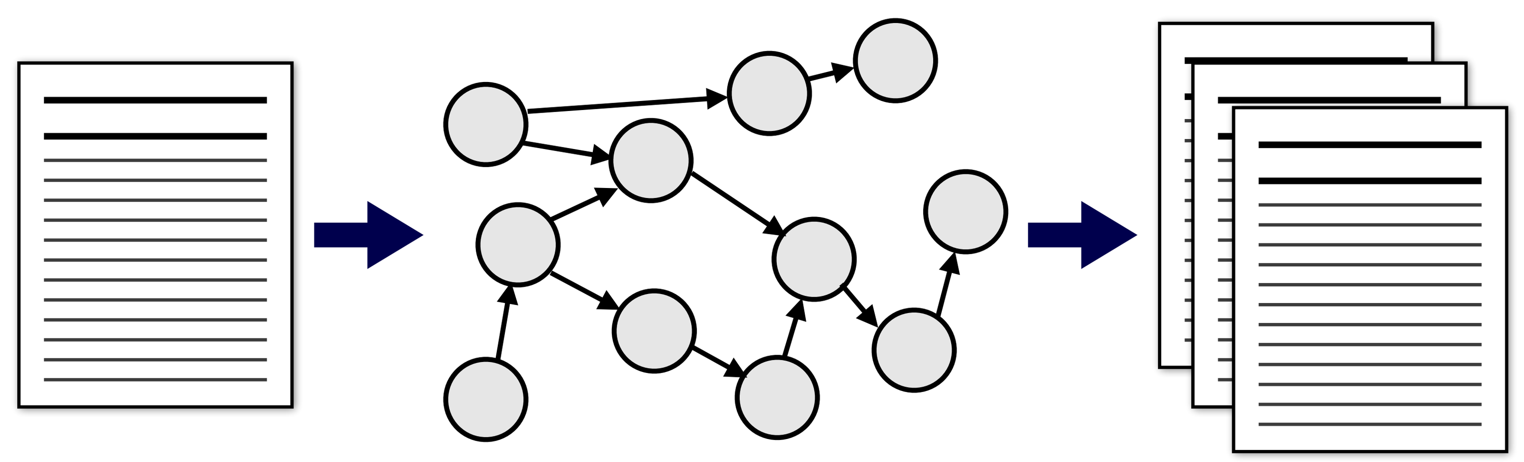
Pub2vec: A Recommender System for Similar Publications via Citation Network Embeddings

Stanford University

Brian K. Ryu (bryu@stanford.edu)

Motivation & Introduction

- Machine-learning based recommender systems have found myriad applications ranging from social networks to advertisements.[1]
- An algorithm for recommending similar scientific publications can be useful for researchers during literature searches.
- Features from a citation network of publications can be learned to recommend related publications.



- A citation network is directed acyclic graph where nodes represent publications and directed edges represent citation relations.
- Goal: Develop an algorithm trained on a citation network that receives a query publication and returns recommended similar publications.

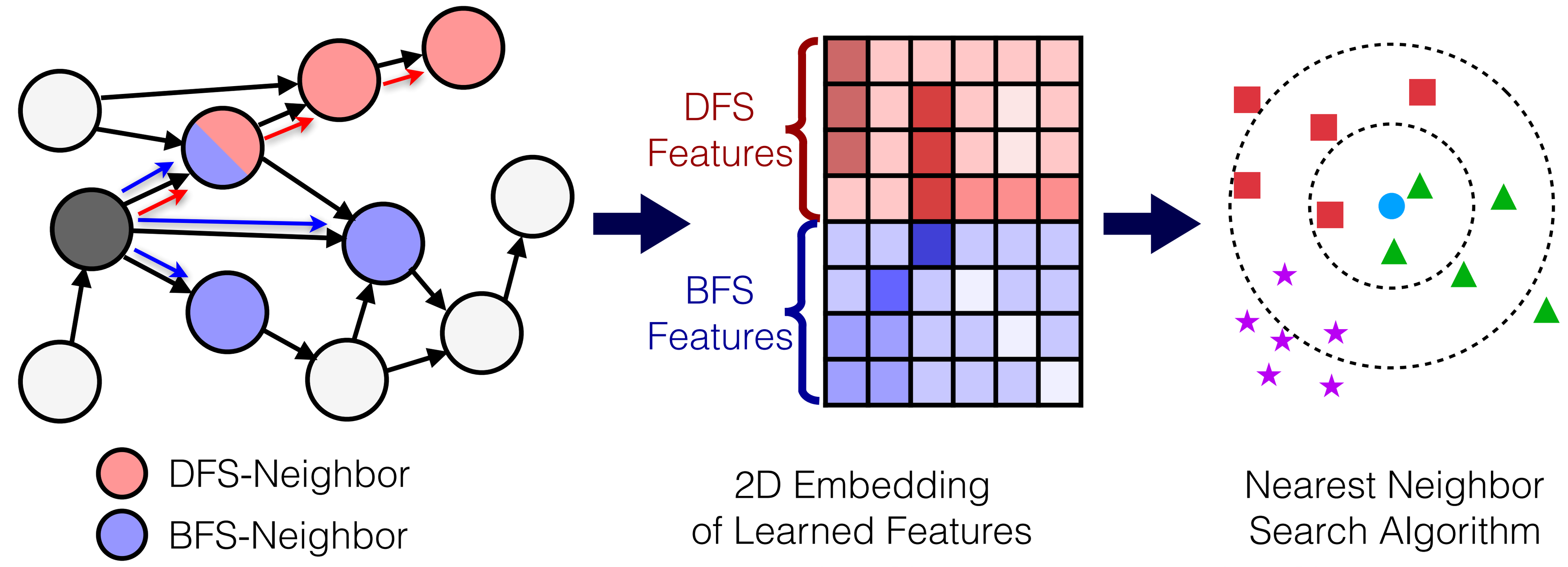
Data & Features

The full DBLP bibliographic dataset: Citation network dataset containing 4,107,340 scientific publications and 36,624,464 citation relationships. Dataset was acquired on May 5th 2019 by Arnetminer.[2]

- Dataset includes all journal publications, conference proceedings, and arXiv preprints in the computer science subject area.
- Dataset additionally contains weighted "field of study" feature vector (e.g. [("Web mining, 0.65"), ("Deep learning", 0.21") …]).

Model

Pub2vec Embedding Algorithm



Random walks, biased for breadth-first (BFS) or depth-first (DFS) walks, are performed to learn neighborhoods.[3]

- Features are learned via stochastic gradient that maximizes log-probability of observing a neighborhood $N(u)$ containing neighbors n_i of node u given its features $f(u)$:

$$\sum_{u \in V} \log Pr(N(u)|f(u)) \quad \text{where} \quad Pr(n_i(u)|f(u)) = \frac{\exp(f(n_i)\cdot f(u))}{\sum_{v \in V}\exp(f(v)\cdot f(u))}$$

Results

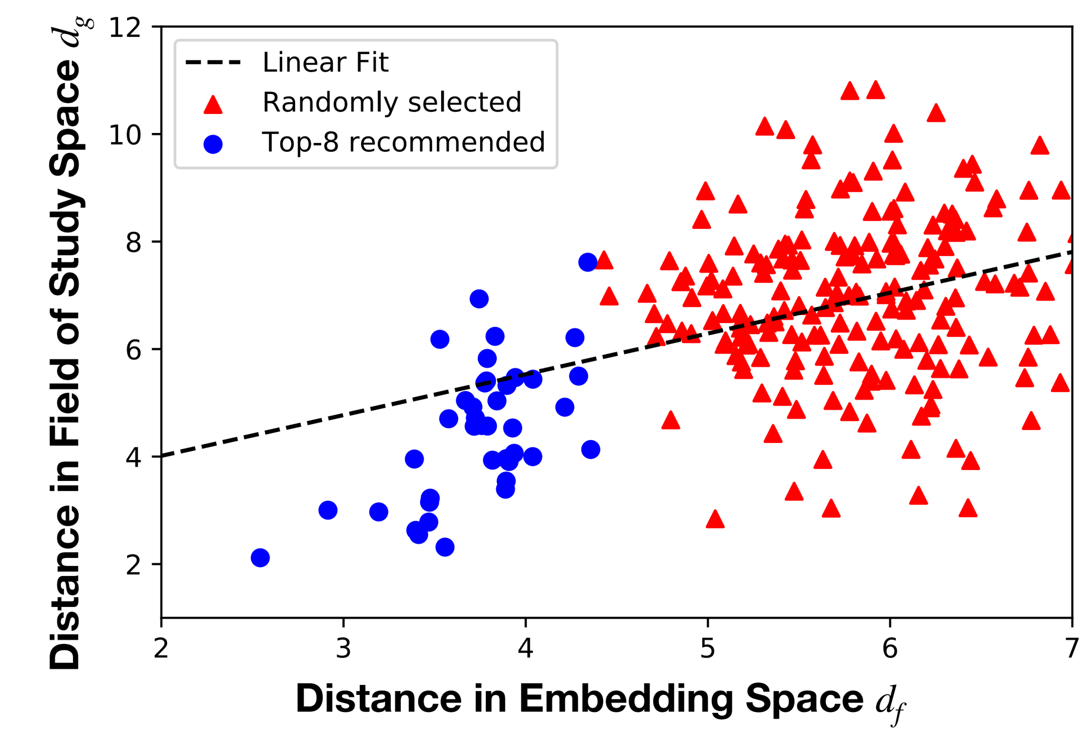
Publications Recommended from Queries

Input Publication	Number of Citations	Top-3 Similar Publications Recommended by *pub2vec*
Training Classifiers with Natural Language Explanations[4]	22	Feasibility study of stochastic streaming with 4K UHD video traces[5] Modeling single event crosstalk speedup in nanometer technologies[6] Cross lifecycle variability anal.: Utilizing requirements and testing artifacts[7]
Learning from untrusted data[8]	97	Fuzzy planar graphs[9] A 65nm std. cell set and flow dedicated to auto. async. circuits design[10] A new table of permutation codes[11]
Parsing with Compositional Vector Grammars[12]	808	Better word representations with RNN for morphology[13] Semi-supervised recursive autoencoders for predicting sentiment dist.[14] Dyn. pooling and unfolding recursive autoencoders for paraphrase detect.[15]
node2vec: Scalable feature learning for networks[3]	2,284	LINE: Large-scale information network embedding[16] A high-performance semi-supervised learning method for text chunking[17] Dependency tree-based sentiment classification using CRFs with hidden vars.[18]
k-means++: the advantages of careful seeding[19]	4,684	Clustering of the self-organizing map[20] Integrating constraints and metric learning in semi-supervised clustering[21] Data clustering: 50 years beyond K-means[22]
A formal basis for the heuristic determination of min cost paths (A*)[23]	9,026	Collision detection and avoidance in computer controlled manipulators[24] A mobile automation: An application of artificial intelligence techniques[25] Heuristics: intelligent search strategies for computer problem solving[26]
Going deeper with convolutions (GoogLeNet)[27]	17,646	Very deep convolutional networks for large-scale image recognition[28] Caffe: Convolutional architecture for fast feature embedding[29] Imagenet classification with deep convolutional neural networks[30]

Recommended publications are colored either as irrelevant (red) or relevant (blue)

Recommendations are more relevant for highly cited (≥ 800) publications due to abundance of "information" from citation relations.

Comparison with Field of Study Features



L_2 distance in field of study space, d_g, vs. embedding space, d_f, as quantitative metric shows positive correlation.

Discussion and Conclusion

- Successfully developed a recommender system for academic publications based on citation networks.
- The model generated pertinent recommendations for sufficiently well-cited articles.
- True performance evaluation is challenging without user feedback.

Future Work

- A/B testing or obtain user feedback to properly evaluate performance for practical use.
- Further tune random walk parameters to obtain optimally learned feature embeddings.

References

[1] Resnick, Paul, and Hal R. Varian. Commun. ACM 40.3 (1997): 56-59.
[2] Tang, Jie, et al. ACM SIGKDD, 2008.
[3] Grover, Aditya, and Jure Leskovec. ACM SIGKDD, 2016.
[4] Hancock, Braden, et al. Proceedings of the conference. ACL, 2018.
[5] Kim, Joongheon, and Eun-Seok Ryu. 2015 IEEE ICTC, 2015.
[6] Sayil, Selahattin, and Li Yuan. Microelectron J. 46.5 (2015): 343-350.
[7] Steinberger, Michal, Iris Reinhartz-Berger, and Amir Tomer. J. Syst. Softw. 143 (2018): 208-230.
[8] Charikar, Moses, Jacob Steinhardt, and Gregory Valiant. ACM STOC 2017.
[9] Samanta, Sovan, and Madhumangal Pal. IEEE Trans. Fuzzy Syst. 23.6 (2015): 1936-1942.
[10] Moreira, Matheus, et al. 2011 IEEE SOCC, 2011.
[11] Smith, Derek H., and Roberto Montemanni. Des. Codes Cryptogr. 63.2 (2012): 241-253.
[12] Socher, Richard, et al. ACL, 2013.
[13] Luong, Thang, Richard Socher, and Christopher Manning. ACL SIGNLL. 2013.
[14] Socher, Richard, et al. ACM EMNLP, 2011.
[15] Socher, Richard, et al. NeurIPS. 2011.
[16] Tang, Jian, et al. ACM WWW, 2015.
[17] Ando, Rie Kubota, and Tong Zhang. ACL, 2005.
[18] Nakagawa, Tetsuji, Kentaro Inui, and Sadao Kurohashi. NAACL-HLT, 2010.
[19] Arthur, David, and Sergei Vassilvitskii. ACM-SIAM, 2007.
[20] Vesanto, Juha, and Esa Alhoniemi. IEEE T. Neur. Net. Lear. 11.3 (2000): 586-600.
[21] Bilenko, Mikhail, Sugato Basu, and Raymond J. Mooney. ACM ICML, 2004.
[22] Jain, Anil K. Pattern recognition letters 31.8 (2010): 651-666.
[23] Hart, Peter E., Nils J. Nilsson, and Bertram Raphael. IEEE T. Syst. Man. Cy. C. 4.2 (1968): 100-107.
[24] Udupa, Shriram Mahabal. Diss. Caltech, 1977.
[25] Nilson, N. J. IJCAI. 1969.
[26] Pearl, Judea. (1984).
[27] Szegedy, Christian, et al. IEEE CVPR, 2015.
[28] Simonyan, Karen, and Andrew Zisserman. arXiv:1409.1556 (2014).
[29] Jia, Yangqing, et al. Proceedings of the 22nd ACM MM, 2014.
[30] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. NeurIPS. 2012.