ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

트위트 이형 정보 망을 이용한 뉴스 기사의 사용자 지향적 클러스터링☆

User Oriented clustering of news articles using Tweets Heterogeneous Information Network

무하마드 쇼아입1 송 왕 철1*
Muhammad Shoaib Wang-Cheol Song

요 약

월드와이드 웹, 특히 web 2.0의 출현과 함께 뉴스 기사들의 양이 엄청나게 증가하면서 독자들이 그들의 요건에 맞춰 뉴스기사를 선택하는데 어려움이 있다. 이러한 문제를 해결하기 위해서 여러 클러스터링 메커니즘이 뉴스기사들을 분별하도록 제안되었다. 하지만, 이러한 기법들은 완전히 기계 지향적 기법들이고, 클러스터링의 멤버쉽을 결정하는 과정에 사용자의 참여가 제외되어 있다. 본 논문에서는 뉴스 기사 클러스터링 처리과정에서 참여문제를 해결하기 위해서, 객체들을 클러스터링하는 뉴스 기사와 트위터에 포스트하려는 사용자의 결정을 조합하므로써 뉴스 기사를 클러스터링하는 프레임워크를 제안한다. 우리는 이를 위해 트위터 해쉬-태그를 이용할 수 있도록 했다. 더욱이, 트윗된 글에 대한 리트윗 빈번도에 기반하여 사용자의 신용도를 계산하므로써, 클러스터링 멤버쉽 함수의 정확도를 개선시키려 한다. 제안된 방법에 대한 성능을 보이기 위해, 2013년도에 파키스탄에서 있었던 선거동안에 발생한 메시지를 이용했다. 우리의 결과를 통해 사용자의 결과를 이용하므로써, 일반 클러스터링보다 더 나은 결과물이 달성될 수 있음을 보였다.

☞ 주제어 : 사용자 지향적 클러스터링, 정보망, 뉴스 기사 클러스터링, 트위트, 마이크로 블로깅

ABSTRACT

With the emergence of world wide web, in particular web 2.0 the rapidly growing amount of news articles has created a problem for users in selection of news articles according to their requirements. To overcome this problem different clustering mechanism has been proposed to broadly categorize news articles. However these techniques are totally machine oriented techniques and lack users' participation in the process of decision making for membership of clustering. In order to overcome the issue of zero-participation in the process of clustering news articles in this paper we have proposed a framework for clustering news articles by combining users' judgments that they post on twitter with the news articles to cluster the objects. We have employed twitter hash-tags for this purpose. Furthermore we have computed the credibility of users' based on frequency of retweets for their tweets in order to enhance the accuracy of the clustering membership function. In order to test performance of proposed methodology, we performed experiments on tweets messages tweeted during general election 2013 in Pakistan. Our results proved over claim that using users' output better outcome can be achieved then ordinary clustering algorithms.

☞ keyword : User Oriented clustering, information network, news articles clustering, tweets, micro-blogging

1. 서 론

Clustering is one of the most important and most used methods in Data Mining that allows users to observe diversity present in any particular dataset. It is also used to view different type of objects in huge dataset using their different properties. Different types of techniques have been studied in clustering text data and documents, however the limitation of these methodologies is that these methodologies only depend on textual information present within a

1 Dept. of Computer Eng., Jeju National University, Jeju, 690-756, Korea.
* Corresponding author (philo@jejunu.ac.kr)
[Received 31 October 2013, Reviewed 4 November 2013, Accepted 2 December 2013]
☆ This research was supported by the 2013 scientific promotion program funded by Jeju National University.
☆ A preliminary version of this paper appeared in APIC-IST 2013, Aug 12-14, Jeju Island, Korea. This version is improved considerably from the previous version by including new results and features.

document or an object and do not utilize external information with whom the document of objet is linked [10].

World Wide Web (WWW) is an example of largest network of documents that is consisted of trillions of documents that are uploaded on different servers by various users, agencies, and corporations. These documents include technical articles, news, current affair articles, and articles containing political or economic analysis etc. As the software technology has been developed, various contents containing those kinds of documents are also posted so much these days in the social networking platforms such as Facebook, Tweeter and so on [18, 19, 20]. As the volume of these articles and documents approaches to hundreds of thousands of documents per day it is not possible for a user to go through all these documents and find the documents and articles related to his/her research areas, education, and business interests. In order to overcome this issue, different mechanisms have been suggested for clustering documents with respect to text they have [2, 6, 7]. However the limitation of these clustering techniques is that they cluster documents without taking users' participation in to account during the process of clustering. As the users' participation in process of clustering is equal to zero, most of time, the documents' clusters that are clustered using these methodologies do not provide information that is needed.

With the emergence of web 2.0, the ways of internet usability has been drastically changed. One major reason of the popularity of web 2.0 is that it encourages users to participate instead of prior generation of web where users were only allowed to watch and read the content that is available on the internet. This participation is of different types like users' comments about some article, video, audio, and their liking or disliking for an article, video, or audio available on the internet. Besides this now social networking sites have become popular applications of web 2.0 where users share their ideas, thoughts and give comments about different objects available on the Internet. One of the most popular applications that have emerged in recent years is micro blogging in which users provide their thoughts and comments using hash-tags. Twitter is one of the popular and widely used applications of micro-blogging. Many efforts have been observed on the analysis of micro blogging data because of the reason that data in micro-blogging sites is relatively very small, well organized and use well defined tags that make the processing of data much easier from traditional blogging data [2].

Users' posts that are being made on the micro-blogging websites have been used for different purposes e.g. to analyze one's political or social prospective generally or during special political and social events. For example recently during the general elections held in Pakistan, twitter and Facebook were widely used to propagate the manifestos of the political parties in order to convince their voters to vote for them. Simultaneously these micro-blogging posts have been used for understanding voters' political views or estimate trends during important events. Clustering micro-blogging posts such as twitter posts has emerged as new research area in recent years but this research fuscous only clustering of tweets by topics or similarity scores.

With the emergence of Information Networks [4] Graph and Networked data [4], intra-domain clustering has been emerged as a new research challenge in which objects are linked among each other for formation of clusters. This has opened new research areas that allow usage of micro-blogging posts in both clustering of other related objects. In this work we have introduced a novel framework for clustering of news articles by taking user generated genres or hash-tags present in micro-blogging data into account for the purpose of enabling users' participation in process of clustering. Consider the example of twitter in which users tweet their opinions in small length text messages. Those tweets contain genres in form of hash-tags. We have utilized these hash-tags of tweets to cluster the other objects e.g. news articles, pictures, video clips that are attached to tweets. The reason for using tweeter data is twofold, firstly tweeter data is small that allow quick processing, and secondly its "#" has-tags allows easy identification of genres in tweets. We call it user-oriented clustering because clustering is done based on tweets and their genres that are defined by the users. We claim that better outcome has been achieved from user-oriented clustering by utilizing users' generated information.

A brief overview of our proposed framework is as follows: We first obtain the list of tweets and news articles published in a specified duration. Among them any tweets that have no link for news articles are discarded. We call

this process of screening tweets, “preprocessing”. We then rank the news articles with respect to the number of retweets with a rule that the most a news article is shared, the higher its rank is. The most highly ranked news articles for each hash-tag, are used as the cluster centroids and other news articles are clustered with respect to those cluster centroids. We have used fuzzy c-mean [6] algorithm as the clustering algorithm in this work. In this framework the results are presented for each cluster sorted with respect to their distance from cluster centroids. Framework also can provide a way to calculate coupling and coherence among news articles based on coupling and coherence values of their respective hash-tags.

Our contributions in this paper are:

1. We have presented tweets and news articles in the form of an information network, in order to apply graph data clustering algorithms on them.
2. We have introduced a novel idea of using micro-blogging genres in clustering of objects or documents in particular for those documents that are present on web.
3. We have observed the effect of Euclidean and ochiai distance formulas on the output.

The rest of paper is organized as follows: section 2 describes a brief introduction of information networks, followed by section 3 in which we have formalized news twitter information network. In section 4 we have presented our clustering framework by introducing mathematical formulation to measure similarity among objects using news tweet information network. This model has been evaluated in section 5 and finally section 6 concludes this paper.

2. Clustering of Graphs and Information Networks

In this section of the article we firstly define the concept of graphs and information network mathematically followed by review of some existing techniques for graph clustering and applications of graph in other domains.

A *graph* is a set of nodes and links (or vertices and edges), where the nodes and/or links can have arbitrary labels, and the links can be directed or undirected (implying an ordered or unordered relation).

Definition: *Information Network: Given a set of atomic types* $T = \{t_1, t_2, t_3, t_4 \dots t_n\}$, *set of objects* $\tilde{O} = \{O_{t=1}^{T}\}$ *where* O_i *is set of objects belonging to type* t_i *and set of relations* $\mathcal{R} = \{r_1, r_2, r_3, r_4 \dots r_n\}$, *a Description Graph* $G = (V, E)$ *is called an* ***information network*** *for* $\tilde{O}$ *if* $V \in \tilde{O}$ *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.*

The clustering problem of an entire graph in a multi graph dataset has been discussed as follows: The reason for presenting here an overview of the graph clustering algorithms is that we have formalized relationship of news article with tweets as an information network, a form of graph dataset. In general graph clustering algorithms have been studied for clustering xml documents, where the full document is represented as one graph including its structure and data. FOAF [16] – friend of a friend – ontology is another example of a graph of graphs in which the graph of a person is entirely linked with graphs of other persons. Most of the algorithms that are used for clustering of the objects use the similarity matrix, so that there exist a need for creation of a mechanism that appropriately use these measurement functions for the clustering of a graphs. In the following, two families of clustering algorithms for clustering the entire graphs have been briefly explained along with their deficiencies. Many of the known approaches for the clustering use the cluster centroids that can be measured using statistical means and median, however for the graphs determination of these values is a challenge as they cannot be computed easily. There are three major approaches in the conventional data mining techniques that have been used in clustering graph and networked data.

Structure distance based approach has been used to compare the XML documents by comparing their structure. In this approach structural distance among different objects is calculated and then is compared with each other using a distance matrix. *XClust algorithm* is one of structure distance based clustering algorithms for clustering the XML document. It undertakes the hierarchical basic 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 algorithms is that they use text matching for the clustering of the documents instead of using the original graphical structure and data set so that these algorithms are more adjacent towards text clustering algorithms rather than document clustering algorithm [7, 9].

Structural Summary Based Approach has been proposed for summarizing the document in the first step 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 category also use tree based comparison methodologies, where a tree structure is firstly created for each document followed by their compression among each other to form a cluster. [10].

Node Clustering Algorithms are widely used in order to cluster multi-domain graph data by defining the distance among 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 partition the graph in such a way that the weights on the edges become minimum. This problem is also known as minimization cut problem [13]

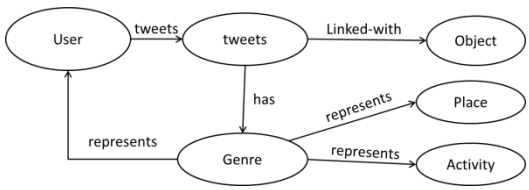
From this short survey it can be observed that current graph clustering algorithms use information from the objects i.e. nodes of the graphs in order to calculate similarity between these objects. Furthermore, we also have observed that all graph clustering frameworks that have been proposed for document clustering to date have been developed for XML and structural documents. This survey also shows that there is no support available for users' input in the current document clustering techniques. In the next section we have formalized graphs for news tweet network and have presented it as an information network whereas the proposed clustering technique is described in section 4.

3. Forming News Tweet Network

Tweets are usually written in an informal language however a good point is that they have 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 easily.

3.1 Tweet Information Network

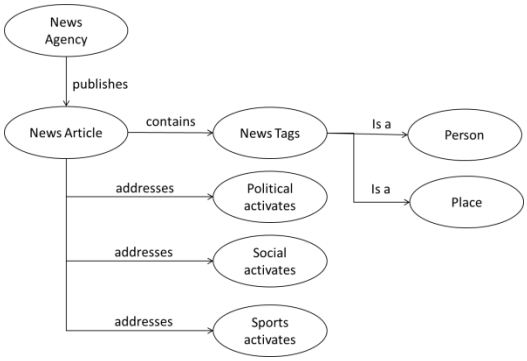
Figure 1 shows a heterogeneous information network of tweeter's tweets. The main objects are tweets, users, tags, pictures and the web URLs. A tweet object has attributes, genres, tags, and list of objects that are linked with it. Tweeter users generate huge amount of tweets consisting huge amount of stories, web URLs and pictures diverse genres. These genres usually are used to recognize the topics addressed in tweets. To keep the simplicity in this work we have used those genres that have been created based on political and social situations. Each genre addresses the thinking of a particular group of users and is also used to analyze the political and social trends in a particular region in a particular period. A sub network (Homogeneous Network) can also be build using intra-genre relationships.



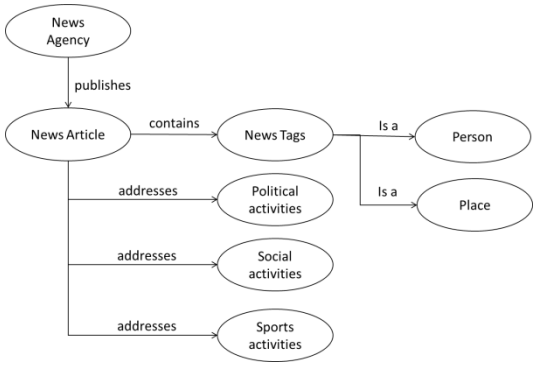
(Figure 1) Tweet Information Network

3.2 News Information Network

News information network is another example of heterogeneous information network in which "news article" is a basic object that is published by news agencies or news websites. News article can provide information about different events, personalities, places, incidents, celebrities, political and social trends. Different news articles can address the same topic, however is usually published by different news publishing agencies which have different level of trust to the audiences. Information that can be extracted from these news articles can be used for the construction of a heterogeneous information network in order to find most reliable news of one's interested topic. Figure 2 presents an overview of news heterogeneous information network.



(Figure 2) New Information Network



(Figure 3) News Tweet Information Network

3.3 Combining News and Tweets Information Networks

In sub-sections 3.1 and 3.2 the tweets and news heterogeneous information networks have been explained individually. In this sub-section we have described that how these two information networks can be combined in order to build a new large heterogeneous information network. Consider Figure 3 that presents the news tweets heterogeneous information network and shows the relationship between tweets and news articles. By comparing Figure 1, 2 and Figure 3 it can be observed that both networks share many common attributes.

As depicted in Figure 3, a genre in tweets represents an activity that is addressed by news articles. This activity can be further divided into sub-activities e.g. political, social, sports activities and so on whereas one link of a news article can also be a part of these tweets.

4. Clustering News using Tweets' Hashtags

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 users' credibility.

Before moving towards the technique we briefly introduce the similarity matrices used in information network to find similarity between objects in an information network as follows.

The similarity of two objects depends on two major characteristics, one characteristic is that two objects share similar schema i.e. they have values for similar relationships or properties. The second characteristic is that those two objects also share the values of shared properties. These characteristic can also be said as schema level similarity and object level similarity respectively.

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 $U_i, U_j \subseteq \tilde{\mathbb{O}}$, 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 $U_i, U_j \subseteq \tilde{\mathbb{O}}$, 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 $\left(\forall\, x \in U_i, y \in U_j \; \exists\, x = y\right) \wedge \left(\forall\, a \in R_i, b \in R_j \; \exists\, a = b\right)$*

Let $T = \{t_1, t_2, t_3, \ldots\ldots t_n\}$ is set of tweets, $U = \{u_1, u_2, u_3, \ldots\ldots u_n\}$ is set of users, $N = \{n_1, n_2, n_3, \ldots\ldots n_n\}$ 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 $T \cap N = \emptyset$

4.1 Simple Clustering

Before the process of clustering, this section presents the ranking algorithm for the ranking of news articles based on tweets and re-tweets about news articles. Firstly the Tweets vs. News Documents (TD) Information Network has been presented followed by the construction of clusters of different news articles from the graph $G_D = \{(T \cup D), E\}$ based on hash-tags that are present in the network. The more, the users will add the tags while sharing of the URL of a news article, the more chances will exist for the news article to become closer to the cluster centroid of the cluster created for representing articles of that particular hash-tag.

Given a set of tweets $T = \{t_1, t_2, t_3, \ldots\ldots t_m\}$ and $D = \{d_1, d_2, d_3, \ldots\ldots d_n\}$ as a set of news articles where for each tweet $t \epsilon T$ there exist a set of hash-tags. Let H represents set of all hash-tags and $H_t \subseteq H$ represents the set of hash-tags that are associated with a tweet t that has a news article $n \epsilon N$ associated with it. Now a news article n can be binary clustered in all of the hash-tags $h \epsilon H_t$. However binary clustering is not enough so that there is a need to define membership function based on the relationship among, h and t on the basis of rationale that the more number of tweets will uses the hash-tags for a news article, the membership value will be awarded to the news-article. Equation 1 defines a formal mathematical model for the membership function $Rank(h_i, n_j)$

$$Rank(h_i, n_j) = \frac{n\left(t(h_i, n_j)\right)}{n\left(t(h_i, n_j)\right) + n\left(t(h_i, n_j)\right)} \quad \textbf{Eq. 1}$$

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

The value of $Rank(h_i, d_j)$ remains between 0 and 1. Once $Rank(h_i, d_j)$ is computed for all $h \in H_i$, d_j is added to all those clusters where $Rank(h_i, d_j) > \alpha$ and α represents the minimum threshold value that is required for joining a cluster. In general value of α can be set to 0.5.

Initially for all $h \in H$, independent, disjoint and non-overlapping clusters where constructed. Therefore it can be said that at the initial level the number of clusters are equal to the number of hash-tags appeared in all tweets.

Using the equation 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 done. As the value for $Rank(h_i, d_j)$ is increased, the news article can 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 $Rank(h_i, d_j)$ or in other words $Rank(h_j, d_i)$ represents the distance between a cluster centroid and d_j.

4.2 Finding overlapping between clusters

In the previous subsection it has been discussed how a news article can be added to one or more clusters when clusters are disjoint from each other and are constructed 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 clusters in order to discover relationships between news articles clustered in the different clusters.

The easiest way to identify overlapping between clusters is by calculating the coupling between 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 identified. In order to find the coupling between two different clusters identified by hash-tags h_i and h_j, the value for overlapping can be defined as a coupling function $Cup(h_i, h_j)$ and compute it a follows

$$Cup(h_i, h_j) = \frac{n\left(d(h_i) \cap d(h_j)\right)}{n\left(d(h_i)\right) + n\left(d(h_i)\right)} \quad \textbf{Eq. 2}$$

And ochiai coefficient can be compute as follows:

$$\text{ochiai}\left(h_i, h_j\right) = \frac{n\left(d(h_i) \cap d(h_j)\right)}{\sqrt{n\left(d(h_j)\right) \times n\left(d(h_j)\right)}} \quad \textbf{Eq. 3}$$

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

4.3 Using User Credibility

In the previous section it has been explained 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 not true. Each tweet and re-tweet has the similar weight without the credibility. In this section it is explained how the weights to the users' profiles are assigned and how those weights can be integrated in order to improve our clustering results.

There exists a very basic method to understand the credibility of users in twitter. One of them is using following-follower network and other one is by using re-tweet network. If a user has more followers, the user has more credibility. Similarly, the number of users who did re-tweeted one's messages also represents his or her credibility.

Relationship described above does not form any layered architecture but it is nested without depth limit. Consider an example of follower-following relationship. Then a way to measure credibility can be how many followers a person have, but another more important thing is how much those followers are credible. This recursive experience can be very simple and even can be endless. 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.

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

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

Here u represents users, $c(u)$ represents the credibility of the user, $t(u)$ represents 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 represent the total number of tweets and retweets respectively.

For an author the more his or her tweets are re-tweeted 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 has been enhanced as following.

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

After computing the user credibility, each tweet is then assigned a weight based on credibility of its author and credibility of those users who retweeted it. When a highly credible a 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.

$$rank(t) = C\big(a(t)\big) + \sum_{u \in rt(u)} \left(\frac{C(u)}{2}\right) \quad \textbf{Eq. 3}$$

where $a(t)$ represents the author of a tweet t. Once rank value for each has been obtained the equation Eq. 1 can be modified for membership function as follows.

$$Rank(h_i, n_j) = \frac{\sum_{t \in \{t(h_i, n_j)\}} rank(t)}{\sum_{t \in \{t(h_i)\}} rank(t) + \sum_{t \in \{t(n_j)\}} rank(t)} \quad \textbf{Eq. 4}$$

5. Experiments and Results

In this section we have applied our technique on a dataset in order to evaluate the accuracy of the proposed framework.

The dataset of tweets that were tweeted by different users during parliamentary election of 2013 in Pakistan was used for the experiments. We used java based twitter API [17] to gather the data from twitter's server. More than 10,000 tweets were recorded from May 5 to May 15 that contained links of different news articles. The performance has been measured on 100 articles selected manually. In the first phase all those tweets were filtered that did not had any news story attached with them. 1000 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 was compared with the already extracted news in order to remove the duplicate tweets. As it is known that a tweet is recognize using a tweet ID therefore it was easy to remove duplicate tweets from the dataset.

After extraction of tweets an information network 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. After links for news URLs were extracted from the tweets hash-tags were linked with extracted news links. After extraction of hash-tags, news URLs and users' information from tweets, a heterogeneous information network of objects as explained in Section 3. We then created an adjacency matrix for creation of clusters. The adjacency matrix has been used to store the relationship between objects of the heterogeneous information network.

(Table 1) Results using Ochiai Distance

Hash-Tag	Total Tweets	Rightly clustered	Wrongly clustered
#hashtag1	200	184	16
#hashtag2	500	430	70
#hashtag3	300	252	48

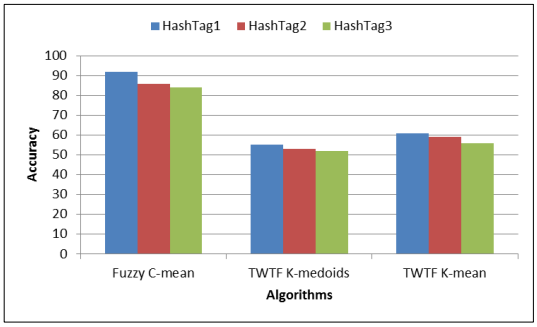
(Table 1) Results using Euclidean distance

Hash-Tag	Total Tweets	Rightly clustered	Wrongly clustered
#hashtag1	200	179	21
#hashtag2	500	423	77
#hashtag3	300	243	57

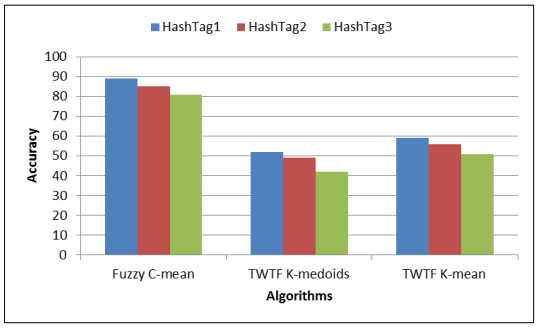
We observed the results in term of false positive – percentage of the objects that should be clustered in a cluster but have not been clustered – and false negative – percentage of the objects that should not have been clustered however they have been clustered – in order to measure the accuracy of the algorithm. Table 1 shows the outcome the experiments. Total 100 news articles were evaluated using algorithm. The overall accuracy was observed 80%, false positive was 4%, and false negative was 16%.

The next parameter that we have studied was the effect of Euclidean distance Eq. 2 with ochiai distance Eq. 3. Table 1 and 2 depicts the results for Eq. 1 and Eq. 3 respectively. The results are almost similar to each other however Ochiai distance gave slightly improved results that depicts that choosing of distance formula does not have a significant impact on results.

Finally we have compared our results with existing clustering methodology known as "TWTF" using k-mean and k-mediates. Figure 4 and 5 present the results of experiments. Our results have shown the significant improvement in the form of the accuracy in clustering process using clustering based on Information Networks.



(Figure 4) Algorithm Comparisons using Ochiai Distance



(Figure 5) Algorithm Comparisons using Ochiai Distance

6. Conclusions

In this paper, we have presented a novel framework for clustering news articles by using users' participation in the process of clustering. The aim in this work was to study the effect of users' participation in the process of clustering. In order to fulfill these requirements we firstly created an information network by combining tweets and news-metadata followed by proposed algorithm for clustering. We performed experiments on 100 news articles using 10000 tweets that were posted during the general election of Pakistan in 2013. In future we plan to enhance our algorithm by combining traditional techniques with user-oriented clustering to get more accuracy in the clustering. Furthermore we also plan to perform more experiments on different other dataset in order to observe the performance of proposed algorithm.

Reference (참 고 문 헌)

[1] Y. Hu, E. M. Milios, J. Blustein, "Interactive Feature Selection for Document Clustering", In. In the 26th Symposium On Applied Computing, 2011, Tiwan, ACM Press

[2] E. Kontopoulos, C. Berberidis, T. Dergiades, N. Bassiliades, "Ontology-based sentiment analysis of twitter posts", In. Expert Systems with Applications, Vol 40, (2013) pp. 4065 - 4074

[3] B. Liu, X. Li, W. S. Lee, P. S. Yu, "Text classification by labeling words" In. AAAI, 2004, ACM Press

[4] M. Newman, "Networks: An Introduction", Oxford Univ. Press, 2010.

[5] 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

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

[7] 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

[8] S. Fortunato, "Community detection in graphs." Physics Reports, 486(3-5):75 - 174, 2010.

[9] Rigutini, L.; Maggini, M., "A semi-supervised document clustering algorithm based on EM," Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on , vol., no., pp.200,206, 19-22 Sept. 2005.

[10] R. Huang and W. Lam. "An active learning framework for semi-supervised document clustering with language modeling," Data & Knowledge Engineering, Vol. 68 Issue 1, pp49-67, 2009 Elsevier

[11] Y. Huang, H. Yeh, V. Soo. "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

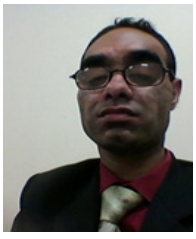
[12] S. S. Ravichandran, D. Sathya, R. Shanmugapriya, G. Isvariyaa, "Rule-base data mining systems for customer queries", 2012 Third International Conference on Computing Communication & Networking Technologies (ICCCNT), pp. 1-5, 2012

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

[14] 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

[15] Y. Dong, D. Shen, T. Nie, Y. Kou, "Discovering Relationships among Data Resources in DataSpace ", Sixth Web Information Systems and Applications Conference, 2009, WISA 2009, pp. 76-81, IEEE Press
[16] D. Brickley, L. Miller, "FOAF Vocabulary Specification", 2010 http://xmlns.com/foaf/spec/
[17] "Twitter API", https://dev.twitter.com/
[18] Jo Hyeon, Hong Jong-hyun, Choeh Joon Yeon, Kim Soung Hie,"A recommendation algorithm which reflects tag and time information of social network," Journal of Korean Society for Internet Information, v.14, no.2, 2013, pp.15-24.
[19] Sam-Yull Hong, Jae-Cheol Oh,"Comparative analysis on Social Network Service users access : Based on Twitter, Facebook, KakaoStory", Journal of Korean Society for Internet Information, v.13, no.6, 2012, pp.9-16.
[20] Tai-Wan Kim, Bumjun Park, Taekeun Park,"An Augmented Memory System using Associated Words and Social Network Service", Journal of Korean Society for Internet Information, v.11, no.6, 2010, pp.41-50.

◐ 저 자 소 개 ◑



Muhammad Shoaib

2010년 National University of Computer and Emerging Sciences Islamabad, Computer Science (Bachelor of Science)
2013년 제주대학교 컴퓨터공학과(공학석사)
관심분야 : 데이터마이닝, 네트워크, 시맨틱웹, Information Retrieval, Data Storage and Querying.
E-mail : muhammad.shoaib@live.com

송 왕 철

1985년 연세대학교 식품공학과 졸업(공학사)
1989년 연세대학교 전자공학과 졸업(공학사)
1991년 연세대학교 본대학원 전자학과 졸업(공학석사)
1995년 연세대학교 본대학원 전자공학과 졸업(공학박사)
1996년~현재 제주대학교 컴퓨터공학과 교수
관심분야 : VANET, 정보공유, 망광리, SDN.
E-mail : kingiron@gmail.com