BAB I
Pendahuluan

1.1 Latar Belakang

Di era bigdata yang sedang populer seperti ini tentunya pengolahan data sangatlah penting dan menjadi hal mutlak dalam merepresentasikan penggunaan teknologi informasi baik di perusahaan maupun instansi di seluruh dunia.

Relational database merupakan sistem penyimpanan dan pengambilan data yang telah populer dan mendominasi selama lebih dari tiga dekade. Banyak aplikasi yang menggunakan relational *database* untuk penyimpanan data. *Relational database* dapat bekerja dengan baik apabila jumlah data sedikit dan memiliki data terstruktur[1]. Ketika terjadinya peningkatan jumlah data dan berbagai pemrosesan data maka relational *database* dengan skema yang kaku sangat tidak cocok untuk kasus data semi terstruktur bahkan tidak terstruktur[1]. Kasus data tidak terstruktur dan semi terstruktur memiliki fleksibilitas dalam hal pemrosesan data seperti halnya dengan kasus social network dengan interconnected data.

Salah satu solusi yang digunakan untuk mengatasi hal tersebut adalah menggunakan *Graph Database*. basisdata graf adalah salah satu metode implementasi dari NoSQL (*Not Only SQL*) yaitu sistem basisdata yang berguna menyimpan data dalam jumlah besar dan direpresentasikan ke dalam graf, berbentuk *Node* dan *Edge*[2]. Hal ini dilakukan karena *Node* dan *Edge* memberi peluang untuk ekstraksi informasi antar user. Kelebihan basisdata graf adalah dalam hal pencarian data bisa dilakukan secara tranversal dengan setiap relasi direpresentasikan dengan suatu *Edge* yang menghubungkan *Node-Node* yang berelasi, sehingga waktu pemrosesan dapat dilakukan dengan efektif[2].

Namun ketika hanya menggunakan model *database* yang berbentuk relational *database* tentunya semakin lama semakin kesulitan karena datanya disini sangatlah banyak sekali. Disinilah penulis akan menggunakan model *database* yang masih tergolong baru, yaitu *Graph Database*. Model ini dapat merepresentasikan banyak data dalam suatu graf yang bisa dianalisis serta diambil kesimpulannya dari banyak *Nodes* serta *Edges* yang penulis peroleh dari dataset NCI (National Cancer Institute) dan Cheminformatics.

Penulis memilih dataset NCI dan Cheminformatics disini adalah karena format *input* yang sudah terstandarisasi sehingga dalam menggali informasi dari ikatan molekul kimia tersebut dapat terlaksana dengan baik. Kemudian format yang dipakai oleh dataset ini adalah SMILES, yaitu bertipe *Simplified Molecular-Input Line-Entry System* (SMILES) yaitu suatu spesifikasi dalam bentuk notasi baris untuk menggambarkan struktur spesies kimia menggunakan string ASCII pendek. String dari SMILES dapat diimpor oleh aplikasi untuk dikonversi menjadi ringkasan molekul[6].

Dengan model basisdata graf tersebut diharapkan dapat meningkatkan kualitas untuk merepresentasikan informasi yang terkandung di dalamnya. Kemudian dilakukan peringkasan graf untuk lebih mencapai tujuan akhir yaitu sebuah ringkasan graf yang menjadi intisari dari dataset basisdata graf tersebut.

Peringkasan basisdata graf penulis ambil sebagai topik dari penulisan tugas akhir ini. Metode peringkasan yang penulis ambil adalah RP-GD Algorithm yang

penulis ambil mempunyai efisiensi dan kualitas yang dapat dengan mudah meringkas suatu basisdata graf[1]. Kualitas hasil peringkasan graf juga diukur berdasarkan cakupan informasi serta rasio peringkasan sebagai parameter kualitasnya.

Proses peringkasan ini mengkombinasikan berbagai keuntungan seperti skalabilitas, konsumsi memory, skema pembangunan basisdata dan kapasitas untuk menguasai berbagai data menjadi data pilihan dari pengguna. Ringkasan yang dihasilkan menyediakan pandangan data yang dapat dibentuk sesuai dengan keinginan pengguna[3].

Selain itu, ada alasan pentingnya melakukan peringkasan basisdata, yaitu kembali kepada definisinya, perubahan penyusutan dari basisdata menjadi bentuk yang ringkas, melalui proses pengurangan isi dengan cara menyeleksi dan/atau menyamaratakan dari apa-apa yang penting di dalam basisdata tersebut[8]. Sehingga tujuan utama penulis melakukan peringkasan adalah untuk memberikan gagasan/ide pokok dari basisdata yang asli namun dalam bentuk yang ringkas.

Dari hasil peringkasan tersebut juga membuang data yang 'tidak diperlukan'. Konsentrasi yang penulis berikan disini adalah dari sudut pandang demi kepuasan pengguna, sehingga dalam meringkas basisdata graf, dapat mengurangi simpul-simpul yang tidak ada hubungannya dengan topik yang pengguna inginkan. Alasan lain adalah, tentunya dapat mengurangi beban memori serta proses query yang dilakukan.

Seringkali *nodes* (simpul) mempunyai atribut yang berhubungan dengan diri mereka. Dalam banyak aplikasi, graf berukuran sangat besar, dengan ribuan atau bahkan jutaan simpul dan tepi. Hasilnya, hamper mustahil untuk memahami informasi yang terkandung di dalamnya hanya dengan melihat sekilas saja. Maka, metode peringkasan graf sangat dibutuhkan agar membantu pengguna menggali dan memahami informasi pokok yang terkandung didalamnya[7].

1.2 Perumusan Masalah

Permasalahan yang dibahas dalam Tugas Akhir ini adalah:

1. Apakah metode algoritma RP-GD dapat digunakan dalam studi kasus peringkasan dataset molekuler kimia dari NCI dan Cheminformatics berdasarkan latar belakang diatas?
2. Bagaimanakah kualitas peringkasan yang dihasilkan dengan algoritma RP-GD untuk studi kasus molekuler ikatan kimia dari dataset diatas berdasarkan rasio peringkasan dan cakupan informasi sebagai parameter kualitasnya?
3. Apa saja rekomendasi yang dapat diberikan oleh algoritma RP-GD ini terhadap molekuler ikatan kimia dalam dataset tersebut?

1.3 Tujuan

Tujuan yang ingin dicapai dari penelitian ini adalah:

1. Mengetahui penggunaan metode algoritma peringkasan RP-GD dalam peringkasan dataset dari NCI dan Cheminformatics.
2. Mengetahui kualitas peringkasan yang dihasilkan untuk studi kasus NCI dan Cheminformatics
3. Memberikan rekomendasi untuk NCI dan Cheminformatics serta kualitas berdasarkan hasil dari tugas akhir.

1.4 Batasan Masalah

Adapun batasan masalah dalam tugas akhir ini adalah:

1. Dataset yang digunakan sebagai data uji berasal dari NCI dan Cheminformatics. Basisdata graf yang dibuat berdasarkan data dari molekul-molekul ikatan kimia yang ditulis kedalam bentuk format SMILES.
2. Algoritma yang digunakan adalah RP-GD.
3. Graf yang digunakan dalam penelitian ini berbentuk graf tak berarah.
4. Graf dataset awal menunjukkan molekul yang terdiri dari atom-atom beserta ikatannya.
5. Setiap *Node* pada graf terhubung dengan minimal satu *Node* lain.
6. Atom H (Hidrogen) tidak diidentifikasi.

1.5 Metodologi Penyelesaian Masalah

Metedologi penyelesaian masalah terbagi dalam beberapa tahapan, yaitu:

1. Studi literatur
   Tahap ini dilakukan pencarian materi-materi yang mendukung dengan dasar teori yang berkaitan erat dengan pembuatan tugas akhir ini. Materi dapat dicari dari buku, jurnal/paper, artikel resmi, tugas akhir, dan tesis yang berhubungan dengan *Graph Database*, *Summarization Graph*, Algoritma RP-GD, *Frequent Subgraph*, dan dataset ikatan kimia..
2. Pengumpulan dan Pengolahan data
   Menggunakan studi kasus pengambilan data-data yang diperlukan untuk penulisan tugas akhir ini, berupa dataset dari situs NCI dan cheminformatics berupa dataset molekuler ikatan kimia.
3. Pembangunan dan Perancangan sistem
   Pada tahap pembangunan sistem, akan dilakukan pendefinisian masalah serta solusi yang diharapkan. Setelah itu, dilakukan analisis untuk pemodelan sistem dan merumuskan langkah-langkah implementasi sistem.
4. Implementasi metode
   Mengimplementasikan metode peringkasan basisdata graf untuk mengetahui hasil pengujian dengan studi kasus yang digunakan.
5. Pengujian
   Pengujian dilakukan dengan mengukur performansi dan mengukur tingkat kompresi dari hasil peringkasan basisdata graf melalui implementasi algoritma RP-GD untuk menghasilkan ringkasan yang efektif.
6. Analisis hasil dan kesimpulan
   Melakukan analisis terhadap hasil pengujian sehingga dapat ditarik kesimpulan berdasarkan hasil pengujian tersebut.
7. Dokumentasi
   Dokumentasi dilakukan untuk mencatat setiap hal penting dalam pelaksanaan tugas akhir ini.

1.6 Sistematika Penulisan

Adapun sistematika penulisan Tugas Akhir ini adalah sebagai berikut:

1. BAB I Pendahuluan
   Pada Bab I diuraikan isi dan rencana pengerjaan Tugas Akhir secara keseluruhan yang meliputi latar belakang, perumusan masalah, tujuan, batasan masalah, hipotesa dan metodologi penyelesaian masalah yang diterapkan.
2. BAB II Tinjauan Pustaka
   Bab II memaparkan dasar-dasar teori yang berkaitan dengan Basisdata Graf, peringkasan graf dan subgraf, algoritma RP-GD, penerapan algoritma RP-GD dan pengukuran kualitas.
3. BAB II Analisis Perancangan dan Implementasi
   Perancangan sistem dari sistem yang dibangun akan dijelaskan pada bab ini. Selanjutnya dilakukan proses implementasi.
4. BAB IV Pengujian dan Analisis
   Pada bab ini dibahas skenario dan hasil pengujian yang dilakukan pada hasil implementasi sistem.
5. BAB V Penutup
   Bab ini berisi kesimpulan dan saran yang didapatkan dari hasil implementasi sistem secara keseluruhan.