SKRIPSI

ANALISIS CO-OCCURRENCE NETWORKS PADA BIG GRAPH

Indra Permana Sugianto

NPM: 2017730008

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2022

LEMBAR PENGESAHAN

ANALISIS CO-OCCURRENCE NETWORKS PADA BIG GRAPH

Indra Permana Sugianto

NPM: 2017730008

Bandung, 20 Januari 2022

Menyetujui,

Pembimbing

Digitally signed by Veronica Sri Moertini

Dr. Veronica Sri Moertini

Ketua Tim Penguji

Digitally signed by Luciana Abednego

Luciana Abednego, M.T.

Anggota Tim Penguji

Digitally signed by Lionov

Lionov, Ph.D.

Mengetahui,

Ketua Program Studi

Digitally signed by Mariskha Tri Adithia

Mariskha Tri Adithia, P.D.Eng

PERNYATAAN

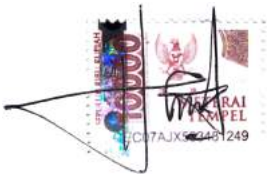
Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

ANALISIS CO-OCCURRENCE NETWORKS PADA BIG GRAPH

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 20 Januari 2022



Indra Permana Sugianto
NPM: 2017730008

ABSTRAK

Salah satu tren analisis data saat ini adalah analisis graf dari data media sosial. Data media sosial merupakan contoh *big data* yang dapat disiapkan menjadi graf dan dianalisis menggunakan teknologi *big data*. Permasalahan seperti deteksi komunitas dan analisis profil komunitas, menjadi salah satu topik yang banyak diteliti saat ini. Berbagai algoritma deteksi komunitas, seperti *Connected Component*, *Strongly Connected Component*, dan *Triangle Count* telah diimplementasikan. Namun, sebagian besar masih untuk graf tidak berarah, atau membutuhkan input jumlah komunitas, atau banyak algoritma yang komputasinya lambat. Oleh karena itu, kurang cocok diimplementasikan untuk *big graph*, yaitu *big data* yang mendeskripsikan relasi antar objek.

Pada penelitian sebelumnya, berhasil diciptakan solusi untuk deteksi komunitas pada *directed big graph* dari data Twitter. Komunitas dapat didefinisikan sebagai sekumpulan pengguna yang aktif berinteraksi pada periode tertentu karena membahas topik tertentu. Solusi yang diusulkan dengan menggunakan *Motif Finding*, yaitu teknik pengenalan pola (*pattern matching*) yang diimplementasikan pada *library* GraphFrames di Apache Spark. Dilakukan perbandingan deteksi komunitas menggunakan algoritma *Strongly Connected Component* (SCC) dan teknik *Motif Finding*. Secara waktu eksekusi, didapatkan bahwa teknik *Motif Finding* memperoleh hasil yang lebih cepat dibandingkan SCC. Namun, teknik *Motif Finding*, perlu mendefinisikan pola-pola komunitas (motif) yang ingin dideteksi.

Skripsi ini berfokus pada deteksi komunitas dan analisis *Co-Occurrence* untuk analisis profil komunitas dari data Twitter. *Co-Occurrence* bermakna kemunculan suatu kejadian dalam frekuensi yang tinggi. Contohnya, kemunculan kata "*trading*" dalam kumpulan teks *tweet* di domain investasi. Deteksi komunitas menggunakan teknik *Motif Finding*, sedangkan untuk analisis profilnya menggunakan analisis *tweet* yang sering dikirimkan di komunitas (*co-occur*). Analisis *Co-Occurrence* dilakukan dengan teknik pemodelan topik.

Pengujian metode deteksi komunitas dan analisis profil komunitas dilakukan pada data *tweet* dengan domain Covid berbahasa Indonesia. Berdasarkan eksperimen yang dilakukan, metode di atas dapat mendeteksi komunitas dan menganalisis profilnya melalui kata-kata yang sering dikirimkan. Kata tersebut dapat digunakan untuk mendeskripsikan konten yang dibicarakan pada komunitas yang terbentuk. Namun, kekurangan dari metode di atas adalah harus mengetahui jumlah topiknya secara pasti, jika ingin hasilnya lebih baik.

Kata-kata kunci: *big graph*, deteksi komunitas, *co-occurrence*, *motif finding*, Apache Spark, GraphFrames, pemodelan topik, Twitter

ABSTRACT

One of the current trends in data analysis is graph analysis of social media data. Social media data is an example of big data that can be prepared into graphs and analyzed using big data technologies. Problems such as community detection and analysis of community profiles (profiling) have become one of the most studied topics today. Various community detection algorithms, such as Connected Component, Strongly Connected Component, and Triangle Count have been implemented. However, most of them are still for undirected graphs, or require input number of communities, or many algorithms that are computationally slow. Therefore, it is not suitable to be implemented for big graphs, namely big data that describes the relation between objects.

In the previous research, a solution was created for community detection on the directed big graph of Twitter data. The community can be defined as a group of users who actively interact during a certain period, because they discuss certain topics. The proposed solution uses Motif Finding, a pattern matching technique that implemented in the GraphFrames library in Apache Spark. A comparison of community detection was carried out using the Strongly Connected Component (SCC) algorithm and the Motif Finding technique. In terms of execution time, it was found that the Motif Finding technique obtained faster results than SCC. However, the Motif Finding technique needs to define community patterns (motifs) to be detected.

This thesis focuses on community detection and Co-Occurrence analysis for community profile analysis from Twitter data. Co-Occurrence can be defined as the occurrence of an event in a high frequency. For example, the occurrence of the word "trading" in a collection of tweet text in the investment domain. The community detection uses the Motif Finding technique, while the profiling uses the analysis of tweets that are often sent in the community (co-occur). Co-Occurrence analysis was carried out using topic modeling techniques.

The evaluation of the community detection and profiling methods were carried out on tweet data with the Indonesian language Covid domain. Based on the experiments, the methods are able to detect communities and analyze the community profiles through frequently sent words. The words can be used to describe the content discussed in the community that is formed. However, the drawback of the methods are that necessary to know the number of topics, if expecting to get better results.

Keywords: big graph, community detection, co-occurrence, motif finding, Apache Spark, GraphFrames, topic modelling, Twitter

KATA PENGANTAR

Puji syukur Penulis panjatkan ke hadirat Tuhan yang Maha Esa, atas berkat-Nya Penulis mampu menyelesaikan skripsi ini. Pada kesempatan ini, Penulis ingin berterima kasih kepada pihak yang membantu dan mendukung selama menjalani kuliah dan pengerjaan skripsi, yaitu sebagai berikut.

1. Kepada orang tua dan paman Penulis yang selalu memberi doa, dukungan dan semangat;
2. Kepada Ibu Dr. Veronica Sri Moertini selaku pembimbing. Terima kasih atas ilmu, pengalaman, semangat, dan bimbingan yang telah diberikan;
3. Kepada Ibu Luciana Abednego, M.T. dan Bapak Lionov, Ph.D. selaku penguji. Terima kasih atas masukan dan saran yang diberikan, sehingga skripsi ini dapat dikembangkan menjadi lebih baik;
4. Kepada seluruh dosen yang mengajar Penulis sejak pertama masuk kuliah. Terima kasih atas bekal ilmu dan pengalaman yang berharga;
5. Kepada teman-teman mahasiswa bimbingan Bu Vero (Fariz, Yoga, Fritz, Yohan), terima kasih sudah berjuang bersama. Akhirnya semua bisa lulus bareng :D;
6. Kepada Fanny Susiani Bunawan, terima kasih atas dukungan yang telah diberikan;
7. Kepada teman-teman mahasiswa angkatan 2017 sampai 2020. Terima kasih atas momen kebersamaannya. Semangat menjalani kuliahnya, bagi teman-teman adik tingkat;
8. Kepada seluruh pihak yang tidak Penulis cantumkan pada bagian ini. Terima kasih atas dukungannya.

Sebagai penutup, Penulis berterima kasih kepada pembaca dokumen skripsi ini. Penulis berharap semoga penelitian yang dilakukan dapat bermanfaat bagi para pembaca, kemajuan ilmu pengetahuan, dan dapat menjadi referensi yang baik bagi penelitian di masa depan. Penulis memohon maaf jika terdapat kesalahan penulisan, sistematika, atau metode yang dipaparkan pada skripsi ini. Penulis sangat terbuka untuk menerima masukan dan saran tambahan dari pembaca.

Bandung, Januari 2022

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR KODE PROGRAM

BAB 1

PENDAHULUAN

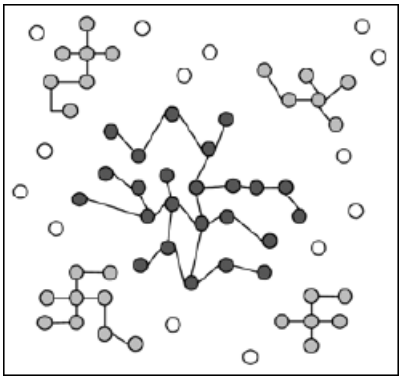
Pada bab ini akan dibahas tentang latar belakang penelitian yang mendeskripsikan gambaran besar permasalahan yang ada. Dilanjutkan dengan pembahasan rumusan masalah, tujuan penelitian, dan batasan penelitian. Pada tiga bagian tersebut, dijelaskan inti permasalahan, tujuan yang ingin dicapai sebagai solusi masalah, serta asumsi yang digunakan untuk membatasi ruang lingkup penelitian ini. Pembahasan tentang metodologi dan sistematika pembahasan menjadi penutup dari bab ini. Dua bagian tersebut menjelaskan tentang tahapan eksperimen dan penjabaran konten setiap bab pada buku ini.

1.1 Latar Belakang

Istilah *Big Data* merujuk pada sebuah himpunan data yang memiliki karakteristik khusus. Umumnya, karakteristik tersebut didefinisikan sebagai 3V, yaitu *Volume*, *Velocity*, dan *Variety*. *Volume* mendefinisikan ukuran data yang sangat besar, *Velocity* mendefinisikan penambahan data yang terjadi terus menerus dalam waktu singkat, sedangkan *Variety* mendefinisikan bentuk data yang sangat beragam [7]. Data dapat berupa data terstruktur seperti tabel basis data, data semi-terstruktur seperti XML dan JSON, dan data tidak terstruktur seperti data teks dari Twitter atau data citra dari satelit. Agar kumpulan data tersebut lebih bermakna, muncul suatu kebutuhan untuk menganalisisnya. Lalu, mencari informasi berharga dari kumpulan data tersebut, dan membuat keputusan/menarik kesimpulan berdasarkan hasil analisis data.

Big data yang datanya mendeskripsikan relasi antar objek, dapat diolah dan dimodelkan dalam bentuk *big graph*. *Big graph* adalah himpunan simpul dan sisi yang dibentuk dari *big data*. Sebagai contoh, data interaksi pengguna Twitter termasuk *big data* dapat dimodelkan dalam bentuk graf. Pemodelan graf menggunakan graf berarah (*directed graph*). Simpul pada graf memodelkan sebuah akun pengguna, sedangkan sisi pada graf memodelkan interaksi antar pengguna. Interaksi yang dimaksud dapat berupa *retweet*, *quote*, atau *reply*. Arah pada sisi graf menunjukkan interaksi sebuah akun terhadap akun lainnya. Sebagai contoh, seorang pengguna me-*retweet* atau membalas *tweet* pengguna lain. Setelah model graf terbentuk, nantinya graf dapat dianalisis menggunakan berbagai teknik. Saat ini, permasalahan yang sering diteliti pada analisis *big graph* adalah masalah deteksi komunitas dan analisis profil komunitas. Komunitas pada data Twitter, dapat didefinisikan sebagai kumpulan pengguna yang aktif berinteraksi pada periode tertentu, karena adanya ketertarikan untuk membicarakan suatu topik[8]. Inti permasalahan deteksi komunitas adalah mencari/menemukan kelompok pengguna yang saling berinteraksi. Jika data pengguna dimodelkan dalam graf, komunitas akan terlihat seperti graf yang padat (memiliki jumlah simpul dan sisi yang lebih banyak). Gambar 1.1 menunjukkan contoh visualisasi komunitas berupa graf yang lebih padat.

Analisis komunitas menjadi banyak diteliti karena melimpahnya data dari media sosial, sehingga ada potensi untuk menggali informasi berharga dari data tersebut. Berdasarkan hasil studi literatur dari Moertini dan Adithia, serta Needham dan Hodler [8, 9], ditemukan bahwa berbagai algoritma deteksi komunitas, seperti *Connected Component* (CC), *Strongly Connected Component* (SCC), dan *Triangle Count* telah berhasil diimplementasikan. Namun, sebagian besar untuk graf tidak berarah,



Gambar 1.1: Contoh pemodelan komunitas [3]

atau membutuhkan input jumlah komunitas. Kedua kendala tersebut menyebabkan implementasi dengan algoritma yang ada cukup sulit untuk kasus komunitas di Twitter, karena komunitas yang terbentuk sangat dinamis dan hanya berlaku pada periode tertentu saja. Bisa saja terbentuk komunitas baru, atau komunitas yang terbentuk pada suatu periode dapat menghilang di periode lain. Terbentuknya atau hilangnya komunitas dipengaruhi oleh interaksi yang dilakukan pengguna. Selain itu, terdapat algoritma yang komputasinya lambat, seperti CC dan SCC, sehingga kurang cocok jika diterapkan untuk *big data*.

Pada penelitian Moertini dan Adithia [8], berhasil diciptakan solusi untuk deteksi komunitas pada *directed big graph* dari data Twitter. Solusi yang diusulkan dengan menggunakan *Motif Finding*, yaitu teknik pengenalan pola (*pattern matching*) yang diimplementasikan pada *library* GraphFrames di Apache Spark. Dilakukan perbandingan deteksi komunitas menggunakan algoritma *Strongly Connected Component* (SCC) dan teknik *Motif Finding*. Secara waktu eksekusi, didapatkan bahwa teknik *Motif Finding* memperoleh hasil yang lebih cepat dibandingkan SCC. Namun, teknik *Motif Finding*, perlu mendefinisikan pola-pola komunitas (motif) yang ingin dideteksi.

Salah satu teknik analisis yang dapat diterapkan pada *big graph* data Twitter adalah analisis *Co-Occurrence Networks*. *Co-Occurrence Networks* bermakna kemunculan suatu kejadian dengan frekuensi tinggi dalam data yang bermodel graf [2]. Contoh *co-occurrence* adalah, kemunculan kata "vaksin" dalam kumpulan teks *tweet* tentang Covid yang dikirimkan pengguna. Analisis *Co-Occurrence Networks*, dapat dilakukan terhadap teks *tweet* sekumpulan pengguna yang sering berinteraksi (komunitas). Tujuannya untuk mencari tahu kata-kata yang sering dikirimkan kumpulan pengguna tersebut saat saling berinteraksi. Hasil analisisnya dapat diinterpretasi dan dimanfaatkan, seperti promosi produk untuk komunitas tertentu, mendeteksi topik dari sebuah komunitas yang aktif berinteraksi, atau menganalisis profil komunitas untuk mencari tahu topik apa yang dibicarakan anggota komunitas, sehingga kumpulan pengguna menjadi aktif berinteraksi.

Salah satu teknologi yang menunjang proses analisis *big graph* adalah *library* GraphX dan GraphFrames pada Apache Spark. Kedua *library* tersebut mengimplementasikan berbagai algoritma graf yang dijalankan secara paralel pada teknologi Apache Spark. Algoritma yang diimplementasikan seperti *PageRank*, *Connected Components*, dan *Breadth First Search*. Spark mampu melakukan komputasi terdistribusi yang menyebabkan proses komputasi berlangsung dengan cepat dan efisien. Oleh karena itu, cocok untuk digunakan pada komputasi yang dilakukan secara iteratif [5], seperti pada analisis *big graph*.

Skripsi ini berfokus pada deteksi komunitas dan analisis *Co-Occurrence* untuk analisis profil komunitas dari data Twitter. Data studi kasusnya berupa data interaksi pengguna Twitter, beserta *tweet* yang dikirim. Data berformat JSON, dikumpulkan secara *near real time* memanfaatkan Apache Kafka, dan disimpan pada *Hadoop Distributed File System* (HDFS). Penjelasan arsitektur sistem pengumpul data dapat diakses pada Gambar 4.1.

Data Twitter yang telah disimpan, nantinya akan dieksplorasi dan ditransformasi hingga dapat dimodelkan menjadi graf. Pembuatan graf memanfaatkan *library* GraphFrames pada Apache Spark.

Setelah graf terbentuk, akan dilakukan deteksi komunitas menggunakan teknik *Motif Finding* yang diimplementasikan pada GraphFrames. Komunitas yang terdeteksi akan dianalisis profilnya dengan cara analisis *tweet* yang sering dikirimkan di komunitas (*co-occur*). Sebelum dianalisis, data teks akan dibersihkan dan diproses dahulu menggunakan *library* Pandas dan PySastrawi di Python. Analisis Co-Occurrence dilakukan dengan teknik pemodelan topik *Latent Semantic Analysis* yang diimplementasikan pada *library* Scikit-Learn di Python.

Perangkat lunak yang dihasilkan terdiri dari tiga bagian, yaitu program untuk melakukan eksplorasi dan penyiapan data, program untuk pembuatan graf dan deteksi komunitas, dan program untuk pemrosesan teks dan pemodelan topik. Program pertama dan kedua diimplementasikan dalam bahasa pemrograman Scala dan dijalankan dengan basis *command line interface*. Sedangkan program ketiga diimplementasikan dalam bahasa pemrograman Python dan dijalankan pada Google Colaboratory, yaitu sebuah teknologi berbasis *cloud* untuk mengeksekusi kode Python.

1.2 Rumusan Masalah

Adapun rumusan masalah dari latar belakang yang telah dipaparkan adalah:

1. Bagaimana cara implementasi deteksi komunitas dan analisis *Co-Occurrence Networks* dari data interaksi pengguna Twitter?
2. Bagaimana cara eksplorasi dan menyiapkan data interaksi pengguna Twitter hingga dapat dimodelkan dalam graf dan dapat dianalisis *co-occurrence*-nya?
3. Bagaimana cara analisis, dan interpretasi hasil analisis *big data* studi kasus?
4. Bagaimana cara membangun perangkat lunak yang mampu melakukan penyiapan dan analisis data studi kasus?

1.3 Tujuan

Adapun tujuan penelitian dari rumusan masalah yang telah dipaparkan adalah:

1. Mempelajari konsep deteksi komunitas dan analisis *Co-Occurrence Networks*;
2. Melakukan eksplorasi dan penyiapan *big data* studi kasus, merancang pemodelan grafnya, dan merumuskan aspek yang dapat dianalisis *co-occurrence*-nya;
3. Melakukan analisis *big data* studi kasus dan menginterpretasikan hasil analisisnya untuk menemukan informasi berharga;
4. Membangun perangkat lunak yang mampu melakukan penyiapan dan analisis *big data* studi kasus.

1.4 Batasan Penelitian

Batasan untuk penelitian ini adalah sebagai berikut:

1. Teks *tweet* yang dianalisis interaksinya, hanya berasal dari domain tertentu dan dalam Bahasa Indonesia;
2. Teks *tweet* yang dianalisis tidak dilakukan pemrosesan tambahan, seperti *n-gram*, lemmatisasi, *named entity recognition*, dan *parsing*, karena keterbatasan *library* untuk pengolahan teks Bahasa Indonesia.
3. Teks *tweet* yang *spam* tidak ditangani.
4. Interaksi antar pengguna yang dianalisis hanya *quote* atau *reply* saja.

1.5 Metodologi

Metodologi yang digunakan dalam pembuatan skripsi ini adalah sebagai berikut:

1. Melakukan eksplorasi dan studi literatur tentang Apache Spark;

2. Melakukan eksplorasi dan studi literatur tentang bahasa pemrograman Scala;
3. Melakukan eksplorasi dan studi literatur tentang *library* GraphX dan GraphFrames;
4. Melakukan studi literatur tentang konsep graf dan deteksi komunitas;
5. Melakukan eksplorasi dan studi literatur tentang konsep *Co-Occurrence Networks*;
6. Melakukan eksplorasi dan studi literatur tentang konsep pemodelan topik, *text pre-processing*, dan fungsi-fungsi terkait pada *library* Scikit-Learn dan PySastrawi;
7. Mengumpulkan *big data* studi kasus yang dapat ditransformasi menjadi *big graph* dan berpotensi untuk analisis *Co-Occurrence Networks*;
8. Melakukan eksplorasi dan penyiapan *big data* studi kasus;
9. Membuat perangkat lunak yang mampu melakukan penyiapan data, analisis data, dan menampilkan hasil analisisnya;
10. Melakukan analisis terhadap hasil yang didapatkan, merumuskan interpretasi hasil analisis, dan memaparkan hasilnya dengan menarik.

1.6 Sistematika Pembahasan

Sistematika penulisan skripsi ini adalah sebagai berikut.

1. Bab Pendahuluan
2. Bab Landasan Teori
3. Bab Studi Eksplorasi Apache Spark, GraphFrames, dan Scikit-Learn
4. Bab Pengumpulan, Eksplorasi, dan Penyiapan Data
5. Bab Analisis Data dan Evaluasi
6. Bab Kesimpulan dan Saran

Bab Pendahuluan membahas tentang pengantar skripsi yang berisi gambaran besar permasalahan yang dibahas, batasan ruang lingkup penelitian, serta tahapan eksperimen yang dilakukan. Bab Landasan Teori berisi pembahasan hasil studi literatur yang mencakup penjelasan berbagai konsep yang digunakan. Bab Studi Eksplorasi Teknologi membahas tentang berbagai eksperimen untuk mempelajari konsep dan teknologi yang digunakan pada skripsi ini. Pada Bab tiga juga dijelaskan tentang contoh penerapan metodologi penelitian (eksplorasi dan penyiapan data, pemodelan graf, deteksi komunitas dengan *Motif Finding*, tahap pra-pemrosesan data teks, dan pemodelan topik dengan *Latent Semantic Analysis*). Bab Pengumpulan, Eksplorasi, dan Penyiapan Data membahas tentang penerapan metodologi penelitian tahap pertama (pengumpulan, eksplorasi, penyiapan data, dan pemodelan graf) pada data studi kasus (detail data studi kasus dijelaskan pada bab terkait). Bab Analisis Data dan Evaluasi membahas tentang penerapan metodologi penelitian tahap selanjutnya (pembuatan graf, deteksi komunitas, pemodelan topik, dan interpretasi hasil analisis). Buku ini diakhiri dengan Bab Kesimpulan dan Saran.