VERTEXVOYAGE: DISTRIBUIRANI SISTEM ZA EMBEDOVANJE čVOROVA U REALNIM MREžAMA

STEFAN NOŽINIĆ
STUDENT II GODINE MASTER STUDIJA RAČUNARSKIH NAUKA
DEPARTMAN ZA MATEMATIKU I INFORMATIKU
PRIRODNO-MATEMATIČKI FAKULTET
UNIVERZITET U NOVOM SADU.
TIP RADA: MASTER RAD
MENTOR: prof. dr MILOŠ SAVIĆ

Sažetak. Ovaj dokument sadrži predlog istraživačkog projekta kao završnog rada na master studijama. Istraživanje bi imalo za cilj implementaciju distribuiranog algoritma za embedovanje čvorova u realnim mrežama gde je broj čvorova moguće skalirati na ogromne skale.

Stefan Nožinić, II godina master studija računarskih nauka
Departman za matematiku i informatiku, Prirodno-matematički fakultet
Univerzitet u Novom Sadu.

---

1. Uvod i opis problema

Dosta istraživanja je urađeno na temu embedovanja čvorova u realnim mrežama sa ciljem dalje upotrebe u algoritmima mašinskog učenja i klasterisanja. [1].

Realne mreže su grafovi koji imaju specifična svojstva takva da oblikuju ponašanje mreža koje se mogu naći u prurodi. Ta svojstva se ogledaju pre svega u tome da takvi grafovi imaju jednu gigantsku komponentu, postojanje habova (čvorova sa visokim stepenom) i postojanjem fenomena malog sveta odnosno svojstva da je prosečna distanca između bilo koja dva čvora mala (često manja od 10). Takođe, izuzetno bitno svojstvo je i postojanje zajednica - delova grafa gde su čvorovi jako povezani međusobno, a slabo povezani van zajednice.

Jedan od najkorišćenijih algoritama za embedovanje čvorova u realnim mrežama jeste node2vec [1]. Postoji dosta verzija paralelne implementacije ovog algoritma kao npr [2], [3]. Node2vec algoritam je baziran na principu po kom funkcioniše word2vec [4]. Algoritam generiše sekvence čvorova pokretanjem nasumične šetnje po grafu, a zatim u drugoj fazi uz pomoć izgenerisanih sekvenci obučava neuronsku mrežu čija je arhitektura jedan skriveni sloj i softmax sloj na izlaznom delu. Cilj obučavanja mreže je maksimizacija verovatnoće predviđanja susednog čvora tokom nasumične šetnje za dati ulazni čvor.

Kako bi se ceo proces paralelizovao potrebno je uraditi paralelizaciju obe faze. Za paralelizaciju nasumične šetnje potrebno je uraditi particionisanje grafa. Neki od pokušaja particionisanja su opisani u [2]. Paralelizacija obučavanja neuronske mreže je rađena u [5] i [6].

U ovom istraživanju biće ispitani različiti modeli particionisanja realnih mreža kao i različiti pristupi obučavanju neuronske mreže.

2. Metode

Metodologija koja će biti sprovedena je eksperimentalno istraživanje. Varijable koje će biti kontrolisane su:

- Metod treniranja neuronske mreže
- Metod particionisanja realnih mreža

Za potrebe istraživanja, biće izmereni sledeći parametri:

- Skaliranje - odnos vremena potrebnog za embedovanje na jednom i na p procesora
- Količina utrošenog vremena na komunikaciju između procesora

Kako bi se verifikovala implementacija, izračunaće se korelacija između klastera dobijenih sa K-means klasterisanjem na embedovanju dobijenom upotrebom standardnog node2vec algoritma i embedovanju dobijenom paralelne implementacije. Ovde je očekivanje da postoji pozitivna korelacije između klastera.

2.1. **Ulazni podaci.**

Algoritam će biti testiran na Zaharijevoj mreži [7] i na većim mrežama kao što je [8].

Pored postojećih mreža, kao ulazni podaci će biti prosleđeni i veštački izgenerisane mreže uz pomoć stohastičkog blokovskog modela sa različitim parametrima [9].

Stefan Nožinić, II godina master studija računarskih nauka
Departman za matematiku i informatiku, Prirodno-matematički fakultet
Univerzitet u Novom Sadu.

Literatura

[1] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[2] Gianfranco Lombardo and Agostino Poggi. A scalable and distributed actor-based version of the node2vec algorithm. In *Proceedings of the 20th Workshop From Objects to Agents*, pages 134–141, 2019.

[3] Peng Fang, Arijit Khan, Siqiang Luo, Fang Wang, Dan Feng, Zhenli Li, Wei Yin, and Yuchao Cao. Distributed graph embedding with information-oriented random walks. *arXiv preprint arXiv:2303.15702*, 2023.

[4] Kenneth Ward Church. Word2vec. *Natural Language Engineering*, 23(1):155–162, 2017.

[5] Otkrist Gupta and Ramesh Raskar. Distributed learning of deep neural network over multiple agents. *Journal of Network and Computer Applications*, 116:1–8, 2018.

[6] Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert Ormandi, George E Dahl, and Geoffrey E Hinton. Large scale distributed neural network training through online distillation. *arXiv preprint arXiv:1804.03235*, 2018.

[7] Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977.

[8] Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

[9] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.