engineering proceedings

Proceeding Paper

Improved Social Network User Recommendation System—The Machine Learning Approach †

Yana A. Bekeneva *,‡ and **Titus U. Eze** ‡

Department of Computer Science and Engineering, Saint Petersburg Electrotechnical University "LETI", ul. Professora Popova 5, 197022 St. Petersburg, Russia; titusugochukwu28@gmail.com

* Correspondence: yabekeneva@etu.ru

† Presented at the 15th International Conference "Intelligent Systems" (INTELS'22), Moscow, Russia, 14–16 December 2022.

‡ These authors contributed equally to this work.

Abstract: In this paper, we propose a recommendation system for social media users which makes recommendations on the basis of the user profile information and the contents posted by users-the bio-aware algorithm. Text mining techniques are used to pre-process the words before feeding it to an LDA model which handles the topic and feature extraction. In this way, we determine the similarity between users based on their interests. We further trained a machine learning model which is able to identify and score the top interests of a particular social media user. Other users who share similar scores are shown as recommendations to each other.

Keywords: social network; people interaction; information searching; information extraction; bio-aware approach

check for updates

Citation: Bekeneva, Y.A.; Eze, T.U. Improved Social Network User Recommendation System—The Machine Learning Approach. *Eng. Proc.* **2023**, *33*, 13. https://doi.org/10.3390/engproc2023033013

Academic Editors: Askhat Diveev, Ivan Zelinka, Arutun Avetisyan and Alexander Ilin

Published: 9 June 2023

1. Introduction

In 2020, an estimated 46% of the world population is active in at least one social network [1]. When they register on these networks, they are asked to follow some people to see their content. Users then interact with each other's content, send messages to each other and so on. To further improve this interaction, there is a need to develop more effective recommendation systems as users are more likely to spend more time online and engage with the content of people they identify better with. Users are often recommended to their fellow classmates, people who live in the same city or friends of their friends, the list goes on. Familiarity relationships are simulated through repeated exposure to profiles [2].

There have been previous attempts which include both content-based and graph-based approaches. The first focuses on measuring the topic similarity among social media users, while the second infers the relationship between users based on a graph. While it takes a lot of data to develop algorithms that would produce these recommendations, this paper explores a peer-to-peer recommendation system with a bio-aware algorithm which measures the similarity interests among users based on their profile biography and their last post content to offer suggestions.

Recommender systems are generally tools that help in filtering and sorting information as it is relevant and interesting to the user. This is as a result of the huge data problem that came with the adoption of the world wide web.

Netflix's recommendation system also plays an important role in the success of the platform. Many times, people come to the site without knowing exactly what to watch but completely believe that the site will recommend what fits in to their taste. When creating an account for the first time, Netflix asks a user to choose about five different movies they like. From this list, Netflix is able to populate their feed with similar movies that match their taste using its recommendation system.

Social media is no different. The basics of social interaction are such that there is a variety of topics to talk about, and even better when the discussions are with people who share similar views with you. When someone registers on a social network such as Twitter, they are asked to follow some people. These are usually celebrities, government officials or new agencies in their country. From this, Twitter is able to decide what kind of users and topics the user is interested in and then populate their feed accordingly. Previously, you could accurately understand what a post on Twitter was about from the hashtags it has. Recently, Twitter introduced topics. People can follow certain topics ranging from sports, celebrity news, particular fields such as cryptography, space and so on. With this, they can accurately suggest relevant posts on the user's timeline.

Traditional approaches to building recommender systems are listed as follows: content filtering, collaborative filtering and hybrid filtering. Other modern approaches include location-content-aware, context-aware, semantic based, cross-lingual and peer-to-peer.

Even though a lot of research has gone into improving the state of recommender systems using machine learning approaches, there are still many challenges that affect these systems to date. Some of them are described below.

Cold start problems occur when there is not enough data. It could be a cold start of items in which case we do not have enough information about the items or users. For example, it would be difficult to recommend movies to a user if they have not watched any movies at all or given any information that will help in profiling them.

Large datasets are often required to develop a commercial recommender system. This leads to the usage of a large and spare user-item matrix for filtering which in turn affects the performance of the recommendation process. The cold-start problem is caused by data sparsity.

As the number of customers and items increases, it poses a scalability problem to traditional collaborative filtering algorithms since the complexity will be too large. Most resources are used for the purpose of determining the similarity in interests between users, and items with similar attributes [3].

This is one of the most commonly encountered challenges, especially with content-based filtering, since it always would recommend items that users are already familiar with. One of the properties of a good system is that it should be diverse.

2. Related Works

Obtaining user profile information can be useful when grouping users and building recommendation systems. Since unlabelled data are considerably more frequent than labeled data in many social network datasets, inferring hidden users attributes and using graph-based semi-supervised learning algorithms are more suitable for this case [4,5].

In order to obtain a good recommendation from a system, there is a need for clear distinctive information on a user's preferences [6]. This is generally achieved by recording feedback from the user; the posts they like, or using sentiment analysis to score their comments on posts [7].

Gurini et al. [8] emphasizes the use of implicit sentiment analysis which further improves the performance of recommendation systems. In this approach, they were able to build more complete user profiles than traditional content-based approaches by defining a novel weighting function that takes into account sentiment and size in relation to the user's interests.

N. R. Vajjhala et al. [9] proposed a solution to identify user interests using IBM Watson platform which makes use of Natural language processing for advanced text analysis. Their solution however does not cover for cases with multiple preferences.

Arru et al. [10] proposed a system that is based on a novel user model, termed bag-of-signal. Guy et al. [11,12] recommend people based on articulated social network information by combining different sources to derive factors that might influence the similarity measure.

Recent studies, e.g., [13], suggest a user's interests are always changing and there is a need to capture that. Their solution combines the user interactions, social features and post history to capture their interests.

Hannon et al.'s proposal [14] recommends users through a hybrid system of both content based and collaborative filtering. In the content-based method, users correlation scores are a function of their followees' posts, their followers' posts or a combination of both. In this case, users with similar posts will be recommended to each other. For the collaborative filtering approach, users are represented by the IDs of their followees, followers or a combination of them. Then tf-idf weighting scheme is used to find users similar followers/followees and recommend them to the target user.

A recurring problem for many of these solutions is that they largely suffer from the cold-start problem. We need information about the user's followers, followees and their posts in order to develop the system. A user needs to have followed a substantial number of users in order to effectively generate relevant suggestions.

3. Approach Description

The bio-aware algorithm proposed in this paper seeks to eliminate this and also solve the data sparsity and scalability problems. It also reduces the computing time and resources involved in working with datasets comprising many features.

This is based on the assumption that most social media users include details about themselves, who they are and what they like or talk about in their biography. When combined with their most recent posts we are able to generate a list of topics that the user is interested in.

To use social media platforms such as Twitter, a registered user needs to describe who they are, their name, short biography, location, etc. A user can follow other users and see information about them, posts they have made, their followers and the people they follow termed 'followees'.

One determining factor as to why a user follows another user can sometimes depend on how their interests align. Interests could be, for example, if they both like football, or they are fans of a celebrity. Studies have shown that people with the most followers are usually new agencies, celebrities or famous organizations. It was also discovered that the lesser the number of followers a user has, the less likely they are to make a tweet while the more followers they have, the less likely they are willing to reciprocate the following.

Our purpose is to present a user with a set of recommendations which they will most likely identify with. If a user already deliberately followed other users, we could assume they already find these people interesting either based on the recent tweets they made, their description of what they do and from how popular they are (Figure 1).

Since we are dealing with a lot of raw user generated data, which are mostly unstructured in nature, we need to pre-process them firstly using about different techniques namely, tokenization, stemming, removing stop words and noise removal.

There are different feature extraction techniques which include bag of words, TF-IDF, word embedding, and Natural Language Processing "NLP". Studies have shown that for machine learning algorithms using text classification.

To find how similar user's interests are to each other, we compare the similarity based on their LDA. We would define the similarity measure as the dot product between the corresponding vectors representing the contributions of each word to a topic.

Merge fields: since we need to feed the LDA model a string text, there is a need to combine the relevant fields which we will be using for our topic extraction.

Preprocessing: in (Figure 1), we have listed the preprocessing techniques and steps that will be used in this work namely; tokenization, lemmatization, removal of stop words, removal of numbers, emails, URLs, punctuations and white spaces.

Topic extraction: after preprocessing, we perform topic extraction using the LDA model to identify the top ten words in each topic. The topics here are not labeled so we use wordcloud to visualize the data.

Topic weight score: here, we obtain the weight score for a user's top topics and assign it as a new corresponding field named 'topic'.

Get recommendations: we retrieve a user's list of recommendations by checking other users with similar weight scores.

To obtain a model that is able to predict a user's topic weight scores, we explore the following steps: split the dataset into train and test data; using random forest classifier to train the topic column; predict the topic weight score for the test data; obtain recommendations.

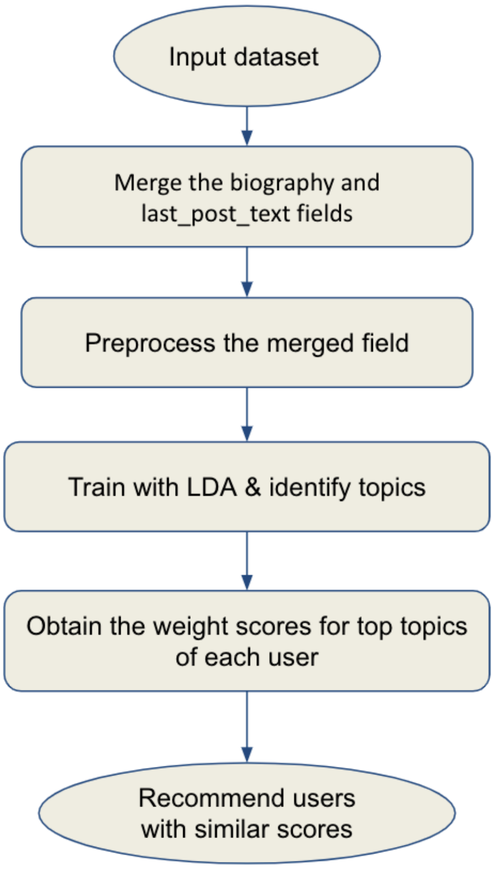
Input dataset
Merge the biography and last_post_text fields
Preprocess the merged field
Train with LDA & identify topics
Obtain the weight scores for top topics of each user
Recommend users with similar scores

Figure 1. Proposed algorithm design for building our classification model.

4. Experiment Results

For the sake of experiment, Twitter was chosen as the social media platform as a case study since it was easier to obtain useful data than other platforms.

We used popular processing tools like numpy for array manipulations, pandas for dataframes, sklearn for feature extraction, re for formatting the string data, nltk library for preprocessing, keras for building models, matplotlib and wordcloud for data visualization.

Using the wordcloud library, we are able to visualize the results of our LDA training model. We can infer that from the most popular words, this topic has to do with world sports since it captures words such as league, game, player, baseball and so on.

We can deduce from Figure 2 that this topic relates more to startups and businesses. People who talk about or are interested in investment, finance, funding, web development and the rest. Figure 3 can be summarized as relating to people who are interested in skill acquisition or building networks.

We built a learning model that can predict the topic weight score of new datasets so as to reduce the build time and save resources. To do this, we split our dataset into training and testing subsets. Our testing dataset was 20% of the entire dataset.

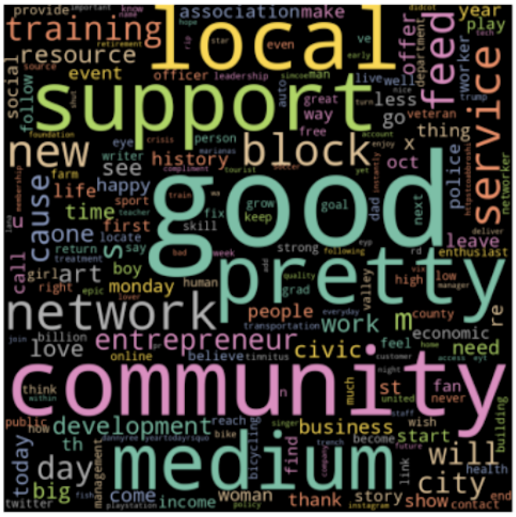

Figure 2. Topic visualization map for posts related to business

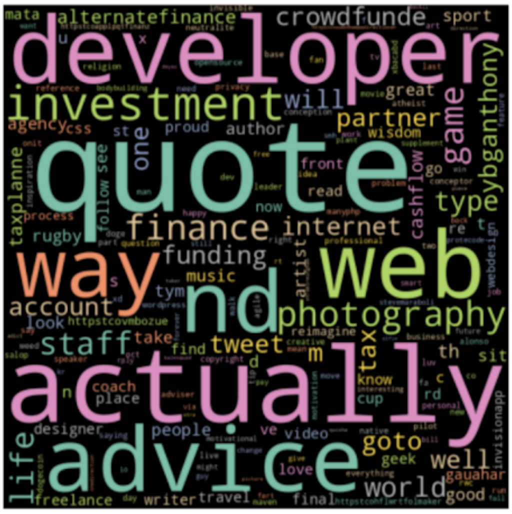

Figure 3. Topic visualization map for posts related to network technologies

5. Comparison of with Other Approaches

When we consider both traditional approaches, e.g., content-based filtering, collaborative filtering and hybrid filtering with our approach, we see a clear improvement (Table 1).

Table 1. Results of comparison of different approaches.

Approaches	Cold Start	Privacy	Multiple Languages	Sparsity
Content-filtering	-	-	-	+-
Collaborative filtering	+-	+	-	-
Hybrid filtering	+	-	-	+
Cross-lingual	-	+	+	-
Bio-aware	+	+	-	+

The bio-aware algorithm proposed in this paper seeks to eliminate most of the challenges listed above such as data sparsity, scalability, cold-start and privacy problems. It also reduces the computing time and resources involved in working with datasets comprising many features since we also develop a model that is able to automatically output the topic weight score for an already preprocessed dataset.

The recommendation system developed with this algorithm is independent of a particular platform and can be used on any social media platform of choice. Twitter was chosen as the use case here because of the availability of relevant dataset.

6. Conclusions

It is difficult to verify the effectiveness of our algorithm without comparing it to the user's actual followee's list, which is not provided in the dataset used for this project. It's also important to acknowledge the limits of the approach used in this work.

From the experiment results, it is obvious that the preprocessing techniques and steps used have a great influence on how the classification performs. Removing mentions from the text could possibly increase the effectiveness. For the purpose of this work, we only chose to process English words which is not the case in the real world. Posts and user bio can be in a user's native language. As much as we tried to remove some English stop words, it did not capture the everyday English words which appeared so many times but do not add to the relevance of our analysis.

Even though this approach tries to eliminate the cold-start problem, there are still few cases where a user does not add anything in their bio or have not made any posts at all. For this case, we suggest that only the popular users be shown as recommendations here. There might also be cases where a user does not capture what they are interested in on their bio or on their last post. To solve this, we propose that only the last five posts with more than 80 characters be considered in the case that there are more.

It is also important to note that our approach does not capture topic classification on the subcategory basis. If someone mentions that they like baseball, we should only offer them suggestions of people who also like baseball and not people who like sports generally. To do this, we might have to increase the number of topics that the LDA model should classify into. A proposed solution to this might be to have a known list of possible topics that is already labeled against each user in the dataset. These days, platforms such as Twitter and Instagram already allow you to choose topics or hashtags based on your preferences and content is recommended around these topics.

Author Contributions: Conceptualization, Y.A.B.; Methodology, Y.A.B., T.U.E.; Software, T.U.E.; Supervision, Y.A.B.; Validation, Y.A.B., T.U.E.; Visualization, T.U.E.; Writing—original draft, T.U.E., Y.A.B.; Writing—review and editing, Y.A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Statista Research Department. Number of Social Network Users Worldwide from 2017 to 2025. 2021. Available online: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (accessed on 1 July 2022).
2. Guy, I.; Inbal, R. Do You Know? Recommending People to Invite into Your Social Network. In Proceedings of the 14th International Conference on Intelligent User Interfaces, IUI '09, Sanibel Island, FL, USA, 8–11 February 2009; pp. 77–86.
3. Jain, S.; Grover, A.; Thakur, P.S.; Choudhary, S.K. Trends, Problems And Solutions of Recommender System. In Proceedings of the International Conference on Computing, Communication and Automation (ICCCA2015), Greater Noida, India, 15–16 May 2015; pp. 955–958.
4. Ding, Y.; Yan, S.; Zhang, Y.; Dai, W.; Dong, L. Predicting the attributes of social network users using a graph-based machine learning method. *Comput. Commun.* **2015**, *73*, 9. [CrossRef]
5. Zheleva, E. To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles. In Proceedings of the 19th Century International Conference of the World Wide Web, Madrid, Spain, 20–24 April 2009; pp. 531–540.
6. Yue, L.; Chen, W.; Li, X.; Zuo, W.; Yin, M. A survey of sentiment analysis in social media. *Knowl. Inf. Syst.* **2019**, *60*, 617–663. [CrossRef]

7. Javed, U.; Shaukat, K.; Hameed, I.A.; Iqbal, F.; Alam, T.M.; Luo, S. A review of content-based and context-based recommendation systems. *Int. J. Emerg. Technol. Learn. (IJET)* **2021**, *16*, 274–306. [CrossRef]
8. Gurini, D.F.; Gasparetti, F.; Micarelli, A.; Sansonetti, G. A Sentiment-Based Approach to Twitter User Recommendation. *RSWeb@ RecSys* **2013**, *1066*, 1–4.
9. Vajjhala, N.R.; Rakshit, S.; Oshogbunu, M.; Salisu, S. Novel user preference recommender system based on Twitter profile analysis. *Soft Comput. Tech. Appl.* **2021**, *1248*, 85–93.
10. Arru, G.; Feltoni Gurini, D.; Gasparetti, F.; Micarelli, A.; Sansonetti, G. Signal-based user recommendation on twitter. In Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 13–17 May 2013; pp. 941–944.
11. Palomares, I.; Porcel, C.; Pizzato, L.; Guy, I.; Herrera-Viedma, E. Reciprocal recommender systems: Analysis of state-of-art literature, challenges and opportunities towards social recommendation. *Inf. Fusion* **2021**, *69*, 103–127. [CrossRef]
12. Gurini, D.F.; Gasparetti, F.; Micarelli, A.; Sansonetti, G. Temporal people-to-people recommendation on social networks with sentiment-based matrix factorization. *Future Gener. Comput. Syst.* **2018**, *78*, 430–439. [CrossRef]
13. Tsou, M.H.; Zhang, H.; Park, J.; Nara, A.; Jung, C.T. Spatial Distribution Patterns of Geo-tagged Twitter Data Created by Social Media Bots and Recommended Data Wrangling Procedures. In *Empowering Human Dynamics Research with Social Media and Geospatial Data Analytics*; Springer: Cham, Switzerland, 2021; pp. 257–273.
14. Sardianos, C.; Ballas Papadatos, G.; Varlamis, I. Optimizing parallel collaborative filtering approaches for improving recommendation systems performance. *Information* **2019**, *10*, 155. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.