Hoang Nguyen - 15M54097

MONTHLY REPORT - OCT 2016

1. Research activities in August and September.
2. Research plan in October.
3. Problems in research (if any).

Collaborators: Choong Jun Jin, Kaushalya, Nukui, Sunil.

1 Research activities in August and September

	AUGUST - 2016	SEPTEMBER - 2016
TOPIC	Network embedding and motifs in network.	Vandalism detection and knowledge graph.
IDEA	Use the graph's motif as a guide for the random walk in network embedding. The biased random walked generated by my model emphasizes on the local motif community of the network. Besides the motif walk, I also proposed the inverse motif walk to discover not-motif-community to use as negative samples.	Detect vandalism in Wikidata by neural random forest, or train a specialized neural network to detect edge-case vandalism output by the traditional random forest model. For triple scoring, I mined the Google's rank score for each person's name using Google Knowledge Graph API. By using the aforementioned Google's score and the similarity score learned by Skipgram model from Wikidata's text corpus, we train a simple feed-forward neural network to classify the popularity of each name-job or name-country on a scale of 0 to 7.

ACTIVITIES	I have conducted extensive experiments on BlogCatalog3 dataset (undirected triangle motif), and some premilinary experiments on other datasets (bipartite and larger datasets). My model (named MAGE) shows superiority compared to previous models. The reason for such performance lies at the number and quality of the negative samples obtained. I also wrote a paper and submitted to AAAI'17. The result for AAAI'17 will be available on December 2016.	I have assembled a team of 4 (Nukui-san dropped out due to his own project) to participate in WSDM Cup 2017. This year WSDM Cup 2017 consists of 2 task: Vandalism detection on Wikidata and Triple scoring. Details are given in the reference.
REFERENCES	Deepwalk [1]; Planetoid [2]; LINE [3]; node2vec [4]; MAGE source code [5].	WSDM Cup 2017 [6]; Vandalism detection [7]; Triple scoring [8].

2 Research plan in October

	OCTOBER - 2016
TOPIC	Rare event detection, knowledge graph, and submodular models on graph.
PLAN	I will continue to develop my ideas for the WSDM Cup 2017 as mentioned in the previous seciton. October 15th, we will submit our first prototype of both systems (vandalism detection and triple scoring). On the other hand, I am studying about set theory and graphical submodular models because I want to conduct a concrete mathematics proof for my motif walk model, which was submitted to AAAI'17 last month. Furthermore, I have great interest in submodularity and random processes, I plan to write my Master thesis on this topic.
REFERENCES	Submodularity [9]; Submodularity in graph [10]; Matroid [11].

3 Problems

Currently I do not have any serious problem in my research. However, There is a few minor problems that I will be grateful if you share some of your comment if possible.

- Rare case prediction with machine learning. Currently, the vandalism takes about 7% of all edit on Wikidata. The simple ZeroR model can easily achieve 93% accuracy by predicting all edit as "valid". We have choosen the Random Forest (and its variant XGBoost) for this problem. However, we want to use some deep architecture (which performs very bad in this task). I wonder if you can give us some comment on this matter.

- Submodularity in set theory. I have a small wonder about how do you think about the future of this branch of discrete mathematics. I want to write my master thesis on the amount of information (Fisher Information) gathered using some random process on a network. Do you think it is possible for me to pursue such mathematical topic? Please tell me if you prefer my master thesis to be more practical-oriented.

Murata-sensei, thank you very much for your time!

References

[1] Perozzi, Bryan and Al-Rfou, Rami and Skiena, Steven. *Deepwalk: Online learning of social representations*. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014.

[2] Zhilin Yang and William W. Cohen and Ruslan Salakhutdinov. *Revisiting Semi-Supervised Learning with Graph Embeddings*. Proceedings of the 33rd International Conference on Machine Learning, 2016.

[3] Tang, Jian and Qu, Meng and Wang, Mingzhe and Zhang, Ming and Yan, Jun and Mei, Qiaozhu. *Line: Large-scale information network embedding*. Proceedings of the 24th International Conference on World Wide Web, 2015.

[4] Grover, Aditya and Leskovec, Jure. *node2vec: Scalable Feature Learning for Networks*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[5] Nguyen, Hoang and Nukui, Shun and Murata, Tsuyoshi. *https://github.com/anonsyuushi/mage*. Anonymous code submitted to AAAI'17.

[6] WSDM Cup 2017. *http://www.wsdm-cup-2017.org/*. Homepage for WSDM Cup 2017 containing two tasks: Vandalism dectection and Triple scoring.

[7] Stefan Heindorf, Martin Potthast, Benno Stein, and Gregor Engels. *Vandalism Detection in Wikidata*. In Proceedings of the 25th ACM International Conference on Information and Knowledge Management, 2016.

[8] Hannah Bast, Bjorn Buchhold, and Elmar HauBmann. *Relevance Scores for Triples from Type-Like Relations*. In SIGIR, 2015.

[9] Krause, Andreas, and Daniel Golovin. *Submodular function maximization.*. Practical Approaches to Hard Problems 3.19 (2012): 8.

[10] Frank, Andras. *Submodular functions in graph theory*. Discrete Mathematics 111.1 (1993): 231-243.

[11] Oxley, James G. *Matroid theory*. Vol 3. Oxford University Press, USA, 2006.