, states, etc., by analyzing the structure of their connections (Scott and Carrington 2011). In an SMA context, SNA may help identify influential users or opinion leaders, and relevant user communities in social media. There are several different measures for the influence of an actor in a network. SNA thereby provides different metrics for the concept of centrality and prestige that can be applied to measure influence (e.g., degree, betweenness, or eigenvector centrality; degree, proximity, or rank prestige) (Stieglitz *et al.*, 2014). SNA might also be useful with different community detection methods and algorithms (e.g., graph theoretical approaches such as the Girvan–Newman algorithm or other clustering methods such as hierarchical, k-means, and fuzzy c-means clustering) (Stieglitz *et al.*, 2014).

Since Stack Overflow is a global community, it is not surprising to receive answers from the other side of the globe. Below fig. 4 illustrates the interaction between questions raised and answered in a different country. The direction of the arrow indicates the number of answers from the source node to questions from the destination nodes. From the below country-specific network graph we can say that activeness of countries like the United States and India is higher comparatively, but other countries also have noticeable contributions.

4.3.3 Trend Analysis

Trend analysis is the third main analysis method that makes use of recent advances in computer science and statistics to predict emerging topics. Many trend-detecting algorithms are based on so-called hidden Markov models where observations of topics are trained by such models which in turn are saved in a library for the topic's prediction (Stieglitz *et al.*, 2014).

To understand the daily changes of activity of the site, the above plot in fig. 5 shows the number of questions asked per day on the Stack Overflow platform. We get an interesting pattern where the number of questions decreases to almost half during the weekends and rise back again during working days. We can also see that the number of questions asked also drops to lower amounts once a year during Christmas and New Year's Eve. This trend can be helpful to determine the frequency of usage of overall users in the platform.

By performing social media analytics, we can gather some useful insights about the user data. This data helps us to understand the trends. Analyzing the data helped determine the preferences of users (platform, technology).

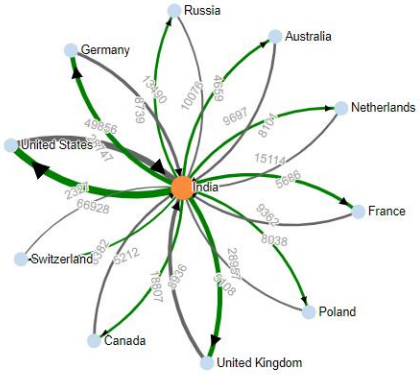

Figure 4:

Reference: [Stack Overflow as a Social Network (stack-overflow-as-a-social-network.github.io)]

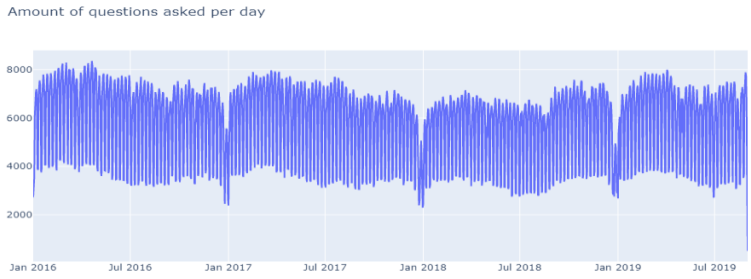

Figure 5

Reference: [Stack Overflow as a Social Network (stack-overflow-as-a-social-network.github.io)]

4.4 User Roles

With the help of analysis done in the Analytics phase, we can determine different possible user roles in the collaborative platform- Stack Overflow. Based on the dimensions described in the Literature review phase we identified below possible user roles in the Stack Overflow platform.

4.4.1 Contributors

A contributor is a user who contributes to existing content, for example by posting frequent questions and ideas on the platform or by providing more answers to the questions posted on the Stack Overflow platform. Contributors play an important role in such question-answer-based platforms by providing solutions or ideas to existing questions and problems. Analysis of the number of questions and number of answers from 2016 to September 2019 by grouping them with the help of unique user id is shown in below fig. 6 and fig. 7.

From the below charts, we can see that both numbers of questions and answers given by users of Stack Overflow roughly follow a power-law distribution. Although a general user only asks a few questions / provide a few answers, some users contribute mainly to the question-and-answer pool. There exists one person who asked over 1200 questions in the 4-year period, which is over one question per day. The phenomenon is even more apparent for the answer distribution, where there are people who gave more than 10000 answers over 4 years (around 6 / 7 answers per day), and one even provided 35219 solutions in 4 years (over 24 answers per day!!!). Us being only visitors of Stack Overflow, we are definitely "carried" by those significant contributors of the community(Xreator, 2019).

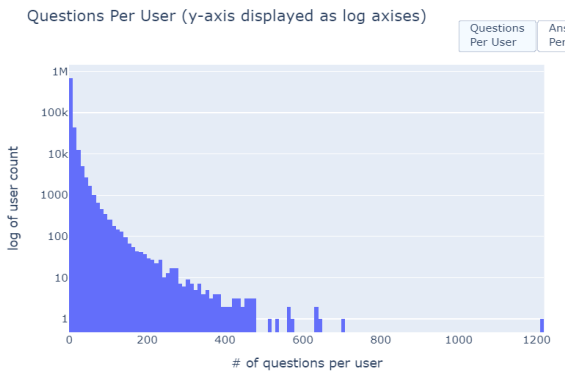

Figure 6

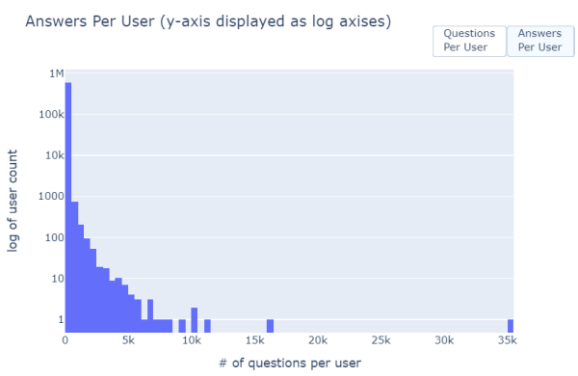

Figure 7

Reference: [Stack Overflow as a Social Network (stack-overflow-as-a-social-network.github.io)]

StackOverflowVisit	A few times per month or weekly	A few times per week	Daily or almost daily	Less than once per month or monthly	Multiple times per day	**Total**
Multiple times per day	8.05%	5.93%	4.56%	11.90%	3.68%	**34.11%**
Daily or almost daily	10.35%	5.78%	2.56%	15.14%	0.23%	**34.06%**
A few times per week	6.46%	2.20%	0.09%	12.30%	0.01%	**21.07%**
A few times per month or weekly	2.33%	0.11%	0.02%	6.97%	0.01%	**9.44%**
Less than once per month or monthly	0.09%	0.02%	0.00%	1.20%	0.00%	**1.32%**
Total	**27.29%**	**14.03%**	**7.24%**	**47.50%**	**3.94%**	**100.00%**

Figure 8

4.4.2 Listener

Figure 8 shows the cross-tabulation analysis of two survey questions- 'StackOverflowParticipate' and 'StackOverflowVisit'. StackOverflowVisit is displayed as rows indicating the number of unique users visiting the platform and StackOverflowParticipate are the columns indicating the users participating in Question-Answer on the platform. Cross tabulation is a method to quantitatively analyze the relationship between multiple variables. It is usually used in statistical analysis to find patterns, trends, and probabilities within raw data. This type of analysis is usually performed on categorical data which can be divided into mutually exclusive groups.

In fig. 8 values shows the percentage of a grand total of that particular attribute. From the cross tab in fig. 8, we can determine that there is a significant percentage (15.14% and 12.30%) who frequently visit the Stack Overflow platform but participate in the question-answer less than once per month. This type of user can also be considered as Lurkers who visit the platform but never contribute to it significantly (Schwade and Schubert, 2019). These are the users who actively consume the data but never contribute. The majority of them can be just knowledge seekers who just visit the platform to gain knowledge and learn from existing data on the platform.

4.4.3 Professional

Figure 9 illustrates the chart of the count of users according to years of experience in coding. Experience plays a crucial role while providing solutions to the questions on the Stack Overflow platform. The reliability of solutions also depends on the experience of a programmer. However, we get interesting insights from the below chart. We can determine from the below chart that the majority of users of the platform are having 0 to 5 years of experience in coding. Similar results can be drawn from cross tabular relation shown in fig. 10 between attributes years of coding as a professional and StackOverflowParticipate. This shows that more experienced people in coding are not a significant part of this platform. We can say that the majority of users of this platform are beginners in coding or enthusiast learners. Hence, this feature – experience of coding can be a driving feature to cluster users based on professionalism.

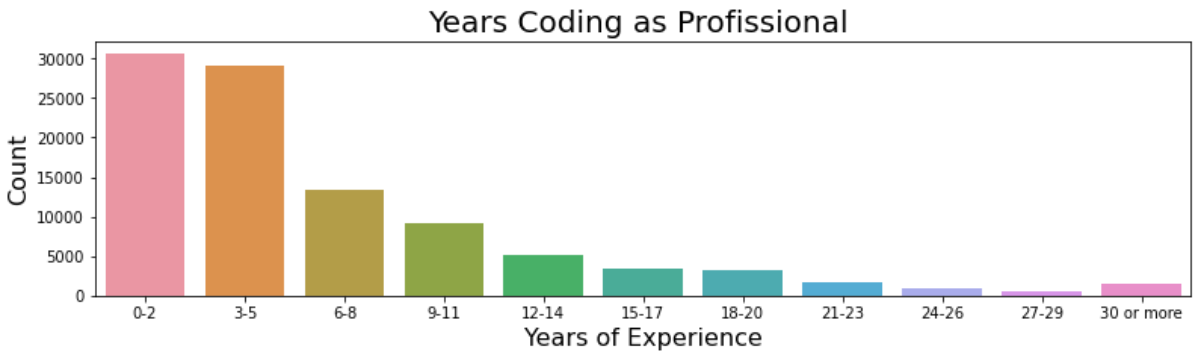

Figure 9

StackOverflowParticipate YearsCodingProf	A few times per month or weekly	A few times per week	Daily or almost daily	I have never participated in Q&A on Stack Overflow	Less than once per month or monthly	Multiple times per day
0-2	0.267153	0.393727	0.259854	0.382988	0.279754	0.233956
12-14	0.060226	0.046187	0.054197	0.040687	0.056345	0.053650
15-17	0.037718	0.027368	0.045985	0.025694	0.037454	0.046520
18-20	0.035444	0.028721	0.038869	0.025996	0.035975	0.044143
21-23	0.018480	0.015006	0.018796	0.010580	0.016071	0.031239
24-26	0.010188	0.007811	0.011496	0.008403	0.009692	0.016638
27-29	0.008434	0.003444	0.006752	0.004836	0.005530	0.010187
3-5	0.296010	0.259041	0.297263	0.311045	0.302533	0.294397
30 or more	0.015116	0.013776	0.020985	0.013119	0.014538	0.030900
6-8	0.147271	0.116421	0.143796	0.110211	0.145845	0.139898
9-11	0.103961	0.088499	0.102007	0.066441	0.096263	0.098472

Figure 10

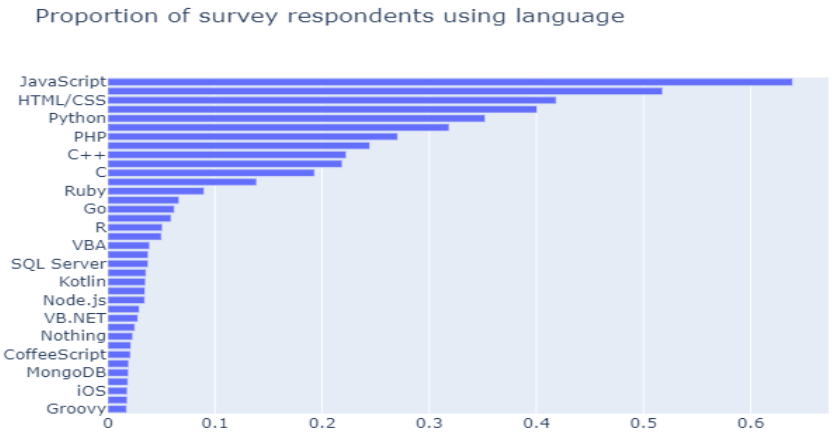

Figure 11

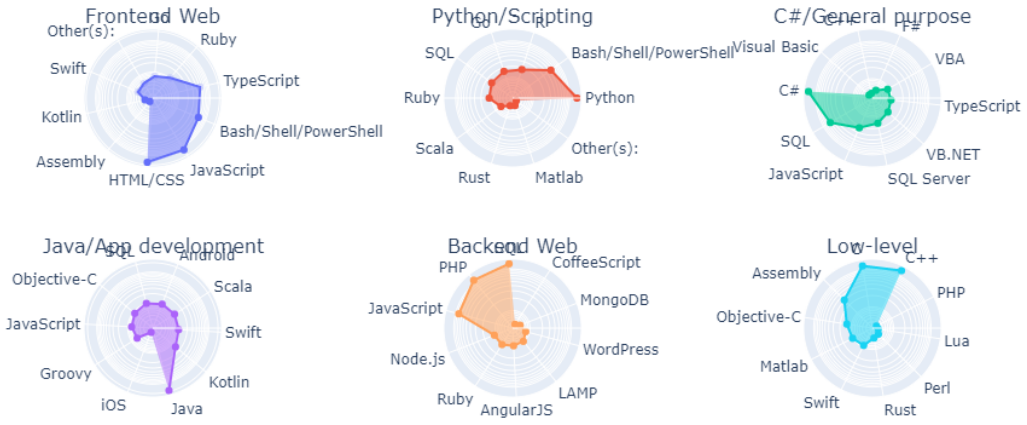

Figure 12

Reference: [Stack Overflow as a Social Network (stack-overflow-as-a-social-network.github.io)](stack-overflow-as-a-social-network.github.io)

4.4.4 Content Specific

Soft Clustering was performed on the users and their programming language use. The idea was to extract programmers' archetypes to which all programmers can be divided, with each programmer having a separate score for each persona. We decided on 6 clusters because that resulted in a good balance of interpretable results and keeping the amount types not too high. In fig. 12 we can see the scores for how much is each programming language associated with each type(Xreator, 2019).

5. Conclusion

Different factors dominate in different types of collaborative platforms. Hence, performing a literature review to identify the objectives and potential features is very crucial. By identifying dimensions and factors for defining roles in the literature review phase and by analyzing and applying clustering techniques on the stack overflow data based on factors defined in the first phase we identified below possible user roles in the Stack Overflow:

- Contributor
- Listener
- Professional
- Content Specific

6. References

Amor, B.R.C. *et al.* (2016) "Community detection and role identification in directed networks: Understanding the Twitter network of the care. data debate," in *Dynamic Networks and Cyber-Security*. World Scientific Publishing Co. Pte. Ltd., pp. 111–136. doi:10.1142/9781786340757_0005.

Borkar, V., Patel, M. and Moparekar, P. (2019) *Social Media Analysis using Optimized Clustering Algorithm*. Available at: http://ijics.com.

Brandtzæg, P.B. (2010) "Towards a unified Media-User Typology (MUT): A meta-analysis and review of the research literature on media-user typologies," *Computers in Human Behavior*, 26(5), pp. 940–956. doi:10.1016/j.chb.2010.02.008.

Charles Rosemburg (2004) *Cluster Analysis for Researchers*.

García, S. *et al.* (2016) "Big data preprocessing: methods and prospects," *Big Data Analytics*, 1(1). doi:10.1186/s41044-016-0014-0.

Javapoint (no date) *Data Preprocessing in Machine learning, Javapoint*. Available at: https://www.javatpoint.com/data-preprocessing-machine-learning#:~:text=%20Data%20Preprocessing%20in%20Machine%20learning%20%201,datasets%20which%20we%20have%20collected%20for...%20More%20 (Accessed: January 1, 2022).

Maniriho, P. and Effendi, A. (2018) *Examining the Performance of K-Means Clustering Algorithm, International Journal of Research in Engineering*. Available at: www.ijresm.com.

Muller, M. *et al.* (2009) *We are all Lurkers: Toward a Lurker Research Agenda*.

Prakash, S. *et al.* (2020) *Characteristic of enterprise collaboration system and its implementation issues in business management, Int. J. Business Intelligence and Data Mining*.

RUI XU, DONALD C. and WUNSCH (no date) *Clustering*. WILEY.

Schubert, P. and Glitsch, J.H. (2016) "Use cases and collaboration scenarios: How employees use socially-enabled Enterprise Collaboration Systems (ECS)," *International Journal of Information Systems and Project Management*, 4(2), pp. 41–62. doi:10.12821/ijispm040203.

Schwade, F. and Schubert, P. (2017) *Social Collaboration Analytics for Enterprise Collaboration Systems: Providing Business Intelligence on Collaboration Activities*. Available at: http://hdl.handle.net/10125/41197.

Schwade, F. and Schubert, P. (2018) *Social Collaboration Analytics for Enterprise Social Software: A Literature Review*.

Schwade, F. and Schubert, P. (2019) *Developing a User Typology for the Analysis of Participation in Enterprise Collaboration Systems*. Available at: https://hdl.handle.net/10125/59486.

Stieglitz, S. *et al.* (2014) "Socialmedia analytics," *Business and Information Systems Engineering*, 6(2), pp. 89–96. doi:10.1007/s12599-014-0315-7.

Stieglitz, S. *et al.* (2018) "Social media analytics – Challenges in topic discovery, data collection, and data preparation," *International Journal of Information Management*, 39, pp. 156–168. doi:10.1016/j.ijinfomgt.2017.12.002.

Wikipedia (no date a) *Cluster analysis*, *Wikipedia*. Available at: https://en.wikipedia.org/wiki/Cluster_analysis (Accessed: January 1, 2022).

Wikipedia (no date b) *Stack Overflow*. Available at: https://en.wikipedia.org/wiki/Stack_Overflow#:~:text=Stack%20Overflow%20is%20a%20question%20and%20answer%20website,a%20wide%20range%20of%20topics%20in%20computer%20programming. (Accessed: January 1, 2022).

Xreator (2019) *Stack Overflow as Social Network*. Available at: https://stack-overflow-as-a-social-network.github.io/ (Accessed: January 1, 2022).

Zhu, H. (2006) "Role mechanisms in collaborative systems," *International Journal of Production Research*, 44(1), pp. 181–193. doi:10.1080/00207540500247495.

Github Link for Python Code: https://github.com/Rahul-Bhanushali/User_Roles_Analysis.git