International Journal of Communication Networks and Information Security
2024, 16(4)
ISSN: 2073-607X,2076-0930
https://ijcnis.org/
Research Article

ADAPTIVE PERSONALIZATION OF SOCIAL MEDIA FEED USING POSTCATEGORIZATION

N. Gopika Rani[1] and M. Swetha[2]
[1]*Department of Computer Science and Engineering, PSG College of Technology, India*
Email:ngr.cse@psgtech.ac.in
[2]*Department of Computer Science and Engineering, PSG College of Technology, India*
Email:swetha2m3@gmail.com

ARTICLE INFO
Received: 10 Aug 2024
Accepted: 17 Sep 2024

ABSTRACT
In today's world, social media constitutes a significant part of everyone's lives. It occupies so much of our time that it even kills our productivity. Social media applications consist of enormous number of posts that can be both informative and entertaining and belong to a wide range of categories. They can be made more user-friendly by the personalization and customization of feeds of users who consume it. Suggestions shown often consist of uninteresting posts, which can make the user experience bad and result in longer scrolling durations. The solution proposed here provides a customized and personalized feed for users based on user interests. Most of the social media applications are proprietary and the algorithms for adaptive personalization are not available publicly. The main objective of this work is to develop a model for deciding how to include or discard a new post from the users' feeds based on user interests collected earlier. The system uses concepts of clustering, informationretrieval and user profiling.
Keywords: ***eigenvalues, user-profiling, personalization, fuzzy-clustering, neural-networks***

1. INTRODUCTION

Social media has seen an incredible growth since the boom of the internet. Research and development in social media are also expanding.Handheld devices are becoming more and more useful. They have all kinds of recommendations and advertisements tailored for us. Social networking and micro-blogging sites have increased in great numbers. Feeds, though entertaining, kill so much of our time. A personalized feed is a way to engage users in a positive manner. Posts recommended these days are not just determined by the people we follow or the topics we show interest to but are also curated specifically by complex algorithms running from behind the scenes.From news feeds to home pages on video players, sites like Instagram, Twitter and Facebook, all applications have their own algorithms to keep users engaged.

Mechanisms by which personalization is offered are like the strategies used by marketing and e-commerce companies. Data which follows a bi-directional pattern plays a key role here. The size of data accumulated over time from an average user is huge. Trained models which use these data and algorithms evolve over time and can make the process of adaptive personalization so
smooth that the user feels that the feed has been tailor made for him.The lack of a complete existing system gives full freedom to experiment and explore personalization algorithms for social media-basedapplications. Homepage is one of the most used features by users. This is in accordance with the Pareto principle where 80% of users use only 20% of the features.

The most suitable approach would be to group users to avoid computing feeds for large number of users.Posts based categorizationis the basis for classification of users into groups. This ensures that like-minded people get into the same group and get a similar feed. Ward and Fuzzy clustering methods have been tested for the data collected. Resultant personalized feeds also make use of text-oriented methods like NLP and Information Retrieval techniques like Term Frequency - Inverse Document Frequency (TF-IDF). This further enhances uniqueness and customization [1]. Cold start problem is a frequently encountered problem in this area as stated in [2]. This is found in the initial stage when the data is too insignificant to run the algorithm and collect user data. Personalization has become so commonplace but the techniques and algorithms for them haven't.

2. RELATED WORKS

Bayesian Models and keyword-based personalization are employed in [1] but it does not give promising results for news-based feeds. [2] gives better evaluation of content and thus good customization results because of the five-factor model for classification and trait-based personalization techniques used. Matrix factorization, CircleCon Model and ContextMF are used in [3] for customization. Social network factors were fused together to improve the accuracy and applicability of the recommender system, but it takes only user historical rating records. Collaborative filtering algorithm based on item category and interest measure is employed in [4]. Lower the MAE value, higher the accuracy of recommendation engine's predictions. Combining different attributes of users and items in the algorithm is challenging. Reinforcement learning is employed in [5] and it uses an approach for examining the features of learning materials or sequences that have not been previously explored. But many possible states/state-action pairs must be considered to avoid complexity and convergence problems.

Ontology Web Language and Resource Description Framework are employed in [6] to personalize feeds. To enrich user profiling, time dimension to capture changes of user-profiles, categorization and opinion analysis are also included. [7] provides an overview of personalized programming, detailing user profile creation methods (behavior, preferences, intent), content modeling (representation, analysis, classification) and recommendation criteria (regulation-based, content-based, participatory, mixed-filtration). It also discusses self-improvement strategies and highlights potential problems and future research indicators to enhance system performance. A profile that reads user access patterns in a dynamic environment such as the RSS feed is proposed in [8]. It uses delay minimization method based on non-homogeneous Poisson process. [9] proposes a secure solution for profile sharing on social networking sites using servers based on homomorphic encryption. It works only as long as at least one of the multiple servers is trustworthy. PrivRank, a customized and continuous social media data publishing framework that maintains user's protection against virtual attacks while allowing recommendations based on personal level, is introduced in [10].

Predictions of Facebook user personality traits based on various aspects and ratings of the Big 5 model are investigated in [11]. It analyses and compares learning models of four machines and makes connections between each set of features and personality traits. In [12], a web-based dynamic model that distinguishes user browsing behavior and predicts areas of interest is introduced. [13] uses splay trees to track changing trends for users. [14] investigates algorithms for consistent responses to personalized multimedia content. In [15], a client-side approach to personalized web search is suggested. Fuzzy variables are described to determine the relevance of each document.

The existing solutions are not available as complete systems and are not completely focused on the homepage. Posts taken from social media sites like Twitter are also focused mostly on sentiment analysis, rumor and fake news detection, etc. The proposed system aims to overcome the drawbacks of these existing systems and develop a model for adaptive personalization of the homepage of the application.

3. METHODOLOGY

The workflow of the entire system is shown in Fig.1. The algorithm summarisesthe process behind coming up with a personalised feed for a user.

Algorithm: Personalized feed
Input: user_data, post_data, user_intention
Output : personalized_feed

```
post_data ← find_post_category(post_data)
cosine_distance ← find_cosine_distance(user_intention)
clusters ← cluster(cosine_distance) /*ward or fuzzy*/
result ← []
for cluster in clusters do
users ← find_users(cluster)
    user_matrix←build_matrix_for_cluster(users,user_intention)
    trans_user_matrix ← get_transpose(user_matrix)
    square_matrix ← trans_user_matrix . user_matrix
    eigen_result ← eigenise(square_matrix)
    result.append(eigen_result)
end
ward_eigen_proportion ← find_proportion(result)

personalized_feed ← []
for user in user_data do
recommended_posts←get_posts_using_TF-IDF(user,ward_eigen_proportion[user])
    personalized_feed.append(recommended_posts)
end
```

display(personalized_feed)

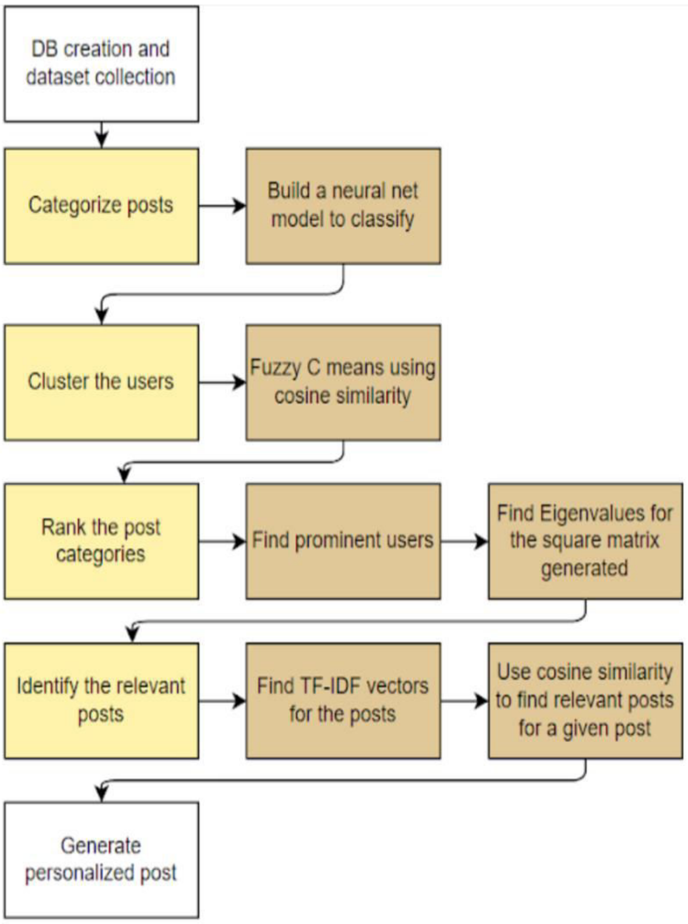
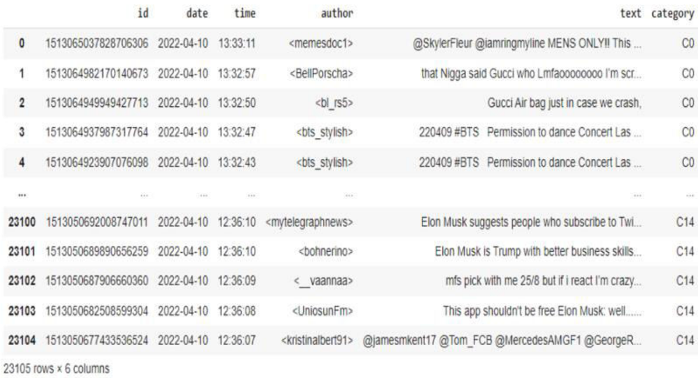


Fig.1. Workflow of the Proposed System.

3.1. DATA COLLECTION

Twint tool has been used to scrape Twitter data. The size of the tweets dataset collected is 23,105. Tweets are collected by specifying categories. Twitter standard categories that are used for user profiling (categories that users are asked to pick when joining Twitter) are used here. The collected dataset has details such as tweet id, tweet content, author of tweet, timestamp and date as shown in Fig.2.The dataset was refined to reflectonly positive likes and no dislikes to avoid issues in eigen decomposition.

Fig.2. Dataset Collected.



	id	date	time	author	text	category
0	1513065037828706306	2022-04-10	13:33:11	<memesdoc1>	@SkylerFleur @iamringmyline MENS ONLY!! This ...	C0
1	1513064982170140673	2022-04-10	13:32:57	<BellPorscha>	that Nigga said Gucci who Lmfaoooooooo I'm scr...	C0
2	1513064949949427713	2022-04-10	13:32:50	<bl_rs5>	Gucci Air bag just in case we crash,	C0
3	1513064937987317764	2022-04-10	13:32:47	<bts_stylish>	220409 #BTS Permission to dance Concert Las ...	C0
4	1513064923907076098	2022-04-10	13:32:43	<bts_stylish>	220409 #BTS Permission to dance Concert Las ...	C0
...	...	...	...	...	...	...
23100	1513050692008747011	2022-04-10	12:36:10	<mytelegraphnews>	Elon Musk suggests people who subscribe to Twi...	C14
23101	1513050689890656259	2022-04-10	12:36:10	<bohnerino>	Elon Musk is Trump with better business skills...	C14
23102	1513050687906660360	2022-04-10	12:36:09	<__vaannaa>	mfs pick with me 25/8 but if i react I'm crazy...	C14
23103	1513050682508599304	2022-04-10	12:36:08	<UniosunFm>	This app shouldn't be free Elon Musk: well......	C14
23104	1513050677433536524	2022-04-10	12:36:07	<kristinalbert91>	@jamesmkent17 @Tom_FCB @MercedesAMGF1 @GeorgeR...	C14

23105 rows × 6 columns

Using all tweets for computation requires high computation cost and time. So, tweets are categorized into 15 standard categories namely Entertainment, Home and Family, Fashion and Beauty, Food, Technology, Outdoors, Music, Art & Culture, Career and Education, Animation, Sports, Business & Finance, Travel, Gaming and Fitness.Totally four collections are created in the database, namely users, posts, user_posts and final_results.

- 'users' has all details about users such as name, mail id, age category and location category.

- 'posts' collection has information about posts such as post id, date, author, tweet and category. Other than category, all other details of posts are collected from Twitter using the Twint tool. The category of each new incoming post will be found by using a neural network model such as Bidirectional LSTM.
- 'user_posts' collection has two values: one is the username and another one is a list of post id. If a user likes the post, then the post id will be added to the list. This data is used to get all the liked tweets of all users.
- 'final_results' collection has the posts for the personalized feeds of users obtained after performing the personalizing feed algorithm.

3.2. CATEGORISING POSTS

To categorize the posts, bidirectional LSTM neural network model is used. We need to nextpre-process the texts. This involves removing stop words from the tweets, shuffling the dataset, removing non-alphabetic characters (such as emojis and numbers) and tokenizing texts. Then the model is trained and tested using the accumulated dataset. Number of epochs used here is 9 and it results in an accuracy of approximately 66%.

3.3. CLUSTERING OF USERS

In a social media-based application, there are a large number of users. Finding personalized information for each user is computationally complex. Thus, a large number of users are clustered into smaller groups. Each cluster contains similar kinds of people who are interested in the same categories of posts. This also helps in suggesting new posts to users which they might like based on the posts liked by other users in the same cluster.Cosine distance is used to calculate similarity between users based on their interests. Here, two clustering methods (Fuzzy C-means and Ward Agglomerative Hierarchical Clustering) are implemented and the results are compared. Using normal clustering methods might lead to providing the same personalized feeds for all users of a given user group. A user may have to belong to more than one cluster. To implement this scenario, the Fuzzy C-means clustering method is used along with cosine similarity. Every user belongs to clusters to a certain degree.

3.4. RANKING OF POSTS

Eigenvalues and eigenvectors are calculated andused to find the relative importance of posts for each cluster. Finding the importance of categories helps in giving more posts from top categories. Using a dynamic threshold, prominent users of each cluster are found and a matrix M of order mxn, where m is the number of prominent users and n is the number of post categories, is obtained. Eigenvalues are obtained for post-categories for each cluster. Largest eigenvalueand thus the corresponding eigenvector is assigned importance and for each category.

3.5. IDENTIFYING RELEVANT POSTS

To find such posts, TF-IDF values are calculated for words in the post data.TF-IDF vectorisation gives a vector for each post in the dataset. Similarities of the TF-IDF vectors of two posts are considered. High similarity between the TF-IDF vectors of new tweets with the user's previous ones indicate that the posts have a high chance of being important and useful to the user in the future. This technique thus helps us differentiate as well as identifyrelationships among various posts. final_results collection is appended with such relevant posts and presented to user.

4. RESULT ANALYSIS

Customers perceive satisfaction while receiving personalized recommendations online

Users perceive satisfactionwhilereceivingpersonalized recommendations online [16]. To improve recommendations, accuracy of system has to be improved. Using mean average score metric, a score of 3.32 is obtained. [17] Precision and recall are the most popular metrics for evaluating information retrieval systems and are defined for our systems using Eq. (1) and Eq. (2).Personalization score is the average of upper triangular matrix of cosine dissimilarity matrix as shown in Eq. (3).

$$precision = \frac{No.of\ relevant\ recommended\ items}{No.of\ recommended\ items} \quad (1)$$

$$recall = \frac{No.of\ relevant\ recommended\ items}{No.of\ all\ possible\ relevant\ items} \quad (2)$$

$$Personalisation\ score = avg(a_{ij}) \quad (3)$$

where $a_{ij} \in A$ and A is the cosine dissimilarity matrix.

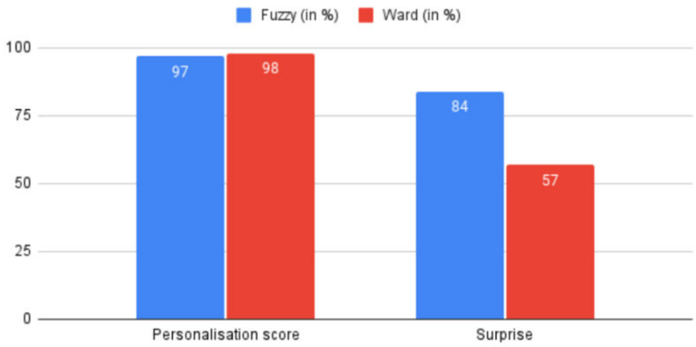
Personalization score obtained after applying the personalizing feed algorithm is 97.965% and if the feed is generated without the algorithm and any sort of personalization by the descending order of time, score obtained is 7%. Personalization scores comparison is tabulated in Table.1.

Table.1. Personalization Scores Comparison.

Cases	Personalisation score (in %)
Fuzzy clustering without personalisation	7
Ward clustering without personalisation	11
Fuzzy clustering with personalisation	97
Ward clustering with personalisation	98

Ward clustering allots each user to a single cluster unlike fuzzy where one user can belong to more than one cluster. However, applying TF-IDF technique ensures that personalisation is good and unique even for Ward clustering. Comparison between ward and fuzzy clustering in terms of both personalization score and surprise score are highlighted inFig.3.

Fig.3. Comparison Between Fuzzy C-means and Agglomerative Clustering.



A good personalized feed should have an equal amount of relevance and surprise. Relevance indicates that the posts categories based on interests are given more significance.Surprise element indicates that new but interesting posts find a place in the feed too. On finding the percentage absolute difference between interests and output, we see that Ward clustering gives 57% and FuzzyC-means clustering gives 84%. Ward clustering is better in this context as it iscloser to being ideal.

MAE (Mean Absolute Error) analysis[17] for the two clustering methods shows that Fuzzy clustering has a score of 3.32 whereas Ward clustering has a score of 4.3. Thus, Ward clustering is better than Fuzzy clustering in terms of surprise-novelty ratio.

5. CONCLUSION

Fuzzy C-means clustering technique and ward agglomerative clustering are compared for social media posts based categorisation. If we need users to belong to different user groups, fuzzy C means clustering algorithm can be employed. But it doesn't matter because of the overall performance of Ward agglomerative clustering algorithm in terms of providing a personalised feed (with a personalisation score of 98%). Other clustering techniquesthat suit the model can be employed too. TF- IDF technique can be replaced by computationally efficient Information Retrieval techniques. The accuracy of the neural network implemented for post categorisation can be improved by tuning the hyperparameters and by adjusting the number of epochs. In this project, only a static dataset is used for generating the result set. Continuous incoming stream of data, which is typical of social media, can also be modelled. To deal with larger number of users, the algorithm can be parallelized with Hadoop or other distributed systems for improving the efficiency. Future scope is vast forsuch social media based applications.

ACKNOWLEDGEMENT

The authors wish to express their appreciation to the reviewers for their helpful suggestions which greatly improved the presentation of this paper. The authors received no specific funding for this study. The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1] Tuck Siong Chung, Michel Wedel and Roland Rust.T, "Adaptive personalization using social networks",Journal of the Academy of Marketing Science, Springer, vol. 44(1), pages 66-87, January 2016.
[2] Stephan Winter, Ewa Maslowska and Anne Vos.L, "The effects of trait-based personalization in social media advertising", Computers in Human Behavior, vol. 114, doi:10.1016/j.chb.2020.106525, August 2020.
[3] Xueming Qian, He Feng, Guoshuai Zhao and Tao Mei, "Personalized Recommendation Combining User Interest and Social Circle",Knowledge and Data Engineering, IEEE Transactions, vol. 26, 2014.

[4] Suyun Wei, Ning Ye, Shuo Zhang,Xia Huang and Jian Zhu, “Item-Based Collaborative Filtering Recommendation Algorithm Combining Item Category with Interestingness Measure”, doi:10.1109/CSSS.2012.507, August 2012.

[5] Doaa Shawky and Ashraf Badawi, “Towards a Personalized Learning Experience Using Reinforcement Learning”, Studies in Computational Intelligence, doi:10.1007/978-3-030-02357-7_8, 2019.

[6] Paula Peña, Rafael del Hoyo, Jorge Vea-Murguía, Carlos González and Sergio Mayo, “Collective Knowledge Ontology User Profiling for Twitter - Automatic User Profiling”, IEEE/WIC/ACM International Conferences on Web Intelligence, 2013.

[7] Min Gao, Kecheng Liu and Zhongfu Wu, “Personalisation in web computing and informatics: Theories, techniques, applications, and future research”, Information Systems Frontiers, doi:10.1007/s10796-009-9199-3, 2010.

[8] Muhammad Ali Zeb and Maria Fasli, “Dynamically Adaptive User Profiling for Personalized Recommendations”, IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 2012.

[9] Xun Yi,Elisa Bertino,Fang-Yu Rao,Kwok-Yan Lam,Surya Nepal and Athman Bou, “Privacy-Preserving User Profile Matching in Social Networks”,IEEE Transactions on Knowledge and Data Engineering, vol. 32, 2020.

[10] Ding qi Yang, Bingqing Qu and Philippe Cudré-Mauroux, “Privacy-Preserving Social Media Data Publishing for Personalized Ranking-Based Recommendation”, IEEE Transactions on Knowledge and Data Engineering, vol.312019.

[11] Michael M. Tadesse, Hongfei Lin, Bo Xu and Liang Yang, “Personality Predictions Based on User Behavior on the Facebook Social Media Platform”, IEEE Access, vol. 6, 2018.

[12] Payal Das and G P Sajeev, “Time- Driven Adaptive Web Personalization System for Dynamic Users”, IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2017.

[13] Payal Das, R. C. Jisha and G. P. Sajeev, “Adaptive web personalization system using splay tree”, International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017.

[14] N.D. Doula mis and P. Georgi Lakis, “Adaptive multimedia content personalization”, IEEE International Symposium on Circuits and Systems, 2004

[15] Mohammad Norouzzadeh .S , Ayoub Bagheri and Mohammad Saraee .H,“Web search personalization: A fuzzy adaptive approach”, 2nd IEEE International Conference on Computer Science and Information Technology, 2004.

[16] JiangZ and BenbasatI,“Virtual product experience: Effects of visual and functional control of products on perceived diagnosticity in electronic shopping”, Journal of Management Information Systems, 21(3), pp: 111–147, doi:10.1080/07421222.2004.11045817, 2004.

[17] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl, “Evaluating collaborative filtering recommender systems”,ACM Transactions on Information Systems, vol. 22, January 2004.