II INTERNATIONAL SCIENTIFIC CONVENTION
"II ICC UCLV 2019"

JUNE 23th – 30th, 2019
CAYOS DE VILLA CLARA. CUBA.

INTERNATIONAL WORKSHOP ON INTERNET OF THINGS AND ARTIFICIAL INTELLIGENCE (IoTAI2019)

Parallel approach for network construction from large purchasing collections

Ivett Fuentes[1,2], Gonzalo Nápoles[2], Leticia Arco[1,3] and Koen Vanhoof[2]

1-Central University of Las Villas, Cuba. ivett@uclv.cu

2-Hasselt University, Belgium. gonzalo.napoles@uhasselt.be

3-Vrije Universiteit Brussel, Belgium. larcogar@vub.be

Abstract: When dealing with problems arising from applications such as customer purchasing interactions, building the network for a dataset comprised of millions of transactions will lead to some computational issues. In this paper, we tackle that computation challenge by using a parallel approach for network representations in presence of massive amount of purchasing data. Numerical simulations using a real-world problem show the advantages of the proposed parallel solution.

Keywords: Complex network construction; Community detection; Parallel computing; Customer purchasing data.

1. Introduction

Complex networks have emerged as a unified representation of complex systems. The process of detecting communities is relevant in many disciplines where systems are often represented as networks and individuals are connected to one other by considering the influence as weight. For example, Customer Purchasing Behaviors (CPBs) are modeled as a complex network from millions of transactions, which describe the customer bags [1]. Dealing with problems arising from real-world scenarios regularly leads to some computational issues as they involve big transaction datasets, which continue to grow. Consequently, there is a pressing need to handle large similarity matrices spread across

many machines. In this paper, we aim to cope the high computational complexity related with the network construction phase by using parallel computing. The basic idea of the methodology for detecting community consists in distributing the calculation of the similarity matrix, which is deemed an extremely time-consuming process when the number of transactions increases.

2. Parallel network construction

Although community detection algorithms are capable of handling large datasets, this does not imply that there is no limit. Aiming at coping the high computational complexity of the network construction phase, we rely on parallel computing. Our overall idea consists in distributing the calculation of the similarity matrix, so each station executes a certain number of cells in the matrix. Figure 1 displays the parallel implementation proposed in this paper.

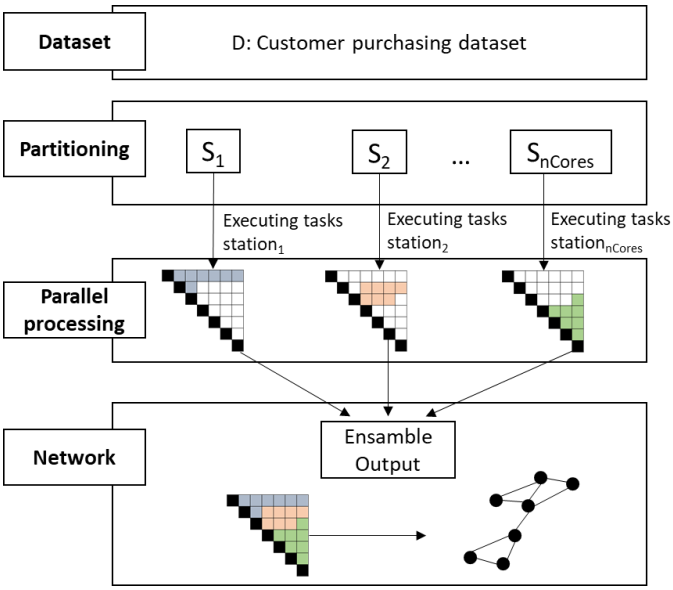

Figure 1. Parallel scheme for computing the similarity matrix.

To remove weak edges created by low fortuitous purchases and improve the quality of the subsequent analysis and the convergence of Community Detection (CD) algorithms, we establish a minimum threshold that can be tuned during the threshold estimation step. Instead of choosing a single alternative, we study the communities that result from networks using different threshold values.

A dataset provided by an anonymous Belgian company was used for evaluating the parallel processing approach for the CPB analysis. The experiments were carried out using a cluster environment powered by the Flemish Supercomputer Centre[1] by using one machine with 20 kernels and two machines with 20 kernels each one. Five CD algorithms[2] are executed by using the parameters suggested in *igraph* R-module. They are computed 10 times on each network with different random seeds. The results show that the CD algorithms reach a better performance (i.e., reach a better modularity value [2]) when deleting weak interactions. As expected, the lowest modularity values are obtained when detecting communities in networks using the thresholds less restrictive (i.e., network with weak interactions) regardless of the similarity function.

3. Conclusions

In this paper we have presented a parallel processing approach for dealing with the high computational complexity of customer network construction, which is a key step in community detection problems. The main idea consists in distributing the calculation of the similarity matrix. This approach was applied on real-life data concerning to a Belgian online shop. Our solution comprises the following advantages: 1) it allows obtaining different segmentation results using multiple instance similarity functions to compare customers, 2) it allows considering different thresholds for improving the convergence of community detection algorithms. The future work will be oriented to obtain a solution of our parallel approach based on the Spark platform.

4. Bibliographical references

[1] Fuentes, I., Nápoles, G., Arco, L., Vanhoof, K.: Customer segmentation using multiple instance clustering and purchasing behaviors. Lecture Notes in Computer Science series. vol. 11047. Springer (2018)

[2] Leskovec, J., Lang, K.J., Mahoney, M.: Empirical comparison of algorithms for network community detection. In: Proceedings of the 19th international conference on World wide web. pp. 631-640. ACM (2010)

[1] https://www.vscentrum.be
[2] Multi-level Modularity Optimization (LV), Walktrap (WT), Label Propagation (LP), Infomap (IM) and Fast Greedy Modularity Optimization (FGO) [9]