Intelligent Automation And Soft Computing, 2020

Vol. 26, no. 1, 87–98
https://doi.org/10.31209/2019.100000130

Finding Temporal Influential Users in Social Media Using Association Rule Learning

Babar Shazad[1], Hikmat Ullah khan[2], Zahoor-ur-Rehman[1], Muhammad Farooq[2], Ahsan Mahmood[1], Irfan Mehmood[3], Seungmin Rho[3], Yunyoung Nam[4]

[1]Dept. of Computer Science COMSATS University Islamabad, Attock, Pakistan
[2]Dept. of Computer Science COMSATS University Islamabad, Wah Pakistan
[3]Department of Software, Sejong University, Seoul, Korea
[4]Department of Computer Science and Engineering, Soonchunhyang University, Asan 31538, Korea

ABSTRACT

The social media has become an integral part of our daily life. The social web users interact and thus influence each other influence in many aspects. Blogging is one of the most important features of the social web. The bloggers share their views, opinions and ideas in the form of blog posts. The influential bloggers are the leading bloggers who influence the other bloggers in their online communities. The relevant literature presents several studies related to identification of top influential bloggers in last decade. The research domain of finding the top influential bloggers mainly focuses on feature centric models. This research study proposes to apply association rule learning for finding the temporal influential bloggers. The widely used Apriori algorithm is applied using Oracle data miner to find the frequent pattern of bloggers having blog activities together and then we find who influences others based on the rules learned from the association rule mining. The use of standard evaluation measures such as accuracy, precision and F1 score verifies the results. This research study uses the standard dataset of TechCrunch which is a real world blog. The results confirm that the association rule mining can produce rules which help to find the temporal influential bloggers in the blogosphere who are consistent on regular basis. The proposed method achieved accuracy as high as 98% for confidence level of 90%. The identification of the top influential bloggers has enormous applications in advertising, online marketing, e-commerce, promoting a political agenda, influencing elections and affect the government policies.

KEY WORDS: Social Media, Association Rule Mining, Influential Users, Bloggers, Apriori

1 INTRODUCTION

THE World Wide Web (WWW) has transformed into the Social web due to content generation facility as the users are not only the content consumers, but also the content generators. The Web 2.0 has attained huge attention, in recent years, as it allows its users to share their views, communicate to each other and also it has expanded the scope of interaction on social issues or topics (Gerlitz & Helmond, 2013; Probst, Grosswiele, & Pfleger, 2013). The social web has created opportunities for researchers to analyze social activities such as e-business(Singer, 2012), online shopping, marketing, blogs and electronic commerce(Van Dijck & Poell, 2013).

In recent times, the more focus regarding the social web is its ability to provide social relations among people all over the world(Appelquist et al., 2010). The Social web sites are designed in a way to foster and support social interaction, and conduct online gaming, shopping, chatting, discussions and seeking guidelines in education (Victor & Britto, 2014). The social aspect of interaction has made it possible to help people to interact with people having similar views, liking and emotions, the concept is known as homo philly (Weber, 2009). Due to people discussion

CONTACT Irfan Mehmood, Yunyoung Nam irfanmehmood@ieee.org; irfan@sejong.ac.kr, ynam@sch.ac.kr

in various blogs and forums, a massive amount of data is being generated on various topics that becomes available for others to learn and comment. The online social communities or networking sites such as Facebook, Twitter, YouTube, Google, Instagram, LinkedIn, Pin interest, Tumblr and Myspace enable the organizations and people to contact each with friendly names and share their views and express their emotions. Nowadays, billions of the social web users are using such websites to connect their friends and even discover some new friends. The number of users is increasing exponentially round the clock.

1.1 Blogosphere

The use of the participating Web has caused more online media channels that converted the previous information users to the current information producers. Wikis, media sharing, blogs, cooperative tagging, and other similar services are main examples. A blog is an assemblage of entries made by different personalities presented in opposite sequential order and such entries are called blog posts, Blog spots can characteristically include multimedia, text and miscellaneous associations to other blog posts web pages or simply blogs. For individuals to express share, debate and communicate through mass web master, blogging is considered as popular source for them. The blogosphere, which is a virtual universe comprises all blogs. Bloggers, also called the blog writers, generally shape up their distinct curiosity groups where different opinions are shared, opinions are expressed and general discussion on different ideas and suggestion is performed interactively(Bruns & Jacobs, 2006). The blogosphere leads to provide favorable and helpful platform in creating virtual communities with specified interests. This thing is taken under observation that blogs are sources of creating new relationships and maintaining the existing one nicely.

According to Vasanthakumar et al (Vasanthakumar et al., 2016), rather than old traditional and classical advertising approach, 83% of people give preference to the family friend or expert for consulting before experiencing new restaurant 71% for purchasing prescription drug and for movie watching it is 61% of such people. It is to be concluded that people tend to make a generic discussion with family members and make a decision on the basis of practices and recommendations. Among such experiences and suggestions those persons having valuable, profitable opinions and experiences are considered as influential. The folks whose participations and opinions, are hunted after are appropriately called as "*the influential*". In blogosphere, there exists two communities naming physical and virtual communities. They are interested in finding out whether the influential, as described above, exist in virtual community or not. In a result of successive finding, who they are and how to locate them(Agarwal, Liu, Tang, & Yu, 2008).

There are two major types of Blogs: individual and community blogs. In an individual blog, the host person is responsible for initiating and leading discussion and automatically taken as influential blogger. While in case of community blog, there is requirement of more than one individual who avails equal opportunities in discussion participation, which consequently creates a chance for them to emerge(Bruns, 2008).

This is the reason for which they focused on blogs. Therefore, blogs refer to community blogs.

1.2 Application of Finding Influential Bloggers

In order to create creative business plans, fabricating political strategies, examining societal and social issues, and prompt numerous stimulating application, identification of influential bloggers can be used. As the bloggers can be related in virtual societies anytime and anywhere. For instance, the influential users are repeatedly market-movers. They are able to impact the choices of the people during the product purchase. Therefore, recognizing them can enable organizations to better comprehend the main concerns and new patterns about items fascinating to them, and sagaciously influence them with extra information and counsel to transform them into informal representatives. As stated in (Jianqiang, Xiaolin, & Feng, 2017), approximately 64% marketing corporations have approved this sensation and are ever-changing their emphasis towards blog advertising. Influential bloggers as a spokesperson of virtual communities, they could also influence opinions in political movements, elections, and disturb the policies of states (Utz & Jankowski, 2016). Beat on the influential can help recognize the fluctuating interests, forecast possible difficulties and likely gains.

1.3 Research Contributions

The major contributions of this article include: firstly, potentials to identify the influential bloggers; secondly, time based performance of top influential bloggers is computed to find out the temporal based influential bloggers; finally, the importance of Association rule learning is find out in the domain of influential bloggers. After finding the top influential bloggers, standard evaluation measure techniques are used to find accuracy of the work along with the proper comparison with the past works by the researchers on the same topic.

The rest of the paper is distributed as follows: Section 2 discusses related work; Section 3 describes the detailed research methodology that is followed in this research. Results are discussed in Section 4 before concluding the paper in Section 5.

2 RELATED WORK

SOCIAL media and blog analysis are popular fields in the research of social network and data mining. Several approaches to measuring influence are centered on the graph (Kwak, Lee, Park, & Moon, 2010). Many methodologies have been used in this manner to find the influential bloggers from the social media and blogs. Mostly the work is divided into two main categories. First of all, there are some authors who work in this field to find the influential bloggers and users from current social media data. On the other hand, there are some researchers who have worked on to predict the social media influential users using the present data. The connection prediction is also one of the fields where researchers have worked a lot (Utz & Jankowski, 2016). UIRank (Use Influence Rank) algorithm is used to identify the influence of users through interaction information and relationship among bloggers in micro-blogs (Jianqiang et al.2017), identifying the influential user weight based on a new metric with respect to time intervals (Mahmoudi, Yaakub, & Bakar, 2018), the prediction from the prediction of evolution (Saganowski et al., 2015), and many more works in the field. Many researchers have used these tools to perform the influential user's identification from the social media and blogs. These approaches give satisfactory results in some cases.

According to Gliva. B et al (Gliwa, Koźlak, Zygmunt, & Demazeau, 2016), these days' social media are present in different forms and social media plays very important role in the society as well as for the individuals. According to the authors, the behaviors of the users usually depend upon the activities of the social media and how they react to the activities of the social media is very important. Moreover, in order to completely understand the activity and working of social media for each of the user the authors' agent based approach for the analysis of social media. And they explained how other users of the social media influence other users and how different kinds of policies and political behavior affect the user of social media. Vasanthakumar g *et al* (Vasanthakumar et al., 2016), proposed an online social media network profiling approach to find the influential bloggers on the social media. They proposed their solution on the basis of top most influential bloggers and performed the content analysis on the system. Vasanthakumar g *et al* (GU, KC, BR, Shenoy, & KR, 2017), proposed another system called PTMIBSS that is based on profiling system. The system was used to find the influential bloggers using synonym substation approach. The authors compare their results with the previous researches on the same topic to find the effectiveness of their work and measured the performance of their system using the standard evaluation system.

Kao JL et al (Kao, Huang, & Sandnes, 2016), work on a new field in which they find the time dependent influential users on social media. The proposed a app that find the influential users in a particular field to find the influential users. according to them the importance of finding time dependent influential users is much more important because with time new users come in the field and with time old users may not be as much influential as they were in the past. this approach was very good and they find different kinds of patterns to recognize the influential users on the social media platform, more specifically on the Facebook.

2.1 Feature based Models

In feature based model the authors use the features of the bloggers to find the influential users on through different approaches. In this model, the features of the users are used to determine the influence of the users and bloggers. A model was proposed to find the posts of blogs Blog post using the activity of the blog posts. Mainly this model uses the length of the blog posts and used it to determine the influential bloggers. Their final results show their aim was novel (Khan, Daud, & Malik, 2015). Another similar model was proposed to find the influential bloggers using the activity of the blogs using eloquence measure of the blogs. Their approach was based on the previous approach and their results were much better than the previous results. They performed clustering on the data to get the clustered results of the bloggers (Khan et al., 2015).

Using TUAW dataset, MEIBI&MEIBIX was proposed to discover the influential bloggers using different aspects of the bloggers. The authors first time used the time as a factor in influential bloggers finding as according to them time is a very important factor in influential users as user influence may change with time.

The contents of blog posts were calculated using the comments of users and out links of the bloggers on that posts. This model was based on HIndex approach and it calculate the activity of the user, and other information related to the users and their quality of posts (Gliwa & Zygmunt, 2015). Researchers find the relation of responses of the users comments when the users post those comments by using post influence methods and determined the influence of the users in those posts where they didn't post any comments (Agarwal, Liu, Tang, & Philip, 2012). Other similar approaches include bloggers popularity measure (Khan & Daud, 2017) using the number of comments and bloggers activity measure (Ishfaq, Khan, & Iqbal, 2016) using the sentimental features of the bloggers. According to Khan HU, blogs use as platform of effective announcement to segment comments or opinions about products, occasions and community matters. According to the author, similar other social web actions, blogging activities extent to a huge amount of individuals. Users influence others in different kind of means, such as purchasing an invention, taking an explicit political or social opinion

or starting new activity. Although there are many kinds of models that are used to find the influential bloggers, the authors considered a new model to find the influential blogger. The proposed model finds the influential bloggers and is called Popularity and Productivity model that is constructed. They discussed the role of different features that they proposed. They used real-world blogs data and performed their analysis on that data. They measured their performance using standard performance evaluation metrics. Their results show they were able to find the influential bloggers in effective way. The reward of recognizing influential bloggers has apparently been so important. On social platforms, like a forum, Wikipedia's etc. are used for virtual communication and expressing opinion about something, products, and experience, a user can influence other users to buy some specific product, have a specific view about some issue and so finding these is surely mean a lot in the online communities (Khan & Daud, 2017), (Khan et al., 2017). If we are capable to find the influential users and bloggers these influential users can be used in different manners and fields to achieve different kinds of tasks. Influential users sometime behave as market setters who work with the other users to set the trends in the market. People take their suggestions and recommendation serious and purchases the products recommended by those users.

Agarwal N et.al.(Khan et al., 2015), present their work and they introduced the tasks of identifying influential bloggers by proposing an initial model to compute and determine influential bloggers. Another work was proposed consist of two separately proposed methods including MEIBI and MEIBIX that consider blogger post inlinks and calculation of blog score respectively. Later the authors proposed another work in the same filed they proposed a framework to identify productive and influential bloggers they used different metrics to identify the influential bloggers by using data gathered from a real-world blog site (Erlandsson, Bródka, Borg, & Johnson, 2016).

Khan et al., (Khan & Daud, 2017) proposed research work by introducing a Metrics for Identification of Influential Bloggers, in which they have formulated various features of bloggers they also calculated bloggers popularity and activity and have used Blog Rank for this purpose. Another model was proposed and called Popularity and Productivity Model (PPM), that was built to discover the top influential users it contains of various features each of the features was based on a model they used the standard measure to evaluate their work. Their model was able to find the influential bloggers ineffective way They compare their results with the previous researches on the same topic to find the effectiveness of their work and measured the performance of their system using the standard evaluation system (Şen, Wigand, Agarwal, Tokdemir, & Kasprzyk, 2016). Khan et al, (Zhang, Huang, He, & Ren, 2017) has worked on the same topic and writes a detailed serve of different past works in this field according to the existing models are categorized into a feature and network-based categories. In the network model, they consider the graph of social media network structure. They also introduced the features with respect to novel and dataset used by each of the authors. They also discussed the application of the relevant literature.

According to Sen F et al (Moh & Shola, 2013), recognizing influential bloggers is a thriving method in mining knowledge from a network. According to them there have been many studies which suggests that calculating the identifiable influential bloggers are very helpful in different fields. In a community, an influential blogger influence people across different categories and backgrounds. The authors developed an approach called focal structure analysis to extract those influential users and they called them focal structures, in social network. According to the authors their offer was unique and such work in finding influential users was not done before. There are many other approaches that have been used in finding influential bloggers and users on the social media networks including Facebook, Twitter, blogs and other platforms. H-Index Family (Khan & Daud, 2017) to study the bloggers using H-Index, MIIB (Khan &Daud, 2017) to identify the influential bloggers based on popularity , and MIBSF (Ishfaq et al., 2016) to analyses the sentiment features and find influential bloggers.

2.2 Network Based Models

Other than using the features of the users to find their influence, different researchers used the network of the users with other users, organizations, companies, etc. to find the influence of those users on the social media and blogs. This approach was also widely used by different researchers and had its own effects on the research in the field. Researchers used pattern discovery from network to find the leaders in the influential networks. The created influential bloggers graphs and find the influence using the movement of the graph. Statistical approaches including post related feature and relationship of the users with bloggers was calculated using CR Algorithm. A Vector based model (Abnar, Takaffoli, Rabbany, & Zaïane, 2014) was proposed by researchers in which the authors used PageRank as their baseline and find the network of the comments on the basis of PageRank. Authors also focuses on discussion related network and performed the analysis of the discussion using PageRank and baseline and find the clustering other users and ranked them accordingly (Takaffoli, Fagnan, Sangi, & Zaïane, 2011). The authors find the interaction of the users in a particular time period and its influence and influence when the interaction changes from one interval to another and when the interaction changes from one user to another.

The researchers also focus other mining features including, different centrality model along with PageRank algorithm and performed the analysis with the previous approaches. Their approach was novel and they were able to find the influential users with different respects and the interaction between users and bloggers in a much better way (Agarwal et al., 2012). LT Model (Kayes, Qian, Skvoretz, & Iamnitchi, 2012), determines the blogs and the influence of the users according to a particular threshold. They ranked the influential users based on a newly proposed method. They tracked the influence of the users alongside different networks and blogs. Ciav & Chen proposed Interest vector (Abnar et al., 2014), an approach to find the bloggers influence with the help of PageRank algorithm in a specific domain.

3 RESEARCH METHODOLOGY

IN a network, the influence of a node (user) on another node can be represented in social network analysis using different centrality metrics. In social media, users follow each other and that it is promising to identify influential bloggers and forecast the involvement of the user. For instance, if users X, Y, Z and A share mutual benefits, there is a chance that if X, Y, and Z previously have liked on a post, A likewise like it. This shows that the users can have a big impact on other users or the influential levels can mean some deduction from other users calculate the influence of users, bloggers, with different bloggers, users, etc. we have different types of algorithms that have been used in this field. As we have discussed earlier in research work the association rule learning has a good output in the calculation of influential users from social websites so we use association rule learning in our research.

The Figure 1 presents proposed framework for the research. According to this framework, in the first step, the dataset is extracted from bloggers data and preprocessing is applied. After data is properly stored in the dataset, the association rule learning is applied on the dataset to extract the frequent itemsets and with the help of different metrics in terms of support, confidence and lift, the association rules are generated that presents the top influential bloggers.

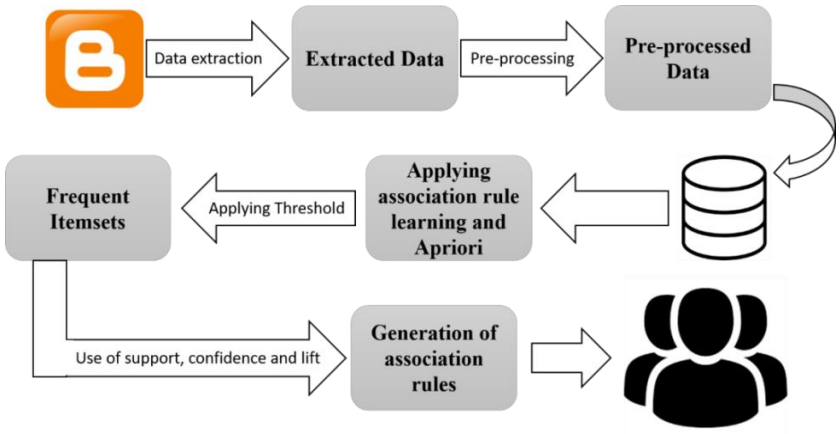

Figure 1 **Proposed Research framework to identify influential bloggers**

3.1 Dataset

TechCrunch, a real blog, data has been used in earlier relevant studies[35]. The statistical facts of the data are given in Table 1. The TechCrunch Dataset contains data about blogs including, bloggers, their posts, users and their comments. The dataset contains the data in the time interval of 2005-2010. The dataset is the most authenticate dataset of bloggers that has been used by many researchers in the past for finding influential bloggers and working on other data mining tasks. In this research, the TechCrunch dataset is used so the authenticity of the work can be compared with the previous works in the same field. Moreover, no latest dataset is available on the same topic and if new data is crawled from the web, it is not possible to compare the work with other works in this way. we divided the dataset into 3 parts and 2-2 years of dataset. In the earlier approaches, when we checked the results against one year parts, we were not getting good results therefore we made parts of 2-2 years and results are far improved.

3.2 Association Rule Learning

Association rule learning is used to find repeated item set by calculating the frequency of items. Association rule learning uses two kind of criteria system, explicitly, confidence and support. Support designates regularity of such items, while confidence shows how numerous times those procedures in the entire data set are correct. While working with association rule learning we have to calculate the evaluation Matric. An example can be given as “Ninety-Eight percent people who buying tires and auto accessories also acquire automotive services done”. We are going to apply it in the field of social

media where we can model the data as follow: Items relate to users on Social media or bloggers and transactions relate to posts.

Table 1. Data Statistics

Parameter	Value
Posts	18,994
Bloggers	1,64,912
Comments	7,46,561
Posts in 80% Data	15,195
Bloggers in 80% Data	1,41,630
Posts in 20% Data	3,799
Total Bloggers with 20% Data	34,522

3.3 Apriori Algorithm

Association rule mining provides us the frequent item sets and rules can be learnt from these item sets. Let us take a look at the formal definition of the problem of association rules given by Agarwal et, al. [54]. Let U = {u1, u2, u3, …, un} be a set of n attributes. These attributes are called items and P = {p1, p2, p3, ……, pn} are the set of transactions. Each of the transaction, p_n in U has as unique ID. A rule can be defined as an implication, A⇨B where A and B are subsets of U (A, B⊂ U), and they have no element in common, i.e., A∩B.

3.4 Evaluation Metric

We find Temporal based user influence using the sliding window approach. Using this approach be feasible for us as this has been a good approach mentioned by different researchers. We can understand the association rule learning with the help of different Evaluation metrics. It consists of multiple measures like Support, Confidence, Lift and Conviction.

3.5 Support

This calculates the popularity of an item in an item-set, as measured by the proportion of transactions (posts) in which an item set appears. That is considered, distributing the frequency of assumed item set, {I}, through over-all transactions (posts) in the dataset, {W}, otherwise the number existences of {X, Y} divided by the n items in {W}. As presented in Equation (1):

$$\text{Support}(\{X,Y\}) = \frac{\{X,Y\}}{W} \quad (1)$$

3.6 Confidence

This calculates the likelihood of one item with another item, expressed as {X -> Y}. Confidence can be calculated as presented in Equation (2). Approximately that {X, Y, Z} contributes on five common posts and {X, Y} contribute in 10 posts in total. This leads to 5/10 = 0.5 or the confidence that Z will participate in a post where X and Y before now are active is 50%.

$$\text{Confidence}(\{X,Y\}\rightarrow Z) = \frac{Support(X,Y,Z)}{Support(X,Y)} \quad (2)$$

3.7 Lift

Lift can be explained as the ratio of Confidence to Expected Confidence. In Equation (3), if lift is equal to 1, it indicates that the rules and for each item the values are independent.

$$\text{Lift}(\{X,Y\}\rightarrow Z) = \frac{Support(X,Y,Z)}{Support(X,Y) * support(Z)} \quad (3)$$

4 RESULT AND DISCUSSIONS

TO calculate the performance of proposed system the accuracy, precision, recall and F1 while their formula can be given as shown in Equation. 4, 5, 6 and 7 as follows:

$$\text{Accuracy} = TP + TN/(TP + TN + FP + FN) \quad (4)$$

$$\text{Precision} = TP/(TP + FP) \quad (5)$$

$$\text{Recall} = TP/(TP + FN) \quad (6)$$

$$F1 = 2 * \text{Precision} * \text{Recall}/(\text{Precision} + \text{Recall}) \quad (7)$$

While TP or True Positive is a condition when the model predict that the user is active and the user is active. The FP or False Positive is a condition when the model predicts that the user is active and it is not active. The TN or True Negative is a condition when the model predicts that the user is not active and user is not active the false negative or FN is a condition when the model predict that the user is not active and it is active. The table 2 shows an example of all these classes. Capital letters indicates user P1-4 corresponds to different posts.

Table 2. Example of true positive and false negative.

Example Rule: {W, X, Y}→ Z		
P1	{W, X, Y, Z} →	true positive
P2	{W, X, Y} →	false positive
P3	{Q, R, S} →	true negative
P4	{Z, U,} →	false negative

Figure 1 shows the result of accuracy, precision and F1 measure. When the confidence is 90% the accuracy of the proposed approach is 98% while the confidence is 80 the accuracy is 80%.

It can be seen in the results that approach is feasible and model achieve good results accuracy while finding the confidence that user will comment on the blog. Table 3 shows a list of top 10 influential bloggers find through proposed model with number of posts and comments.

After finding the overall influential bloggers of the whole dataset, the model finds temporal influential bloggers. For this purpose, whole dataset is divided into 3 parts, each of 2 years as data of 2005-2006, data

of 2007-2008 and data of 2009-2010. This way, it becomes easy to explore the change of the dataset and how the influence of the bloggers changes with time. In order to find the influential bloggers with respect to time, the results of each of the part of the dataset are discussed separately.

Table 3. Top 10 Influential Bloggers Overall in the dataset

S. No	Bloggers	No Of Posts	No Of Comments
1	Michael Arrington	4846	6750
2	Erick Schonfeld	2670	918
3	Nick Gonzalez	1440	1177
4	MG Siegler	1423	2692
5	Jason Kincaid	1810	656
6	John Biggs	988	533
7	Duncan Riley	1270	1051
8	Robin Wauters	1351	1335
9	Mark Hendrickson	566	341
10	Marshall Kirkpatrick	447	515

4.1 Dataset from 2005-2006

First of all, the influential users of 2005-2006 dataset are explored. With these attributes of support, confidence, lift and rule length dataset related to 2005-2006 is processed. During the experiments, the results of the influential bloggers against a different set of rule and computed the consequent and antecedent's names are checked

In the Table 4, it can be seen that the model achieves good confidence levels against each of the influential bloggers. As the data is huge in volume, the support levels are not good, but considering the huge amount of the dataset, the confidence values are very good.

4.2 Time period of 2007-2008

After exploring the dataset of 2005-2006, the influential users of 2007-2008 dataset are explored. This way the change from the previous time period the year 2007-2008 is explored. Considering the support, confidence, lift and rule length the dataset of 2007-2008 is processed. During evaluation process, influential bloggers' result is assessed and checked alongside different length of rules and extracted names of consequent and antecedent part of rules. Values of support, confidence and lift are calculated for each influential user and recorded the top bloggers on the basis of this computation and rules. Table 5 shows top bloggers by confidence.

In the Table 5, it is shown that model achieved high confidence levels against respective influential bloggers. Perde is considered as top blogger with enhanced confidence as compared to other bloggers. Level of support is not effective because of huge volume of data On the other hand, huge data set amount lead to the better confidence values.

4.3 Time period 2009-2010

After exploring the dataset of 2005-2006 and 2007-2008, the influential users of 2009-2010 dataset are explored. By applying features of support, confidence, lift and rule length, the model process the dataset associated to 2009-2010. While experimenting, the results of influential bloggers against a different length rules and computed the consequent and antecedent's names are monitored. The value of support, confidence and lift against each of the rule is calculated and top influential bloggers related to each of the rules are find out. Table 6 shows top bloggers by confidence.

In the Table 6, it is evident that against each of the influential bloggers a good confidence level is achieved. As of huge volume of the data, the support levels are not good, but it is contrary to confidence values.

4.4 Comparison with baseline methods

In Table 7 performed comparison of the proposed model with the baseline models so they result can be evaluated. For instance, the proposed model is compared with baseline methods MISBF(Ishfaq et al., 2016) and H-Index model (Bui, Nguyen, & Ha, 2014). The results show that the similarity between the proposed model and MISBF is 90%. Similarly, the similarity with the H-Index model and proposed model is 80%. Likewise, the proposed model performs better and finds the influential bloggers in efficient way with respect to the time and is able to consider the time as an important factor. The comparison shows that the proposed model is efficient and conclusively finding the top influential bloggers from the dataset.

4.5 Findings

During the experimental setup, the model explores the influential bloggers through the help of association rule learning and apriori algorithm. The model identified the influential bloggers, i.e. the bloggers who influence other users on the basis of their posts and other people's place comments on their posts. During all the experiments, the standard dataset and evaluated the performance of the projected method using standard measures. Influential bloggers with respect to different categories are found. Initially, influential users are computed through the help of ARL from the overall dataset of 2005-2010-time period. This way the overall influential bloggers are calculated. After finding the influential bloggers from overall dataset, in order to show the evidence of change of influence of the users from different time perspective, the whole dataset is divided into 3 parts. The dataset is divided into the time period of 2005-2006, 2007-2008 and 2009-2010. In case of support, support levels are between 0.03 to 0.13 due to the huge amount of posts during the experimental setup. Move rover, as the users make so many posts during the particular time period, it is not possible that each

of the user is always following the other user on each of their posts. Support, confidence, and lift among all the time period is computed, this way the results of the most influential bloggers are computed from different time perspective and computed the difference between all results. For all the time periods, we computed the most influential bloggers by using the values of support, confidence and lift. The values of support, confidence and lift of all these influential bloggers are computed. During our experiments, the influence of the users is changing with time and if a user influences another user at a particular time period or in a particular scenario it is not compulsory that the influential blogger carries out to influence the other user in the overall time period. Therefore, when the results of influential bloggers are computed, there is a big difference in the results of different time period. According to our experiments, 2 years are enough for a user to have his/ her influence.

When the model computed the results of 2005-2006-time period, according to the confidence, the model finds John, David Mackey, Ray, Startups.in, Michael Arrington as the most influential users, however, when model computed the results of 2007-2008-time period with respect to the confidence, the system computed the top influential bloggers as: perde, Michael Arrington, Michael Arrington, Mike, and Perde. Now it is clearly showing the difference in influence with respect to time. Finally, when the model computed the influence of bloggers of 2009-2010 with respect to the confidence, John, Mike, Chris, Chris, and Alex. Now the results clearly show that time does matter when it comes to calculate the influence of the users with respect to the time.

As the model computed the results of influential bloggers in different time periods, the statistics show that the confidence level of different time period varies from 0.5 to 1.0. This shows that our approach is efficient and is able to find the influential users with good confidence level. The average confidence level of 2005-2006 is 0.94 while it comes to a 0.62 to 0.63 during the time period of 2007-2008 and 2009-2010 that is critical considering the nature of the data. The model computed the influential users with respect to different levels of rules of association rule learning. The model chooses the length of rules from 1 to 4 that means the number of users who are influenced by the other users were between 1 and 4. Although rules of length 1 to 4 are computed, the length of average rule 3.4, 2.1 and 2.9 for the year 2005-2006, 2007-2008, 2009 and 2010 respectively. This shows that the best results gather was during the year 2005-2006 because the confidence level at that time period was good, the support was good and the overall rules were of high length. The empirical analysis show that time is an important factor while calculating the influence of the users because due to the difference in time the users' influence changes. This shows that a user may have higher influence of lower influence, but it is possible that their influence varies with time. Now this can help companies and organizations finding the right influenced person at the right time that they can use to take specific decisions or aim to change the course of their decision.

As the blog data is different for different period of time, the influential bloggers are varying. The results of our research suggests that the variance in the influence of a blogger may change with time due to their quality and interest in the platform. It also highly depends upon their activity. Therefore, the results are changing with time as presented in the Table 4, Table 5 and Table 6.

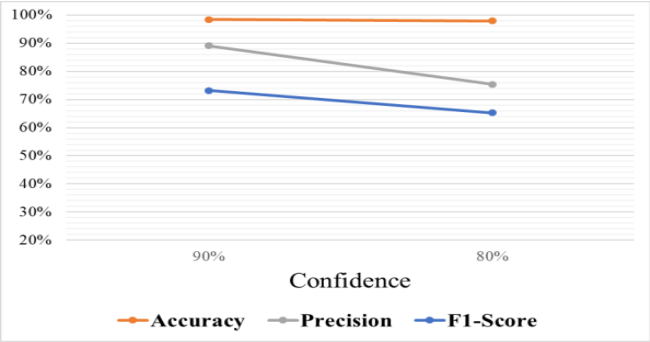

Figure 2. **Results showing of Accuracy, Precision and F1-Score of different levels of confidence**

Table 4. **Top blogger by confidence 2005_2006 data**

antecedent	consequent	support	confidence	Lift
Nick Gonzalez	perde	0.007	1.0	3.177
Matt, perde	Michael Arrington	0.006	1.0	3.177
perde, Mike	Michael Arrington	0.006	1.0	3.177
Nick Gonzalez, Sotek	Mike	0.004	1.0	3.177
Michael Arrington, James	perde	0.004	1.0	3.177

Table 5. Top Rules of 2007-08 by confidence

antecedent	consequent	support	confidence	lift
Michael Arrington	John	0.023	1.0	1.718
Michael Arrington, Sean	David Mackey	0.022	1.0	1.718
Michael Arrington	Ray	0.018	1.0	1.718
Michael Arrington	Startups.in	0.017	1.0	1.718
Joe, Mike	Michael Arrington	0.017	1.0	1.718

Table 6. Top Rules of 2009-10 confidence wise

Antecedent	consequent	support	confidence	lift
Mark,Andrew,Brian	John	0.004	1.0	6.413
Jeff,John,Daniel,Chris	Mike	0.004	1.0	6.616
Alex,Andrew,Jack	Chris	0.004	1.0	7.242
Eric,Michael,Paul	Chris	0.003	1.0	7.242
Tom,Marc,Brian	Alex	0.003	1.0	8.460

Table 7. Comparison with baseline methods

S. No	OUR APPROCH	MISBF	H INDEX
1	Michael Arrington	Michael Arrington	Michael Arrington
2	Erick Schonfeld	John Biggs	Erick Schonfeld
3	Nick Gonzalez	Erick Schonfeld	MG Siegler
4	MG Siegler	Jason Kincaid	Duncan Riley
5	Jason Kincaid	Duncan Riley	Jason Kincaid
6	John Biggs	Robin Wauters	Mark Hendrickson
7	Duncan Riley	MG Siegler	Robin Wauters
8	Robin Wauters	Mark Hendrickson	Leena Rao
9	Mark Hendrickson	Nick Gonzalez	Marshall Kirkpatrick
10	Marshall Kirkpatrick	Leena Rao	Guest Author

5 CONCLUSION

TO explore the influential bloggers using apriori, a well-known association rule mining algorithm. We use TechCrunch bloggers dataset for this purpose. The proposed methodology computed the influential bloggers, on the basis of their posts and other people posts upon their comments or posts. During the empirical analysis, we gathered standard dataset and evaluated the performance of the system through standard measures. The proposed methodology computed the influential users from different perspective and time periods. We computed influential users of overall time period, 2005-2006, 2007-2008 and 2009-2010. In order to show the evidence of change of influence of the users from different time perspective, we divide the whole dataset into three parts. During experiments, we explore the influential users through three diverse approaches: by applying confidence, support and lift. The final results show that the proposed approach is efficient and effective as it is capable to get good confidence level with high number of rules showing that there are many influential users in the blogosphere. The potential future work include applying the concept of association rule learning and apriori algorithm on the dataset of Facebook and Twitter to find out the influential Facebook and Twitter users. As Facebook and Twitter are active social sites, finding influential bloggers on those social sites can have deep social, financial and political impacts.

6 ACKNOWLEDGMENT

THIS research was funded by the Soonchunhyang University Research Fund and the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2019-2014-1-00720) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation)

7 REFERENCES

Abnar, A., Takaffoli, M., Rabbany, R., & Zaïane, O. R. (2014). SSRM: Structural social role mining for dynamic social networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on* (pp. 289–296). IEEE.

Agarwal, N., Liu, H., Tang, L., & Philip, S. Y. (2012). Modeling blogger influence in a community. *Social Network Analysis and Mining*, *2*(2), 139–162.

Agarwal, N., Liu, H., Tang, L., & Yu, P. S. (2008). Identifying the influential bloggers in a community. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 207–218). ACM.

Akritidis, L., Katsaros, D., & Bozanis, P. (2011). Identifying the productive and influential bloggers

in a community. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *41*(5), 759–764.

Appelquist, D., Brickley, D., Carvahlo, M., Iannella, R., Passant, A., Perey, C., & Story, H. (2010). A standards-based, open and privacy-aware social web. *W3C Incubator Group Report*, *6*.

Bruns, A. (2008). *Blogs, Wikipedia, Second Life, and beyond: From production to produsage*. Peter Lang.

Bruns, A., & Jacobs, J. (2006). Uses of blogs.

Bui, D.-L., Nguyen, T.-T., & Ha, Q.-T. (2014). Measuring the influence of bloggers in their community based on the H-index family. In *Advanced Computational Methods for Knowledge Engineering* (pp. 313–324). Springer.

Erlandsson, F., Bródka, P., Borg, A., & Johnson, H. (2016). Finding Influential Users in Social Media Using Association Rule Learning. *Entropy*, *18*(5), 164. https://doi.org/10.3390/e18050164

Gerlitz, C., & Helmond, A. (2013). The like economy: Social buttons and the data-intensive web. *New Media & Society*, *15*(8), 1348–1365.

Gill, K. E. (2004). How can we measure the influence of the blogosphere. In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*.

Gliwa, B., Koźlak, J., Zygmunt, A., & Demazeau, Y. (2016). Combining agent-based and social network analysis approaches to recognition of role influence in social media. In *Advances in Practical Applications of Scalable Multi-agent Systems. The PAAMS Collection* (pp. 109–120). Springer.

Gliwa, B., & Zygmunt, A. (2015). Finding influential bloggers. *ArXiv Preprint ArXiv:1505.06926*.

GU, V., KC, V. R., BR, A. R., Shenoy, P. D., & KR, V. (2017). PTMIBSS: PROFILING TOP MOST INFLUENTIAL BLOGGER USING SYNONYM SUBSTITUTION APPROACH. *ICTACT Journal on Soft Computing*, *7*(2).

Ishfaq, U., Khan, H. U., & Iqbal, K. (2016). Modeling to find the top bloggers using sentiment features. In *Computing, Electronic and Electrical Engineering (ICE Cube), 2016 International Conference on* (pp. 227–233). IEEE.

Jianqiang, Z., Xiaolin, G., & Feng, T. (2017). A New Method of Identifying Influential Users in the Micro-Blog Networks. *IEEE Access*, *5*, 3008–3015.

Kao, L.-J., Huang, Y.-P., & Sandnes, F. E. (2016). Mining time-dependent influential users in Facebook fans group. In *Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on* (pp. 000718–000723). IEEE.

Kayes, I., Qian, X., Skvoretz, J., & Iamnitchi, A. (2012). How influential are you: detecting influential bloggers in a blogging community. In *International Conference on Social Informatics* (pp. 29–42). Springer.

Khan, H. U., & Daud, A. (2017). Finding the top influential bloggers based on productivity and popularity features. *New Review of Hypermedia and Multimedia*, *23*(3), 189–206.

Khan, H. U., Daud, A., Ishfaq, U., Amjad, T., Aljohani, N., Abbasi, R. A., & Alowibdi, J. S. (2017). Modelling to identify influential bloggers in the blogosphere: A survey. *Computers in Human Behavior*, *68*, 64–82.

Khan, H. U., Daud, A., & Malik, T. A. (2015). MIIB: A Metric to Identify Top Influential Bloggers in a Community. *PLOS ONE*, *10*(9), e0138359. https://doi.org/10.1371/journal.pone.0138359

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web* (pp. 591–600). ACM.

Mahmoudi, A., Yaakub, M. R., & Bakar, A. A. (2018). New time-based model to identify the influential users in online social networks. *Data Technologies and Applications*. https://doi.org/10.1108/DTA-08-2017-0056

Moh, T.-S., & Shola, S. P. (2013). New factors for identifying influential bloggers. In *Big Data, 2013 IEEE International Conference on* (pp. 18–27). IEEE.

Probst, F., Grosswiele, L., & Pfleger, R. (2013). Who will lead and who will follow: Identifying Influential Users in Online Social Networks. *Business & Information Systems Engineering*, *5*(3), 179–193.

Saganowski, S., Gliwa, B., Bródka, P., Zygmunt, A., Kazienko, P., & Koźlak, J. (2015). Predicting community evolution in social networks. *Entropy*, *17*(5), 3053–3096.

Şen, F., Wigand, R., Agarwal, N., Tokdemir, S., & Kasprzyk, R. (2016). Focal structures analysis: identifying influential sets of individuals in a social network. *Social Network Analysis and Mining*, *6*(1), 17.

Singer, Y. (2012). How to win friends and influence people, truthfully: influence maximization mechanisms for social networks. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 733–742). ACM.

Takaffoli, M., Fagnan, J., Sangi, F., & Zaïane, O. R. (2011). Tracking changes in dynamic information networks. In *Computational Aspects of Social Networks (CASoN), 2011 International Conference on* (pp. 94–101). IEEE.

Utz, S., & Jankowski, J. (2016). Making "friends" in a virtual world: The role of preferential attachment, homophily, and status. *Social Science Computer Review*, *34*(5), 546–566.

Van Dijck, J., & Poell, T. (2013). Understanding social media logic.

Vasanthakumar, G. U., Priyanka, R., Raj, K. V., Bhavani, S., Rani, B. A., Shenoy, P. D., & Venugopal, K. R. (2016). PTMIB: Profiling top

most influential blogger using content based data mining approach. In *Data Science and Engineering (ICDSE), 2016 International Conference on* (pp. 1–6). IEEE.

Victor, S. P., & Britto, C. (2014). Analytical Study of Collective Behavior in Web Social Network Using Variant Datasets. *IJITR*, *2*(4), 1101–1106.

Weber, L. (2009). *Marketing to the social web: How digital customer communities build your business*. John Wiley & Sons.

Zhang, B., Huang, G., He, H., & Ren, J. (2017). Approach to mine influential functions based on software execution sequence. *IET Software*, *11*(2), 48–54.

8 NOTES ON CONTRIBUTORS

Babar Shazad, Department of Computer Science, COMSATS University, Attock, Pakistan. Babar Shazad is a student of master's degree in computer science from the COMSATS University, Attock Campus, Pakistan. He is currently serving as a Lecturer with the Department of Computer Science, Punjab Collage, Attock, Pakistan. He has also served as a Lecturer in University of Education, Attock, Pakistan. He has over 7-year experience in teaching and research in various well known educational and research institutes. His research interests include Data Mining, Social Media Analysis and Applied Artificial Intelligence.

Hikmat Ullah Khan, Department of Computer Science, COMSATS University, Wah Cantt, Pakistan. Hikmat Ullah Khan received the master's degree in computer science and the Ph.D. degree in computer science from International Islamic University, Islamabad. He has been an Active Researcher for the last ten years. He is currently an Assistant Professor with the Department of Computer Science, COMSATS Institute of Information Technology, Wah Cantt, Pakistan. He has authored a number of research articles in top peer-reviewed journals and international conferences. His research interests include Social web mining, Semantic Web, data science, information retrieval, and scientometrics. He is a member of the Editorial board of a number of prestigious Impact Factor Journals.

Zahoor-ur-Rehman has experience both in academia and research. He has received his educational and academic training at university of Peshawar, Foundation University Islamabad and UET Lahore, Pakistan. He joined COMSATS institute of information technology as assistant professor in the early 2015. Along with teaching responsibilities, he is an active researcher and reviewers of various conferences and reputed journals.

Muhammad Farooq, Department of Computer Science, Government College Rehmatabad, Rawalpindi, Pakistan. Muhammad Farooq received the master's degree in computer science from the COMSATS Institute of Information Technology, Attock Campus, Pakistan. He is currently serving as a Lecturer with the Department of Computer Science, Government College, Rehmatabad, Rawalpindi, Pakistan. He has over ten-year experience in teaching, software development and research in various well known educational institutes and software organizations. He is an Oracle Certified Professional. He is currently serving as a Lecturer with the Department of Computer Science, Government College, Rehmatabad, Rawalpindi, Pakistan. His research interests include software engineering, software development, and scientometrics.

Ahsan Mahmood, Department of Computer Science, COMSATS University, Attock, Pakistan. Ahsan Mahmood received the master's degree in computer science from the COMSATS University, Attock campus, Pakistan. His research interests include Data Mining, Social Media Analysis, Sentiment Analysis and Machine Learning.

Irfan Mehmood, Department of Media, Design and Technology, University of Bradford, Bradford, UK. Irfan Mehmood is currently a Lecturer in Applied Artificial Intelligence, School of Media, Design and Technology, Faculty of Engineering and Informatics, University of Braford, United Kingdom. He has been involved in IT industry and academia in Pakistan, South Korea and UK for a decade. His sustained contribution at various research and industry-collaborative projects gives him an extra edge to meet the current challenges faced in the field of multimedia analytics. Specifically, he has made significant contribution in the areas of video summarization, medical image analysis, visual surveillance, information mining, and deep learning in industrial applications. He is an active member of IEEE. He has also provided editorial services in various special issues in top ranked Journals of reputed publishers: Elsevier, IEEE, Springer and Wiley. He is also serving as a professional reviewer for numerous journals and conferences.

Yunyoung Nam received the B.S., M.S., and Ph.D. degrees in computer engineering from Ajou University, South Korea, in 2001, 2003, and 2007, respectively. From 2007 to 2010, he was a Senior Researcher with the Center of Excellence in Ubiquitous System. From 2010 to 2011, he was a Research Professor with Ajou University. He also spent time as a Post-Doctoral Researcher at the Center of Excellence for Wireless and Information Technology, Stony Brook University, NY, USA, from 2009 to 2013. From 2013 to 2014, he was a Post-Doctoral Fellow with the Worcester Polytechnic Institute, Worcester, MA, USA. In 2017, he was the Director of the ICT Convergence Rehabilitation Engineering Research Center, Soonchunhyang University, where he is currently an Assistant Professor with the Department of Computer Science and Engineering. His research interests include multimedia database, ubiquitous computing, image processing, pattern recognition, context-awareness, conflict resolution, wearable computing, intelligent video surveillance, cloud computing, biomedical signal processing, rehabilitation, and healthcare system.

Seungmin Rho is now a faculty of Department of Software at Sejong University in Korea. During 2013-2018, he was an Assistant professor at Department of Media Software at Sungkyul University. In 2012, he was an assistant professor at Division of Information and Communication in Baekseok University. In 2009-2011, he had been working as a Research Professor at School of Electrical Engineering in Korea University. In 2008-2009, he was a Postdoctoral Research Fellow at the Computer Music Lab of the School of Computer Science in Carnegie Mellon University. He gained his B.Science. (2001) in Computer Science from Ajou University, Korea (South), M.Science. (2003) and Ph.D. (2008) in Information and Communication Technology from the Graduate School of Information and Communication at Ajou University. He visited Multimedia Systems and Networking Lab. in Univ. of Texas at Dallas from Dec. 2003 to March 2004. Before he joined the Computer Sciences Department of Ajou University, he spent two years in industry. His current research interests include database, big data analysis, music retrieval, multimedia systems, machine learning, knowledge management as well as computational intelligence.