Citation Network Analysis to Predict Citation Trends

Presented By:

Shreya Sunkam Ramaprasad
Department of Computer Science, UCLA
UID: 204946268

Advised By:

Prof. Yizhou Sun
Department of Computer Science, UCLA

Citation Network Analysis to Predict Citation Trends

1. Abstract

Citation Network analysis is the exploration of reference patterns in scholarly literature and has numerous applications in understanding research impact, knowledge flows, and knowledge networks. In this project, the problem of predicting the citation trend for a research paper has been explored. In particular, given graph snapshots of a citation network, we are interested in predicting the number of citations that the paper would receive over the years. The problem is approached as a sequence prediction problem where we are given time series data for a time period in the form of graph snapshots of the citation network and we are required to predict the citation sequence for the following years. This approach is unique in the sense that we do not make use of citation histories of the research paper to predict the number of citations in the future.

2. Literature Review

The problem of predicting the citation count for a given paper by modelling it as a link prediction problem has been a widely researched topic. [1] shows the statistical analysis of five link prediction models: Adamic/Adar, Weighted Common Neighbors, Jaccard's Coefficient and Preferential Attachment; using well-defined statistical measures, such as precision, accuracy, sensitivity and specificity on a Wikipedia citation network. [2] explores link prediction algorithms to predict potential interactions/friendships in social networks.

The incorporation of time-dependent information to enhance predictions has also gained considerable attention. [3] introduces the time-series link prediction problem, taking into consideration temporal evolutions of link occurrences to predict link occurrence probabilities at a particular time on high-energy particle physics literature co-authorship data.

Link prediction on a heterogenous citation network which is a hyper-graph consisting of paper citation network, author citation network and author collaboration network has been shown in [4]. [5] proposes a ranking factor graph model (RFG) for predicting links in social networks for a friend recommendation system.

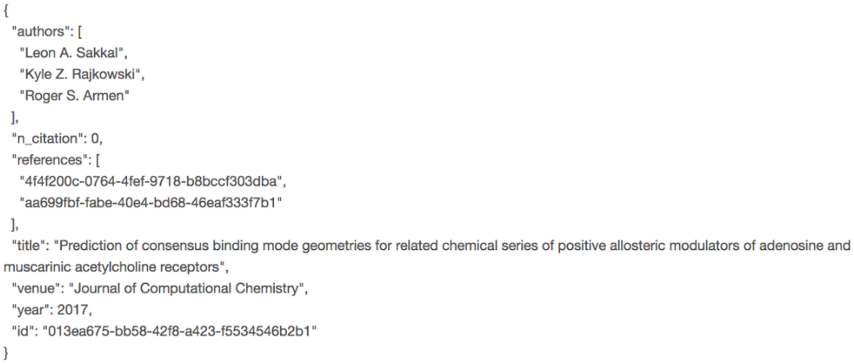
In recent years, deep learning models are being used for link prediction replacing models with hand engineered features. [6], [7] and [8] present the state-of-the-art graph embedding techniques Deepwalk, node2Vec and LINE respectively. These embeddings capture different latent factors of the nodes of the network corresponding to first order and second order proximity along with structural equivalence. The embeddings then serve as feature vectors for training a neural network. [11] evaluates several embedding techniques on a few common datasets and compares their performance against one another. [10] describes a general framework for incorporating temporal information into network embedding methods which is relevant to capture time dependencies in a fast-evolving network.

3. Implementation

3.1 Dataset

The citation network used for the analysis consists of data extracted from DBLP, ACM, MAG (Microsoft Academic Graph) and other sources for research purposes. The dataset consists of around 3 million papers and 25 million relationships.

In addition to the citation relationships, the dataset consists of additional data: name of the authors of the papers, the venue where the paper was presented, the year of publication and the title and abstract of the paper in json data format. Fig. 1 shows a snapshot of the json data structure for one such paper.



```
{
 "authors": [
  "Leon A. Sakkal",
  "Kyle Z. Rajkowski",
  "Roger S. Armen"
 ],
 "n_citation": 0,
 "references": [
  "4f4f200c-0764-4fef-9718-b8bccf303dba",
  "aa699fbf-fabe-40e4-bd68-46eaf333f7b1"
 ],
 "title": "Prediction of consensus binding mode geometries for related chemical series of positive allosteric modulators of adenosine and
muscarinic acetylcholine receptors",
 "venue": "Journal of Computational Chemistry",
 "year": 2017,
 "id": "013ea675-bb58-42f8-a423-f5534546b2b1"
}
```

Fig. 1: Json data structure for a paper in the Aminer Citation Network

Table 1 summarizes the number of papers, authors and venues in the dataset.

Number of papers	**3000000**
Number of authors	**1728168**
Number of venues	**5030**
Number of years	**82 (1937-2017)**

Table 1

- The papers in the dataset were sorted based on the year they were published.

The plot of the number of papers published as per the dataset for the years 1936-2018 is shown in Fig. 2 . It can be observed that more than 80% of the papers in the dataset were published during the years 2000-2017. Hence more attention was given to predict the citation counts for papers published in this period.

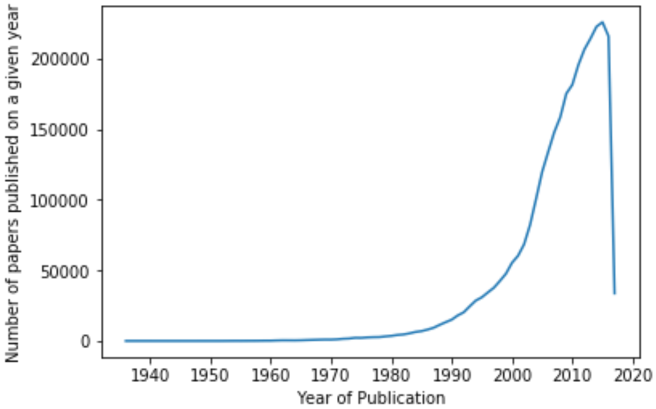


Fig. 2: Number of papers published across different years

- The dataset contains papers published in over 5000 different conferences. The top conferences are shown in Table 2. The venue information is missing for over 48000 papers in the dataset.
  A histogram of the number of papers published across different venue locations is shown in the Fig. 3.

Conference	**Number of Papers**
Lecture Notes in Computer Science	28128
International conference on acoustics, speech, and signal processing	26539
International conference on robotics and automation	19809
International conference on image processing	18333
International conference on communications	17642

Table 2

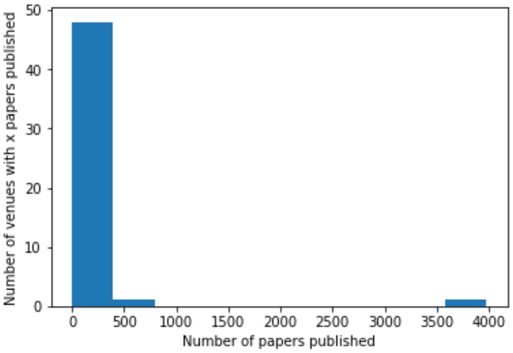


Fig. 3: Histogram of the number of papers published across different locations

- The dataset contains papers published by over 1 million authors.
  The top authors in the dataset are shown in Table 3. The high paper count for some of these authors is due to the fact that there are multiple authors publishing papers in top Computer Science conferences with the same name.

Author	Number of Papers
Wei Wang	2426
Wei Zhang	1599
Lei Zhang	1565
Yang Liu	1522

Table 3

3.2 Graph Creation

- A citation network was created for the papers in the dataset.For each paper 'P' in the dataset, a directed edge is drawn between papers 'P' and 'P1' , 'P2' …'Pn' where 'P1', 'P2' …'Pn' are the references of 'P'.

Snapshots of the network were stored for the years 2000, 2002, 2004, 2008, 2010, 2012, 2014, 2016 and 2018. The snapshot of the network at 2000 consists of all the citation relationships for papers published until and including 2000.

Figures 4 and 5 show the evolution of the graph in terms of the number of edges and number of nodes across graph snapshots.

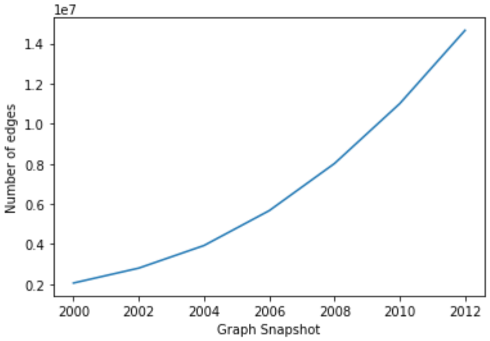


Fig. 4: Number of edges across graph snapshots.

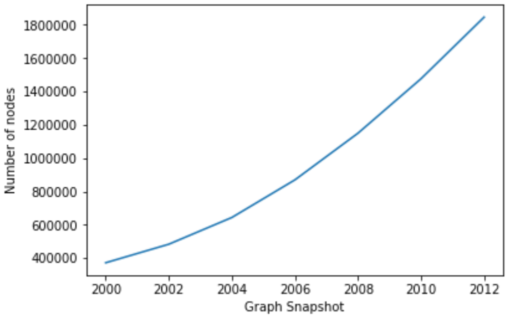


Fig. 5: Number of nodes across graph snapshots

- The degree distribution of the citation graph is shown in Fig. 6 . The degree distribution of a network provides an insight about the overall connectedness of the network. The degree distribution of the citation network indicates that the network is largely distributed.

The maximum number of citations received by a paper is 15850 and the minimum number of citations is 0.

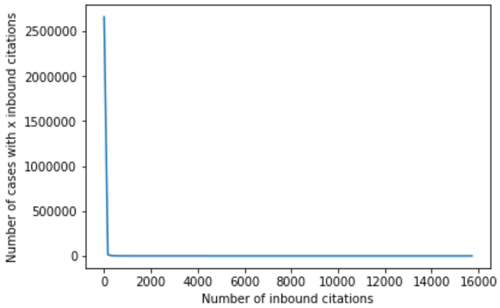


Fig. 6: Degree distribution of the citation network

- The average number of citations received by papers published in 2000,2002 and 2004 are shown in Figures 7a 7b and 7c respectively.

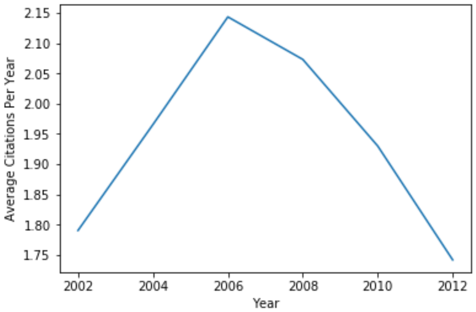


Figure 7a

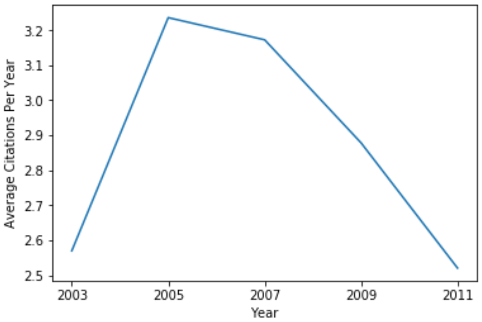


Figure 7b

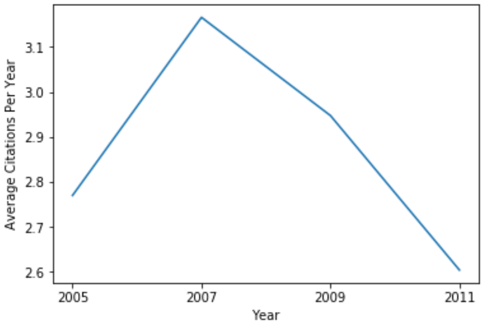


Figure 7c

We can observe that, **on an average citations follow an increasing trend for a few years after a paper is published and then follow a decreasing trend.** The problem addressed in this project is to learn such a trend by observing time series data in the form of embeddings of Graph snapshots. The idea is to predict the citation count of the papers for the coming years without using the citation history just by observing the evolution of the graph over time.

3.2 Graph Embeddings

The graph embedding technique used for obtaining the embeddings of the graph snapshots is called DeepWalk. DeepWalk uses local information obtained from truncated random walks to learn latent representations by treating walks as the equivalent of sentences. DeepWalk [6] preserves higher-order proximity between nodes by maximizing the probability of observing the last k nodes and the next k nodes in the random walk centered at
vi, i.e. maximizing $Pr(v_{i-k}, \ldots, v_{i-1}, v_{i+1}, \ldots, v_{i+k}|Y_i)$, where 2k + 1 is the length of the random walk. The model generates multiple random walks each of length 2k + 1 and performs the optimization over sum of log-likelihoods for each random walk. A dot-product based decoder is used to reconstruct the edges from the node embeddings.
The following parameter settings were applied for the DeepWalk algorithm to obtain the graph embeddings:

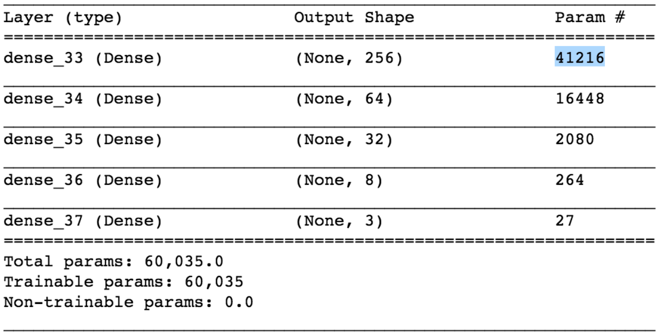
Parameter	Value
Window size	10
Embedding size	32
Walks per node	60
Walk length	60

3.3 Dataset Creation

The feature set for the modelling problem consists of graph embeddings of the citation network for an 8 year period. The labels for the corresponding features are the citation counts for the next 6 year period. The length of the input feature vector is 128 . (The embedding size * Number of graph snapshots = 32 * 4). The length of the label vector is 3.
An 80%-10%-10% split is used to split the dataset into training, validation and test set. The dataset consists of samples from different timeframes.

3.4 Model

A simple 5 layer neural network was used for modelling. The architecture of the neural network is given in the figure:



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
dense_33 (Dense)             (None, 256)               41216     
_________________________________________________________________
dense_34 (Dense)             (None, 64)                16448     
_________________________________________________________________
dense_35 (Dense)             (None, 32)                2080      
_________________________________________________________________
dense_36 (Dense)             (None, 8)                 264       
_________________________________________________________________
dense_37 (Dense)             (None, 3)                 27        
=================================================================
Total params: 60,035.0
Trainable params: 60,035
Non-trainable params: 0.0
_________________________________________________________________
```

The loss function used for optimization was **'Logarithmic Mean Squared Error'**. The logarithmic error loss function was used to mitigate the effect of certain papers which receive an unusually large number of citations.

The following parameters were used for training:

Parameter	**Value**
Batch Size	4096
Optimizer	adam
Number of epochs	100

3.5 Results

3.5 a Predicting the citation counts for the year 2010 using graphs from 2000-2008

In this experiment, graph embeddings of the citation network for the period 2000-2008 were used to train the network to predict the citation count for the year 2010.
The dimension of the input layer of the neural network was set to 160 and the dimension of the output layer to 1.

The learning curves while training the model are shown in Fig. 9.

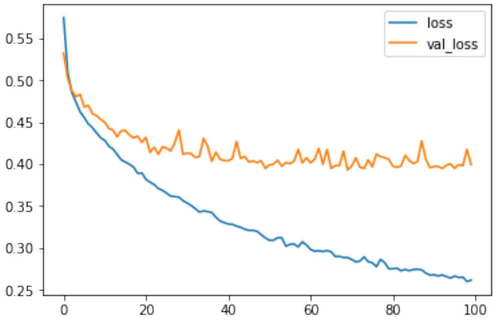


Fig. 9: Learning curves for predicting citation counts for the year 2010

The mean absolute error and mean squared logarithmic error are shown in the table below:

Metric	**Training Set**	**Test Set**
Mean Squared Logarithmic Error	0.26	0.4
Mean Absolute Error	1.57	1.9

3.5 b Predicting the citation counts for the years 2008-2012 using graphs from 2000-2006

In this experiment, graph embeddings of the citation network for the period 2000-2006 were used to train the network to predict the citation counts for the years 2008-2012.
The dimension of the input layer of the neural network was set to 128 and the dimension of the output layer to 3.

The learning curves while training the model are shown in Fig. 10.

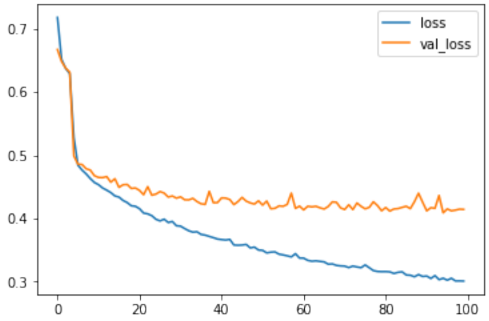


Fig. 10: Learning curves for predicting citation counts for the year 2008-2012

The mean absolute error and mean squared logarithmic error are shown in the table below:

Metric	Training Set	Test Set
Mean Squared Logarithmic Error	0.3012	0.414
Mean Absolute Error	1.56	1.84

3.5 c Predicting the citation counts for the papers published in 2008 (New Papers) for the years 2008-2012 using graphs from 2000-2006

In this experiment, graph embeddings of the citation network for the period 2000-2006 were used to train the network to predict the citation counts for the years 2008-2012.
Since each paper is a newly published paper and the citation network for the years 2008-2012 is not present in the citation network of graph 2000-2006, we take the average of the embeddings of the references of the paper for the corresponding years.
The dimension of the input layer of the neural network was set to 128 and the dimension of the output layer to 3.

The learning curves while training the model are shown in Fig. 11.

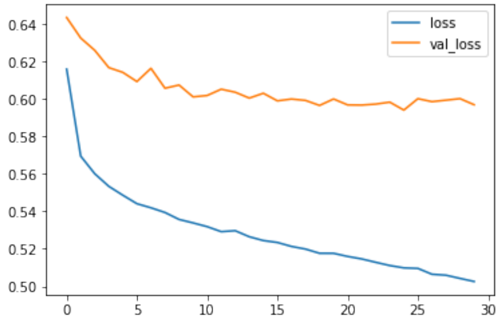


Fig. 11: Learning curves for predicting citation counts for the year 2008-2012

The mean absolute error and mean squared logarithmic error are shown in the table below:

Metric	Training Set	Test Set
Mean Squared Logarithmic Error	0.5	0.6
Mean Absolute Error	1.78	2.14

4. Future Work

- Explore the efficiency of other graph embedding techniques: LINE [7] and node2vec [8] These embeddings capture lower order proximity information in the graph embeddings in addition to higher order proximity information captured by DeepWalk.
- Incorporate author and venue information into the citation network to obtain a richer heterogenous citation network. Heterogeneous graph embedding techniques such as metapath2vec can be used to obtain the embeddings.

5. Acknowledgement

I would like to thank Prof. Yizhou Sun for giving me the opportunity to be involved with her research and giving me the problem statement for this project. I would also like to thank Song Jiang, her PhD student for guiding me at every step through this project.

6. Conclusion

In this project, the problem of predicting the citation trend of a research paper was explored as a sequence-in-sequence-out time series problem. The input sequence provided to the model is the sequence of graph embeddings at different snapshots. The output sequence is the citation counts for the following years.

The model developed performs well as compared to a baseline average citation count predictor but is still not able to properly capture the entire variance present in the dataset. Specifically, temporal information is not made use of in the training phase as the graph embeddings across different snapshots are supplied at once as a feature vector to a deep neural network. This shortcoming can be addressed by making use of a recurrent neural network which is more suitable for sequence training and prediction.

7. References

[1] Ferenc Molnar, Link Prediction Analysis in the Wikipedia Collaboration Graph
[2] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, J. Am. Soc. Inf. Sci., 58(7), 1019–1031, 2007.
[3] Z. Huang, D. Lin, The Time-Series Link Prediction Problem with Applications in Communication Surveillance, INFORMS Journal on Computing, 2008.
[4] Jingyu Cui,Fan Wang,Jinjian Zhai, Citation Networks as a Multi-layer Graph: Link Prediction and Importance Ranking
[5] Yuxiao Dong et. al. Link Prediction and Recommendation across Heterogeneous Social Networks, 2012 IEEE 12th International Conference
[6] Bryan Perozzi et, al. ,DeepWalk: Online Learning of Social Representations
[7] Aditya Grover,Jure Leskovec, node2vec: Scalable Feature Learning for Networks
[8] Jian Tang et. al. LINE: Large-scale Information Network Embedding
[9] Shiyu Chang et. al. Heterogeneous Network Embedding via Deep Architectures
[10] Giang Hoang Nguyen et. al. Continuous-Time Dynamic Network Embeddings
[11] Palash Goyal and Emilio Ferrara, Graph Embedding Techniques, Applications, and Performance: A Survey