Role of Natural Language Processing in Community Structure Detection

Vivek Bhola¹, Mahesh Grover² and Madhavi Sinha³
Birla Institute of Technology, Mesra(Ranchi) Extension Center Jaipur.
27, Malviya Industrial Area, Malviya Nagar, Jaipur-302017, E-Mail:¹vivek.vatts007@gmail.com

ABSTRACT
In this paper, relationship as a metric in the community is introduced for community detection in a social network. The relationship metric is inspired by Newman’s Edge Betweenness metric but an attempt to look at the community detection problem from the angle of Natural Language Processing. In this approach, a social network is developed on the basis of blog data or email archives and later it’s lexical and semantic analysis is done to find out the relationships between each and every individual. Finally a social network structure is achieved where vertices represent individuals and edges between vertices represent their relationships. To detect the communities, it is needed to decide what type of community is of our interest. In other words we need to decide the relationship metric of our interest and then relationships (edges) which are not of our interest are removed. Thus a social network of edges having the same type of relationship is achieved. On this achieved social network, a divisive algorithm proposed by Girvan and Newman is applied that uses edge betweenness as a metric to identify the boundaries of communities.

KEYWORDS
Social Networks, Edge Betweenness, Relationship metric, Natural Language Processing, Lexical Analysis, Syntactic analysis, Semantic Analysis, Word Sense Disambiguation.

1. INTRODUCTION
Social Network explicitly exhibit relationships among individuals and groups and it play important roles in our daily lives. People conduct communications and share information through social relations with others such as friends, family and collaborators. An important fact is that a single person can have different relationships with different people, for example, a person can share professional thoughts with some other people and in the same network that person can also share religious thoughts with another group of people. Thus, a single person can be involved in multiple relationships with different groups of people. These relationships in a network can be of Professionals, friends, same religion followers etc. Social networks can be obtained from various sources such as e-mail archives, message boards and Social Networking services etc. Hence, a social network might have multiple small or large communities or can have a single community in its own on the basis of relationships exist between individuals. A social network, as stated by Krebs and Holley (2004) [1], is generally built in four phases, each is with its own distinct topology (as shown in figure 1).

1. Scattered clusters,
2. Single hub and spoke,
3. Multi-hub small-world network, and
4. Core/periphery

Figure 1. Four phases of social structures (from Krebs and Holley 2004)

Social Network Structures are the outcome of the relationships between individuals interacting within a group. Initially, multiple small groups of people with different relationships are formed. Keeping the phases of social networks by Krebs and Holley (2004) [1] in mind, these small groups can be called as scattered clusters of single relationship. Gradually over a period of time these small scattered clusters form themselves into single hub and spokes and later convert themselves into multi-hub networks and finally into core/periphery. The core/periphery model is the most stable structure. In the core/periphery model, the network core encompasses key group members who are strongly connected to each other, while the periphery contains members who are usually weakly connected to each other as well as to the core members. Hence in a large group of people, patterns of various types of relationships between individuals may coexist simultaneously. These relationships can be analysed using Natural Language Processing and can be used as metric to extract a subset of social network having the same kind of relationship between individuals. Finally, Girvan and Newman’s algorithm of edge betweenness [2] is applied on achieved subset of social network to detect the communities.
2. ROLE OF ‘RELATIONSHIP’ METRIC IN COMMUNITY DETECTION

Girvan and Newman proposed a divisive algorithm that uses edge betweenness as a metric to identify the boundaries of communities. This algorithm focuses on finding edges with the highest “betweenness”, where betweenness is some measure that favors edges those lie between communities and disfavors those that lie inside communities. Relationship metric also works with a divisive algorithm that favors those edges having relationship similar to selected relationship metric and disfavors those edges having not similar to selected relationship metric. In Girvan and Newman algorithm [2], edges having highest betweenness are removed from the social network to identify the communities. Similarly in this approach also, dissimilar edges compared to relationship metric are removed from the social network. Thus both “Edge betweenness” and “Relationship” metrics are used to identify an edge that needs to be removed from the network. As we discussed earlier, multiple relationships may coexist simultaneously in a social network, so if initially a social network of a similar edges (relationship) is achieved then the complexity of Girvan and Newman algorithm can be minimized because the total number of edge traversals are reduced.

3. ROLE OF NATURAL LANGUAGE PROCESSING

In this approach, a relationship metric is selected on the basis of a subset of social network having similar kind of edges (similar relationship) achieved from a large Social Network Structure. Therefore, for the establishment of that metric, first of all, blogs or email archives of individuals are analyzed to find out the relationships which exist between them. To analyze the data of blogs or email archives, lexical analysis, syntactic analysis, and semantic analysis [7] of the corpus of blog communication or email archives is done.

1. Lexical Analysis
   - First, lexical Analyser reads blog data or email archives data in a stream of characters.
   - Second, identifies the lexemes (string of characters) in the stream through keyword matching.
   - Third, categorizes these lexemes into tokens (lexeme’s type combined with its value).

2. Syntactic Analysis
   - Syntactic analyser checks that whether the tokens form an allowable expression.

3. Semantic Analysis
   - In this step, meaning making of data of blogs or email archives are done and relationships exists between individuals are found.

For example,

*I want to talk about the materials that marketers provide our sales teams.*

Lexical analysis of this sentence tag these bold faced words after matching with words of domain specific database. After lexical analysis, a semantic analysis of these tagged words is performed with Word Sense Disambiguation algorithm to extract the actual sense in which this sentence has been delivered. Resulting sense indicates a particular relationship between these two people.

4. DETECTION OF COMMUNITY STRUCTURES IN A NETWORK

First, a social network is generated by analyzing the blog or email archives. This social network shows only the connectivity between all the vertices. Second, blog or emails archives are parsed by Natural Language Processing techniques to find out the relationships between the individuals. It results with a social network having edges marked with various relationships. Third, social network structure is passed to community detection algorithm where social network is filtered with an already decided relationship metric and processed according to Girvan and Newman algorithm.

![Figure 2: Community detection system.](image-url)

Here U1, U2, U3…..Un are individuals who are interacting with each other through Blog or E-Mail application. D1, D2, D3…..Dn are individual’s blog data or E-Mail archive. Let’s assume, there exist multiple relationships, two of them are Professional and Religious. So, relationship can be termed as Rp for professionals, Rr for same religion followers and other relationships are represented as Ro.

Following are the steps for community detection system:

1. Develop a social network by making interaction between all the individuals. Or a blog or email data can also be used to develop a social network.
2. Infer the relationships (edges) between individuals (vertices) by analysing the blog communication or email archives with Natural Language Processing Techniques.
3. Mark the edges on the basis of relationship between vertices (people).
4. Decide the relationship metric or community type to extract community from the network.
5. Attempt to find the edges or relationships which are not of interest.
6. Remove the edges which are not of interest.
7. Apply the Girvan and Newman algorithm of edge betweenness on achieved social network.

**Figure 3**: Social Network structure developed after analyzing data of blog.

Showing the connectivity between the vertices.

- Religious relationship
- Other Relationships
- Professional Relationship

**Figure 4**: Social Network structure developed after Natural Language processing of blog data.

Showing the relationships between the vertices.

**Selected Measure**: Communities of Professionals

**Figure 5**: Achieved subset of Social Network on the basis of Selected Relationship metric.

**Figure 6**: Detected Communities on the basis of Selected Relationship metric after Girvan and Newman algorithm.

**5. CONCLUSION**

In this paper, community detection in social network structure has been seen from a different point of view. Blogs or email archives are generic in nature and they might have those relationships between individuals which are neither practically relevant nor of our interest. Therefore, a metric for relationship in the community is introduced for community detection in a social network. This metric for relationship removes all the irrelevant edges from the social network in early stages of this approach. Thus, total number of edge traversals in Girvan and Newman algorithm is reduced. Hence complexity of algorithm is minimized.

**6. FUTURE SCOPE**

In future, more stress will need to be given on Relationship extraction from large corpus of Email archives or blogs. An experiment would be performed in a virtual environment. In that virtual environment, keywords will be thrown by the individuals and Natural Language Processing methods will try to extract various relationships between the individuals. We will also try to get concrete results in the form of communities at the minimized complexity.

**7. ACKNOWLEDGEMENT**

This paper work has been carried out under the sponsorship of Department of Science and Technology of India for the project “Soft Modeling of Group Dynamics: A Cognitive Psychology Perspective”.

**REFERENCES**


[6]. Deo. N, “Graph Theory with Applications to Engineering and Computer Science”, Prentice Hall India Pvt., Limited, India, 1974