https://dx.doi.org/10.22161/ijaers.4.2.11 *ISSN: 2349-6495(P) | 2456-1908(O)*

Email Analysis for Community Detection - Multi-User Perspective

S. Karthika, P. Saranya

Department of Information Technology, SSN College of Engineering, Chennai, Tamil Nadu, India

Abstract— *Email classification has the capacity to group emails and user as community which is based on communication arrangement. Personalized network is used to understand the behavior of each user in an email and it analyze the various or different aspect of community structure e.g. the person who is having the same likeliness in a various social relations. The proposed system extracts single user or multi user from the email corpus using statistical analysis. This methodology uses a multi-user personalized email community detection method and tracks the email user it should be grouped. It also depends on their structural and semantic intimacy. Multi-user personalization concept used to find out the structure of community with fuzzy information i.e., an incomplete set of email details. The interactions are visualized as social graphs.*

Keywords— Community Detection, Multiuser Community, Pattern of Interest, PI-Net

I. INTRODUCTION

A social network is a social structure, which inter-relates nodes that are commonly individuals or organizations (such as friends, and co-workers) connected by interpersonal relationships. It is a platform that is used by people to build social relation with other people. Multi-user (or multi-account) represents the person who is having more than one email account. Multi-user personalization concept provides an approximation to the entire network community structure with an email corpus.

The focus of this paper is used to group more than one email account information using related communication behavior [2]. It uniquely constructs an Undirected Weighted Graph (UW-graph) using emails meta-data from more than one account. These user grouping is done based on their similar communication patterns [12]. Community detection is used to detect the behavior of the person from multi-user email account. Each user personal email is represented as undirected weighted graph for structural and semantic intimacy. It analyzes community structures using multi-user information, i.e., email from multiple accounts, and it can be used to understand the network.

People belonging to the same community are expected to have similar community behavior. The identified communities can be used to classify emails and determine prominent users [10]. It reflects similar neighborhood structure of email communication, e.g., frequent email exchanges with neighbors. The integration of semantic and structural information is necessary for community analysis [13].

The meta-data from an email's content includes subject length, text size, and attachment size, and TAG is the set of attribute and their labels. In order to extract the Communication Patterns of Interest (CPI) an email network is constructed using outgoing email i.e., email that contains sender information in metadata [9]. Using this meta-data it identifies the communication pattern behavior, for example, usually users exchanging the email in a certain time period.

The rest of this paper is structured as follows. In Section 2, the different supporting work for community detection is presented. In Section 3, the framework for multi-user personalized email community is discussed and Section 4 explains the methodology for multi-user personalization. In Section 5, experimental study is discussed and in the Section 6 the paper is concluded with the future work.

II. RELATED WORK

The Communities are a union of nodes in a dark network which are identified to have common properties like interests between each other with denser connectivity than to the other nodes of the network. Such communities are likely to form a functional unit of a network and exhibit some interactions and knowledge exchange with each other as discussed by [1]. The evolution of communities [2] and various approaches opted for dealing with overlapping communities [3] [4] are important for the analysis of communities in social network.

The existence of hierarchical structures in networks becomes a challenging issue in community detection where there is a possibility of a community being a part of another larger community. [5] Introduced a measure for evaluating the goodness of partitioning known as modularity which states that it is better to investigate community structure by provisioning nested hierarchy rather single community partitioning. The most prevailing and predominant technique which sociologists use in their

https://dx.doi.org/10.22161/ijaers.4.2.11 *ISSN: 2349-6495(P) | 2456-1908(O)*

analysis of social network and community identification is hierarchical clustering [7] [8].

Hierarchical clustering methods are implemented by discovering natural partitions in a social network which is identified by similarity metrics [9]. The concept of stochastic block modelling is also adopted for detecting communities from social network [10] [11]. The topological properties of nodes define the equivalence within a class such as structural equivalence [12] and regular equivalence [13]. Extending the approach defined by [6] that uses general stochastic block modelling approach with Gibbs sampling for inferring object positions, [3] proposed an approach that allows an object to attain multiple positions and belong to multiple category which can be modelled with individuals having multiple roles at various context within the same social network. Another approach proposed by [9] defines the generalized stochastic block model which detects groups of individuals whose activities are focused on a particular topic that are identified by monitoring the demographic properties and relationship between the individuals participating in the activity.

Community detection can be done through various other methods and techniques like maximum likelihood [11], mathematical programming [13], inference and latent space clustering [12]. To find the clusters, hubs and outliers in large networks based on structural similarity, the DBSCAN algorithm [Ester et al., 1996] is extended by [10] as SCAN (Structural Clustering Algorithm for Networks) so that it can be applied for undirected and unweighted graph structures by using the neighbourhood of vertices as clustering criteria. Likewise, [1] proposed an extension of DBSCAN algorithm to weighted interaction graph structures for online social networks by considering only the weighted interaction graph of the network.

In [3] a method which reduces the number of possible values that must be considered by obtaining only the edge weights of a Core Connected Maximal Spanning Tree (CCMST) which is formulated from the underlying network structure is proposed. A two stage parameter free extension of density based clustering is proposed by [5] where the first stage focuses on finding micro communities using the highest local structural similarity value and a constant, whereas the second stage focuses on iteratively joining the identified micro communities based on the value of gain in modularity.

III. FRAMEWORK FOR MULTI-USER PERSONALIZED EMAIL COMMUNITIES

The E-mail data source is a data set, which contains user mail information. Multi-user is defined as the person who is having more than one account. Email data discover a multi-user personalized community to represent the grouping of individuals with similar neighborhood and communication behavior. Initially, multi-user information is acquired from more than one email account.

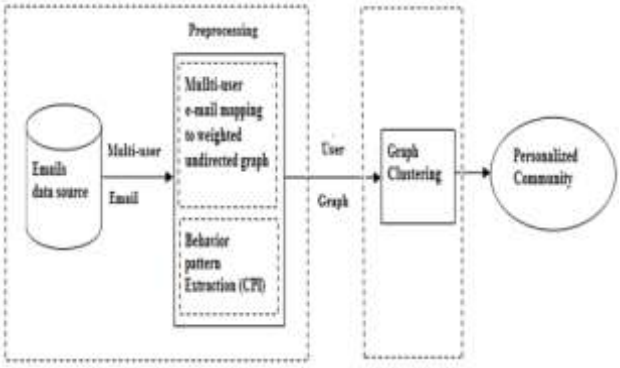

Fig. 1: Proposed architecture for Pi- Net system

In the preprocessing phase, the interactions among the users are represented in the form of an undirected weighted graph (UW-Graph) as illustrated in Fig. 1. Emails from multiple user accounts are provided as an input to the system. This information is directly extracted from personal emails of corresponding user under the constraint of privacy. The UW-Graph describes and analyzes the behavioral patterns. The CPI is analyzed for communication behavior extraction with the help of attribute or meta-data, which is subject length, text size, email size, time.

IV. MULTI-USER PERSONALIZATION

This section discusses the concept of personalization with respect to the user roles and behaviors. Personalized emails are defined as structured mails from a set of users having a header structure as From, To, CC and BCC sections.

User Roles are defined based on the either the user is a sender or a receiver. User Behavior defines the structure of the mails sent and received by the user. Pi-Net is a network constructed form the personalized emails Communication Behavior Pattern (CPI) is used for arbitrary user.

$$CPI = \{a_{ij} | i = 1, \ldots, N_a\} \quad (1)$$

$$Influential\ CPI = argmax_j \{Frequency\ (a_{ij})\} \quad (2)$$

where $j = 1, \ldots, n_{ai}$.

$$TAG = \{T_a | a \in \{SubLen, TxtSize, EmailSize, Time\}\} \quad (3)$$

where $T_a = \{tag_i | i = 1, \ldots, n_a\}$

Each user has a set of labels, T_a represented as tag_i where n_a represents the total number of labels. The value of i^{th} attribute a_i is expressed as a_{ij}, N_a refers total number of attribute, and n_{ai} represents possible values for each attribute ai, which is used to find the behavior aspect of appropriate user through attribute values.

https://dx.doi.org/10.22161/ijaers.4.2.11 *ISSN: 2349-6495(P) | 2456-1908(O)*

V. EXPERIMENTAL STUDY

This section elaborates on the formation of the Pi- Net.

Data source:

The Enron corpus is a collection of 600,000 emails received from 158 employees. The Enron data was originally collected by Enron Corporation [14].

Result on Multi-user Personalization behavior:

The Fig. 2 depicts the pattern of communication among multi-users but holding the account only in a single extension ie., only in enron mail account.

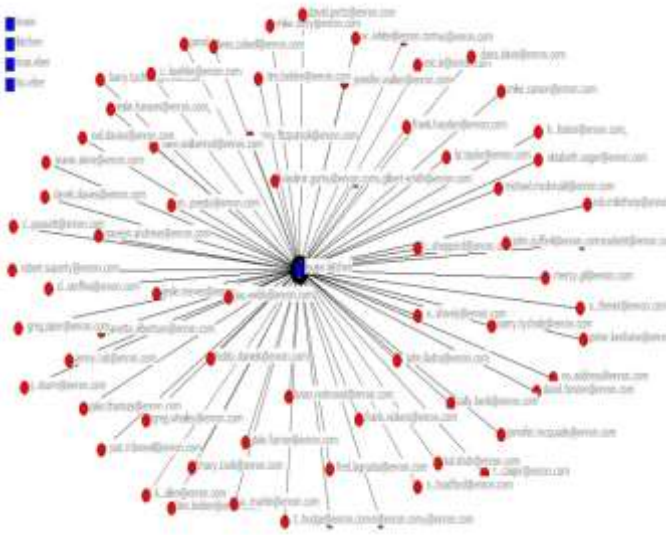

Fig.2: Multi-user to single account

The sample e-mail in Table 1 defines the enron to enron communication among different individuals. In the table, the value 1 represents the conversation presence between the two enron e-mail accounts.

Table 1. Sample index Table

Other enron mail accounts	**louise.kitchen@enron.com**
f..calger@enron.com	**1**
no.address@enron.com	**1**
liz.taylor@enron.com	**1**
david.oxley@enron.com	**1**
mercy.gil@enron.com	**1**
tim.belden@enron.com	**1**
eddy.daniels@enron.com	**1**

The Fig. 3 illustrates multi- user fused Pi- Net for three uni - account communications.

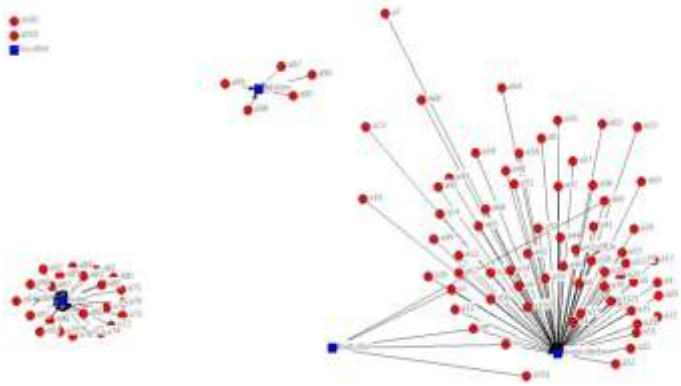

Fig. 3: Multi-user fused Pi-Net

The sample fused network is shown in Table 2. It represent the communication between enron to different user, C4 represents the louise and C5 represent the lou.eber where 1 represent the conversation occur between multi-user to different user. C1 represents Cluster1 communication from louise@enron.com to other people (for example polo@polo.com). C2 represent Cluster 2 is like to communicate with people from lktichen@enron.com to info@intellor.com. C3 represent Cluster 3 to communicate with people from, louise.eber@enron.com and louise.kitchen@enron.com to s.bradford@enron.com.

Table.2. Fused network table

Other mail accounts	**C4**	**C1**	**C2**	**C3**	**C5**
s.bradford@enron.com		**1**		**1**	
no.address@enron.com		**1**			
liz.taylor@enron.com		**1**			
david.oxley@enron.com		**1**		**1**	
polo@polo.com	**1**				
orders@gymboree.com	**1**				
info@intellor.com			**1**		
eblastoff@rocketball.com			**1**		
john.sherriff@enron.com					**1**

The TAG is formed based on the attributes from metadata like email size, text size, time and subject length. This meta-data forms the semantics of an email. User community with homogeneous behavior and their mutual analysis requires similarity in interpretation, which can be achieved through scaling. The time stamp of an email is categorized and then assigned labels on regular basis. Every attribute is scaled for similarity and each attribute has a set of label. The scaling is performed based on the minimum and maximum approximation. The scaling is as shown below:

https://dx.doi.org/10.22161/ijaers.4.2.11 *ISSN: 2349-6495(P) | 2456-1908(O)*

Subject Length: It reflect the number of characters which is present in the subject content in a user email.

Text Size: Content size of the e-mail. The email in a loop to many people will be inflated as it will include the forwarded contents also.

Email Size: Email size depicts the overall size of email and attachments. The users can also send non-textual content in the email that cannot be easily found by other properties.

Date and Time: An email has given a time/date stamp where the users exchange it with other users.

The range is depicted under three categories as shown below:

Table.3: Subject length

Categories	Subject Length
Too Short	0-20
Medium	21-40
Minimum	41-82

Table.4: Text size

Categories	Text Size
Little	164-455
Small	456-700
Low	701-900
High	901-1395
Maximum	1396-1995
Large	1996-3000
Too high	3001-7339

Table.5: Email size

Categories	Email Size
Less	646-900
Average	901-1499
Moderate	1500-3902
Short	3903-5858
Huge	5859-8522
Extra large	8533-17525

Table.6: Time

Categories	Time
Morning	4am-12pm
Afternoon	12pm-4pm
Evening	4pm-8pm
Night	8pm-12am

The above attributes are analyze and scaled according to the multi-user email conversation.

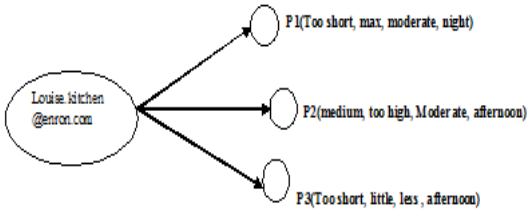

Fig.4: UW-graph with CPI for C3

The UW graph shows the conversation between multi-user to different user according to the above attributes. C3 represents cluster 3 and the graph is depict for louise kitchen conversation to different users. Here P1 represents f.calger@enron.com, P2 represents david.oxley@enron.com and P3 represents jeff.golden@enron.com.

V. CONCLUSION

The email network is uniquely represented as a social graph using the email interactions. This social graph is clustered using communication patterns and the behavioral bonding is identified. The community dynamics was experimented using the enron email dataset. The TAG attributes are used to determine the structural and semantic similarity.

REFERENCES

[1] Biswas, A., & Biswas, B, "Investigating community structure in perspective of ego network", Expert Syst. Appl., 42 , 6913-6934, 2015.

[2] Boykin, P., & Roychowdhury, V, "Personal Email Networks: An Elective Anti-Spam Tool", Technical Report University of California, Los Angeles, 2004.

[3] Boykin, P. O., & Roychowdhury, V. P, "Leveraging social networks to fight spam Computer", 38, 61-68, 2005.

[4] Cheng, H., Zhou, Y., & Yu, J. X. (2011), "Clustering large attributed graphs:A balance between structural and attribute similarities", ACM Trans. Knowl. Discov. Data, 5 , 12:1-12:33, 2011.

[5] Fortunato, S, "Community detection in graphs", Physics Reports, 486, 75-174. doi:10.1016/j.physrep.2009.11.002, 2010.

[6] Hu, P., & Lau, W. C, "Localized algorithm of community detection on large-scale decentralized social networks", CoRR, abs/1212.6323, 2012.

[7] Johnson, R., Kovcs, B., & Vicsek, A, "A comparison of email networks and off-line social networks: A study of a medium-sized bank", Social Networks, 34, 462-469, 2012.

https://dx.doi.org/10.22161/ijaers.4.2.11 *ISSN: 2349-6495(P) | 2456-1908(O)*

[8] Johansen, L., Rowell, M., Butler, K., & Mcdaniel, P, "Email communities of interest", In Proceedings of the 4th Conference on Email and Anti-Spam (CEAS) CEAS '07, 2007.

[9] Martin, S., Sewani, A., Nelson, B., Chen, K., & Joseph, A. D, "Analyzing behavioral features for email classication", In Berkeley, CA: University of California at Berkeley, 2005.

[10] Nagwani, N. K., & Bhansali, A, "An object oriented email clustering model using weighted similarities between emails attributes", International Journal of Research and Reviews in Computer Science (IJRRCS), 1 , 1-6, 2010.

[11] Nawaz, W., Lee, Y.-K., & Lee, S, "Collaborative similarity measure for intra graph clustering", In DASFAA Workshops (pp. 204-215), 2012.

[12] Nawaz, W., Han, Y., Khan, K.-U., & Lee, Y.-K, "Personalized email community detection using collaborative similarity measure", CoRR, abs/1306.1300, 2013.

[13] Papadopoulos, S., Kompatsiaris, Y.,Vakali, A., & Spyridonos, P,"Community detection in social media". Data Mining and Knowledge Discovery, 24.3: 515-554, 2012.

[14] Shetty, J., & Adibi, J, The enron email dataset database schema and brief statistical report, 2004.