International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-5S3, July 2019

# Understanding users Display-Name Consistency across Social Networks

**Waseem Ahmad, Rashid Ali**

***Abstract:*** ***Online users create their profiles on numerous social platforms to get benefits of various types of social media content. During online profile creation, the user selects a username and feeds his/her personal details like name, location, email, etc. As different social networking services acquire common personal attributes of the same user and present them in a variety of formats. To understand the availability and similarity of personal attributes across various social networking services, we propose a method that uses the different distance measuring algorithms to determine the display-name similarity across social networks. From the experimental results, it is found that at least twenty percent GooglePlus-Facebook and Facebook-Twitter users select the same display name, while forty five percent Google and Twitter user select identical name across both the social networks.***

***Index Terms: Cross link posts, Personal Information, Social Account, User Identity***

## I. INTRODUCTION

The emergence of smart mobile phone has proved to be boon for the rapid proliferation of online social networks in rural areas.A recent survey reveals that out of total 4.388 billion Internet users around 3.484 billion users have created at least one account on social media [1]. Another survey [2], states that the majority of users prefer Facebook (2.23 billion active users) and YouTube(1.9 billion users) for entertainment and learning purposes. This survey further describes that seventy three percent Twitter users are also the member of Instagram, while ninety one percent Twitter users have an account on Facebook, similarly ninety five percent Twitter users prefers YouTube. Therefore, people use several social media platforms simultaneously for different purposes, for instance; user shares personal posts on Facebook, get latest news updates via Twitter, publishes own photographs on Instagram, upload videos on YouTube, and select favorite fashionable items from Pinterest. Finding the same user across the different social networks could be useful in the business, marketing, crime detection and item recommendation fields.

Different social media services collect the user's personal trait to create a user profile. During online profile creation an online user gives his/her identity, i.e. username, real-name, birth place, birth-date, hobby, preferred items, emailaddress, and personal contacts. Many impure minded people, create their identity on social networks by feeding fake personal attributes. To overcome with such issues, recently, many social media services started cross verification of suspected profiles by sending OTP (One time password) on registered mobile numbers. Therefore, in this work we assume that users shared personal attributes are real and authentic. Nowadays, social media users are treated as a customer and they get the recommendationsfrom various agencies via email or personal message (marketing, hotel, tours and travel). User's revealed personal information can be classified into two main categories, namely; nonsensitive and sensitive information. Name, username, profile pictures, and hobbies etc.,can be considered as non-sensitive personal information. Whereas the personal contact number, email-address, birth-date and office or home addresses are considered as sensitive information. Leakage of both sensitive and non-sensitive information can be exploited by the suspected person to hurt an individual.

Basically a user's social network profile contains three types of information, namely; network, profile and the content. In this paper, we make use of profile based information to find the attribute similarity across the networks.Many online users share their profile information via posting URLs on a particular social media (many social media-advertisersshares their other social account information on Twitter). In the profile section, display-name is assumed as one of the most authentic attributes and it plays a very crucial role in username generation and internetwork formation. In the social network field, even though an individual has not revealed his/her identity personally, in spite that his personal life may be at risk due to his/her social connections. In this paper, we explore the problem of attribute homogeneity by exploiting the usersdisplay-name across three different social networks.

To find the user's display name similarity across different social networks in this paper, we study theuser's display-name attribute and select the appropriate feature using different distance measuring algorithm to discriminate matching attributes with non-matching attributes. The set of algorithmsis applied on the different datasets: Twitter-Google Plus, Twitter-Facebook and Facebook-Google Plus. The applied algorithms find exact display-name, similar display-name, display-name with prefixes and suffixes and lastly, display-name with reversing the order of first and the last name.

In section II, we describe the recent work done by the researchers and practitioners in user account matching across various online social networks.

Section III, presents the problem formulation and its description, while section IV gives the overview of the statisticaldescription of data along with data gathering and cleaning procedure.Section V demonstrates the results and its discussion. Finally, section VI concludes the work.

**Manuscript published on 30 July 2019.**
***** Correspondence Author (s)
**Waseem Ahmad**, Computer Engineering, Aligarh Muslim University, Aligarh, India.
**Rashid Ali**, Computer Engineering, Aligarh Muslim University, Aligarh, India.

## II. Related Work

Social networking services encourage people to expose more and more personally identifiable information (PII) [3] in a variety of formats. There are two main reasons for personal information leakage in online social media; (1) user'slack of knowledge of privacy mechanism provided by the site (new and rural area user) (2) User who joins media for own reputation (like celebrities or social media marketing vander). Personally identifiable information is classified into three main categories, namely; Profile based (personal attribute), content-based (user-generated text and content) and structure-based information (friendship and interaction networks).

### A. Profile based Information

In a social environment user often divulgeself identity like username, name, gender, birth-date, etc. Many authors used such attributes across the different sites to find identical users across the networks. Vosecky et al. [4], exploits profile information to identify users across social networks. Esfandyari et al. [5], extracted profile attributes present on the three most popular social media sites; Twitter, Facebook, andgoogle+. Thereafter they apply supervised machine learning algorithms on the common attributes available on these media to find identical users. Zafarani et al. [6], study the profile attribute 'username' and describe that a user generally selects maximum up to four user-names across all the sites due to human limitation, exogenous and endogenous factors. They further suggested that around 59% of the users select the same username across the sites. Malhotra et al. [7], suggested that user-ID and name are the most discriminating features for finding the user's identity across Twitter and LinkedIn. In the recent work based on profile information, researchers have to cope with two difficulties such as privacy and profile utility [3,8]. Motoyama and Verghese [8] matched the profile attribute to find the user's identity across Myspace and Facebook. Further, Raad et al. [9], studies social networks and compare the profile using similarity measures. Profile-based information retrieval from different social networks in the current scenario is a very crucial task due to users' privacy and roped security mechanism. As far as authors' knowledge is concerned, vectorattributes similarity is often used by the researchers, but it requires lots of effort during data retrieval again, it's pairwise matching across the different social networks using different similarity measure algorithms.

### B. Content based Information

Social media services support the construction and exchange of online user-generated contents [10]. Social media researchers use, user-generated online content as another intuitive approach for user account linkage on social networks by exploiting time of post, writing style, social tagging, Geo-location, etc. It is assumed that a user has a behavioral property that same content may be posted by the same user on different social networks at different interval of time. Lei et al. [11], discusses the roles of user-generated content in different online social networks like tagging, question answering, and micro-blogging services. Li et al. [12], studied the different types of information present on social networks and recommended that single point information may be efficient for finding users identity across the networks. In [13-15] authors used user-generated content like the spatial-temporal location to find the unique user's identity shared on different social networks.

### C. Structure based Information

The structure of the social network is formed by the collection of users identity present on a particular network. Friend, followers, following, likes, connection, etc. are considered as structural information. In literature,the authors used the friend relationship present on the user's profile page and exploited the relationships across the sites to find identical user accounts across the sites. Structural information is of two types, depends on the nature of sites like display name by username or realname. Facebook, LinkedIn, Pinterest, etc. use real-name based display while Twitter, Instagram, etc. use both username and real-name but the mandatory field is the username which is used to get user profile on social media. In this context, Narayan and somatic [16] study Twitter and Flickr network structure and propose a re-identification algorithm to match the anonymized user identity with around 12% error rate. Further Bartunov et al. [17] used user profile along with a social link and applied a conditional random field to find the user identity across Facebook and Twitter, the proposed method performs better than single attributes used in the literature. Korula et al. [18] proposed a reconciliation algorithm that improves the efficiency of user identification by minimizing the error. Liu et al. [19] exploit two heterogeneous social networks Twitter and Foursquare to determine identical user's profile by using the follower /followee relationship. Tan et al. [20] investigated that username based matching works well when a user put the same username across the networks. It is difficult to find the users' identity across the networks when the user chose different user-name and same username hold by the different individuals. Zhou et al. [21] proposed an efficient framework for user identification using network-based information (Friend relationship). Before the application of the give, the framework requires priory user matched pairs as input. Later Zhou et al. [22] modifiedthe friendrelationship-based method in which prior seed users were required [21] and it requires lots of effort. The proposed a new framework without using prior used seed users, it uses pure network relation in the absence priory user matched pairs. Li and Su [23] critically analyzed the FRUIalgorithm proposed in [21] and suggested that if the proper 'seed user' is not selected in the experiment then several controversial nodes are generated and the algorithm stops working without finding mode identical nodes. They suggested the use of closeness centrality in the supplement and termed a method known as FRUI with the proposed suggestion p-FRUI. As we conclude from the above study that profile and structure-based information is more liable for user identification. But in the current scenario, both types of information are the main concern for user privacy and most of the social media sites recommend a new friend with the friendship recommendation method. Social networking sites like LinkedIn and Facebook restricted public availability of connection and friendship relation for unauthorized users. Therefore, only structure-basedinformation research is limited to those sites which display the username publicly otherwise result may be biased towards Known relationship up to first-degree connection (neighborhood).

To overcome the above limitation, we propose a hybrid method which exploits the publicly available text content and publicly available network features.

## III. Problem Formulation

Social network users generally reveal their other used social network information (URLs or Username) in two ways. Firstly, they share their URLs during online conversation with friends or colleagues; secondly, they provide their other used site information as domain URLs section in the profile field section. For example, a Twitter user reference his or her other social account information as by the set of keywords "follow me on ..",Find me on.. " or "connect with me on .." on Twitter. Where ".." represents the name of social media about which users are talking about. In such a way, we can collect many useful seed users. The obtained seeds information is used as input in the proposed method. In this section, we propose an approaches based on display names available on different sites. The proposed approach exploit seed node to obtain display name and their matching using different distance measuring algorithms like Lavenshtein, Jaro Winkler, cosine and Jaccard distance.

### A. Cross-link Posts

To strengthen the friend relationships, many new users or social media marker often leaves some footprints on public media to attract his known people to become a friend. Such footprints may be the username of a particular site or profile URLs, etc. There may be the probability that a user may leak his all social account information on a particular site.

**Example1**: Suppose a user U have an account on Twitter and have a certain number of followers there. After sometime, he has created a new account on any other social network like Instagram and he desires that his friends should follow him on Pinterest. Therefore, he is forced to post his profile URLs on Twitter, which provide a link between the user's account on Twitter and Pinterest (fig. 1). It is the individual's tendency that his network should be large enough to communicate with friend, colleges and the family.

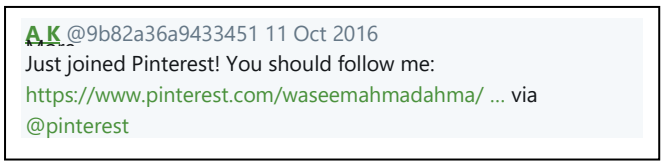
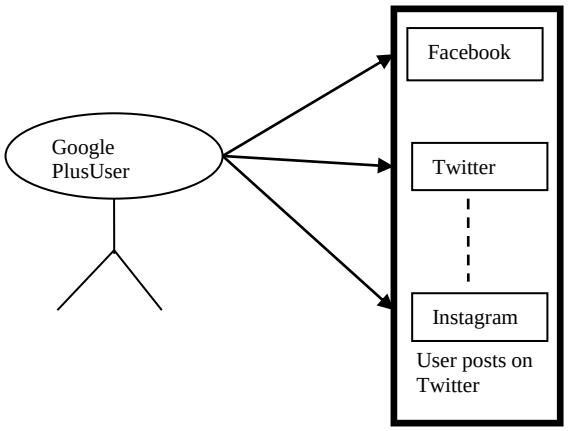


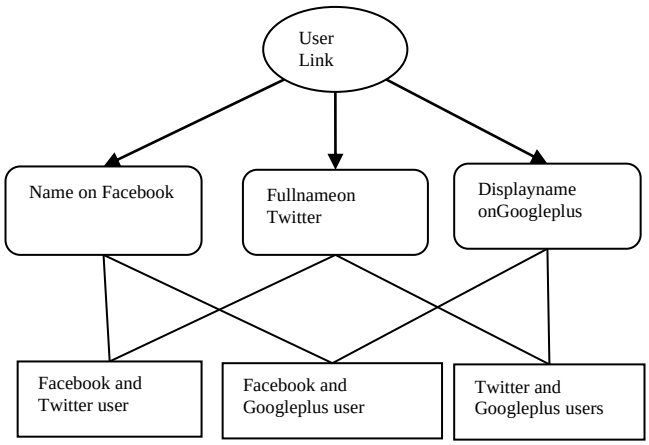

Fig. 1: Single user broadcasts post

**Example 2:** In another scenario, some social marketing people broadcast social accounts (URLs) on public media Twitter to increase the number of followers all over the worlds. Further, to improve the business they post interactive content on his/her different social media account. Hence, a user is compelled to reveal his social account information to them. By which they learn the user's behavior and recommend the desired item to a particular user.


Fig. 2: Single user broadcasts post




(a)




(b)

Fig. 3.(a) A Google Plus user shares his account link (b) User personal identity, similarity using self-disclosed information

Assume that a user U reveal his identity on social networks S viashared links (profile URLs) as depicted in fig. 3(a). While in the fig. 3 (b) represents the extracted profile attributes from the links and the matching of common attributes across the social networks. Social network display-name often include a username and realname. Twitter and Instagram are two popular sites, generally represents friend relation by displaying the username while Facebook, Googleplus and Pinterest only show real-name.

To understand the similarity among the profiles of a user across different social networks, we matched the common features like name using distance measure algorithms (two tokens based and two characters based) namely; cosine, Jaccard, Levenshtein and JaroWrinkler distance [25]. Levenshtein distance is defined as the minimum edit operation required to transform one given string into another. Jaro Winkler algorithm is used to measure the distance between two strings by taking the concern about the character penalty. Cosine measures the angular separation between the strings using turf and IDF. Jaccard distance measures the overlap between the given strings. It also accounts for the reverse order matching of the two given strings.

## IV. DATASET

We borrow the dataset from [4], which is freely available on [26]. We extracted the data from the given website and found that profile is data is available in ".json" format. We converted the given file in the .csv format using R studio platform. Further, we extract two most important features of the Twitter i.e. screenname and name. Screenname attribute of Twitter is equivalent to the username on Facebook while the name attribute of Twitter is equivalent to full-name on Googleplus. Thereafter, we extract Facebook users' real-name. These two attributes are most often present in the network relationship of a user across different social networks. The description of the username based dataset is shown in table 1. We extract only the English name and username available in the given dataset. We carefully align all the data available and found that 9520 Screenname available for Twitter accounts while 13790 surnames available for Facebook. The name or fullname gave in the dataset is depicted in table 2. We extract the name available on Twitter and Googleplus. The total number of name obtained on Twitter and Googleplus are 8570 and 9151 respectively.

Table 1. Dataset: Twitter-GooglePlus-Facebook (Name/Fullname)

| S. No. | Social Network | Number of Users |
|---|---|---|
| 1. | Twitter | 8570 |
| 2. | Googleplus | 9151 |
| 3. | Facebook | 13118 |

## V. RESULT AND DISCUSSION

We match the corresponding features available across the social networks. The matching algorithm that we use in this paper are categorized into two types

### A. Character based distance measuring algorithm

In this form of algorithms, two given strings are matched character by character if characters on respective positions are same then the distance is zero otherwise weight of the distance measure is increased. The well known algorithms are Levenshtein distance (minimum edit operation required to transform one string into another), JaroWrinkler algorithm.

### B. Token based distance measuring algorithm

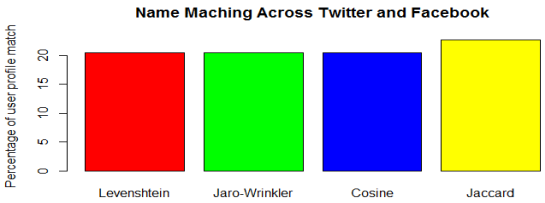
In this type of algorithms, the whole word of the given string is matched with the corresponding word in the other string. Such string strings do not check the words in lexicographic order. The well known distance measure algorithms are cosine, Jaccard, etc.In this paper, we use two characters based distance measure algorithm, namely; Levenshtein and JaroWrinkler. The other two are token based (Cosine and Jaccard). The result of our experiment corresponding to username and name matching is demonstrated in table 4 and 5 by using different distance measuring algorithms.

### C. Display-Name matching across the Twitter, Facebook and Googleplus

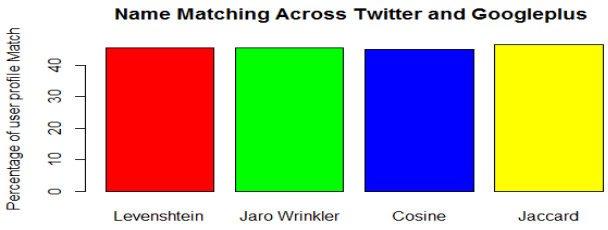
To find the user identity across these three social networks, we select the fullname which is the most consistent feature available across the site. To attribute consistency across the sited we exploit four different distance measuring algorithms, namely; Levenshtein distance, JaroWrinklerdistance,cosine distance and Jaccard distance measure. Corresponding results of each algorithm across the sites are shown in table 2. and its percentage wise representation in depicted in Fig.4. (a), (b) and (c). From the observation of fig. 7, we found that exact name similarity across Twitter and Facebook ranges from 20.42% to 22.65%, while Twitter and Googleplus range from 45.69% to 46.53 % and across Facebook and Google Plus ranges from 22 to 24 %.

Table 2. Results: Exact name matching across the networks

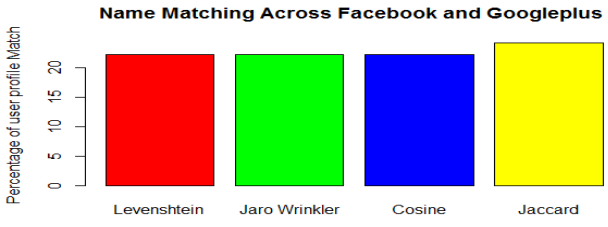
| S. No. | Network pairs | Levenshtein | JaroWrinkler | Cosine | Jaccard |
|---|---|---|---|---|---|
| 1. | Tw-Gp | 3916 | 3916 | 3927 | 3988 |
| 2. | Tw-Fb | 1864 | 1864 | 1869 | 2067 |
| 3. | Fb-Gp | 2038 | 2038 | 2046 | 2220 |




(a)




(b)




(c)

Fig. 4. Exact Name match across (a) Twitter-Facebook (b) Twitter-Googleplus (c) Facebook-Googleplus

From the observation of the results, we found that Jaccard distance consistently outperforms than the Levenshtein, Jaro Winkler and cosine measures. From the observation of Table 2 we, found that Around 20 to22 percent of the users have selected the same name across Facebook and Google plus social media service.

From the study of all the three results we observe that Twitter and Google Plus users have common goals join the social networks and they open to disclose their real-name. Facebook-Twitter and Facebook-Googleplus users are reluctant to share their real-name due to their privacy.

## VI. CONCLUSION AND FUTURE WORK

Social networking services are becoming more popular among the people due to the decreasing cost of the Internet and the popularity smart handheld devices (Mobile Phone and Tablet), Consequently, the strength of social media users is growing rapidly. In this paper, we investigated the user's selected name attribute across the three different social networks; Twitter Facebook and GooglePlus. From the observation of the experimental results we obtained that 20 to 22 percent users selected the same name across Twitter and Facebook, while the name similarity across Twitter and Google plus increases from 45 to 46 Percent. From the observation, we find that users' name similarity across Twitter-Facebook and GooglePlus-facebook pairs are in approximately in the same ratio. In the future, we will try to find the accuracy, Precision, recall, and F- measure by using machine learning algorithms like Multilayer perceptron, Random Forest and LibSVM, Etc.

## REFERENCES


1. https://www.smartinsights.com/sarocial-media-marketing/social-media-strategy/new-global-social-media-resech/ (last accessed 05/18/2019).
2. https://www.pewinternet.org/2018/03/01/social-media-use-in-2018/.(last accessed 05/18/2019).
3. L. Humphreys, P. Gill, B. Krishnamurthy “Privacy on Twitter: how much is too much? Privacy issues on Twitter,”The annual meeting of the international communication Association, Singapore, pp. 1-29, 2010.
4. J. Vosecky, D. Hong, and V. Y. Shen, User identification across multiple social networks. In First International Conference on Networked Digital Technologies, pages. 360-365. 2009.
5. A. Esfandyari, M. Zignani, S. Gaito, G. P. Rossi, “User identification across online social networks in practice:Pitfalls and solution,” Journal of Information Science, Vol. 44, no. 3, pages, 377-391,2018.
6. K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, User Identity Linkage across Online Social Networks: A Review. ACM SIGKDD Explorations Newsletter vol. 18, no. 2: pages 5-17, 2017.
7. A. Malhotra, L. Totti, W. J. Meira, P. Kumaraguru, and V. Almeida, Studying user footprints in different online social networks. In International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pages 1065-1070, 2012
8. M. Marti, and G. Varghese, "I seek you: searching and matching individuals in social networks," Proceedings of the eleventh international workshop on Web information and data management, ACM.67-75, 2009.
9. E. Raad, R. Chbeir, & 2010.
10. Z. Cheng ,J. Caverlee and K. Lee “You are Where you Tweet: a Content-Based Approach to Geo-locating Twitter Users’,” In Proceedings, of the 19th ACM International Conference on Information and Knowledge Management, Toronto, Canada, pp. 759–768, 2010.
11. O. Goga, H. Lei, S.H.K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira, "Exploiting innocuous activity for correlating users across sites," In Proceedings of the 22nd international conference on World Wide Web, 447-458, 2013.
12. Li, G. A. Wang, and H. Chen, "Identity matching using personal and social identity features," Information Systems Frontiers, vol. 13, no. 1, pages 101-113, 2011.
13. O. Goga, , D. Perito, H. Lei, R. Teixeira, and R. Sommer, "Large-scale correlation of accounts across social networks," University of California at Berkeley, Berkeley, California, Tech. Rep. TR-13-002, 2013.
14. O. Peled, M. Fire, L. Rokach, and Y.Elovici, "Entity matching in online social networks," In International Conference on Social Computing (SocialCom), pages 339-344, 2013P.
15. O. Goga, H. Lei, S.H.K. Parthasarathi, G. Friedland, R. Sommer, and R.Teixeira, Exploiting innocuous activity for correlating users across sites.In Proceedings of the 22nd international conference on World Wide Web, pages 447-458, 2013
16. A. Narayanan and V. Shmatikov, De-anonymizing social networks. In Proceedings Of the 30th IEEE Symposium on Security and Privacy, pages 173-187, 2009
17. S. Bartunov, A. Korshunov , S. Park, W, Ryu., and H. Lee, Joint link-attribute user identity resolution in online social networks. The 6th SNA-KDD Workshop, 2012
18. N. Korula, and S. Lattanzi, An efficient reconciliation algorithm for social networks.In proceedings of the VLDB Endowment, vol. 7, no. 5, pages377-388, 2014
19. K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, User Identity Linkage across Online Social Networks: A Review. ACM SIGKDD Explorations Newsletter vol. 18, no. 2: pages 5-17, 2017
20. S. Tan, Z. Guan, D. Cai, X. Qin, J. Bu, and C. Chen, Mapping users across networks by manifold alignment on hypergraph, In AAAI, vol. 14, pages 159-165. 2014
21. X. Zhou, X. Liang, H. Zhang, and Y. Ma, "Cross-platform identification of anonymous identical users in multiple social media networks," IEEE transactions on knowledge and data engineering, vol. 28, no. 2, pages.411-424, 2016.
22. X. Zhou, X. Liang, X. Du, and J. Zhao, "Structure based user identification across social networks," IEEE Transactions on Knowledge and Data Engineering, 30(6),1178-1191,2018.
23. Yongjun L.I., and Su, Z. A Comment on “Cross-Platform Identification of Anonymous Identical Users in Multiple Social Media Networks,” IEEE Transactions on Knowledge and Data Engineering, 30(7),1409-1410,2018.
24. P. Jain, P. Kumaraguru and A. Joshi, @ iseek'fb.Me': "Identifying users across multiple online social networks," In Proceedings of the 22nd international conference on World Wide Web ,pages 1259-1268, 2013.
25. W. Cohen, P. Ravikumar, and S.Fienberg, A comparison of string metrics for matching names and records. In Kdd workshop on data cleaning and object consolidation Vol. 3, pages. 73-78, 2003.
26. http://nptlab.di.unimi.it/?page_id=360. (last accessed 02/29/2019).


## AUTHORS PROFILE

**WaseemAhmad** received the B. Tech from VBS Purvanchal University, Jaunpur, UP, India in 2007 and M. Tech degree from MD University, Rohtak,Haryana,india in 2011. He is currently working toward the PhD degree in the Department of Computer Engineering, Aligarh Muslim University, India, and is an assistant professor at the Al-Falah University, Faridabas Haryana. His research interests include web mining and information retrieval.

**Rashid Ali**Rashid Ali obtained his B.Tech. andM.Tech. from A.M.U. Aligarh, India in 1999 and 2001 respectively. He obtained his PhD in Computer Engineering in February 2010 from A.M.U. Aligarh. His PhD work was on performance evaluation of Web Search Engines. He has authored about 125 papers in various International Journals and International conference proceedings. He has presented papers in many International conferences and has also chaired sessions in few International conferences. He has reviewed articles for some of the reputed International Journals and International conference proceedings. He has supervised 19 M.Tech Dissertation and three PhD Thesis. Currently, he is supervising five PhD candidates. His research interests include Web-Searching, Web-Mining, soft computing Techniques (Rough-Set, Artificial Neural Networks, fuzzy logic etc), Recommender Systems and Online Social Network Analysis.