

ABSTRAK

ALGORITMA FACETNET PARALEL MENGGUNAKAN TEKNIK VERTEX CUT SEBAGAI METODA UNTUK MENDETEKSI EVOLUSI KOMUNITAS PADA GRAF BESAR

Oleh

Imelda Atastina

NIM: 33214013

(Program Studi Teknik Elektro dan Informatika)

Model evolusi komunitas adalah model yang dapat digunakan untuk memahami dinamika interaksi berbagai komunitas dari waktu ke waktu. Model ini dapat diperoleh dengan melakukan deteksi evolusi komunitas pada graf. Sementara itu, perkembangan teknologi informasi telah mengakibatkan berbagai organisasi dapat menyimpan data dalam jumlah yang sangat besar, yang mengakibatkan munculnya fenomena graf berskala besar. Sejauh ini metoda yang sudah dikembangkan untuk mendeteksi evolusi komunitas baru mampu memproses graf dengan ukuran sekitar maksimal sekitar satu juta simpul. Oleh karena itu dibutuhkan metode deteksi evolusi komunitas untuk graf berskala besar. Penelitian-penelitian yang berkaitan dengan graf skala besar menunjukkan bahwa cara terbaik untuk memproses graf yang sangat besar adalah dengan memanfaatkan sistem terdistribusi, karena pada umumnya graf yang sangat besar membutuhkan memori yang sangat besar pula dan umumnya tidak mampu diproses dengan komputer tunggal dengan kapasitas memori yang tentunya terbatas pula. Berdasarkan studi awal penelitian ini, algoritma Facetnet adalah salah satu algoritma deteksi evolusi komunitas yang berpotensi untuk dikembangkan, karena memiliki kompleksitas algoritma linier sehingga dianggap cocok untuk memproses graf berskala besar, dan mampu menghasilkan model evolusi yang lengkap. Sehingga dipandang sangat bermanfaat untuk mengembangkan algoritma Facetnet ini agar dapat memproses graf yang sangat besar menggunakan sistem terdistribusi.

Untuk dapat memproses graf berskala besar menggunakan sistem terdistribusi ada beberapa komponen yang harus dipenuhi yaitu, teknik untuk mempartisi graf serta algoritma yang dapat dijalankan secara paralel. Teknik mempartisi graf diperlukan, agar terbentuk sub graf yang lebih kecil sehingga dapat diproses secara paralel pada masing-masing komputer yang menjadi bagian dari sebuah sistem terdistribusi. Penelitian ini telah menganalisis teknik-teknik pemecahan graf dan menemukan teknik *Vertex Cut* adalah teknik yang tepat untuk dikombinasikan dengan algoritma Facetnet yang dimodifikasi skema iterasinya, sehingga menghasilkan algoritma Facetnet Paralel yang dapat dijalankan pada sistem terdistribusi. Namun dengan hanya memecah graf yang besar menjadi subgraf dan mengkombinasikan dengan algoritma Facetnet Paralel tidak cukup menjadi sebuah metoda yang mampu memproses graf yang sangat besar karena adanya matriks yang menyimpan nilai peluang keanggotaan sebuah simpul pada kluster yang ternyata juga berukuran sangat besar dan tidak dapat disimpan pada memori sebuah komputer saja. Penelitian ini juga telah berhasil menemukan representasi graf dengan simpul beratribut yang mampu mengatasi masalah ini.

Dengan struktur graf yang memiliki simpul beratribut algoritma Facetnet Paralel tidak lagi perlu mendefinisikan atau menyimpan matriks ketetanggaan maupun matriks keanggotaan simpul pada klaster secara terpisah. Kedua matriks tersebut sudah dapat digantikan dengan subgraf-subgraf dengan simpul beratribut yang dapat disimpan secara paralel, sekaligus diproses secara paralel oleh algoritma Facetnet Paralel menggunakan memori lokal pada masing-masing komputer pada sebuah sistem terdistribusi. Dengan demikian metoda deteksi evolusi komunitas menggunakan ketiga unsur penting tersebut mampu menghasilkan model evolusi komunitas untuk graf berskala besar. Keterbatasan memori pada sebuah komputer untuk memproses graf yang besar dapat diatasi dengan menambah komputer saja, tanpa perlu mengubah logika proses yang didefinisikan pada algoritma Facetnet Paralel. Sehingga untuk metoda ini diprediksi mampu memproses graf dengan jumlah simpul yang jauh lebih besar dari jutaan simpul bahkan sampai milyaran simpul.

Penelitian ini juga menemukan bahwa Facetnet memiliki kelemahan, yaitu menentukan jumlah komunitas sebagai nilai tebakan di awal iterasi, sehingga menghasilkan model evolusi dengan jumlah komunitas yang sama dari waktu ke waktu. Hal ini tentunya kurang sesuai dengan kondisi riil yang memungkinkan perbedaan jumlah komunitas yang terkandung pada sebuah graf dari waktu ke waktu. Oleh karena itu telah dicoba untuk menghasilkan metoda penentuan banyaknya komunitas secara otomatis, yaitu menggunakan nilai Nulitas dari Matriks Laplacian dari Matriks Ketetanggaan (NML) graf yang dimaksud. NML terbukti mampu menunjukkan banyaknya komunitas yang terkandung pada sebuah graf secara otomatis, walaupun teknik ini belum dapat dikombinasikan dengan algoritma Facetnet Paralel karena masih perlu dikembangkan algoritma yang dapat menghitung nilai NML ini secara paralel menggunakan matriks ketetanggaan yang sudah dipartisi.

Secara bertahap penelitian ini dimulai dari menganalisis berbagai metoda yang telah ada, menganalisis dan mendesain metoda usulan, melakukan eksperimen untuk menguji kemampuan metoda yang diusulkan hingga akhirnya dapat ditarik kesimpulan akan keberhasilan penelitian. Hasil eksperimen membuktikan metoda deteksi evolusi komunitas yang terdiri dari algoritma Facetnet Paralel menggunakan teknik *Vertex Cut* dan struktur graf dengan simpul beratribut mampu menghasilkan model evolusi komunitas dari graf yang berukuran hingga sekitar dua jutaan simpul dan 15 juta sisi menggunakan sistem terdistribusi berupa klaster Spark yang terdiri dari sembilan komputer. Hal ini sama artinya klaster Spark tersebut dapat memproses dengan maksimal 32 prosesor dan maksimal total memori worker sebesar 72 GB, karena masing-masing komputer memiliki empat buah prosesor dan 10 GB memori. Algoritma Facetnet Paralel terbukti merupakan algoritma yang memiliki skalabilitas tinggi, karena memiliki kompleksitas waktu yang linier. Berdasarkan eksperimen menggunakan infrastruktur klaster Spark yang ada dapat pula ditunjukkan bahwa setiap komputer mampu memproses graf yang nampaknya membutuhkan memori lebih besar dari total memori yang tersedia pada klaster, selama kebutuhan memori untuk memproses setiap sub graf tidak melebihi dari sekitar 55% dari memori pada masing-masing komputer.

Kata kunci: deteksi evolusi komunitas, sistem terdistribusi, teknik *Vertex Cut*, algoritma Facetnet Paralel, struktur graf dengan simpul beratribut

ABSTRACT

PARALLEL FACETNET ALGORITHM WITH VERTEX CUT TECHNIQUE AS A METHOD FOR COMMUNITY EVOLUTION DETECTION ON THE BIG GRAPH

By

Imelda Atastina

NIM: 33214013

(Program Studi Teknik Elektro dan Informatika)

The community evolution model is one of the most valuable information to support business activities because it can be used to understand community dynamics. This model can be obtained by detecting the evolution of the community on the graph. Meanwhile, the development of information technology has enable various organizations to store very large amounts of data as well as has resulted in the emergence of large-scale graph phenomena. To date, the method that has been developed to detect the evolution of communities is only able to process the graph with a maximum size of around a million vertices. Therefore, we need to develop the community evolution detection method for large-scale graphs. Studies related to the large graph processing show that the best way to process enormous graphs is by utilizing a distributed system because in general, the large graphs will require a large computer's memory as well. The size of the single computer's memory is generally unable to process the large graph because of the limited memory size. Based on the preliminary study of this research, the Facetnet algorithm is one of the community evolution detection algorithms that has the potential to have linear algorithm complexity so that it is considered suitable for processing large-scale graphs, also to produce a complete evolutionary model. So, it is considered very useful for developing this Facetnet algorithm to be able to process very large graphs using a distributed system.

To be able to process the huge graph using a distributed system there are several critical components, namely, graph partitioning technique and an algorithm that can be run in parallel. Graph partitioning technique is needed so that the large graph can be divided to produce some smaller subgraphs. In consequence, each subgraph can be processed in parallel using a computer that is part of a distributed system. This research has analyzed graph splitting techniques and found that the Vertex Cut is the best technique to be combined with the modified Facetnet algorithm resulting in the Parallel Facetnet algorithm that can be run on a distributed system to develop a community evolution model. However, simply dividing large graphs into subgraphs and combining them with the Parallel Facetnet algorithm is not adequate to be a method to process a tremendous graph. This is because of the probability matrix that has a huge size. This research has also succeeded in finding graph representation with attribute vertices to overcome this problem. With a graph structure that has an attribute node, the Parallel Facetnet algorithm is no longer necessary to define or store the adjacency matrix and probability membership matrix separately. Both matrices can already be replaced by

subgraphs with attribute nodes that can be stored in parallel, as well as processed in parallel by the Parallel Facetnet algorithm using local memory on each computer in a distributed system. Thus, the community evolution detection method using these three important elements be able to process a community evolution model for large-scale graphs. The limitation of memory capacity on a computer to process large graphs can be overcome by adding one or more computers in the distributed system, without changing the logic of the process defined in the Parallel Facetnet algorithm. Remarkably, this method is predicted to be able to process graphs with a million and even up to billion vertices.

This research also found that Facetnet has a weakness, that is the algorithm needs to determine the number of communities as conjecture at the beginning of the iteration, resulting in the community evolution model with the same number of communities for different graphs. This is certainly not in accordance by the real conditions that allow the various number of communities contained in a graph from time to time. Therefore, it has been tried to develop a method to determine the number of communities automatically, which uses the Nullity of the Laplacian Adjacency Matrix (NML) of the graph. NML is proven to be able to show the number of communities contained in a graph automatically, although this technique cannot be combined with the Parallel Facetnet algorithm yet, because it still needs to develop an algorithm that can calculate this NML value in parallel using a partitioned adjacency matrix.

This research was carried out in stages starting from analyzing various existing methods, analyzing and designing the proposed method, conducting experiments to test the ability of the proposed method until finally a conclusion can be drawn on the success of the study. Experimental results prove the community evolution detection method consisting of the Parallel Facetnet algorithm using the Vertex Cut technique and graph structure with attribute vertices capable of producing community evolution models of graphs of up to about two million vertices and 15 million edges using a distributed system of Spark clusters consisting of nine computers. This means that the Spark cluster can process the graph with a maximum of 32 processors and a maximum total worker memory of 72 GB, because each computer has four processors and 10 GB of memory. The Parallel Facetnet algorithm is proven to be an algorithm that has high scalability because it has linear time complexity. Based on experiments using the existing Spark cluster infrastructure it can also be shown that each computer is able to process graphs that appear to require memory from the total memory available in the cluster, as long as the memory requirements for processing each subgraph do not exceed about 55% of the memory on each computer.

Keyword: *community evolution detection, distributed system, Vertex Cut Partitioning, Parallel Facetnet Algorithm, graph with attribute node structure*