Received: May 19, 2023. Revised: June 10, 2023.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

QST-ClustNet: Quantum Inspired Sooty Tern Optimization for Clustering user Profiles in Social Network

Rashmi C[1]* **Mallikarjun M Kodabagi[1]**

[1]*Faculty of Computing and Information Technology, REVA University, 560064 Bengaluru, India*
* Corresponding author's Email: rashmi.c.reva@gmail.com

Abstract: The exponential growth of social media platforms has led to a demand for applications related to social network analysis. One major challenge in social network analysis is identifying communities within the network. In this work, a novel QST-ClustNet is proposed in which the sooty tern optimization algorithm is integrated with the quantum behavior to identify the communities within the social network by clustering the user profiles. Initially, the LinkedIn dataset is pre-processed using natural language processing (NLP) techniques such as data extraction, removal of stop words handling the missing data or values, and data stemming for removing irrelevant data. After preprocessing the feature are extracted using bag of words for the creation on user profiles. These created features are given as an input to frequency probability-based similarity measures to find similarities between users. Finally, quantum-inspired Sooty tern optimization is utilized for clustering the user profile in social networks. The effectiveness of the proposed strategy is examined using several parameters, such as Davies Bouldin score, Calinski Harabasz score, and Silhouette score. The proposed QST-ClustNet improves the silhouette score 22.48, 16.25, and 11.081 better than K mean clustering, C mean clustering, and fuzzy C mean clustering respectively. This approach helps examine and comprehend the structure of professional networks of online social networks, which can be applied for multiple applications such as job recommendation, talent acquisition, and many more.

Keywords: Bag-of-words, User profile, Clustering, Modularity, Community detection, Social network analysis, Clustering coefficient.

1. Introduction

An essential technique for social network analysis is community discovery, which will be created to find groups of nodes that share characteristics. In particular, community detection is a crucial component of network analysis that has been used in numerous practical applications, such as fraud detection, community detection, recommendation systems, influential node detection, and link prediction, among others [1-4]. Common community detection methods look for strong intra-community and low inter-community similarity, which indicates that nodes inside a community have high pairwise similarity scores while nodes in nearby communities have low pairwise similarity scores, to locate node clusters [5].

Different community-detecting methods are used for various applications. The traditional method concentrated only on the social network's network structure, or the relationships between nodes. Such methods base their analysis of social networks and community detection on node structural similarity. On the other hand, clustering is a technique where a set of instances are divided into smaller groups known as clusters, with each cluster's members being very similar to one another, and quite dissimilar from those of the other clusters based on the node features or attributes [6-8]. A user's profile information frequently includes some node traits, such as age, gender, and interests on several social networks like Facebook, and Twitter. The network is known as an attributed social network when this information is available, where each node is given a set of attributes. Even when fundamental information is provided, the problem with this data can result in incomplete data.

Few social network platforms block unauthorized users' access to their services for

security concerns such as LinkedIn [9, 10]. It is crucial to take into account the important characteristics of each user profile to connect the user profile with similar features by clustering, which can be used in a variety of applications. Hence, this paper introduces a technique for grouping user profiles of LinkedIn users with a comparable set of abilities is proposed. The method uses the bag of words feature to calculate the similarity between users and connects all the nodes that are a part of the same community.

The main contributions of this work are as follows:

- The key contribution of this work is to present a novel QST-ClustNet to identify the communities within the social network by clustering the user profiles.
- Initially, the dataset is pre-processed using NLP languages such as Data extraction, Removal of stop words, Handling the missing data or values, and Data stemming for removing irrelevant data.
- The Features are extracted using Bag of Words for the creation on user profiles. These created features are given as an input to Frequency Probability-based similarity measures to find similarities between users.
- The quantum-inspired Sooty tern optimization is utilized for clustering the user profile in social networks.
- Several factors were assessed to evaluate the proposed method based on the Shilhouette index, Calinski harabasz index, and Davies Bouldin index.

The remainder of the research paper was structured as follows: the Literature survey is summarized in section 2. Section 3 discusses the proposed QST-ClustNet methodology. Section 4 includes the result and discussion of the proposed model. Section 5 holds the conclusion and future work.

2. Literature survey

Some numerous applications and domains can be networked and linked to all social user profiles. Some of them are described below,

A graph compression-based community detection was developed for analyzing large-scale social networks [11]. The Compressed graphs are constructed by first merging a vertex of degree 1 or 2 with its higher-degree neighbors. Quality of vertices and density are used to assess the probability of vertices becoming the seeds of communities. The suggested algorithms are evaluated on 14 real social networks and provide a better performance, yet this model can extend to multilayer networks.

The genetic algorithm has been developed based on a similarity matrix for community detection on large-scale social networks [12]. In this algorithm, the similarity score between each pair of nodes is calculated on multiple computing nodes simultaneously. The suggested model doesn't suffer from resolution limit problems. However, this method does not perform in overlapping communities.

A density-based clustering technique was presented that made use of spatial-textual data from Twitter and other social media [13]. Initially, the technique discriminated between tweets with geological tags that were relevant to points of interest and those that weren't as texts that, respectively, included or did not contain a point of interest name or its semantically consistent modifications. In terms of the F1 score and its variants, the suggested approach performs better than the DBSCAN scenario. Yet this model requires more information to jointly utilize spatial and textual information.

Community detection on social networks was developed based on the divide and agglomerate (DA) algorithm [14]. Networks are divided into small groups according to the similarities of their node pairs and then merged with the group with the strongest attraction until the community criteria are satisfied. AA indexes with similarity constraints capture both local and global information for optimal community detection. The DA algorithm is useful in identifying the structure of data, yet overlapping can also occur.

The meta-clustering ensemble approach is developed to tackle the associated difficulty with model selection [15]. This scheme uses the bi-weighting policy and the similarity between the instances is estimated to create primary clusters. Then the Model Selection approach is applied which entails the re-clustering of main clusters to produce meta-clusters. By combining similar clusters and accounting for a threshold after clustering, the number of ideal clusters is determined. This model works well in various descriptive methods. However, reliability in model selection for configuration is not performed.

For incomplete data, a novel Bayesian theory-based density-based clustering method is presented in [16] that utilizes intermediate clustering results while performing imputation and clustering

simultaneously. The local imputation clustering method shown here enhances the clustering outcome while reducing the influence of low-density areas within non-convex clusters. The suggested is performing effectively and has its characteristics, yet it cannot cluster data with higher differences in density.

Quad tournament optimizer based on four searches was presented in [17]. The four searches are neighborhood searches around the corresponding solution and the global best solution, searches relative to a randomly chosen solution, searches toward the center between the randomly selected solution and the global best solution, and searches toward the global best solution. This algorithm effectively finds the optimal solution, yet can add more methods in a tournament rather than four.

The multiple interaction optimizer was developed for solving order allocation problem [18]. Initially, agents interact with a few randomly chosen agents from the population. Every contact includes a guided search. Each agent does a local search in the second phase, which linearly shrinks the search space for the iteration. The suggested algorithm is implemented with low cost and latent, however, it cannot solve multiple objective order allocation problems.

A guided pelican algorithm was developed for the enhancement of the pelican optimization algorithm (POA) that replicates the hunting behavior of pelican birds [19]. The global best solution is used by GPA as a deterministic target in the beginning, replacing the randomized target. In addition, when calculating local search space size, GPA swaps out the pelican's present location for the size of the search space. Thirdly, GPA uses numerous candidates in both phases, as it did in the original POA. This algorithm efficiently solves the portfolio optimization problem. Yet this implementing algorithm in real-time is challenging.

A stochastic komodo algorithm was developed which is derived from the behavior of Komodo during foraging and mating calls [20]. Three different komodo species make up this algorithm: big male, female, and little male. While female komodos carry out diversification based on the search space radius, males concentrate on intensification. At the start of the iteration, the sorting mechanism is removed and a random distribution of the komodo is undertaken. This algorithm is very competitive in both unimodal and multimodal functions, yet it can add a more comprehensive evaluation.

A literature survey of clustering user profiles in social networks for community detection reveals several promising future directions for research. One area of future research is to improve the accuracy of clustering algorithms by incorporating more sophisticated user profiling techniques using NLP and sentiment analysis. Future research should also examine how different network topologies affect the precision of community discovery since networks can either be sparse or dense. Additionally, there is a need for more research on the practical applications of community detection in social networks. For example, clustering user profiles can be useful for targeted advertising, recommendation systems, and identifying potential influencers. Overall, the literature survey highlights several promising avenues for future research on clustering user profiles in social networks for community detection.

3. Proposed QST-ClustNet

In this work, a novel quantum inspired Sooty tern optimization for clustering is proposed based on user profiles in social networks. Initially, the LinkedIn dataset is pre-processed using NLP techniques such as data extraction, removal of stop words handling the missing data or values, and data stemming for removing irrelevant data. After preprocessing the feature are extracted using bag of words for creation of user profiles. These created features are given as an input to frequency probability-based similarity measures to find similarities between users. Finally, quantum-inspired Sooty tern optimization is utilized for clustering the user profile in social networks. The overall workflow proposed QST-ClustNet for user profile clustering is shown in Fig. 1.

3.1 Dataset description

In this paper, experiments are performed on a familiar social network dataset named the LinkedIn user profile dataset. However, the proposed method has considered only the skills attribute, and its features, as the method aims to cluster the user's profile using their features. The user profile skills can vary from domain to domain, and in each domain into multiple categories. The details of the datasets are described below:

LinkedIn user profile dataset: LinkedIn is a professional networking platform with more than 500+ million users. Exploring LinkedIn profiles gives enormous information about each user profile which can be extracted to obtain meaningful information. To authenticate users and authorize members, the LinkedIn API employs OAuth 2.0. The dataset contains more than 1000 user profiles with multiple attributes such as name, positions,

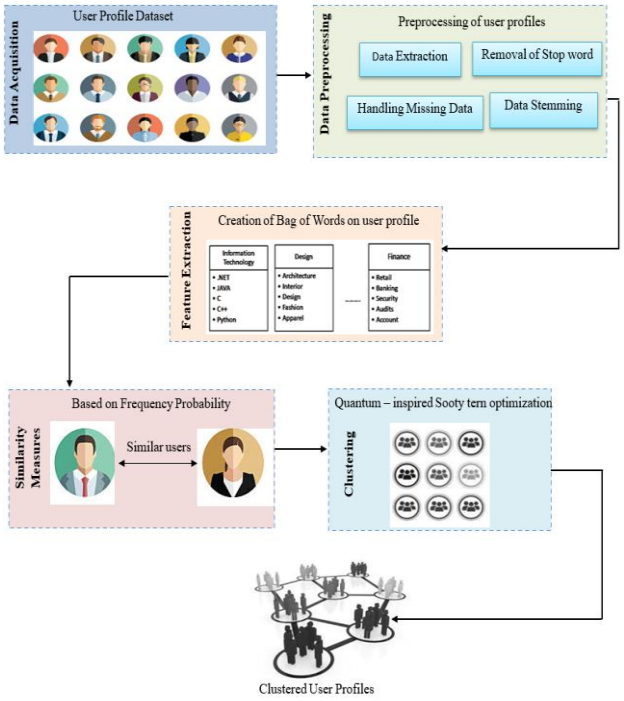

Figure. 1 The overall workflow of the proposed QST-ClustNet for user profile clustering

summary, skills, location, education, and many more.

3.2 Data pre-processing

Data pre-processing is required to modify the format of the data into the required format. Pre-processing is required because the user profile data that was retrieved from LinkedIn is in unstructured form. Pre-processing is carried out in multiple stages such as data extraction, removal of stop words, handling the missing data or values, and data stemming.

3.2.1. Data extraction

In data extraction, the user profile data is extracted from the LinkedIn website. There are more than 850 million user profiles on LinkedIn, which is a highly well-known and widely utilized professional network. To extract the user profile data set, LinkedIn has REST application programming interface (API) through which data can be extracted. To connect REST API, it uses OAuth 2.0 to access LinkedIn API for user permission and API authentication. Before they may access members, data, or retrieve data from LinkedIn, applications need to be allowed and authenticated. After authentication, all user profile data can be extracted.

3.2.2. Removal of stop words

A huge amount of unnecessary information can

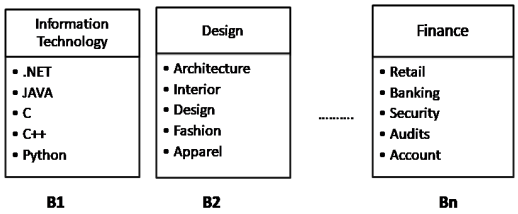

Figure. 2 Creation of a bag of words of skill-based user profiles of social network

be described as stop words. ‘‘The’’, ‘‘and’’, ‘‘a’’, ‘‘an’’, ‘‘as’’, ‘‘about’’, ‘‘at’’, etc. are few examples of stop words. Due to the enormous amount of superfluous information, they return, these terms are filtered out.

3.2.3. Handling the missing data or values

The user profile data set extracted from LinkedIn may contain missing values, so those profiles are eliminated from the dataset by removing the corresponding rows containing NULL values thus extracting valuable information for further processing.

3.2.4. Data stemming

Inflected words are shortened through stemming. The written form of inflected words is typical. The morphological root of the term need not correspond to the stem. The term "consult" serves as the root word that a stemming algorithm uses to condense the words "consultant," "consults," and "consulting".

3.3 Creation of bag of words

The bag of words is a statistical language model that analyzes text and documents according to their word count. In this model, categorizing user profiles based on their skill set, a different bag of words is created as shown in Fig. 2.

Each bag of words contains a set of common words which is calculated by the total appearance of the word count and categorized into the corresponding feature of a bag of words. Sometimes the user profile of some can be matched to more than one bag of word sets. Each user profile that is categorized is identified by the user index number and the skill set it contains. The creation of a Bag of words of skill-based user profiles is shown in Algorithm 1.

3.4 The similarity between social network user's profiles

Similarity measures can be used to identify clusters in social networks. The interests of users in

Algorithm 1: Creation of bag of words

Input: Skill-set feature of Social Network User's Profile data set

Output: Creation of different Bag of Words

1: Consider Feature Sets $S_i \longleftarrow \{S_1, S_2, S_3 \ldots . S_n\}$ where $i \in (1, n)$
3: $D \longleftarrow$ Skill-set feature of a user's profile
4: for x in D do
5: if $x \in S_i$ then
7: $B_i \longleftarrow x$
8: end if
9: end for
10: Creation of a Bag of words, $B_i \longleftarrow \{B_1, B_2, B_3 \ldots . B_n\}$ where $i \in (1, n)$

the same cluster are similar in social networks. Based on several characteristics, it is possible to compute the similarity between any two users. These variables might be influenced by topological data or shared activities. The primary goal of the proposed technique is to leverage frequency probability to identify shared interests among users. The Hamming distance gives differing category values a distance of one and identical values a distance of zero.

The distance between any two categorical users is by using Eq. (1)

$$Dist(u_i, u_j) = \sum_{s=1}^{d} [\delta(u_{is}, u_{js}) L(u_{is} = u_{is})] \quad (1)$$

Where $\delta u_{is,} u_{js}$ denotes the hamming distance between two users and it is calculated as follows:

$$L(u_{ir} = u_{jr}) = L(X_s = u_{is}|U).L'(X_S = u_{is}|U) + L(X_S = u_{ja}|U).L'(X_S = u_{js}|U) \quad (2)$$

where $P(X_s = u_{is}|U)$ and $P'(X_s = u_{is}|U)$ are computed using Equ. 3 and 4 respectively.

$$P(X_s = u_{is}|U) = \frac{\sigma X_s = u_{ir}(U)}{\sigma X_s \neq NULL(U)} \quad (3)$$

$$P'(X_s = u_{is}|U) = \frac{\sigma X_s = u_{ir}(U) - 1}{\sigma X_s \neq NULL(U) - 1} \quad (4)$$

Where $\sigma X_r = u_{ir}(U)$ indicate the number of users. whose $X_s the$ index comprises the value u_{ir} While the NULL symbol denotes the lack of data.

The greatest distance between any two occurrences must be determined to normalize the distance between them. Let's look at the next two scenarios to find the maximum distance between two users.

Scenario 1: All users except $u_{js,}$, have similar values as $u_{is,}$

Let's consider the scenario, the dataset of m users, let $m-1$ user have a similar value that is it can be either 0 or 1 one case has a value $u_{js,}$

$$P(u_{is} = u_{js}) = \frac{m-2}{m} \quad (5)$$

Assuming that the same situation holds for each attribute in our dataset (d number attribute), i.e., each attribute contains a numerical value of u_{is} values and one number of u_{js}, then the value of the distance between the two cases$(u_{is} = u_{js}) = \frac{m-2}{m}$

Scenario 2: About Half of the cases have $u_{is,}$ values and about half of the cases have $u_{js,}$ values.

Let's Consider the scenario, the dataset of m users, let $\frac{m}{2}$ user has the same value, which is $u_{is,}$ (it can be0 or 1) and $\frac{m}{2}$user has value $u_{js,}$ (its value must not be the same as $u_{is,}$) on attribute X_s.

$$P(u_{is} = u_{js}) = \frac{m-2}{2(m-1)} \quad (6)$$

Assuming that the same situation holds for each attribute in our dataset (d number attribute), i.e., each attribute contains $m-1$number of u_{is} values and one number of u_{js}, then the distance value between two instances $(u_{is} = u_{js}) = \frac{d(m-2)}{2(m-1)}$

To calculate the normalized distance between two users, divide the computed distance by the maximum distance between the two users. Mathematically it can be shown as follows:

$$D_m(u_i, u_j) = \frac{Dist(u_i, u_j)}{d(\frac{m-2}{m})} \quad (7)$$

The relationship between distance and similarity measure must be followed to determine how similar two instances are, and it is represented as follows.

$$fpsm(u_i, u_j) = 1 - D_m(u_i, u_j) \quad (8)$$

3.5 Clustering of social network users profiles

In the proposed QST-ClustNet, the clustering is done by quantum-inspired sooty tern optimization to find out the clusters of users based on their similarity measures such as word frequency, the strength of words between user profiles, word count between user profiles, and total word count. The QSTO algorithm is motivated by the natural

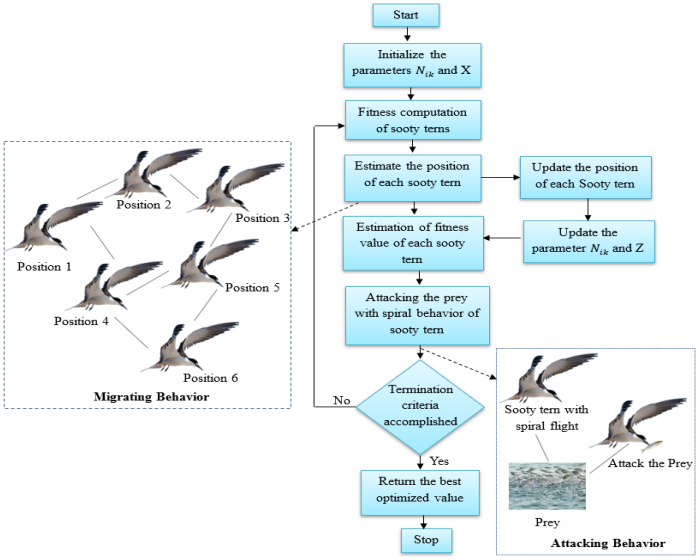

Figure. 3 Flowchart of QSTO algorithm

behavior of sea birds and their lifestyle which is used as a clustering algorithm to find out the clusters of users. Based on similarity measures, the STO algorithm selects the relevant communities. In this QSTO algorithm, the number of sooty terns is represented as the similarity from the similarity measure phase, the similarity measures are defined with the best fitness levels and the irrelevant features indicate the worst fitness value of the sooty tern. The two major steps in the STO algorithm are migrating behavior and attacking behavior. Sooty terns travel in groups during their journey. A variety of starting positions was used by sooty terns to prevent collisions or conflicts. Even with lower fitness levels, sooty terns in a group can still cover the same distance as the fittest individuals. A sooty tern with low fitness (irrelevant features) can update its initial position based on a sooty tern with high fitness (relevant features). The flowchart of the QSTO algorithm is shown in Fig. 3.

3.5.1. Migrating behavior

A sooty tern has to satisfy the following three criteria to successfully migrate: *Collision avoidance:* A new search agent (SA) position is computed to prevent conflict between its nearby SA (e.g., sooty terns).

$$a_{ij} = N_{ik} \times m_{ij}(p) \tag{9}$$

where a_{ij} is the position of SA that avoids colliding with other SA, m_{ij} indicates the current position of the search agent, p defines the actual iteration, and N_{ik} signifies the moment of SA in a given search space.

Converge in the direction of best neighbor: To avoid collisions, the search agents move in the direction of their best neighbor.

$$b_{ij} = X * (Y_{best}(P) - m_{ij}(p)) \tag{10}$$

Where b_{ij} defines the various positions of search agents, m_{ij} towards the best fittest search agent (relevant features), X is a random variable that represents the better exploration. X is derived as follows:

$$X = 0.5 \times R_{and} \tag{11}$$

Where R_{and} is a random number that lies between the ranges from [0, 1].

Update corresponding to best search agent: Finally, the search agent can become the best search agent by updating its position.

$$C_{ij} = S_{ij} + T_{ij} \tag{12}$$

where C_{ij} defines the gap between the search agent and the best fittest search agent with a high fitness value.

3.5.2. Attacking behavior

The sooty tern can change its speed as well as angle of attack during migration. The wings of these birds help them reach higher altitudes. In the air, sooty terns exhibit spherical behavior when attacking prey.

This behavior is mathematically derived as,

$$l = S_{radius} \times sin(i) \quad (13)$$

$$m = S_{radius} \times cos(i) \quad (14)$$

$$n = S_{radius} \times j \quad (15)$$

$$D = a \times e^{kb} \quad (16)$$

where S_{radius} stands for the radius of each spiral round, i indicates that the variable within an interval of $\{0 \leq k \leq 2\pi\}$. a and b are constant variables to denote the spiral form, and e is the base of the natural logarithm. Therefore, the updated position of SA when the constant values of a and b as 1 and is derived as,

$$V_{ij}(p) = (C_{ij} \times (l + m + n)) \times Y_{best}(P) \quad (17)$$

where $V_{ij}(p)$updates the positions of other SA and returns the best solution with clusters. Afterward, the inputs are multiplied by the feature vectors; the randomly selected clusters are summed. Finally, the STO algorithm is evolved based on the clusters gathered from migrating and attacking operations, and it can be fed into a fully connected (FC) layer for further clustering. As the result of the previous layer, the FC layer uses selected clusters of the STO algorithm to group the relevant user.

4. Results and discussions

The results of the proposed QST-ClustNet are discussed in this section. Experiments are conducted on NetworkX tool for graph-based mining that makes it simple to analyze the structure of social networks. To evaluate communities of social network structure the most popular metric used is modularity and average clustering coefficient (ACC).

The clustering of the user profiles using skills is done based on the maximum similarity count which in turn takes into account the similarity of user profile skills by considering word frequency, the strength of words between the user profile skills, word count between user profile skills, the total word count of the user profile skills, and with the empirically chosen threshold value which eventually leads into multiple communities. The clustering of user profiles is plotted in the form of a graph using the NetworkX tool as shown in Fig. 4.

4.1 Performance metric:

The effectiveness of the proposed method was carried out using Shilhouette index, Calinski Harabasz index, and Davies-Bouldin index. The parameters are evaluated statistically below.

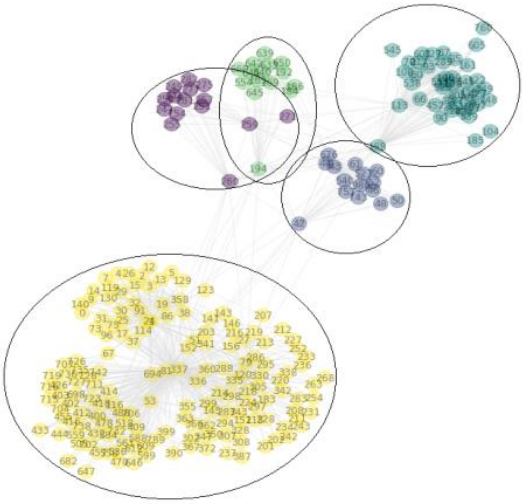

Figure. 4 Clustering of social network user profiles

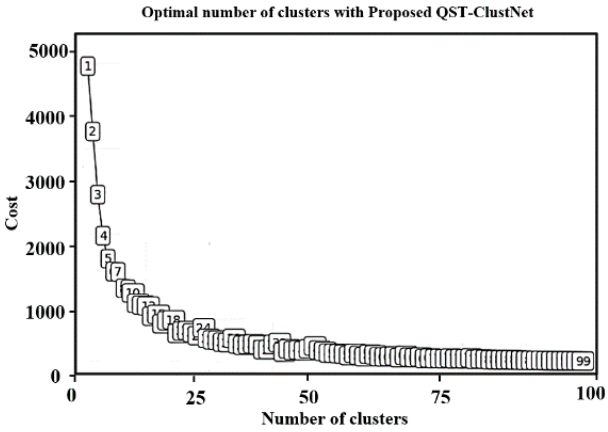

Figure. 5 Cost of x on FPSM attribute of LinkedIn

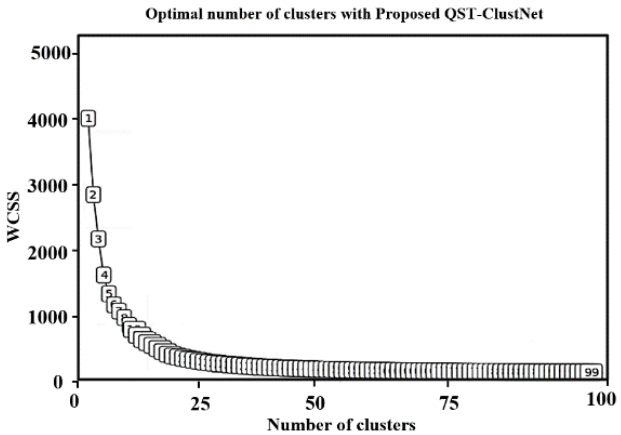

Figure. 6 Optimal values of WCSS on FPSM for LinkedIn dataset

Shilhouette index: The silhouette score measures a user's similarity to their cluster to other clusters. The Silhouette index's ranges range from -1 to 1. The Silhouette index is described mathematically in Eq. (18).

Table. 1 Comparative analysis of LinkedIn user profile dataset

Algorithm	Number of clusters generated	Silhouette score	Davies Bouldin score	Calinski Harabasz score
K mean clustering	13	0.7345	0.66571	24153.07584
C mean clustering	17	0.7935	0.54785	20145.41477
Fuzzy C mean clustering	21	0.8425	0.44257	163485.18975
Proposed	27	0.9475	0.31724	114586.75824

$$SI_{(g_i)} = \frac{(q(g_i)-p(g_i))}{\max\{p(g_i),q(g_i)\}} \quad (18)$$

where $p(g_i)$ denotes the average distance between the user g_i and each additional user in the cluster, and $q(g_i)$ indicates the average distance between the user g_i and a different user from the nearby cluster.

Davies-Bouldin index: The average similarity between each cluster and its closest neighbor is calculated for this index. If the DBI score is low, indicating that the clusters the clustering algorithm produces are well separated, the clustering algorithm is deemed to be doing well.

The Davies-Bouldin Index is represented mathematically as follows:

$$D_I = \frac{1}{x}\sum_{m=1}^{x} max_{m\neq n}\left(\frac{Q_m+Q_n}{\mathrm{f}(ce_m,ce_n)}\right) \quad (19)$$

Where every user of the cluster m has an average distance of Q_m from the centroid of the same cluster, and the distance between the centroids of clusters m and n is denoted as $\mathrm{f}(ce_m, ce_n)$. x value represents the total number of clusters generated by the clustering algorithm. Cost of x on FPSM attribute of LinkedIn as shown in Fig. 5. Optimal values of WCSS on FPSM for the LinkedIn dataset as shown in Fig. 6.

4.2 Comparative analysis

The performance of each clustering algorithm was assessed to validate the results of the proposed QST-ClustNet. The competence evaluation was carried out between the proposed QST-ClustNet with four clustering techniques such as K mean, C mean, and fuzzy C mean. The performance estimation was made with different metrics such as Calinski Harabasz score Silhouette score, and Davies Bouldin score of each algorithm. From the experimental result, the proposed method performs better than the existing methods. Table. 1 Comparative analysis for LinkedIn user profile dataset as shown in Table 1.

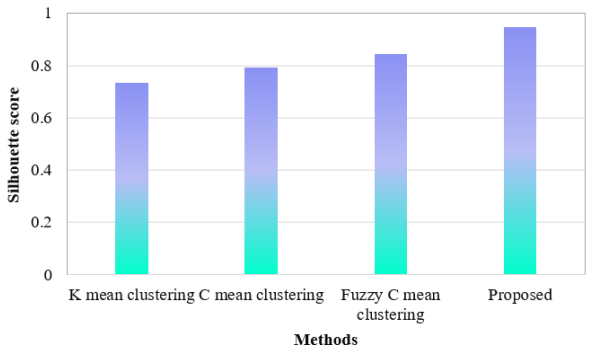

Figure. 7 Silhouette score for the proposed and the existing method

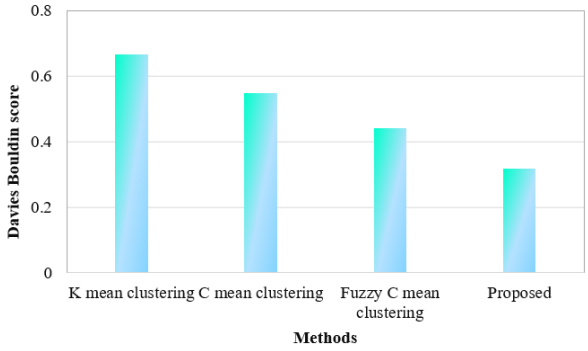

Figure. 8 Davies Bouldin score for the proposed and existing method

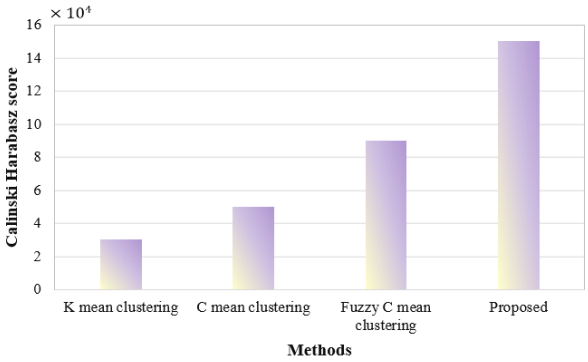

Figure. 9 Calinski Harabasz score for the proposed and the existing method

In Figs. 7, 8, and 9, three performance measures are compared for finding values of x (calculated using the proposed method). Due to the categorical nature of our datasets, the K-Mean approach does not work well in these datasets. It only works with numerical data. But our approach fixes the local optimum issue. The study of these graphs demonstrates that our proposed method outperforms the other existing techniques.

Table 2. Comparing modularity and clustering coefficient obtained by different community detection algorithms on user profiles of social network

Community detection Algorithms	Modularity value	Average Clustering Coefficient
Girvan-Newman (GN)	0.451	0.467
Spectral Clustering (SC)	0.532	0.588
Louvain Method (LM)	0.601	0.694
Hierarchical Clustering (HC)	0.551	0.611
Proposed model	0.791	0.851

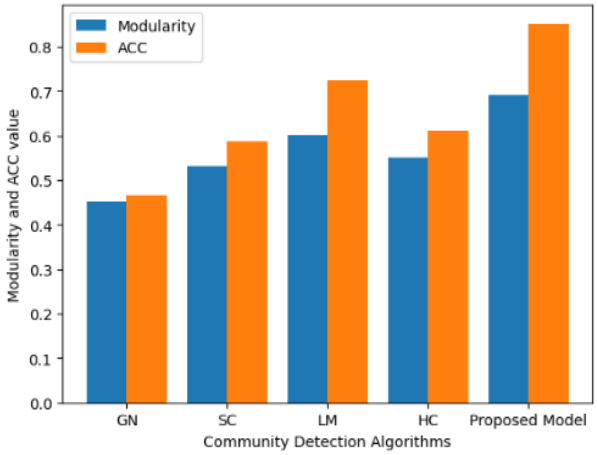

Figure. 10 Comparison of modularity and average clustering coefficient values for different community detection algorithms

Table 3. Comparison between the existing and proposed technique

References Number	Methods	Modularity value
[11]	Graph compression	0.508
[12]	Genetic Algorithm	0.720
[13]	DBSTexC	0.701
[14]	DA Algorithm	0.49
Proposed QST ClustNet		0.791

A network's degree of modularity, a property of networks or graphs, is used to determine how well it is segmented into modules. Networks with high levels of modularity feature large connections between nodes inside modules as opposed to sparse connections between nodes in various modules.

It accepts values between -1 and 1, where 1 indicates extremely good clusters, 0 indicates clustering is not superior to randomness, and a large negative number would indicate anti-clusters [21-23]. Good clusters were discovered when the modularity was between 0.3 and 0.7. Our proposed model had a modularity value of 0.691. The other metric used to gauge the network structure communities is the clustering coefficient.

An indicator of how closely nodes in a graph tend to cluster together is called a clustering coefficient. In other words, the clustering coefficient tells us how closely related a vertex's neighbors are to one another. To compute it more precisely, divide the total number of edges that could join neighbors by the total number of edges that could connect neighbors at each vertex. Because it represents the percentage of feasible edges that are realized, the number will always fall between 0 and 1 [24, 25]. When applied to the entire social network it is termed the average clustering coefficient. The modularity and average clustering coefficient were computed for other existing community detection algorithms as shown in Table 2. Comparison of modularity and average clustering coefficient values for different community detection algorithms as shown in Fig. 10.

Table 3 compares the proposed model with other existing methods. The comparison of existing models (i.e., Graph compression, Genetic Algorithm, DBSTexC) DA Algorithm) with the proposed QST-ClustNet. The overall modularity value of the proposed method is 35.77, 8.975, 11.37, 38.05, better than existing techniques. The proposed QST-ClustNet achieves 0.791 which is better than the existing techniques.

5. Conclusion

In this paper, a novel quantum inspired sooty tern optimization for clustering is proposed based on user profiles in social networks. Initially, the LinkedIn dataset is preprocessed using NLP techniques such as data extraction, removal of stop words handling the missing data or values, and data stemming for removing irrelevant data. After preprocessing the feature are extracted using bag of words for creation on user profiles. These created features are given as an input to frequency probability-based similarity measures to find similarities between users. Finally, the quantum-inspired sooty tern optimization is utilized for clustering the user profile in social networks. The performance of the proposedQST-ClustNet is examined using several parameters, such as Silhouette score, Davies Bouldin score, and Calinski Harabasz score, that attains 0.9475, 0.31724, and 114586.75824 respectively. The technique has been used on user profiles on LinkedIn and has produced encouraging outcomes. Our proposed methodology for identifying communities based on users' profile

skills outperforms other popular community detection algorithms in terms of modularity and average clustering coefficient. In the future the proposed QST-ClustNet can be applied in emerging deep learning technique and also reduce time consumption.

Conflicts of interest

The authors declare that they have no conflict of interest.

Author contributions

The authors confirm contribution to the paper as follows: Study conception and design: Rashmi C and Mallikarjun M Kodabagi; Data collection: Rashmi C; Analysis and interpretation of results: Mallikarjun M Kodabagi; Draft manuscript preparation: Rashmi C and Mallikarjun M Kodabagi. All authors reviewed the results and approved the final version of the manuscript.

Acknowledgments

The authors would like to thank the reviewers for all of their careful, constructive and insightful comments in relation to this work.

References

[1] Y. Yang, Y. Xu, Y. Sun, Y. Dong, F. Wu, and Y. Zhuang, “Mining Fraudsters and Fraudulent Strategies in Large-Scale Mobile Social Networks”, *IEEE Trans. Knowl. Data Eng.*, Vol. 33, No. 1, pp. 169–179, Jan. 2021, doi: 10.1109/TKDE.2019.2924431.

[2] W. Luo, D. Zhang, L. Ni, and N. Lu, “Multiscale Local Community Detection in Social Networks”, *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2019, doi: 10.1109/TKDE.2019.2938173.

[3] S. Kumar, A. Kumar, and B. S. Panda, “Identifying Influential Nodes for Smart Enterprises Using Community Structure With Integrated Feature Ranking”, *IEEE Trans. Ind. Informatics*, Vol. 19, No. 1, pp. 703–711, Jan. 2023, doi: 10.1109/TII.2022.3203059.

[4] R. Aliakbarisani, A. Ghasemi, and M. A. Serrano, “Perturbation of the Normalized Laplacian Matrix for the Prediction of Missing Links in Real Networks”, *IEEE Trans. Netw. Sci. Eng.*, Vol. 9, No. 2, pp. 863–874, Mar. 2022, doi: 10.1109/TNSE.2021.3137862.

[5] M. Rostami and M. Oussalah, “A novel attributed community detection by integration of feature weighting and node centrality”, *Online Soc. Networks Media*, Vol. 30, p. 100219, Jul. 2022, doi: 10.1016/j.osnem.2022.100219.

[6] P. Chunaev, “Community detection in node-attributed social networks: A survey”, *Comput. Sci. Rev.*, Vol. 37, p. 100286, 2020, doi: 10.1016/j.cosrev.2020.100286.

[7] R. C, and M. M. Kodabagi, “CONNECTING USER PROFILES OF SOCIAL NETWORKS USING PROXIMITY-BASED CLUSTERING”, *Malaysian J. Comput. Sci.*, pp. 1–15, 2022, doi: 10.22452/mjcs.sp2022no2.1.

[8] J. Zhu, C. Wang, C. Gao, F. Zhang, Z. Wang, and X. Li, “Community Detection in Graph: An Embedding Method”, *IEEE Trans. Netw. Sci. Eng.*, Vol. 9, No. 2, pp. 689–702, 2022, doi: 10.1109/TNSE.2021.3130321.

[9] S. Li *et al.*, “Deep Job Understanding at LinkedIn”, In: *Proc of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2145–2148, 2020, doi: 10.1145/3397271.3401403.

[10] M. A. Johnson and C. Leo, “The inefficacy of LinkedIn? A latent change model and experimental test of using LinkedIn for job search”, *J. Appl. Psychol.*, Vol. 105, No. 11, pp. 1262–1280, 2020, doi: 10.1037/apl0000491.

[11] X. Zhao, J. Liang, and J. Wang, "A community detection algorithm based on graph compression for large-scale social networks", *Information Sciences*, Vol. 551, pp. 358-372, 2021.

[12] R. K. Behera, D. Naik, S. K. Rath, and R. Dharavath, "Genetic algorithm-based community detection in large-scale social networks", *Neural Computing and Applications*, Vol. 32, pp. 9649-9665, 2020.

[13] M. D. Nguyen and W. Y. Shin, “An Improved Density-Based Approach to Spatio-Textual Clustering on Social Media”, *IEEE Access*, Vol. 7, pp. 27217–27230, 2019, doi: 10.1109/ACCESS.2019.2896934.

[14] Z. Liu and Y. Ma, "A divide and agglomerate algorithm for community detection in social networks", *Information Sciences*, Vol. 482, pp. 321-333, 2019.

[15] T. Li, A. Rezaeipanah, and E. M. T. E. Din, “An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement”, *J. King Saud Univ. - Comput. Inf. Sci.*, Vol. 34, No. 6, pp. 3828–3842, Jun. 2022, doi: 10.1016/j.jksuci.2022.04.010

[16] Z. Xue and H. Wang, “Effective density-based

clustering algorithms for incomplete data", *Big Data Min. Anal.*, Vol. 4, No. 3, pp. 183–194, 2021, doi: 10.26599/BDMA.2021.9020001.

[17] P. D. Kusuma and M. Kallista, "Quad Tournament Optimizer: A Novel Metaheuristic Based on Tournament Among Four Strategies", *International Journal of Intelligent Engineering & Systems*, Vol. 16, No. 2, 2023, doi: 10.22266/ijies2023.0430.22.

[18] P. D. Kusuma and A. Novianty, "Multiple Interaction Optimizer: A Novel Metaheuristic and Its Application to Solve Order Allocation Problem", *International Journal of Intelligent Engineering & Systems*, Vol. 16, No. 2. 2023, doi: 10.22266/ijies2023.0430.35.

[19] P. D. Kusuma and A. L. Prasasti, "Guided Pelican Algorithm", *International Journal of Intelligent Engineering and Systems*, Vol. 15, No. 6, pp. 179-190, 2022, doi: 10.22266/ijies2022.1231.18.

[20] P. D. Kusuma and M. Kallista, "Stochastic Komodo Algorithm", *International Journal of Intelligent Engineering and Systems*, Vol. 15, No. 4, pp. 156-166, 2022, doi: 10.22266/ijies2022.0831.15.

[21] J. Chellig, N. Fountoulakis, and F. Skerman, "The modularity of random graphs on the hyperbolic plane", *J. Complex Networks*, Vol. 10, No. 1, 2021, doi: 10.1093/comnet/cnab051.

[22] P. Parraguez, S. A. Piccolo, M. M. Perisic, M. Storga, and A. M. Maier, "Process Modularity Over Time: Modeling Process Execution as an Evolving Activity Network", *IEEE Trans. Eng. Manag.*, Vol. 68, No. 6, pp. 1867–1879, Dec. 2021, doi: 10.1109/TEM.2019.2935932.

[23] S. Zhu, L. Xu, and E. D. Goodman, "Hierarchical Topology-Based Cluster Representation for Scalable Evolutionary Multiobjective Clustering", *IEEE Trans. Cybern.*, Vol. 52, No. 9, pp. 9846–9860, Sep. 2022, doi: 10.1109/TCYB.2021.3081988.

[24] D. L. Hansen, B. Shneiderman, M. A. Smith, and I. Himelboim, "Calculating and visualizing network metrics", *Analyzing Social Media Networks with NodeXL*, pp. 79–94, 2020.

[25] R. E. Kooij, N. H. Sørensen, and R. Bouffanais, "Tuning the clustering coefficient of generalized circulant networks", *Phys. A Stat. Mech. its Appl.*, Vol. 578, p. 126088, 2021, doi: 10.1016/j.physa.2021.126088.

Appendix

symbols	Descriptions
$\delta u_{is,}\, u_{js}$	the hamming distance between two users
$\sigma X_r = u_{ir}(U)$	the number of users
X_s	index comprises the value u_{ir}
m	users
NULL	the lack of data
$D_m(u_i, u_j)$	normalized distance between two users
a_{ij}	the position of SA that avoids colliding with other SA
m_{ij}	the current position of the search agent
p	the actual iteration
N_{ik}	the moment of SA in a given search space.
b_{ij}	the various positions of search agents
m_{ij}	the best fittest search agent (relevant features)
X	a random variable that represents the better exploration
R_{and}	a random number that lies between the ranges from [0, 1].
C_{ij}	the gap between the search agent and the best fittest search agent with high fitness value.
S_{radius}	radius of each spiral round
i	variable within an interval of $\{0 \leq k \leq 2\pi\}$
a and b	constant variables to denote the spiral form
e	base of the natural logarithm
$V_{ij}(p)$	updates the positions of other SA and returns the best solution with clusters.
$p(g_i)$	the average distance between user g_i and every other user in the same cluster,
$q(g_i)$	the average distance between user g_i and another user from nearest cluster
f(ce_m, ce_n)	the centroids of clusters $m\ and\ n$
$Inter - Cl_{dp}$ and $Intra - Cl_{dp}$	the inter cluster dispersion and intra cluster dispersion