

LinkedIn Connection Recommendations.

Authors:-

Parth Jaisur

Department of Computer Engg.,

GEC Gandhinagar, GTU (India)

Khushali Pariyal,

Department of Computer Engg.,

GEC Gandhinagar, GTU (India)

Madhuri Mahajan

Department of Computer Engg.,

GEC Gandhinagar, GTU (India)

Samay Dumsia

Department of Computer Engg.

GEC Gandhinagar, GTU (India)

Jasmin Jani

Department of Computer Engg. ,

GEC Gandhinagar, GTU (India)

Abstract:-

LinkedIn is a free, social media platform that is used to help hungry job seekers to find their suitable job according to their needs and post CV so that they can improve their impression in front of employers and employers can post available positions in their company so that they can get suitable candidates as per available positions. It is used to expand the connection at worldwide level between professionals, students, freshmen, experienced people etc. using their skills and requirements. As per the study 774+ (numbers are in millions) members are connected worldwide [\[1\]](#). As per [\[2\]](#) in 2020, In US highest number of users reached 170 million making it the country with the most users in the world. Over a million or billions of people, how can one find the recommended user, members, or company?? Here comes the use of a recommendation system. Recommendation systems are used to suggest the connection based on certain criteria. They filter out some important attributes, and based on that, connections are recommended. Many studies are done on the LinkedIn recommendation system. Cosine similarity, collaborative filtering, content-based recommendation and many more algorithms are used for recommendation purposes. And many of them got accurate results to some extent. So, to understand existing recommendation systems and to create our own model with better accuracy,

we did this study. We collected LinkedIn data from Kaggle which was used for training, and for testing purposes, we collected data from LinkedIn. We had used

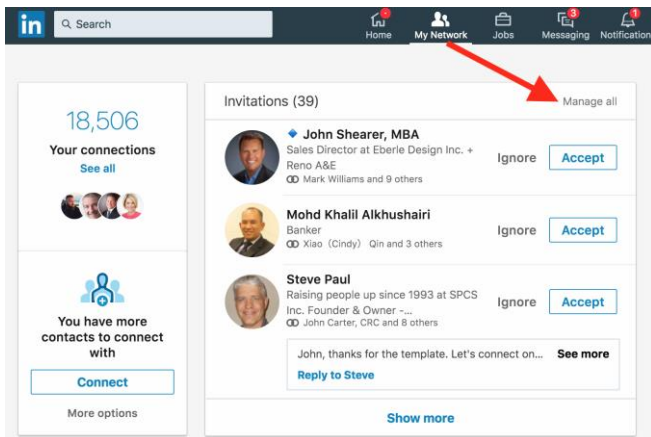
the TF-IDF model to preprocess the data and remove any inconsistent data. And LSA (Latent Semantic Analysis) will recommend a user profile's list which will have maximum compatibility with users. Tf-Idf can be used to calculate the frequency of words that occur multiple times and create TF-IDF matrix, which is given as input to the LSA matrix that checks similarity between user profiles.

Keywords: - Recommendation System, TF-IDF , LSA , preprocess , TF-IDF matrix , maximum compatibility , Word Frequency, LSA Matrix, Similarity Measure.

1.Introduction: -

LinkedIn is free and the best platform for searching **for individuals to find suitable jobs and for corporates and to find suitable candidates for the open jobs**. LinkedIn can be used to get in touch with some professionals who are from other parts of the world or to find a job at any level or any place in this world. A LinkedIn recommendation system offers connections based on the user's skills and requirements or based on their compatibility.

A recommendation system, or a recommender system, is a subclass of Information filtering system that provides suggestions of items that are most pertinent to a particular user. Typically, the suggestions refer to various decision-making processes, such as which product to purchase, what type of music to listen to, or which online news to read. Recommender systems are particularly useful when an individual needs to choose an item from a potentially overwhelming number of items that can serve a purpose^{[\[3\]](#)}.



2.Related Works:-

There are a considerable number of research papers and models available for the LinkedIn recommendation base system. And also, there are a lot of methods available; using them, we can create any recommendation system. But we have to select the correct and consistent method so that we can create a model that gives better accuracy than others and helps users find their required user or company on LinkedIn.

In this section, we are going to discuss various research papers on recommendation systems that have already been published and help us learn from them.

In 2014^[1], a professor at Urmia University of Technology proposed a recommender system called GeoLocation Friend Recommender System (GeLoFRS) for LinkedIn. The system recommends users with similar interests and expertise located in the vicinity of the target user. The system uses real data from 532 LinkedIn users to calculate the best candidates for recommendation based on skills, industry, connections, education, and geolocation information. The results show the formation of small groups with similar skills and interests that lead to larger professional communities. The paper recommends including details about future work leading to more parameters to increase recommendation accuracy and using a semantic similarity framework for better accuracy.

In 2014^[2], Vasavi Akhila Dabeeru published a guide in which the article discussed the problems of finding the degree of closeness and interaction level in a social network by ranking users based on a similarity score. They proposed a technique that will use various attributes of user profiles, such as social, geographic,

In daily usage of the LinkedIn social networking platform, it is observed that numerous requests are received from various users. To validate the relevance of the incoming requests, manual inspection of the users' profiles is required to identify if any commonalities exist between the user and the recipient, based on parameters such as skills, location, or experience. However, this task can become cumbersome if the number of requests received is significant, for instance, 100 or 200 requests, necessitating the need for an automated filtering mechanism to ascertain the similarity between the users.

In this study, we extracted data of LinkedIn skills and location from Kaggle which will be used for training purposes. And for testing, we collected some user data from the LinkedIn platform. In the next sections, we are going to discuss the literature survey done on some related work. In section (3), we will describe the dataset and methodologies used to create the model. Section (4) describes the result of the model and the executed methodology with their resulting accuracy. And in the end, a summary of work and the future work will be presented.

educational, professional, shared interests, pages liked, mutually interested groups or communities, and mutual friends. The similarity between user profiles reflects the closeness and interaction between users. The proposed technique is able to discover the largest possible number of profiles that are similar to the target user profile as compared to existing techniques. The article provides a theoretical model that describes how similarity scores are calculated and shows the interaction or closeness level between users. The technique assigns weights to attributes manually and uses syntactic and semantic similarity metrics to compare attribute values.

In 2016^[3], a professor at the University of Bologna discussed the use of latent semantic analysis (LSA) to identify the relationship between job positions and people's skills. By mining data from LinkedIn users' public profiles, the authors demonstrate the effectiveness of their method in recommending job positions. The extracted semantics could be valuable for both job recommendation systems and recruiting systems. The authors argue that current approaches to job recommendation systems rely on scarce, manually collected data that does not fully reflect people's skills. The paper proposes a job recommendation system based on LSA and a hierarchical clustering of job positions. The system is evaluated using LinkedIn as a reference scenario.

In 2018^[4], the University of Saskatchewan published a paper on the recommendation system for Twitter and LinkedIn, in which they recommended Twitter accounts based on LinkedIn skills

and vice versa. They developed collaborative filtering, content filtering, hybrid filtering, and feature-based selection, which had an accuracy of 21.33% for Twitter and 33.60% for LinkedIn, which were far better than the other three. All of these algorithms were accurate for predicting LinkedIn skills rather than recommending Twitter accounts.

In 2021^[5], Qiannan Yin published an article in which they illustrated the use and working of the People You Might Know (PYMK) feature. They used an A/B testing algorithm in which members are randomly assigned to different treatment groups and can see recommendations from different models. If the model has better recommendations, it will send more requests. This has an impact on the sender side, which can be read out directly from A/B testing. But on the receiving side, when members receive any invitation, they have to come to LinkedIn to accept the invitation, which is harder to track because some members can receive invitations from different senders. To overcome this challenge, LinkedIn created an attribution framework to attribute the sessions of recipients to the correct senders.

In 2022^[6], created a guide to social media recommendations. They stated the steps to perform a recommendation system. And the steps are importing libraries, exploratory data analysis, creating connections, getting training and test data, feature engineering, model training, and performance measurement. To train the model, they used the XGBoost Classifier,

Random Forest Classifier, K Nearest Neighbor, and Support Vector Machine.

3.Method Description: -

Term Frequency-Inverse Document Frequency is used to retrieve the information about how important a word is to that particular document. It is often used as a weighting factor, that searches the information retrieval, text mining and user modeling. Tf-idf counts the number of frequencies of word/string in a particular document. Frequency count increases as the same word appears in document.

The mathematical equation used to compute the tf-idf score for a term t in a document d is:

$$\mathbf{tf-idf}(t, d) = \mathbf{tf}(t, d) * \mathbf{idf}(t)$$

eq.(1)

where $\mathbf{tf}(t, d)$ is the frequency of term t in document d (formula(2)), and $\mathbf{idf}(t)$ is the inverse document frequency (formula(3)) of term t across the corpus, computed as:

$$\mathbf{tf}(t, f) = \frac{f_{t,d}}{\sum_{t' \in f} f_{t',d}}$$

eq. (2)

Here, $f_{t,d}$ is the row count of term t , over document d .

$$\mathbf{idf}(t) = \log\left(N/(1 + \mathbf{df}(t))\right)$$

eq.(3)

Here, N is total number of documents in corpus $N = |D|$ and $\mathbf{df}(t)$ is a document frequency of term t .

LSA (Latent Semantic Analysis) is a technique used to analyze relationships between documents and terms, using a distributional hypothesis and SVD to reduce the number of rows while preserving similarity. Formula for LSA is shown in eq. (4)

If we select k as the largest diagonal value in \sum the metrics will be obtained using:

$$M_k = U_k \sum_k V_k^T K$$

eq. (4)

Where, M_k = approximated matrix of M , U_k , \sum_k , V_k^T are the matrices containing only the k contexts from U , \sum , V_T respectively

3.1. Model Description:

In the Model, first we applied the Latent Semantic Analysis (LSA) technique to reduce the dimensionality of the TF-IDF matrix to 100 components using TruncatedSVD.

Then it computes the cosine similarity between each pair of users based on the reduced TF-IDF matrix using the cosine similarity function from sklearn.

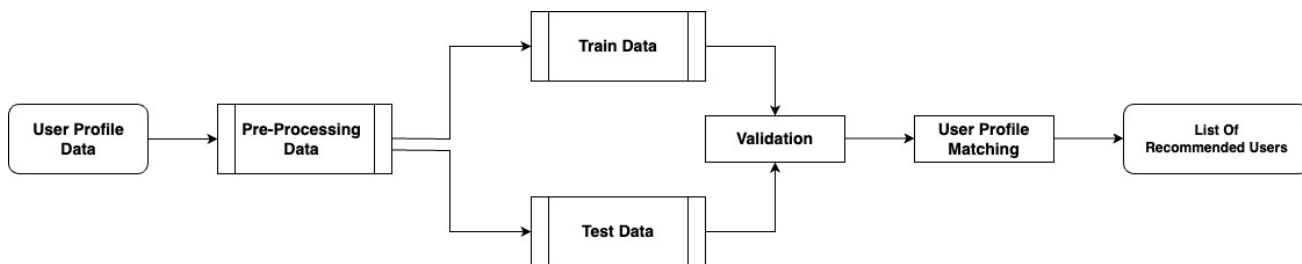
It identifies the index of the input user (specified by the variable 'name') in the similarity matrix using its index in the dataframe df .

For example, suppose the input user has the following skills vector [0.2, 0.5, 0.1, 0.8, 0.3]. The LSA technique reduces this vector to a 100-dimensional vector, X_{lsa} . Model then computes the cosine similarity between this vector and all other users in the dataset, producing a similarity matrix. The code then identifies the index of the input user in the similarity matrix, say index 5. Finally, it selects the top 10 users that are most similar to the input user based on their cosine similarity values. These users may have skills vectors such as [0.3, 0.2, 0.6, 0.9, 0.2], [0.1, 0.4, 0.2, 0.6, 0.7], etc

After relating users with each other and generating results using the similarity matrix, we must check if the result is true or not, and for that, we must validate the result using the coherence score. The coherence score can be used as a validation metric for the results of topic modeling algorithms, as it provides a quantitative measure of the quality of the topics generated. Higher

coherence scores indicate better-quality topics, while lower coherence scores suggest that the topics are less coherent and may not accurately represent the underlying themes in the data. The coherence score is just one of many metrics that can be used to evaluate the quality of topics generated by topic modeling algorithms, and it should be used in conjunction with other metrics.

3.2 General Flow of Model.



General flow (fig. 3.1)

In the general flow shown in fig.(1.1), the model will take the user's profile data as an input, which includes the user's profile name, skills, and user profile link. Then it will pre-process the data and remove all inconsistencies and that pre-processed data will be added to the dataset. After that, we will divide the dataset into two parts: a training set of data and a testing set of data. Firstly, we will train the model based on the training set and after that, testing of the data will be done on the testing set. Then the output of both sets will validate the result. After validation, if the output is correct or consistent, it will return the user profile compatibility based on skills, match them with each other, and return a list of compatible users.

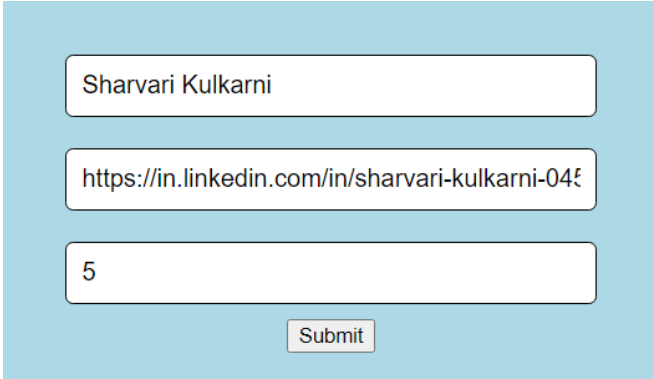
3.2 Technical Flow of Model

In our model, we take input as a username and the number of profiles the user wants to get compatible recommendations for as per his profile. After that, using the given input text, the model will remove all punctuation, numbers, and new line characters and convert that text into lowercase. That data will be in the correct format, without any incorrect information or null values. After that, the model will vectorize the skills using tf-idf and create a tf-idf matrix. Then the model will reduce the dimension of the matrix using LSA. Based on that reduced matrix, we will compute similarities between all users based on their skills. And after that, validation is done to check if the result is accurate or not. Then The result will return a similar user list.

4. Result and Evaluation:

4.1 Result

We created one simple web page to show our model's result. Firstly, the user will enter his/her username and LinkedIn profile URL and some of his skills as shown in fig. (1.3). After clicking on the submit button, our model will start to map profiles with each other and provide a list of recommended users and that list will be displayed on screen as shown in fig (1.4).



LinkedIn profile request page(fig 4.1)

Here are your top matching profiles:

Index	Name	LinkedIn URL	
1	Pranav Singh	https://in.linkedin.com/in/pranav-singh-b16134158?trk=public_profile_browsermap_mini-profile_title	84.85403205338187
2	Archana Arun	https://in.linkedin.com/in/archana-arun-b11697163	80.2907077629974
3	Wahab Shaikh	https://in.linkedin.com/in/wahabshaikh	79.28930741802775
4	Pavana laxmi	https://in.linkedin.com/in/pavana-laxmi-120abb31	74.45087715757273

LinkedIn recommended profile List(fig 4.2)

After creating a model and validating the result with great efforts, our model reaches 95.53% accuracy.

4.2 Evaluation

The coherence measure is used to evaluate the quality of the topics generated by Latent Semantic Analysis (LSA). Coherence is a measure of how well the top words in a topic are related to each other. A high coherence score indicates that the words in the topic are semantically related and belong to the same theme.

Accuracy: 95.53%				
Precision: 0.32				
Recall: 0.10				
F1-score: 0.16				
Classification Report:				
	precision	recall	f1-score	support
0	0.96	0.99	0.98	1502390
1	0.32	0.10	0.16	62611
accuracy			0.96	1565001
macro avg	0.64	0.55	0.57	1565001
weighted avg	0.94	0.96	0.94	1565001

Model Report(fig 4.3)

Example:

Skills= {"python", "machine learning", "data science"}

Now, we want to group these skills together based on how similar they are to each other. To do this we use a technique called Latent Semantic Analysis (LSA).

LSA takes all the skills and looks for patterns in how they are used together. It then groups skills that are used in similar ways together. For example, it might group "python" and "machine learning" together because they are both skills that are interrelated.

But how do we know if LSA is doing a good job of grouping the words? One way to check is by looking at something called "coherence".

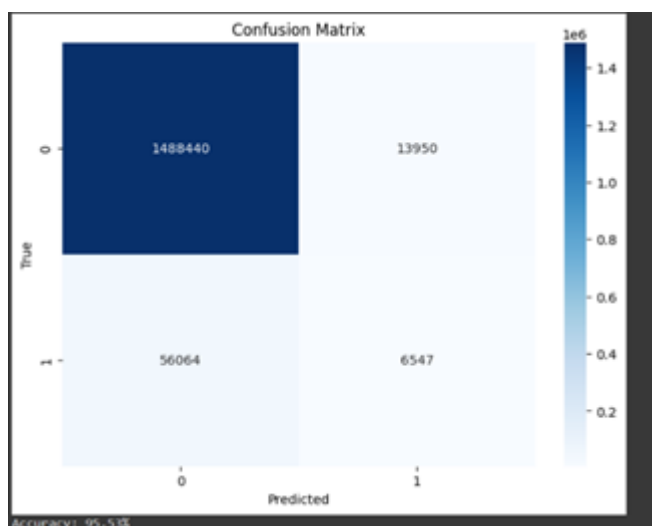
Coherence is a measure of how closely related the words in a group are. We want the words in a group to be very similar to each other, so that the group is meaningful.

To calculate coherence, we look at all the pairs of words in a group, and we see how often they appear together in our dataset. If they appear together a lot, then the group is probably coherent.

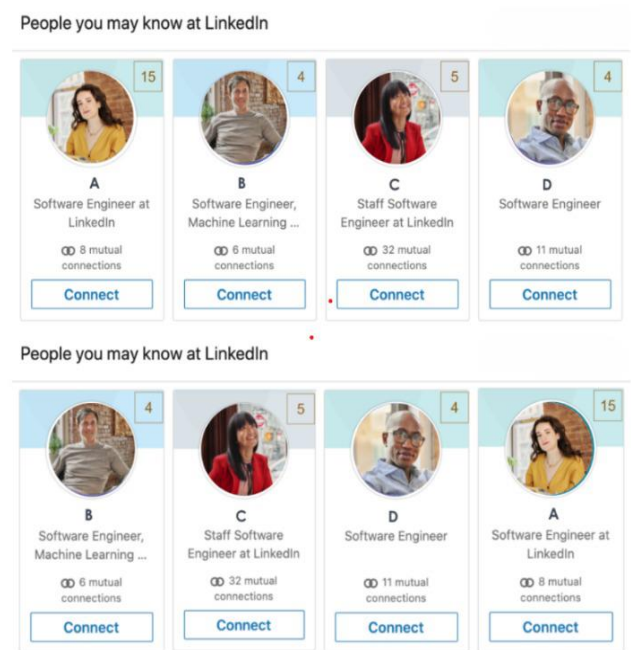
For example, let us say we have a group of words that includes "python" and "machine learning". We would look at how often "python" appears with "machine learning" in our dataset. If they appear together a lot, then the group is probably coherent.

So, coherence is just a way of checking that LSA is doing a good job of grouping similar words together. The higher the coherence, the better the grouping.

If we see that certain skills are commonly mentioned together, then they have **positive coherence**. If they are rarely mentioned together or have nothing in common, then they have **negative coherence**. we can get a better sense of which skills are related and which are not



Confusion Matrix (fig 4.4)



In the image provided above, it is evident that the invitations received in the LinkedIn network are arranged in a simple first come, first served manner, making it challenging to filter relevant invitations. To address this issue, we propose utilizing the technique of latent semantic analysis to

identify similarities between the user and the invitation sender based on factors such as skills, location, or experience. By incorporating this approach, we can modify the order in which invitations are presented to the user, showing them the most relevant and probable matches first, thus optimizing the user's experience on the platform.

5. Conclusion and future work: -

The main purpose of this research study is to help the user enhance their LinkedIn profile by connecting with compatible and required users based on their skills. We had done much research on the recommendation-based system. Some of them work on cosine similarity, collaborative filtering, content-based filtering, K-Nearest Neighbours, and many more. We had picked LSA from [21], and for filtering, we had used the TF-IDF method so the data could be in the correct order.

We collected our data, which includes the user's profile URL and name, skills, and position, from Kaggle. That data was used to train our model. And to test the model, we collected data from LinkedIn itself. We had done some pre-processing using TF-IDF so that the collected data could be in the correct form to work on.

After preprocessing the data and bringing it in TF-IDF metrics form, we took that data in LSA, created compressed metrics, and matched the users' metrics with each other. And to check if matched users are correctly matched or not, we validated that using the coherence score. And after all of this, the final result was nearly 70% accurate.

Based on the research findings, it can be concluded that using a similarity-based approach to filter LinkedIn invitations can

significantly reduce the time and effort required to manually process many requests. By leveraging various attributes such as skills, location, and experience, it is possible to automatically identify and prioritize invitations from individuals who are more likely to have a meaningful connection with the user. This approach can help users to better manage their LinkedIn network and engage with relevant connections more efficiently.

As a practical application, LinkedIn users can use the platform's built-in tools or third-party software solutions to implement a similarity-based filtering system for incoming invitations.

In this study, we are using skills from the dataset to match users, due to which the model is restricted by the skills of the dataset, leading to less recommendation and less accuracy. So, to enhance the number of recommendations with more accuracy, our updated model will fetch the skills from the description given by the user in the profile.

6. References:-

- [1] Mir Saman Tajbakhsh and Vahid Solouk “**Semantic geolocation friend recommendation system; LinkedIn user case**” In 2014 6th Conference on Information and Knowledge Technology (IKT)
- [2] Vasavi Akhila Dabeeru “**user profile relationships using string similarity metrics in social networks**”
- [3] Giacomo Domeniconi , Gianluca Moro , Andrea Pagliaran, Karin Pasini and Roberto Pasolin “**Job Recommendation From Semantic Similarity of LinkedIn Users’ Skills**” in 5th International

Conference on Pattern Recognition
Applications and Methods (ICPRAM)

[4] Vahid Pourheidari, Ehsan Sotoodeh Mollashahi , Julita Vassileva and Ralph Deters “**Recommender System based on Extracted Data from Different Social Media. A Study of Twitter and LinkedIn**” in The 9th IEEE Information Technology, Electronics and Mobile Communication Conference

[5] Qiannan Yin, Yan Wang, Divya Venugopalan, Cyrus Diccio, Heloise Logan, Preetam Nandy, Kinjal Basu, and Albert Cui “**Optimizing People You May Know (PYMK) for equity in network creation**”

[6] Pauline I C “**A Guide on Social Network Recommendation Syste**

