UNIVERSITY OF PRISHTINA

FACULTY OF MATHEMATICS AND NATURAL SCIENCES

DEPARTMENT OF MATHEMATICS, COMPUTER SCIENCE

LITERATURE SURVEY

CLUSTERING OF WORDS FROM SOCIAL NETWORKS

ZANITA RAHIMI, VALON HALITI

PRISHTINA, MAY 2017

Table of Contents

CLUSTERING OF WORDS FROM SOCIAL NETWORKS

Abstract:

Clustering is a widely studied data mining problem in the text domains. K-means is one of the most commonly used clustering algorithms. It is widely used for text clustering as it converges fast to a local optimum, even when it works with a large sparse matrix. In this paper we improve text understanding by using frequent item (term) sets for text clustering.

Keywords:

Social networks, text clustering, k-means, text mining, text classification.

1 Introduction:

Social networking websites create new ways for engaging people belonging to different communities. Social networks allow users to communicate with people exhibiting different moral and social values.

Online social media has emerged as a medium of communication and information sharing. Status updates, blogging, video sharing and social networking are some of the ways in which people try to achieve this. Popular online social media sites like Facebook and Twitter allow users to post short message to their homepage. These are often referred as micro-blogging sites and the message is called a status update [7].

Text clustering methods can be applied to structure the large result set such that they can be interactively browsed by the user [12].
The clustering methods algorithms are mostly of two types:

1. hierarchical methods
2. partitioning methods (non-hierarchical).

The hierarchical algorithms for clustering represent data sets as a cluster tree and are of two types:

1. agglomerative
2. divisive hierarchical clustering methods.

Partitional clustering algorithms are of two types:

1. iterative
2. single pass methods

K-means and its variants etc. are the popular partitioning methods [17].

K-means algorithm is one of the most used unsupervised learning algorithms for clustering tasks. It uses a pre-defined number of clusters k to partition the introduced data to this number of clusters. The main advantages of k-means are the ease of implementation, fast computational cost compared to other techniques, and independence of data ordering [18].

Text mining is a form of data mining that deals with text resources. It is the process of discovering by learning machines new or previously unknown hidden information from text resources [18].

The problem of classification has been widely studied in the database, data mining, and information retrieval communities. In reality, the frequency of words also plays a helpful role in the classification process, and the typical domain-size of text data (the entire lexicon size) is much greater than a typical set-valued classification problem [10].

Social networks require text mining algorithms for a wide variety of applications such as keyword search, classification, and clustering.

The rest of the paper is organized as the following; section 2 presents the methodology of the research, how the data were collected and processed. Section 3 provides a background and general information about text mining and clustering, it also surveys literature about related work. Section 4 discusses the experimental analysis and results, how the experiments were conducted and what are the outcomes. In section 5 are the references.

2 Research methodology

This section of the paper describes the steps of this research methodology.
We collected papers about the topic "Text clustering in social networks" using Google Scholar[1].
First we searched with these keywords: **text mining**, **text clustering**, and **k-means**.
We got 2750 results, and that is a huge number of papers to read. So we decided to be more specific. We searched with these new keywords: **unsupervised learning**, **text clustering**, **text learning**, and **k-means**. This combination of keywords turned out to be the best, because we got 65 results, then we started to download these papers. Most of them were in pdf format, somewhere around 50.
Some papers were duplicate, they had different titles on different websites, but the content was the same. So then the number of papers was around 35.
Also, some papers had nothing to do with our topic, and some of them did not have enough information. So the number of papers was reduced to 20.
We have read all of these papers and have made the classification below:

Paper	Article	Survey	Journal	Study	Conference paper	**Total**
5	3	4	2	5	1	20

[1] Google Scholar- provides a simple way to broadly search for scholarly literature.

We saw that most of these papers use the same clustering algorithm. In 13 out of 20 papers is used **k-means** clustering algorithm to cluster text.
Distance measures used in these papers are: Euclidian distance, Jaccard similarity, cosine similarity, RSS distance[2], relative entropy[3] etc.

2.1 Text Representation

In order to prepare the text data for clustering analysis, the text strings in documents should be presented in a way that the learning machine can understand and process it. In this research, the text documents (posts) will be presented as vectors. One of the most used models to represent the text is the Bag of Words (BOW). The text is broken down into tokens (words), where each token is considered as feature vector for the text documents.
Each document or post in our case is presented by a bag of its own words. Each word or term in feature space is weighted by a weighting scheme.
Most popular schemes are term frequency (TF) and term frequency-inverse document frequency (TF-IDF).

Term frequency will look into each term in a document and measure how frequent it appears in this document by dividing the number of appearances of each term on the total number of terms in this document.
Inverse document frequency will be the determining factor for scoring a weight for each term. Document frequency is the number of documents which contain a specific term [18].

3 Background

This section delivers background information about text clustering and clustering technique used in this study. It also reviews literature which have used text clustering on social networks.

3.1 Text clustering

The major challenge in handling short text documents is to deal with the sparsity of the words in them. Typically, documents are represented as a TFIDF feature vectors, where a document represents a data point in d-dimensional space where d is the size of the corpus vocabulary [3].

3.2 K-means algorithm

We begin with K-means clustering. The important factors in K-means is defining the distance measure between two data points and defining the number of clusters [3].
This algorithm first uses mean to calculate the cluster center, and then partitions the data sets into

[2] RSS Distance- is the squared distance of each vector from its cluster centroid summed over all vectors in the cluster.
[3] Relative entropy- is a measure of distinguishability between two quantum states.

different clusters by minimizing the distances between the objects and the cluster centers [19]. K-means clustering algorithm works through the following number of processes:

1. Initialize a number of k centroids
2. Assign each point randomly to a cluster to have all points assigned to cluster
3. Calculate each point to its closest centroid through a distance measurement and assign those points to their closest clusters
4. Recalculate the mean of the points in each cluster to assign new centroids
5. Repeat the calculation and assignment of points until reaching maximum number of iterations or no change occurs in cluster assignment [18].

3.3 Minibatch k-means

The k-means optimization problem is to find the set C of cluster centers $c \in R^m$, with |C| = k, to minimize over a set X of examples $x \in R^m$ the following objective function:

$$min \sum_{x \in X} \left| |f(C, x) - x| \right|^2$$

Here, f (C, x) returns the nearest cluster center c ∈ C to x using Euclidean distance. It is well known that although this problem is NP-hard in general, gradient descent methods converge to a local optimum when seeded with an initial set of k examples drawn uniformly at random from X.

Minibatch k-means clustering algorithm works through the following number of processes [20]:

1. Given: k, mini-batch size b, iterations t, data set X
2. Initialize each $c \in C$ with an x picked randomly from X
3. $v \leftarrow 0$
4. for i = 1 to t do
5. $M \leftarrow b$ examples picked randomly from X
6. for $x \in M$ do
7. $d[x] \leftarrow f(C, x)$
8. endfor
9. for $x \in M$ do
10. $c \in d[x]$
11. $v[c] = v[c] + 1$
12. $\eta \leftarrow \frac{1}{v[c]}$
13. $c \leftarrow (1 - \eta)c + \eta x$
14. endfor
15. endfor

3.4 Related work

Mining Facebook is a challenging task because of the nature of the status, its users would be able to share status updates up to 63,206 characters long, that's a little harder to visualize than

Twitter's 140-character limit. Researchers have done some studies and experiments on twitter data with different findings and outcomes.
In [1] were presented a number of feature transformation methods such as Latent Semantic Indexing (LSI), Probabilistic Latent Semantic Analysis (PLSA), and Non-negative Matrix Factorization (NMF). These techniques are available to improve the quality of the document representation and make it more amenable to clustering. Therefore, it is critical to select the features effectively, so that the noisy words in the corpus are removed before the clustering.

In another study [2], was estimated the probability of concepts using a naive Bayes model:

$$P(c_k|E) = \frac{P(E|c_k)P(c_k)}{P(E)}$$

In this case, the concept with the largest posterior probability is ranked as the most possible concept to describe the observed instances.

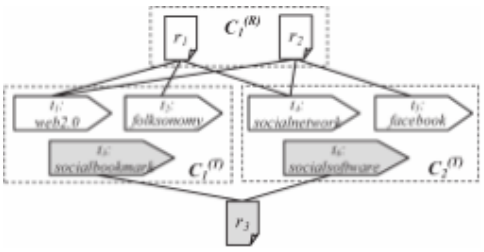
Researches in [3] begins with K-means clustering algorithm, they used two variations of distance measures: one is derived from cosine based similarity and the other is derived from Jaccard similarity coefficient. They used propagation to get the clusters. The comparison of two proposed distance measures is given in the table below. Jaccard based measure does better than cosine based measure, especially in K-means.

	Cosine-based	Jaccard-based
K-means	10.25%	6.61%
Affinity Propagation	2.95%	3.29%

A conference paper [4], presents a hybrid recommendation system which is indeed a social network-based collaborative strategy. Also in this paper is used the bag of words approach just like in ours.

They use the advantage of Wikipedia ontology to embed semantic information into social networks profiles.

Another clustering method was performed on Facebook data in article [5]. They propose an extended K-means approach based on the tripartite network of social tagging. The cluster assignment of resources, users, and tags can be represented by three matrices, where each row of the matrix represents a node, and each column of the matrix represents a cluster.



Distance between a resource node and the centroid of a resource cluster is influenced by the cluster structure of the tag nodes.

An interesting idea in paper [6] is the use of bridges in order to further improve the classification accuracy. The work in [6] proposes a method which can perform the clustering with the use of both content and structure information. Specifically the method constructs a new graph which takes into account both the structure and attribute information.

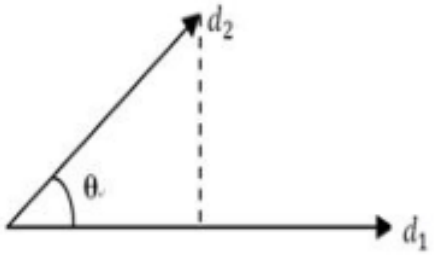
On the other hand [8] combines lexical correspondence analysis and clustering methods to better visualize and highlight the similarities between words/texts. All the samples in their dataset are compared with each center by means of the Euclidean distance and assigned to the closest cluster center.

Paper's [9] methods involve the following three steps:

1. Partition the documents into n possibly overlapping sub-collections with fixed or variable time interval
2. Extract the most salient themes $\theta_i = \{\theta_{i,1}, \theta_{i,2}, \dots, \theta_{i,k}\}$ from each sub-collection $C_i$using a probabilistic mixture model
3. For any themes in two different subcollections, $\theta_1 \in \Theta_i and\ \theta_2 \in \Theta_j$ where i < j decide whether there is an evolutionary transition based on the similarity of $\theta_1 and\ \theta_2$.

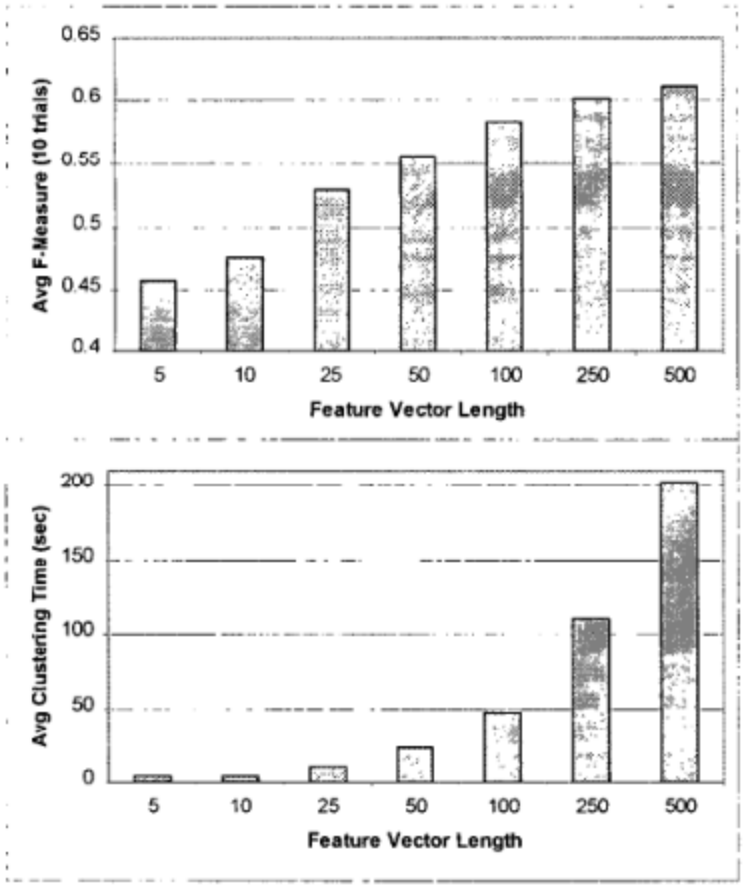
A similar survey [11] provides two basic methods to calculate feature vectors: (a) term frequency (TF) and (b) inverse document frequency (IDF). This survey just like ours uses k-means algorithm.

Paper [13] is somewhat similar with paper [4]. The bag of word model is used in information retrieval and text mining. With documents presented as vectors, they measure the degree of similarity of two documents as the correlation between their corresponding vectors, which can be further quantified as the cosine of the angle between the two vectors.



Angle between documents

Another study on clustering was done by [14]. They use k-means algorithm, and remove stop words automatically, just like us. To compare a document to a cluster, they simply calculate the cosine between the document vector and the cluster's centroid vector.



The effects of feature vector length on clustering quality and time

The only difference between paper [15] and other papers is the path algorithm. The use Dijkstra's shortest path algorithm to calculate the shortest between v_i and all other vertices.

Study [16] is almost like ours, the only difference is that they cluster and classify twitters messages (tweets). They also used k-means algorithm, and have had the same problem as us. They observed that a large fraction of the tweets in their dataset contain embedded URLs. Surprisingly, they find that the inclusion of URL-based text actually hurts the performance of our classifiers, in terms of precision as well. Most of the time URLs happen to be off-topic, and a lot of them are simply spam, so they decided to remove data which contains URLs.

4 Experimental Analysis

This chapter of the report delivers a clear discussion and analysis about the results of the experiments on the Facebook posts datasets, and how are the results addressing the purpose of this research.

4.1 Conclusion

Facebook is an exciting and emerging field of research, conducting text mining techniques on Facebook posts is a challenging task.
The task of developing perfect strategies for classification of varied forms and types of documents (in our case Facebook posts) for a near optimal solution or finding accurate ways of assessing the quality of the performed clustering though is impossible and is increasing in its complex nature, the field today deals with extraordinary tasks.
By using text mining combined with k-means clustering algorithm and cosine similarity measure or Euclidian measure, and after conducting a number of experiments, these papers showed a process for clustering Facebook posts.
We noted that k-means (and minibatch k-means) are very sensitive to feature scaling and that in this case the IDF weighting helps improve the quality of the clustering by quite a lot. Also as k-means is optimizing a non-convex objective function, it will likely end up in a local optimum. Several runs with independent random in it might be necessary to get a good convergence.

4.2 Future work

We believe that there is a lot of scope for interesting research directions people can take with respect to this problem.
A comparative study could be conducted in comparing the results generated by k-means algorithm and minibatch k-means algorithm with other clustering algorithms. Also, previous studies mentioned in related work have proven that Jaccard similarity measure can generate competitive results on text mining, so it will be interesting to compare the results obtained by Jaccard with the results of cosine similarity.
In the future, researchers aim to develop both quantitative and qualitative approaches or evaluating the clustering results of users and tags.
We feel that clustering Facebook users could be extended in a way where a new user could decide to follow only those users that post content that is of any interest to him. Similarly the work could be extended to cluster hashtags. This will help a user to follow the topics he is interested in.
For future work, we would like to use a faster implementation of k-means algorithm such as the one that uses coresets to quickly determine clusterings of the same point set for different values of k.

5 References

[1] Charu C. Aggarwal, ChengXiang Zhai. A survey of text clustering algorithms. 77-128

[2] Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hongsong Li, Weizhu Chen. Short Text Conceptualization Using a Probabilistic Knowledgebase. *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence.*

[3] Aniket Rangrej, Sayali Kulkarni, Ashish V. Tendulkar. Comparative Study of Clustering Techniques for Short Text Documents. *March 28–April 1, 2011, Hyderabad, India.*

[4] Elnaz Davoodi, Mohsen Afsharchi, and Keivan Kianmehr. A Social Network-based Approach to Expert Recommendation System.

[5] Caimei Lu, Xiaohua Hu, Jung-ran Park. Exploiting the Social Tagging Network for Web Clustering. *IEEE transactions on systems, man, and cybernetics-Part A: Systems and Humans, Vol.41, No.5, September 2011.*

[6] Charu C. Aggarwal, ChengXiang Zhai. Text mining in social networks. 353-378

[7] Anand Karandikar. Clustering short status messages: A topic model based approach.

[8] Domenica Fioredistella Iezzi. A new method for adapting the K-means algorithm to text mining. *Statistica Applicata - Italian Journal of Applied Statistics Vol. 22 (1)*

[9] Qiaozhu Mei, ChengXiang Zhai. Discovering Evolutionary Theme Patterns from Text - An Exploration of Temporal Text Mining.

[10] Charu C. Aggarwal, ChengXiang Zhai. A survey of text classification algorithms. 163-222

[11] Rizwana Irfan, Christine K. King, Daniel Grages, Sam Ewen, Samee U. Khan, Sajjad A. Madani, Joanna Kolodziej, Lizhe Wang, Dan Chen, Ammar Rayes, Nikolaos Tziritas, Cheng-Zhong Xu, Albert Y. Zomaya, Ahmed Saeed Alzahrani and Hongxiang Li. A survey on text mining in social networks. *The Knowledge Engineering Review, Vol. 30:2, 157–170. Cambridge University Press, 2015 doi: 10.1017/S0269888914000277.*

[12] Florian Beil, Martin Ester, Xiaowei Xu. Frequent Term-Based Text Clustering.

[13] Anna Huang. Similarity Measures for Text Document Clustering.

[14] Bjornar Larsen and Chinatsu Aone. Fast and Effective Text Mining Using Linear-time Document Clustering.

[15] Aron Culotta, Ron Bekkerman, and Andrew McCallum. Extracting social networks and contact information from email and the Web.

[16] Kevin Dela Rosa, Rushin Shah, Bo Lin, Anatole Gershman, Robert Frederking. Topical Clustering of Tweets.

[17] Annaluri Sreenivasa, S. Ramakrishna. Nomenclature and Contemporary Affirmation of the Unsupervised Learning in Text and Document Mining. *Global Journal of Computer Science and Technology: C Software & Data Engineering. Volume 15 Issue 2 Version 1.0 Year 2015. Type: Double Blind*

Peer Reviewed International Research Journal. Publisher: Global Journals Inc. (USA). Online ISSN: 0975-4172 & Print ISSN: 0975-4350.

[18] Hamadeh, Moutaz Wajih. Using Text Mining and Clustering Techniques on Tweets to Discover Trending Topics in Dubai.

[19] Chan, Yat-ling. An optimization algorithm for clustering using weighted dissimilarity measures

[20] D.Sculley. Web-Scale K-Means Clustering.