**A THESIS**

**FOR THE DEGREE OF MASTER OF SCIENCE**

# Relation Discovery Mechanism in Heterogeneous Information Networks

Using clustering for discovering relationships in Heterogeneous Information Networks

MUHAMMAD SHOAIB

DEPARTMENT OF COMPUTER ENGINEERING

GRADUATE SCHOOL

JEJU NATIONAL UNIVERSITY

2013.08

석사학위 논문

# 이질적 정보 네트워크 에서의 관계 발견 메커니즘

## 클러스터링 기반의 이질적 정보 네트워크에서의 관계발견

무하마드 소아입

제주대학교 대학원

컴퓨터공학과

2013.08

# Relation Discovery Mechanism in Heterogeneous Information Networks

**Muhammad Shoaib**

**(Supervised by Professor Wang-Cheol Song)**

A thesis submitted in partial fulfillment of the requirement for the degree of Master of Science in Computer Engineering

2013. 08

This thesis has been examined and approved.

……………………………………………………………………

**Thesis Director, Khi-Jung Ahn, Professor, Jeju National University**

……………………………………………………………………

**Thesis Director, Jung-Hoon Lee, Professor, Jeju National University**

……………………………………………………………………

**Supervisor, Wang-Cheol Song, Professor, Jeju National University**

**Department of Computer Engineering**
**GRADUATE SCHOOL**
**JEJU NATIONAL UNIVERSITY**
**REPUBLIC OF KOREA**

To my parents

# Acknowledgements

All praise and glory belong to Allah, Who has blessed me with the power, health, knowledge, intellect, and abilities in order to complete this thesis. This work would not have been possible without His blessings. I ask Allah to send Peace and Blessing upon all Prophets particularly on Prophet Muhammad, his pure household and his noble companions.

I would like to extend my thanks to my advisor Prof Wang-Cheol Song for his kind support, guidance and valuable advices during my studies at Jeju National University. I am very grateful to Prof Song for giving me a chance of working in his lab as a Master Student. I also would like to extend my thanks to Prof. Khi-Jung Ahn and Prof. Jung-Hoon Lee, for their value able suggestions, which helped me a lot in order to improve this thesis. I also would like to thank to Dr. Farrukh Aslam Khan for his kind recommendation at Jeju National University. I also would like to say thanks to Mr. Kazim Ali Syed for his kind help and support during my studies.

I would like to offer humble gratitude to my mother who not only put her all effort in her capacity for my future but she always encouraged me during my studies and research. Without her prayers, support and love any of my achievement would not have been possible. I also would like to present my humble gratitude to my father, who have been a great source of inspiration for me, and who taught me patience, honesty, courage and importance of knowledge. I also thank my brothers Ghasan, Raheeq, Talha and Awais for their love.

I would like to extend my thanks to my friends Safdar Ali and Rashid Ahmad for their kind help, support, healthy discussions and valuable suggestion. I also thank Chin Shu for his help and support during my stay in Jeju, without him my stay in Jeju would not have been joy able. Furthermore, I would like to thank all of my friends in Korea, in particular Dr. Muhammad Naeem Awais Dr. Muhammad Nauman Malik and Dr. Naveed Ejaz for their kind support during my stay at Jeju National University. I would like to say thanks to Mrs. Eun Gyoung Joung for her kind help, support and assistance during my early days stay at Jeju National University. In the end I would like to say thanks to all my lab mates in particular Jin-Hyeok Kang, Ji-Hoon Hong and Dong-Seok Jang for their help. It was really a great time with them. Thank you all.

Last but not the least; I would like to present my special thanks and gratitude to Ms. Amna Basharat for her kind support, guidance encouragements, and prayers that have been always with me during my study and research. She has been always a great source of inspiration for me and she always thought me to be patient, persistent sincere and honest in the life.

Muhammad Shoaib

June 2013

# Table of Contents

# List of Figures

# List of Tables

# Abstract

Graph data structure is being widely used for modeling various problems of the real life. As it is a fact that data of activities that people generate in our daily life is connected with each other and these connections form a network. Therefore study of network has been highly encouraged during last two decades particularly in the domain of computer networks and information systems. Any problem that has a graph orientation can be modeled as a network. These networks consist of vertices and edges, vertices are nodes that represent daily life objects either physical or meta-physical and edges represent the relationship among these vertices. This relationship can be numeric or descriptive.

As huge amount of networked applications have been marketed to date therefore Graph Mining has also obtained large attention in the data mining research community. Because of its structural difference from conventional data, conventional data mining techniques cannot be applied for graph mining directly that leads towards the need for the development of specific techniques for mining graph datasets by taking into account the properties of graph data.

Finding relationship among different pieces of data has been remained an interesting challenge in data mining domain to date. Based on different datasets researchers always keep trying to discover relationship among different objects. However all efforts that are made to date in this domain have been made on homogenous networks as it is easy to deal with the homogenous networks particularly those datasets that can easily be converted into adjacency matrices.

Another limitation that has been observed is that the most of relationship discovery techniques work on the numeric datasets. In order to overcome these drawbacks, this thesis presents a mechanism for discovering relationship among the objects in a heterogeneous network by utilizing the clustering concept. The proposed method consists of three major steps

- Clustering objects present in a heterogeneous information network
- Discovering relationship between objects present in one cluster by taking cluster centroids into account
- Using probabilistic method to find relationship among objects in different clusters

In order to cluster objects this thesis introduces a new hybrid technique that find the similarity between objects not based on structure only but also utilizes the data and relationships that are present on the edges and the values of the nodes of that a network is composed. In this way firstly objects based on structure are clustered to their respective clusters and then in the second iteration those cluster objects are clustered based on their relationships with the other objects. After performing the clustering this thesis has introduced a probability based relationship discovery mechanism to identify the hidden relationship among different clusters using the similarity matrix. For this the creation of new graph comprises of clusters as the nodes and common relationships among these clusters as the edges has been proposed in order to discover the relationship among different clusters. The proposed has been examined on news-twitter dataset and results shows the presented techniques for clustering and relation discovery perform better then present k-mean and k-medoids methods.

# 개요

그래프 데이터 설계는 실생활에 많은 문제들을 모델링 하기 위해 널리 사용되고 있다. 우리가 일상생활에서 만드는 활동에 대한 데이터는 서로 연결되어 있고, 이러한 연결은 하나의 네트워크를 형성한다. 그러므로 네트워크 연구는 지난 20년동안 컴퓨터네트워크와 시스템 정보의 영역에서 매우 장려되어 왔다. 그래프 성향을 가진 어떤 문제던지 네트워크로 모델링을 할 수 있다. 이 네트워크는 꼭지점(vertices) 와 에지(edges)로 구성되어 있어, 꼭지점(vertices)은 일상 생활 객체를 물리적 혹은 메타-물리적 객체로 나타내는 노드이고 에지(edges)는 꼭지점(vertices) 들의 관계를 나타낸다. 이 관계는 수적이거나 서술적 일 수 있다.

지금까지 엄청난 양의 네트워크 어플리케이션이 세상에 나왔고, 따라서 그래프 마이닝은 또한 데이터 마이닝 연구 커뮤니티에서 많은 관심 얻었다. 일상적 데이터와의 구조적 차이로 인해서, 일상적인 데이터 마이닝 기술은 그래프 마이닝에 직접적으로 적용될 수 없었고, 그래프 데이터 셋을 마이닝 하기 위해 그래프 데이터의 속성들을 고려하여 특정한 기법을 개발할 필요가 생기게 되었다.

다른 데이터 조각들 사이의 관계를 찾는 것은 지금까지 데이터 마이닝 영역에서 흥미 있는 도전으로 남아있다. 다른 데이터 셋을 기반으로 연구원들은 항시 각각의 개체들 사이에서의 관계를 찾으려 시도하고 있다.

그러나 이 영역에서 지금까지 이뤄진 모든 노력 은 동질의 네트워크에서 이루어져왔다. 이는 동질의 네트워크를 다루는 것이 쉽기 때문인데, 그 데이터 셋이 인접 매트릭스를 쉽게 변환 될 수 있어서이다. 관찰 된 또 다른 한계점은 대부분의 관계성 발견기법들을 수적인 데이터 셋에 작용하고 있다는 것이다.

본 논문에서는 이러한 단점을 극복하기 위하여, 클러스터링 개념을 활용하여 이질적 네트워크에서 객체들 사이의 관계성을 발견하기 위한 메카니즘을 제시한다.

- 이종정보 네트워크 에서 객체들의 클러스터링
- 클러스터 센터로이드를 고려하므로써 하나의 클러스터에 있을 객체들 사이의 관계성을 발견하기
- 다른 크러스터에 있는 개체들의 관계를 찾기 위해 확률론적 방법을 사용하기

크러스터 개체를 위하여 우리는 구조에만 기반하여 객체들 사이의 유사성을 찾을뿐 아니라 네트워크가 구성되는 노드들에 대한 에지(edge)와 값들에 나타나는 관계 및 데이터를 이용하는 하이브리드 테크놀로지를 소개한다. 이렇게 해서 우리는 먼저 구조에 기반으로하여 객체를 클러스터링 하고, 두 번째 반복으로서 객체들을 다른 객체들과의 관계에 기반하여 클러스터링을 한다. 클러스터링을 수행한 후에 우리는 유사성 메트릭스를

이용하여 다른 클러스터들 사이의 숨겨진 관계를 정의하기 위해 확률에 기반한 관계 발견 메커니즘을 소개 하였다.

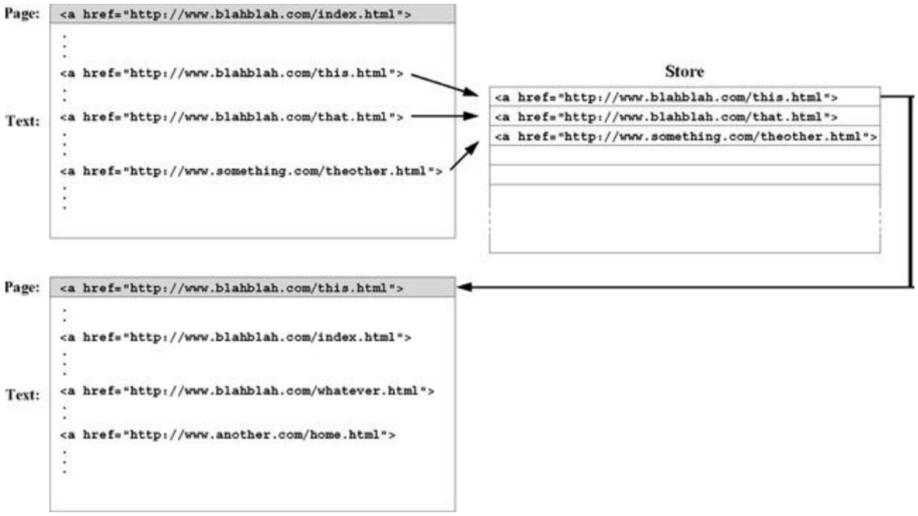
우리는 새로운 그래프를 만들었다. 이는 노드들로서의 크러스터들과, 다른 크러스터들 사이의 관계를 발견하기 위한 클러스터들 사이의 공통 관계들을 에지(edge)로 나타내는 그래프이다. 우리는 news-twitter에 우리의 기법을 적용하여 실험하였고, 그 결과로 클러스터링과 관계 발견을 위한 기법들이 기존의 k-mean과 k-medoids 메소드 보다 더 나은 성능을 보였다.

# Chapter 1 Introduction

In the past decade public appealing with the complex connectedness has been observed in various terms beginning from the social network where people try to connect themselves with each other. This connectedness leads towards the study of network analysis and mining to obtain useful information. The aim of network science is to study the behavior of real-world networked systems in order to make its data in representable format for the reader so that he can grape the information of its interest.

In general a network can be described as a system that accept the mathematical representation as a graph, whose objects can be represented as the vertexes that are connected with each other though the set of connecting links known as edges. These links represent the presence of relations or interactions among the objects. The study of Network is as old as the study of graph theory. Biological relationship, educational and social relationship can be described as examples from our daily life that can be modeled as graph or network.

Internet is one of the best known examples of the physical networks where the computers are employees as vertices and the edges are physical data connections between these vertices. Another more common example of the network is World Wide Web where web pages are described as the vertices of the network and links among those web pages are described as the edges between these vertices. These edges are known as the hyperlink WWW is an excellent example of bi-directional network in which vertices carry every type of relationships i.e. one-one, one-to-many, many-to-one and many-to-many among themselves. WWW is different from the normal network as it is a good example of a vertical network where no physical link i.e. fiber, wireless waves etc. exists between these vertices of the network.



```
<a href="http://www.blahblah.com/index.html">
:
<a href="http://www.blahblah.com/this.html">
:
<a href="http://www.blahblah.com/that.html">
:
<a href="http://www.something.com/theother.html">
:
```

```
<a href="http://www.blahblah.com/this.html">
<a href="http://www.blahblah.com/that.html">
<a href="http://www.something.com/theother.html">
```

```
<a href="http://www.blahblah.com/this.html">
:
<a href="http://www.blahblah.com/index.html">
:
<a href="http://www.blahblah.com/whatever.html">
:
<a href="http://www.another.com/home.html">
:
```

**Figure 1.1: An example of World Wide Web Networked Structure**

Similar to other networks Information networks [16] are compound of different objects belonging to one or more than one type. These objects are connected with each other though different relationships. These Information Networks are used to represent the information gathered from real-world networked in its real format using networking science. [16] An Information network is very simple when its objects belong to same type. On the other side an information network can be so complex that its objects are connected with objects belong to various other domains.

A network that is built upon items of data, linked with various relationships is known as Information Networks or Network of Information. Social Network, and E-Commerce Networks, Bibliographical Networks, Citation networks and Web-Blog Networks are some worthy example of Information Networks. Recommender Networks are another examples of networks in which people are represented as the vertices and their preferences and interests are represented as the edges for those vertices. In social networking twitter is one of the best examples of Networking In which people tweets and hash tags are represented as the vertices and follower, following relationship and citation of hash-tags are edges for these vertices.

Wikipedia is one of the best examples of Information Network in which each entity is represented by an article describing its details. All entities are linked with other entities via HTML hyper link. Recently Wikipedia has started creation of entity based network that allows linking of the information items among themselves though their properties. For example a scientist may have properties or scientific interests and similarly a politician may have property of association with some political party and author may have a relationship with his books. This all can be exampled as creation of a huge international information network in which information may be searched and presented in the form of graphs.

## 1.1 Open Linked Information Network

In early of the 21st century Tim Burners Lee presented the idea of open linked data that means data from all over the world will be linked each other through the internet and common data sharing protocols. This was the first step towards the creation of an open information network that would be in the access of every one. DBPedia is a project that converts data from Wikipedia to a semantic information network that can be accessed through endpoints using SPARQL query language.

## 1.2 Homogeneous vs. Heterogeneous Information Networks

An Information Network consists of objects and vertex that represents links i.e. relations among these objects. An Information Network that consists of one type of objects is called homogeneous Information networks whereas network is called heterogeneous information network. In chapter 3 this work presents the formal definitions for Homogeneous and Heterogeneous Information Networks. World Wide Web can be considered as an example of homogenous Network in which webpages are linked with each other Bibliographical Information Network; Citation Networks are examples of heterogeneous Information Networks. Social network is an example of very simple heterogeneous Information network in which people are not only connected with people but they are connected with groups and communities as well. Traditional Social Network has two type objects, people and groups or communities.

## 1.3 Example of Recommender Networks

A Recommender Network is an excellent example of heterogeneous Information Network in which people works as central entity that has edges with the different type of objects based on interested and is used to represent the preferences for things for the people, such as for certain products sold by a retailer or a book read by certain type of people or a movie watched by certain group of people. Now a days many organization keep records of the sold and widely used objects in their database in order to keep a custom centered approach.

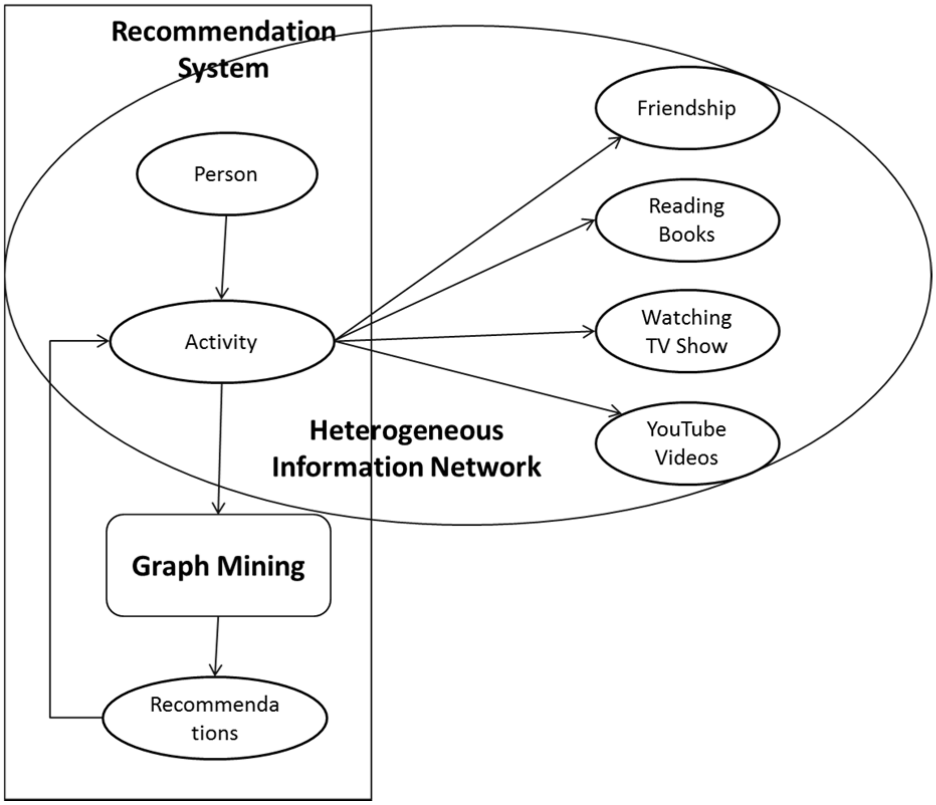


**Figure 1.2: An Example of Heterogeneous Information Network as part of Recommendation System**

The basic representation of a recommender network is as a "bipartite network," where vertex is of two type's people and items they consumes with edges connecting people to the items where all items are similar type e.g. books. However it becomes heterogeneous network from the bipartite network when items in the network are no more of one type but they are of multiple types. These types of Networks can be used to classify people with the different nature and are most useful in the supermarket where different types of objects are need to be coupled with each other.

## 1.4 Example of Twitter Information Network

The detailed architecture of twitter information have been discussed in chapter 4, here this section just briefly introduces the basics for twitter information network. The vertices of a twitter Information network are composites of users that are follower, followings, tweets, hash-tags and attached objects. As object or vertices are of multiple types therefore it can be classified as a heterogeneous information network. These objects are connected with each other through relationships follow, following, tweets, retweet, and has-link etc. These relationships can be classified as the set of edges in the network.

## 1.5 Problem Statement

Research in graph data mining has obtained a great attention from data mining research community because largely growing amount of graph data in various domain. One major intention behind study of graph mining is to understand the overall relationship among the objects of graph. It is easy when there are few hundreds number of vertices but this task becomes challenging when the number of vertices of graph that represents the objects acceded from few thousands and if belong to various types i.e. more than one type the problem become more challenging. One solution for this problem is to first partition the graph using some existing clustering algorithm before finding the relationship among objects.

However, this action requires maintenance of the original structure of graph after partition of graphs.

Recently, many techniques for mining graph data has been proposed however, there are two major limitations of those approaches 1) they consider only structure of graph for graph mining 2) they consider only values given on the edges i.e. ardency matrix in order to find similarity between different objects. Therefore these techniques cannot be used for the information network as in information networks nodes are not connected with each other with the weights only but they have some specific non-numeric properties that are listed on the edges of the graph / network. These limitations of current algorithms limit the objectives of the mining task in a way that each network is needed to be modeled in such a way that vertices must be connected with each other using a single property and on the edges there must be numeric values as weights. Modelling network as described way is unrealistic because of the reason that mining objects in graph with their original values have a different impact on the results from the impact of the results that is obtained as the result of mining network having mapped values. Furthermore current clustering techniques do not use the original structure for structure mining but they exploit the text based clustering by identifying the common number of tags particularly when it is considered about the XML documents that is a major stock holder of these techniques when it is talked about the Graph Mining techniques.

When the clusters are created by encoding the values for edges or by using textual techniques it does not remain possible to create relationship among different clusters. Therefore, an overall picture after the clustering cannot be views that is very important in many of cases to understand the overall scenario. It is because that the original edges are not maintained during the process of clustering.

This all discussion leads towards the need for development of method for clustering objects in a heterogeneous information network based on the original relationships that the objects of a network have with each other in order to get better clustering results. So, that the original relationships of the data are preserved and new relationships can be discovered.

### 1.6 Analyzing Heterogeneous Information Networks

Information networks can be analyzed at two different levels, structure level and data levels. In structure level, objects are analyzed based on their relationship with other objects, where at data level objects are analyzed based on specific values for relations [13, 16, and 14]. However, analyzing them separately does not generate as much useful results as their combination can produce [13]. In this work proposes an algorithm for analyzing objects in heterogeneous information networks by combining structure level and object level analytic techniques. This thesis proposes an extension of the fuzzy c-mean algorithm for automatic extraction of data relationships from multi-dimensional datasets. The proposed technique is based on following two steps.

1. Firstly the objects are clustered based on their type that is identified using relationship that these objects have with other objects in a heterogeneous

information network. In each cluster, objects with the similar relations are organized. As the result of process the objects belonging to similar type are placed in one cluster.

2. In the second step, the clustering technique is applied on each cluster obtained in result of step 1. However, this time objects are clustered based on the values for the relationship they have.
3. After clustering of objects in their respective clusters, in the third step, the process is defined for discovering relationship among those clusters. These new relationship can play an important role for understanding of the overall picture of the data after the clustering.

This theism defines the clustering similarity function in term of ratio of common attributes for the structure level clustering. It has been observed that the proposed function behaves differently from the statistical algorithms and it has produced better results from the present structure clustering algorithms.

Performing clustering on two different levels has its own benefits as well. From these benefit the worthy mentioning is its facility for distributed computing that is very important in term of computational and processing point of view particularly when it is talked about large datasets as the clustering of those dataset is not possible on one machine at one time. Identifying graph structure first and portioning the graph using its structure play an important role in such scenarios.

The benefits of our approach are twofold. Our proposed method is more efficient when the information about the network structure is either hidden or cannot be analyzed manually. Secondly the relations that exist between objects still remain same after the clustering algorithm is applied as the proposed technique does not convert heterogeneous networks into homogeneous networks but it treat them as a heterogeneous networks at all stages of entire process of mining. Furthermore, the proposed technique uses fuzzy membership function in order to create soft and overlapping clusters so that objects with multiple relationships can be classified in multiple clusters.

Our technique can be extended for different domains including analyzing social networking graphs, open link data networks, information networks structure mining, graph and ontology matching and all other graph matching where object of a graph belong to more than one category i.e. Heterogeneous. Because of its applicability in various domains this thesis claims that our technique is applicable in graph based applications to understand the relationship between vertices of a graph.

The proposed technique has been evaluated on two different graphs datasets. First graph dataset was constructed for an agriculture information network with the different objects and relationship among them. For this random dataset of about 1 million vertices and 5 million edges was created. The purpose of this was to examine the stability of proposed algorithm on the large scale dataset. The results discover that more than 98% of the objects were clustered properly in their respective clusters and only 2% were wrongly clustered.

Secondly twitter dataset about 5000 tweets that contain links for the news items has been used for experiments. The news article objects have been clustered based on the tweets and discovered the commonality among those clusters in order to evaluate the performance of our proposed approach.

## 1.7 Applications of Proposed Research

The proposed research is applicable and can be extended in various ways. Proposed method can be used in analyzing the user's behavior in social networking sites, clustering objects for advertisement recommendations, disease and clinical information systems. One of main stream research area of semantic computing and semantic web is ontology alignment and is one of the major research areas, proposed clustering technique can also be extended for ontology alignment and matching.

## 1.8 Document Structure

The document is organized as follows

**Chapter 2: Data Mining:** explains basic concept of data mining. It also presents a survey for general data mining techniques in general and clustering in particular for graph and non-graph data that have been considered in the research area to date.

**Chapter 3: Using Clustering for Discovering Relations p**resents the proposed mechanism for 1) clustering objects in heterogeneous information network based on their relationship and 2) discovering relationships among created clusters in order to make overall understandability of the data easy for the users.

**Chapter 4: Case Study of tweeter-news information network** draw a conceptual model for heterogeneous information model for the tweets-news network, presenting a real life application of our proposed technique in order to find the highly credible news articles based on the tweets hash-tags and user credibility,

**Chapter 5: Experiments and Results** presents the experiment results on two different datasets, 1) agriculture information network and 2) tweets-news heterogeneous information network in order to demonstrate the performance of proposed approach.

**Chapter 6: Conclusions** presents a brief summary of the dissertation and draws conclusions.

# Chapter 2    Data Mining

Endeavoring for extracting patterns from the data is as old as it is ubiquitous, has witnessed the research in the methodologies for identifying patterns though out the years.  Weather it was tried to model harvest evolution, performing sales analysis, finding publications related to our research topic, searching news articles on internet, analyzing thinking of the human being or doing diagnoses analysis it is all about that the ultimate object is to reach on the conclusion that the ultimate goal is to identify hidden patterns from the huge amount of raw data.

Traditional approaches that were used for extraction of knowledge from the data were strongly relied on the human understanding of the data, human capability of finding knowledge and human's analysis of data and human interpretation of extracted knowledge from that raw data. These techniques were slow, applicable on small amount of data, and strongly used to reply on the experience of the person who was extracting knowledge from the data along with the large amount of chances for human errors.

As the traditional approaches are applicable for a small amount of data, the rapidly growing volume of data has made these approaches irrelevant and old for the data analysis, Infect when it is said that these techniques are no more appropriate for the knowledge extraction from the data it will not be wrong.  Within the rapidly growing volume of data in science, health and business, traditional approaches are far away to find hidden patterns from these huge volume datasets. The development of information technology particularly the usage of internet has more make difficult for the traditional apaches to extract knowledge from the billions of web pages, billions of tweets and billions of Facebook messages. More than one million new are published daily on different web pages by different news agency it is impossible for the user to get list of all news published on one topic. Similarly thousands of tweets are posted on tweeter in a minute that made a impossible for a human skills to identify the tweets of his or her interests. From the thousands of the patients in the hospital it is very difficult to treat all the patients in a difficult and small amount of time. From thousands of research publications it is difficult to find the best publications or research work that were published or done in a particular research area. All these problems lead towards the need for the computational technique that allow finding of knowledge within the sea of raw data that is not possible for the human beings.

Data Mining refers to the computational techniques and methodologies that are used in order to extract knowledge from the large scale data and information. Therefore data mining can also be said as "Knowledge Mining from the Data". From the statically point for view data mining might be considered s the summering the data however data mining goes beyond the concept of summarizing or data cleansing. Formal definition of data mining is given below

> **Definition:** *Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.*

Though the definition presented above used the word database i.e. data mining is specified for data stored in the relational database however it is no more true because of huge amount of research that is going in the domain of link discovery, structure mining of internet and mining from the graphical data. The term *process* shows that data mining is a complex activity, comprised of several steps, while *non-trivial* implies that some search or inference is necessary, every time the straightforwardly extraction of the hidden the patterns is not possible.

## 2.1 Clustering

The process of grouping the objects of the similar type sharing similar properties is called clustering and created group is called cluster. The process of clustering put the object sharing similar properties i.e. of similar type into one cluster and different object into different clusters. Clustering overcome the problem classification of knowing class labels before processing therefore it is useful when data objects are not known. Clustering can be used as classification when it first divide data objects into different groups and then assign lavels to those groups. Adoptability of the groups or cluster is one of the major advantages that clustering has over the classification.

Clustering different objects is a very basic human activity that human learns in the early childhood in order to find difference between different things i.e. between male and female, between humans and animals, between book and notebooks and similarly between different colors, and different toys, and so on. In data mining cluster analysis has been widely used in business analysis, market analysis pattern recognition, data analysis, and image processing the benfit of the using clustering over the classification is that it not only construct different groups but using clustering sparse and dense regions can also be found. This sparse and dense region has been widely utilized in ad-hoc networking where a routing protocol can decide the time interval for broadcasting the routing packets. Broadcasting routing information after long interval in a dense traffic mode is more useful and save a lot of bandwidth as well. Similarly vehicle clustering allows finding the sparse and dense road in order to find more optimal path and also help in creating a balancing in the traffic on the roads. In business cluster helps identifying the custermer group and to analyze that what type of customers is more and what type of clusters are less. Similarly clustering I social media plays in important role where it help in gathering people from the same mindset, same interests and performing similar activates. On the World Wide Web clustering helps in clustering the documents of similar topic, addressing the same issue, read by same group of users in order to perform an efficient indexing for the documents. In twitter tweet clustering is an extremely interesting task that is if based on genre or hash-tags it give many interesting results to the researchers in order to understand the recent trends that are going on in the world and it also help in identifying that how people belonging to a specific region specific area think on a particular issue.

Clustering is also known as data segmentation in some applications because clustering partitions large data sets into small groups according to their *similarity*. This process it use in performing the indexing in large datasets or in creation of distributed databases where information of same type are stored on one machine so that it can be accessed very easily. The important use of clustering is finding the exception or outliers in the data that helps in identifying the fraud and a criminal activity. As a data mining function, cluster analysis can be used as a stand-alone tool to obtain an overall understanding about the distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis. Alternatively, it may serve as a preprocessing step for other algorithms, such as characterization, attribute subset selection, and even classification where after clusters creation different class labels can be assigned to them.

## 2.2 Clustering Methods

In the following sections a broad understanding of the clustering algorithm has been presented that has been followed by their limitation for the Heterogeneous Information Networks.

### 2.2.1 Partitioning methods

Given a dataset of objects Partitioning is used to partition it into *k* different partitions based on the different properties and similarities. This type of method is used in creation of distributed datasets and performing indexing on huge data in order to ensure that it can be accessed quickly by the users. Partitioning methods have two basic properties that each partition must have one membered  object and each object must be partitioned only in one group i.e. overlapping of groups or partitions is not allowed. This type of clustering is also known as hard clustering. The limitation of this approach is that it requires overall picture of data. The *k-means* algorithm, where each cluster is represented by the mean value of the objects in the cluster, and the *k-medoids* algorithm, where each cluster is represented by one of the objects located near the center of the cluster are two most widely used method for partitioning data.

### 2.2.2 Hierarchical methods

Hieratical clustering methods are used to create or find tree bases structure from the data. Top down and bottom up are two different techniques that are used in the hieratical clustering. In top down approach the data is first divided into all possible sub-clusters and then from each clusters and those clusters are then merged with each other. Where in bottom up approach items are firstly divided into one cluster and then that particular cluster are split into different sub clusters.

General Hierarchal clustering faces the problem for decision about when the criteria for splitting and merging i.e. when the clusters should be split and when they should be merged. In order to overcome this problem Clustering Feature (CF) and Clustering Featuring TREE (CF Tree) was introduce. CF and CF Tree helps in overcoming the issue of scalability and it improve the speed of clustering. CF works as following

CF is useful only for the numeric values where it first identified the center for each cluster followed by calculation of mean wise radius for each cluster centroids. Cluster centroid is calculated by following equation.

$$x_0 = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \textbf{Eq. 2.1}$$

Each cluster centroid has a specific radius that is used for calculation of membership value for a value that either a value should be part of a particular cluster or not.

$$R = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x_0)}{n}} \qquad \textbf{Eq. 2.2}$$

Similarly the value of pairwise distence is computed using the following formula

$$D = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)}{n(n-1)}} \qquad \textbf{Eq. 2.3}$$

Both R and D are used to identify the tightness of the cluster around the cluster centroid. Clustering Feature time complexity is $O(n)$ where n is the number of objects. Therefore CF Tree is fast in processing however because of its limited capacity for each cluster and most of the time it does not provides what user wants. The reason for this is that it uses radius for controlling the objects of the clusters.

## 2.3 Graph Mining

Modeling complex problem using graph has become very frequent. An Introduction to Graphs has been discussed in chapter 1 during discussion on networks. Images, chemical compounds, protein structures, biological networks, social networks, the Web, workflows, and XML documents are some of the examples of those problems that can be represented in term of graph. Along with the research n the searching algorithm for Graphs mining has also become an active field of research become rapidly growing amount of graph data on internet particularly in World Wide Web, social media websites and medial domain. Along with the example of social networks another close example for semi structured graph data is XML datasets. XML data is represented in graphical format in which the attributes along with their values are repented as nodes and relationship among those attributes is represented as the edges.

Discovering the *frequent substructures* among the graphs is one of the basic pattern discoveries that can be done in the graphs. Identifying the frequent substructure of the graphs helps in identifying the clusters, building graph's indexing and providing facility in the searching techniques for searching elements from the graph databases/datasets.

The domain of Graph mining is currently under research and extremely diverse in term of its nature. There exists many opportunities in mining graph data. Many already propose techniques for mining of social network data and clinical data needs to be taken into account in order to extend them for the general use in mining graphical datasets.

## 2.4 Clustering Graphical Data

The problem of creating clusters in graphical data arises in two different scenarios. In this subsection first the problem of clustering object in the Graph data has been defined followed by the discussion about some of the existing approaches

The problem of clustering is defined as follows: "For a given set of objects, different objects of are supposed to be divided into groups such as objects in each groups are *similar objects*". The similarity between objects is usually defined by using of a mathematical objective function. The problem of clustering is very useful in many of application as explained in section 2.3. The following section explains two methodologies from the many frequent used methodologies for Graph Clustering. Graph clustering is applicable and very useful in many applications domains.

### 2.4.1 Node Clustering Algorithms

Node Clustering algorithms are widely used in order to cluster multi-domain graph data by defining the distances of multi-dimensional data points. In graph data the values presented on the edges of nodes i.e. associated with the objects that depict the relationship strengths between nodes are mainly taken into account while performing object clustering. Therefore it is desired to petition the graph in such a way that the weights on the edges become minimum. This problem is also known as minimization cut problem.

In minimization cut problem is that a graph is partitioned in such a ways that if a graph $G$ with the node set $N$ is partitioned into two different graphs $G_1$ and $G_2$ such that for the set of edges $E$ one end of $e \in E$ should be lies in $G_1$ and other should be lies in $G_2$. This can be achieved by minimizing the function $\sum_{(i,j)\in\{G_1 \cup G_2\}} u_{ij}$. Multi-way Graph Partitioning is an extension of minimize cut problem in which the graph is divided in more than two sub graph in such a way so that the total weights between different partitions is minimized.

k-mean and k-medoids are two very famous algorithms for clustering data points based on statistical means and defined number of k seeds. Based on characteristics of k-mean and k-medoids an algorithm for network structure was presented that differ with the basic algorithm in terms of objective functions. An object function is used to find the distence between two objects in k-mean and k-medoid algorithm. Similar to k-medoids algorithm in the start number of seeds are selected randomly and based on those seeds clusters are created. For modified k-means algorithm a *local closeness centrality* based technique is used to find the cluster centroids. The challenge in this method of using k-mean and k- medoids is that finding distence between objects is a challenging task that has not been addressed yet in a proper terms.

Spectral clustering makes use of similarity matrix and statistical operation upon that matrix to create the clusters of different datasets. Similarity matrix is provided as the input to the algorithm on which it performs the statically computation and eventually clusters containing the similar nodes are found. Similarity matrix represents the node-to-node adjacency matrix. If $A_{ij}$ is an adjacency matrix for $m \times n$ nodex $(i, j)$ represents the value presented on the edge between node $i$ and $j$. The similarity value is then denoted by $w_{ij}$ and corresponding matrix is then denoted by $W$. However the problem is that this technique is god for homogenous graphs in which the entities are of one type are connected with same type object though a uni-relational way that can be Boolean on weighted.

### 2.4.2 Clustering Graphs as Objects

In this subsection, the clustering problem of an entire graph in a multi graph dataset has been discussed. This scenario usually occurs in the xml documents where the full document is represented as graph including its structure and data. FOAF friend of friend ontology is another example of graph of graphs in which graph of a person is entirely link with the graph of other persons. Most of the algorithms that are used for clustering of the objects uses the similarity matrix therefore there exists a need for creation of an mechanism that appropriate use these measurement functions for the clustering of a graphs. In the following, some of the approaches presented in order to cluster the entire graphs have been briefly explained along with the deficiencies they have. Many of the known approaches for the clustering uses the cluster centroids that can be measured using statistical means and mediums however for the graphs determination of these values is a challenge as they cannot be computed easily. There are two major approaches in the conventional data mining techniques that have been used in clustering graphs objects.

**Structure distence based approach** was used to compare the XML documents by comparing their structure. In this this approach structural distence between different objects are calculated and then is compare with each other. *XClust algorithm* is a clustering algorithm for clustering the XML document that undertake the herarical clustering algorithm and works on the basis of DTD schema of XML documents in order to efficiently cluster documents with the similar schema. The problem with these agorithms is they use text matching for the clustering of the document instead of using the original graphical structure and data set therefore these algorithms have more interaction towards text clustering algorithms rather than document clustering algorithm.

**Structural Summary Based Approach:** was proposed to first summarizing the document and then clustering these documents or objects. Though this idea also seems interactive however summarizing the graphs or documents are itself another challenge. Some approaches of this categories uses tree based comparison, in that particular scenarios tree structure is firstly created for the document and then it is compare with each other.

From this short survey this can be analyzed that graph clustering techniques are highly immature yet and therefore there exists a need for the improvement in the graph clustering techniques particularly when it is graph of heterogeneous objects known as heterogeneous information network. In the next section a very brief and summarized overview of using clustering in different homogenous networks applications have been presented and in chapter 3 of this thesis introduces proposed technique for clustering graph based objects from a huge graph based dataset, database or data warehouse.

## 2.5 Clustering the Homogeneous Networks

With the increasing amount of data on the internet and methodologies was data sharing, most of the data shared on World Wide Web can be formed as heterogeneous information network. Beyond this sensor network also create heterogeneous information networks. Research in information networks and graphs clustering has been studied widely in recent years. Clustering is considered as one of the most efficient way for summarizing data and information. The purpose of clustering is creating blocks of information based on its different but particular characteristics and gathering spares objects holding same characteristics [1].

Hierarchical clustering has been studied particularly for image processing domain [2] to find the hierarchical structure and those patterns that are hidden in the images. Other than image processing it has also been studied in organizing the statistical and textual data when data aggregation is needed. However this is our first approach to apply hierarchical clustering on graph of heterogeneous objects. In some recent works done on clustering Zheng et.al [3] proposed a Hierarchical Ensemble Clustering for tree structured data using top-down approach. However their proposed algorithm works for the simple data type objects and it doesn't define the way how to cluster the object belonging to different domains. In [19] Serban et.al proposed Hierarchical Core Based Incremental Clustering (HCBIC) clustering technique by pertaining the large objects in homogenous network in short objects when set of attributes increase from a specific length. In our views this is not a good approach for dealing with the objects because dividing the objects into different components can cause loss its real meaning also it is too hard to define which set of attributes will be combined which set of attributes. In [10] Zeng et.al applied hierarchical clustering for filtering of topics to provide a meaningful topic description. The algorithm uses a top-down approach in order to extract subtopics and arrangements of relation among different topics in neighbor levels based on common documents number.

Different methods for assignment of cluster membership have been proposed in order to classify the objects in different clusters. Fuzzy C-mean is one of those methods. In soft clustering fuzzy logic has also been widely studied an applied in various domains [11]. These domains include homogeneous clustering [15] and classifications, network data analysis, medical image segmentation and segmentation of brain MRI. The key advantage of fuzzy logic is that is very fast and required low processing power while it gives very efficient results. In the following it has been explained that how fuzzy have been applied as membership function in creation of homogeneous clusters various domains.  In [8] Deng et.al presented text clustering technique based on fuzzy c-mean by modifying the similarity measurement for the calculating distence for assignment of clusters to objects. This work is similar to our approach but on homogeneous network.  In [9] Szabo et.al presented FaiNet an algorithm for clustering using Artificial Immune Network. Authors used fuzzy logic as membership function where AIN was used as an algorithm for formation of cluster using iterative way. In order to enhance performance of network intrusion detection Ceccarelli et.al [12] present a framework based on fuzzy C-mean Clustering Algorithm in order to perform semi supervised clustering for biological datasets. From the successful results of this work semi supervision for clustering the information network has also been applied and as expected improvements in the results have been obtained from unsupervised clustering.

Another method for clustering homogenous information networks has been proposed in [18] in order to find outliers from in the networks. This technique is also very similar to our technique in which a neighborhood based analysis has been used to find the outliers based on structure of data. Although all of these methods are effective in their respective domains i.e. homogeneous networks therefore there still exists limitation of understanding the data structure in order to work with the dataset with unknown attributes. Particularly in heterogeneous networks where objects are of various types and are connected with different objects.

# Chapter 3 ClusReD: Clustering based Relation Discovery Mechanism

Information in real life is organized in hierarchal way and is described as summarized, semi-detailed and detailed formats [16]. Hierarchal clustering can be used to understand un-seen information system automatically [4]. The algorithm is called hierarchal because it creates hierarchy of clusters for a given information network. There are several new challenges while working with the heterogeneous information networks particularly when information schema is not known, or new set of heterogeneous information network is needed to be created using some XML documents or CSV files. This thesis claims that hierarchal clustering not only helps in decision making but also can be used for understanding the information schema as well.

## 3.1 Example of agriculture information network

In this chapter agriculture information network has been used in order to make proposed technique more illustrated. In this subsection the basics of an agriculture information network have been explained.

An agriculture information network is compound of objects containing rich information about crops, soil, crops, pests, herbs and fertilizer. Figure 3.1 shows the relationship between these objects. A crop is main object in the information network and it is connected with those herbs and pests, which affect its productivity, as well as soil and fertilizers that are utilizes in order to increase the productivity of the crop. In this information network effect ends with negative outcomes and utilization ends with positive outcomes.

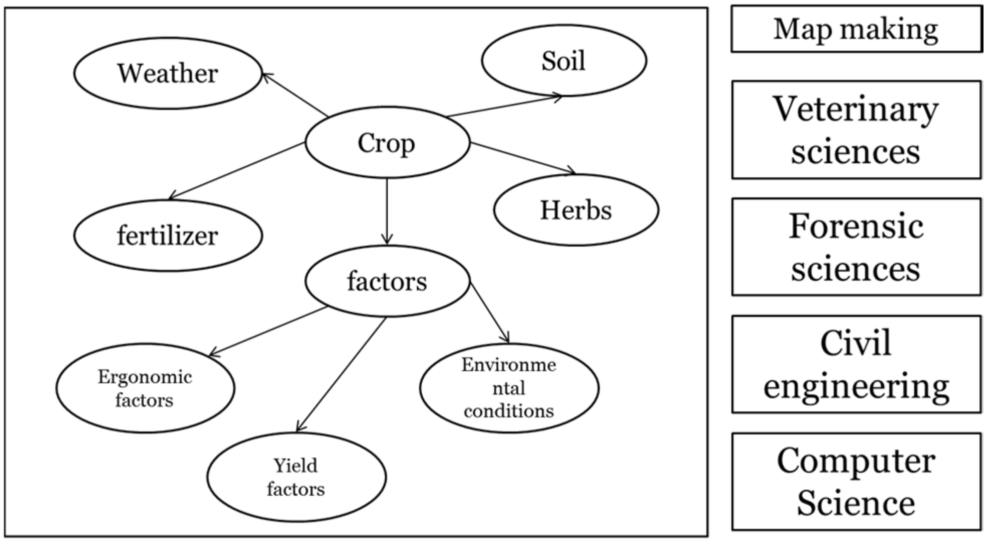


**Figure 3.1: agriculture information network**

## 3.2 Clustering Heterogeneous Information Networks using Fuzzy C-Mean

In a Heterogeneous Information Network, object types are not supposed to have flat structure but they are organized in a hierarchal way. Therefore the clusters should be created in a hierarchal way. Consider the example of an agriculture information network where an object can have an object type “crop” and this crop can be rice crop, wheat crop, cotton crop etc. and so on. When it is talked about large networks belonging to various domains it is not necessary that all objects types in the network are well known that leads towards the need of hieratical clustering. Rest of this section explains an idea for building hieratical clusters heterogeneous information network.

Here firstly the concepts related to heterogeneous information network have been defined in order to formalize the concept of clustering.

**Definition 1:** *Information Network: Given a set of atomic types* $T = \{t_1, t_2, t_3, t_4 \ldots t_n\}$, *set of objects* $\bar{\tilde{O}} = \{O^T_{t=1}\}$ *where* $O_i$*is set of objects belonging to type* $t_i$ *and set of relations* $\mathcal{R} = \{r_1, r_2, r_3, r_4 \ldots r_n\}$, *a Description Graph* $G = (V, E)$ *is called an* ***information network*** *for* $\bar{\tilde{O}}$ *if* $V \in \bar{\tilde{O}}$ *and* $E$ *is a semantic relation on* $V$ *and* $E \in \{V \times \mathcal{R} \times V\} \cup \{V \times \mathcal{R} \times I\}$ *where* $I$ *belongs to class of literal values i.e. data type values.*

Let $G = (V, E)$ a simple information network, a graph $G = (V, E, W)$ is defined as a weighted information graph such as $E = \{ e \in E \wedge e = (u, r, v, w) : u, v \in V\}$. Weight is used in an edge to define the importance of connection among two connections and is a bi-side property. Using weight the strength of relationship that exists between two different objects have been analyzed i.e. how much these two objects are necessary for each other. Consider an example of *crop* and *seed* in an *agriculture information network.* Both are necessary for each other for example there cannot be a crop without seed, and there can be no seed that can't grow and become a crop. In this case, weight on edge that connect crop with seed should be high with respect to other edges that go out from crop or come in to crop.

**Definition 2:** *Similar Objects: Given objects $O_i$ and $O_j$connected with set of objects $\{(R_i, U_i)\}$ and $\{(R_j, U_j)\}$ respectively, where $R_i, R_j \subseteq \mathcal{R}$ and $U_i, U_j \subseteq \bar{\tilde{\mathrm{O}}}$, $O_i$ and $O_j$ are said to be similar objects if and only if there exists direct mapping $U_i \rightarrow U_j$ and $\left(\forall\, x \in U_i, y \in U_j \exists\, T(x) = T(y)\right) \wedge \left(\forall\, a \in R_i, b \in R_j \exists\, a = b\right)$*

**Definition 3:** *Equal Objects: Given objects $O_i$ and $O_j$connected with set of objects $\{(R_i, U_i)\}$ and $\{(R_j, U_j)\}$ respectively, where $R_i, R_j \subseteq \mathcal{R}$ and $U_i, U_j \subseteq \bar{\tilde{\mathrm{O}}}$, $O_i$ and $O_j$ are said to be equal objects if and only if there exists direct mapping $U_i \rightarrow U_j$and $\left(\forall\, x \in U_i, y \in U_j \exists\, x = y\right) \wedge \left(\forall\, a \in R_i, b \in R_j \exists\, a = b\right)$*

### 3.2.1 Calculating Similarity between two objects

The similarity of two objects depends on two major characteristics, one characteristic from these characteristics is that two objects share similar schema i.e. they have values for similar relationships or properties. The second characteristic is that either those two objects also share the values of shared properties.

Let's O1 and O2 are two objects residing in the heterogeneous information network there are following probabilities exists

1. O1 and O2 belong to similar object type.
2. O1 and O2 belong to two different object types i.e. O1 is not similar to O2.
3. O1 is associated with O2 using some relation that exists between them.
4. O1 belongs to same object type as O2 but it has also some additional attributes.

Similarity in usual is measured in term of attributed values of the objects [1]. However this type of similarity is not efficient in heterogeneous information networks where objects are belonging to different types, in particular if similarity measures are used without weighting scheme. When clustering is performed on a dataset without considering the schema the huge amount of wrong classification can take place. This is because of when the objects of two different classes share same properties, the distance function of the algorithm may treat them as single type of objects. For example objects belonging to rice, and cotton crop can be classified into same cluster In order to overcome this major issue, firstly in section 3.2.2 similarity based on those relationships that objects carry among themselves has been defined.

### 3.2.2 Finding Schema Level Similarity ($SSim$)

The schema level similarity of two objects depends on number of common relationships. Two relationships and is called common attribute if and only if it has same domain and range. For example if an attribute share textual name with some attributes but their domain or range is not same to each other, they are not same relations. Similarly if the relationship has same domain and range but not the same name, these relations are still not same.

Consider another example of $formar_1$ and $formar_2$ they both are objects of class person; however $formar_1$ might be former of rice crop and $formar_1$ might be former of cotton crop. As both objects will be sharing more common values however will be connected with different objects. Therefore these formers needs to be classified as person first then rice and cotton former respectively in a hieratical way.

*Given* $O_i$ *and* $O_j$ *are two different objects, where* $O_i$ *has connected nodes* $X = \{N_{out}(O_i)\}$ *and* $O_j$ *has values for a set of attributes* $Y = \{N_{out}(O_j)\}$. *Firstly consider a set of types for* $X_t = \{\forall x \in X \mid x = (\, r, T(v))\}$ *and* $Y_t = \{\forall y \in Y \mid y = (\, r, T(v))\}$.

Now consider the definition for Schema Level Similarity of objects in term of set operations. *For* $x \in X_t$ *and* $y \in X_t$ *they are called them equal if and only if* $(r(x) = r(y)) \wedge \left( T(v(x)) = \; T(v(y)) \right)$ Equation 1 explains the mathematical model for $SSim$

$$SSim\,(O_i, O_j) = \frac{2(|X_t \;\cap Y_t|)}{|X_t| + |Y_t|} \qquad \textbf{Eq. 3.1}$$

The maximum value for $SSim\,(O_i, O_j)$in this scenario will be 1 when the two objects $O_i$ and $O_j$ are similar, and minimum value can be 0 when two objects are disjoint.

**Example:** In an agriculture information network, Soil and Crop are two different objects that share some of properties with each other. Figure 3.2 shows the placement of crop and soil in the information network. By considering *crop as object* $X$ and *soil as object* $Y$ following sets $X_t$ and $Y_t$ can be formed

$$X_t = Crop_t = \{\, TEXT, Color, Crop - Family, Session\}$$

and

$$Y_t = Soil_t = \{TEXT, Organisms, Texture, Soil - Family\}$$

According to the Eq. 3.1 the similarity between both objects can be measured as follows

$$SSim\ (\mathrm{Crop}, \mathrm{Soil}) = \frac{2(|Crop_t\ \cap Siol_t|)}{|Crop_t| + |Soil_t|}$$

$$\therefore\ SSim\ (\mathrm{Crop}, \mathrm{Soil}) = \frac{2(1)}{4+3} = 0.286 \qquad \blacksquare$$

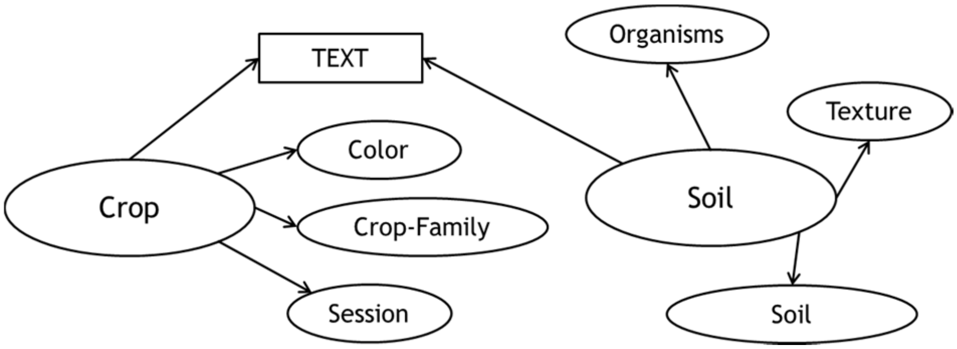


**Figure 3.2: Placement of soil and crop in an Agriculture Information Network**

When network is weighted information network $SSim$can is being found using the formula presented in Eq. 3.2

$$SSim\ (O_i, O_j) = \frac{2(\sum_{e\in(X_t\cap Y_t)} w(e))}{\sum_{e\in X_t} w(e) + \sum_{e\in Y_t} w(e)} \qquad \textbf{Eq. 3.2}$$

**Example:** Consider the above example again with the weight suppose that weight “1” is given to TEXT and rest all connection have given weight of 3. The equation 4.2 can be solved as following

$$SSim\ (Crop, Soil) = \frac{2(1)}{10+10} = 0.02 \qquad ■$$

Here, now, I define the Schema Level Difference between $O_i$ *and* $O_j$. In order to compute the difference only neighboring nodes have been taken into account and measure the difference is measure in term of set difference as shown in Eq. 3.3.

$$SDiff\ (O_i, O_j) = \frac{|X_t - Y_t| + |Y_t - X_t|}{|X_t| + |Y_t|} \qquad \textbf{Eq. 3.3}$$

Set $\{X_t - Y_t\}$ represents those object types that are connected to object $X$ but not connected to object $Y$. Similarly set $\{Y_t - X_t\}$ represents those object types that are connected with $Y$ but not connected with object $X$.

When network is weighted information network $SDiff$ can be find using formula that is explained in Eq. 3.4.

$$SDiff\ (O_i, O_j) = \frac{\sum_{e \in (X_t - Y_t)} w(e) + \sum_{e \in (Y_t - X_t)} w(e)}{\sum_{e \in X_t} w(e) + \sum_{e \in Y_t} w(e)} \qquad \textbf{Eq. 3.4}$$

### 3.2.3 Object Level Similarity ($OSim$)

Once clusters of different objects have been found the next step is to find clusters within the clusters i.e. when there is a need to find similar type of objects who share attributes with each other. In order to find the similarity between two objects irrespective of object type, for example when there is a need for finding two patients having two different diseases; it becomes necessary that original values for properties of that object should be taken into account.

Let $O_i = \{(r_i , x_i , w_i)\}$ $and$ $O_j = \{(r_i , x_i , w_i)\}$ $where$ $r_i, r_j \in \mathcal{R}$, $x_i, x_j \in O \cup I$ $and$ $w_i, w_j \in \mathbb{R}$. The Object Similarity *($OSim$)* can be defined as Eq. 3.5

$$OSim\ (O_i, O_j) = = \sum_{i=1}^{n} |x_i w_i - x_j w_j| \qquad \textbf{Eq. 3.5}$$

$$Diff\ (O_i, O_j) = \frac{1}{2}\Big(\ OSim\ (O_i, O_j) + SDiff\ (O_i, O_j)\Big) \qquad \textbf{Eq. 3.6}$$

Next, the problem of assigning weights to the attributes has been addressed. Weight assignment is really a tough task in clustering or classification of objects because clusters that are made to highly depend on the weights that are assigned to different objects in a dataset. In general the process of feature selections, it is done manually but here the already made clusters have been utilized to select the weights. Low weights have been assigned to those attributes whose values are changed frequently and high weights to those attributes that values change less frequently. If it is assumed that the weight is a value between 0 and 1, $O_r$ represents the set of objects that are linked with though relation $r$ and $O_r^t$ represents the set of object types for $O_r$ the of weight for an attribute can be defined computed using Eq. 3.7

$$\forall\, r \in \mathcal{R}: w(r) = \begin{cases} 1 & if\,(|O_r| = 1 \,\wedge\, |O_r^t| = 1 \\ \dfrac{2}{|O_r| + |O_r^t|} & elsewise \end{cases} \qquad \textbf{Eq. 3.7}$$

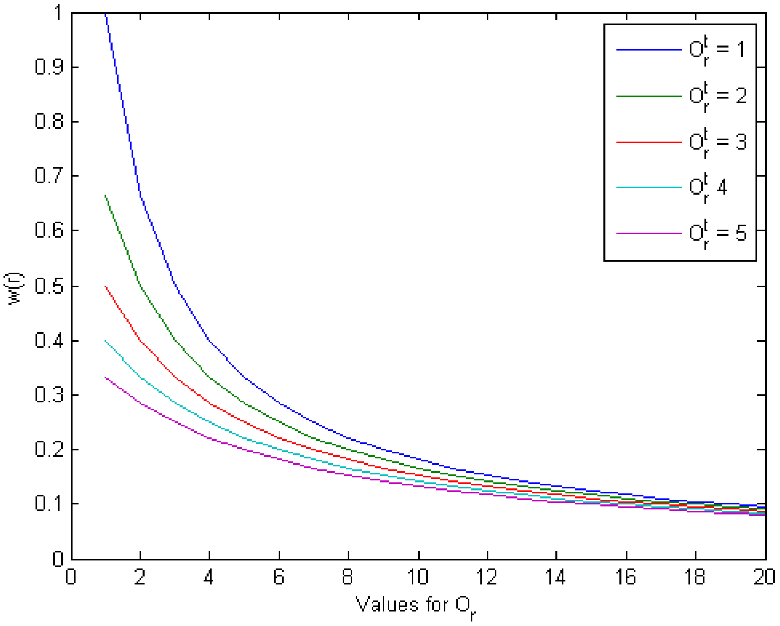


**Figure 3.3: Relationship between $|\boldsymbol{O_r}|, |\boldsymbol{O_r^t}|$ and $\boldsymbol{w(r)}$**

**Example:** consider that a crop name changes more frequently than crop type, therefore crop type have more weight. Similarly the crop session change less frequently then crop type, thus crop session has more weight. Figure 3.3 depicts the relationship between impacts of frequent changing attributes on weights for 20 crops of five different types where $\{O_r^t\}$ represents the set of crop types and $\{O_r\}$ resents the set of the crop names. ■

### 3.2.4 Construction of clusters

This section discusses two types of clustering for heterogeneous information network using fuzzy c-mean clustering algorithm. 1) Autonomous clustering in which there are no center points. This can also be called un-supervised clustering. 2) Manual Clustering in which this section firstly defines the disjoint objects by ourselves as cluster centroids. In autonomous clustering the centroids was selected randomly and then was evolved these centroids for each cluster iteratively. As the results of this iterative process, best cluster centroids are chosen for each cluster. A cluster centroid is called best centroids if it has maximum common relationships or common objects with respect to other objects that are present in the same cluster.

Here again the idea has been illustrated using an Agriculture Information Network; consider a sub-network having objects of different crops. If there are three different types of crops i.e. rice, cotton and wheat are presented in the information network. There will be three different clusters that will be formed as a result of clustering and for each cluster; those objects that will have values for more attribute will become cluster centroids. It is because of the reason that a perfect cluster centroid can form a well-organized cluster. In the following a formal definition of cluster is given.

**Definition 4:** *Information Network Cluster: Given a sub graph* $C(V^*, E^*) \subseteq G(V, E, W)$ *where* $V^* \subseteq V(G)$ *and* $E^* = \{e \in (E(G) \mid e = (u, r, v\,)\; with\; u, v \in V^*\; and\; r \in \mathcal{R}\,\}$ *can be said an* ***information network cluster*** *if and only if there exists no object* $O_i, O_j \in\; V^* \wedge\; Sim(\,O_i, O_j) <\; \tau$ *where* $\tau \in\; \mathbb{R}$

### 3.2.5 Fuzzy C-mean algorithm

In fuzzy c-mean algorithm each object is made part of some cluster based on membership function. Distance for each object from the center of each cluster is measured. As the measurement of difference between two objects has already been defined therefore here just briefly introduction of the fuzzy c-mean membership function has been presented. If C represents the set of cluster centroids, the main function for fuzzy c-mean that is needed to be minimized has been explained in Eq. 3.8 and Eq. 3.9.

$$J(\,O_i,V_j) = \sum_{j=1}^{N}\sum_{i=1}^{c}\mu_{ij}\ \ Diff(\,O_i,V_j) \qquad \textbf{Eq. 3.8}$$

$$Diff(\,O_i,V_j) = \ 1 - \left|\frac{1}{\big(SSim(O_i,V_j)\big) + \big(OSim(O_i,V_j)\big)}\right| \qquad \textbf{Eq. 3.9}$$

here $V_j$ is the cluster centroid of the cluster $j$ and belongs to set $C$ that can be computed using Eq. 9 and Eq. 3.10 and Eq. 3.11 and $c$ represents the total number of elements in set $C$

$$\mu_{ij} = \ 1\Big/\sum_{k=1}^{c}\frac{Diff(\,O_i,V_j)}{Diff(\,O_i,V_k)} \qquad \textbf{Eq. 3.10}$$

$$V_j = \frac{\sum_j^N \left(\mu_{ij} \left(\sum\{(x,w)\}_j\right)\right)}{\sum_j^N \mu_{ij}} \quad O_j = \{(r_i \,, x_i \,, w_i)\} \qquad \textbf{Eq. 3.11}$$

The difference method that has been used here in order to compute the distance between clustering centroids and the current object has been described in equation 4.6

## 3.3 Framework for classification

Once the fuzzy membership function have been defined the next step is to define the framework for clustering, by illustrating the steps that are needed to be taken into account in order to find suitable clusters for the objects.

1. In the first step autonomous clusters based relationships have been constructed. Objects with different schemas are put into different clusters. Thus one cluster contains only the object with same relationships.
2. In the next step for each cluster some objects are randomly made cluster centroids. Two different methods have been used, one was choosing centroids randomly, and other was defining the cluster centroid manually. This was to study the effect of learning on our proposed algorithm.
3. Each object present in the cluster is matched with the cluster centroids, if the distence between cluster centroid is less than threshold value the object become member of the cluster.

When cluster centroids are chosen randomly, there exists a need to know either these cluster centroids are good centroids or not. After the step 3 is completed for each object, an object is marked as best object in the cluster if and only if it comprises with condition of having maximum object with minimum distance. In other words It can also be said that the object that has more common relation will become the cluster condition.

Finally this section explains how to deal with the problem of outliers i.e. those objects that had not become part of any cluster in result of execution of above process. For this the three possibilities have been considered and all three have been processed differently. They are called non-connected-mode outliers, semi-connected outliers, and connected outliers.

1. If outliers are not connected among themselves also they are called sparse outliers.
2. If the outliers made more than one groups i.e. cluster among themselves they are called semi-connected outliers.
3. If all outliers are connected with other, they are called connected outliers.

Disconnected outliers are ignored as the high probability exists that they will be mistyped values or some noise etc. semi connected and connected outliers are technically not outliers but they have been classified as outliers because of not having their respective cluster centroid in set of cluster centroids. This scenario usually occurs with a large datasets. In order to overcome this situation a simple method in this paper has been implemented. Once an outlier group has found after running the fuzzy c-mean algorithm on original dataset, this work treats that outlier group as a separate dataset and run fuzzy c-mean algorithm separately on each dataset by defining the cluster centroid randomly. This process continuous until there exists no outlier cluster that members are more than δ a numeric value that can be differ from scenario to scenario.

### 3.4 Constructing Relationship among the clusters

Once clusters have been created from a heterogeneous network the next step is to find the relationships among those clusters in order to enhance the understandability of the data. This section explains the algorithm for defining the relationship among the clusters. Remember that algorithm does not create new relationship between objects but the relationships are already existed in the data that went hidden with the creation of clusters our algorithm will discover that relationship again. Structure level similarity measurements have been used for discovering of relationship among the clusters.

The proposed technique for discovering relationship among different heterogeneous information network's clusters is consists of two different step, In the first step each cluster – that is a sub-graph of the original graph – computed its similarity with the other clusters for edges. This allows understanding of coupling and coherence among the clusters with each other based coupling and coherence for each on the relationship – edge —. In the second step the overall similarity between clusters has been measured.

**Definition 7:** *Let* $C_1 and C_2$ *are two different clusters extracted from a graph* $G,\ \ E(C_1)$ *and* $E(C_2)$ *are set of edges for each cluster,* $R_1\big(E(C_1)\big)$ *and* $R_2\big(E(C_2)\big)$*is set of distinct relationship in cluster* $C_1$ *and* $C_2$ $fo$ *is a function that represents set of object type for each* $R_1\big(E(C_1)\big)$ *and* $R_2\big(E(C_2)\big)$*. Firstly the relationship strength between* $r_1 \in R\big(E(C_1)\big)$ *and* $r_1 \in\ R\big(E(C_2)\big)$ *has been compute separately then all values calculated for* $R_1\big(E(C_1)\big)$ *and* $R_1\big(E(C_1)\big)$ *have been added*

$$S(r_1, r_2) = \frac{n\big(f(r_1) \cap f(r_2)\big)}{n\big(f(r_1)\big) + n\big(f(r_2)\big)} \qquad \textbf{Eq. 3.12}$$

In order to measure the strength between two clusters values of all possible pairs are calculated using Equation 3.13.

$$S(C_1, C_2) = \ \forall\, r_1 \in\ R\big(E(C_1)\big), r_2 \in R\big(E(C_2)\big) \sum S(r_1, r_2) \qquad \textbf{Eq. 3.13}$$

In equation 3.13, $C_1$ and $C_2$ represent two clusters between them the relationship is being found, $R\big(E(C_1)\big)$ and $R\big(E(C_2)\big)$ represents the sets of inter-cluster relationships for $C_1$ and $C_2$ respectively.

# Chapter 4 Case Study – Twitter and News Documents

This chapter presents a case study of ranking news based on ontology of tweet' hash tags and a heterogeneous information network of tweets, hash tags, news documents and twitter users. Figure 5.3 shows an overview of our constructed heterogeneous information network.

Tweets are usually written in an informal language however the most benefited thing is hash tags and URL of the news documents. These has tags allows understanding the tweets' subject i.e. the topics it addresses therefore the links between hash tags and news documents can be built very easily.

## 4.1 Tweets Heterogeneous Information Network

Figure shows the heterogeneous information networks of tweeter's tweets. The main roles/ objects are tweets, users, tags, pictures and the web URLs. Tweeter's users generates huge amount of tweet text that are consist of huge amount of diverse genres, web URLs and pictures. These genres usually are used to recognize the topics addressed in tweet. Genres are usually created based on political, social situations or two address particular number set of users and are used to analyze the political and social trends in a particular region. An sub network can also be build based using intra-genres relationships.

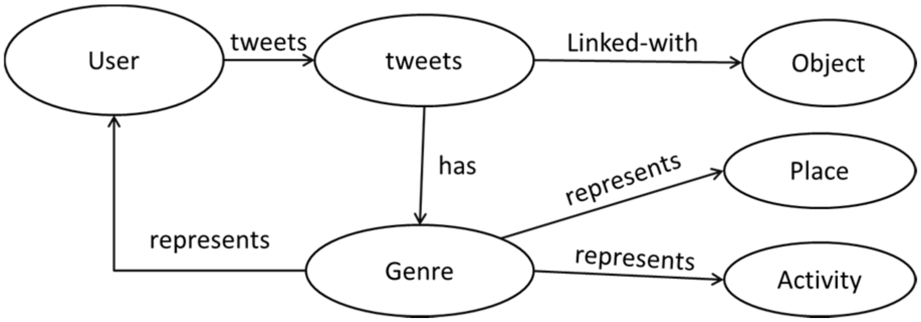


**Figure 4.1: Heterogeneous information network of Tweets**

## 4.2 News Heterogeneous Information Network

News Network is another example of heterogeneous Information Network in which "news article" is the basic object that is published by the news agencies or news website. A news article provides information about different personalities, places, incidents, celebrities, political and social activities. News articles are addressing the same topic published by different news publishing agencies which have different level of trust among the audiences. Information that can be extracted from these news articles can be converted into an heterogeneous information network in order to find most reliable news of one's interested topic. Figure 3 presents an overview of news heterogeneous information network.

## 4.3 Combining News and Tweet Networks

In sub-sections 5.1 and 5.2 the tweets and news heterogeneous information networks separately have been explained. This sub-section describes that how these two information networks can be combined in order to build a new big heterogeneous information network. Consider Figure 4 which presents the News-Tweets Heterogeneous Information that shows the relationship between tweets and news networks. By comparing figure 2, 3 and 4 it can be observed that both networks share many common attributes.

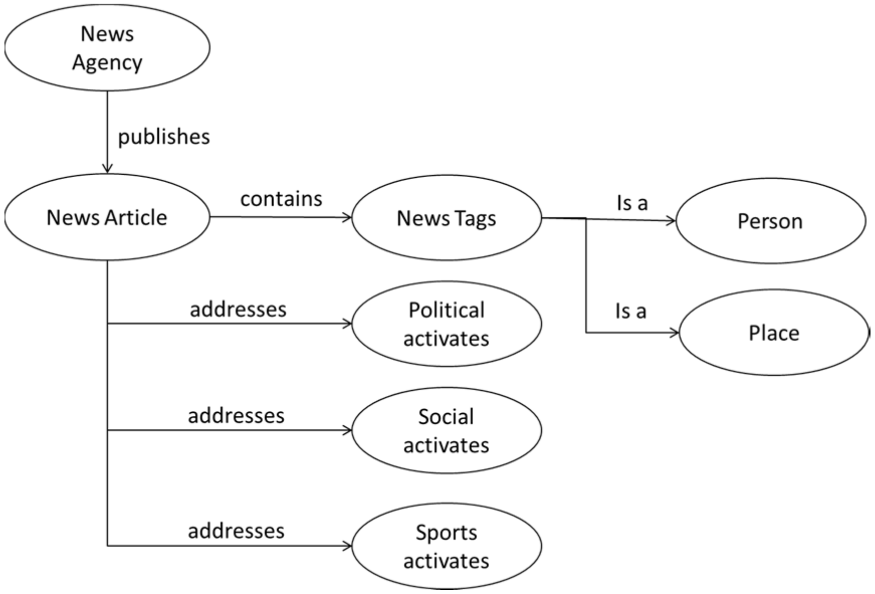


**Figure 4.2: News Heterogeneous Information Network**

As depicted in figure 4, genre in tweet represents an activity that is addressed by news articles. This activity can be divided into sub-activities political, social, sports activities and so on whereas one link of news arable can also be part of these tweets. ■

In the first step those tweets are filtered from tweets datasets that has no hash-tags and documents URL. Because the bases of our clustering documents are based on hash-tags therefore all those tweets that do not containing the hash-tags were filtered out in order to improve the quality of the clustering. Similarly all tweets that have no document URL are also filtered. Those tweets are known as informal tweets. In the second step the tweets are divided into two different categories, first category contains those tweets that has hash-tags and documents URL and second category contains those tweets that have hash-tags but not URLs. Because of simplicity of the technique 99% tweets were filtered and categorized correctly. Finally hash-tags were used to build an ontology that have been used in order to improve our clustering results.

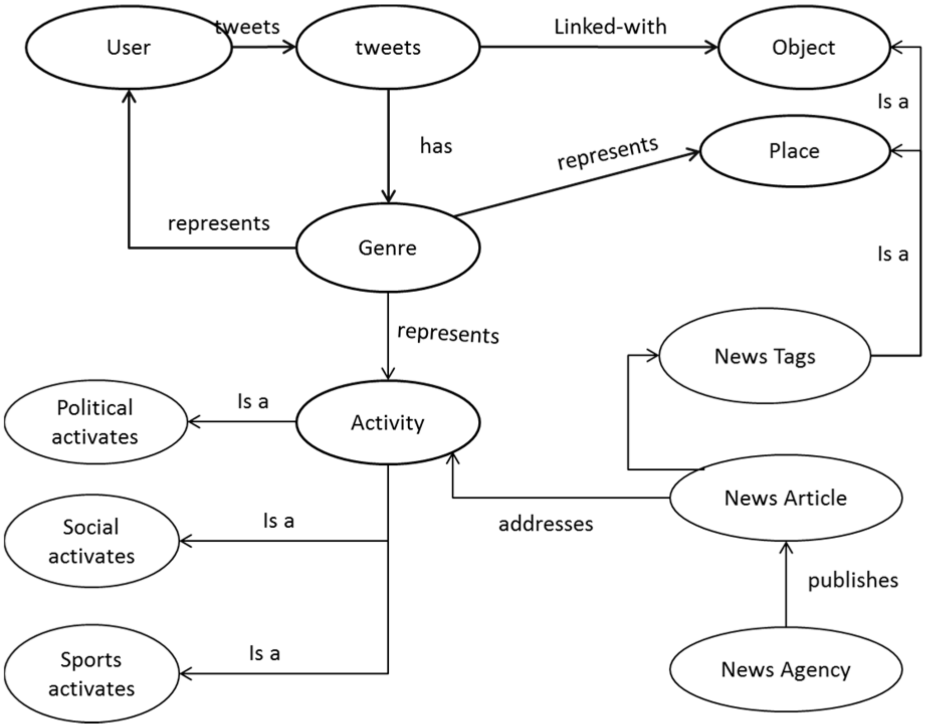


**Figure 4.3: News Tweets Heterogeneous Information Network**

Our framework for clustering of news consists of three different steps

1. Filtering the tweets and categorizing them
2. Clustering documents based on relationship with hash-tags
3. Improving the clustering based on credibility of tweets
4. Improving the clustering using the credibility of the users

## 4.4 Constructing a Heterogeneous Information Network

*Let* $T = \{t_1, t_2, t_3, \dots \dots \dots t_n\}$ *is set of tweets,* $U = \{u_1, u_2, u_3, \dots \dots \dots u_n\}$ *is set of users,* $N = \{n_1, n_2, n_3, \dots \dots \dots n_n\}$ *are set of news articles* $E$ *represents the set of edges or relationships there exists a Heterogeneous Information Network* $G = \{(T \cup U \ \cup N), E\}$ *because* $T \cap U = \emptyset$ , $N \cap U = \emptyset$ *and* $T \cap N = \emptyset$ ■

## 4.5 Simple Clustering

Similar to [14] firstly this section presents the ranking algorithm for the ranking of news articles based on tweets and re tweets of the articles. Firstly the Tweets vs. News Documents (TD) Information Network has been presented followed by the construction of clusters of different news from the graph $G_D = \{(T \cup D), E\}$ based on hash-tags in the network. This thesis claims as more as the users would have added same tags along with the sharing of the URLs, the more chances exists for the news article to become part of the cluster.

**Table 4.1: Relationship between News document and Tweets**

| $TD$ | $d_1$ | $d_2$ | ...... | $d_i$ | ...... | $d_n$ |
|---|---|---|---|---|---|---|
| $t_1$ | $x(t_1, d_1)$ | $x(t_1, d_2)$ | ...... | $x(t_1, d_i)$ | ...... | $x(t_1, d_n)$ |
| $t_2$ | $x(t_2, d_1)$ | $x(t_2, d_2)$ | ...... | $x(t_2, d_i)$ | ...... | $x(t_2, d_n)$ |
| $t_3$ | $x(t_3, d_1)$ | $x(t_3, d_2)$ | ...... | $x(t_3, d_i)$ | ...... | $x(t_3, d_n)$ |
| .... | ...... | ...... | ...... | ...... | ...... | ...... |
| $t_j$ | $x(t_j, d_1)$ | $x(t_i, d_2)$ | ...... | $x(t_j, d_i)$ | ...... | $x(t_j, d_n)$ |
|  | ...... |  | ...... | ..... | ...... | ...... |
| $t_m$ | $x(t_m, d_1)$ | $x(t_m, d_2)$ | ...... | $x(t_m, d_i)$ | ...... | $x(t_m, d_n)$ |

The value of $x\left(t_j, d_i\right)$ is 0 if news article $d_i$ has no link with $t_j$ and 1 otherwise.

**Table 4.2: Relationship between Hash-Tags and Tweet**

| $TH$ | $h_1$ | $h_2$ | . . . . . . | $h_i$ | . . . . . . | $h_n$ |
|---|---|---|---|---|---|---|
| $t_1$ | $x(t_1, h_1)$ | $x(t_1, d_2)$ | . . . . . . | $x(t_1, h_i)$ | . . . . . . | $x(t_1, h_n)$ |
| $t_2$ | $x(t_2, h_1)$ | $x(t_2, h_2)$ | . . . . . . | $x(t_2, h_i)$ | . . . . . . | $x(t_2, h_n)$ |
| $t_3$ | $x(t_3, h_1)$ | $x(t_3, h_2)$ | . . . . . . | $x(t_3, h_i)$ | . . . . . . | $x(t_3, h_n)$ |
| . . . . | . . . . . . | . . . . . . | . . . . . . | . . . . . . | . . . . . . | . . . . . . |
| $t_j$ | $x\left(t_j, h_1\right)$ | $x(t_i, h_2)$ | . . . . . . | $x\left(t_j, h_i\right)$ | . . . . . . | $x\left(t_j, h_n\right)$ |
|  | . . . . . . |  | . . . . . . | . . . . . | . . . . . . | . . . . . . |
| $t_m$ | $x(t_m, h_1)$ | $x(t_m, h_2)$ | . . . . . . | $x(t_m, h_i)$ | . . . . . . | $x(t_m, h_n)$ |

The value of $x\left(t_j, h_i\right)$ is 1 if tweet $t_j$ includes (has a link with) hash-tag $h_i$ and 0 otherwise.

Given a set of tweets $T = \{t_1, t_2, t_3, \ldots \ldots \ldots t_m\}$ and $\mathrm{D} = \{d_1, d, d_3, \ldots \ldots \ldots d_n\}$ as a set of news articles where for each tweet $t \in T$ there exist a set of hash-tags. Let $H$ represents set of all hash-tangs and $H_t \subseteq H$ represents the set of hash-tags that are associated with a tweet $t$ that has a news article $n \in N$. Now a news article $n$ can be binary clustered in all of the hash-tags $h \in H_t$ However binary clustering is not enough there exist a need to define membership function based on the relationship between $n$, $h$ and $t$. With means as more number of tweets will uses the hash-tags for a news article, the membership value will be awarded to the news-article. Equation xxx defines the formal mathematical model for the membership function $Mem(h_i, n_j)$

$$Mem(h_i, n_j) = \frac{n\left(t(h_i, n_j)\right)}{n\left(t(h_i, n_j)\right) + n\left(t(h_i, n_j)\right)} \qquad \textbf{Eq. 4.1}$$

Where $|t(h, n)|$ represents the number of tweets that has a link with (or contains) hash-tag $h$ and news $n$ ,$|t(n)|$ represents number of tweets that has a link with news $n$ and $|t(h)|$ represents the number of tweets that has link with the hash-tag $h$. ■

Once $Mem(h_i, d_j)$ is computed for all $h \in H_i$, $d_j$ is added to all those clusters where $Mem(h_i, d_j) > \alpha$ where $\alpha$ represents the minimum threshold value is required for joining a cluster. In general it can be set the value of $\alpha$ set to 0.5.

Initially for all $h \in H$, independent, disjoint and non-overlapping clusters where constricted. Therefore it can be said that at the initial level no of clusters are equal to number of hash-tags appeared in all tweets.

Using the Eq. 4.1 not only the binary membership for the news articles to some clusters can be computed but its closeness to the cluster centroids can also be computed. As the value for $Rank(h_i, d_j)$ will be increased the news article will be placed to the nearest to the cluster centroid. Therefore it can be said that the closeness of the cluster is directly proportional to the value of $Mem\left(h_i, d_j\right)$ or in words $Mem\left(h_j, d_i\right)$ represents the distence between cluster centroid and $d_j$.

## 4.6 Finding overlapping between clusters

In the previous subsection it has discussed that how a news article can be added to one or more than one clusters when clusters are disjoint from each other and are made based on hash-tags of tweets. Next task is to identify the overlapping of the hash-tags clusters i.e. for two hash-tags how many common news articles exist. This is very important to find overlapping of the cluster in order to discover relationship between news articles clustered in the different clusters.

The easiest way to find it is by identifying the overlapping of different clusters. In order to find coupling or cluster overlapping between two clusters the number of those news articles that are member of both clusters have been calculated. Consider there exists a need to find the coupling between two different clusters that are identified by hash-tags $h_i$ and $h_j$ the value for overlapping can be defined as a coupling function $Cup\left(h_i, h_j\ \right)$ and compute it a follows

$$Cup\left(h_i, h_j\right) = \frac{n\left(d(h_i) \cap d(h_j)\right)}{n\left(d(h_i)\right) + n\left(d(h_i)\right)} \qquad \textbf{Eq. 4.2}$$

And ochiai coefficient can be compute as follows

$$\text{ochiai}\left(h_i, h_j\right) = \frac{n\left(d(h_i) \cap d(h_j)\right)}{\sqrt{n\left(d\left(h_j\right)\right) \times n\left(d\left(h_j\right)\right)}} \qquad \textbf{Eq. 4.3}$$

Where $\left|\left(d(h_i) \cap d(h_j)\right)\right|$ represents the number of news that has a link with hash-tag $h_i$and, $h_j$, $|d(h_i)|$ represents number of news articles that has a link with hash tag $h_i$and $\left|d(h_j)\right|$ represents the number of number of news articles that has link with the hash-tag $h_j$.

## 4.7 Using User Credibility

In the previous section it has been explained that how the proposed algorithm can work on a tweet-news network to create clusters using heterogeneous information network in order to provide useful information to the users. The limitation of the previously explained work is each user has a similar level of credibility that is there in the real case. Each tweet and re-tweet has the similar weight without the credibility In this section it has been explained that how the weights to the users profiles have been assigned and how those weights have been integrated in order to improve our clustering results.

There exists very basic method for understanding of credibility of users in twitter one of them is using following, follower network and other one is using re-tweet. If a user has more followers it shows that user has more credibility and similarly number of users who did re-tweeted one's messages also represents the person has credibility otherwise why other people should have re-tweeted him.

Relationship described above is not layer but it is nested that ends no depth limit. Consider an example of follower-following relationship then one way to measure credibility is how man followers a person have, but another important thing that is more important is how much those follower are credible i.e. or what is the credibility of those followers. This recursive experience can be very simple and even can be endless. Similarly when a person's tweet is re-tweeted it is also important that what is the credibility of the person who has re-tweeted the once tweet.

$$\forall\, u \in U;\; C(u) = \frac{1}{2}\left(\left(1 - \frac{n\left(\overleftarrow{f}(u)\right)}{n\left(\vec{f}(u)\right)}\right) + \left(1 - \frac{n\big(t(u)\big)}{n\big(rt(u)\big)}\right)\right) \qquad \textbf{Eq. 4.4}$$

Here $\overleftarrow{f}(u)$ represents set of users followed by $u$ and $\vec{f}(u)$ represents set of users that follows $u$

For an author the more his or her tweets are retweeted the more he or she is credible similarly the more he or she is followed by other, the more he or she is credible.

The credibility of the author increases when he or she is followed by highly credible users. Using this rule equation Eq. 4.4 has been enhanced as folloiwng.

$$\forall\, u \in U;\; C(u) = \frac{1}{2}\left(\left(1 - \frac{\sum_{f \in \overleftarrow{f}(u)} c(f)}{\sum_{f \in \vec{f}(u)} c(f)}\right) + \left(1 - \frac{n\big(t(u)\big)}{n\big(rt(u)\big)}\right)\right) \qquad \textbf{Eq. 4.5}$$

here $u$ represents users, $c(u)$ represents the credibility of the user, $t(u)$ repsresents the tweets tweeted by user u and $rt(u)$ represents those tweet that are tweeted by $u$ and re-tweeted by other users as well. Notice that $t(u)$ and $rt(u)$ are global values and they representing the total number of tweets and retweets respectively.

After computing the user credibility each tweet is then assigned a weight based on credibility of its author, credibility of those who retweeted it. When a highly credible person retweets an already tweeted message its ranking increases with respect to credibility of the person who have retweeted it.

Let $\{rt(u)\}$ is the set of users who retweeted a tweet t. The overall ranking of that tweet can be computed using the following equation.

$$rank(t) = C\big(a(t)\big) + \sum_{u \in rt(u)} \left(\frac{C(u)}{2}\right) \qquad \textbf{Eq. 4.6}$$

where $a(t)$ represents the author of a tweet $t$. Once rank value for each t have been obtained the equation Eq. 4.6 can be modified for membership function as following

$$Mem\big(h_i, n_j\big) = \frac{\sum_{t \in \{t(h_i,n_j)\}} rank(t)}{\sum_{t \in \{t(h_i)\}} rank(t) + \sum_{t \in \{t(n_j)\}} rank(t)} \qquad \textbf{Eq. 4.7}$$

# Chapter 5 Experiments and Results

This chapter presents results of application of the proposed algorithm on two different datasets, agriculture information network in order to test the scalability of the clusters and on twitter-news information network in order to test the performance of the algorithm.

## 5.1 Agriculture Information Network

This section applies our proposed clustering technique on agriculture information network, a heterogeneous information network connecting 5 different object types among each other. The simple matrix in the form of successful and unsuccessful classification was used for performance measurement.

### 5.1.1 Schema Level Similarity

Firstly, this section discusses how the clustering algorithm behave while creation of clusters of objects in order to find different types of objects present in the information network. Table 1 presents the number of classification and misclassification. Obtained results can be described as good results as all more than 90% of the objects belonging to all classes were classified in their correct object type. In order to cross check the classification of the objects in different clusters a hidden object type in the dataset was added for each object. Once the clustering algorithm created all clusters using the step 1 in section 3.3 those hidden labels were used to match the accuracy of the clustered.

**Table 5.1: schema representing of Agriculture Information Network**

| Type | Nos | Attributes |
|---|---|---|
| Crop | 200000 | Name, Size, Color, Family, season |
| Soil | 20000 | Name, Family, Organisms, Texture |
| Fertilizer | 10000 | Name, Family, Soil-acidification |
| Herb | 500000 | Name, Family, Type, Color |
| Pests | 10000 | Name, Family, Season, Control-Method |

The threshold value for assigning the membership was kept 75%.Two major reasons were identified for misclassification has been noticed. First reason was missing relation types, for example all crops have relationship with soil, herbs, fertilizer, and pests. If two are more than two relationships were missed the membership function failed to assign the membership for cluster of crops' objects.

### 5.1.2 Object Level Experiments

This section presents the study for clustering of objects with in their parent clusters. For object level experiments the proposed algorithm was examined using different no of cluster centroids, and different values for membership threshold and using weight for the relations and without using weights.

In chapter 3 the clustering by combining schema and object similarity has been explained. In this section presents the study of the impact of using schema in the object clustering. Results presented in Figure Table 5.1 shows the improvement in the clustering performance suing schema and depicts the impact of using schema with object similarity on the results. It can also be seen as a linier relationship that exists among the membership threshold and ration of successful classifications.

**Table 5.2: Schema level cluster creation**

| Object Type | | Classified | | Misclassified | |
|---|---|---|---|---|---|
| | Total | No | % | No | % |
| Crop | 200000 | 181145 | 90.5725 | 18855 | 9.4275 |
| Soil | 20000 | 18586 | 92.93 | 1414 | 7.07 |
| Fertilizer | 10000 | 9667 | 96.67 | 333 | 3.33 |
| Herb | 500000 | 479621 | 95.9242 | 20379 | 4.0758 |
| Pests | 10000 | 9857 | 98.57 | 143 | 1.43 |

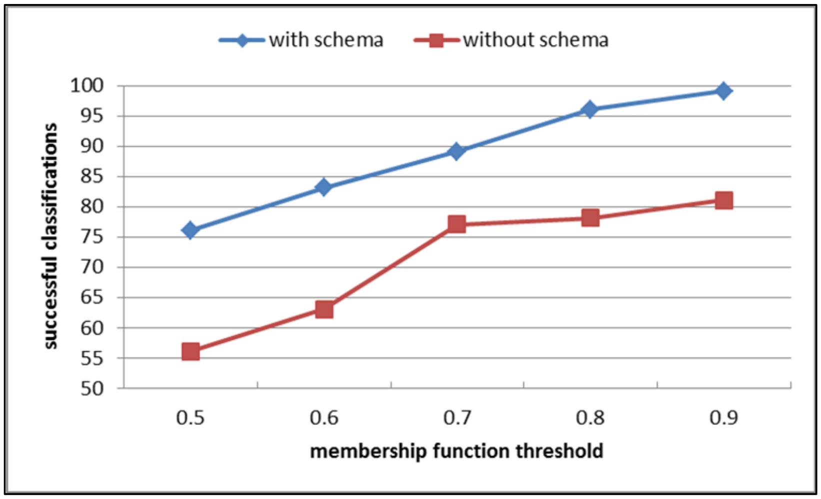


**Figure 5.1: Impact of threshold, membership function vs. accuracy of classification**

Figure 6.1 presents the classification behavior with respect to no of cluster centroids. An increasing number of cluster centroids increase the performance and optimization of the clusters. When there exists more cluster centroids this means more accurate clustering can be done because of having more membership functions. For the experiments chose 0.1 to 0.5 percent objects were chosen as cluster centroid. from the total objects as the cluster centroids randomly. However these cluster centroids was updated iteratively.

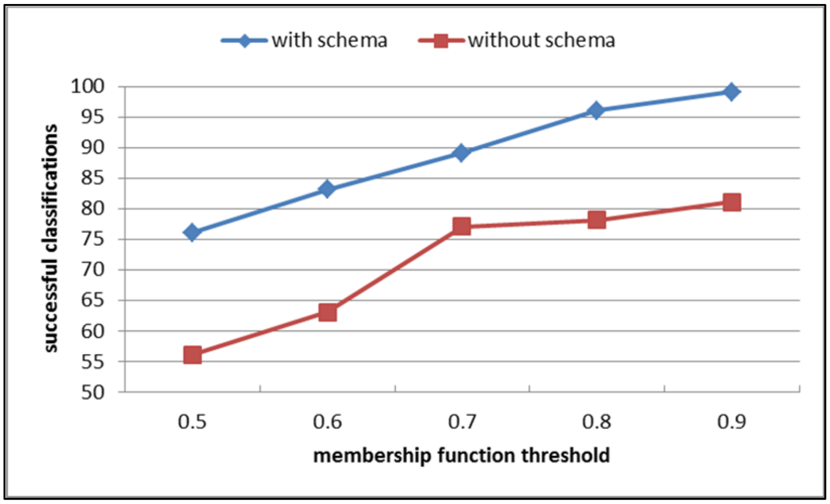


**Figure 5.2: Impact of threshold, membership function vs. accuracy of classification**

Learning has always played an important role in impeding the performance of computational systems, as described in Chapter 3. This section presents the results for learning and non-learning clustering. Through comparison of figure 2, 3 and figure 4 a cluster performance improvement can be observed with and without learning. It also was observed as a linear relationship between learning ratio and classification output. As expected learning showed huge impact on the clustering process. Another interesting finding was relationship between number of centroids and learning. It was discovered that in those scenarios when the cluster centroids cannot be defined because of any reason, increasing number of centroids can fulfill this limitation.

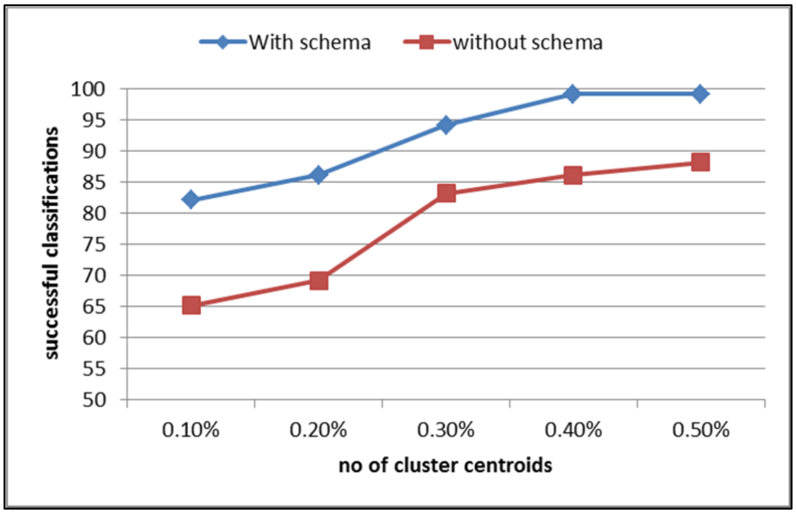


**Figure 5.3: Impact of cluster centroids**

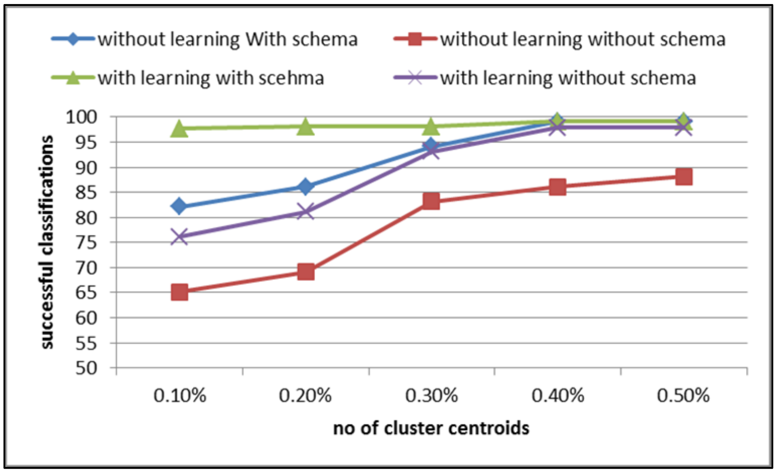


**Figure 5.4: Impact of Learning, cluster centroids vs. accuracy of classification**

## 5.2 News Tweet Dataset

The previous section had explained the scalability of our algorithm in which it was examine on the huge dataset of Agriculture Information Network. This section discusses the performance of the proposed system using news-tweet dataset. Chapter 4 has explained a heterogeneous information network for news-tweet as case study; this section has used the similar network.

### 5.2.1 Dataset

The dataset of tweets that were tweeted by different users during parliamentary election of 2013 in Pakistan was used for the experiments. More than 10000 tweets were recorded from May 5 to May 15 that contained links of different news articles. In the first phase all those news were filtered that did not had any news story attached with them. 10000 tweets and re-tweets are 30% of total tweets recorded from May 5 to May 15. Twitter API was used for recording of live tweets. API fetched the tweets after every 5 seconds and compare with the already extracted store in order to remove the already stored tweets. As it is known that a tweet is recognize using a tweet id therefore it was easy to remove the duplicate tweets from the dataset.

### 5.2.2 Constructing Heterogeneous Information Network

After extraction of tweets an information networks were constructed by using those extracted tweets. As the very first step the users from the tweets where extracted in order to create the following-follower graph for the users. This graph was created in order to measure the credibility of the users. In the next step the hash-tags were extracted from the tweets. Figure shows the list of more popular hash-tags that were obtained from the tweets. After links for news URLs were extracted from the tweets. After extraction of tags, news URLs and user's information a heterogeneous network for tweets news and users was constructed as explained in chapter 4 followed by the creation an adjacency matrix for creation of clusters.

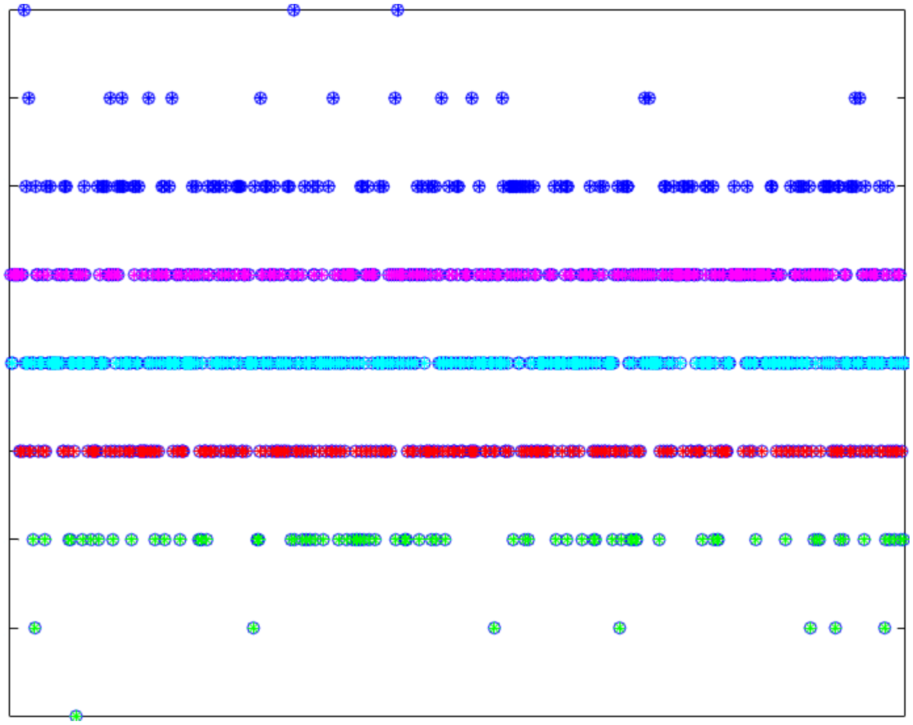


**Figure 5.5: Clustering News objects Without Weights**

### 5.2.3 Experimental Results

Firstly the tweets were clustered based on the different hash-tags with and without assigning them weights. Figure shows the behavior of the clusters with and without weighted values. Experimented were performed by assigning different weights to different attributes in order to examined behavior. In the end it was concluded from the obtained results that formation of a cluster is highly depended on the correct assignment of weights to the relationship own the edges. Good assignment of weights did produce good results where bad assignment of weights produce bad results.

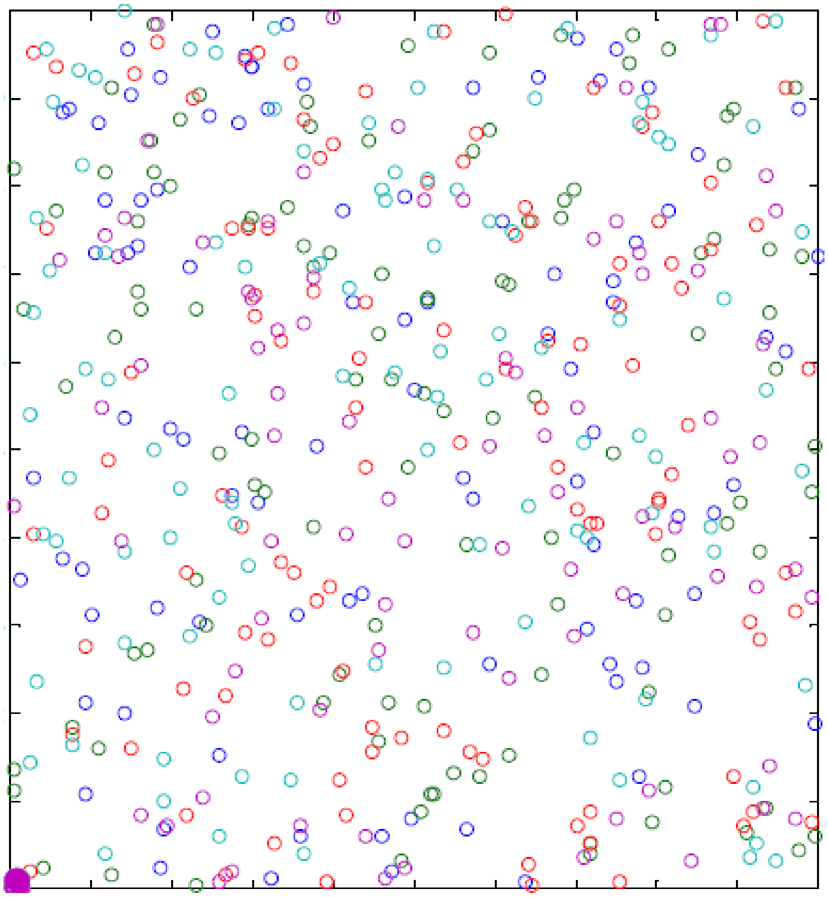


**Figure 5.6: Weighted News Objects before Clustering**

Clustering without weights resulted as gathering of number of common attributes and their values treated each value as equal, therefore the clusters were observed in the straight lines. It is because that either a parameter ahs an value or does not have a value. As in the tweeter information the data is binary without assigning weights to users. Therefore this type of results has been observed. While in figure the results are quite different and more realistic that support our argument that more better results can be obtained by mining heterogeneous information networks in many domains instead of mining homogenous information network by showing that user networked graph has an significant effect on the clustering of tweets.

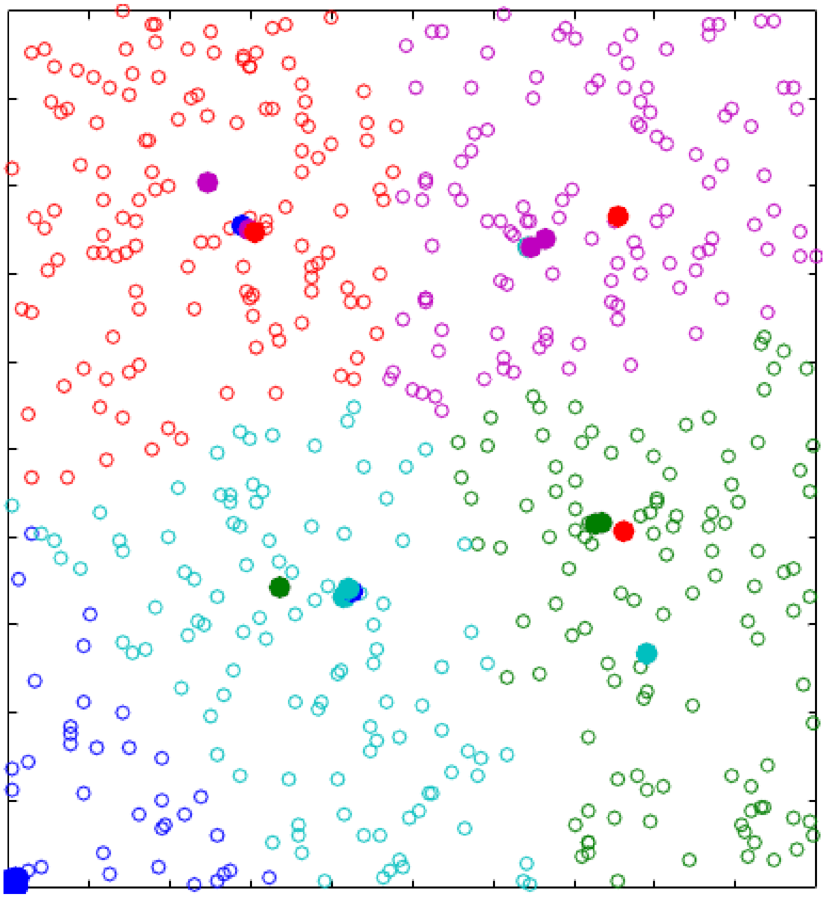


**Figure 5.7: Weighted News Articles Objects after Clustering**

In the following, the performance of the algorithm on the dataset has been studied. There were overall 250 news documents that were examined by choosing them randomly as a sample space to examine the performance. The performance of our proposed algorithm on this particular dataset was 95% as some false positive and false negative clustering objects has been observed. The percentage of false positive was high and it was 3.5% and where the percentage of false negative was observed 1.5. Overall 5% objects were wrongly related with hash-tags as extracted by the algorithm.

# Chapter 6  Conclusions

This thesis has presented a clustering algorithm based on fuzzy c-mean clustering algorithm in order to perform clustering on large scale unknown heterogeneous information networks. The aim behind was to preserve the original graph structure.  For this intention this work have proposed methodology by combiing  structure and object level clustering to improve quality of the clustering particularly when the objects are connected with each other and network structure is not available. Our technique allows finding of structure of heterogeneous networks based on connections that different objects hold among themselves. Proposed technique has used the clustered, resulted the clustering process for discovering relationship among the clustered heterogeneous information networks. To make the easy understandability of the proposed approach a tweeter-news case study has also been presented in this dissertation. Experiments have been performed on an agriculture information network in order to test quality of the proposed algorithm. It was discovered that working with schema and objects together result as creation of more precise and relation preserve clusters then creation of clusters separately. It also has been observed increasing of random cluster centroids in fuzzy c-mean algorithm almost result same as pre-defined cluster centroids.

# References

[1]. R. Xu and D. Wunsch, “Survey of clustering algorithms,” IEEE Trans. Neural Netw., vol. 16, no. 3, pp. 645–678, May 2005.

[2]. L. Wang, C. Leckie, K. Ramamohanarao, and J. Bezdek, “Automatically determining the number of clusters in unlabeled data sets,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 3, pp. 335–350, Mar. 2009.

[3]. L. Zheng, T. Li, C. Ding, 'Hierarchical Ensemble Clustering", In. 2010 IEEE International Conference on Data Mining, 20010, pp. 1199-1204

[4]. G. Serban, A. Campan, "A New Core-Based Method For Hierarchical Incremental Clustering", In. Proc. The Seventh International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC’05), IEEE, 2005, DOI: 10.1109/SYNASC.2005.9

[5]. K. Tasdemir, P. Milenov, B. Tapsall, "Topology-Based Hierarchical Clustering of Self-Organizing Maps", 474 IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 22, NO. 3, MARCH 2011

[6]. J. Zeng, L. Gong, Q. Wang, C, Wu, "Hierarchical Clustering for Topic Analysis Based on Variable Feature Selection", In Proc. 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009, IEEE, pp. 477-481

[7]. E. G. Mansoori, M. J. Zolghadri, and S. D. Katebi, “A weighting function for improving fuzzy classification systems performance,” Fuzzy Sets Syst., vol. 158, no. 5, pp. 583–591, 2007.

[8]. J. Deng, J. Hu, H. Chi, J Wu, "An Improved Fuzzy Clustering Method for Text Mining", In Proc. 2010 Second International Conference on Networks Security, Wireless Communications and Trusted Computing, 2010, IEEE, pp. 65-69

A. Szabo, L. N. Castro, M. R. Delgado, "FaiNet: An Immune Algorithm for Fuzzy Clustering", In Proc. WCCI 2012 IEEE World Congress on Computational Intelligence, 2012, IEEE press

[9]. L. Zheng, T. Li, "Semi-supervised Hierarchical Clustering", In Proc. 2011 11th IEEE International Conference on Data Mining, 2011, pp. 982-991

[10]. D. Park, "Intuitive Fuzzy C-Means Algorithm", IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2009, pp. 83-88

[11]. M. Ceccarelli, A. Maratea, "Semi-supervised fuzzy c-means clustering of biological data", In. Proc. Proceedings of the 6th international conference on Fuzzy Logic and Applications 2005, Springer-Verlag, pp. 259-266

[12]. J. Han, Y. Sun, X. Yan, P. S. Yu, "Mining knowledge from databases: an information network analysis approach", In proc. SIGMOD '10: 2010 ACM SIGMOD International Conference on Management of data, 2010, pp. 1251-1252

[13]. D. Saez-Trumper, G. Comarela, V. Almeida, R. Baeza-Yates, F. Benevenuto, "Finding trendsetters in information networks", In. Proc. KDD '12: 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, ACM, pp. 1014-1022

[14]. T. Wang, M. Srivatsa, D. Agrawal, L. Liu, "Modeling data flow in socio-information networks: a risk estimation approach", In Proc. 16th ACM symposium on Access control models and technologies, 2011, pp. 113-122

[15]. M. Newman, “Networks: An Introduction”, Oxford Univ. Press, 2010.

[16]. D. Easley and J. Kleinberg, Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge Univ. Press, 2010.

[17]. X. Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger, “SCAN: A structural clustering algorithm for networks,” in Proc. 2007 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD’07), San Jose, CA, Aug. 2007.

[18]. G. Serban, A. Câmpan, “A new core-based method for hierarchical incremental clustering ", In Proc. SYNASC 2005. Seventh International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, 2005, IEEE, DOI: 10.1109/SYNASC.2005.9

[19]. Chen, J., Liu. J., Yu, W., "Discovering Semantic Relationships for Knowledgebase", Third International Conference on Pervasive Computing and Applications, 2008, ICPCA 2008.

[20]. Dong, Y., Shen, D., Nie, T., Kou, Y., "Discovering Relationships among Data Resources in DataSpace ", Sixth Web Information Systems and Applications Conference, 2009, WISA 2009, pp. 76-81, IEEE Press

[21]. K, G., Potamias, M., Terzi, E., "Clustering Large Probabilistic Graphs", IEEE Transactions on Knowledge and Data Engineering, Volume: 25 , Issue: 2, pp. 325 - 336, 2013

[22]. Raheja, V., Rajan, K.S.,"Comparative Study of Association Rule Mining and MiSTIC in Extracting Spatio-temporal Disease Occurrences Patterns", IEEE

12th International Conference on Data Mining Workshops (ICDMW), pp. 813 - 820, 2012

[23]. Huang, Y., Yeh, H., Soo. V., "Network-based inferring drug-disease associations from chemical, genomic and phenotype data", 2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1-6, 2012

[24]. Ravichandran, S.S., Sathya, D., Shanmugapriya, R., Isvariyaa, G.,"Rule-base data mining systems for customer queries", 2012 Third International Conference on Computing Communication & Networking Technologies (ICCCNT), pp. 1-5, 2012

[25]. Li, W., Xu, Y., Yang, J., Tang, Z., "Finding structural patterns in complex networks", 2012 IEEE Fifth International Conference on Advanced Computational Intelligence (ICACI), pp. 23 - 27, 2012