International Journal of Electrical and Computer Engineering (IJECE)
Vol. 12, No. 2, April 2022, pp. 1881~1892
ISSN: 2088-8708, DOI: 10.11591/ijece.v12i2.pp1881-1892

Cosine similarity-based algorithm for social networking recommendation

Shaha Al-Otaibi, Nourah Altwoijry, Alanoud Alqahtani, Latifah Aldheem, Mohrah Alqhatani, Nouf Alsuraiby, Sarah Alsaif, Shahla Albarrak
Department of Information Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia

Article Info

Article history:

Received Jun 9, 2021
Revised Sep 8, 2021
Accepted Oct 2, 2021

Keywords:

Cosine similarity
Feature extraction
Social media
TF-IDF
User profiling
Virtual community

ABSTRACT

Social media have become a discussion platform for individuals and groups. Hence, users belonging to different groups can communicate together. Positive and negative messages as well as media are circulated between those users. Users can form special groups with people who they already know in real life or meet through social networking after being suggested by the system. In this article, we propose a framework for recommending communities to users based on their preferences; for example, a community for people who are interested in certain sports, art, hobbies, diseases, age, case, and so on. The framework is based on a feature extraction algorithm that utilizes user profiling and combines the cosine similarity measure with term frequency to recommend groups or communities. Once the data is received from the user, the system tracks their behavior, the relationships are identified, and then the system recommends one or more communities based on their preferences. Finally, experimental studies are conducted using a prototype developed to test the proposed framework, and results show the importance of our framework in recommending people to communities.

Corresponding Author:

Shaha Al-Otaibi
Department of Information Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University
Riyadh, Saudi Arabia.
Email: stalotaibi@pnu.edu.sa

1. INTRODUCTION

Social networks are a set of platforms that enable online users to communicate with each other and exchange their thoughts and experiences. Users join these platforms and create profiles that identify their interests and preferences. Several approaches have been used to establish these profiles. The most commonly used way is creating a profile by letting the user answer targeted questions asked by the system [1]. Social networking platforms have recently started having a critical impact on people and society. They can help users create their own community based on interest, age, and ideas, and hence, users from diverse backgrounds and different countries have access to each other [2].

The vast volume of data circulating in the various social networks results in a rich social structure that encourages users to join various communities and helps them make appropriate decisions [3]. However, this huge volume of data may cause many problems, such as the individual being unable to find their interest or intended purpose easily and spending too much time to find what they are looking for. Establishing communities on social networking platforms based on people's interests is a possible solution for such problems. In these communities, people can introduce themselves in certain ways and join those that fit their interests based on system recommendations. The individuals who use such networking thus become part of

communities whose members share the same interests and preferences. Through conversations, people connect and form relationships in convenient ways. They exchange knowledge, experiences, hobbies, comments, and video streams in organized ways. Finally, such networking saves people time, lets them interact easily with fewer distribution, and increases awareness among the community members [4].

In this article, we will present a framework using which communities can be recommended to users based on their preferences. It utilizes a feature extraction algorithm based on user profiling and combines the cosine similarity measure with term frequency. It helps establish relationships on social networks according to user preferences. The framework aims to let people join communities according to features extracted from user profiles and their interaction data. It also enables users to exchange text and media with people in the same community. We developed a prototype with a user-friendly interface based on the proposed framework to perform system testing.

The rest of this article is organized as follows: section 2 describes the problem statement and motivation. Section 3 demonstrates the research background, following which section 4 presents our findings from the literature review to explore this problem and determine the important related aspects. Then, we present the proposed framework that was applied to recommend communities to users. Finally, the proposed framework is tested using a prototype that was developed to evaluate the algorithm's accuracy.

2. PROBLEM STATEMENT AND MOTIVIATION

The existing social networking platforms are huge channels that are available for everyone, with no restrictions on location, time, or interests. Therefore, individuals cannot find their interests or intended purpose easily and spend long periods on social networking sites, which may affect their health and waste their time. It can negatively affect mental health in various ways, including symptoms of anxiety and depression. Additionally, people sometimes cannot find the appropriate community to share the needed knowledge or tips, and this may frustrate them or distract them from their primary goal [4]. Moreover, several problems may arise from conflicts of interest, such as bullying and receiving negative messages. Moreover, there is a limitation for people with regard to sharing their interests or thoughts on social media; just because they feel they are different does not mean others will also care about their interests. In fact, making social networking more interactive by effectively enabling people to collaborate and exchange knowledge will play a crucial role in sharing interests and useful thoughts in order to solve problems and disseminate knowledge between community members.

3. BACKGROUND INFORMATION

Social networking platforms emerged a few years ago, and they have numerous positive impacts on the lives of community members. One of the most essential features is that they use web-based social media to create highly interactive platforms that help community members share various kinds of content that can be discussed and modified directly [5]. Although the key technological feature is consistent, there is a remarkable observation that the diverse cultures in social networking help promote contact between various people based on the beliefs, views, and activities of common interests. Most platforms engage various categories of people according to common languages, religion, ethnicity, or even nationality [6].

Social networking platforms have evolved over time; many programs have emerged to help create communities, while the collective use of social networking has facilitated the interaction and reduced the time required to search, collect, and share information. As more members participate actively in the online community, more benefits are accumulated for each member and the community at large. At the beginning, the idea of an online community emerged because of people's needs for information, support, and relationships. Additionally, depending on the type of need, interested individuals or a group of friends can begin forming a vision for a community where people can exchange information, communicate, and interact [7], [8].

Building virtual communities is unlike those in the real world; in real communities, people connect with each other face to face, then get to know them, and if they are in the same mood, they form relationships. Whereas, in an online community, people know others after they form relationships and then, if they are in the same mood, go to meet them. Furthermore, online community sharing is driven by optional choices that are different from real communities where relationships may be enforced involuntarily by geographical location [7].

Many studies on online communities confirm that many people are attracted to the internet for social interactions. There are many benefits when people are part of online communities. The first class of benefits involves the inherent benefits derived from the establishment of a social group, such as opportunities to exchange information, send and receive social and emotional support, enhance friendships, and have fun. For

instance, a community of researchers exploring asynchronous learning networks uses the platform to exchange information and comment on effective ways to achieve their goals through the network. Moreover, people of online communities receive and provide social and keen endorsements for each other, and online communities present interactive fun opportunities for participants. The second class of benefits comes from media and technology. The Internet and its applications offer, worldwide access, asynchronous interactions, typing correction capabilities, and permanent storage capacity. In online communities, people can communicate and interact comfortably with other people who live far away at any time. People also have the ability to view profiles of people and leave comments so that they can take part in any ongoing conversations directly after they join a group, thereby finding self-satisfaction and fulfilling their interests [8].

4. RELATED WORK

This section introduces the research that related to build communities in social networking. We will present extant research and justify the technologies used as well as clarify some of the other related aspects. The virtual community (VC) concept was developed early decade and is defined as a group of people who communicate via electronic group discussions. In fact, building VCs in social networks is considered a multi-disciplinary concept, which is hard to define, thus resulting in many definitions identifying the core attributes of VC as a shared goal, interest, need, or activity, which is the main reason for joining such communities. These communities offer repeated active participation and, often, intense interactions, strong emotional ties, and shared activities among participants, as well as access to shared resources and policies determining access to those resources. From a technological perspective, VCs enhance communication by providing ubiquitous, inexpensive or free, and fast communication, file sharing, public access services, voice chat facilities, audio and video conferencing, and virtual reality experiences. Finally, a VC is mainly formed for four purposes: transaction, fantasy, interest, and relationship [9]. Transaction mainly enables selling and buying of products and services; interest gives participants ability to interact on specific topics; fantasy establishes new characters, environments, and stories; additionally, relationships are founded around specific life experiences such as death or threatening disease.

People are increasingly using computer-based communications, which have been applied for a long time in some areas, such as healthcare services, where the Internet can change the way users carry out their roles and communicate with each other as well as with service professionals. Hence, patients and healthcare providers have increasingly turned to peer-to-peer communication by building communities and connecting with their online fellows to share experiences and knowledge. Professional service meetings that entail interaction between a patient and their healthcare provider can reach virtual communities to disseminate healthcare information and obtain support. Thus, people gather digitally to speak and listen, satisfy their curiosity, or form and maintain relationships. It also provides an opportunity to compare healthcare systems, diagnoses, and treatment regimens and provide mutual support and advice. Virtual communities are abundant in the healthcare sector. Therefore, virtual health care leads to knowledge building and active participation in e-health, e-medicine, and e-detailing activities [10].

In the same vein, the education sector has delivered various online communities, for instance, to link teachers from 10 different schools online and enable collaboration in curriculum planning and delivery, utilizing a site for resource sharing and communication. Such communities are built within reliable, familiar, and supportive environments where teachers can participate and develop a view of what the practice of sharing online requires [11]. Furthermore, gaming applications have found massive online social communities by encouraging new users, helping them develop new strategies, and offering novel content. They have a significant impact on the members and enjoy the active support of the engaged individuals. Hence, the participants can gain benefits from their membership in online communities where ideas and hints are exchanged, strategies discussed, stories heard, and experiences shared. The authors of [12] stated that game developers have realized how online communities associated with their games are important to their success. For example, this can be an opportunity to join like-minded people and discuss shared interests. They offer an abundance of resources, such as knowledge, opinions, and personal experience, as well as they allow members to share and obtain higher status in their community. Hence, the motivation to play these games is intrinsic to the creation of a comfortable community [12]. Moreover, friends' influence on social networking plays an important role in product marketing. However, it has rarely been considered in traditional recommendation systems. A new paradigm has been utilized for recommendations in social networking platforms based on peer influence, user preferences, and an item's general acceptance [13].

4.1. User profiling

The goal of many Internet-related research activities is to transform the web into a user-friendly environment where people can easily find what they are looking for. In many cases, this has translated into

the task of finding useful information quickly and efficiently. With the growing amount of information available online, this task becomes a challenge. User profiles are considered a mean of filtering information and providing the most desirable and effective ways for delivering information that fits user needs and requirements. A very important aspect of that approach is the process of matching user profiles against information retrieved from the web. User profiling is employed to identify a user's interests and preferences. Hence, there are several suitable approaches for constructing a profile. The most intuitive one is developing a profile by asking a user several targeted questions. However, the user may not be willing to provide information frequently, and moreover, a user's interests change constantly. The process of constructing user profiles can be divided into two categories: the knowledge-based approach and the behavior-based approach. The knowledge-based approach constructs user profiles by including server-side accounts and the identity of user profiles. The behavior-based approach models a user profile in a binary fashion and develops it based on the user evaluation of pages as interesting or uninteresting via machine learning techniques. These approaches can utilize and integrate information about the user [1]. For example, an approach that proposed for mining the differences in customer's behavior to maintain a strong profiling model over period in time in B2C relationship [14]. This approach captures the user's interests from their online activities rather than force them to explicitly give their preference. It applied a decision tree generation method which accepts a set of variables: continuous, categorical, and fuzzy variables as input, as well as considers the customer's reviews rates as classes.

Building any recommendation system based on user profiles should consider the social interactions between users and their personal preferences, which help develop friendships. The increasing data available through online social networking create obstacles for mining user interests and behaviors for recommendation for example, the dynamic nature of the generated data that reflects changes in user activities [15]. Moreover, profiling users on online social networks requires data related to their online behaviors. Their behaviors may depend on their interests, or they can be affected by other influences on them. To profile users based on their behaviors on various social networks, we need the entire accumulated data of all social networking sites, which is hard for researchers to collect. However, based on the activities performed through one particular site for a significant period, the user can be profiled using separate approaches [16].

Many researchers have developed similarity algorithms between users and items due to the ease of computation [17]. For instance, a similarity algorithm developed in [18] that calculates the neighborhood similarity based on collaborative filtering of all the ratings produced by a couple of users. Personality similarity has been examined in some research. For instance, similarity is addressed in the couple context, where researchers have examined personality similarity and dissimilarity as dyadic predictors of spousal quality, aiming to detect the compatibility issues between spouses. Couples may be similar in some personality traits but dissimilar in others. The studies observed positive correlations between spousal quality and spouses' similarity in several domains. Some of these studies have observed the positive effects of personality similarity on spousal quality. For example, couples with similar personality patterns reported better marital quality; to measure personality similarity, the researchers used different methods. The commonly used method is profile-level similarity which represents the similarity between two profiles across multiple characteristics. Previous studies have also applied various measures to obtain profile similarity in personalities from various perceptions. Another new measure of profile similarity, the double-entry intraclass correlation, is selected by researchers, because it is more sensitive to different features of profile similarity [19].

4.2. Recent studies in social networking recommendation

Many studies have been conducted for different recommendation problems based on user profiles and preferences. In this section, we will present some recent studies in this domain. There are two well-known clustering algorithms that have been used to identify communities: hierarchical clustering and k-means [20]. The hierarchical algorithm is very slow in handling large-scale datasets, such as users on Twitter, whereas k-means is quick and effective. Therefore, the authors apply k-means for clustering users. Additionally, an online study was conducted to analyze pictures on Instagram and their contents, which were clustered with labels applying the k-means clustering approach. Using a total of 17 clusters, their relationship with users' personality traits was analyzed. A relationship between picture content and personality traits was identified. The study suggested a new way to extract personality traits from social media trials [21]. The web performance optimization system proposed to enhance the user experience when browsing a website by applying the users' ratings. The different user experience factors were identified using various techniques and given ratings. All ratings of different factor are combined to determine the overall rating of the website that can be used to optimize the user experience [22]. Moreover, the detection method based on a scalable

framework identifies related communities on social networks. It is based on a multilevel clustering algorithm that uses structural and textual information to determine local communities [23].

An event recommendation algorithm was proposed to combine the data collected by the collaborative filtering (CF), the internet of things (IoT) devices and social influence. The algorithm recommends an event to the user with the help of a prediction score for IoT-based factors and CF based on social influencers. Hence, the event with the highest prediction score is recommended to the user [24]. A travel recommendation technique based on a machine learning classifier that processing user activities in social networking is presented [25]. It presents personalized recommendations for a place of interest that satisfies the user's specific needs and preferences. It was applied on Twitter data, considering their followers in a timely manner, to pattern the latest travel interests. Additionally, a session-based recommendation algorithm for social networking that understands the users' preferences and provides an accurate recommendation based on user interests. The algorithm predicts the next session for the user based on many other sessions of the user's friends to consider for the social influences. It predicates the upcoming session by utilizing a graph neural network and item representations that extract knowledge from social networking. Then, the prediction is generated [26]. A hybrid algorithm for community detection was developed by fusing topology and a non-negative matrix factorization algorithm. It used the attribute similarity matrix and the attribute feedback change method to fuse the topological matrix and the attribute matrix. Then, it inferred the iteration formula of each matrix by applying a special mathematical method to be used in a community detection process [27]. Furthermore, the approach of a hybrid microblog recommender system utilized a deep neural network and a group of diverse features associated with user interests. Moreover, a collaborative filtering was used to find the candidate microblogs for final recommendations. Then, a deep neural network with multiple hidden layers was developed to predict and rank the microblogs [28]. The study proposed in [29] that analyzed the collective user behavior on Facebook using the visible communication. It studied the progress of human behavior through visible detection and non-volatile interactions. It applied the breadth first search method and a semi-supervised crawler agent.

Social networking needs to get benefits from innovative technologies that utilize every human sense [30]. The work suggests the using of augmented reality to connect people through various human senses, such as vision, hearing, touch, and taste. These advanced solutions will produce novel trends in information technology and social networking in a unified and multidisciplinary manner. For example, a deep learning framework trains convolutional neural networks as a part of an artificial intelligent characteristic of a cooperative game intending to inspire the white-collar workers to exercise their hands and wrists regularly by playing a game. This network aims to classify a still image into one of the six predefined classes of gestures, and it looks to cope well with small variations in size, skin tone, position, and orientation of the hand [31]. An architecture focuses on finding the overall sentiments and feelings of related topics with reference to a given topic was developed [32]. It combines sentiment analysis and community detection to obtain the overall sentiment of related topics. The authors applied that model on shopping, politics, COVID-19, and electric vehicles to understand emerging issues, trends, and their potential business, marketing, and political implications [32]. Moreover, a similarity-based algorithm that integrates user rating value and user behavior has been proposed [33]. The user behavior is acquired from the user probability's value when evaluating an individual's data. This study proposes a new similarity algorithm that explores data types and user profile data, such as gender, age, location, and occupation. The user profile data are applied to compute the weights of the similarities of user behavior and user rating values. The weights of both similarities are acquired by computing the correlation coefficients between the user profile data and the user behavior or rating values.

5. COSINE SIMILARITY-BASED ALGORITHM FOR RECOMMENDING COMMUNITIES

In this section, we will demonstrate the system architecture for the proposed solution, representing the processes that define the framework of the solutions against business requirements. It ensures that the application design is reliable, manageable, and scalable [34]. Because we planned to develop a prototype based on a mobile application, we chose an application architecture comprising three layers: the presentation, business, and data access layers. Layered architecture has many advantages, such as easy assignment of different roles, updating and improvement of layers separately, and it provides reliability and independence of the underlying servers. Figure 1 represents the proposed system architecture. The process can be described as follows:

- The user subscribes to the system, and the form to be filled out contains some information, including personal data, such as name, gender, location. Additionally, the system prompts the user to give a description of their interests or preferences.

- The user profile will be constructed using appropriate representation, such as an object of the term-frequency-inverse document frequency (TF-IDF) vector (TF-IDF is used in this study and will be explained in the below section).
- According to the initial data, the communities that are appropriate for the user will be recommended using the matching algorithm. In addition to the user profile information, the proposed algorithm explores the user's upcoming activities, such as posts, messages, emails, and any interaction to optimize the recommendation process. More accurate communities are suggested on the basis of future user behavior.

The following paragraphs provide an overview of how the proposed system was implemented.

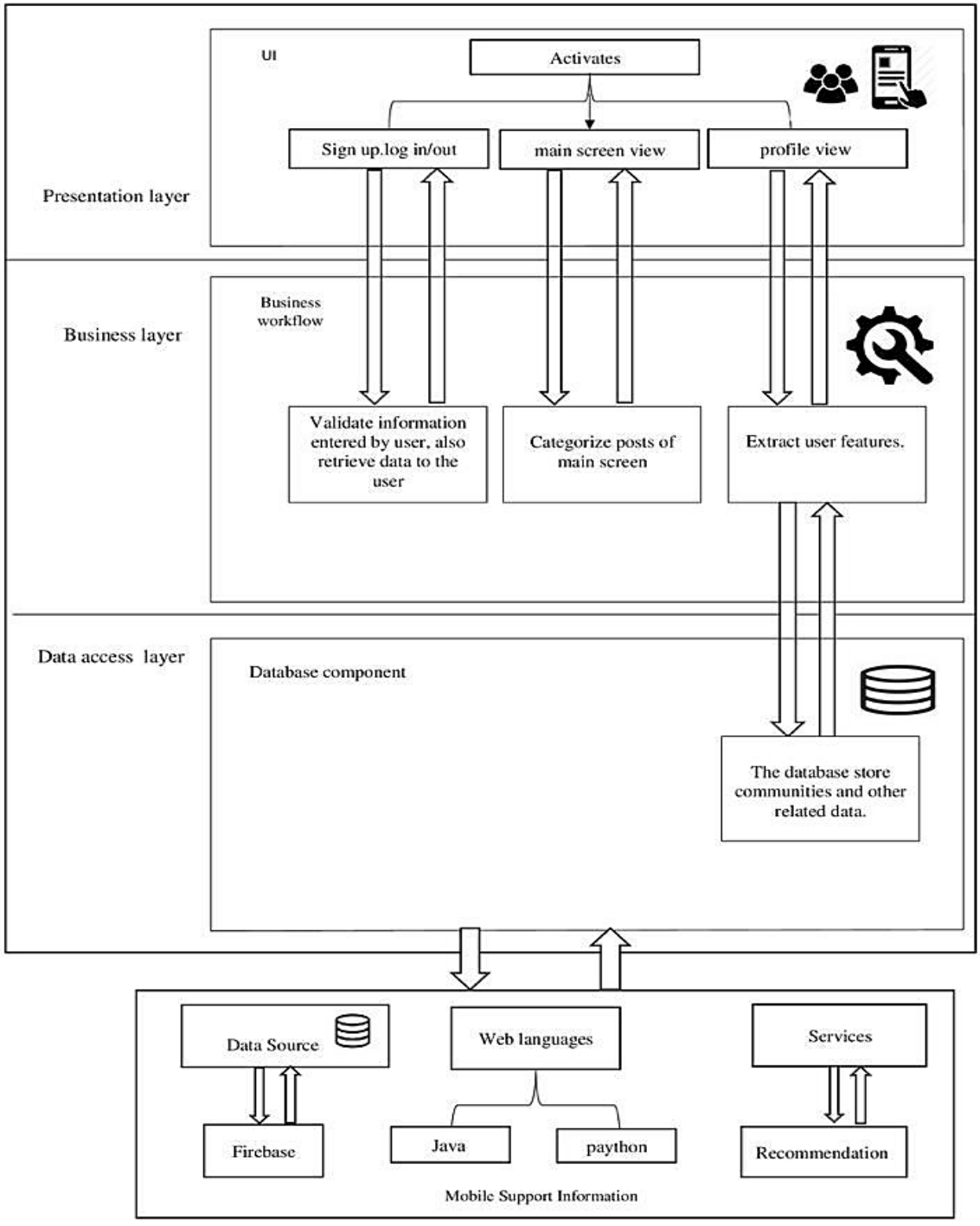

Figure 1. The proposed system architecture

5.1. Profiling and user behavior

In this section, we present an overview of how to profile users and take their behavior into consideration. As mentioned above, profiling refers to the process of building user profiles through systematic data analysis [35]. It is accomplished using algorithms or mathematical techniques that enable pattern discovery in a large amount of data. Then, these correlations or patterns are used to define or represent user profiles. One of the most challenging problems of the information domain is the increasing data overload. It becomes important for companies, governments, and individuals to differentiate information from noise to detect useful and interesting data. Therefore, we need to use technologies to analyze the collected data and explore the knowledge in that data [36]. The processes that are used for profiling include the following steps:

- Initial stage: the profiling process starts with the specification of a problem domain and the identification of the purpose of analysis.
- Data collection: the target dataset is created by choosing the relevant data that matches the users' needs.
- Data preparation: the data are pre-processed to eliminate noise and reduce complexity by removing some features.
- Data mining: a data mining algorithm that fits the data, model, and goals is developed.
- Interpretation: the mined patterns are evaluated to assess their relevance and validity.
- Application: the constructed profiles are applied to solve certain problems, for example, community recommendations.
- Decision: this determines the actions that can be applied to individuals or groups whose data match a relevant profile [37].

In this work, profiles are classified according to the users' preferred areas of interest. When a profile is built with the data of a single person, it is called individual profiling. This type of profiling is used to discover specific characteristics of one individual and helps enable unique identification or the establishment of personalized services. However, personalized services most often also dependent on group profiling, which is represented by communities; this facilitates community recommendations to a certain user depending on the extent to which their profile matches the community's specifications. A group profile can refer to a category of people that share certain patterns of behavior or characteristics [36].

Individual user profiles keep updating based on their behaviors. Hence, user behavior is acquired in an accumulative fashion by monitoring their interactions. User behaviors such as their posts that reflect their interests and preferences make for an efficient and useful way to suggest or recommend a new community to users. Three important issues listed below can be studied to realize the practicality of the recommendation algorithm.

- Accumulating user behaviors: the information must be accumulated in advance (profiling) depending on the users' active participation.
- Profile representation: the representation of the profile that stores the information is an important concern.
- Matching algorithm: this is the direct product of the suggested framework.

5.2. Term frequency inverse document frequency

A well-known and broadly utilized weight method is term frequency- inverse document frequency. Denoted as $TF_{t,d}$, it gives number of occurrences of a certain term t in a document d. The best method is to assign a weight for the number of occurrences of the term t in document d. The more frequent it is, the more relevant it should be. Term frequency (TF) introduces a crucial issue: all terms are considered equally significant. In reality, some terms have small or no discriminatory power to determine the relevance. To resolve this issue, the document frequency (DF) method is modified and is indicated as DF_t. DF_t denotes the number of documents within the collection including the term t and N is the number of documents in the collection. To alter the term weight utilizing the measure DF, the inverse of the frequency in the documents is defined (IDF_t) as shown in (1) and (2).

$$\mathrm{IDF_t} = \log \frac{\mathrm{N}}{\mathrm{DF_t}} \quad (1)$$

$$\mathrm{TF - IDF_{t,d}} = \mathrm{TF_{t,d}} \times \mathrm{IDF_t} \quad (2)$$

TF-IDF computes the relative frequency of terms in a certain document compared to the reverse extent of the same term over all documents in the corpus. This calculation decides how important a given term t is in a certain document d, represented by (2). TF-IDF denotes the term *t* as a weight in document d that is: (i) most elevated once t happens many times in a small number of documents, (ii) low when the term

happens fewer times in a document or happens in many documents, and (iii) the least when the term happens in all documents [17].

5.3. Cosine similarity measures

Cosine similarity is a vector-based scheme that can be used to measure the similarity between two strings. The primary concept in cosine similarity is to convert the two strings into vectors in the multi-dimensional space. The cosine of the angle between the two vectors is measuring how "similar" they are, which in turn is measuring of the similarity of those strings. If the vectors are of unit length, the cosine of the angle between them is simply the dot product of the vectors. There are numerous ways to convert a string into a vector. The TF-IDF vector is a common option for this representation, as it is made up of the product of TF and IDF for each term that is shown in the string [38].

The cosine similarity is a metric that qualifies the similarity between two vectors. It is computed the cosine of the angle between two vectors and decides whether two vectors indicate the same direction. It is utilized to measure the similarity between documents. Hence, each document is an object represented by a TF-IDF vector. We can calculate the cosine similarity using (3):

$$\text{sim}(\text{x},\text{y}) = \frac{\text{x}\cdot\text{y}}{\|\text{x}\|\|\text{y}\|} \quad (3)$$

where ||x|| is the Euclidean distance of vector $x = (x_1, x_2, x_3 \ldots, x_p)$, defined as $\sqrt{x_1^2 + x_2^2 + \cdots 1 + x_p^2}$. Conceptually, it is measured the length of the vector. Similarly, ||y|| is the Euclidean distance of the vector y. Then, the similarity is measured by calculating the cosine of the angle between vectors x and y. A cosine value 0 implies that the two vectors are at 90 degrees from each other (orthogonal) and have no match. The closer the cosine value is to 1, the smaller the angle and the greater the match between the vectors, as shown in Figure 2 [39].

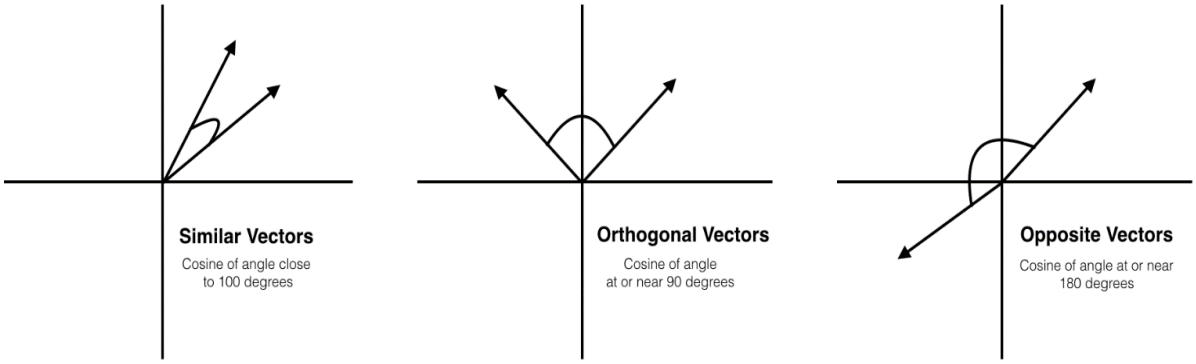

Figure 2. Type of matches

In this article, the cosine similarity was used to measure the similarities between the two vectors, the user and the community. Additionally, TF-IDF was used to build these vectors to be matched. The user profiles and their behaviors represent one vector for each user, which was then matched with all the vectors of information related to certain social networking communities. Finally, the similarity was calculated based on the angles between these vectors. Where the angle was small, the similarity was higher.

The cosine similarity algorithm was developed using Python supported by the chaquo library. In Figure 3, we present the Python code used to run the cosine similarity with the TF-IDF code. Initially, we called the Sklearn library, and then we started defining the cosine similarity parameters (*pairwise_similarity*). First, however, we had to compute the $tfidf$ using $TfidfVectorizer()$, which is a function used to calculate the TF-IDF and convert the given document into vectors. Next, we used $fit\_transform()$ to join the $fit()$ and *transform()* methods to transform the dataset. The $fit()$ function computes the μ and σ parameters and saves them as internal objects. The $transform()$ function uses these calculated parameters and applies the transformation to a dataset. After computing the *tfidf*, we compute the *pairwise_similarity* using the dot product between $tfidf$ and $tfidf.T$.

5.4. Multi-language support

Android is a popular mobile operating system and has millions of users in over 190 countries [40]. We aim to build solutions supporting both the Arabic and English languages. Android considers English the primary language by default and loads the string resources from res ⇒ values ⇒ strings.xml. Adding a new language in this case, Arabic needs to create new resource file strings.xml/ar to initiate support for the Arabic language and allow the translation of all strings into Arabic. Hence, the text feeds from strings.xml can benefit the ability of translating messages into the specified language or what we define as a synonym that appears to the user. The communities will be recommended using the proposed algorithm, as shown in Figure 4. After joining the communities, the user can create a post within any community by sending text, photos, or media.

```
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer


def pairwise_similarity(descriptions):
    documents = descriptions
    print(descriptions)
    tfidf = TfidfVectorizer().fit_transform(documents)
    # no need to normalize, since Vectorizer will return normalized tf-idf
    pairwise_similarity = tfidf * tfidf.T
    arr = pairwise_similarity.toarray()
```

Figure 3. Cosine similarity and TF-IDF in Python

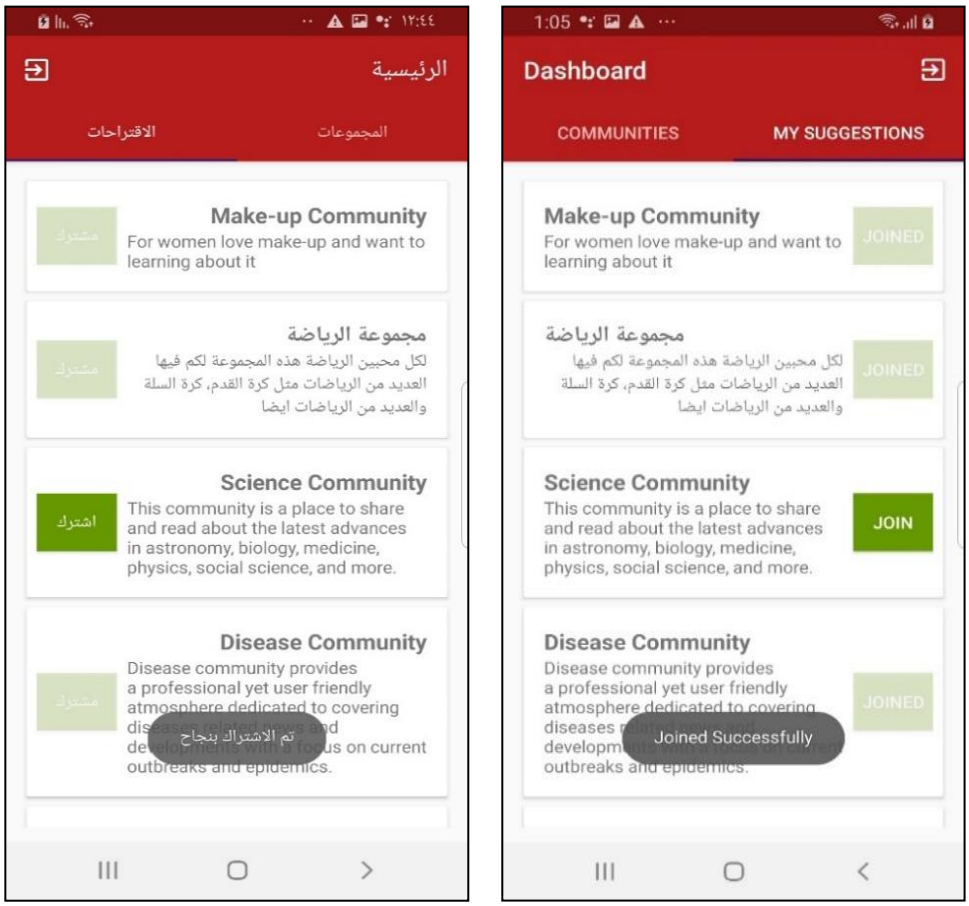

Figure 4. Successful subscription to a social networking community in Arabic and English

6. EXPERIMENTAL RESULTS

The proposed framework was tested using the developed Android prototype. Testing the algorithm's accuracy required a training dataset, a testing dataset, and an independent piece of code as a benchmark to run the algorithm. The experiment was applied to 80 user profiles, and their behaviors were tracked manually against 10 communities. Table 1 represents the matching communities for 6 users as an example. Therefore, the same method was applied to all users in the test dataset. The algorithm is tested in terms of its ability to classify the correct community assigned in the manual test. Table 2 shows the cosine similarity representing the result of calculation between the users and the recommended communities' vectors that appear in Table 1. The confusion matrix was used to measure the algorithm accuracy as shown in Table 3. It constitutes information on actual and predicted classifications performed via a proposed system.

Table 1. The matching communities for 6 users

User	Suggested Communities
1	programming, disease, science, biking, make-up
2	make-up, food community, video games, football, disease
3	biking, مجموعة مدينة الرياض
4	مجموعة مدينة الرياض، مجموعة الصحة.
5	football, video games, biking, science, programming
6	مجموعة مدينة الرياض، مجموعة الصحة

Table 2. Cosine similarity between the users and communities' vectors

User	Cosine similarity
1	[0.16], [0.11], [0.11], [0.05], [0.04]
2	[0.14], [0.11], [0.09], [0.05], [0.04]
3	[0.25], [0.11]
4	[0.15], [0.08]
5	[0.26], [0.18], [0.10], [0.10], [0.10]
6	[0.05], [0.47]

Table 3. The confusion matrix

	Predicted: YES	Predicted: NO
Actual: YES	TP	FP
Actual: NO	FN	TN

(TP) is the rate of positive cases that were correctly identified
(FP) is the rate of negatives cases that were incorrectly classified
(TN) the rate of negatives cases that were correctly classified
(FN) is the rate of positive cases that were incorrectly classified as negative

As mentioned above, we conducted a sample test on the selected communities and fetched approximately 80 user profiles manually. As shown in Table 4, we classified them into 4 sections; the actual side represents the result of the manual test, and the predicted side represents the result of the proposed system. Table 4 shows the calculated accuracy and error rate using (4) and (5). Accuracy rate (AC) is the rate of the total number of predictions that were true.

$$(AC) = \frac{TP+TN}{TP+FP+FN+TN} \quad (4)$$

$$(Error\ rate) = \frac{FP+FN}{TP+FP+FN+TN} \quad (5)$$

Table 4. The resultant confusion matrix

	Predicted: Yes	Predicted: No
Actual: Yes	72	6
Actual: No	2	0

Error rate is the rate of the total number of predictions that were wrong [41]. The experiment shows that the algorithm can be interactive and offers a high level of accuracy. By applying (4) and (5), we get a high accuracy rate of around 90% and an error rate of 10%, which reflects a high performance.

7. CONCLUSION

This article presented a framework that can be used to recommend communities to users based on their preferences. It applied a feature extraction algorithm that utilized user profiling and combined the cosine similarity measure and TF-IDF vector method to recommend groups or communities. The user profiles and their behaviors when interacting with the system or inside communities were applied to achieve the goal. Therefore, social networking communities were formed based on people's interests; hence, the users are able to introduce themselves in an adequate way and choose communities that fit their interests. People can use such social networks to become part of communities with members who share the same interests and preferences. Then, these social communities, the resulting conversations, and subsequent relationships will connect people in a convenient way. Finally, the algorithm created based on this framework was tested using a developed prototype. It achieves a high accuracy rate of around 90% and error rate of around 10%, which indicates high performance.

ACKNOWLEDGEMENT

This research was funded by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University through the Fast-track Research Funding Program.

REFERENCES

[1] M. Z. Reformat and S. K. Golmohammadi, "Rule-and OWA-based semantic similarity for user profiling," *International Journal of Fuzzy Systems*, vol. 12, no. 2, pp. 87–102, 2010, doi: 10.30000/IJFS.201006.0001.

[2] S. Edosomwan, S. K. Prakasan, D. Kouame, J. Watson, and T. Seymour, "The history of social media and its impact on business," *The Journal of Applied Management and Entrepreneurship*, vol. 16, no. 3, pp. 79–91, 2011.

[3] S. Al-Otaibi *et al.*, "Customer satisfaction measurement using sentiment analysis," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 2, pp. 106–117, 2018, doi: 10.14569/IJACSA.2018.090216.

[4] A. S. Kümpel, V. Karnowski, and T. Keyling, "News sharing in social media: a review of current research on news sharing users, content, and networks," *Social Media and Society*, vol. 1, no. 2, Jul. 2015, doi: 10.1177/2056305115610141.

[5] C. Oh, S. Sasser, and S. Almahmoud, "Social media analytics framework: the case of Twitter and super bowl ads.," *Journal of Information Technology Management*, vol. 26, no. 1, pp. 1–18, 2015.

[6] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, Oct. 2007, doi: 10.1111/j.1083-6101.2007.00393.x.

[7] H. Jang, L. Olfman, I. Ko, J. Koh, and K. Kim, "The influence of on-line brand community characteristics on community commitment and brand loyalty," *International Journal of Electronic Commerce*, vol. 12, no. 3, pp. 57–80, Apr. 2008, doi: 10.2753/JEC1086-4415120304.

[8] A. Iriberri and G. Leroy, "A life-cycle perspective on online community success," *ACM Computing Surveys*, vol. 41, no. 2, pp. 1–29, Feb. 2009, doi: 10.1145/1459352.1459356.

[9] S. Gupta and H. Kim, "Virtual community : concepts , implications , and future research directions," *Tenth American Conference on Information Syatem*, no. August, pp. 2679–2687, 2004.

[10] R. Misra, A. Mukherjee, and R. Peterson, "Value creation in virtual communities: the case of a healthcare web site," *International Journal of Pharmaceutical and Healthcare Marketing*, vol. 2, no. 4, pp. 321–337, Nov. 2008, doi: 10.1108/17506120810922358.

[11] J. Parr and L. Ward, "Building on foundations : creating an online community," *Teacher*, vol. 14, no. 4, pp. 775–793, 2006.

[12] C. Ruggles, G. Wadley, and M. R. Gibbs, "Online community building techniques used by video game developers," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3711 LNCS, Springer Berlin Heidelberg, 2005, pp. 114–125.

[13] J. He and W. W. Chu, "A social network-based recommender system (SNRS)," in *Data Mining for Social Network Data*, Springer {US}, 2010, pp. 47–74.

[14] H. Zaim, A. Haddi, and M. Ramdani, "A novel approach to dynamic profiling of e-customers considering click stream data and online reviews," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 1, pp. 602–612, Feb. 2019, doi: 10.11591/ijece.v9i1.pp602-612.

[15] I. Guy, U. Avraham, D. Carmel, S. Ur, M. Jacovi, and I. Ronen, "Mining expertise and interests from social media," in *Proceedings of the 22nd international conference on World Wide Web - WWW '13*, 2013, pp. 515–526, doi: 10.1145/2488388.2488434.

[16] V. G. U., S. K., P. Deepa Shenoy, and V. K. R., "An overview on user profiling in online social networks," *International Journal of Applied Information Systems*, vol. 11, no. 8, pp. 25–42, Jan. 2017, doi: 10.5120/ijais2017451639.

[17] M. A. Hoshiba Pimentel, I. Barreto Sant'Anna, and M. Didonet Del Fabro, "Searching and ranking educational resources based on terms clustering," in *Proceedings of the 20th International Conference on Enterprise Information Systems*, 2018, vol. 1, pp. 507–516, doi: 10.5220/0006647305070516.

[18] B. K. Patra, R. Launonen, V. Ollikainen, and S. Nandi, "A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data," *Knowledge-Based Systems*, vol. 82, pp. 163–177, Jul. 2015, doi: 10.1016/j.knosys.2015.03.001.

[19] S. Wang, K. Kim, and K. Boerner, "Personality similarity and marital quality among couples in later life," *Personal Relationships*, vol. 25, no. 4, pp. 565–580, Dec. 2018, doi: 10.1111/pere.12260.

[20] Y. Zhang, Y. Wu, and Q. Yang, "Community discovery in twitter based on user interests," *Journal of Computational Information Systems*, vol. 8, no. 3, pp. 991–1000, 2012.

[21] B. Ferwerda and M. Tkalcic, "You are what you post: What the content of instagram pictures tells about users' personality," in *CEUR Workshop Proceedings*, 2018, vol. 2068.

[22] A. Mistry and A. P. Rajan, "Evaluation of Web Applications based on UX Parameters," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 4, pp. 2564–2570, Aug. 2019, doi: 10.11591/ijece.v9i4.pp2564-2570.

[23] I. Inuwa-Dutse, M. Liptrott, and I. Korkontzelos, "A multilevel clustering technique for community detection," *Neurocomputing*, vol. 441, pp. 64–78, Jun. 2021, doi: 10.1016/j.neucom.2021.01.059.
[24] P. Mahajan and P. D. Kaur, "Harnessing user's social influence and IoT data for personalized event recommendation in event-based social networks," *Social Network Analysis and Mining*, vol. 11, no. 1, Dec. 2021, Art. no. 14, doi: 10.1007/s13278-021-00722-6.
[25] P. Nitu, J. Coelho, and P. Madiraju, "Improvising personalized travel recommendation system with recency effects," *Big Data Mining and Analytics*, vol. 4, no. 3, pp. 139–154, Sep. 2021, doi: 10.26599/BDMA.2020.9020026.
[26] T. Chen and R. C.-W. Wong, "An efficient and effective framework for session-based social recommendation," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021, pp. 400–408, doi: 10.1145/3437963.3441792.
[27] Q. Xu, L. Qiu, R. Lin, Y. Tang, C. He, and C. Yuan, "An improved community detection algorithm via fusing topology and attribute information," in *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2021, pp. 1069–1074, doi: 10.1109/CSCWD49262.2021.9437681.
[28] J. Gao, C. Zhang, Y. Xu, M. Luo, and Z. Niu, "Hybrid microblog recommendation with heterogeneous features using deep neural network," *Expert Systems with Applications*, vol. 167, Apr. 2021, Art. no. 114191, doi: 10.1016/j.eswa.2020.114191.
[29] J. Idrais, Y. El Moudene, and A. Sabour, "Characterizing user behavior in online social networks: Analysis of the regular use of Facebook," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 4, pp. 3329–3337, Aug. 2021, doi: 10.11591/ijece.v11i4.pp3329-3337.
[30] C. Kerdvibulvech, "An innovative use of multidisciplinary applications between information technology and socially digital media for connecting people," in *Communications in Computer and Information Science*, vol. 540, Springer Berlin Heidelberg, 2015, pp. 60–69.
[31] M. Rungruanganukul and T. Siriborvornratanakul, "Deep learning based gesture classification for hand physical therapy interactive program," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12198 LNCS, Springer International Publishing, 2020, pp. 349–358.
[32] S. Bhatnagar and N. Choubey, "Making sense of tweets using sentiment analysis on closely related topics," *Social Network Analysis and Mining*, vol. 11, no. 1, Dec. 2021,Art. no. 44, doi: 10.1007/s13278-021-00752-0.
[33] T. Widiyaningtyas, I. Hidayah, and T. B. Adji, "User profile correlation-based similarity (UPCSim) algorithm in movie recommendation system," *Journal of Big Data*, vol. 8, no. 1, Dec. 2021, Art. no. 52, doi: 10.1186/s40537-021-00425-x.
[34] "What is Application Architecture?," *OrbusSoftware*. [Online]. Available: https://www.orbussoftware.com/solutions/enterprise-architecture/application-architecture (Accessed: 21-Nov-2019).
[35] V. Ferraris and F. Bosco, "Defining Profiling," *SSRN Electronic Journal*, 2013, doi: 10.2139/ssrn.2366564.
[36] A. Vedder, "KDD: The challenge to individualism," *Ethics and Information Technology*, vol. 1, no. 4, pp. 275–281, 1999, doi: 10.1023/A:1010016102284.
[37] B. H. M. Custers, "The power of knowledge ethical, legal and technological aspects of data mining and group profiling in epidemiology," Wolf Legal Publishers (WLP), Tilburg, 2004.
[38] S. Tata and J. M. Patel, "Estimating the selectivity of tf-idf based cosine similarity predicates," *ACM SIGMOD Record*, vol. 36, no. 2, pp. 7–12, Jun. 2007, doi: 10.1145/1328854.1328855.
[39] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*, 3rd ed. Elsevier, 2012.
[40] R. Tamada, "Android building multi-language supported app," *Androidhive*, 2017. [Online]. Available: https://www.androidhive.info/2014/07/android-building-multi-language-supported-app/ (Accessed: 23-Mar-2020).
[41] I. R. Management Association, Ed., *Bioinformatics*, vol. 1–3. IGI Global, 2013.