iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.43176

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

TD score: Time Aware Domain Similarity based Link Prediction

Yugchhaya Galphat
Assistant Professor, Computer Engineering, Vivekanand Education Society's Institute of Technology, Mumbai, India

Abstract: Online Social Network has gained immense traction of users in past decade. Link prediction across social networks has become a new exploration area for researchers, where existing links are investigated and new links are anticipated among billions of online customers. Majority of work in this area focusses on exploring the current status of a particular network at a specific time, without exploring the behavior of the network links as time goes by. Only a Small amount of work has been performed with the consideration of temporal aspect of network. As the interests and interactions of user change over time, the links among nodes become weaker or noisy which affects the prediction accuracy. This paper intend to explore a new integrated temporal method TD score which includes time stamp of interaction and domain similarity information for each pair of unconnected nodes to predict links. Experiment over co-authorship network reveals that link prediction covering time aware domain similarity is effective and efficient approach than traditional ones.

Keywords: Co-authorship network; link prediction; node similarity; global feature; temporal feature; domain similarity.

I. INTRODUCTION

The concept of online social network (OSN) is introduced by J.A. Barnes in 1954 and from 1967 it has been taken as new research area. OSNs are ubiquitous in nature, some social networking sites are application specific while some of them are designed for social interaction. It provides platform that brings people together with common interest. A social network represent a relationship amongst a set of entity joined together by some kind of relationship, such as co-authorship in which two author are connected if they co-authored any paper(published paper together).These networks can be represented as graph or hyper graph. Link prediction has become a growing research focus in network analysis domain having wider application areas such as recommender System, viral marketing, communication surveillance, information integration. An entity in given network can be linked by some relation with another entity in the future, even when no past relation has been observed between them. Link prediction issue has mainly been addressed as inferring new links only by exploring state of network at particular moments of time.

Immense work has been carried out on the static link prediction like: authors [20] suggested a way to predict future link by exploring a location feature along with supervised learning using place-of-friend and friends-of-friends attribute along with supervised learning. Various methods have been evaluated which consider the topology of a given network without focussing on attributes for individual nodes [21]. Clustering and hierarchical structure was also used as basis for prediction [22] where nodes of a graph presented as leaves of a tree representing a community with a recursive structure. However some authors also applied probabilistic relational model [23] [24]. Still there is a need to temporally evaluate network structure to perform link prediction. *Potgieter et.al* [12] had shown the benefits of temporal feature in their earlier research.

Most of the social networks are time evolving and dynamic in nature where strength of link varies over time. Consideration of only static view of network is not perfect measure for prediction, it is also equally important to evaluate the network behaviour across time domain. Traditional approaches for link prediction fails in exploring the network evolution because they consider network data up to the present time without giving any consideration when links were developed in the network. Temporal feature is not explored fully up till now for the prediction. Only small amount of work has been done using time base information like history of network evolution, data on time of interaction across various entities which results in more accurate predictions. For the co-authorship network, links and its strengths vary over time because of authors, who were active in particular field, have lost their interest in that or became inactive after some years. So considering the static view of network (single time snapshot) will give noisier information. It is inefficient to use such information for link prediction. So that this paper propose an integrated novel approach which calculates similarity Score for each pair of non-connected node, based on common neighbour, time of interaction of users and their area of interests (research areas) with respect to time**.** The learning algorithm in this approach takes the dynamic view of network as input which is efficiently utilize time aware domain information of individual authors. In order to verify the feasibility of the proposed approach, we executed number of experiments on arXiv dataset covering network data for co-authors). The experimental results showed that by including this temporal feature, performance accuracy is increased than the traditional ones.

©IJRASET: All Rights are Reserved

II. LINK PREDICTION ACROSS NETWORKS

A. Online Social Network

Social networks have emerged as a significant and promising field of study within computer science with the advent of Web 2.0. There has not been a single well-established and standard definition for OSNs; rather there exist a number of definitions given by different researchers according to their individual viewpoints. *Mitchell (1969)* [25] defined in the following way: "social network is a specific set of linkages among a defined set of persons, with the additional property that the characteristics of these linkages as a whole may be used to interpret the social behaviour of the persons involved". Katarzyna Musiał classified social networks into three classes: Homogeneous (HSN), system–based (SSN) and internet multisystem social network (ISN). The social network may be used for many purposes like business or socializing. Examples of websites that are primarily used for socializing includes Facebook and MySpace. Wikipedia, Flickr, Facebook, etc., representing a growing set of Community Based OSNs are becoming increasingly popular due to increased internet penetration as well as growing set of users well versed with web. Web users could obtain and contribute knowledge through Wikipedia. Facebook can connect people with their communities sharing similar interests while Flickr acts as a platform for Web users to seek and post favourite photos.

B. Link Prediction

Link prediction is subfield of social network analysis which focuses on mainly two types of problem - link detection and future link prediction. In link detection, network structures are analyzed to know the existence of links among various nodes in current time t. Link detection can further be expanded to know *link-type* "which type of association or attribute(s) on the basis of link has formed?", and *cardinality* " number of associations of node pair?". The task of accurately predicting the upcoming edges in the network in future interval is called Link prediction. Some problem occurs in the selection of training and testing data, when the link prediction is treated as a binary classification task due to random pairs of vertices in Online Social Networks.

According to authors [26] several techniques, including structural and temporal techniques, have been employed to address the issue of link prediction. Structural technique consider node similarity which is a local features such as - Common neighbour [13], Admin/Adar [14], Jacquard's coefficient [15] etc., and also includes topological feature which is a Global similarity such as - Katz coefficient [16], Shortest path algorithms, page rank [17], Random walk with restart [18], Sim rank [19] etc., and Probabilistic model based approaches. Models like PRM (Probabilistic Relational Models) and DAPER (Directed Acyclic Probabilistic entity relationship), or discriminative model like SRM (Stochastic Relational Models) can be categorized as generative graphical models. These models are applied to build link trait which are redirected into one of the Learning algorithms or binary classification algorithms to predict future links in a given network. Probabilistic models and approaches can be power full tool to extract connectivity information among all. However, such link prediction approaches once applied to the big and complex network will be computationally unproductive. Temporal approach solves this problem and provides a better solution to overcome the above issues raised. Proposed Temporal technique applies time based structural data as time of link creation, duration of interaction etc. for calculating future links. Temporal graph can be created by encoding temporal data into graphs. This will act as a valuable tool for temporal analysis and prediction. This description helps us to look at the active temporal properties of data. Previous link prediction methods are based only on graph structure, without giving any weightage to a critical aspect covering evolution history. Since social networks represent two kinds of link: one is persistence relation e.g. friendship between two people another is a discrete event e.g. co-authorship of academic and research publications. Former relation can easily analyze with simple graph but later case requires time aware network information to predict new as well as repeated links among nodes. Temporal metrics have demonstrated extremely significant features in terms of accuracy and new contribution to link prediction. Weighted time aware maximum entropy method [7] can predict both new as well as repeated link. Relations at distance two or higher can also be predicted by this method. Ryan Rossi et.al [13] had proposed time varying relational classification model for such kind of dynamic relationship.

III. PROPOSED WORK

Machine learning approaches remain an immense challenge for link prediction just because of dynamically changing networks [26]. Huge size of these networks makes it extremely hard to study and analyze the structure of social network activity graphs. The temporality in social network can be caused by various factors depending on the nature of networks. Link prediction can be effectively performed by using these factors. Some links which are strong, after a definite period become weaker and fade. Strong links have a greater influence over link evolution than weaker links, so that the behavior of link increases complexity in prediction. It would be more successful if higher weights are allocated to more recent publications. *Munasinghe and Ryutaro Ichise* [3] introduced an additional feature called *time Score* to carried out temporal link prediction.

This time Score damps the old interaction and assigns higher weights to recent interactions with the help of damping factor *β*.
This paper introduced new approach which exploits temporal information which is different from previous one. Here combination of three different attributes is used for feature generation and link prediction. First two are local attributes and third attribute is time based. To find out similarity between node pair, common neighbor technique is used.

$$score(x,y)=\Gamma(x)\cap\Gamma(y) \qquad (1)$$

Conceptually, it defines common neighbors of a pair of authors x and y from the neighbor sets of x and y. Higher Score represents high number of common neighbors and vice versa. High score means link is more probable between x and y. It may be possible that authors who are having common neighbors will not collaborate in future because they perform research in different areas. So other Parameter *domain similarity* has also been used as a second attribute that restricts links between those authors, who do not have even a single common research area so that the false predictions is avoided from the results.

$$Ds(x,y)=D(x)\cap D(y) \qquad (2)$$

Where D(x), D(y) represents domain areas of authors x, y respectively. To apply time awareness in the prediction task time score which given by [3] is used as third attribute, illustrated in equation 3. Basic idea behind this index is that authors who are active in the past in particular field have lost their interest in that, or has become inactive. So creating link with them is worthless. Collaboration of node pair with central node (common neighbour) with respect to time is an important aspect. If one node recently collaborated with common node and other long time ago that will affect the prediction.

$$TS=\sum_{n}\frac{H_m.\beta^k}{|t_1-t_2|+1} \qquad (3)$$

$$k=currenttime-\max(t_1,t_2)$$

Where t_1 and t_2 are recent interactions of common neighbor z with x and y. $\beta \in [0, 1]$ is used as damping factor which damps the older interactions and gives higher weight to the recent one. h_m show harmonic mean of publications of x and y with its common neighbors. Fig.2 describes this phenomenon where six authors (A, B, C, D, E, F) works on different domains (D1, D2, D3, D4, D5 and D6).solid lines represent actual interaction of authors, t1 and t2 shows time of interactions While Dashed lines shows the predicted link at future time tc, according to our approach. Authors pair (AC, DF, EC) will be link in future because they have at least a common domain of research (D5, D4, D5) and also common neighbor (B, B and E, D) who is active at time t2(recent interaction)respectively. Authors pair AD and BE having a recently active common neighbor B and D but not have even a single domain in common. There is neither recently active common neighbor nor common research field between of author pair AE. Whereas author pair CF have domain similarity but not any recent interaction with common neighbor. So that links (AD, BE, AE, CF) will not form by improved integrated methodology. By including these features, performance accuracy will be increased in comparison to work done before. To incorporate this temporal aspect of network we define TD score which find out strength of link on basis of time aware domain similarity. This is an integrated approach which includes combination of three different features. We formulated it to calculate similarity score among pair of node called TD score (Time domain score).

TD score: For a network G (n, e), where e represents a set of edges (links) and n stands for the set of nodes node set respectively. For a pair of node x and y having n common neighbors TD score can be defined as:

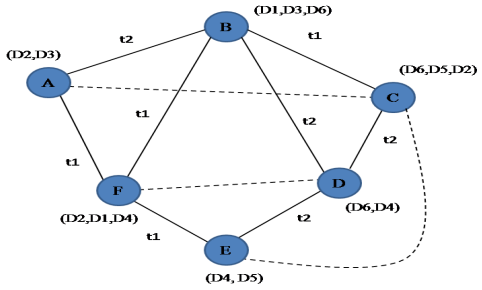


Fig 1: Time Aware Domain Similarity Based Link Prediction

$$TDScore=TS*(1+\log_{10}(1+C\{Ds(x,y)\})) \qquad (4)$$

Where TS is the time score given in equation 3 and Ds is a set, representing domain similarity of x and y given in equation 2. C is the cardinality of that set. The log term used here gives more weight to TD score as domain similarity increases. Where 1 inside the log term is used to avoid indefinite value of log (0) in case of dissimilar domain or C=0. Since value of log (1) is 0, outside 1 is used to preserve non-zero value of TD score. In case of single similar domain TD score will b equal to TS.

If D(x) = {IT, ME, DC, CN} and D(y) = {NA, IT, DC} then

Ds = {IT, DC} and C {Ds} =2

If z represent common neighbor of x and y, z co-authored papers with x in 2009, 2010,2011 and with y in 2009 and 2010 and current year is 2013 then-

K= 2013-max(2011,2010)

$$h_m = \frac{2}{\frac{1}{3}+\frac{1}{2}} = 2.5 \qquad (5)$$

$$TDscore = \frac{2.5*(.5)^2*(1+\log_{10}3)}{|2011-2010|+1} = 0.4616 \qquad (6)$$

IV. EXPERIMENTAL EVALUATON

A. Dataset and Background

Proposed method on co- authorship network, Cond-mat (arXiv) dataset have chosen for understanding author's interaction and their publications in recent years with the help of collaboration graph. Authors who are close in collaboration graph repeatedly interact, so that some redundant links are obtained from the dataset. Further the dataset is cleaned by eliminating such redundant links. Publications of single author were also excluded since they do not help in finding links (prediction) anyway. Three years of data is used to build collaboration graph over which TD score method is applied. Last two years of data yields true growth in the network, on which prediction accuracy is calculated. Some problem occurs in the selection of training and testing data, when the link prediction is treated as a binary classification task due to random pairs of vertices in Online Social Networks. Weka data mining tool is used to perform classification and link prediction.

B. Evaluation Parameter

Experimental analysis is performed on both temporal and non temporal methods. Common neighbor and common neighbor along with domain similarity are the time unaware methods. Time score and integrated time aware domain similarity (TD score) are having temporal aspect. With the use of training and testing datasets ten-fold cross validation is performed. To evaluate the effectiveness of all four methods, we have considered four performance measures: Precision, Recall, F- measure, Accuracy given by the following formulas:

$$\mathrm{Pr}ecision = \frac{|TP|}{|TP|+|FP|}$$

$$\mathrm{Re}call = \frac{|TP|}{|TP|+|FN|}$$

$$F-measure = \frac{2*\mathrm{Pr}ecision*\mathrm{Re}call}{\mathrm{Pr}ecision+\mathrm{Re}call}$$

$$Accuracy = \frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|}$$

Where TP, FP, FN, TN shows True Positive, False positive, False Negative, True Negative respectively.

C. Results

Experimental result shows that non temporal methods are poor in their prediction because false positive and false negative both values are too high so that they have accuracy only (64.72%). It is also seen that the Performance accuracy was slightly improved (77%) after incorporating domain similarity attribute with common neighbor. Both Time score and TD score performed very well. TD score method has done less false predictions than the Time score and has highest accuracy (85.19%). Table.1 illustrates performance matrices across: Precision, Recall F-measure and Accuracy for all four methods. Fig 2 and Fig 3 depicts the performance of all comparative method under evaluation. It is revealed from the result that the proposed integrated time aware method is best among all.

Methods / Measures	CN	CN with DS	TD Score
Accuracy	64.72 %	77%	85.19%
Precision	0.694	0.788	0.891
Recall	0.647	0.77	0.852
F-Measure	0.61	0.748	0.862

Table1. Performance Matrix

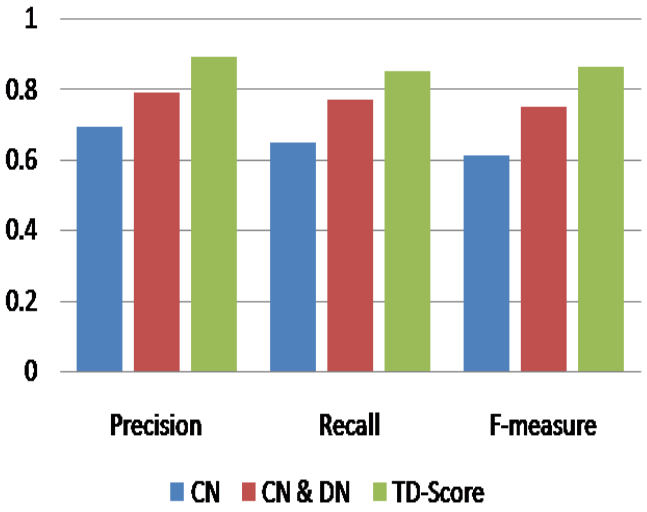


Fig 2: Performance chart1

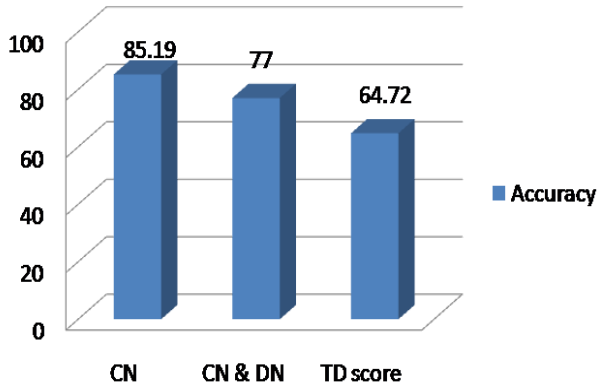


Fig 3: Performance Accuracy

V. RELATED RESEARCHS

A lot of research work related to link prediction has been performed till now, but temporal feature was not given much attention [26]. Different authors exploited temporal feature in their own way. In [7] authors *Tomasz Tylenda, Ralitsa Angelova, Srikanta Bedathu* introduced maximum entropy with time aware feature and weighted time aware methods. They considered temporally extended probabilistic model that makes use of only recent interactions, avoiding the old ones for predicting new as well as recurring links. While Danial M. Dunlavy et.al [6], focused on periodic link prediction rather than normal temporal prediction by focusing on truncated singular value decomposition (TSVD). The method proposed by them was able to effectively answer queries like: "Who is most likely to publish at the IEEE conference next year?" or "Where is Alice most likely to publish next year?" using O (M + N) memory. In addition, they demonstrated that the Tensor based method, though requiring higher space and time complexity, demonstrated prediction accuracy than matrix factorization for link prediction.

Satoshi Oyama et.al [5] experimented for the communication system such as email and predicted cross sectional links, among nodes in different time frames. Sheng Ga et.al [4] proposed a matrix factorization model and resulting an optimization method GRJMF (Graph Regularized Joint Matrix Factorization) to realize the underlying factors of this model. They used time evolving past history of interaction among users to predict links. Furthermore, *Yizhou Sun* [2] applied temporal phenomenon on heterogeneous social network where objects and link varies over time. Paulo Ricardo da Silva Soare et.al [1] also considered time series snapshots and local similarity for link prediction using supervised and unsupervised approach. Jia Zhu et.al, [10] developed 'DynamicNet', a hybrid prediction framework. Where time series was applied with the advantages of topological pattern and PRM for link prediction. To find new friend in the location based social discovery network *Terence Chen et.al* [9] used spatio-temporal information.

VI. CONCLUSION AND FUTURE WORK

Temporal link prediction is gaining immense research interest, as prominence of dynamically changing networks is projected to significantly increase in future. It is inefficient to consider single snapshot of network, rather multiple snapshots at particular time interval are needed to accurately predict association among users. Incorporating temporal information available on evolving social networks is valuable feature to predict actual growth of network. This paper proposed a method that considers temporal aspect of network for future link prediction. Common neighbor is a useful feature to calculate the similarity among node pairs. Area of interest along with common neighbor is more useful than considering only common neighbor to predict link. Including interaction time of authors with their common neighbour and their respective fields of research together is efficient approach to evaluate the probability of future interaction. The accuracy of link prediction is increased by our proposed method TD score which exploits time aware domain similarity information. False positive results are less than other experimented method so that the performance accuracy is increased.

Since temporal feature extremely valuable feature in term of accuracy so that it should be apply in various aspect to evolving social network. The proposed time aware method applied only on co-author ship network, experiment on other networks can be the future work. Dataset used here, is having un-directional links. The work can be extended by considering directional dataset such that link from main authors to all its co-author(s). For the location based social network, locations of the node pair and the central node with respect to time can be another temporal aspect to evaluate future association.

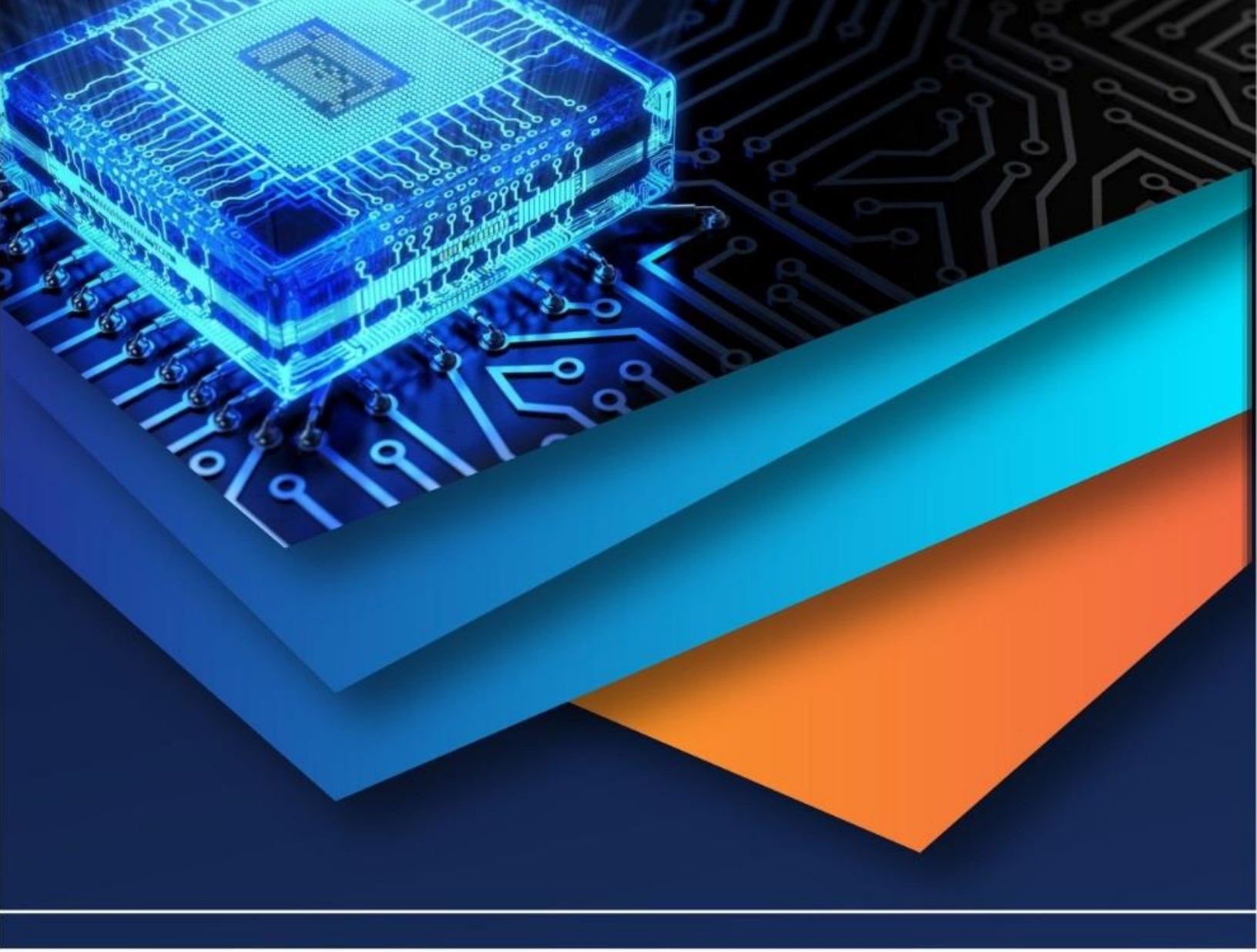
REFERENCES

[1] Paulo Ricardo da Silva Soares, Ricardo Bastos Cavalcante Prudencio, "Time Series Based Link Prediction"IJCNN the 2012 International Joint conferences, pp. 1-7, IEEE.
[2] Yizhou Sun, Jiawei Han, Charu Aggarwal, Nitesh V. Chawla, "when Will It Happen? Relationship Prediction in Heterogeneous Information Networks", WSDM'12, Seattle, Washington, USA, pp. 663-672 February 8–12, 2012.
[3] Lankeshwara Munasinghe and Ryutaro Ichise, "Time Aware Index for Link Prediction in Social Networks," A. Cuzzocrea and U. Dayal (Eds.): DaWaK 2011, LNCS 6862, pp. 342–353, 2011.
[4] Sheng Gao, Ludovic Denoyer, Patrick Gallinari, "Temporal Link Prediction by Integrating Content and Structure Information", CIKM'11, Glasgow, Scotland, UK, pp. 24–28, October 2011.
[5] Satoshi Oyama, Kohei Hayashi and Hisashi Kashima, "Cross-temporal Link Prediction," Data Mining (ICDM), 2011 IEEE 11th International Conference, pp. 1188-1193.
[6] Danial M. Dunlavy and Tamara G. Kolda, "Temporal Link Prediction Using Matrix and Tensor Factorizations", ACM Journal Name, Vol. V, No. N, Month 20YY, pp. 111-137.
[7] Tomasz Tylenda, Ralitsa Angelova, Srikanta Bedat hur "Towards Time-aware Link Prediction, in Evolving Social Networks," the 3rd SNA-KDD Workshop 09, Paris, France June 28, 2009.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

[9] Terence Chen, Mohamed Ali Kaafar, Roksana Boreli, "where and when of finding new friend: analysis of a Location-based Social Discovery Network" 7th international AAAI conference on weblogs and social media 2012 NICTA.
[10] Jia Zhu,Qing Xie and Eun Jung Chin,"A Hybrid Time Series Link Prediction Frame work for large Social network"S.W.Lidell al.(edi) DEXA 2012 Part II,LNCS 7447, Verlag Berlin Heidelberg, pp. 345-359.
[11] R Rossi, Neville, "Time-evolving relational classification and ensemble methods" PAKDD 2012, Part I, LNAI 7301,Springer-Verlag Berlin Heidelberg 2012ime.Paris, France pp. 1–13, 2012.
[12] A. Potgieter, Kurt April, R.J.E. Cooke, I.O. Osunmakinde "Temporality in Link Prediction: Understanding Social Complexity" PAKDD 2012, Part I, LNAI 7301, Verlag Berlin Heidelberg 20 (2007) pp. 1–13, 2012.
[13] J. Chen, W. Geyer, C. Dugan, M. Muller, I. Guy, "Make new friends, butkeep the old: Recommending people on social networking sites", CHI 2009, Boston, MA, USA, pp. 3–9, April 2009.
[14] Lada A. Adamic and Eytan Adar, "Friends and neighbors on the web Social Networks", 2003.
[15] Gerard Salton and Michael J. McGill. "Introduction to Modern Information Retrieval'. McGraw-Hill, Inc. New York, NY, USA1986.
[16] Leo Katz. "A new status index derived from sociometric analysis" Psychometrika", 18(1): March 1953 pp. 39–43.
[17] Sergey Brin and Lawrence Page. "The anatomy of a large-scale hyper textual web search engine," Journal Computer Networks and ISDN Systems, 30(1–7), 1998 pp. 107–117.
[18] H. Tong, C. Faloutsos, J. Pan, "Fast random walk with restart and its applications, in": Proceedings 6th International Conference on Data Mining(ICDM'2006), Hong Kong, 2006, pp. 613–622.
[19] Glen Jeh and Jennifer Widom. "Simrank: a measure of structural-contexts similarity", In KDD, Proceedings of the eighth ACM SIGKDD, 2002 pp. 538–543.
[20] Salvatore Scellato, Anastasios Noulas, Cecilia Mascolo, "Exploiting Place Features in Link Prediction on Location-based Social Networks" ,KDD '11 Proceedings of the 17th ACM SIGKDD, pp. 1046-1054.
[21] Michael Fire, Lena Tenenboim, Ofrit Lesser, Rami Puzis, Lior Rokach and Yuval Elovici, "Link Prediction in Social Networks using Computationally Efficient Topological Features,"2011 IEEE Third International Conference on Social Computing, pp. 73 – 80.
[22] Elham Hoseini, Sattar Hashemi, Ali Hamzeh "Link prediction in social network Using Co-clustering based Approach", proceeding 2012 on international conference IEEE, pp. 795 – 800.
[23] Hisashi Kashima, Naoki Abe "A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction", ICDM 06'sixth international conference Data Mining IEEE, pp. 340 – 349.
[24] Lise Getoor, Nir Friedman, Daphne Koller, Benjamin Taskar, "Learning Probabilistic Models of Link Structure" 2002 Lise Getoor, Nir Friedman, Daphne Koller and Benjamin Taskar.
[25] Richter, Alexander, Koch, Michael (2008). Functions of Social Networking Services. In: Proc. 8TH International Conference on the Design of Cooperative Systems, pp. 87-98.
[26] Y. Dhote, N. Mishra and S. Sharma, "Survey and analysis of temporal link prediction in online social networks," 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Mysore, 2013, pp. 1178-1183, doi: 10.1109/ICACCI.2013.6637344.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)