www.ijecs.in

International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 6 Issue 2 Feb. 2017, Page No. 20221-20225
Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i2.10

Survey on Matching Of Users in Social Networks Using Friend Book

***Ms. D.Saral Jeeva Jothi*[1], *Ms.R.Ramyadevi*[2]**

[1]M.E., Department of Computer Science and Engineering, Velammal Engineering College, Chennai
[2]Assistant Professor, Department of Computer Science and Engineering, Velammal Engineering College, Chennai
saranive23@gmail.com, ramyakathir@gmail.com

Abstract:
People use various social network sites for different purposes. Social networks provide an important source of information regarding users and their interactions which is very valuable for identifying the identical users and recommender systems. In this survey paper we aims to address the identical user identification problem and recommending friends based on lifestyle of the users in social networking sites (SNS). A methodology called MOdeling Behaviour for Identifying Users across Sites (MOBIUS) is used for finding a mapping among identities of individuals in social media sites. Recommender systems or recommendation systems are a subclass of information filtering system that search for to predict the 'preference' or 'rating' that a user would give to an person/item/place/thing. Social Networking services focuses towards suggesting friends based on Users Social Graph or Geo-location based, which does not take user's liking, disliking etc. This survey also investigates about an app that utilizes the information of the user and makes recommendations by considering user's point of interest and calculating the similarities between each user, thus recommending the friends to the user in heterogeneous sites.

Keywords: Social network analysis, User identification, Friend relationship, MOBIUS, Friend recommendation, Mobile sensing, User interest.

1. INTRODUCTION

Social Media are computer-mediated tools that allow people or companies to create, share, or exchange information, career interests, ideas, and pictures/videos in virtual communities and networks. Social media is defined as "a group of Internet-based applications that build on the ideological technological foundations of Web 2.0, and that allow the creation and exchange of user generated content. Social networking is expanding the number of social contacts by making connections through individuals. While social networking has gone on almost as long as societies themselves have existed, the unparalleled potential of the internet to promote such connections is only now being fully recognized and exploited, through Web-based groups established for that purpose. It establishes interconnected Internet communities that help people make contacts that would be good for them to know. Websites dedicated to social networking include Linkedin, Facebook etc. Social networks are now visited more often than personal email is read. Some Social networks have grown to such enormous proportions that they rival entire countries in terms of population- If Facebook, for example, was a country it would be the fifth-most populated in the world. Facebook is one among the most popular sites in this era of communication and sharing. According to the statistics, Facebook has 1.4 billion active users!

User Identification refers to now a days more and more people have their virtual identities on the Web. It is common that people are users of more than one social network and also their friends may be registered on multiple web sites. A facility to aggregate our online friends into a single integrated environment would enable the user to keep up-to-date with their virtual contacts more easily, as well as to provide improved facility to search for people across different websites. A method was developed to identify users based on profile matching and Network structures.

Figure 1.1 Social Network

Twenty years ago, people typically made friends with others who live or work close to themselves, such as neighbours or colleagues. With the rapid advances in social networks, services such as Facebook, Twitter and Google+ have provided us revolutionary ways of making friends. According to Facebook statistics, a user has an average of 130 friends, perhaps larger than any other time in history. One challenge with existing social networking services is how to recommend a friend based on the user according to his needs. Most of them rely on pre-existing user relationships to pick friend candidates. For example, Facebook relies on a social link analysis among those who already share common friends and recommends symmetrical users as potential friends. Unfortunately, this approach may not be the most appropriate based on recent sociology findings.

Traditional way of making friends (G-Friend)

➢ Geographical location based:

Neighbours, colleagues

Pros: be familiar with each other

Cons: Number of friends is limited
Emerging social networks

- Facebook, Twitter, Google+, etc.

Pros: unlimited number of friends
Cons: “Friends” are not the expected friends

But the real fact in the world is:
People tends to make friends with people having similar interests

Existing system recommends based on social graphs which may not be the most appropriate to reflect a user’s preferences. So, a new method is devised which recommends friends based on lifestyles of users which may include the following parameters: 1) activities, 2) location, etc.

Different Problems are:

- How to identify friend candidates based on interests rather than pre-existing relationships?
- How to automatically get one’s interests without one’s specification?
- How to help people find friends at any time and any place?
- How to measure the similarity of interests between different users?

There are lots of solutions to solve the all above issues. The aim of this paper is to find the identical user’s profile in SMN, eliminate those fake users and build Friend Recommendation System by using activities, interest of users to suggest friend as required.

Different objectives achieved in this paper are as follows:

- Identification of identical user in social network sites (SNS) and elimination of fake user account from the SNS.
- Designed and implemented a friend recommendation system that allows users with similar interests to be quickly identified and recommended.

2. LITERATURE REVIEW

User identification is also called user recognition, user identity resolution, user matching, and anchor linking. Although no solution can identify all identical anonymous SMN users, some SMN elements may be used to identify a portion of users across SMNs. Many studies have addressed the user identification problem by examining public user profile attributes, including screen name, birthday, location, gender, profile photo, etc.[1][2][4]. Since these attributes do not require exclusivity and are easily faked by users for different purposes (including malicious users), these schemes are quite fragile. Some researchers have leveraged public user activities to recognize users using post time, location and writing style. Since location data is difficult to obtain and writing style is difficult to extract from short sentences, these techniques are plagued by limitations. A screen name is the publically required profile feature in almost all SMNs. It has been widely explored as a way to recognize users across different SMNs.

2.1 Username Uniqueness

Perito, et al.[1] proposed how unique and traceable are usernames by calculated the similarity of screen names and identified users using binary classifiers. Profiling unique identities from multiple public profiles is a challenging task, as information from public profiles is often incorrect, misleading or altogether missing. First, introduce the problem of linking multiple online identities relying only on usernames. Second, devise an analytical model to estimate the uniqueness of a username, which can in turn be used to assign a probability that a single username, from two different online services, refers to the same user. Based on language models and Markov Chain techniques. Third, extend this model to cases when usernames are different across many online services. Finally, by applying the technique to subsets of usernames they extracted from real cases scenarios, validate and discuss the technique in the wild. This paper explores the possibility of linking user’s profiles only by looking at their usernames. The problem is different web services has different username policies. All the methods tried have high precision in linking username couples that belong to the same users.

2.2 A Behavioral-Modeling Approach

Reza Zafarani, et al.[2] proposed connecting users across social media sites A behavioural modelling approach by an algorithm called learning algorithm. The proposed behavioural modelling approach exploits information redundancy due to these behavioural patterns. An alternative solution addressing the age verification problem by exploiting the nature of social media and its networks. The information available on all social media sites (usernames) to derive a large number of features that can be used by supervised learning to connect users across sites. Users often exhibit certain behavioural patterns when selecting usernames. It includes analysing these possibilities and discovering features indigenous to specific sites, beyond those constricted to usernames, and incorporating them into MOBIUS for future needs.

2.3 User Identification across Social Media

Cortis**,** et al.[3] proposed user identification across
social media a framework for authorship identification using the writing style of online messages and classification techniques. Proposed a methodology for connecting individuals across social media sites (MOBIUS). MOBIUS takes a behavioural modelling approach. MOBIUS employs minimal information available on all social media sites (usernames) to derive a large number of features that can be used by supervised learning to effectively connect users across sites. Users often exhibit certain behavioural patterns when selecting usernames. The proposed behavioural modelling approach exploits information redundancy due to these behavioural patterns. Categorized these behavioural patterns into (1) human limitations, (2) exogenous factors, and (3) endogenous factors. MOBIUS employs supervised learning to connect users. The empirical results show the advantages of this principled, behavioural modelling approach over earlier methods. Identifying users across social media sites opens the door to many interesting applications. Future work includes incorporating MOBIUS along with the username.

2.4 Identities across Communities

Reza Zafarani[4] proposed connecting corresponding identities across communities by a technique called link analysis algorithm. The relationship between usernames

DOI: 10.18535/ijecs/v6i2.10

selected by a single person in different communities, and on some of the web phenomena regarding usernames and communities. The unrevealing nature of the web and the fact that most communities preserve the anonymity of users by allowing them to freely select usernames instead of their real identities and the fact that different websites employ different username and authentication systems. Nevertheless, if there exists a mapping between usernames across different communities and the real identities behind them, then connecting communities across the web becomes a straight forward task.

2.5 Identifying users across different sites

Yubin Wang, et al.[5] proposed identifying users across different sites using usernames by a technique called abbreviation detection method and self-information vector model. 59% of individuals prefer to use the same username(s) repeatedly, mostly for ease of remembering. Self-information vector model integrate content and pattern features extracted from usernames into vectors. It uses cosine similarity. Abbreviation detection method to identify the abbreviations in usernames and helps to increase user identification results.

2.6 Personalized Recommendation

Xueming Qian, et al.[6] proposed personalized recommendation combining user interest and social circle by considering three social factors personal interest, interpersonal interest similarity and interpersonal influence. These factors are fused into a new recommendation model based on probabilistic matrix factorization. They tried to solve the cold start problem of the user. But the experiments were performed by collecting the three to four months old historical data from three shopping sites Yelp, Douban and Epinions. The main contributions of the paper is to propose a personalized recommendation system combining the social factors which would the direct connections between the user and item vectors. They also proposed a personalized recommendation approach by enforcing user personal interest, modelled to get an accurate model for cold start and sparsity problem of the user.

2.7 A Semantic based Friend Recommendation System

Zhibo Wang, *et al*.[7] proposed a semantic-based friend recommendation system for social networks by modelling recommendation system using collaborative filtering (CF).The author used users work profession or daily activities like walking, shopping, sitting, typing etc. as life style activity. Gathering this data author tries to extract relevant data and using pattern matching algorithms author recommends candidate friends to the user. But using only professional data may not be the best case to suggest friend.

2.8 Ranking model for Online Activity

Sachin V Jose[8] proposed incremental iterative time spent based ranking model for online activity based friend group recommendation systems by using an algorithm called latent Dirichlet allocation (LDA). LDA is a probabilistic model which is used in text mining. It consists of Document which is then collection of topics and topic which is consisting of words. In this firstly need to decide the topic. For e.g. in LDA suppose topic is Cat related then it will have words with their probabilities such as milk, kitten, and meow. The author propose a system that recommends based on the daily activities of users. Here a semantic based friend recommendation is done based on the user's life styles such as posting, chatting, searching, commenting etc. By connecting users with similar professional background or similar interests, online social networks open up a new platform for information sharing and social networking.

2.9 Recommendation Based on Collaborative Filtering

Changchun Yang, et al.[9] proposed personalized recommendation based on collaborative filtering in social network. This paper analysed the disadvantages of the traditional collaborative filtering, and proposed an improved algorithm of personalized recommendation based on collaborative filtering in SNS, which increase the efficiency of the recommendation. This system recommend users information, services or products by data mining based on the user's preferences, interests, behaviour, or needs.

2.10 A Collaborative Filtering Recommendation Based on User Profile and User Behaviour

Yang, et al.[10] proposed a collaborative filtering recommendation based on user profile and user behaviour in Online Social Networks that discovers life-style of users and calculates the similarity between them, and recommends friends to users if their life-styles have high similarities. They have implemented it on android based smart-phones. It uses collaborative filtering algorithm to identify similarity based on three features user rating on items, user's profile and activity. The user's profile, user's activity and user's social friend information can help to improve the prediction accuracy of recommender systems.

3. METHODOLOGY

3.1 Maximum Likelihood Estimation (MLE)

To measure the username uniqueness first estimate the probability 'P(u)' for each username 'u'. Given a dataset of usernames from different services contains username, email address and password. If there are 'N' usernames then estimate the probability of each username 'u' as $\frac{count(u)}{N}$ if u belongs to the given dataset, and 0 otherwise. The drawback of the MLE approach is that it cannot be used to give any estimation for the usernames not in the given dataset sample. The estimation given is very rough [1].

3.2 MOBIUS

MOdeling Behaviour for Identifying Users across Sites (MOBIUS) is used for finding a mapping among identities of individuals across social media sites. It consists of three key components:

- Identifies user's unique behavioural patterns that lead to information redundancies across sites
- Constructs features that exploit information redundancies due to these behavioural patterns
- Employs machine learning for effective user identification.

MOBIUS is effective in identifying users across social media sites.

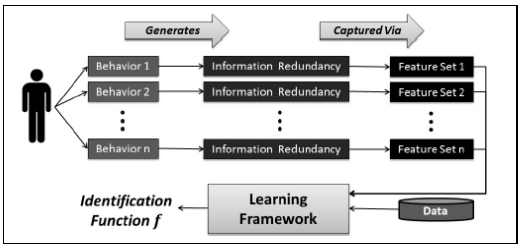


Figure 3.2.1 MOBIUS: Modeling Behavior for Identifying Users across Sites [2]

MOBIUS uses behavioural patterns among individuals while selecting usernames. These behavioral patterns can be categorized as follows:

1. Patterns due to Human Limitations
2. Exogenous Factors
3. Endogenous Factors

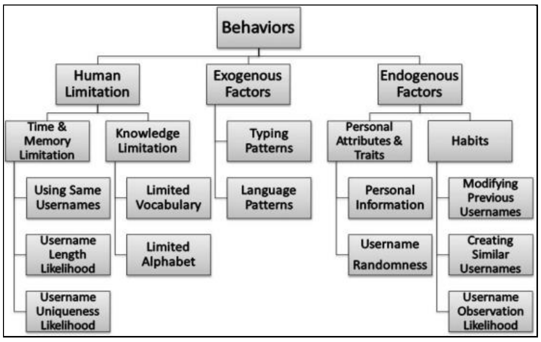


Figure 3.2.2 Individual Behavioral Patterns when Selecting Usernames [3]

Constructed features contain sufficient information for user Identification, importance or relevance of features can be assessed thus features can be selected based on particular application needs and adding more features can further improve learning performance. MOBIUS can help solve the problem of age verification.

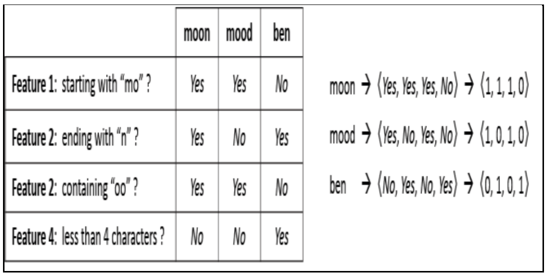
3.3 Self-information Vector Model

The self-information vector model is used to integrate multiple features extracted from each User name into a vector, then the problem of quantifying the similarity between two usernames is translated into the calculation of similarity between their self-information vectors. Given a feature λ and a username u, feature indicator function $I_\lambda(u)$ is defined to indicate that whether u satisfies feature λ:

$$I_\lambda\ (u) = \begin{cases} 1\ if\ u\ satisfies\ the\ feature\ \lambda \\ \qquad\qquad 0\ otherwise \end{cases}$$

Then extract *m* features $\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ from username *u*, then we can construct a binary vector *Bu* of *m* members for *u*:

$$Bu = \{I_{\lambda 1}(u)\ ,\ I_{\lambda 2}\ (u)\ ,\ \ldots\ ,\ I_{\lambda m}\ (u)\}$$



	moon	mood	ben	
Feature 1: starting with "mo"?	Yes	Yes	No	moon → (Yes, Yes, Yes, No) → (1, 1, 1, 0)
Feature 2: ending with "n"?	Yes	No	Yes	mood → (Yes, No, Yes, No) → (1, 0, 1, 0)
Feature 2: containing "oo"?	Yes	Yes	No	ben → (No, Yes, No, Yes) → (0, 1, 0, 1)
Feature 4: less than 4 characters?	No	No	Yes	

Figure 3.3.1 Example of representing usernames as binary vectors [4]

Usernames moon, mood and ben can be represented as <1, 1, 1, 0>, <1, 0, 1, 0> and <0, 1, 0, 1>. Then we can get the number of common features between two usernames by counting the number of positions in two binary vectors both with value 1. Self-information is derived to measure the quantities of information. The self-information of a specific message *m* is defined as *I* (*m*) = − log Pr (*m*). Then it can represent each username as a self-information vector.

3.4 Probabilistic topic model

Given "documents", the probabilistic topic model could discover the probabilities of underlying "topics". The probabilistic topic model is used to discover the probabilities of hidden "life styles" from the "life documents".

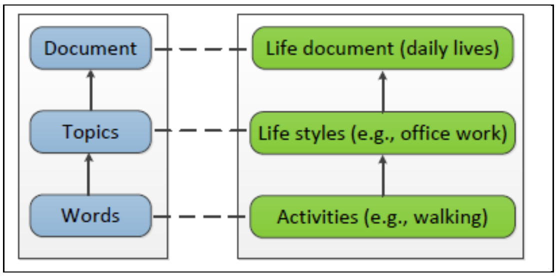

Document
Topics
Words
Life document (daily lives)
Life styles (e.g., office work)
Activities (e.g., walking)

Figure 3.4.1 Analogy between word documents and people's daily lives [7] [8]

In probabilistic topic models, the frequency of vocabulary is particularly important, as different frequency of words denotes their information entropy variances. It uses the "bag-of-activity" to replace the original sequences of activities recognized based on the raw data with their probability distributions. Thereafter, each user has a bag-of-activity representation of his/her life document, which comprises a mixture of activity words.

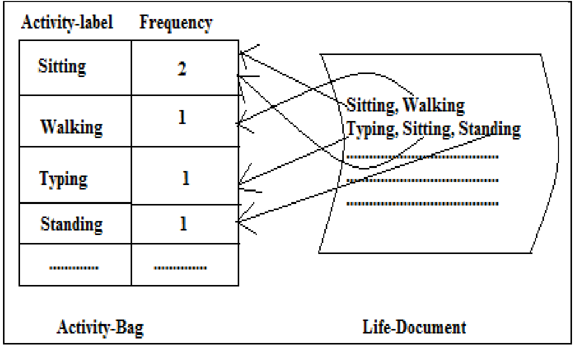


Figure 3.4.2 Activity-Bag representation [7]

Let w = [$w_1, w_2, \ldots\ldots, w_W$] be the set of 'W' Activities
z = [$z_1, z_2, \ldots\ldots\ldots, z_Z$] be the set of 'Z' Lifestyles
d = [$d_1, d_2, \ldots\ldots\ldots, d_n$] be the set of life-documents
Where n = Number of users and

$p(w_i | d_k)$ is the Probability of activity w_i in a life-doc d_k
$p(w_i | z_j)$ is the Probability of how much the activity w_i contributes to the lifestyle z_j
$p(z_j | d_k)$ is the Probability of lifestyle z_j embedded in a life-document d_k

By using this, we will automatically get one's interest without one's specification.

Let $f_k(w_i)$ be the frequency of occurrence of activity w_i in life-doc d_k.

$$p(wi | dk) = \frac{fk(wi)}{\sum_{i=1}^{W} fk(wi)}$$

where $f_k(w_i)$ denotes the frequency of w_i in d_i.

4. CONCLUSION AND FUTURE WORK

The main focus of this survey is identifying the identical users across social networking sites and recommending friends to the users. This survey provides a study about component model, techniques and algorithms used for identifying identical users across sites and recommendation system. Maximum Likelihood Estimation uses only the user name to find the identity of the user. User name format may vary in various social networking sites. MOBIUS uses username along with profile attributes, such as gender, location, interests, profile pictures, language, etc., help in better identify individuals. Social Networking services focuses towards suggesting you friends based on Users Social Graph or Geo-location based, which neither take users life style into account nor user's interest liking, disliking etc. Suggesting friends using the user's link analysis may not be the best preference of suggestion for the users. In order to overcome these problems friend circle can be used to identify identical users which is highly reliable and with user's interest recommend friends to the users. By connecting users with similar interests, online social networks open up a new platform for information sharing and social networking.

5. REFERENCES

[1] Perito, Castelluccia, Kaafar, and Manils, "How unique and traceable are usernames?" in Proc. 11th Int. Conf. Privacy Enhancing Technol., 2011, pp.1–17.

[2] Zafarani and Liu, "Connecting users across social media sites: a behavioral-modeling approach," in Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2013, pp. 41–49.

[3] Reza Zafarani, Lei Tang, and Huan Liu. 2015. "User identification across social media," ACM Trans. Knowl. Discov. Data 10, 2, Article 16 (October 2015), 30 pages.

[4] Zafarani and Liu, "Connecting corresponding identities across communities," in Proc. 3rd Int. ICWSM Conf., 2009, pp. 354–357.

[5] Yubin Wang, Tingwen Liu,"Identifying Users across Different Sites using Usernames",in Proc. ICCS ,Volume 80, 2016, Pages 376–385.

[6] Qian, Feng, Zhao, and Mei, "Personalized recommendation combining user interest and social circle," IEEE Trans.Knowl. Data Eng., vol. 26, no. 7, pp. 1763–1777, Jul. 2014.

[7] Zhibo Wang, Jilong Liao, Qing Cao, Hairong Qi, and Zhi Wang, "Friendbook: A Semantic-based Friend Recommendation System for Social Networks," IEEE Transactions On Mobile Computing, 1536-1233 (c), 2013.

[8] Sachin V Josef, MinuLalitha Madhavu2, " Incremental iterative time spent based ranking model for online activity based friend-group recommendation systems," 6th ICCCNT – 2015 July ,13 - 15, Denton, U.S.A.

[9] Changchun Yang, Jing Sun and Ziyi Zhao, "Personalized Recommendation Based on Collaborative Filtering in Social Network,"978-1-4244-6789-1110/ ©2010 IEEE.

[10] Lu Yang, Anilkumar Kothalil Gopalakrishnan," A Collaborative Filtering Recommendation Based on User Profile and User Behavior in Online Social Networks ," 2014 International Computer Science and Engineering Conference (ICSEC) 978-1-4799-4963-2/14/ ©2014 IEEE.

[11] Face book statistics. http://www.digitalbuzzblog.com/ facebook-statistics-stats-facts-2011/.

[12] RenRen. http://www.renren.com/.

[13] Amazon. http://www.amazon.com/.

[14] Netfix. https://signup.netflix.com/.