IJICTR
International Journal of Information &
Communication Technology Research

Volume 13- Number 3 – 2021 (48 -57)

A Graph-Based Content Similarity Approach for User Recommendation in Telegram

Davod Karimpour
Department of Computer Engineering
Yazd University
Yazd, Iran
dkarimpoor@stu.yazd.ac.ir

Mohammad Ali Zare Chahooki*
Department of Computer Engineering
Yazd University
Yazd, Iran
chahooki@yazd.ac.ir

Ali Hashemi
Department of Computer Engineering
Yazd University
Yazd, Iran
alihashemi@stu.yazd.ac.ir

Received: 4 March 2021 - Accepted: 13 May 2021

Abstract—**Telegram is a cloud-based instant messenger with more than 500 million monthly active users. This messenger is very popular among Iranians, as more than 50 million Telegram users are Iranians. Telegram is used as a social network in Iran because it offers features beyond a simple messenger, but does not offer all the features of social networks, including user recommendation. In this paper, investigating a real dataset crawled from Telegram, we have provided a hybrid method using the user membership graph and group characteristics to recommend the user in Telegram. The membership graph connects users based on membership in the same groups. Also, the characteristics for each group are indicated by the name and description of that group in Telegram. We created a bag of words for each group using natural language processing methods, then combined the bag of words for each group with the results of the membership graph processing. Finally, users are recommended based on the list of groups obtained by the combination. The data used in this paper include more than 900,000 groups and 120 million users. Evaluation of the proposed method separately on two categories of Telegram specialized groups shows the model integration and error reduction for the first category to 0.009 and the second category to 0.016 in RMSE.**

Keywords: Recommender systems; Telegram; Social networks; Membership graph; group's characteristics.

Article type: Research Article

I. Introduction

Today, the activity of users in social networks and messengers has become more prominent than before. This topic has spread to such an extent that many companies and factories are trying to promote their products and services among users through these environments. In recent years, instant messenger softwares have become very popular and has become one of the most important communication tools in various operating systems. In these environments, a lot of information is generated every day by users' activity, and analyzing this information is very valuable for researchers and marketers [1].

One of the advantages of messengers is the impact on business prosperity that can be used to market products. Today, due to the expansion of businesses and the inability to maximize face-to-face advertising, a

* Corresponding Author

huge wave of businesses have turned to messengers. Advertising based on sending messages has long been of interest to marketers since the advent of mobile messaging services. In this type of advertising, finding target users to send advertising to them is very important; because if unrelated users are found, it will cause users to feel dissatisfied after receiving the ad and block the sender of the message (Kyuyong Shin and et al. [2] provided a large-scale framework for targeted advertising in Line Messenger). According to the mentioned need, recommender systems for modeling users according to their interests and also finding target users are very useful. In these systems, an attempt is made to find the most appropriate and closest items to recommend the user by guessing how he thinks. Recommender systems include many filtering methods that model users based on their interests. These methods are divided into different categories based on the amount of information extracted from users. One of these methods is content-based filtering, which depends only on a single user's information. It also depends on the content, including keywords and text analysis of the user message. Another method is called collaborative filtering, which depends on the information of multiple users. This method uses other users' information for more accurate recommendations. If we combine two or more filtering methods, the combined filtering method is obtained. This method tries to reduce the limitations of other methods. This paper is based on hybrid filtering because it uses the information of all users (collaborative filtering) and combines the membership graph with the characteristics of the groups (content-based filtering).

Telegram is a cloud-based instant messenger with more than 500 million monthly active users (MAU). This messenger doubled its MAU in two years [3]. Telegram offers different features such as creating a supergroup, channel, bot, secret chat, voice and video calls, and finding groups and users based on the location. Users in each group discuss a specific topic. Of course, some groups have a lot of spam messages. The channel in this messenger is a one-way notification. Channel members are not allowed to send posts and can only comment on each post. Bots are like telegram accounts that are managed virtually by software and often use artificial intelligence features. For example, a bot can delete spam messages in a group.

In fig. 1, the features of Telegram are compared to Facebook Messenger. Telegram is similar to Facebook Messenger in many features. The channel feature in Telegram has made it unique compared to Facebook Messenger. The feature of creating a group in Telegram is possible with an infinite number of members, and this amount is a maximum of 250 members in Facebook Messenger. The number of forwards of a message can be displayed in Telegram, while Facebook Messenger does not display the number of forwards of a message. Message editing is possible in Telegram, but Facebook Messenger does not offer message editing. Also, file sharing in Telegram is 1.5 GB and in Facebook Messenger is 25 MB.

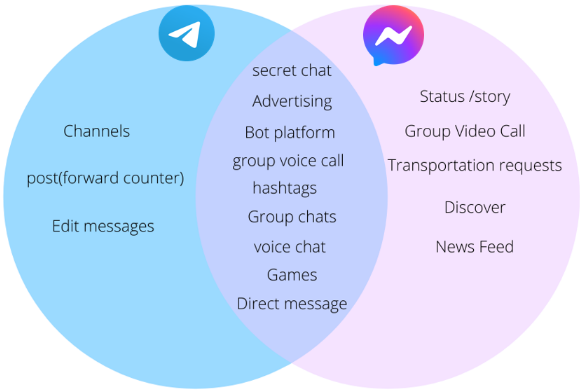

Fig 1. Comparison of Telegram and Facebook Messenger

Recently, a lot of research has been done with data extracted from Telegram groups and channels. Some papers such as [4] and [5] have collected and offered data in the context of this messenger. Hashemi and Chahooki [6] proposed a way to the ranking of groups. In another study, Hashemi and Chahooki [7] have measured groups' quality based on the behavior of the users. Karimpour et al. [8] have proposed a method for group recommendation by modeling users' records and analyzing their migration between groups. Furthermore, in another study, Karimpour et al. [9] improved the ranking of the recommendation list groups compared to the article [8].

Telegram is used as a social network in Iran, but does not offer all the features of a social network, including user recommendation. The social network search engine offers the ability to find users by first and last name and bio. But in messengers, users often communicate with a small number of people at their audience level and are not able to find users like social networks. Of course, the Telegram search engine can only find users by having the exact ID of each user. Also, Telegram does not do any analysis of user groups.

In this paper, we get a list of ranked groups by combining membership graph and keywords extracted from groups name and description. Then, users from these groups are recommended in order of listing. In general, the proposed method consists of two phases, offline and online. Each of the phases is summarized as follows:

- Offline phase: In this phase, there is a membership graph and a word bag (one bag of words for each group). The membership graph indicates the membership of users in Telegram groups and also this graph is heterogeneous and has two types of nodes (group and user). The sack of words contains a bag of words for each group. To make a bag of words from each group, we convert the group name and descriptions into keywords using natural language processing methods in eight consecutive steps.
- Online phase: This phase receives the user set (input), and using the membership chart (offline phase), it obtains a set of ranked groups based on the most common members. In the list of obtained groups, the bag of words of each group (offline phase) is combined with the bag of words of the

previous groups and compared with all the bag of words of the groups (the whole bag of words in the sack of words obtained in the offline phase). Then the groups are listed based on the maximum number of words common. The members of the new groups are extracted from the groups in order of the list to reach the number of target users.

The dataset of this paper was obtained through the Telegram API by Idekav[1] system, and this data contains more than 900,000 supergroups and 120 million users. In this paper, Telegram specialized groups have been used to evaluate the proposed method. For evaluation, we have considered two categories of groups separately. Each category includes 25 specialized groups in Telegram, obtained by an expert. In order to evaluate the proposed method by each of our specialized groups, we have divided the users of each group into two sets of test and train. The proposed method in this study is not limited to telegram messengers, but it can be examined on messengers and social networks that have the ability to create groups.

The rest of the paper is organized as follows; Sect. 2, provides related works. Sect. 3, demonstrates the proposed method. Sect. 4, analyzes the experimental results. Finally, Sect. 5 renders conclusions and future work.

II. Related Work

In this paper, we briefly review related work from two perspectives. First, we will explain the user recommendation in social networks, and then we will explain the document similarity methods.

A. User recommendation in social networks

Social networks use different filtering methods for user recommendation. The following are explain three of the most widely used filtering methods.

- Content-based filtering: This type of filtering uses only the user's own information to recommend similar users and this method does not take into account other users' information. Features of this filtering include user messages, user gender, user favorite color, etc [10].
- Collaborative filtering: This filtering is one of the most popular filtering methods in recommender systems, which is also widely used on Amazon and Netflix sites. This method tries to make more accurate recommendations by searching and finding users who have similar interests to the target user, and assumes that users who have had similar interests in the past will have similar interests in the future [11]. Collaborative filtering is divided into two categories: memory-based and model-based. The memory-based method is based on user feedback, and the model-based method uses a graph that models user activity and behavior for recommendations [12].
- Hybrid filtering: This method, by combining other filters, tries to reduce their limitations [10].

Considering that this research has considered the graph and all users' information to recommend the user, and also has used the groups' characteristics for the recommendation, it can be said that this research is a method based on hybrid filtering. Many studies have been done in relation to recommender systems based on different filters, some of which are described in this subsection based on the type of filtering and social network used in Table 1.

B. Document similarity

The similarity of the document has been highly regarded for the past two decades, and so far much research has been done on the similarity of the document. There are many ways to display texts and vector modeling, including display as a word bag and vector space model [19]. Many algorithms such as cosine similarity, jaccard similarity and dice similarity are the basic methods in this field (see [20] for a review and comparison of all these methods). In addition to these methods, there are popular methods such as GloVe [21] and word2vec [22] for embedding words in this field. In the following, we will describe some studies that have examined the similarity of the document.

The proposed method by R. Singh and S. Singh [23], could efficiently recognize the best news reports and measure the similarity among them. This study checked the best report items on the news sites and measures the similarity in two related report items in two languages (English and Hindi) relating to the corresponding event. They created a link extractor to obtain the best report for Hindi and English from Google. First, the Hindi report is translated into English by Google Translator and then matched to the English report. Lastly, they used the cosine similarity, Jaccard similarity, Euclidean distance measure to determine the report similarity rate.

The proposed method by I. Rushkin [24], is a computational way for computing similarities among text documents. The name of this method is the density similarity, or DS for short, because it describes documents as possibility densities in the embedding space. This way is based on a word embedding in a high-dimensional Euclidean space and on kernel regression, and considers into account semantic associations between words.

III. Proposed Method

The general workflow of the proposed method is shown in fig. 2. The proposed method consists of two phases, offline and online. Each of the two phases has two separate steps. In the offline phase, we create membership graph and sack of words. In the online phase, the groups of incoming users are checked through the membership graph (offline phase) and then its results are combined with the sack of words (offline phase).

A. Section 1: offline phase

In this section, the membership graph of users is created. A bag of words is also created for each group.

1) Step1: Membership Graph

In this step, the membership graph, models users based on their membership in groups. Each user is a

[1] idekav.com/

member of at least one group. Bottom left part of fig. 2, shows a schematic of the membership graph.

TABLE I. COMPARISON OF PREVIOUS USER RECOMMENDATION STUDIES BASED ON THE TYPE OF FILTERING AND SOCIAL NETWORK

Paper	Filtering	Social Network (Dataset)	Explain
[13]	Hybrid	LinkedIn	This paper presents a hybrid method for user recommendation based on enterprise communication and SCM. The proposed method used a hybrid approach that combines collaborative filtering and demographic recommendation systems, using data mining, artificial neural networks, and fuzzy ways. This system works like a demographic recommender system, with the difference that the people's distinctive features in the SCM are considering into account rather than personal specification. This study used specific features of users such as function, industry, work level, and work experience to recommend people to each other.
[14]	Collaborative	Twitter Facebook	In this paper, two separate algorithms for friend recommendation using model-based collaborative filtering are presented. The first algorithm takes into account the number of mutual friends of each user and the second algorithm is designed to prioritize users and influence different users. So that each user is assigned an impact rating. For example, if a user has an impact factor of 1, this factor is shared among his friends.
[15]	Hybrid	Movie-lens	This paper discusses the problem of recommendation performance for groups of users. The proposed method concentrate on the performance of very Top-N recommendations, which are necessary during recommending long-lasting items. This article provides a hybrid recommendation for groups to develop existing group recommenders by combining content-based collaborative filtering. The results of this study showed that candidates who are recommended with both approaches at the same time are more suitable for the group than the candidates with individual approaches.
[16]	Content	Flickr	This paper uses the characteristics of gender, color, age, and user interest. The friend recommendation in this study is based on a two-layer method. The first layer is for examining the graph of friendship between users and the second layer is for the tagged graph of each user's characteristics.
[17]	Hybrid	Instagram	This paper offered user-to-user recommendation utilizing a user similarity metric calculated and analyzing the pictures shared by users on their Instagram account. In this method, some users with a large audience and a well-established reputation are called "influencers". The main idea is that if a pair of influencers share pictures including similar content it is possible that they have similar interests. Also, users that follow other users sharing similar content are more related. This method is a hybrid recommendation that combines collaborative filtering and results from pictures content.
[18]	Hybrid	Yelp	The method proposed in this paper, is hybrid filtering that combines user-based collaborative filtering with semantic and social recommendations. The semantic section recommends friends based on the calculation of the similarity among the user and his/her friends. The social section is based on social-behavior features such as friendship and credibility degree. This method explains the user's credibility based on his/her trust and commitment in the social network.

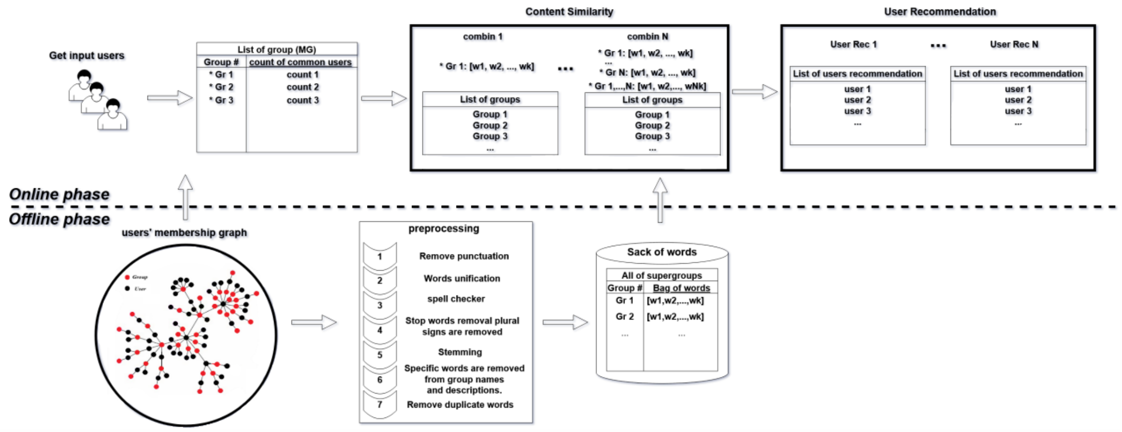

Fig 2. Workflow of the proposed method

2) Step 2: Sack of Words

In this step, we make a word bag for each group. The bag of words is derived from the name and description of each group. The bag of words for all the groups is specified in fig. 2 as sack of words. All the groups studied in this paper are in Persian and English. Furthermore, many Persian groups have an English name and description. We have processed all Persian and English words. In the following, the data preprocessing indicates the methods of extracting keywords from the name and description of each group.

a) Data Preprocessing

In this part, we extract keywords from the names and descriptions of groups. The following methods extract these keywords.

- Remove punctuation: All punctuation marks in Persian and English are removed. For example: '":;,?.! . The elimination of punctuations is because we have seen the integration of these symptoms unusually in many of the groups' descriptions.
- Words unification: For all words in the name and description of English groups, uppercase letters were converted to lowercase letters. Most groups had links in their descriptions, including site URLs, group links, and emails, all of which were removed. Also, some letters that had the same text in Persian and Arabic letters were edited.
- Spell checker: The spelling in the name and description of the groups is checked. In addition, some words found in the description of many groups, such as the word "ssaallaamm" is converted to "salam". Of course, this causes some words, such as "Address" is converted to "Adres" in English, which does not pose a problem for our purpose because it makes all the words the same.
- Stop words and Plural signs are removed: Prepositions and English and Persian pronouns are removed from the name and description of the groups. For example, the word "از" in Persian and "from" in English. Also, pronouns like "ما" in Persian and "we" in English. Also, plurals in words are removed. For example, replacing "users" with "user". Of course, in Persian, some words are not plurals that are mistakenly identified as plurals. For example, the word "تهران" in Persian is changed to the word "تهر", which is not true. To solve this problem, we have replaced these words with their correct spelling.
- Stemming: All words in group name and description are replaced with their stems. In this case, the different ways of writing words are reduced and many words become one form.
- Specific words are removed from groups name and description: Some annoying words in the groups' description such as "telephone", "address" in Persian and English were removed. Also, some annoying words in the groups' name such as the word "group", "chat" in English and some words such as "تبلیغ","گروه", and "چت" in Persian were removed.
- Remove duplicate words from each group's word bag: Finally, duplicate keywords are removed from each group's word bag.

B. *Section 2: online phase*

In this section, firstly, we extract the groups of input users from the membership graph (offline phase). Then, the groups that have the most common members with incoming users are listed in order (maximum number of common users). We call the list of obtained groups MG which stands for Membership Graph. In the content similarity step, the bag of words of each group in the MG list is extracted from the sack of words and combined with each other. In the user recommendation step, users are recommended from the groups obtained in the content similarity step.

1) Step 1: Content Similarity

In this step, the bag of words of the MG groups are extracted from the sack of words. Then, at each stage of the combination, the bag of words of each group (from the MG list) is gathered with the bag of words the previous groups. Finally, the new bag of words is compared to the bag of words of all groups (in the sack of words), and the groups that have the most common words are ranked accordingly. At all stages of the combination, groups that their bag of words is used will be removed from the new recommended group list.

a) Combination 1

In this combination, the first group is extracted from the list of MG groups. Also, the bag of words of this group is extracted from the sack of words. Then, this bag of words is compared to the word bag of all the groups (in the sack of words) and new groups are obtained based on the maximum word commonality. New groups (based on the number of common words) are ranked in descending order. In this combination, the first group that bag of words has been used is removed from the recommended new list.

b) Combination 2

In this combination, the first and second groups are extracted from the list of MG groups. The bag of words of these groups is extracted from the sack of words and merged. Then, this new bag of words is compared to the word bag of all the groups (in the sack of words) and new groups are obtained based on the maximum word commonality. New groups (based on the number of common words) are ranked in descending order. In this combination, two groups that their bag of words have been used is removed from the recommended new list.

c) Combination N

In this combination, the first group, the second group, the third group, ..., group N are extracted from the list of MG groups. The bag of words of these groups is extracted from the sack of words and merged. Then, this new bag of words is compared to the word bag of all the groups (in the sack of words) and new groups are obtained based on the maximum word commonality. New groups (based on the number of common words) are ranked in descending order. In this combination, N groups that their bag of words have been used is removed from the recommended new list.

2) Step 2: User Recommendation

In this step, we recommend users. In general, users' recommendations are made through the groups obtained in the Content Similarity step. In the content similarity step, a list of groups is obtained from each combination. We gather users of these groups in the order of the recommended list in each combination. After each stage of the combination, this will continue until we reach the target (desired number of users). Therefore, for each combination in the content similarity step, we obtain separate users' recommendation. The number of target users can be considered as different values. It can be considered 10, 20, and 30 times the number of incoming users or any other value.

IV. Experimental Results

In this section first, the data used is described and then the evaluation method and its results are explained.

A. Experimental Dataset And Implementation Environment

The data used in this paper is a real-world dataset from Telegram Messenger and were obtained accurately by Idekav system. This dataset contains only general information of Telegram and includes 900,000 supergroups and 120 million users. For all supergroups, in addition to group member information, we have considered the group name and description. The exact statistics of this dataset are shown in Table 2. All users are obtained from the membership graph. This means that each user is in at least one member group.

The implementation and evaluation environment of all these methods is performed on a 64-bit core-i7 system with 8 GB of RAM. To implement the proposed method, we have used MySQL database installed on the server, using mysql.connector library in Python.

TABLE II. The Dataset Statistics

Count of Supergroups	Count of users	Average count of members of Supergroups
920810	125269522	1135.553

B. Evaluation Method

In this paper, specialized Telegram groups have been used to evaluate the proposed method. Specialized groups are groups in which no spam or advertising messages are sent. Users in these groups discuss a specific topic. Also, in choosing these groups, we tried to keep the number of group admins as small as possible. If the number of admins in a group is more than usual, the multifaceted administration leads to decrease in group quality. The reason for choosing specialized groups for evaluation is that all members of these groups are users who are really interested in the topic of the group and do not send messages that are not related to the topic of the group. This indicates that the members of such groups tend to have discussions appropriate to the topic of the group and agree with each other on a particular topic. To evaluate the proposed method, we chose two separate categories of groups, each of which includes 25 specialized groups. The reason for choosing two categories of 25 groups is to show that the result was not accidental and the results in the other 25 are not different. The number of groups' members in each category is between 2,000 and 10,000. We have named these two categories with *A* and *B*. Category *A's* information is given in Table 3 and Category *B's* information in Table 4. We have evaluated the proposed method on each specialized group separately. The evaluation method is that for each group we give 80% of the group members to the proposed method and evaluate the results on the remaining 20%. For evaluation and comparison, the target of all methods is to reach 10 times the number of input users chosen (Or reaching 10 times the 80% set). Each of the user recommendation methods (in the proposed method) will continue until reaching the target set. In this paper, RMSE (Root-Mean-Square Error) is used to evaluate the proposed method. Equation (1) demonstrates this error. This method is used to check the model prediction error. According to (1), $Predicted_i$ is prediction set that includes a set of zeros and ones. Zero indicates that the model prediction was correct and one indicates the opposite. $Actual_i$ is actual set and contains a set of zeros that represent the set of users stored for testing. N is the set of errors in the suggested list.

$$Rmse = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}} \quad (1)$$

C. Evaluation Results

The evaluation results of the proposed method are shown for category *A* in Table 5 and category *B* in Table 6. According to Tables 5 and 6, the number of groups, number of users, and RMSE are considered for each combination. To explain these three topics, we start with an example in Table 5. Consider group number 6 in Table 5. This group, in the first combination, has reached 59,599 users by adding users of 55 groups. Given that, the target of each combination is to reach 10 times the number of incoming users. The number of incoming users of this group is 6005 in table 3 and the target is 60050. This group has been able to reach a maximum of 59,599 users in the first combination with an RMSE of 0.871.

TABLE III. Information of Categoty A

Category A			
GR #	***Number of members***	***Number of inputs (80%)***	***Number of predictions (20%)***
1	9919	7935	1948
2	9191	7353	1838
3	8841	7073	1768
4	8564	6851	1713
5	8167	6534	1633
6	7506	6005	1501
7	7031	5625	1406
8	6791	5533	1358
9	6741	5393	1348
10	6351	5081	1270
11	6111	4889	1222
12	6014	4811	1203
13	5811	4649	1162
14	5579	4463	1116
15	5318	4254	1064
16	4630	3704	926
17	4557	3646	911
18	4379	3503	876
19	3725	2980	745
20	3377	2702	675
21	3271	2617	654
22	2828	2262	566
23	2298	1838	460
24	2067	1654	413
25	2038	1630	408
Average	5644.2	4519.4	1128.8

TABLE IV. Information of Categoty B

Category B			
GR #	***Number of members***	***Number of inputs (80%)***	***Number of predictions (20%)***

1	10000	8000	2000
2	9308	7446	1862
3	8993	7194	1799
4	8031	6425	1606
5	8023	6418	1605
6	7720	6176	1544
7	7573	6058	1515
8	7163	5730	1433
9	6935	5548	1387
10	6713	5370	1343
11	6697	5358	1339
12	6552	5242	1310
13	5924	4739	1185
14	5675	4540	1135
15	5350	4280	1070
16	4950	3960	990
17	4948	3958	990
18	4627	3702	925
19	3360	2688	672
20	3288	2630	658
21	3255	2604	651
22	2749	2199	550
23	2400	1920	480
24	2116	1693	423
25	2024	1619	405
Average	5774.96	4619.9	1155.1

We performed our experiments by combining bags of words of 1 to 5 groups. In addition, we checked the bags of words combination of 20 groups to assess changes in RMSE. The mean RMSE results for 25 groups A and 25 groups B are shown in fig. 3. In fig. 3, the horizontal axis represents the number of groups their words are used (Each of the combinations in content similarity step of the online phase). The vertical axis represents the mean of the RMSE. According to fig. 3, the results obtained by combining the bag of words of the 4 groups reduced the RMSE compared to the other combinations in both categories A and B.

According to the results obtained in fig. 3, the combination of 20 groups has a significant increase in RMSE compared to other combinations. The general conclusion is that as the number of groups (combination the bags of words) increases, the prediction accuracy decreases, and the RMSE increases. Of course, given that in fig. 3, combination 4 has the best combination and the least RMSE, we can say that the slope of the RMSE diagram is not ascending all the time; there is rise and fall.

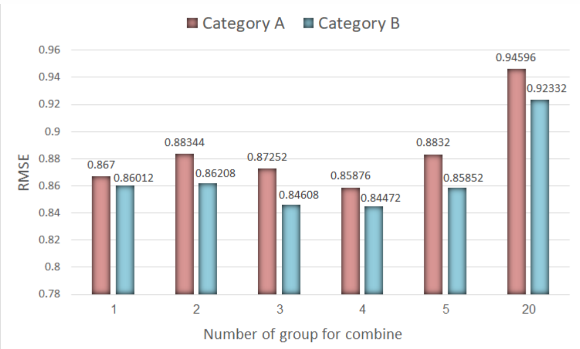

Fig 3. Average RMSE of each combination in categories A and B

D. Further Analysis

In this subsection, the results obtained from Tables 5 and 6 are analyzed.

Some groups, such as groups 10, 15, and 17 in Table 5 and groups 3, 11, 14, 18, and 21 in Table 6, achieved a higher RMSE than other combinations by considering the bag of words one group. The reason is that, the bag of words of the first group did not have related words or there were no groups according to the bag of words of that group.

In the groups marked with an + in Table 5 and Table 6, the RMSE of combination 1 is lower than the other combinations. The reason is that, the bag of words of combination 1 has better keywords than the other combinations, and also in these groups, as the bag of words expands, the number of unrelated users increases.

For groups 5, 15, 17, 18, and 20 in Table 5 and Groups 3, 8, 11, 13, 14, 17, 21, 24, and 25 in Table 6, the RMSE of combination 3 is less than combination 1 and 2. The reason is that, in these groups, combinations 1 and 2 are unable to find important keywords. This refers to the same rise and fall of the diagram in fig. 3, which here is the initial rise in combination 3.

In groups 6, 7, 10, and 24 in Table 5 and group 15 in Table 6, the RMSE of combination 2 and combination 3 are equal. In these groups, given that the number of users and the number of groups in the combination of 2 and 3 are different, but the RMSE is equal. When the bag of words of Group 3 is added to the bag of words combination 2, no increase or decrease in RMSE occurs. The reason is that the bag of words of the third group does not have related words or all the bag of words of the third group are in the bag of words of combination 2. This indicates that no raise or fall occurs and the diagram continues steadily.

In the groups marked with an * in Tables 5 and 6, combination 4 has less RMSE than all combinations. This indicates that most related words are created in combination 4.

According to Tables 5 and 6, in addition to the error, each of the combinations (each step of the combination) is also shown based on the number of groups required and the number of users obtained. According to the operation of each of the steps of the combination, which was explained in the section of the proposed method, the output of each combination is a list of ranked groups. Users of these groups are merged in the order of list to reach the end user (target) group. Each combination merges a different number of groups to achieve a list of (target) users. Fig. 4 shows the combinations based on the number of groups needed to achieve the target. If a combination with the least number of groups required achieves a list of end users (less than the target), there is no reason for the combination to be good or bad; because the number of members in each group varies.

TABLE V. EVALUATION RESULTS FOR CATEGORY A (NG: NUMBER OF GROUP, NU: NUMBER OF USER AND RE: RMSE)

GR #		Combine 1			Combine 2			Combine 3			Combine 4			Combine 5			Combine 20		
		NG	*NU*	*RE*	*NG*	*NU*	*RE*	*NG*	*NU*	*RE*	*NG*	*NU*	*RE*	*NG*	*NU*	*RE*	*NG*	*NU*	*RE*
1	+	7	79102	0.747	13	18503	0.83	30	47766	0.799	17	15988	0.836	37	79211	0.816	46	51017	0.947
2	*	41	69323	0.863	34	72164	0.856	28	70562	0.859	31	72315	0.854	13	69354	0.864	65	70594	0.986
3	*	11	49246	1	28	65805	0.999	22	49386	1	22	65588	0.886	5	65270	0.999	9	65656	1
4		33	64455	0.892	43	67127	0.869	47	62514	0.906	26	62185	0.911	25	68496	0.922	18	64689	0.952
5	*	31	64233	0.89	21	59179	0.889	30	62701	0.868	43	63199	0.863	48	64025	0.927	28	61732	0.966
6	+	55	59599	0.871	36	55221	0.949	28	59050	0.949	28	59366	0.874	21	59954	0.905	28	57632	0.968
7		7	46254	0.906	24	56032	0.876	19	52647	0.876	17	50487	0.88	8	54959	0.895	26	53436	0.933
8	+	35	53783	0.728	18	52144	0.757	26	53282	0.780	26	46853	0.782	26	51280	0.758	31	48146	0.927
9	+	73	52186	0.718	40	53580	0.801	20	53915	0.761	17	41470	0.753	26	50837	0.726	13	18523	0.974
10		24	50649	0.996	5	43291	0.868	8	45073	0.868	11	49488	0.868	13	50807	0.868	2	31228	0.942
11	+	34	47365	0.849	39	48679	0.91	53	47905	0.877	19	44569	0.893	25	48568	0.892	41	47537	0.924
12	+	48	45584	0.716	63	47858	0.756	65	47699	0.754	69	45838	0.761	45	47481	0.761	37	45615	0.901
13		3	42362	0.874	3	46118	0.873	5	39668	0.881	9	46248	0.876	10	44729	0.92	5	36413	0.956
14		5	44272	0.859	9	38226	0.901	8	36169	0.877	11	36957	0.876	12	44252	0.846	9	14320	0.936
15		16	30785	0.94	27	41196	0.931	35	34119	0.884	34	41144	0.899	34	38378	0.926	38	40439	0.925
16	+	33	36406	0.867	15	36228	0.979	12	36642	0.985	10	36020	0.999	5	11596	0.999	16	36829	0.935
17		35	36435	0.997	19	32287	0.9	22	34247	0.863	31	31843	0.867	26	35698	0.847	13	14870	0.937
18	*	3	12635	0.955	9	24931	0.926	11	28359	0.909	20	33812	0.891	18	33854	0.904	4	31484	0.974
19	*	24	29596	0.799	29	14492	0.934	20	23015	0.931	14	22885	0.775	9	21951	0.93	35	29083	0.976
20		18	26630	0.922	24	26959	0.9	28	26819	0.842	26	23514	0.91	9	11332	0.945	10	10188	0.914
21	*	25	25103	0.826	12	26101	0.808	9	14015	0.867	10	13200	0.667	28	23674	0.799	18	20338	0.906
22		3	3098	0.912	3	20529	0.784	3	21204	0.788	11	20972	0.851	8	18276	0.869	3	16957	0.937
23	+	1	52930	0.789	4	14863	0.962	5	15987	0.882	7	14256	0.9	7	16252	0.91	4	7592	0.938
24		5	3359	0.942	6	13002	0.904	5	12907	0.904	5	11743	0.923	8	14109	0.961	11	15487	0.993
25	+	1	52930	0.817	4	14863	0.924	4	7518	0.903	7	14256	0.874	5	9014	0.891	7	15331	0.902
AVG		22.8	43132.8	0.867	21.1	39575	0.883	21.7	39327	0.872	20.8	38568	0.858	18.8	41334	0.883	20.6	36205	0.946

TABLE VI. EVALUATION RESULTS FOR CATEGORY B (NG: NUMBER OF GROUP, NU: NUMBER OF USER AND RE: RMSE)

GR #		Combine 1			Combine 2			Combine 3			Combine 4			Combine 5			Combine 20		
		NG	*NU*	*RE*	*NG*	*NU*	*RE*	*NG*	*NU*	*RE*	*NG*	*NU*	*RE*	*NG*	*NU*	*RE*	*NG*	*NU*	*RE*
1	*	72	78872	0.836	71	78327	0.9	44	79451	0.847	58	75269	0.833	63	75465	0.843	45	72543	0.952
2		12	72955	0.911	12	64756	0.894	13	73448	0.936	19	73395	0.932	5	41422	0.941	14	74409	0.946
3		46	68452	0.818	30	62643	0.798	35	67611	0.766	30	70186	0.783	36	69717	0.773	54	70776	0.789
4		48	54502	0.966	55	63289	0.961	27	63533	0.97	28	57659	0.962	21	42646	0.968	15	61203	0.988
5		12	48126	0.762	20	63376	0.804	23	63794	0.764	25	62521	0.702	28	64155	0.701	53	60647	0.743
6	+	38	61301	0.809	16	57145	0.849	25	60966	0.835	31	59417	0.858	26	61061	0.889	35	39022	0.963
7	+	22	58255	0.760	6	59280	0.841	6	46516	0.84	5	59194	0.841	5	59194	0.841	11	53686	0.893
8	*	34	56246	0.939	37	56017	0.814	28	41998	0.735	28	46439	0.722	20	19448	0.822	47	55469	0.897
9	*	7	46254	0.877	9	38226	0.891	8	55024	0.884	17	50487	0.875	15	49643	0.916	34	52753	0.923
10	+	72	52385	0.796	50	51995	0.867	51	50913	0.86	33	44402	0.852	26	50837	0.857	28	53161	0.965
11		37	52593	0.996	49	39400	0.971	45	52423	0.941	46	52250	0.937	45	49277	0.936	39	52865	0.959
12		51	50985	0.775	56	39437	0.77	36	41850	0.774	46	52369	0.776	44	51683	0.792	45	49730	0.843
13		24	45192	0.774	16	45365	0.775	8	47235	0.735	10	46958	0.812	17	46258	0.754	39	36551	0.961
14	*	8	22495	0.983	23	45330	0.982	25	45340	0.954	19	38013	0.825	15	33322	0.825	23	40486	0.828
15	+	12	26199	0.895	15	42125	0.947	25	25721	0.947	26	42476	0.962	35	42541	0.951	4	22428	0.987
16	+	61	39591	0.681	30	37162	0.742	18	38811	0.797	32	38725	0.697	31	27094	0.77	26	39416	0.924
17		61	39085	0.891	29	24492	0.921	35	34680	0.819	28	39036	0.852	14	37660	0.889	5	35101	0.957
18		8	35097	1	13	31522	0.84	16	29488	0.882	17	28977	0.88	25	31884	0.845	32	34374	0.916
19		16	22340	0.851	10	26510	0.842	9	19094	0.853	11	18837	0.863	9	21352	0.881	2	20192	0.995
20		14	24536	0.787	14	25864	0.686	5	15272	0.728	5	19724	0.728	10	25392	0.789	1	6220	1
21		19	25968	1	17	25838	0.961	18	22876	0.866	17	22424	0862	19	25831	0.843	27	25856	0.926
22	*	15	20371	0.638	20	16200	0.802	21	21897	0.804	9	20862	0.783	16	20977	0.981	10	10131	0.989
23		8	16875	0.85	10	16471	0.795	13	18347	0.816	16	19123	0.88	15	19170	0.82	16	19199	0.802
24		4	8951	0.995	6	5498	0.998	11	13586	0.955	9	4736	0.968	11	7313	0.922	15	16498	0.983
25		17	12423	0.913	10	16062	0.901	16	8464	0.844	5	6418	0.933	4	14556	0.914	8	10577	0.954
AVG		28.7	41601.9	0.860	24.9	41293	0.862	22.4	41534	0.846	22.8	41996	0.844	22.2	39516	0.858	25.1	40532	0.923

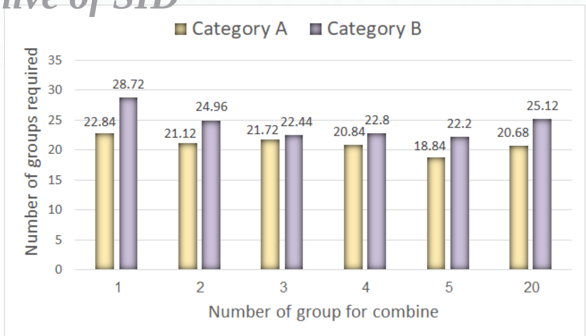

Fig 1. The average number of groups required for each combination in categories A and B

As shown in fig. 4, in all combinations, the two categories A and B acted equally, so that as category A increased or decreased, category B acted the same. As a result, the integrity of the model is shown based on this comparison. Of all the combinations, combination 5 in categories A and B requires fewer groups to achieve the target.

A more accurate comparison of the average number of groups required is the average number of users obtained by each of the proposed method combinations. According to fig. 5, Combination 1 and Combination 5 acted differently than the other combinations in the two categories A and B. But in other combinations, categories A and B have been integrated. In general, among all combinations, combination 1 in category A and combination 4 in category B were better able to achieve the number of target users. According to fig. 5, as the number of groups for word combinations increases, the number of end users (target) decreases, although this decrease is ascending and descending.

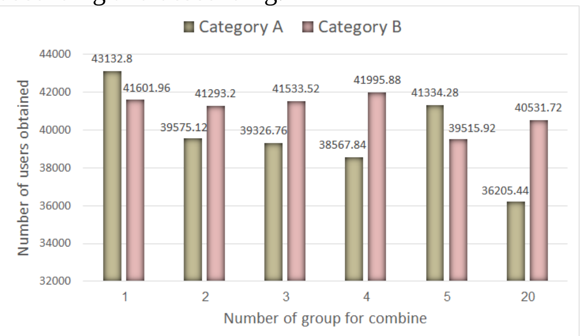

Fig 5. Average number of users obtained for each combination in categories A and B

V. CONCLUSION AND FUTURE WORK

In this paper, a method based on hybrid filtering by combining user membership graph and extracting keywords from groups' characteristics for users' recommendation is presented. The membership graph indicates the membership of users in Telegram groups. Also, the characteristics for each group show the name and description of that group in the telegram. The proposed method has two phases, offline and online. In the offline phase, there is a membership graph and a sack of words for the groups. We have created a bag of words for each group in the sack of words based on natural language processing methods. In the online phase, a set of users are first given to the system. Then, from the membership graph, a list of ranked groups of incoming users is obtained. The list of ranked groups obtained from the graph is combined with the results obtained in the offline phase. Finally, users are recommended from the end groups list. To evaluate the proposed method, we selected two categories of groups called A and B, each category consisting of 25 separate specialized groups. Also, these groups had between 2,000 and 10,000 members. The results of the evaluation indicate that the proposed method is able to provide accurate recommendations with low error and similar to incoming users. After analyzing the evaluation results, we found that if the incoming users to the recommender system are ranked based on the list of most members in the groups and then the keywords of the first 4 groups are combined, the system will have less error than other combinations. This shows that most related words are formed in the combination of words of 4 groups. In general, as the number of groups for word combinations increases, the average RMSE increases, the average number of groups required decreases, and the number of users obtained in each combination decreases. Of course, the diagram of these values is not always ascending or descending, there are rise and fall.

The proposed method focuses on the information of more than 120 million users and 900,000 supergroups. In order to develop and improve this study in the future, more users and groups can be considered. In the future, we can consider a separate score for the group's name and description. Furthermore, to improve the efficiency of the user recommendation, the content of the groups can be increased and the users' messages, the date and time of sending messages in the groups can be used. In the first step of the online phase, the initial groups can be considered based on the percentage of common members instead of the number of common members with incoming users.

REFERENCES

[1] P. Dashtizadeh, and Ali Harounabadi. "Recommending Friends in Social Networks By Users' Profiles And Using Classification Algorithms." International Journal of Information and Communication Technology Research, vol. 10, no. 1, pp. 56-61, 2018.

[2] K. Shin, Y. J. Park, K. M. Kim, and S. Kwon, "Multi-Manifold Learning for Large-scale Targeted Advertising System," arXiv preprint arXiv:2007.02334 (2020).

[3] P. Durov. (2020, Oct.) 400 Million Users, 20,000 Stickers, Quizzes 2.0 and €400K for Creators of Educational Tests. [Online]. Available: https://telegram.org/blog/400-million.

[4] J. Baumgartner, S. Zannettou, M. Squire, and J. Blackburn, "The Pushshift Telegram Dataset," In Proceedings of the International AAAI Conference on Web and Social Media, Vol. 14, 2020, pp. 840-847.

[5] A. Jalilvand and M. Neshati, "Channel retrieval: finding relevant broadcasters on Telegram," Soc. Netw. Anal. Min. vol. 10, no. 23, pp. 1-16, March 2020.

[6] A. Hashemi, and M. A. Z. Chahooki. "GroupRank: Ranking Online Social Groups Based on User Membership Records." Journal of AI and Data Mining, vol. 9, no. 1, pp. 45-57, 2021.

[7] A. Hashemi and M. A. Z. Chahooki, "Telegram group quality measurement by user behavior analysis," Soc. Netw. Anal. Min. vol. 9, no. 1, p. 33, July 2019.

[8] D. Karimpour, M. A. Z. Chahooki, and A. Hashemi. "Telegram group recommendation based on users' migration." In 2021 26th International Computer Conference, Computer Society of Iran (CSICC), pp. 1-6. IEEE, 2021.

[9] D. Karimpour, M. A. Z. Chahooki, and A. Hashemi. " GroupRec: Group Recommendation by Numerical Characteristics of Groups in Telegram." In 11th International

Conference on Computer and Knowledge Engineering (ICCKE), Accepted, 2021.

[10] M. Salehi, I. N. Kamalabadi, and M. B. G. Ghoushchi. "A New Adaptive Hybrid Recommender Framework for Learning Material Recommendation." International Journal of Information and Communication Technology Research, vol. 5, no. 3, pp. 25-33, 2013.

[11] F. S. Gohari, and M. J. Tarokh. "A New Hybrid Collaborative Recommender Using Semantic Web Technology and Demographic data." International Journal of Information and Communication Technology Research, vol. 8, no. 2, pp. 51-61, 2016.

[12] X. Cai, M. Bain, A. Krzywicki, W. Wobcke, Y. S. Kim, P. Compton, and A. Mahidadia. "Learning collaborative filtering and its application to people to people recommendation in social networks." In 2010 IEEE International Conference on Data Mining, pp. 743-748. IEEE, 2010.

[13] A. Zare, M. R. Motadel, and A. Jalali, "Presenting a hybrid model in social networks recommendation system architecture development," AI & SOCIETY, vol. 35, no. 2, pp. 469-483, June 2020.

[14] P. Kumar, and G. R. M. Reddy. "Friendship recommendation system using topological structure of social networks." In Progress in Intelligent Computing Techniques: Theory, Practice, and Applications, pp. 237-246. Springer, Singapore, 2018.

[15] O. Kaššák, M. Kompan, and M. Bieliková, "Personalized hybrid recommendation for group of users: Top-N multimedia recommender," Information Processing & Management, vol. 52, no. 3, pp. 459-477, May 2016.

[16] S. Huang, J. Zhang, L. Wang, and X-S. Hua. "Social friend recommendation based on multiple network correlation." IEEE transactions on multimedia, vol. 18, no. 2, pp. 287-299, 2015.

[17] M. Bertini, A. Ferracani, R. Papucci, and A. D. Bimbo, "Keeping up with the Influencers: Improving User Recommendation in Instagram using Visual Content," In Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization, 2020, pp. 29-34.

[18] L. Berkani, "A semantic and social‐based collaborative recommendation of friends in social networks," Software: Practice and Experience, vol. 50, no. 8, pp. 1498-1519, August 2020.

[19] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," Communications of the ACM, vol. 18, no. 11, pp. 613-620, November 1975.

[20] W. H. Gomaa, and A. A. Fahmy, "A survey of text similarity approaches," International Journal of Computer Applications, vol. 68, no. 13, pp. 13-18, April 2013.

[21] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532-1543.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781 (2013).

[23] R. Singh and S. Singh, "Text Similarity Measures in News Articles by Vector Space Model Using NLP," Journal of The Institution of Engineers (India): Series B, pp. 1-10, November 2020.

[24] I. Rushkin, "Document Similarity from Vector Space Densities," In: K. Arai, S. Kapoor, R. Bhatia (eds) Intelligent Systems and pplications. IntelliSys 2020. Advances in Intelligent Systems and Computing, vol 1251. Springer, Cham, 2020, pp. 160-171.

Davod Karimpour received his B.Sc. degree in Computer Engineering from Birjand University, Birjand, Iran. And his M.Sc. degree in Software Engineering from Yazd University, Yazd, Iran. He did his M.Sc. in AI lab (with Dr. Chahooki) at Yazd University. His research interests include Data Analysis, Big Data, Recommender Systems, Social Network Analysis, and Deep Learning.

Mohammad Ali Zare Chahooki, Currently is Associate Professor in Faculty of Computer Engineering at Yazd University. His research interest is Machine Learning in Software Engineering, Image Retrieval, and Text Mining. From 2015, he is Idekav funder which is a platform for marketing on Telegram. Telegram is a messenger and social network with more than 550 million users. From Idekav, Now there are requests from customers for finding target peoples among 200 million users which are found from Telegram groups.

Ali Hashemi is a Software Engineering Doctoral Graduate from Yazd University. He received his B.Sc. degree in Software Engineering and M.Sc. degree in Computer Networks both from Yazd University. He is interested in Large-Scale Distributed Systems, Recommender Systems, and Search Engines.