Penerapan Algoritma Anonimisasi Data pada Lingkungan Big Data

Stephen Jordan–2016730018

1 Deskripsi

Pertumbuhan data yang semakin pesat, mendorong munculnya teknik baru bagi pengolahan data untuk menemukan informasi yang tepat pada sekumpulan data. Konsep ini sering dikenal dengan nama *data mining*. Selain keuntungan yang ditawarkan, penggunaan teknik *data mining* juga menimbulkan masalah baru bagi keamanan privasi data. Masalah ini terjadi ketika informasi tersebut disalahgunakan untuk merugikan orang lain. Keuntungan yang didapat dari teknik *data mining* dinilai tidak seimbang dengan konsekuensi yang diterima saat ada privasi yang terlanggar. Oleh karena itu, diperlukan pendekatan baru agar perlindungan privasi masih tetap terjaga meskipun data diolah menggunakan teknik *data mining*.

Perlindungan privasi dapat dicapai dengan 2 metode, yaitu enkripsi dan anonimisasi. Enkripsi adalah metode perlindungan privasi dengan memanfaatkan pola atau kunci untuk mengubah bentuk data menjadi bentuk lain yang tidak mudah dikenali. Anonimisasi adalah metode perlindungan privasi dengan cara menyamarkan sebagian atribut data agar tidak dapat dikenali dengan nilai dari atribut data lainnya. Anonimisasi bertujuan untuk menyamarkan data agar tidak ada seseorang yang dapat mencari kemungkinan tertinggi antara hubungan nilai atribut dengan entitas yang dimaksud. Pada kasus tertentu, keamanan enkripsi dapat diretas melalui penalaran hubungan antar nilai atribut yang dapat mengungkapkan identitas individu sesungguhnya.

Pemanfaatan teknologi informasi saat ini, menimbulkan pertumbuhan data yang sangat pesat. Dalam waktu singkat, data yang diperoleh dapat mencapai ukuran yang sangat besar. Ukuran terbesar saat ini dapat mencapai *Petabyte*, yakni satu juta kali lebih besar dari ukuran *Gigabyte*. *Big data* adalah himpunan data dalam jumlah yang sangat besar dan kompleks sehingga sulit untuk ditangani atau diproses apabila hanya menggunakan manajemen basis data biasa. *Big data* dipilih, karena di masa mendatang data yang diolah akan berukuran besar, sehingga rawan terjadinya pelanggaran privasi saat dilakukan teknik *data mining*.

Keamanan privasi data dapat ditingkatkan dengan menyembunyikan lebih banyak nilai data, tetapi menurunkan nilai informasi pada data tersebut, berlaku juga sebaliknya. Karena itu, diperlukan pendekatan untuk menyeimbangkan kebutuhan nilai informasi dan privasi. Pendekatan ini disebut *k-anonymity*. Pada penelitian sebelumnya, *k-anonymity* dimodelkan dengan 2 algoritma, yaitu *hierarchy based generalization* and *hierarchy-free generalization*. Akan tetapi algoritma ini memiliki kelemahan, yaitu hilangnya nilai informasi yang relatif tinggi. Solusinya adalah memandang *k-anonymity* menjadi permasalahan *clustering*, dikenal sebagai *K-member clustering problem*. *K-member clustering problem* mendorong penggunaan algoritma *Greedy K-member clustering* karena memiliki performa yang cukup baik pada penelitian sebelumnya.

Pada penelitian ini, dibuat perangkat lunak yang dapat memproses anonimisasi pada lingkungan *big data* melalui penerapan algoritma *Greedy K-member clustering* menggunakan Spark. Spark adalah *framework* yang mendukung penerapan komputasi kompleks pada lingkungan *big data*. Spark membutuhkan mekanisme penyimpanan Hadoop, karena Spark tidak memiliki mekanisme penyimpanan tetap. Mekanisme penyimpanan Hadoop dikenal sebagai Hadoop File System (HDFS). Mekanisme penyimpanan Hadoop dibutuhkan, agar hasil pemrosesan data dapat disimpan pada *hardisk* komputer. Kelebihan dari Spark adalah waktu pemrosesan data yang lebih cepat, karena hasil pemrosesan data dapat disimpan sementara pada memori untuk diambil lagi pada iterasi selanjutnya. Tujuan akhir dari penelitian ini adalah membandingkan kualitas informasi yang didapat dari penggunaan teknik *data mining*, sebelum dan setelah data dianonimisasi.

2 Rumusan Masalah

Berdasarkan deskripsi diatas, rumusan masalah pada skripsi ini adalah sebagai berikut:

1. Bagaimana cara kerja algoritma anonimisasi *Greedy K-member clustering* ?
2. Bagaimana implementasi algoritma anonimisasi *Greedy K-member clustering* pada lingkungan Spark?
3. Bagaimana perbandingan kualitas informasi, sebelum dan setelah data dianonimisasi?

3 Tujuan

Berdasarkan rumusan masalah di atas, tujuan dari skripsi ini adalah sebagai berikut:

1. Mempelajari cara kerja algoritma *Greedy K-member clustering*.
2. Mengimplementasikan algoritma *Greedy K-member clustering* pada lingkungan Spark.
3. Melakukan analisis kualitas informasi terhadap hasil teknik data mining, sebelum dan setelah data dianonimisasi.

4 Deskripsi Perangkat Lunak

Perangkat lunak akhir yang akan dibuat memiliki fitur minimal sebagai berikut:

- Perangkat lunak dapat menerima masukan data XML dan CSV.
- Perangkat lunak dapat melakukan modifikasi data input menjadi data yang sudah dianonimisasi
- Pengguna dapat memilih atribut data yang ingin dianonimisasi.
- Pengguna dapat memperoleh data yang sudah dianonimisasi.
- Pengguna dapat membandingkan kualitas informasi, sebelum dan setelah data dianonimisasi.

5 Detail Pengerjaan Skripsi

Bagian-bagian pekerjaan skripsi ini adalah sebagai berikut:

1. Mempelajari teknik-teknik dasar *data mining*.
2. Mempelajari algoritma *Greedy K-member clustering*.
3. Mempelajari konsep Hadoop, Spark, dan Spark MLlib.
4. Melakukan instalasi dan konfigurasi Spark pada *cluster* Hadoop.
5. Mempelajari bahasa pemrograman Scala pada *framework* Spark.
6. Melakukan studi dan eksplorasi teknik-teknik dasar *data mining* pada Spark MLlib.
7. Mencari dan mengumpulkan data studi kasus.

8. Mengimplementasikan algoritma *Greedy K-member clustering* pada Spark.
9. Melakukan perancangan dan implementasi perangkat lunak menggunakan *library* Scala-swing.
10. Mengimplementasikan teknik-teknik dasar *data mining* menggunakan *library* Spark MLlib.
11. Melakukan pengujian fungsional dan experimental.
12. Melakukan analisis kualitas informasi, sebelum dan setelah data dianonimisasi.
13. Menulis dokumen skripsi.

6 Rencana Kerja

Rincian capaian yang direncanakan di Skripsi 1 adalah sebagai berikut:

1. Mempelajari teknik-teknik dasar *data mining*.
2. Mempelajari algoritma *Greedy K-member clustering*.
3. Mempelajari konsep Hadoop, Spark, dan Spark MLlib.
4. Melakukan instalasi dan konfigurasi Spark pada *cluster* Hadoop.
5. Mempelajari bahasa pemrograman Scala pada *framework* Spark.
6. Melakukan studi dan eksplorasi teknik-teknik dasar *data mining* pada Spark MLlib.
7. Menulis dokumen skripsi.

Sedangkan yang akan diselesaikan di Skripsi 2 adalah sebagai berikut:

1. Mencari dan mengumpulkan data studi kasus.
2. Mengimplementasikan algoritma *Greedy K-member clustering* pada Spark.
3. Melakukan perancangan dan implementasi perangkat lunak menggunakan *library* Scala-swing.
4. Mengimplementasikan teknik-teknik dasar *data mining* menggunakan *library* Spark MLlib.
5. Melakukan pengujian fungsional dan experimental.
6. Melakukan analisis kualitas informasi, sebelum dan setelah data dianonimisasi.
7. Menulis dokumen skripsi.

Bandung, 03/02/2020

Stephen Jordan

Menyetujui,

Nama: _____

Pembimbing Tunggal