International Journal of Electronics, Communication & Soft Computing Science and Engineering
ISSN: 2277-9477, Volume 3, Issue 6

Temporal Link Prediction using Neighborhood Integrated Matrix Factorization

Ms.Trishna Jadhav Prof. Satish S. Banait

Abstract:- **The ability to predict links among data objects is central to many data mining tasks such as social network, business analytics, and recommendation system. Link Mining and temporal link prediction is emerging trend in recent years. Link mining deals with heterogeneous and homogeneous data sources that generates link data and this link data provides scope for collaborative filtering tasks, which is a prime requirement in recommending systems and a significant role is played in predictive analytics .The proposed introduction of Neighbourhood Integrated Matrix Factorization method improves the accuracy of missing value predictions as pre-heuristic task of Neighbourhood Similarity Computation which produces object profiles. In general, link prediction problem is modelled as, given link data for times 1 through T, the task is to predict links at time T+1, And if data has underlying periodic structure, up to what extend predictions can be made in future time T+2,T+3……T+k.(k>0).**

Keywords- **Collaborative Filtering, Neighbourhood Integrated Matrix Factorization, Predictive Analytics., Temporal link prediction.**

I. INTRODUCTION

Many objects and entities in the world are dependent, and linked to many other objects through a diverse set of relationships: people have friends, family and co-worker's: scientific papers have authors , venues , and references to other papers; web pages links to other web pages and have hierarchical structures; proteins have locations and functions, and interact with other proteins. In link mining, the connections among objects are explicitly modeled to improve performance in task such as classification, clustering, and ranking, as well as enabling new applications, such as link prediction. Most link mining problems have a great deal of uncertainty as well. Link data is typically very noisy and incomplete [2]. The data in different analysis applications such as social networks, communication networks, web analysis, and collaborative filtering consists of relationships, which can be considered as links, between objects. For example, two people may be linked to each other if they exchange emails or phone calls. These relationships can be modelled as a graph, where nodes correspond to the data objects (e.g., people) and edges correspond to the links (e.g., a phone call was made between two people). The link structure of the resulting graph can be exploited to detect underlying groups of objects, predict missing links, rank objects, and handle many other tasks [1].Link mining refers to data mining techniques that explicitly consider these links when building predictive or descriptive models of the linked data. Commonly addressed link mining tasks include object ranking, group detection, collective classification, link prediction and sub graph discovery. This is actually an exciting, and rapidly increasing area. There is not yet comprehensive framework that can support a combination of link mining tasks as needed for many real applications. Link prediction is a sub-field of social network analysis. It is concerned with the problem of predicting the (future) existence of links amongst nodes in a social network. The link prediction problems is interesting in that it investigates the relationship between objects, while traditional data mining tasks focuses on objects themselves [3]. Dynamic interactions over time add another dimension to challenge of mining and predicting link structure. This study focuses on basic link mining tasks and the problem of temporal link prediction. In this problems given link data for T time steps, can one predict the relations among data objects at time T+1,T+2,…..,T+k. where (k >0), is a point of interest. Section 2 of this review is to introduce the various link mining tasks and addresses the categories of link mining taxonomy that includes object, link and graph. Various time series analysis approaches for statistical data and proposed "Neighbourhood Integrated Matrix Factorization" method in state space approach for matrix factorization is been discussed in section 3. [1]

II. LINK MINING TASK

Eight link mining tasks that are broadly categorized as tasks that focus on objects, links, or graphs.

Table I. Link Mining tasks taxonomy [32].

Sr. No	Task of Interest	Goal	Approach	Example	Work Contributed
Object Related Tasks					
1	Link based object ranking	Prioritize set of objects by exploiting link structure of graph	Page Rank Hits	Web Information retrieval	Leskovec et.Al.[7] J.Kleinberg et.al .[8]
2	Link based object classification	Predicting the class or label of an object based on its attributes and its links and attributes of linked objects	Machine learning classifier	Web:predict the category of a web page ,based on words that occur on the web page,links between pages . Cite:Predict the topic of a paper , based on word occurrences ,citations, cocitations.	Lu and Getoor et .al [9]
3	Object Clustering	Predicting when a set of entities belong to the same group based on clustering both object attribute values and link structure	Group detection	Sna: identifying communities and tribes. Cite: identify research communities.	S.waasserman et.al [10]
4	Object identification	Predicting when two objects are the same,based on their attribute and links	Entity resolution	Web: predict when two sites are copy of each other . Cite: predicting when two citations are referring to the same paper.	Li et.al.[11]
Link Related task(link prediction)					
5	Link type classification	Predicting type or purpose of link based on properties of the participating objects	Model based probabilistic approaches	Web: predict advertisement link or navigational link: predict an advisor-advisee relationship. Epi : predicting whether contact is familiar,coworker or acquaintance.	Getoor et.al [12]
6	Predicting Link Existence	Predicting whether a link exits between two objects.	Collaborative filtering approaches	Web:predict if there will be a link between two pages Cite: predicting the participation of objects/authors in events of co-authorship. Predicting whether a paper will cite other paper.	Libennowell et.al [13]
Graph related tasks					

7	Sub graph discovery	Find characteristics sub graph	Focus of graph base data mining	Bio-protein structure discovery Chem-chemical structure discovery	Cook and holder et.al [14]
8	Graph classification	Goal is to categorised an enter graph as positive or negative instance of a concept	Schema mapping, schema discovery, schema reformulation	Cite-matching between two bibliographic sources Web-discovering schema from unstructured or semi-structured data Bio-mapping between two medical ontology	King et .al [15]

*web data (web) bibliographic data (cite) social network analysis (sna), biological data (bio) chemical data (chem), epidemic data (epi)

III. TEMPORAL LINK PREDICTION

Combined approach of static graph link prediction algorithms and time series model produced significantly improved predictions than static graph link prediction methods, demonstrating the great potential of integrated methods that exploit both interlink structural dependencies and intra-link temporal dependencies. This section introduces the time series link prediction problem, taking into consideration temporal evolutions of link occurrences to predict link occurrence probabilities at a particular time and predicting the missing values to fill the sparse user item matrix.

A. *Time Series Analysis*

Most of the literature on link prediction has been formulated based on a static network setup, where a partial network structure is known and the objective is to predict the hidden links of the underlying complete network. In such a static network, link occurrence is modelled as a onetime event and the primary interest is on the existence of the linkages rather than the timing of the occurrence or frequency or occurrences. However, in many application settings that involve dynamic evolving networks/graphs, link occurrence is preferably modelled as a sequence of binary states or occurrence frequencies. In this scenario, if occurrence of a link at a particular time represented using a random variable (binary or real-valued depending on whether binary occurrences or occurrences are modelled), we are dealing with a multivariate time series data with large number of variables. Time series analysis is a well-established field in statistics which provides systematic approaches to modelling of data with time correlations. There have been many recent studies on dynamic or evolving networks that consider temporal connectivity data. Most such studies convert the temporal connectivity data into a sequence of non over lapping network snapshots by aggregating links within each discrete unit of time period into a graph. The majority of such studies extend from the main body of network science literature to characterize the time-varying structure of the graph series, such as density and diameter [16], sub graph and cycle structures [17],and cluster formation patterns [18], based on empirical temporal network data.

The link prediction literature is largely based on this fundamental assumption that the network structure itself has predictive information regarding the hidden or future links. Theoretically, an ideal time series link analysis framework would need to integrate temporal and structural link dependencies simultaneously [4].

B. *State Space Approach*

The state space system models are useful concepts to represent relations, interactions and to build up tools for predicting the collective behaviour of entities. An input output model expresses the behaviour of an open system that interacts with its environment, in a causal manner(the past and the present of the system shape its future),which takes from its environment(input ,or stimulus) and gives to its environment(output ,or response).If one is interested in the behaviour of an observable(measurable) open system ,then one should observe the system for some time for stimulus-response relations and construct an input-output model that suitably fits to these observations. A similar causal model (local in time) is the state-space representation of an open system, where the underlying system's behaviour is represented by its state and its response to an input at that state [5].The collection of all possible (admissible) states of a system is known as the state-space. In order to predict the future states of a dynamic system, with minimum mean square error, one can design a recursive predictor based on the current estimate and the current filtered estimation error. The good estimate of the state of a linear system from partial observation is known as kalman filter [4], [5].

C. *Collaborative Filtering Method*

Collaborative filtering methods are widely adopted in commercial recommender systems [19], [20], [21].There are two types of collaborative filtering approaches which are widely studied: neighbourhood-based (memory based) and model-based. The most analysed example of neighbourhood-based collaborative filtering include user-based approaches [22], [23], item-based approaches [24],

[25], and their fusion [26], [27].User based approaches predict the missing values of a
user based on the values of similar users. Item based approaches predict the missing values of a current user based on the computed information of items similar to those chosen by current user. Neighbourhood based approaches often use the PCC algorithm [21] and the VSS algorithm [24] as the similarity computation methods. PCC-based collaborative filtering approaches generally can achieve higher prediction accuracy than VSS-based algorithms, since PCC considers the differences of the user value characteristics.

In model based approaches on the other hand training data sets are used to train a predefine model. Examples of model-based approaches include the clustering model [28], aspect model [29], and so on. Recently several matrix factorisation methods [30], [31], [32] have been proposed for collaborative filtering. These methods focus on fitting the user-item matrix with low-rank approximations, which is engaged to make further predictions. The premise behind a low dimensional factor model is that there is only a small number of factors influencing the values in the user-item matrix, and that a user's factor vector is determined by how each factors applies to that user. The neighbourhood-based methods utilize the values of similar users or items for making value prediction, while model-based methods, like matrix factorization models, employ all the value information of the matrix for making value prediction. Different from previous work this approach takes advantage of both local information of similar users and global information of the whole matrix to achieve prediction accuracy [1].

D. ***Neighbourhood Integrated Matrix Factorisation***

NIMF approach makes best utilization of both the local information of similar users and the global information of all the available values in the user-item matrix to achieve better prediction accuracy. This approach is designed as a two-phase process. In phase 1 user similarity is calculated using PCC and a set of Top-K similar users. Then based on neighborhood information NIMF approach is proposed to predict the missing values in user-item matrix.

Our proposed approach for temporal missing value prediction contains following steps of execution [1].

1. Convert available time series data into proper format (user-item matrix format)

$$PCC(i,k)=\frac{\sum_{j\in J}(R_{ij}-\overline{R}i)(R_{kj}-\overline{R}_k)}{\sqrt{\sum_{j\in J}(R_{ij}-\overline{R}_i)^2}\sqrt{\sum_{j\in J}(R_{kj}-\overline{R}_k)^2}}$$

where following two things are known that what data is available and what data is missing. One dimension of this matrix is users which represent rows and other dimension of the matrix is item which represents columns. Use this user-item matrix made in previous step, as input to "Neighbourhood Integrated Matrix Factorization method" (NIMF Method).Execute NIMF method as a two phase process as mentioned in below steps.

2. In phase 1, calculate the user similarities using PCC (Pearson correlation coefficient) and determine a set of Top-K similar users for each user.PCC can be calculated by using following formula,

where J is the subset of items where this item information is available for both user i and user k, Rij is the item value j observed by service user i , and Ri and Rk are the average of different item values observed by service.user i and k, respectively. From this definition, the similarity of two users i and k, PCC (i, k), is in the interval of [1:1], where a larger PCC value indicates higher user similarity.

3. In Phase 2, (let user-item matrix is of 'm * n'), introduce a new dimension 'l' and Factorize user-item matrix into two matrices U and V where Matrix U is of 'l * m' and Matrix V is of 'l * n'. In Matrix U, for every user out of 'n', there will be list of 'l' users which are similar to this user. In Matrix V, correspond to every user in Matrix U, there will be item value in Matrix V.

More specifically, each column of U performs as a "factor vector" for a user, and each column of V is a linear predictor for item value, predicting the entries in the corresponding column of the user-item matrix R based on the "factors" in U.

4. Set of Top-K similar users of user ui and Sik is the normalised similarity score between user ui and Sik is the normalised similarity score between user ui and user uk, which can be calculated by

$$S_{ik}=\frac{Pcc(i,k)}{\sum_{k\in T(i)}PCC(i,k)}$$

To minimize root mean square error and in order to complete missing values in prediction matrix.

CONCLUSIONS

The link prediction problem is a central to many data mining task and has gained attention for many predictive analytic tasks. The proposed work of temporal link prediction has focused on capturing temporal dynamics so as to provide better accuracy in prediction tasks. The proposed method of Neighborhood Integrated matrix factorization has scope for improving scalability issue over traditional link prediction approaches. Neighborhood Integrated Matrix Factorization approach systematically fuses the neighborhood-based and model-based collaborative filtering approaches to achieve higher prediction accuracy.

REFERENCES

[1] Zibin Zheng,Hao Ma,Michael R.Lyu,"Collaborative Web Service QoS Prediction via Neighbourhood Integrated Matrix Factorization"."IEEE Trans. On services computing" Vol 6. No 3.July-September 2013

[2] Dunlavy D. M., Kolda, T. G., and Acar E. "Temporal link prediction using matrix and tensor factorizations." ACM Trans. Knowl. Discov. Data 5, 2, Article 10 (February 2011), 27 pages.

[3] "Markov Logic: A Language and Algorithms for Link Mining" Pedro Domingos,Daniel Lowd,et.al.

[4] "State Space Model for Link Mining".Kushal P.Birla,Prof.S.M.Kamlapur,IJETTCS.ISSN 2278-6856.

[5] Neural Networks and learning machines(3rd ed.).(2011) by simon Haykin,PHI,ISBN 978-81-203-4000-8

[6] Sun, J.Z.; Varshney, K.R.; Subbian, K, "Dynamic matrix factorization: A state space approach," Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference, vol., no., pp.1897-1900, 25-30 March 2012.

[7] Leskovec, J., J. Kleinberg and C. Faloutsos. 2007. Graph evolution: Densification and shrinking diameters. ACM Trans. on Knowledge Discovery from Data.

[8] J. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604632, 1999

[9] Q. Lu and L. Getoor. Link-based classi_cation. In International Conference on Machine Learning, 2003.

[10] S. Wasserman and K. Faust. Social Network Analysis: Methods and Applications .Cambridge University Press, Cambridge, 1994.

[11] X. Li, P. Morie, and D. Roth. Semantic integration in text: From ambiguous names to identifiable entities. AI Magazine. Special Issue on Semantic Integration, 2005.

[12] L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models of link structure. Journal of Machine Learning Research, 3:pg679-707,2003.

[13] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In International Conference on Information and Knowledge Management (CIKM), pg 556-559, 2003.

[14] N. Ketkar, L. Holder, and D. Cook. Comparison of graph-based and logic-based multi-relational data mining. SIGKDD Explorations, 7(2), December 2005.

[15] R. D. King, S. H. Muggleton, A. Srinivasan, and M. J. E. Sternberg. Structure-activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. National Academy of Sciences, 93(1):pg438-442, January1996.

[16] Leskovec, J., J. Kleinberg and C. Faloutsos. 2007. Graph evolution: Densi_cation and shrinking diameters. ACM Trans. on Knowledge Discovery from Data, 1.

[17] Vazquez, A., J. G. Oliveira and A.-L. Barabasi. 2005. The inhomogeneous evolution of subgraphs and cycles in complex networks. Phys. Rev. Lett. E, 71 025103

[18] Holme, P., S. M. Park, B. J. Kim and C. R. Edling. 2007.Korean university life network perspective: Dynamics of a large affiliation network. Physical A, 373 821-830

[19] R. Burke, "Hybrid Recommender Systems: Survey and Experiments," User Modeling and User-Adapted Interaction, vol. 12, no. 4,pp. 331-370, 2002.

[20] X. Su, T.M. Khoshgoftaar, X. Zhu, and R. Greiner, "Imputation-Boosted Collaborative Filtering Using Machine Learning Classifiers,"Proc. ACM Symp. Applied Computing (SAC '08), pp. 949-950,2008.

[21] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," Proc. ACM Conf. Computer Supported Cooperative Work,pp. 175-186, 1994

[22] R.Salakhutdinov and A.Mnih "Probabilistic Matrix Factorization",Proc.Advances in Neural Information Processing Systems,pp.1257-1264,2007.

[23] J.L. Herlocker, J.A. Konstan, A. Borchers, and J. Riedl, "An Algorithmic Framework for Performing Collaborative Filtering," Proc. 22nd Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '99), pp. 230-237, 1999

[24] G. Linden, B. Smith, and J. York, "Amazon.com Recommendations:Item-to-Item Collaborative Filtering," IEEE Internet Computing, vol. 7, no. 1, pp. 76-80, Jan./Feb. 2003.

[25] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms," Proc. 10th Int'l Conf. World Wide Web (WWW '01), pp. 285-295, 2001.

[26] J. Wang, A.P. de Vries, and M.J. Reinders, "Unifying User-Based and Item-Based Collaborative Filtering Approaches by Similaritym Fusion," Proc. 29th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 501-508, 2006.

[27] G. Xue, C. Lin, Q. Yang, W. Xi, H. Zeng, Y. Yu, and Z. Chen, "Scalable Collaborative Filtering Using Cluster-Based Smoothing,"Proc. 28th Int'l ACM SIGIR Conf. Research and Development inInformation Retrieval (SIGIR '05), pp. 114-121, 2005.

[28] T. Hofmann, "Latent Semantic Models for Collaborative Filtering,"

AUTHOR'S PROFILE

Ms.Trishna J. Jadhav is post graduate student of computer engineering at K.K Wagh college of engineering ,Nashik under Savitribai Phule University of Pune.She completed her undergraduate course of engineering from Savitribai Phule University of Pune. Her areas of interest include Data Mining,Link Mining.

Prof Satish S. Banait completed his post-graduation from Government Engineering college Aurangabad,BAMU university Maharashtra. Presently he is working at K.K.Wagh college of engineering ,Nashik,Maharashtra,India as a Assistant professor .He has presented papers at National and International conferences and also published paper in national and international journals on various aspect of the computer engineering and Data mining and Data Warehouse .His research of interest include Data Warehouse,Software testing and quality assurance ,Software design and architecture,parallel computing.