

**Pune Institute of Computer Technology  
Dhankawadi, Pune**

**A SEMINAR REPORT  
ON**

Social Recommendation Engines Using Machine Learning

**SUBMITTED BY**

**Name Archit Prasad Kane  
Roll No. 31206  
Class TE 2**

**Under the guidance of  
Prof. Dr. A. S. Ghotkar**



**DEPARTMENT OF COMPUTER ENGINEERING  
Academic Year 2020-21**



DEPARTMENT OF COMPUTER ENGINEERING  
**Pune Institute of Computer Technology**  
**Dhankawadi, Pune-43**

**CERTIFICATE**

This is to certify that the Seminar report entitled

**“Social Recommendation Engines Using Machine Learning”**

Submitted by

Archit Prasad Kane      Roll No. 31206

has satisfactorily completed a seminar report under the guidance of  
Prof. Dr. A. S. Ghotkar towards the partial fulfillment of third  
year Computer Engineering Semester II, Academic Year 2020-2021  
of Savitribai Phule Pune University.

Prof. Dr. A. S. Ghotkar  
Internal Guide

Prof. M.S.Takalikar  
Head  
Department of Computer Engineering

Place:

Date:

## **Abstract**

Billions of multimedia messages and posts are shared among users. We can use machine learning algorithms to analyse these multimedia messages and assign tags. These tags are then used to analyse user connections in a more meaning full way. Users share similar posts irrespective or origin. These massive amount of data can be used to discover new users and forums on social media platforms using ML.

## **Keywords**

Big data , Machine Learning, Social Graphs, Recommendation, Social Network Analysis.

## ACKNOWLEDGEMENT

I sincerely thank our Seminar Coordinator Prof. B.D.Zope and Head of Department Prof. M.S.Takalikar for their support.

I also sincerely convey my gratitude to my guide Prof. Dr. A. S. Ghotkar, Department of Computer Engineering for her constant support, providing all the help, motivation and encouragement from beginning till end to make this seminar a grand success.

# Contents

<b>1</b>	<b>INTRODUCTION</b>	<b>1</b>
<b>2</b>	<b>MOTIVATION</b>	<b>2</b>
<b>3</b>	<b>LITERATURE SURVEY</b>	<b>3</b>
3.1	Survey On Papers . . . . .	3
3.1.1	Characterizing User Connections in Social Media through User-Shared Images . . . . .	3
3.1.2	Recommendation Framework for Online Social Networks . .	3
3.1.3	Characterizing User Behaviour and Information Propaga- tion on a Social Multimedia Network . . . . .	3
<b>4</b>	<b>PROBLEM DEFINITION AND SCOPE</b>	<b>4</b>
4.1	Problem Definition . . . . .	4
4.2	Scope . . . . .	4
<b>5</b>	<b>METHODOLOGY</b>	<b>5</b>
5.1	Data Collection . . . . .	5
5.1.1	Dynamic Data . . . . .	5
5.1.2	Static Data . . . . .	6
5.2	Data Analysis . . . . .	7
5.3	Work Flow . . . . .	8
5.3.1	Recommendation Process . . . . .	8
<b>6</b>	<b>CONCLUSION</b>	<b>10</b>
	<b>REFERENCES</b>	<b>11</b>

## List of Figures

1	Types Of User Data . . . . .	5
2	Recommendation Process Diagram . . . . .	8

# 1 INTRODUCTION

The number of users using online social media platforms has increased over the past decade. With this has come large social communities on online platforms. Social Media has now transitioned from a simple text messaging platform to a richer multimedia experience.

Users interact with online communities and users. Interaction is established on topics of their interests and their roles. The roles can be either that of a creator or content consumer. These online connections are known as social graphs. Connections between users happen when there is a shared interest between them. With the increasing number of users on the internet, it has now become inadequate to use user connection discovery to find new connections on the platforms. The variety in content creation has made the interests of the users diverse. This amount of content to be shown to the user has now become incredibly small. Herein comes social recommendation systems. Content is tailored by looking into the user's past interests and activities. This helps to increase the engagement levels for the user.

In general recommendation systems use approaches: Collaborative filtering, content-based filtering and hybrid recommendations. Collaborative based filtering is based on the opinions of the user. For example in a review users rate the product after buying it on amazon. Content-based filtering is based on the previous posts liked by the user. For example, Amazon recommends products similar to the one you recently bought. Hybrid recommendations combine the above two specified methods.

This report summarises how hybrid recommendation engines can improve engagement levels for users with the help of machine learning. We can use machine learning algorithms to segregate the users and recommend new users based on their interests and behaviours. We shall discuss the usage of psychological tools to find out behavioural traits. Openness, conscientiousness, extraversion, agreeableness, and neuroticism are the five dominant traits we are using to gather behavioural data. Machine learning can be used to extract the user's interests by looking into the posts he interacts with and the frequency of conversational topics. It is to be noted that machine-generated labels have not proved to improve the user recommendation system. It has so far been able to prove that there are similar interests among users. We inspect the demographic and psychological aspects of users on social media networks by extracting data. The methods of collecting user data from sources are discussed (section 3). We shall also detail our use of clustering algorithms to similar individuals or groups. Using the above-mentioned techniques it has been proved an effective alternative to the traditional approach which used user described tags or social graphs.

## 2 MOTIVATION

The number of users using online social media platforms has increased over the past decade. With this has come large social communities on online platforms. Social Media has now transitioned from a simple text messaging platform to a richer multimedia experience.

This lead to an increase in the amount of content that can be accessed by the user. Giving us the need to improve our social recommendation algorithms to provide more engagement to the user using machine learning. Content is tailored by looking into the user's past interests and activities. This helps to increase the engagement levels for the user. Also it can recommend people of similar interests. These techniques have been proved an effective alternative to the traditional approach which used user described tags or social graphs.



## **3 LITERATURE SURVEY**

### **3.1 Survey On Papers**

#### **3.1.1 Characterizing User Connections in Social Media through User-Shared Images**

Ref . [1] Images analysed using Machine learning algorithms generate labels to identify the user's interests. They analysed the data to find out that users who share similar images follow each other. This is irrespective of the source or medium of exchange between the users. It shows better performance than the said friends of friends recommendation approach.

#### **3.1.2 Recommendation Framework for Online Social Networks**

Ref . [3] The algorithms needed to measure the strength of relationships, user activity and similarities are discussed. The data required for analysis and recommending users is filtered through these algorithms. These algorithms together are used to recommend new users on the social media platform.

#### **3.1.3 Characterizing User Behaviour and Information Propagation on a Social Multimedia Network**

Ref . [5] The roles among users are inferred from the behaviours on the social platform. Using data extracted from user messages they can be clustered based on these behavioural patterns. The messages shared between users can be used to find out the roles between users.

## **4 PROBLEM DEFINITION AND SCOPE**

### **4.1 Problem Definition**

Understand how social recommendation engines work with the help of machine learning.

### **4.2 Scope**

Extracting user-generated data and examining it to find out the users interests.

The data includes text and images shared or interacted with by users.

This derived data is used to recommend users or forums of similar interests to the user on the social graph.

## 5 METHODOLOGY

### 5.1 Data Collection

The data to be collected can be mainly classified into two broad divisions: Dynamic and Static.

#### 5.1.1 Dynamic Data

Dynamic data is gathered and analysed by the system. Dynamic data is analysed to determine the strength of user-user or user-forum connection. For this, we will require the recent post-interaction history. This process is now sub-divided into the following parts:

1. **User to user using chat analysis:** Using the recent conversations between them, we can analyse the common topics of interest and the roles between the users. Multi-Media shared on the platform is used to generate tags by training the system.
2. **User to forum:** Tracking the recent posts (likes, comments, awards, etc) that the user has interacted with is necessary to find out the topic of interest of the user. In this case, the posts interacted with are assigned their respective labels which are then added to the user's topics of interest.

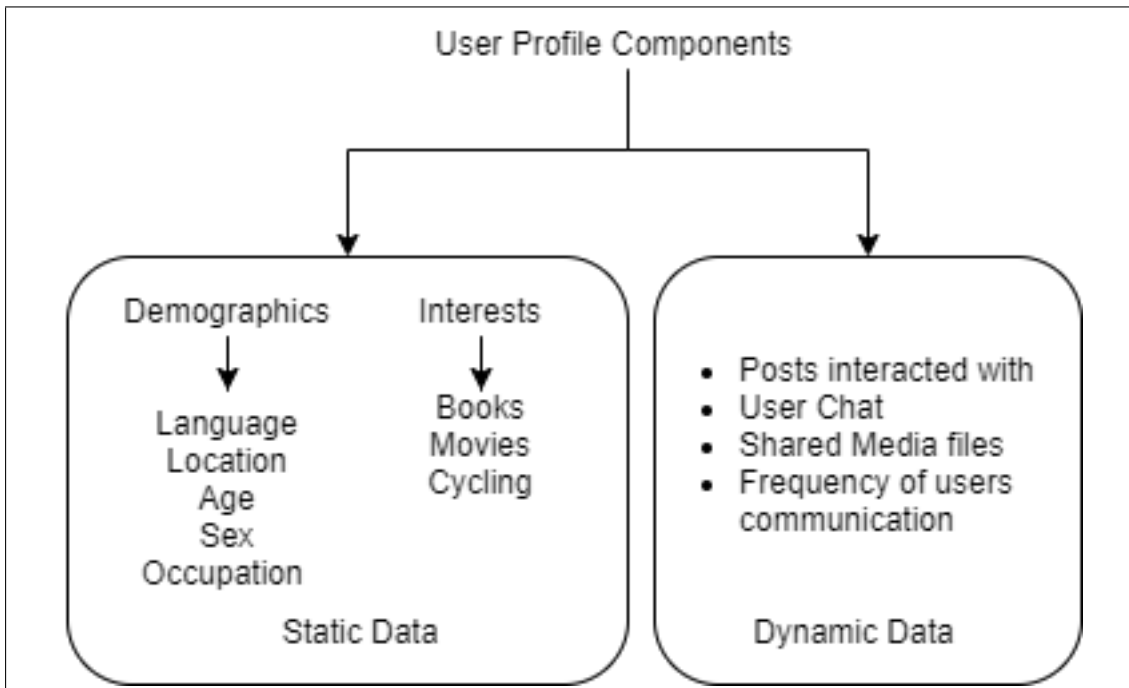


Figure 1: Types Of User Data

We shall now look into how the data is extracted from the above sources:

1. Text-based media can be analysed by using natural language processing to find out the topics of interest, openness, honesty, extroversion, agreeableness, and neuroticism . Natural Language Processing is used to extract the

keywords in the text. Rapid Automatic Keyword Extraction is a keyword extraction algorithm used to extract keywords in each chat bubble. The RAKE algorithm consists of a list of stop words ('and', 'of', 'the'). These words are used to partition the text into an array of keywords. We now use this array of candidate keywords to create a graph of word co-occurrence and calculate the word score. Several metrics are used to determine the word scored using Word Frequency, Word Degree, Ratio of the degree to the frequency, words that occur often, and the words that occur the most. The keywords extracted from this can be assigned as a topic of interest to the user.

2. Image-based media can be analysed by using developed object detection algorithms to generate labels. These labels are assigned a topic of interest to the user profile.

### 5.1.2 Static Data

Static data is delivered by the user by filling out forms and from their bio-data. The data collected from the user are Language, Location, Age, Gender, number of followers, following which users, tagged posts, likes and views per post and forums.

Using the above data the openness of an individual is estimated by the number of times the user is tagged in other users' posts, the number of likes to view ratio of posts created by other users and the number of times they are tagged in status posts. The conscientiousness of a user is determined by the frequency of responding to messages. Extraversion is determined by the no of times conversations are initiated, analysing the images posted or tagged in. Agreeableness is calculated by the strength of the social graph links with other users over some time. neuroticism can be determined by analysing the chats or messages that the user send to others.

## 5.2 Data Analysis

Based on the above-gathered data we can now create a user recommendation system framework. By combining these multiple sources of data we can create a richer engagement experience. Our goal is to use data-driven algorithms to find out new user connection by behaviour. We are going to use the k-means algorithm to extract entity grouping based on their interests. We then examine user profiles within the clusters. The following are the user features that have been identified:

1. Avg. no. of broadcasts per day (f1)
2. Number of friends (f2)
3. A common time for broadcasts (f3)
4. Diversity of topics within messages(f4)
5. Diversity of topics within posts (f5)

We now compute the above features into the k-means algorithm. K-means depends on equal scaling between feature elements. Then the feature values are normalized to 0 - 1. This is done because each feature value have different units.

We then apply the k-means algorithm to divide the users into clusters based on similar behaviour. Each user within the feature cluster goes through demographic and behavioural profiles to identify common characteristics between the users.

We then filter the list of recommended users by using the strength of relationship calculation equation to chose the top few.

## 5.3 Work Flow

### 5.3.1 Recommendation Process

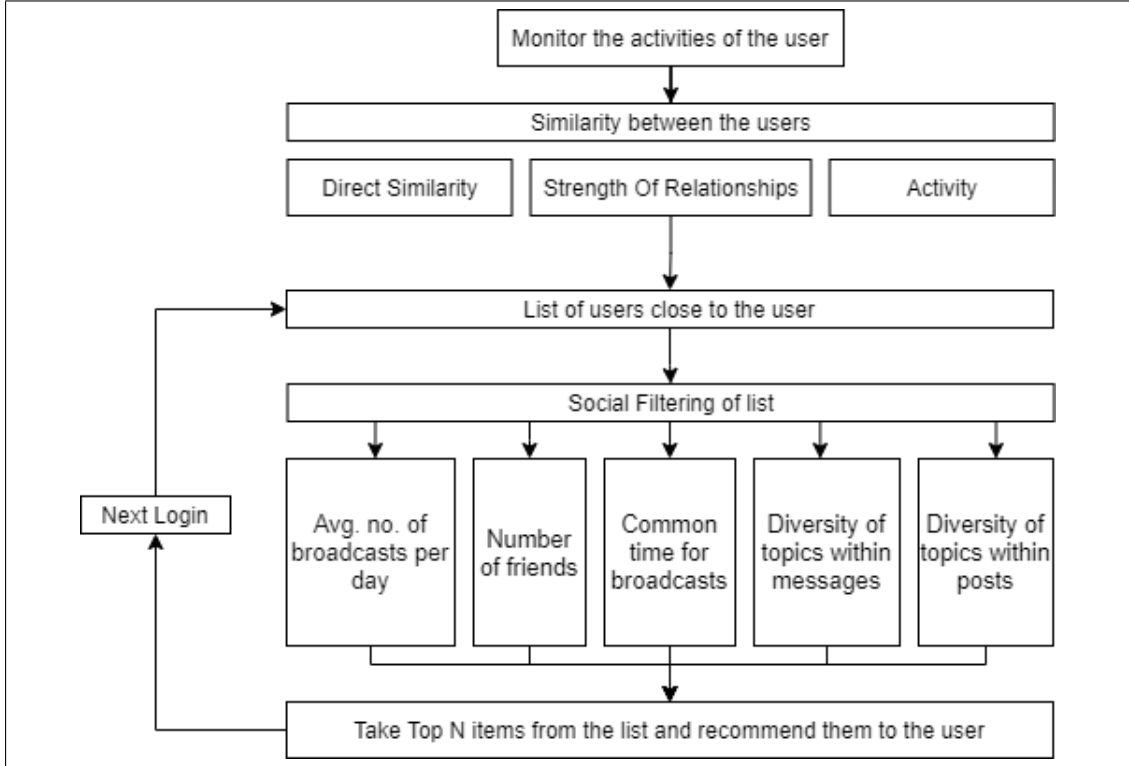


Figure 2: Recommendation Process Diagram

Based on the accumulated information we have developed the recommendation framework that supports the creation of the recommendation for the social network. The main purpose of the system is to present the most appropriate recommendations to users. By merging several, various sources of data, this method helps a new user to join the platform and engage him/her with content.

Based on user profiles, particularly their static components, all users are clustered separately in groups using clustering algorithms. Next, depending on the aim of the network, the connectivity social filtering is used to improve the connections. The users from the same group will be suggested if strong similarities are existing,

The purpose of the recommendation is to find out whether the user ought to be recommended to person  $x$ . It is accomplished by utilising the final similarity function:

$$r(x - y) = As(x, y) + Bc(x, y) + Ga(y) + Dsr(y)$$

where:  $s(x, y)$  – the direct correlation connecting user  $x$  and  $y$  that is derived from the comparison of all attributes from demographic and interests.  $c(x, y)$  – the complementary of connection initiation function defines the social behaviour of the users in relationships.  $a(y)$  – the activity of the user.  $sr(y)$  – the strength of the relationship between two users.

A,B,G,D are the importance coefficients which are assigned values from 0 -1. Coefficients are used to simulate and adjust the progression of the social network. The algorithm to calculate  $s(x, y)$ ,  $c(x, y)$ ,  $a(y)$ ,  $sr(y)$  are explained in [3]

## 6 CONCLUSION

This report examined the social recommendation system and the variety of data required along with the techniques of analysing and suggesting users and forums.



## References

- [1] M. Cheung, J. She and N. Wang, "Characterizing User Connections in Social Media through User-Shared Images," in *IEEE Transactions on Big Data*, vol. 4, no. 4, pp. 447-458, 1 Dec. 2018, doi: 10.1109/TBDATA.2017.2762719.
- [2] <https://medium.datadriveninvestor.com/rake-rapid-automatic-keyword-extraction-algorithm-f4ec17b2886c>
- [3] Kazienko, Przemysław Musiał, Katarzyna. (2006). Recommendation Framework for Online Social Networks. 10.1007/3-540-33880-2-12.
- [4] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.
- [5] F. T. O'Donovan et al., "Characterizing user behavior and information propagation on a social multimedia network," 2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2013, pp. 1-6, doi: 10.1109/ICMEW.2013.6618395.


## Document Information

---

Analyzed document	PlagarismCheck_31206_Archit-7-16.pdf (D106496475)
Submitted	5/26/2021 12:25:00 PM
Submitted by	Aaghotkar
Submitter email	aaghotkar@pict.edu
Similarity	1%
Analysis address	aaghotkar.pict@analysis.urkund.com

## Sources included in the report

---

SA	<b>Pune Institute of Computer Technology / TE_Seminar_2019_20_report_format.pdf</b>		1
	Document TE_Seminar_2019_20_report_format.pdf (D67797150)		
	Submitted by: ddkadam@pict.edu		
	Receiver: ddkadam.pict@analysis.urkund.com		

---