Analisis Hasil Sequential Pattern Mining Menggunakan Eliminasi Gauss

Analisis Matriks

Fitrandi Ramadhan
Program Studi Magister Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 30132, Indonesia
23515050@std.stei.itb.ac.id

***Abstrak*—Persoalan data mining menjadi sebuah persoalan yang menjadi populer saat ini dikarenakan kemampuanya untuk mendapatkan informasi yang tidak didapatkan secara tradisional yaitu dengan mencari informasi dari informasi lainya. Akan tetapi kemampuan manusia untuk menelaah hasil dari data mining ini yang terbilang relative besar sangatlah terbatas. Disini akan dibahas mengenai apakah hasil dari data mining ini dapat diproses oleh computer dengan menggunakan metode eliminasi gauss dan mendapatkan hasil yang memberikan nilai lebih dari data yang didapat dari proses tersebut.**

Kata kunci—Data mining; Gaussian Elimination; customer analysis; matriks; paralel.

I. LATAR BELAKANG

Banyak persoalan di dunia pengetahuan saat ini tidak dapat lagi diselesaikan dengan menggunakan perhitungan yang dilakukan oleh manusia. Manusia memiliki kemampuan berfikir yang sangat tinggi, mampu melakukan analisis non-linear pada sebuah persoalan. Konteks persoalan yang akan digunakan saat ini adalah persoalan penyelesaian analitik metode numerik. Satu contoh yang akan diberikan adalah eliminasi gauss (Gaussian Elimination), persoalan ini membutuhkan kemampuan analisis non-linear dimana seseorang dapat memprediksi apakah setiap persamaan pada matriks dapat diselesaikan (satu solusi). Proses prediksi terhadap suatu persamaan untuk dilakukan pertukaran (pivoting) pun dapat dilakukan jauh sebelum proses perhitungan dilakukan. Akan tetapi kemampuan analitik manusia juga memiliki batas, pada persoalan yang membutuhkan proses panjang, kemampuan manusia semakin menurun. Hal yang dapat dijadikan contoh adalah pada metode newton-raphson dimana pada metode ini dibutuhkan banyak proses analitik untuk mendapatkan turunan dari sebuah fungsi, Tentunya fungsi yang dilakukan analitik bukanlah fungsi sederhana seperti

$$2x + 3y = c \qquad (1)$$

Berapakah rata-rata waktu yang dibutuhkan manusia untuk menyelesaikan persamaan diatas. Di lain sisi mesin (komputer) dapat melakukan perhitungan ekstensif dalam waktu yang sangat cepat. Di lain sisi mesin tidak memiliki kemampuan analitik yang tinggi. Kemampuan analitik mesin terbatas pada himpunan aturan yang telah didefinisikan pada mesin tersebut. Dengan kemampuan terbatas tersebut, perubahan atau penambahan aturan perlu dilakukan agar mesin dapat melakukan perhitungan yang diinginkan.

Data mining telah lama menjadi persoalan di dunia bisnis intelijen. Data mining sendiri adalah proses penemuan informasi baru dari himpunan informasi lainya. Pada hasil dari proses data mining ini diperlukan sebuah analisis kembali untuk memeriksa apakah hasil tersebut memberikan informasi baru tersebut. Beberapa dari proses data mining hanya dapat memberikan hasil data yang bagi pengguna data tersebut belum dapat diartikan menjadi sebuah informasi. Disinilah banyak penggunaan metode-metode khusus untuk menganalisis hasil tersebut. Saat ini data mining banyak digunakan untuk menganalisis proses jual beli yang terjadi di internet, dimana setiap transaksi dan aksi yang dilakan penjual dan pembeli dapat tercatat dengan mudah dan rinci. Ide ini dimunculkan dari sebuah penelitian mengenai perilaku pembeli yang didasarkan dari data aksi pembeli/calon pembeli dari suatu website jual beli tertentu. Proses analisis yang digunakan saat itu adalah memeriksa secara manual berapa panjang aksi yang dilakukan oleh calon pembeli sebelum akhirnya membeli suatu produk dari laman tersebut. Akan tetapi setiap laman tidak diberikan poin atau kecenderungan atas pembelian tersebut yang artinya hasil data mining tidak dapat dijadikan dasar mengenai apakah sebuah laman lebih cenderung menghasilkan pembelian produk atau tidak.

Disini penulis memberikan hipotesis bahwa apabila sebuah hasil dari data mining tersebut diproses lebih lanjut dengan metode gauss dimana setiap laman menjadi sebuah variable unik yang dipetakan pada hasil data mining tersebut, maka akan dapat dibuktikan bahwa laman tersebut merupakan laman

$$\alpha + \beta = \chi. \qquad (1888) \quad (1)$$

yang memberikan kecenderungan pembeli untuk membeli produk dari toko online tersebut. Disini akan didapatkan sebuah informasi yang memberikan nilai tambah atau bahan bagi analisis ekonomi lebih lanjut bagi pemilik toko.

II. STUDI PUSTAKA

A. *Paralelisasi*

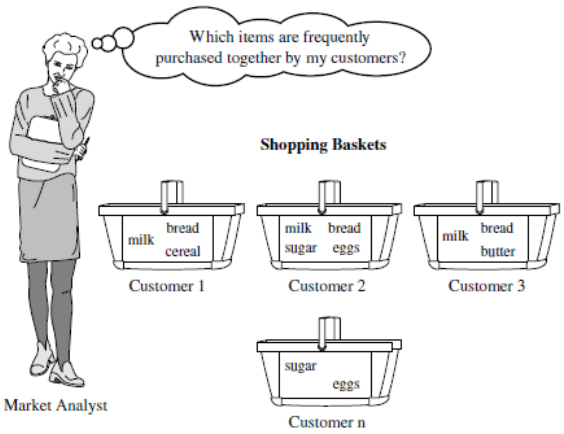
Sebuah contoh penyelesaian persoalan matekmatika adalah deret taylor telah menjadi salah satu metode penting dalam dunia matematika. Deret Taylor dirancang oleh James Gregory, seorang matematikawan Skotlandia dan diperkenalkan lebih lanjut oleh Brook Taylor pada tahun 1715. Untuk mendapatkan sebuah fungsi kontinyu sebuah pendekatan diskrit dapat dilakukan dengan melakukan interasi terbatas dengan setiap iterasi mendekati fungsi kontinyu. Hal ini menjadi penting dikarenakan setiap fungsi yang dilakukan pada mesin harus dilakukan dalam waktu yang terbatas. Sesuai dengan aturan yang berlaku pada mesin turing dimana apabila tidak ada mesin turing yang dapat menyelesaikan suatu persoalan, dalam kasus ini mesin turing akan melakukan looping tanpa akhir, maka tidak ada mesin (komputer) yang dapat menyelesaikan persoalan tersebut. Untuk itu diperlukan sebuah penyederhanaan persoalan menjadi sebuah persoalan yang diskrit. Pada aspek ini komputer memiliki kemampuan yang jauh lebih tinggi daripada manusia dimana komputer dapat melakukan iterasi penyelesaian persoalan dengan jumlah iterasi yang besar dalam waktu yang cepat.

Pada kasus saat ini yang akan dibahas lebih lanjut adalah penyelesaian persoalan nirlanjar. Penyelesaian persoalan ini dapat dilakukan dengan berbagai metode dengan setiap metode klasik dapat diaplikasikan kedalam mesin untuk mengoptimalkan penyelesaian persoalan dalam skala besar. Pada paper ini akan dibahas proses eliminasi gauss. Eliminasi gauss melakukan proses eliminasi matriks dari satu row ke row lainya sampai dengan seluruh row telah dieliminasi menjadi matriks segitiga bawah. Matriks segitiga ini akan memberikan solusi untuk setiap persamaan dengan N jumlah variabel dengan minimal N jumlah persamaan.

B. *Data Mining*

Data mining pun terus berkembang dengan lingkungan yang telah ada sekarang. Seperti yang telah dikemukakan diatas bahwa lingkungan yang saat ini banyak menjadi perhatian dari pelaku bisnis saat ini adalah lingkungan dari e-business itu sendiri, yaitu web. Manusia perlu tahu bagaimana cara mengekstrak informasi dari web-web ini. Apa saja yang dapat diketahui dengan melakukan mining pada web ini.

Setelah berbagai kemudahan yang ditawarkan oleh teknik data mining diatas, muncul pula berbagai metode-metode data mining yang dikenal pada saat ini. Artificial neural networks, Decision trees, dan Nearest-neighbor method adalah 3 contoh teknik data mining.



Gambar I.2 Market basket analysis

Salah satu contoh lainya adalah Sequential Pattern Mining. Sequential Pattern Mining adalah salah satu metode yang mengaplikasikan teknik data mining kepada basis data sekuensial untuk menemukan korelasi antara relasi yang ada diantara kejadian yang terjadi. Sequential pattern mining seringkali digunakan untuk mengekstraksi data yang terdapat di web. Sedangkan dengan banyaknya jenis data yang tersimpan didalam web itu sendiri perlu diketahuilah apakah sequential pattern mining ini dapat menjadi jawaban dari kebutuhan data mining saat ini, merupakan suatu pertanyaan yang perlu dijawab dan perlu diketahui pula jenis data apa yang cocok untuk digali menggunakan metode ini. Maka, yang menjadi latar belakang utama adalah kebutuhan akan data mining untuk mendapatkan suatu informasi tersebunyi yang mana pada akhirnya akan membantu penggunanya untuk meningkatkan bisnisnya.

III. ANALISIS DAN PERANCANGAN

Pada persoalan pola akses pembelian pada suatu laman jual beli dapat dianalisis apakah seorang pembeli cenderung membeli produk pada laman tersebut atau tidak. Hal ini dapat ditunjukkan dengan hasil implementasi

Pada implementasi normal algoritma persamaan direpresentasikan dalam bentuk matriks M x N (tinggi x lebar). M adalah jumlah variabel yang dicari dan N adalah panjang persamaan + dengan konstanta persamaan tersebut. Proses eliminasi dapat dilakukan secara sekuensial akan tetapi apakah proses sekuensial tersebut cukup menjadi pertanyaan untuk pengembangan algoritma dari metode eliminasi gauss ini kedepan. Algoritma dimulai dengan melakukan eliminasi dari kolom pertama atau variabel pertama untuk setiap row kecuali row pertama. Kemudian dilakukan eliminasi untuk kolom kedua pada seluruh row kecuali row pertama dan row kedua dst. Proses ini dilakukan sampai dengan row terakhir tereliminasi dan hanya menyisakan satu variabel dan satu konstanta. Dengan variabel ini dapat ditemukan maka satu per satu variabel dapat ditemukan. Pada jumlah yang sangat besar proses ini tidaklah efektif lagi dimana pengembangan multithreading pada proccessor menjadi berkembang pesat, muncul ide untuk melakukan paralelisasi pada algoritma-

algoritma tradisional. berikut pseudocode untuk eliminasi gauss secara sekuensial

```
//forward elimination
Matrix M;
  for k=0 to height
    for i=k+1 to height
      for j=i to width
        M[i][j] = M[i][j]-M[k][j]*M[i][k]/M[k][k]
      for j=0 to i
        M[i][j] = 0
//backward substitution
vector result = 0;
for i=height to 0
  sum = 0;
  for j=(width-2) to 0
    sum = sum + result[j]*M[i][j]
  result[i] = (M[i][width-1] - sum) / M[i][i]

return result
```

Pada pseudocode tersebut dapat terlihat bahwa proses k+1 membutuhkan matriks yang merupakan hasil dari proses k. Untuk itu secara analitik proses ini tidak dapat dilakukan paralelisasi. Sedangkan untuk proses i dan j, setiap proses dapat dilakukan paralelisasi dikarenakan tidak adanya konflik penulisan kedalam shared memory yang sama oleh lebih dari satu thread.

Pada penelitian ini akan digunakan sebuah library paralelisasi program berbasis bahasa C/C++ yaitu openmp. Proses penambahan proses parallel pada bahasa pemrograman yang digunakan bukan menjadi sesuatu yang sulit dimana hanya dibutuhkan penambahan 2 line of code dikarenakan library parallel yang digunakan.

```
//forward elimination
  Matrix temp(A);
  for(int k=0;k<temp.get_height()-1;k++) {
#pragma omp parallel for
   for(int i=k+1;i<temp.get_height();i++) {
   #pragma omp parallel for
        for(int j=k+1;j<temp.get_width()-1;j++) {
        temp.set_data(i,j,(temp.get_data(i,j)-
(temp.get_data(k,j)*temp.get_data(i,k)/temp.get_data(
k,k))));
                        }
        for(int j=0;j<k+1;j++) temp.set_data(i,j,0);
        }
   }
```

```
//backward substitution
vector<float> result;
   for(int              i=0;i<temp.get_width()-1;i++)
result.push_back(0);
        for(int i=temp.get_height()-1;i>=0;i--) {
        float sum = 0;
        for(int j=temp.get_width()-2;j>i;j--) {
                sum += result[j]*temp.get_data(i,j);
        }
        result[i]                                          =
(temp.get_data(i,temp.get_width()-1)-
sum)/temp.get_data(i,i);
   }

return result;
```

Kemudian muncul pertanyaan lebih lanjut apakah hasil data mining ini memiliki konstanta yang dapat dijadikan acuan untuk mendapatkan hasil yang diinginkan. Disini diperlukan beberapa asumsi atau penilaian mengenai hasil tersebut. Hasil data mining tersebut adalah proses navigasi pembeli pada suatu website jual-beli. Dan pada data tersebut telah tercatat hasil dari masing-masing penelusuran apakah pelanggan yang dianalisis melakukan transaksi atau tidak. Sehingga dapat dilakukan pemberian nilai pada setiap row dari matriks. Nilai satu diberikan untuk row dimana transaksi terjadi dan nilai 0 diberikan kepada row dimana transaksi tidak terjadi.

Dari hasil analisis ini diharapkan didapatkan nilai dari masing-masing variable dimana setiap variable merepresentasikan halaman yang diakses dan nilai dari masing-masing variable merupakan nilai kemungkinan terjadinya transaksi apabila pelanggan masuk ke halaman tersebut.

IV. IMPLEMENTASI

Terdapat 2 langkah utama untuk menyiapkan penggunaan metode ini dapat dilakukan yaitu.

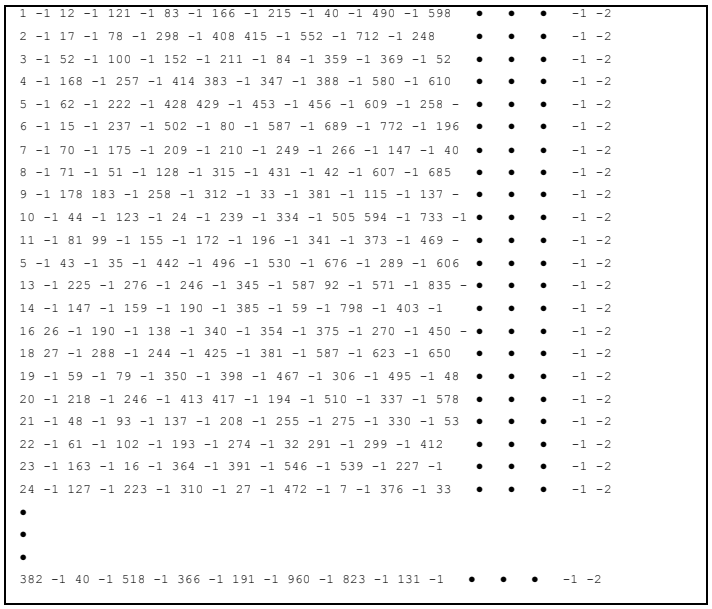
Setiap langkah dari proses tersebut adalah sebagai berikut.

- Pembangkitan file output.
- Setiap record (dalam konteks ini line pada file teks), akan dipindai dan dikenali 3 elemen utamanya yaitu indentifier (alamat IP), waktu (timestamp), dan tautan. Pemindaian dilakukan secara sekuensial (line per line) karena sifat dari web logs sendiri yang sudah berbentuk sekuensial dan setiap record (line) satu piece data yang berdiri sendiri.
- Pemeriksaan EoF dilakukan untuk memastikan bahwa data belum mencapai akhir dari file sebelum proses selanjutnya dilakukan. Apabila pembacaan telah sampai kepada EoF (end of file) maka proses akan diterminasi karena dianggap tidak ada lagi data yang dapat diproses. Apabila pembacaan belum mencapai EoF maka dapat diasumsikan bahwa file masih mengandung record yang

signifikan dalam proses transformasi ini dan proses dilanjutkan pada proses berikutnya.

- Dilakukan proses pemeriksaan identitas dari setiap record yang terdapat pada weblogs. Setiap IP address yang sama yang terdapat pada setiap record weblogs akan diasumsikan sebagai seorang pengguna yang sama. Dilakukanlah pemeriksaan apakah setiap dari IP Address, apakah pengguna yang terdapat pada record yang sedang diproses telah dimasukkan kedalam sequential database sebelumnya.
- Pengguna (IP) yang belum tercantum pada sequential database akan diinsersi sebagai sebuah record baru pada sequential database sebagai SID (Sequence ID). Sequence yang terdapat pada pengguna yang baru diinsersi ini masih kosong pada tahap ini.
- Setiap pengguna yang telah terdaftar pada sequential database tentunya telah memiliki tautan yang telah tercatat pada sequence nya. Setiap akses tentunya perlu dikelompokkan apakah setiap item dari sequence ini merupakan item dalam kelompok yang sama. Pada sequence akses ini setiap item pada sequence yang dilakukan pada waktu yang kurang dari 1 hari dari sequence sebelumnya akan dikelompokkan menjadi item-item dalam tuple yang sama.
- Seperti yang telah dijabarkan pada poin sebelumnya akan terdapat item-item yang terkelompokkan kedalam tuple-tuple yang sama dan berbeda. Untuk itu untuk representasi lebih lanjut pada sequence diperlukan separator atau pembatas antar item-item yang berada pada tuple yang berbeda satu dengan lainya. Pada implementasi ini separator didefinisikan sesuai dengan definisi separator pada SPMF library yaitu diimplementasikan dengan bilangan <-1>.
- Terlepas dari setiap tautan yang terdapat pada satu tuple yang sama ataupun berbeda satu dengan lainya, setiap tautan akan diinsersi secara mandatory pada sequence dari penggunanya masing-masing. Hal ini tidak akan berpengaruh pada pengelompokkan tautan-tautan karena separator pada poin sebelumnya telah menjadi representasi yang valid.
- Sebelum setiap tautan diinsersi kedalam setiap sequence perlu dilakukan mapping kepada represetasi yang sederhana dan sesuai dengan input yang diterima oleh SPMF library yaitu bilangan positif, -1, -2. Proses ini lah yang diselipkan pada transformasi weblogs ini untuk mengurangi proses pembacaan data lebih lanjut pada look-up table dan sequential database.
- Proses dilakukan dengan pembacaan record berikutnya pada weblogs sampai dengan pemindaian pada weblogs mencapai EoF (end of file). Seperti yang telah dijelaskan pada poin no.4, apabila proses pemindaian record weblogs telah sampai pada akhir dari file weblogs tersebut akan dilakukan terminasi.
- Penulisan file dilakukan dengan melakukan flush dari setiap record yang terdapat pada sequential database kedalam file teks yang telah dibangkitkan sebelumnya. Penulisan ini tetunya sudah disesuaikan dengan format setiap end of record dari sequential database yang dapat diterima oleh SPMF library.

Untuk penulisan file tersebut akan dilakukan kembali transformasi kedalam sebuah persamaan. Setiap matriks dibentuk dengan jumlah dari setiap variable



Setiap hasil yang didapatkan dapat symbol (215) atau (609) adalah sebuah variable tersendiri. Sehingga dapat dibentuk sebuah matriks dimana tinggi dari matriks tersebut adalah jumlah variable yang ada dan lebar adalah Jumlah variable dengan masing-masing row merepresentasikan kemunculan dari variable yang dimaksud pada setiap row yang ada.

Nilai yang menjadi representasi terjadinya pembelian atau tidak adalah nilai boolean dimana pada awal proses diberikan nilai true pada sekuens terjadinya pembelian dan nilai false pada row yang tidak memberikan pembelian. Hal ini akan memberikan hasil nilai yang dimiliki oleh masing-masing laman yang tercatat pada hasil sequential pattern mining menggunakan SPMF ini.

V. KESIMPULAN

Kesimpulan yang didapatkan dari hasil penelitian ini adalah dimana dapat dilakukan operasi Gaussian elimination pada hasil proses data mining menggunakan sequential pattern mining pada data aksi pengguna laman jual beli. Akan tetapi proses terhambat oleh banyaknya data dan kelengkapan dari setiap variable seringkali tidak memenuhi persyaratan dilakukan nya eliminasi gauss tersebut. Pada kasus dimana terdapat banyak variable yang bernilai nol pada suatu row nya hal ini akan menyebabkan variable tersebut memiliki banyak nilai yang memenuhi sehingga tidak dapat ditentukan secara pasti apakah laman tersebut merupakan laman yang signifikan atau tidak.

Secara teori apabila hasil dari data mining ini memenuhi keseluruhan syarat sebuah himpunan persamaan dilakukan operasi gauss akan dapat dihasilkan nilai dari masing-masing variable yang ada dalam hal ini laman tersebut. Sehingga laman tersebut dapat diberikan poin atas kecenderunganya untuk membuat transaksi

REFERENCES

[1] Chapra, Steven C. "Applied Numerical Method,with MATLAB 3rd Edition".

[2] Atkinson, Kendall A. (1989), *An Introduction to Numerical Analysis* (2nd ed.), New York:

[3] Calinger, Ronald (1999), *A Contextual History of Mathematics*, Prentice Hall.

[4] G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

[5] Gottlieb, Allan; Almasi, George S. (1989). Highly Parallel Computing. Redwood City, Calif.: Benjamin/Cummings.

[6] Han, Jiawei and Micheline Kamber (2006). Data Mining : Concepts and Techniques, San Fransisco: Morgan Kaufmann Publishers.

[7] Han, Jiawei, dkk. Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. Hawlett-Packard Labs. Palo Alto, California

[8] Agrawal, Rakesh, Srikant, Ramakrishnan. Mining Sequential Patterns. San Jose California. IBM Almaden Research Center

[9] Michael McCool; James Reinders; Arch Robison (2013). *Structured Parallel Programming: Patterns for Efficient Computation*. Elsevier.