



FRIEND RECOMMENDATION USING GRAPH MINING ON SOCIAL MEDIA

Kosaraju Naren Kumar¹, Kanakamedala Vineela²

¹ Student & Sathyabama University.

² Assistant Professor & NRI Institute of Technology.

¹ naren010898@gmail.com.

² vineela.nriit@gmail.com.

Abstract:

Recommendation system is an important type of machine learning algorithm that provide precise suggestions to the users. Recommendation systems are used in innumerable types of areas such as generation of playlists, music and video services like Jio savaan, wynk, amazon prime music etc., and products recommendation for users in e-commerce applications and commercial applications. The recommendations that are provided by various types of applications increases the speed for identifying and makes easier to access the products that users are interested in. For each user, the recommendation system is capable of envisaging the future predilections on a set of items and recommend the top items. In several industries, recommendation systems are very useful as they generate huge amount of income and this type of industries can stand uniquely from competitors. Due to cumbersome number of items that each user can find in the web, the impact of recommendation system has been increased in the internet. Recommendation systems are used for custom-made navigation by getting huge amount of data particularly in social media domain for recommending friends. A recommendation system act as a subclass for the information filtering system that pursue to predict the rating. The similarity measures that are calculated in this research are Jaccard distance and Otsuka-Ochiai coefficient. The feature extractions that are used in this paper are Adar index, PageRank, Katz centrality, Hits score. Now a days many

research people are implementing different types of algorithms in various domains for recommendation systems.

Keywords: Adar Index, In-degree, Jaccard Distance, Katz Centrality, Out-degree, Social Network Recommendation system.

1. INTRODUCTION

The recommendation system is used to recommend the user based on their preferences. In recent times social media is enjoying a great deal of success with a million of users visiting many sites like Facebook, Twitter etc. for social networking. By using computational methods such as natural language processing, data mining machine learning etc., social recommendation system involves the investigation of collective intelligence of data from wikis, query locks, Q&A communities etc. The information that is very much interested by the users is suggested by the recommendation system by using information filtering techniques. Social recommendation system is a system that recommends the friends in social media applications such as the Facebook, Twitter, Instagram etc.

Over the last couple of years, for social media personalized recommendation systems are came into existence. For example, StumbleUpon is a customized recommendation system which suggests web pages for



the users based on the ratings given by the users, rating given by the user of users with similar interests and topics. It recommends the friends to the users by knowing the followers and followee of a particular user. Using a dataset of 9437519 nodes of both source and destination, in that entire volume of data 80% is used for training and 20% is used for testing for future predictions. In this model the classification algorithm used is Random Forest Classifier. Random forest algorithm is a tree based algorithm, so it will work well for dimensional data and nonlinear separable data. Accuracy score is calculated by the precision, recall and F1 score. For describing the performance of the test data in a classification model confusion matrix is used.

II. STATE OF ART

1). Snigdha Luthra et.al (2019) approach the social hub network is dynamic because it changes the structure at completely different timestamps. The network obtained at time t is varied at time $t+1$. so as to predict the continuing changes on network, graph embedded techniques square measure won't to acquire associate unattended graph with completely different parameters of nodes and edges which might be utilized in machine learning ways. For this experiment community detection, formula performs bunch technique to cluster nodes along in the same cluster that has a similar edge betweenness issue. The projected framework decides that more connections may be established supported nodes happiness to a similar cluster.

2). Ivana Andjelkovic et.al (2019) proposed a recommendation system for musical artists which acquaint with a novel collaborating visualization of moods and the artist. Within the visualization via manipulation of the avatar the system provisions control and explanation of the recommendation system. Implementation and design of an online experiment is

obtainable to estimate the system through four conditions with interaction, control and unpredictable degrees of visualization. The results has shown that a certain combination of cooperative features and interface design progresses perceived and objective recommendation accuracy. As there is self-conveyed user fulfillment with the recommendation system which it makes the people to know the mood the artist's music which combined with the relevant interactivity in the music recommendation, can change the way for the accuracy of the recommendation.

3). Imane Belkhadir et.al (2019) presented an amalgamation of social regularization approach that incorporates the trust information and social network information to categorize a comprehensive trust path in social graph. The proposed recommendation system recommends the friends to the users based on the users who having the similar favors and tastes and recommends the experts to users in some field. Based on these conditions the system proposes matrix factorization frame work. The correlation between the users and items and the shortest path is calculated between the appropriate groups of friends that are huddled to get an accurate friends' recommendation. To restrict the framework of matrix factorization, tags and friendships are joined as the regularization terms.

4). Neha Verma et.al (2019) proposed a recommendation system to understand the user in the e-commerce websites like Flipkart, Amazon, and Netflix etc. The work flow of the recommendation system is divided into two segments such as the gathering information segment and the analysis segment. The proposed system builds the recommendation system using various techniques like traditional techniques, Hybrid techniques and the modern techniques and the information of the users is collected



from the buying, searching and habits etc. The collected information then used for analysis segment. Based on this analysis about the user and the recommendations are generated.

5). Swathi Sambangi et.al (2019) proposed a recommendation system for the user to choose the customized products of the users. The primary goal the proposed system is to endorse a one to one text-based review recommendation. The proposed system takes the data that was obtained from the Amazon API. The amazon recommendation system is one of the effective technologies that has a huge accomplishment on the information available on the internet. This technology supports the users to choose the customized product which increases the development rate of user satisfaction.

5)Peng Liu et.al (2019) dynamic graph based embedding model is proposed a real time social recommendation. The dynamic graph based embedding model (DGE) takes the user behaviour pattern, social relationships and temporal semantic effects in an amalgamated way. To take the semantic effect of the edges, the probability matrix is formulated. To generate a top recommendation in large scale social media, an incremental learning algorithm and query processing techniques are developed. The proposed recommendation system is based on the contiguity of related users and items although considering the sparkle of the items. Estimation of the toe real world datasets demonstrate the efficiency of the proposed approach.

III.ARCHITECTURE OF PROPOSED SYSTEM

A). DATASET OVERVIEW

Dataset is taken from the Facebook's recruiting challenge on Kaggle and the dataset comprehends two columns such as Source and Destination. Dataset link -

<https://www.kaggle.com/c/FacebookRecruiting>.

The volume the of dataset is approximately 94Lakhs, later on the data is fragmented into two types such as training data and testing data. 80% and 20% of the data is used for training and testing the data respectively.

Total no of nodes presents in the data: 1862220

Total no of edges presents in the data: 9437519

Data Columns Data type

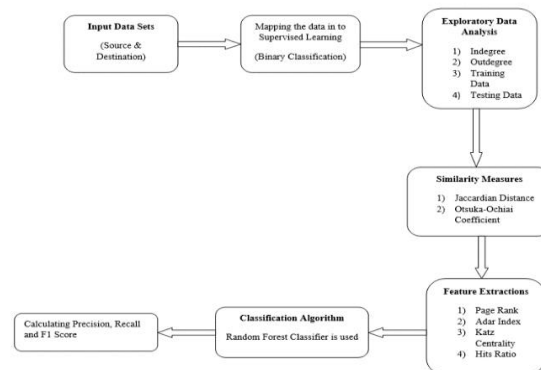
Source Node int 64

Destination Node int 64

The number of people that are common in both test and train data are: 1063125

The number of people that are not present in test data but present in train data are: 717597

The number of people that are not present in train data but present in test data are: 81498



FigA. Process for proposed Architecture.

B). MAPPING THE PROBLEM INTO SUPERVISED LEARNING PROBLEM

Generated training samples of good and bad links from given directed graph and for each link got some features like number of followers, is he followed back, page rank, katz score, adar index, some Svd features of adjacent matrix, some weight features etc. and trained ml model based on these features to predict link.



C). IN-DEGREE –

The number of edges directed into a vertex in a directed graph is called in-degree. In other words, it can be defined as number of incoming nodes for a particular node. In-Degree is used to find the number of followers for each user

The average In-Degree for the total data: 5.0679

D). OUT-DEGREE –

The number of edges directed away from the vertex in a directed graph is called out-degree. In other words, it can be defined as number of outgoing nodes for a particular node. Out-Degree is used to find the number of people each user is following.

The average Out-Degree for the total data: 5.0679

E). JACCARD DISTANCE –

$$j = \frac{|X \cap Y|}{|X \cup Y|}$$

Formula E)– Jaccard Distance

Jaccard distance is nothing but a measure of similarity between two data nodes ranging from 0% to 100%. As the Jaccard distance increases then there is a high chance of

existing edge between the two nodes. It is defined as the size of the intersection of two sets to the size of the union of the two sets. It's very sensitive towards small amount of data and gives flawed results.

F). COSINE-DISTANCE(OTSUKA-OCHIAI COEFFICIENT)

$$K = \frac{|A \cap B|}{\sqrt{|A| \times |B|}}$$

Formula F) – Otsuka-Ochiai coefficient

Otsuka-Ochiai coefficient is nothing but an intersection of the no of elements to the square root of the no of elements in A multiplied by number of elements in B.

G). PAGE RANK –

PageRank is an algorithm that was designed to rank the importance of web pages. Given a directed graph the PageRank algorithm will give each vertex (U_i) a score. The score represents the importance of the vertex in the directed graph. Networkx library is used to compute the PageRank.

H).CHECKING FOR SAME WEAKLY CONNECTED COMPONENTS-

If two users belonging to the same weakly connected components that gives the higher probability or higher chance of similar edge being present. Weakly connected component acts as a subgraph for the given directed graph. Weakly connected component acts as a strongly connected component in case of undirected graph.

I).ADAR INDEX–

$$A(x, y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{\log(|N(u)|)}$$

Formula I) – Adar Index

Adar index or Adamic index is nothing but an inverted sum of degrees of common neighbors for given two vertices. Networkx library is used to compute the Adar index.



J). KATZ CENTRALITY –

Based on the centrality of its neighbors Katz Centrality computes the centrality of a node. It is an inductive reasoning of the eigenvector centrality. The Katz centrality for node i is

$$x_i = \alpha \sum_j A_{ij}x_j + \beta,$$

Formula J) – Katz Centrality

Where A is the adjacency matrix of the graph G with eigenvalues λ. Networkx library is used to compute the Katz centrality.

K). HITS SCORE –

The HITS(Hyper Induced Topic Search) algorithm computes two no for a node. Based on the incoming links authorities estimates the node value. Based on outgoing links hubs estimates the node value. Networkx library is used to compute the HITS score.

L). WEIGHT FEATURES –

An edge weight value was calculated between nodes in order to find the similarity of nodes. As the neighbor count goes up edge weight decreases. Intuitively, consider one million people following a celebrity on a social network then chances are most of them never met each other or the celebrity. Whereas on the other hand, if a user has 30 contacts in his/her social network, the chances are higher that most of them know each other.

As it is directed graph, weighted in and weighted out are calculated separately.

$$w = \frac{1}{\sqrt{1 + |X|}}$$

Formula L) – Weight Features

M). SINGULAR VALUE DECOMPOSITION –

For factorization of matrix into singular values and singular vectors SVD (Singular Value Decomposition) algorithm is used. SVD features for both source and destination nodes SVD is widely used in machine learning reduction techniques and matrix calculations.

IV. RESULTS AND DISCUSSION

1) SUBGRAPH

The below figure 5.1, represents the directed graph obtained from the dataset. By this digraph number of edges and nodes are calculated.

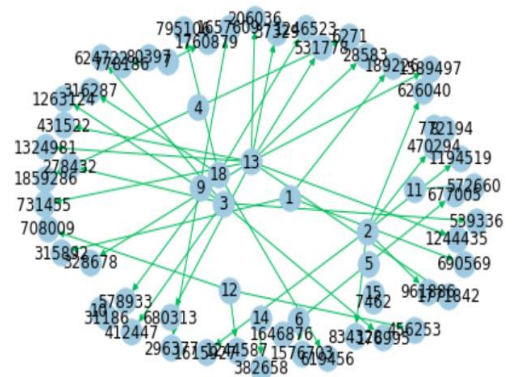


Fig Subgraph

The no of nodes in the subgraph are: 66

The no of edges in the subgraph are: 50

The average In-Degree of the above subgraph: 0.7576

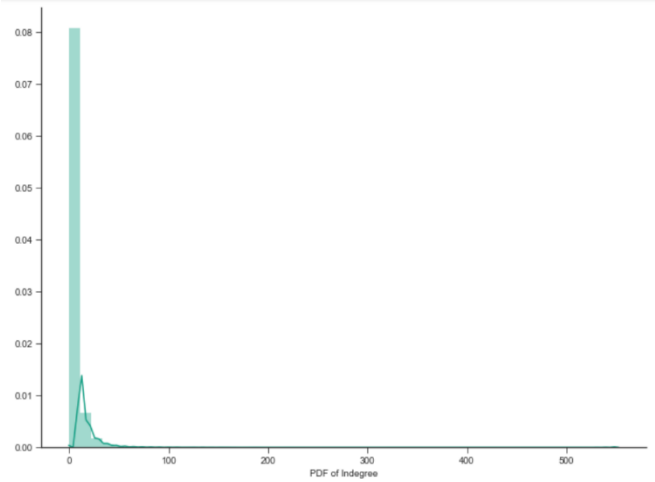
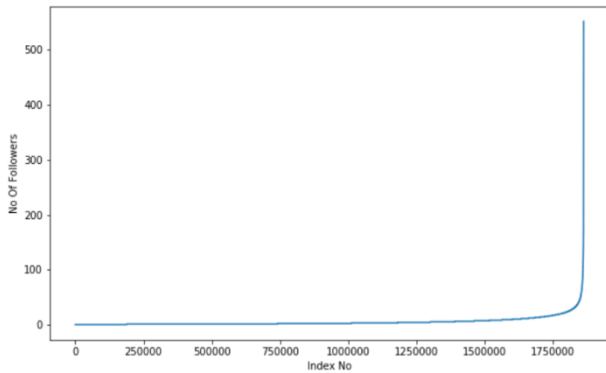
The average Out-Degree of the above subgraph: 0.7576

2). IN-DEGREE GRAPH

The below figure 5.2, represents the graph of In-Degree. With help of In-Degree graph we calculate the number of followers for each person.



Fig In-Degree



- 90% of the people are having only 12 followers
- 95% of the people are having only 18 followers
- 99% of the people are having only 40 followers
- 99.9% of the people are having only 112 followers
- 100% of the people are having only 552 followers

Probability density function is represented in the form of bar graph that is histogram

3).PDF of In-Degree Graph

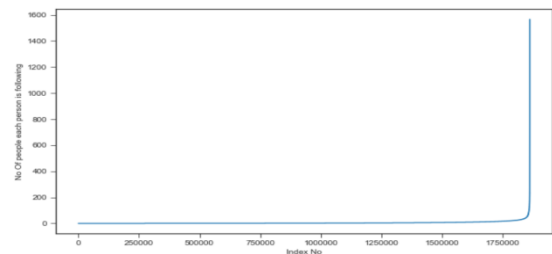
The below probability density function graph shows that very less number of people will be getting followed more number of followers.

Fig PDF of Indegree

4). OUT-DEGREE GRAPH

The below Out-Degree graph shows that number of persons that each person is following.

Fig Out-Degree

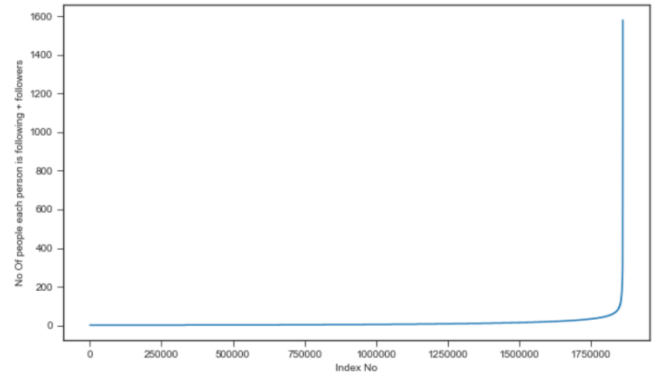


- 90% of the people are following only 12 persons
- 95% of the people are following only 19 persons
- 99% of the people are following only 40 persons
- 99.9% of the people are following only 123 persons



100% of the people are following only 1566 persons

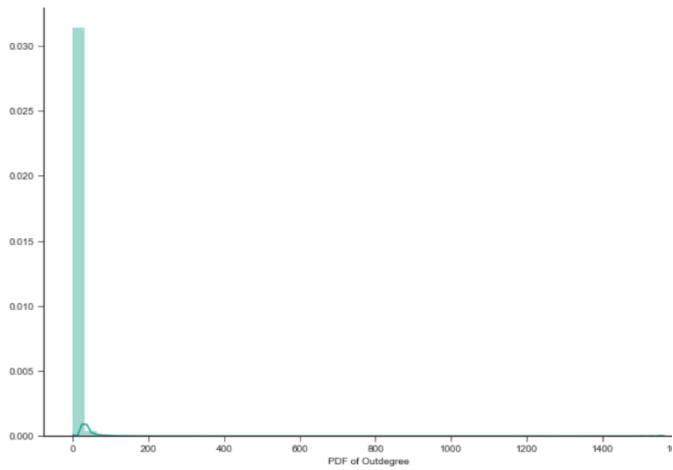
Fig In-Degree and Out-degree



5). PDF of Out-Degree Graph

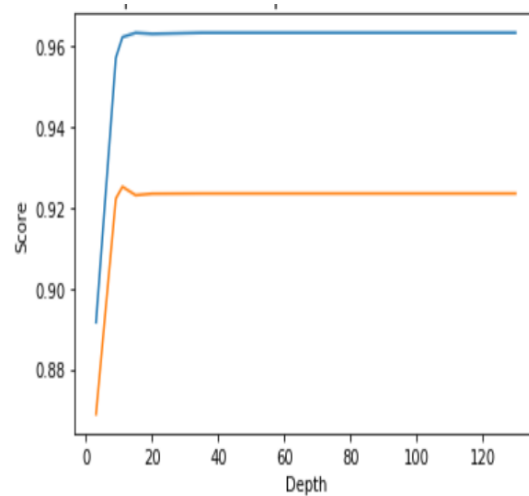
The below probability density function shows that very less number of people will follows the more number of people

Fig PDF of Out-Degree



- 90% of the people are having only 24 persons
- 95% of the people are having only 37 persons
- 99% of the people are having only 79 persons
- 99.9% of the people are having only 221 persons
- 100% of the people are having only 1579 persons

7) Depth and Estimator Graph



6). In-Degree – Out-Degree Combined Graph

The below In-degree and Out-Degree graph shows that number of people that each person is following and the number of followers that each person is having.

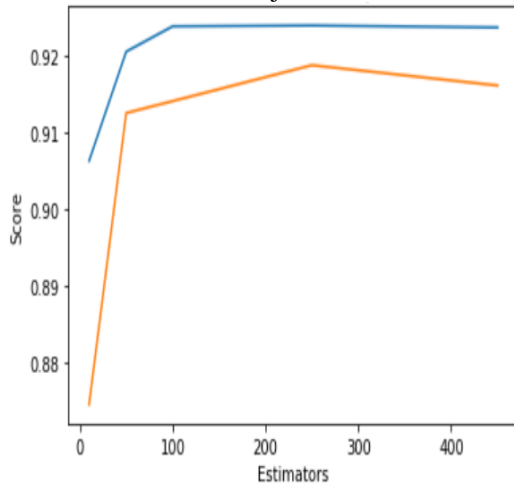
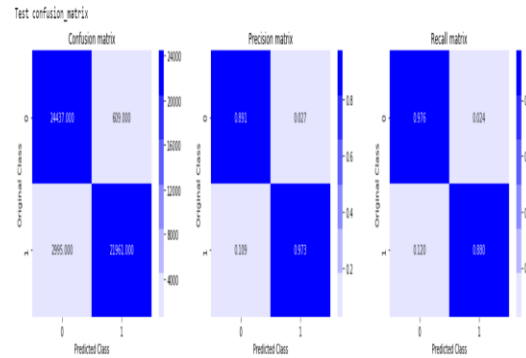


Fig Graphs after calculating F1 Score



From the above confusion matrix we infer that the accuracy score of the train matrix is 95% and accuracy score of test matrix is 89%.

8). Confusion Matrix Graph

Fig Train Confusion Matrix

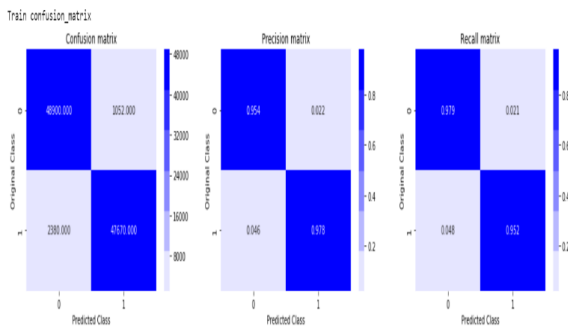
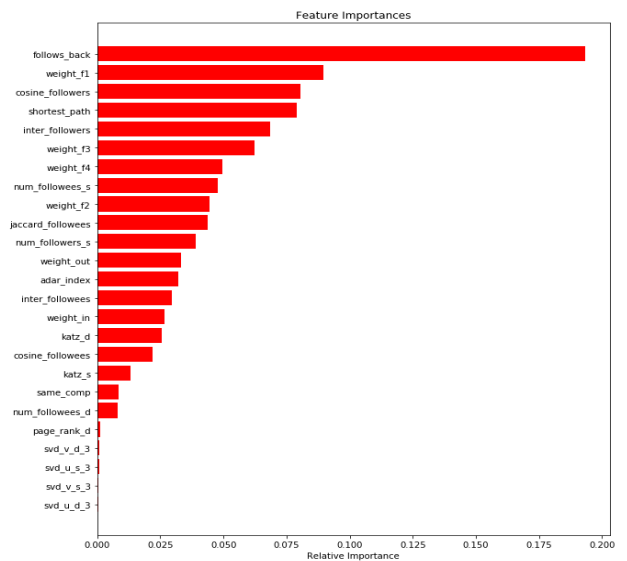


Fig Test Confusion Matrix.

9). Feature Extraction Graph

The below graph shows the differences and importance of the feature extractions that are calculated and obtained in a form of the bar graph

Fig Feature Extractions



From the graph we can say that follows_back feature is



most important extraction compared to other features and SVD(Singular Value Decomposition) is the least preferred one.

V. CONCLUSION AND FUTURE WORK

The proposed system develops a friend recommendation system to suggest friends to the users using Random Forest Classifier. With the help of the proposed system this recommendation model is working with the accuracy of 89%. This is quite reasonable for the hardware and the volume of data that we have. Our data set consists of 94 lakhs nodes. Performance metrics for this model is obtained by calculating Precision, Recall and F1 score. The most important feature extraction that we calculated is follows_back. feature. For the better results and accuracy Preferential attachment, SVM classifier and Graph neural networks can be used. This can improve the performance of the model in future.

REFERENCES

- [1] D Davis, R Lichtenwalter, N V. Chawla, "Multi-relational link prediction in heterogeneous information networks[C]," (Advances in Social Networks Analysis and Mining (ASONAM) 2011 International Conference on, pp. 281-288, 2011).
- [2] D Liben – Nowell, J. Kleinberg, "The link – prediction problem for social networks[J]," (Journal of the American society for information science and technology, vol. 58, no. 7, pp. 1019-1031, 2007).
- [3] M E J Newman, "Clustering and preferential attachment in growing networks[J]," (Physical Review E, vol. 64, no. 2, pp. 025-102, 2001).
- [4] P. Jaccard, "Etude comparative de la distribution florale dans une portion des Alpes et du Jura[M]," (Impr. Corbaz, 1901).
- [5] LA Adamic, E. Adar, "Friends and neighbors on the web[J]," (Social networks, vol. 25, no. 3, pp. 211-230, 2003).
- [6] Y B Xie, T Zhou, B H. Wang, "Scale-free networks without growth[J]," (Physica A: Statistical Mechanics and its Applications, vol. 387, no. 7, pp. 1683-1688, 2008).
- [7] L. Katz, "A new status index derived from sociometric analysis[J]," (Psychometrika, vol. 18, no. 1, pp. 39-43, 1953).
- [8] R N Lichtenwalter, J T Lussier, N V. Chawla, "New perspectives and methods in link prediction[C]," (Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 243-252, 2010).
- [9] M Al Hasan, V Chaoji, S Salem et al., "Linkprediction using supervised learning[C]," (SDM'06: Workshop on Link Analysis Counter-terrorism and Security, 2006).
- [10] N Benchettara, R Kanawati, C. Rouveirol, "Supervised machine learning applied to link prediction in bipartite social networks[C]," (Advances in Social Networks Analysis and Mining (ASONAM) 2010 International Conference on, pp. 326-330, 2010).

