# COMMUNITY PROFILING FOR SOCIAL NETWORKS

**Ch. Archana[1], K. P. Supreethi[2]**

[1]*CSE, JNTUCEH, Hyderabad, Andhra Pradesh, India*
[2]*CSE, JNTUCEH, Hyderabad, Andhra Pradesh, India*

## Abstract

*A Community Profile is a combined picture of the members data in the community and the process of creating it is called community profiling. It uses a variety of different techniques to build up a picture of the community from a number of different perspectives. The members in the Community can interact with other members in the same community using the community profile. The communities are created dynamically based on the users' interests and one user can be in more than one community. The Community Profile of a particular community can be accessed by only the members of that associated Community.*

***Keywords:*** *Community Profiling, Social Networks, Clustering, Cliques, Summarization algorithms, Optimal Summarization algorithm.*

----------***----------

## 1. INTRODUCTION

Social media are tools and technologies that enable you to communicate with other people. Social media includes blogs, wikis, video, photo sharing. Communication is very important in human life for proper interaction and for the development of our society. It brings people with common interests together and also helps people to interact with other people comfortably via the Internet. Online media have the ability to distribute the information as fast as the traditional news sources.

Social Network is a Social media through which users can communicate with other users via Network. The communication can be in the form of messages or content sharing. Now – a – days, Social Networking Sites have become very popular as people can be in contact with each other, even though there is no face – to – face communication. People can discuss the relevant topics, ask the questions and get the answers.

Community is a group of members who have same interests in a certain area. Profile is a term given for information of a particular thing. For example, a Person's Profile includes age, gender, contact information, address, qualification, experience, interests, etc. Profiling is a process of analyzing the people's profiles for identifying their respective categories(called communities).

"Community profiling" is a process of analyzing the profiles of the Communities in order to find out the data or information that is relevant to each member in a particular Community. The Community Profile is the analyzed and reported results of the data collected by a Community survey, which describes a combined/overall picture or profile of the population/members in the Community.

## 2. RELATED WORK

The detection of communities(also called clusters) is implemented using clustering concepts.

### 2.1 Clustering

Clustering is a data mining (machine learning) technique used to group the related data elements without any knowledge about the group definitions(details).

#### 2.1.1 Clustering Techniques

- **Agglomerative:** Considers each element as one cluster and compares with all other elements(clusters) iteratively to combine the clusters having similar features[1].
- **Divisive:** Considers all elements as one cluster and the whole cluster is divided into two parts after comparing the elements similarity within the cluster and this is repeated iteratively [1].
- **Hierarchical:** Organizes all elements into a tree in which leaves represent genes and the length of the paths between leaves represents the distances between genes. A subtree represents the similar genes[1].

The profiling of the detected communities(called community profile) is implemented using summarization concept.

### 2.2 Summarization:

*Summarization* is a concept of creating a summary with necessary components from huge information. Summarization is a data mining concept which involves techniques for creating a summary/description of each cluster (community).

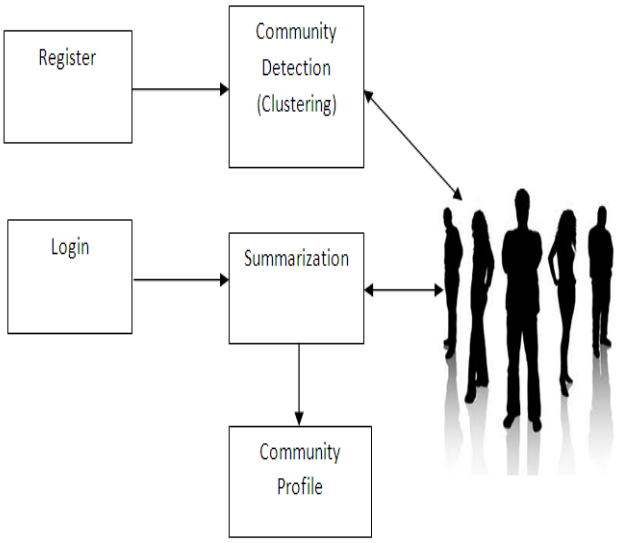
Multi-document summarization [6] is the process of creating a summary from multiple documents. There are two approaches (Extraction and Abstraction) to produce concise summary. Extraction approach takes the most important information from the original source documents to produce summary while abstraction involves paraphrasing sections of the original source documents.

### 2.2.1 Summarization Techniques:

- Optimal summarization algorithm[7].
- Two – step approach to summarization [7].
- The BUS algorithm[7].

## 3. PROPOSED WORK

We detect the communities using *clustering concept* and we create the community profile for each community using *summarization concept*. The block diagram of the project is shown in the following Figure (a).




**Fig(a)**Block Diagram

### 3.1 Process of Clustering

Define the domain for the clustering effort. It is the determination of the set of items to be clustered. The domain can be a subset of the database or the complete database. If domain is defined, determine the attributes of the objects to be clustered.

Determine the strength of the relationships between the attributes which suggest those objects should be in the same class (cluster).

### 3.1.1Types of relationships:

Equivalence relationships are the most common and represent synonyms (Ex: photograph, print). Hierarchical relationships are the very common technique where the class name is a general term and the entries are specific examples of the general term (Ex: “computer” is the class name and “microprocessor”, “Pentium” are the entries). Non – hierarchical relationships are the “object-attribute” relationship (Ex: employee-job title) [2].

After determining the total set of objects and the strengths of the relationships between the objects, the final step is applying some algorithm to determine the classes(clusters or communities) to which each item will be assigned[2].

### 3.1.2 Types of algorithms:

- Cliques
- Single link
- Stars[2]

The detection of communities is implemented using Cliques algorithm [2].

### Cliques Algorithm:

Cliques require all items in a cluster to be within the threshold of all other items. The clique technique:

- Produces classes having terms with the strongest relationship.
- Produces more classes.
- Most costly to compute.

Algorithm:
1. Let $i = 1$
2. Place Term $_i$ in a new class
3. $r = k = i + 1$
4. Validate if Term $_k$ is within the threshold of all terms in the current class and add to the current class
5. $k = k + 1$
6. If $k > n$ (number of items)
the n$r = r + 1$
if $r >= n$ then go to 7
else
$k = r$
Create a new class with Term $_i$ in it
go to 4
else go to 4
7. If current class has only Term $_i$ in it and there are other classes with Term $_i$ in them
then delete the current class and $i = i + 1$
else $i = i + 1$

8. If $i = n + 1$ then go to 9
else go to 2
9. Eliminate any classes that duplicate or are subsets of other classes.

The basis for determining the clusters [2] is the calculation of the similarity between every term pair. Considering the vector model is the easiest way to understand this approach. The vector model is represented by a matrix where the rows represent individual items and the columns represent the unique words (processing tokens) in the items. The values in the matrix represent how strongly that particular word represents concepts in the item. Figure 3.1.3 is an example of a database with 5 items and 8 terms.

| | T1 | T2 | T3 | T4 | T5 | T6 | T7 | T8 |
|---|---|---|---|---|---|---|---|---|
| I1 | 0 | 4 | 0 | 0 | 0 | 2 | 1 | 3 |
| I2 | 3 | 1 | 4 | 3 | 1 | 2 | 0 | 1 |
| I3 | 3 | 0 | 0 | 0 | 3 | 0 | 3 | 0 |
| I4 | 0 | 1 | 0 | 3 | 0 | 0 | 2 | 0 |
| I5 | 2 | 2 | 2 | 3 | 1 | 4 | 0 | 2 |

**Fig. 3.1.3**Vector Example

Where I and T represent the Items and Terms respectively

The similarity function is performed between rows of the item matrix. Using figure 3.1.3 as the set of items and their terms and similarity equation:

$$SIM (Item_i, Item_j) = \sum( Term_{i,k} )( Term_{j,k} )$$

as k goes from 1 to 8 for the eight terms, an Item-Item matrix is created as shown in Figure 3.1.4

| | I1 | I2 | I3 | I4 | I5 |
|---|---|---|---|---|---|
| I1 | - | 11 | 3 | 6 | 22 |
| I2 | 11 | - | 12 | 10 | 36 |
| I3 | 3 | 12 | - | 6 | 9 |
| I4 | 6 | 10 | 6 | - | 11 |
| I5 | 22 | 36 | 9 | 11 | - |

**Fig3.1.4**Item/Item Matrix

The Item Relation matrix shown in Figure 3.1.5 is produced using the threshold of 10.

| | I1 | I2 | I3 | I4 | I5 |
|---|---|---|---|---|---|
| I1 | - | 1 | 0 | 0 | 1 |
| I2 | 1 | - | 1 | 1 | 1 |
| I3 | 0 | 1 | - | 0 | 0 |
| I4 | 0 | 1 | 0 | - | 1 |
| I5 | 1 | 1 | 0 | 1 | - |

**Fig. 3.1.5**Item Relationship Matrix

The following classes/clusters/communities are created when the Cliques algorithm is applied to Figure 3.1.5.
Class1 = I1, I2, I5
Class2 = I2, I3
Class3 = I2, I4, I5

Mining process [4] includes the sharing of large scale amount of data from various sources, which gets concluded at the mined data.

Continuing growth of World Wide Web and on-line text collections makes a large volume of information available to users. The user may not get the effective required information as the user need to browse all the information available which is a time consuming process and some useful information may be missed out. Using automatic text summarization, the above problem can be solved and therefore, the required information can be displayed to the user effectively. Text summarization is the process of automatically creating a compressed/brief version of a given text that provides useful information to users, and multi-document summarization is to produce a summary consisting the majority of information content from the original documents [5].

The profiling of the detected communities is implemented using Optimal Summarization algorithm[7].

## 3.2 Optimal Summarization Algorithm:

Optimal Summarization algorithm [7] generates an optimal summary of given size *l* for a given set of transactions T The first step of this algorithm involves generating the power set of T (=all possible subsets of T), denoted by ζ. The size of ζ will be $2^{|T|}$. The second step involves searching all possible subsets of ζ to select a subset, S which has following properties:

1. $|S| = l$, the size of this subset is equal to desired compaction level.
2. The subset S covers all transactions in T (a set cover of T)
3. The total information loss for S with respect to T is minimum over all other subsets of T which satisfy the properties 1 and 2.

*Information loss [7]* is defined as the total amount of information missing over all original data transaction in the summary.

Algorithm:
**Input**: T: a transaction data set.
W: a set of weights for each feature.
*l*: size of final summary.

**Output**: S: the final summary.
**Method**:

1. Generate $\zeta$ = power set of T
2. Let *current_min_loss = inf*
3. Let S = {}
4. For each $\zeta_i \varepsilon \zeta$
5. if $|\zeta_i| = l$ And Information Loss for $\zeta_i$<*current_min_loss*
6. *current_min_loss* = Information Loss for $\zeta_i$
7. $S = \zeta_i$
8. End if
9. End For each
10. Return S
11. End.

## 4. EXPERIMENTAL EVALUATION

The Community detection of the users' data and Community profiling for each detected community is implemented using JSP (as Front End), Java (for Business Logic) and Oracle (as Back End).

For the detection of communities, we have collected the users' data in the form through Facebook. The form consists of user's full name, email and password, age, qualification, designation, area and interests. The user can select several interests provided in the form.

Based on the users' interests, the users are categorized into different groups (called communities/clusters) using Cliques algorithm and are detected at the time of user's Registration. The detected communities are displayed to the user (as hyperlinks), when the user logs in. The Community Profiling for the detected communities is implemented using Optimal Summarization algorithm. It is created when the user clicks on the communities (Hyperlinks) and the overall profile of the user clicked community is displayed to the user.

## 5. CONCLUSIONS

The detection of communities and creation of community profile for each community has been implemented. A Community Profile is that if anyone joins in a particular Community, then the person can view the overall Profile of the Community through which the person can communicate with others about the topics, related to the interests provided in the profile. The Cliques algorithm has been implemented for the detection of communities as a member can be in more than one community and Optimal Summarization algorithm has been implemented for creating the profiles for the detected communities. We plan to evaluate our approach with otherreal world datasets and to improve our approach based on the obtained results.

## REFERENCES

[1].Charu C. Agarwal. "Social Network Data Analytics", ISBN 978-1-4419-8461-6, Springer, USA, 2011.

[2]. Gerald J. Kowalski, Mark T. Maybury. "Information Storage and Retrieval Systems, Document and Term Clustering", Second Edition, ISBN 978-81-8128-497-6, Springer, India,1997

[3].RaghunathKar&Susant Kumar Dash. "A Study On High Dimensional Clustering By Using Clique".International Journal of Computer Science & Informatics, Volume-I, Issue-II, 2011

[4].SM. Meena. "Dynamic Peer-to-peer Distributed Document Clustering and Cluster Summarization". Chennai and Dr.MGR University Second International Conference on Sustainable Energy and Intelligent System (SEISCON 2011) , Dr. M.G.R. University, Maduravoyal, Chennai, Tamil Nadu, India. July. 20-22, 2011

[5].ZHANG Pei-ying, LI Cun-he. "Automatic, text summarization based on sentences clustering and extraction". 978-1-4244-4520-2/09/$25.00 ©2009 IEEE.

[6].A.Kogilavani, Dr.P.Balasubramanie. "Clustering Based Optimal Summary Generation Using Genetic Algorithm".Proceedings of the International Conference on Communication and Computational Intelligence – 2010, Kongu Engineering College, Perundurai, Erode, T.N.,India.27 – 29 December,2010.pp.324-329.

[7].VarunChandola and Vipin Kumar."Summarization - Compressing Data into an Informative Representation". Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05) 1550-4786/05 $20.00 © 2005 IEEE