Link Prediction in a Semi-bipartite Network for Recommendation

Aastha Nigam[1] and Nitesh V. Chawla[1,2(✉)]

[1] University of Notre Dame, Notre Dame, IN 46556, USA
{anigam,nchawla}@nd.edu
[2] Wroclaw University of Technology, Wroclaw, Poland

Abstract. There is an increasing trend amongst users to consume information from websites and social media. With the huge influx of content it becomes challenging for the consumers to navigate to topics or articles that interest them. Particularly in health care, the content consumed by a user is controlled by various factors such as demographics and lifestyle. In this paper, we use a semi-bipartite network model to capture the interactions between users and health topics that interest them. We use a supervised link prediction approach to recommend topics to users based on their past reading behavior and contextual data associated to a user such as demographics.

1 Introduction

Internet has become the biggest and the most preferred medium for consuming information. As a user we can discover information about a breadth of topics. But, with large amounts of information present across different mediums it becomes extremely laborious for users to find all the content pertinent to them. There are various systems that aim at predicting the rating or preference a user would give to an item. These predictions are based on the user's interaction with the item or similar items.

Reading habits of a user are driven by various factors such as interests, lifestyle, city they are residing in and their age. Particularly, the health content a user consumes is an outcome of many variables. Health content can be defined as information relating to one's health. It could be an article describing the symptoms of a disease or a recipe for healthier eating. Moreover, the articles the user reads may or may not be indicative of an illness they are suffering from or a problem pertaining to them. Given the diversity in the content and other above listed challenges, it becomes difficult to model the user's interest.

Capturing relevance of a particular article for a user becomes extremely granular, therefore we wanted to understand interests of users at a broader level using topics. Every article is associated with a topic or a theme which can be used to cluster them together and be used to understand user behavior. In this paper, we model the interests of users based on the articles they have previously shown interest in and then leveraging other user attributes particularly the city they reside in to predict other topics that would be relevant to them. We propose the

N.T. Nguyen et al. (Eds.): ACIIDS 2016, Part II, LNAI 9622, pp. 127–135, 2016.
DOI: 10.1007/978-3-662-49390-8_12

use of a semi-bipartite network to model this phenomena and identifying missing links in the networks to make recommendations at a user level.

We firstly describe research related to this work in Sect. 2 and then discuss the data used for our model development and validation in Sect. 3. We then present our methodology in Sect. 4 where we discuss the network structure and describe the approach used. Next, we present results obtained by applying our model on the data in Sect. 5. Lastly, we conclude with a discussion in Sect. 6.

2 Related Work

There has been a lot of work on link prediction in general. Liben et al. [1] presented a survey on various methods for link prediction in homogeneous networks where all the nodes are of the same type. They experimented with various measures such as Graph Distance, Common Neighbors, Jaccard's Coefficient, Adamic/Adar Score, Preferential Attachment, Katz, Hitting Time, Page Rank and Sim Rank for predicting new edges in a social network. Backstrom et al. [2] proposed an approach based on supervised random walks that combined information from both nodes and edges. However, many real world systems form complex networks with varied node and interactions types therefore, there has been work on link prediction in heterogeneous network. Davis et al. [3] proposed a supervised approach for link prediction in heterogeneous information networks where they used a modified Adamic Adar measure and compared its performance across various data-sets. They examined the measure on a disease-gene bipartite network. In this paper, we model the health care data using a semi-bipartite network which to the best of our knowledge has not been done before. We then try to understand topical preferences of users using link prediction.

3 Data Description

The data was provided by a digital media company, Everyday Health (EDH)[1] producing content related to health and wellness. EDH owns multiple companies addressing a varied set of topics such as pregnancy, diseases and healthy eating.

Being a web-based platform for health related information they are available across the globe. EDH allows users to sign up to their websites and select the topics they will be most interested in reading about. EDH also sends out weekly newsletters over the email to the users who have signed up and captures the user's reading behavior in terms of when they received the email, when they opened the newsletter, when they read the articles and which health topic category did the each article belong to. Along with health topic information, EDH also collects demographic details for each user such as city they reside in, gender and age group.

In this study, we utilize the data for Saint Joseph County, Indiana from June 2012 to June 2014. Saint Joseph County data comprises of 8 cities: Mishawaka, New Carlisle, South Bend, Osceola, Granger, Walkerton, Lakeville, and North Liberty.

[1] http://www.everydayhealth.com/.

4 Methodology

In this section, we present our method to construct the network and then illustrate the algorithm to perform link prediction.

4.1 Network Model

We propose a semi-bipartite network [4] to model our data for topic recommendation. A semi-bipartite graph can be defined as $G = (V_1, V_2, E_1, E_2)$ where V_1 and V_2 are the two set of nodes, E_1 denotes the edges between V_1 and V_2 whereas E_2 depict the edges (interactions) amongst the nodes in V_1. In our network, the two set of nodes are the user (V_1) and topic (V_2) nodes. As described earlier, the two ways to understand user's interests are firstly when he signs up at the website and selects the topics and secondly when he reads a particular article in the newsletter. Using these two sources, we get an aggregate of which topics the user is interested in. An edge (E_1) between the user and topic node signifies the user's interest. The other set of edges (E_2) in the network are amongst the users. Two users are connected if they come from the same city. This results in cliques of users registering from the same city. This might add noise but we want to leverage this demographic information using topological attributes.

The network was constructed following the above explained methodology. The network was undirected and unweighted. As a result, the network consisted of 4240 nodes and 2,780,874 edges as shown in Fig. 3. As can be seen from Fig. 3, users coming from the same city form cliques and all user cliques are connected to the topics. In the nodes set, 4076 were user nodes and 164 were topic nodes. Similarly, in the edge set, 33,269 edges were between users and topics and rest were between users. Figure 1 illustrates how the users are distributed across the 8 cities in Saint Joseph County. We also study how each city as a whole consumes the EDH content. Figure 2 captures the activity of each city.

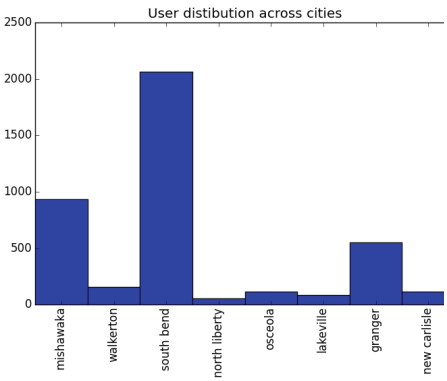

Fig. 1. Captures the user distribution across 8 cities in St. Joseph County.

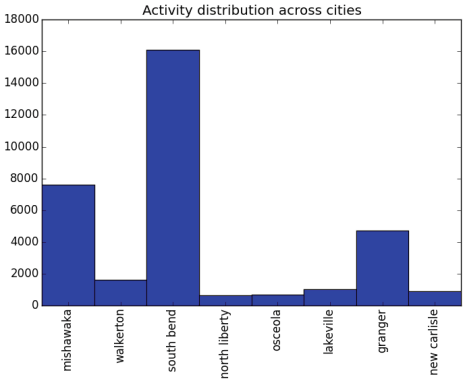

Fig. 2. Captures how active each city is in terms of consuming information from EDH.

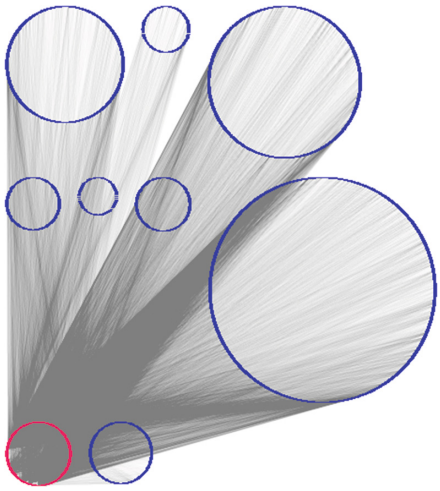

Fig. 3. Representation of the network where the blue nodes signify the user nodes and the red nodes signify the topic nodes. Only the interactions between the user and topics nodes are illustrated here. Since we have 8 cities we have 8 cliques of color blue all connected to the topic nodes (Color figure online).

Activity is calculated based on number of links clicked by the users in their respective cities. It can be seen from Figs. 1 and 2 that the overall activity of users in a city is correlated with the number of users in the city.

Table 1 lists the top 20 topics being consumed in each of the 8 cities. We see that weight management, diet and nutrition and exercise and fitness seem to be popular amongst most cities but articles related to depression seem to be consumed more in North Liberty. Similarly, diabetes seems to be a bigger concern in South Bend, Granger and Walkerton compared to other cities.

4.2 Topic Recommendation

Once the network was constructed, we deal with the problem of topic recommendation using link prediction. The network captures heterogeneous information and we only want to predict links between the user and topic nodes. We calculate various features for links (or nodes) of interest. The topological attributes that we consider can be broadly categorized into two categories of neighborhood methods and path methods. Neighborhood methods are: Common Neighbors, Jaccard's Coefficient, Adamic Adar and Preferential Attachment whereas path methods are: PageRank. For a node u in our network, we denote the set of direct neighbors as $\Gamma(u)$. Using this as the notation, we define the features as follows:

1. **Common Neighbors**: It captures the similarity between the two nodes by identifying the common nodes in their neighborhood [1]. Since, the network is semi-bipartite, the common neighbors can only be the user nodes at path length of 2. It can be calculated using Eq. 1.

Table 1. Top 20 topics being selected by the users across cities

Mishawaka	New carlisle	South bend	Osceola	Granger	Walkerton	Lakeville	North liberty
Weight management	Weight management	Weight management	Weight management	Weight management	Weight management	Weight management	Weight management
Diet and Nutrition	Exercise and Fitness	Diet and Nutrition	Exercise and Fitness	Diet and Nutrition	Diet and Nutrition	Exercise and Fitness	Depression
Exercise and Fitness	Diet and Nutrition	Exercise and Fitness	Diet and Nutrition	Exercise and Fitness	Exercise and Fitness	Diet and Nutrition	Diet and Nutrition
High blood pressure	High blood pressure	Diabetes	Depression	High blood pressure	Depression	High blood pressure	High blood pressure
Depression	Depression	High blood pressure	High blood pressure	Diabetes	Diabetes	Allergies	Exercise and Fitness
Diabetes	Allergies	Depression	Diabetes	Depression	High blood pressure	Diabetes	Diabetes
Heart disease	Diabetes	Heart disease	Allergies	Heart disease	Heart disease	High cholesterol	Arthritis
Allergies	Heart disease	Allergies	Heart disease	Allergies	Arthritis	Depression	Heart disease
High cholesterol	Arthritis	Arthritis	Sleep disorders	Beauty	Pain	Arthritis	High cholesterol
Arthritis	High cholesterol	High cholesterol	High cholesterol	High cholesterol	High cholesterol	Heart disease	Allergies
Anxiety	Menopause	Beauty	Arthritis	Menopause	Anxiety	Migraines	Anxiety
Sleep disorders	Anxiety	Sleep disorders	Anxiety	Arthritis	Allergies	Menopause	Beauty
Beauty	Sleep disorders	Anxiety	Menopause	Anxiety	Sleep disorders	Headache	Menopause
Pain	Beauty	Menopause	Pain	Sleep disorders	Beauty	Beauty	Pain
Menopause	Pain	Diabetes type 2	Beauty	Cancer	Menopause	Sleep disorders	Sleep disorders
Migraines	Cancer	Pain	Migraines	Digestive health	Digestive health	Pain	Digestive health
Diabetes type 2	Migraines	Sexual health	Diabetes type 2	Migraines	Migraines	Sexual health	Sexual health
Cancer	Skin conditions	Migraines	Headache	ADD/ADHD	Osteoarthritis	Anxiety	Osteoperosis
Smoking	Sexual health	Headache	Cancer	Sexual health	Diabetes type 2	Cancer	Cancer
Sexual health	Smoking	Digestive health	Sexual health	Diabetes type 2	Skin conditions	Digestive health	Headache

$$|\Gamma(u) \cap \Gamma(v)| \tag{1}$$

2. **Jaccard's Coefficient**: Number of common neighbors divided by the total combined number of neighbors of both nodes [1]. Instead of considering the raw number, it looks at the ratio of common nodes. It is given by Eq. 2.

$$\frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|} \tag{2}$$

3. **Adamic Adar**: It weights the impact of neighbor nodes inversely with respect to their total number of connections [5]. It is based on the assumption that rare relationships are more specific and have more impact on similarity. It can be calculated using Eq. 3.

$$\sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{log\Gamma(z)} \tag{3}$$

4. **Preferential Attachment**: It emphasizes on the number of neighbors a node has [6]. The higher the degree of a node, the more probable that a new node attaches to it. As shown in Eq. 4, it multiples the number of common neighbors.

$$|\Gamma(u) \times \Gamma(v)| \tag{4}$$

5. **PageRank**: The significance of a node in a network is based on the significance of other nodes that are linked to it. We take the product of the page rank scores for node u and v [7].

We then applied a supervised machine learning approach [8,9]. Each of the above listed attributes contribute to a feature vector. We take the presence of a link as class 1 and the absence of a link as class 0. Since we have a static network, to evaluate the performance of our model we divide the given data into train data and test data. Since there are fewer instances of links compared to absence of links we perform stratified sampling to divide our data set. This ensures that both train and test data sets follow the same class distribution as the original data set. As a result, we had 334,233 train instances which constituted of 317,598 samples from class 0 and 16,635 samples from class 1. Similarly, the test data had 334,231 instances of which 317,597 were class 0 and 16,634 were class 1. This was then evaluated using different machine learning algorithms.

5 Results

We firstly study the degree distributions over the user and topics nodes. The degree distribution of the user nodes was calculated using only the edges between the user and topic nodes. The user-user edges were not considered as they would add noise for the degree distribution. As can be seen from Figs. 4 and 5, they both follow a power law distribution indicating that there are more nodes with fewer links versus fewer nodes with more links. It essentially means that most of the users have presented a very sparse set topic choices whereas there are only a smaller number of users which have indicated their topic interests comprehensively. Similarly, from Fig. 5 we can say that there are lesser topics read by all users but most topics have fewer readers.

In our experiments, we applied three machine learning algorithms: Support Vector Machines (SVM) [10], decision trees [11] and logistic regression [12]. Area under the curve [13] and confusion matrix has been used as the evaluation metric. ROC curves for all three algorithms can be seen in Fig. 6. The area under the curve for SVM, decision trees and logistic regression is 0.5737, 0.5466 and 0.5401 respectively. We can see that SVM performs slightly better than the rest. In Fig. 7, we see we have high number of false positives and false negatives. Due to the inherent imbalance in the dataset, we see that the model is able to predict the absence of the links but due to few positive samples the model suffers.

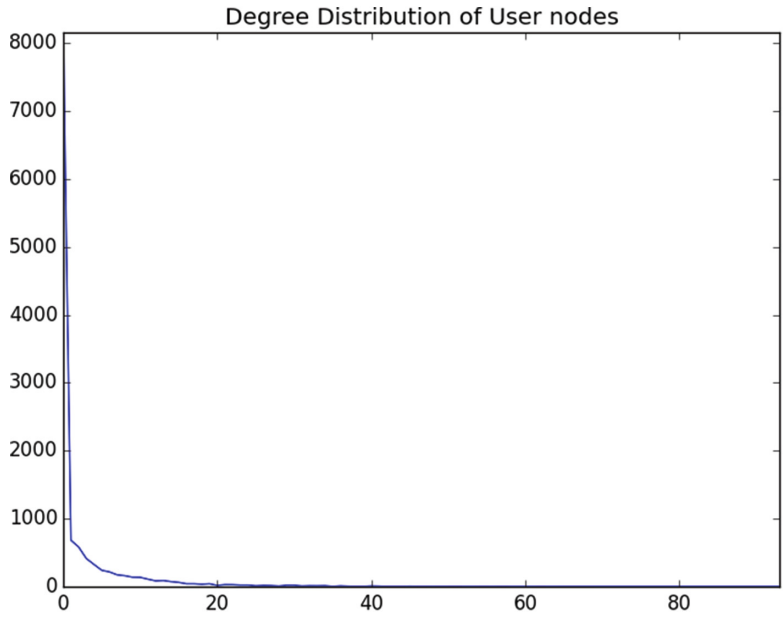

Fig. 4. Captures the degree distribution of user nodes with respect to topic nodes only.

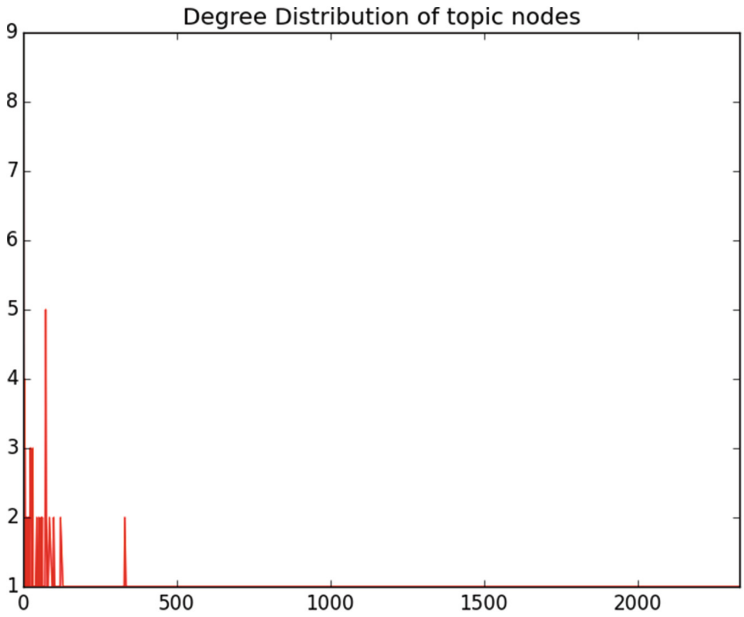

Fig. 5. Captures the degree distribution of topic nodes.

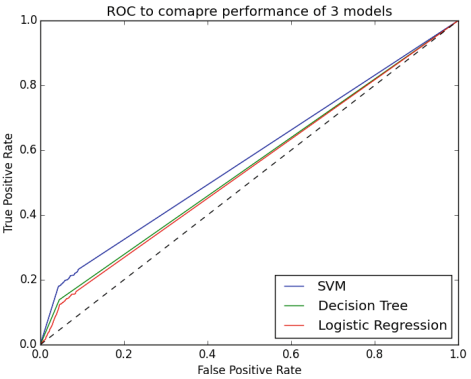

Fig. 6. ROC performance for SVM, decision tree and logistic regression.

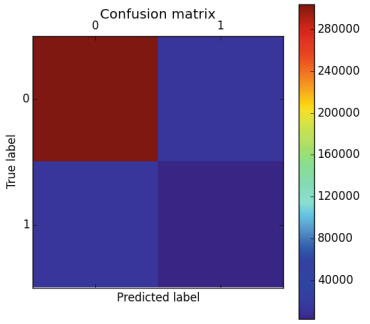

Fig. 7. Confusion matrix for SVM

6 Conclusion

In this work, we have proposed the construction of a semi-bipartite network from the user data. We approach the topic recommendation problem using link prediction in a supervised machine learning framework. We would like to include other features to incorporate the heterogeneity of the data. We evaluated various machine learning algorithms and found that SVMs was most effective on our dataset. It is challenging for a user to navigate through the entire website to find content relevant to him, therefore it becomes important for us to present the user with information he might be most interested in reading about. But many-a-times, data about user and his interests is very sparse. We have tried to capture the user's interest through demographic and reading habits.

As extension to this work we would like to incorporate more features and study the effect of each feature and analyze their effectiveness for predicting a new link in a semo-bipartite network.

Acknowledgements. We would like to thank Everyday Health for providing us their user data.

This research was supported in part by National Science Foundation (NSF) Grant OCI-1029584 and IIS-1447795.

References

1. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. J. Am. Soc. Inf. Sci. Technol. **58**(7), 1019–1031 (2007)
2. Backstrom, L., Leskovec, J., Supervised random walks: predicting and recommending links in social networks. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM 2011, pp. 635–644. ACM, New York, NY, USA (2011)

3. Davis, D., Lichtenwalter, R., Chawla, N.V.: Multi-relational link prediction in heterogeneous information networks. In: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011, pp. 281–288. IEEE Computer Society, Washington, DC, USA (2011)
4. Xu, K., Williams, R., Hong, S.-H., Liu, Q., Zhang, J.: Semi-bipartite graph visualization for gene ontology networks. In: Eppstein, D., Gansner, E.R. (eds.) GD 2009. LNCS, vol. 5849, pp. 244–255. Springer, Heidelberg (2010)
5. Adamic, L.A., Adar, E.: Friends and neighbors on the web. Soc. Netw. **25**, 211–230 (2001)
6. Barabasi, A.-L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)
7. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: bringing order to the web (1999)
8. Benchettara, N., Kanawati, R., Rouveirol, C.: Supervised machine learning applied to link prediction in bipartite social networks. In: Memon, N., Alhajj, R. (eds.) ASONAM, pp. 326–330. IEEE Computer Society (2010)
9. Kunegis, J., De Luca, E.W., Albayrak, S.: The link prediction problem in bipartite networks. In: Hüllermeier, E., Kruse, R., Hoffmann, F. (eds.) IPMU 2010. LNCS, vol. 6178, pp. 380–389. Springer, Heidelberg (2010)
10. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
11. Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**(1), 81–106 (1986)
12. Hosmer, D.W., Lemeshow, S.: Applied Logistic Regression (Wiley Series in Probability and Statistics), 2nd edn. Wiley-Interscience Publication, New York (2000)
13. Bradley, A.P.: The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recogn. **30**(7), 1145–1159 (1997)