Data Mining and Community Detection in Social Networks

*Benhaddouche Djamila[1], Bensaoula Hadjer Imene[2], Foulani Fatima Zohra[3]
[1,2,3] *Computer Science department, Faculty of Mathematics and Computer Science, University of Science and Technology of Oran Mohamed BOUDIAF, Oran, Algeria*
[1]*benhaddouche.djamila@gmail.com*
[2]*hadjerimene95@gmail.com*
[3]*foulanifatimazohra@gmail.com*

Abstract

Social media analysis is a widespread area that attracts the attention of many data mining experts. In this context, community is one of the most common and important properties to reveal the hidden structures of a network. Girvan and Newman define a community as a set of entities that have internal relationships more than external relationships. Radicchi and al. Improve this definition by the constraint that each individual of a community has more neighbors inside his community than outside. They call these structures strong communities. The main problem related to the detection of communities are the definition of these groups, their detection in practice which is a more algorithmic problem and how to validate that the results are indeed relevant. Several definitions exist, but one, based on a function called modularity, is more generally used today than the others. In this work we had two visions, at first we aimed to compare two of the most famous community detection modularity based algorithms which are LOUVAIN and GIRVAN&NEWMAN, in which we found that Louvain is superior than GIRVAN&NEWMAN in terms of modularity. Secondarily, we intended to employ the community detection in a decision making context by applying the RFM (Regency, Frequency, Monetary) method.

Keywords: *Social network analysis, community detection, Modularity, Louvain, Girvan Newman, K-means, Regency Frequency Monetary Method*

1. Introduction

To evaluate the quality of a partition, one generally uses a so-called quality function giving a score to a partition and which formally captures the intuition given previously. Then, it "remains" to maximize this function to find the best partition for a given graph. There are several quality functions, the most used being modularity. Definition For a partition of the set of nodes of a graph G = (V, E), noting L =—E—the number of links of the graph, ds the total degree of a part s of and ls the number of links inside s, the modularity Q is defined by:

$$Q\ \underset{S}{(\pi)} = (ls/L - (ds/2L)2)$$

This magnitude is the sum, over all the communities, of the differences between the proportion of links inside the community s (ls/L) and the proportion of links that a community should have in a random graph with the same degree distribution (i.e. (ds/2L)2). Modularity is between -1 and 1, so a partition will be good if there are significantly more links inside the communities than expected and therefore fewer links outside the communities. Finding the partition(s) that maximizes modularity is an NP-hard problem and many heuristics.

Modularity is used by several algorithms as a function of quality by optimizing or modifying it such as the Algorithm of Girvan and Newman, the Algorithm of Louvain which is based on greedy techniques to provide acceptable solutions in computing time.

2. Louvain

2.1 Principle

This approach is composed of a set of passes, each composed of two phases, which are repeated iteratively until a local maximum of modularity is obtained. The algorithm starts from an undirected weighted graph with N vertices.

2.1.1. First phase: The initial partition consists in placing each vertex in a distinct community; this partition is therefore composed of N communities for a graph of N vertices. We then consider the first vertex X, and calculate the modularity variation that is obtained by removing X from its community and placing it in the community of one of its neighbors. This variation is calculated for each of X neighbor's .By the end vertex X will be placed in the community for which this variation is maximum, but only if it is positive. If all the gains are negative, then vertex X is placed back in its community of origin. This process is applied sequentially on all vertices and this is called iteration. The process is then re-applied to all vertices repeatedly until no more vertex is moved during a full iteration. After this first phase, the graph will be divided into a partition P having C communities. If this first phase grouped vertices, the algorithm moves on to the second phase, otherwise the algorithm is terminated and the result is the partition P.

2.1.2. Second phase: The second phase is all about building a new graph whose vertices are the communities discovered during the first phase. For this, the weight of the links between these new vertices that represent the obtained communities is given by the sum of the weights of the links that existed between the vertices of these two communities. The links that existed between vertices of the same community create loops on this community in the new graph. Once this second phase is finished, it is possible to apply the first phase of the algorithm again on the weighted graph and to iterate. This constitutes a pass and with each pass the modularity grows by aggregating nodes into larger and larger communities. The last partition found is the one with the best modularity

2.2. Communities detection in the deezer social network using the Louvain Algorithm

2.2.1. Identifying the datasets: The used data was collected from the music streaming service Deezer (November 2017). These data sets represent user friendship networks in the form of a list of mutual friendship links in a csv file where each link is defined by the departure node and the arrival node. The nodes used are indexed (starting from 0) in order to achieve a certain level of anonymity.

2.2.2. Applying the Algorithm: We applied the Louvain algorithm on a graph with 1000 friendship links, using python. The main steps were:

1. Load data from a csv file.

2. Create the initial graph from the loaded data.

3. Target the best partition by invoking the ”Louvain best partition function”.

4. Calculate the modularity for the obtained partition

2.2.3. Visualize the obtained results: The obtained modularity is: 0.8066410456982845, the number of community is: 27

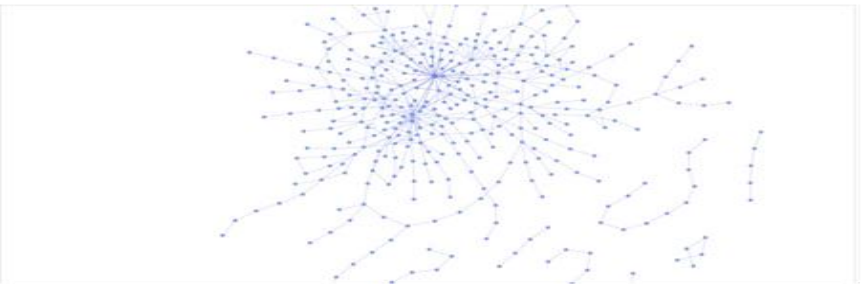

Figure 1. Results of the Louvain algorithm on a Twitter dataset

2.3. Communities detection in the Twitter social network using the Louvain Algorithm

2.3.1. Identifying the datasets: In this example, we will apply the Louvain algorithm on a real datasets where we will target the comments that were published on Twitter in order to determine the topics that were mostly discussed and identify the ones that relate to them.

2.3.2. Applying the Algorithm:

1. Load data
2. Unlike the dataset used in the previous example let’s first go through “data preprocessing”
3. Apply the Louvain Algorithm
4. Calculate the modularity.

Pre-processing is implemented to avoid incomplete data, data breaks, and inconsistent data. The text pre-processing steps in our study include:

1. Removal of URLs (http://www.situs.com) and emails (name@situs.com).
2. Removing special characters in Twitter, this process is done by removing the hashtags, usernames (@user-name) and special characters (e.g. RT, which indicates that the user is retweeting something).
3. Removal of symbols. This step is performed to remove symbols and punctuation marks from the tweet.
4. Removal of stop words which are words that do not affect the classification process such that (i, herself, if, such ...).

5. The tokenization process consists of cutting the input string according to each letter. An example of a tokenization method is the N-gram which is a probability model that was originally devised by Russian mathematicians in the early 20th century and then developed to predict the next word or character in a sequence of strings. The strings can be in the form of characters or words depending on the needs of the application. In this study, the uni-gram character is used as a tokenization method.

3. Girvan Newman

3.1. Principle

The GN algorithm is a community detection algorithm based on dividing hierarchical clustering. Typically, this method first views the network as a community and then divides the community into multiple smaller communities through hierarchical segmentation. The GN algorithm is one of the most classical methods, which uses edge mediators. Since the edge with high edge mediators will be removed in each iteration, the speed is much faster than the random edge removal speed.

3.1.1. Concept of betweenness: Betweenness was introduced by Newman and Girvan (2004) and has been widely used. The degree of betweenness of an edge is proportional to the number of shortest paths in the graph (for all pairs of nodes) that pass through this ridge. This quantity can be computed for all edges in O (mn) time on a graph of n nodes and medges. There are other definitions of the degree of betweenness based on random walks or on flows which give similar results.

3.1.2. Modularity: The most widely used quality measure (Pons (2007)) is modularity, introduced by Newman and Girvan (2004). We distinguish the edges internal to the communities from the edges connecting nodes of different communities. If we havec communities, we can define D as the matrixcxc whose elements dij give the proportion of edges connecting nodes of community i to those of community j. The diagonal terms dii give the proportion of edges internal to cluster i among all the edges of the graph. Modularity is then defined

As:

$$M = \sum_{i} (d_{ii} - (\sum_{j} d_{ij})^2)$$

3.2. Algorithm of Girvan and Newman (GN)

Girvan and Newman’s algorithm is a divisive method operating according to the following principle:

1. Reading the input dataset and creating the starting graph GN on Deezer dataset
2. Compute the betweenness of all edges
3. Delete the edge with the strongest betweenness.
4. Recalculate the betweenness between all nodes affected by the deletion.
5. If there are any bones left, start over at step 2.

3.2.1. Visualization of the results obtained:

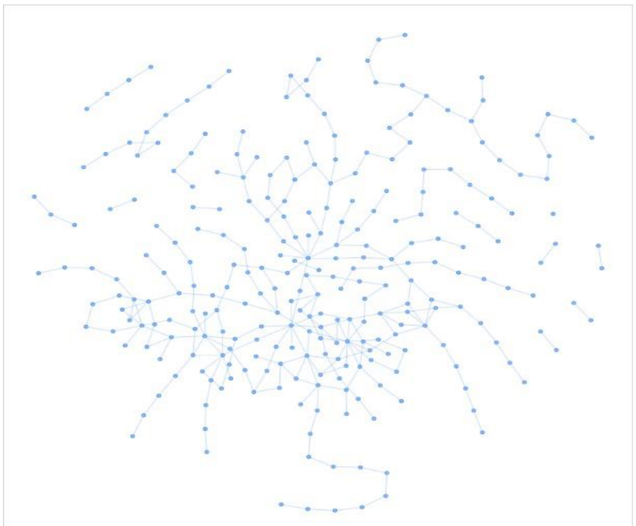

Figure 2: Results of the Girvan Newman algorithm for Twitter

The negative point: modularity is used to determine the optimal number of partitions. However, this method is relatively slow; the worst-case complexity is O (m2n), or O(n3) for a weakly connected network (m∞not).

3.3. GN Algorithm Improvements

3.3.1. Step 01: 1. The only difference arises on the reading of the input file and the creation of the starting graph

3.3.2. Step 02: 2. We start first by computing the betweenness of the edges of each edge of the graph.
3. Select a node X and run BFS to find the node's shortest path count X to each node, and assign the numbers as a score to each node.
4. Starting from the leaf nodes, calculating the edge credit (w) by: 1 + (sum credits of edge at nodes) *(score of destination node / score of departure node).
5. Calculate the edge credits of all the edges of the graph G.
6. Repeat from step 1, until all nodes are selected.
7. Sum all the credits (w) calculated in step 2 and divide by 2 and the result is betweenness ridges.

3.3.3. Step 03: Next, let's remove the edges with the highest betweenness and repeat until we find the right community distribution.
8. Delete the edge that has the highest edge betweenness
9. Calculate the modularity Q of the divided communities
10. Compare Q obtained with the Qbest.
11. Repeat from step 1 until the graph contains no edges.

3.3.4. **Visualization of the results obtained:** The obtained modularity is: 0.7514096325218268, the number of community is: 25

4. Community detection and decision making

The challenges of community detection methods are major for marketing, and particularly for digital marketing. With the development of techniques for producing, retrieving and storing data, the mass of information has never been so large and paradoxically has never been so hard to interpret. Segmentation is a process that makes it possible to classify customers based on similarities, clustering aims to find similarities within customers in order be able to group them. The principle of segmentation is to separate customers according to deterministic criteria. The importance of these models is crucial for marketers, since the discovery of a new cluster of customers

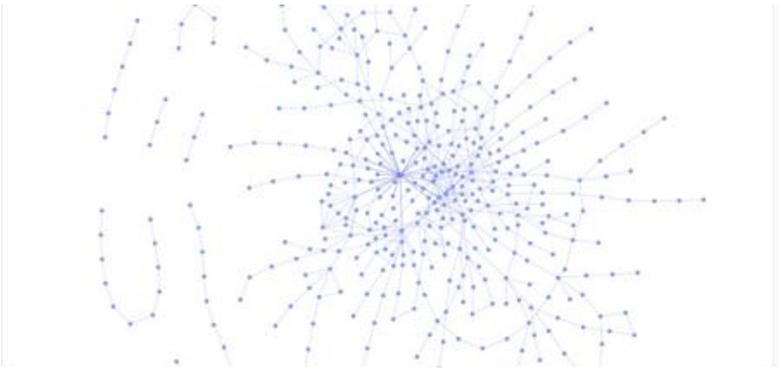

Figure 3: Results of the Girvan Newman (improved) algorithm for Twitter

And the associated purchasing habits can make it possible to save money by targeting this cluster more finely and thus acquiring a new market and better knowing its customers. Among the segmentation methods used, there is RFM segmentation.

4.1. RFM Segmentation:

RFM segmentation for Regency, Frequency and Amount, is widely used in the field of direct sale. Indeed, this approach makes it possible to take purchasing behavior into account. It permits establishing homogenous customer segments which will be useful for targeting offers.

4.2. Principe (RFM)

Regency value: This is the time elapsed since a customer’s last interaction with a brand, which can include visiting a website, using a mobile app, etc. Regency is a key metric because customers who have interacted more recently are more likely to respond to new marketing offers.

Frequency value: This is the number of times a customer has made a purchase or otherwise interacted with the site during a given period. Frequency is a key metric because it shows how engaged a customer is. A higher frequency indicates a higher degree of customer loyalty.

Monetary value: This is the total amount that a customer has spent on products and services over a given period.

Each of these RFM metrics is important for predicting future customer behavior and increasing revenue.

Who made a purchase in the recent past are more likely to do so in the near future. Those who interact with the brand more frequently are more likely to do so again soon. And those who spent the most are more likely to be big spenders in the future.

4.3. Online Retail (K-Means)

4.3.1. Preview (Database): E-Retail is a cross-national dataset that contains all transactions between 01/12/2010 and 09/12/2011 for registered UK-based non-store e-retail. The company mainly sells unique gifts for all occasions. Many of the company's customers are wholesalers.

4.3.2. Calculates RFM values: M (Money): Total amount of transactions To calculate the amount, we multiplied the quantity with the unit of space for each purchase and after that we did the grouping by customer ID in order to add up the total amount of all purchases for each customer.

F (Frequency): Number of traces to calculate the frequency, we count the number of visits to the website by grouping the data by identifier and doing the count.

R (Regency): number of days since last purchase
To calculate the regency, we start by calculating the last transaction date and after that we calculate the difference between the max date and the transaction date of each customer. And finally we grouped all the differences for each consumer to take the minimum value that represents the regency.

4.4. Data cleaning

4.4.1. Detection and removal of outliers: An outlier is an extreme value that is abnormally different from the distribution of a variable. In other words, the value of this observation differs greatly from other values of the same variable.

4.4.2. Importance of detecting an Outlier: Several machine learning algorithms are sensitive to training data as well as their distributions. Having Outliers in the Training Set of a Machine Learning algorithm can make the training phase longer. Not to mention that the learning will be biased. Consequently, the predictive model produced will not perform well, or at least, far from optimal. In this case, we end up with an extreme and rare data, but plausible. Deciding whether to keep the observation or not is trickier. Indeed, it is necessary to know in which context this information will be used.

4.4.3. Outlier detection – IQR approach: IQR stands for Interquartile Range. It measures the statistical dispersion of data values as a measure of the overall distribution. The IQR is equivalent to the difference between the first quartile (Q1) and the third quartile (Q3) respectively. Q1 refers to the first quartile or 25% and Q3 refers to the third quartile or 75%.

Using IQR, we can follow the approach below to remove outliers:

1. Calculate the first and third quartiles (Q1 and Q3).
2. Also, evaluate the interquartile range, IQR = Q3-Q1.
3. Estimate lower bound, lower bound = Q1*1.5
4. Estimate upper bound, upper bound = Q3*1.5
5. Delete data points outside the lower and upper bounds.

4.5. Community detection using K-means

4.5.1. Dendrogram and hierarchical clustering:The dendrogram is a hierarchical grouping diagram, allowing to organize data in a tree structure according to their similarities. Hierarchical clustering involves creating a cluster tree to represent data. Within this tree, each group or “node” is related to two or more successor groups.

The groups are nested together and organized in the form of a tree. Each node in the tree contains a group of similar data, and the nodes are grouped based on their similarities.

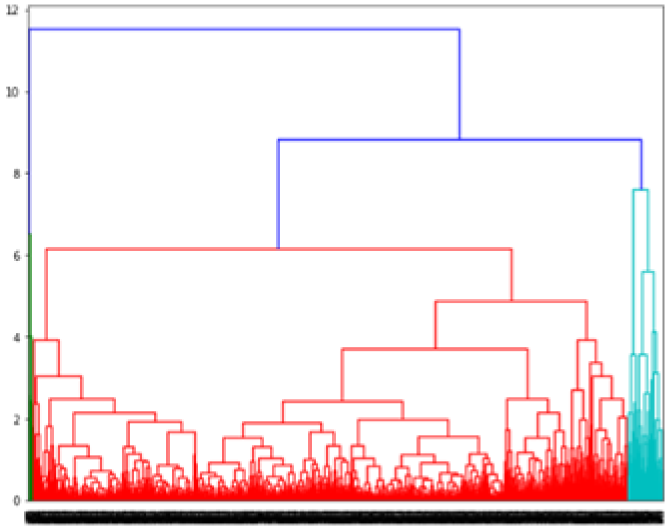

Figure 4: Dendrogram Capture

4.5.2. Kmeans: It is one of the most popular clustering algorithms. It allows you to analyze a set of data characterized by a set of descriptors, in order to group “similar” data into groups (or clusters). The similarity between two data can be inferred thanks to the “distance” separating their descriptors; thus two very similar data are two data whose descriptors are very close. This definition makes it possible to formulate the data partitioning problem as the search for K “prototype data, around which other data can be grouped. These prototype data are called centroids; in practice, the algorithm associates each datum with its nearest centroid, in order to create clusters.

On the other hand, the means of the descriptors of the data of a cluster, define the position of their centroid in the space of the descriptors: this is at the origin of the name of this algorithm (K-means).

After initializing the centroids by taking random data from the dataset, K-means alternates these two steps several times to optimize the centroids and their clusters:

1. Group each object around the nearest centroid.
2. Replace each centroid according to the average of the descriptors of its group.

After a few iterations, the algorithm finds a stable breakdown of the dataset: the algorithm is said to have converged.

4.5.3. Important factors we need to consider while using K-means algorithm: Number of clusters (K): The number of clusters in which we want to group the data points must be predefined and that is why we used the dendrogram to define the number of communities to be used in K Means.

Outliers: Community formation is very sensitive to the presence of outliers. Outliers pull the cluster towards itself, thus affecting the optimal formation of the cluster, which is why we used the IQR method to remove outliers.

4.6. Results

Clients with cluster ID 1 are the clients with a high number of transactions compared to other clients.

Customers with cluster ID 1 are frequent buyers.

Customers with cluster ID 2 are recent but infrequent buyers and therefore less commercially important.

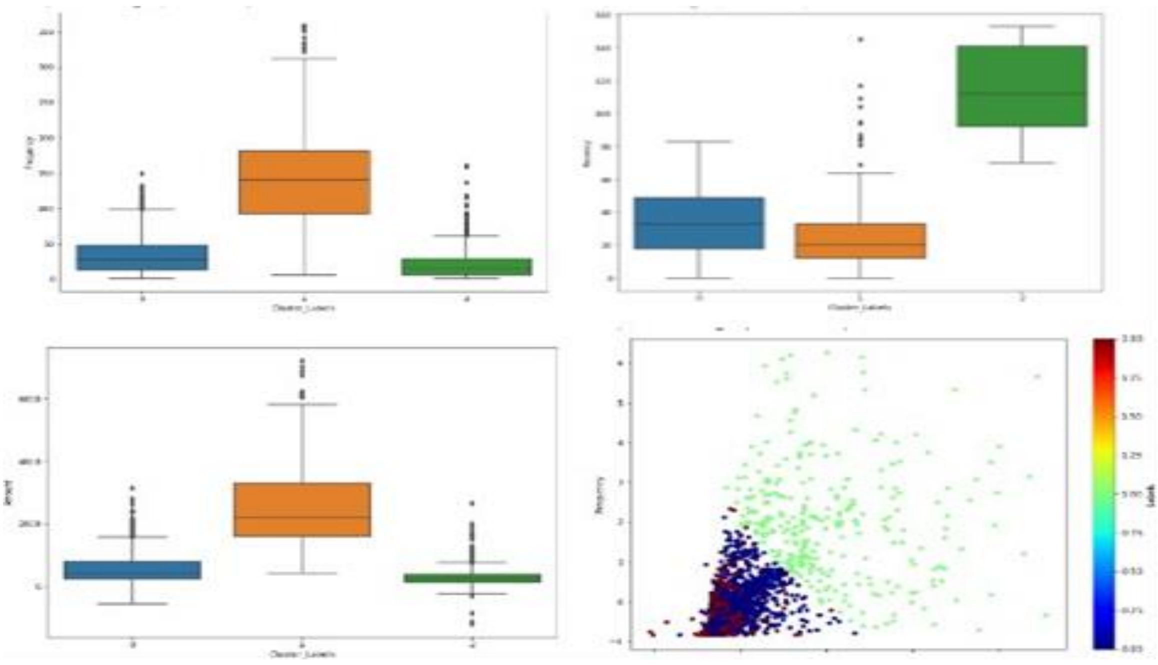

Figure 5. Community Display after Detection using Kmeans.

References)

[1] *W.Loua and N. Loua , “Analysis of social networks and detection of communities.” these Master, Univ 08 Mai 45,Guelma, 2019.*

[2] *M. E. J. Newman and M. Girvan, ”Finding and evaluating community structure in networks,” Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys., vol. 69, no. 2 2, pp. 115, 2004.*

[3] *A. Lancichinetti, S. Fortunato, and F. Radicchi, ”Benchmark graphs for testing community detection algorithms,” Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys., vol.78,no.4,pp.15,2008.*

[4] *X. Ma and D. Dong, Evolutionary Nonnegative Matrix Factorization Algorithms for Community Detection in Dynamic Networks,” IEEE Trans. Knowl. Data Eng., vol. 29, no. 5, pp. 10451058, 2017.*

[5] *S. Gupta and P. Kumar, An overlapping community detection algorithm based on rough clustering of links,” Data Knowl. Eng., vol. 125, p. 101777, 2020.*

[6] *S. Fortunato, V. Latora, and M. Marchiori, Method to nd community structures based on information centrality,” Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top., vol. 70, no. 5, p. 13, 2004.*

[7] *Y. Li, C. Sha, X. Huang, and Y. Zhang,”Community detection in attributed graphs: An embedding approach,” 32nd AAAI Conf. Artif. Intell. AAAI 2018. 30*