ANALISIS DAN IMPLEMENTASI SEQUENCE PATTERN MINING PADA DATA SEQUENTIAL MULTIDIMENSIONAL MENGGUNAKAN ALGORITMA PREFIXMDSPAN

Alinda Prima Damayanti[1], Kemas Rahmat Saleh Wiharja[2], Intan Nurma Yulita[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Knowledge Discovery from Database (KDD)berfungsi untuk menghasilkan pengetahuan dari sekumpulan data yang terdapat di dalam satu atau lebih basis data. Salah atu prosesnya adalah data mining. Terdapat beberapa jenis pola hasil proses data mining, diantaranya sequential pattern. Sequential pattern menggambarkan keterurutan suatu peristiwa yang terjadi beberapa kali dalam kurun waktu tertentu. Terdapat beberapa Algoritma untuk melakukan Sequential Pattern Mining, diantaranya PrefixSpan.

Tugas Akhir ini mengkaji Algoritma PrefixMDSpan yang merupakan turunan dari Algoritma PrefixSpan (Prefix-projected Sequential pattern mining) yang digunakan pada data sequential multidimensional.

Hasil pengujian menunjukkan bahwa jumlah minimum support mampu membatasi menarik tidaknya suatu pola yang dihasilkan dan dari pola-pola tersebutdidapatkan informasi yang bisa digunakan untuk menentukan sebuah kebijakan baru pada periode berikutnya atau yang akan datang untuk item-item tertentu.

Kata Kunci : Knowledge Discovery from Database, data mining, Sequential pattern, PrefixSpan, PrefixMDSpan, sequential multidimensional, minimum support.

Abstract

Knowledge Discovery from Databases (KDD) function to generate knowledge from a collection of the data contained in one or more databases. One of the process is data mining. There are several types of patterns the data mining process, including sequential pattern. Sequential patterns describe pattern an event that occurred several times within a certain time. There are several algorithms for Sequential Pattern Mining, such as PrefixSpan.

This final project review PrefixMDSpan algorithm which is derived from Algorithm PrefixSpan (Prefix-projected Sequential pattern mining) that are used in sequential multidimensionaldata.

Test results showthat theminimum amount ofsupportcanlimit thedrawwhether or not apattern thatis producedandfromthe patternsobtainedinformation that couldbe usedtodetermineanewpolicyin the next periodor willcometocertain items.

Keywords : Knowledge Discovery from Database, data mining, Sequential pattern, PrefixSpan, PrefixMDSpan, sequential multidimensional, minimum support.

1. PENDAHULUAN

1.1 Latar Belakang

Kegiatan manusia semakin berkembang dan memunculkan kebutuhan akan data semakin berkembang pula. Penggunaan data tumbuh begitu cepat, mengakibatkan semakin besarnya jumlah data yang harus dikumpulkan dan disimpan. Basis data yang digunakan juga semakin berkembang dari segi ukuran yang akan semakin besar dan semakin beragam jenisnya. Dengan tersediannya data yang sangat melimpah tersebut ternyata informasi yang dapat digali terbatas karena keterbatasan manusia untuk melihat seluruh data dan menjadikannya informasi yang lebih bermanfaat secara langsung. Diperlukan sebuah aplikasi yang mampu mengubah data yang begitu besar menjadi informasi yang bernilai, yaitu melalui *data mining*.

Data mining dideskripsikan sebagai proses untuk melakukan ekstraksi pengetahuan dari data yang jumlahnya besar. Banyak pihak yang menggunakan data mining sebagai suatu langkah *knowledge Discovery in Database* atau KDD. *Data mining* adalah sebuah proses essensial dimana metode-metode intellegent digunakan untuk mengekstraksi pola-pola(pattern) yang terbentuk dari kata.

Terdapat beberapa jenis pola hasil proses *data mining*, diantaranya *sequential pattern*. *Sequential pattern* menggambarkan keterurutan suatu peristiwa yang terjadi beberapa kali dalam kurun waktu tertentu. Contoh dari penggunaan *sequential pattern* adalah pada bisnis retail. Apabila terdapat suatu pola, yaitu 'seorang pelanggan yang membeli raket tenis bulan ini akan membeli bola tenis bulan depan',pola tersebut adalah Sequential Pattern. *Sequential Pattern* dapat digunakan untuk memeprediksi kejadian yang dapat terjadi berikutnya ataupun mengidentifikasi pengulangan kejadian dari data yang dinamis.

Dengan struktur data multidimensional, pengguna dapat menampilkan data dengan dimensi yang lebih luas, contohnya data pembelian berdasar jumlah penjualan per-counter (lokasi toko) karakteristik customer dan juga, data dapat ditampilkan per harian, bulanan, tahunan atau tingkatan waktu yang lain yang dibutuhkan pengguna.

Pada kondisi tertentu, data multidimensional akan memunculkan juga data *Sekuensial Multidimensional*. Sekuensial ini muncul ketika dimensi yang digunakan dalam menampilkan data adalah beragam segi dan kriteria. Misalkan setiap atribut yang digunakan diasumsikan memiliki 1 dimensi, dan di dalam atribut tersebut memiliki dimensi lagi sesuai dengan apa yang kita inginkan.Dari contoh kejadian inilah, diperlukan suatu usaha tersendiri untuk dapat melihat keterurutan berdasarkan kriteria yang berbeda.

Oleh karena itu dibutuhkan sebuah program yang dapat digunakan pada kasus data *sequential multidimensional* sehingga dapat dilihat keterurutan berdasarkan kriteria yang berbeda. Pada kasus ini kriteria yang ingin dibahas adalah mengenai perbedaan dimensi waktu, sehingga didapatkan pengetahuan dari setiap dimensi waktu tersebut.

Terdapat beberapa algoritma untuk melakukan *sequential pattern mining*, antara lain *Apriori-All*, *Apriori-Some*, *FreeSpan*, *PrefixSpan* dan lain-lain. Namun yang digunakan dalam tugas akhir ini adalah *PrefixMDSpan* (*Prefix-projected Sequential pattern mining*), karena terbukti bahwa algoritma *PrefixMDSpan* cenderung lebih baik dari algoritma lainnya [10].

1.2 Perumusan masalah

Terdapat banyak literatur yang membahas mengenai pengembangan *sequential pattern*. Namun beberapa hanya membahas teori dan algoritma umumnya saja. Selain itu, karena data yang digunakan multidimensional, diperlukan strategi implementasi agar waktu proses *sequential pattern* menjadi lebih singkat dan *sequential pattern* yang dihasilkan menjadi sebuah informasi yang bermanfaat.

Pembahasan masalahnya sebagai berikut :

1. Mengimplementasikan teknik *sequential pattern mining* pada data *sekuensial multidimensional* berdasarkan algoritma *PrefixMDSpan* dan mengimplementasikannya menjadi sebuah program yang dapat digunakan.
2. Mendapatkan informasi yang bermanfaat dan dapat digunakan dari sequential pattern yang dihasilkan oleh program.

Sebagai bagian dari teknik ekstraksi pengetahuan dari data, proses yang dilakukan dalam tugas akhir ini meliputi proses data mining dengan pendekatan sequential pattern mining pada kumpulan data. Batasan masalahnya antara lain :

1. Aplikasi ini menggunakan sumber data yang telah siap digunakan (data fix).
2. Implementasi fokus pada data sequensial multidimensional.
3. Tidak membahas mengenai teknik-teknik optimasi program.

1.3 Tujuan

Tujuan yang ingin dicapai dalam tugas akhir ini adalah memperoleh pemahaman terhadap data *sequensial multidimensional* dan teknik yang dapat digunakan untuk melakukan *sequential pattern mining* pada data *sequensial multidimensional*.

Tujuan umum yang ingin dicapai adalah :

1. Melakukan pengujian dengan parameter waktu eksekusi algoritma yang dipengaruhi oleh beberapa karakteristik seperti jumlah*record*, jumlah *dimens*i, panjang *sequence* rata-rata yang terlibat dan *minimum support*.
2. Mendapatkan pengetahuan berupa informasi yang dari *sequential pattern* yang dihasilkan sehingga memberi suatu pengetahuan bermanfaat dan dapat diterapkan pada periode waktu selanjutnya ataupun yang akan datang.

1.4 Metodologi penyelesaian masalah

Untuk menyelesaikan pembuatan tugas akhir yang merupakan tujuan dari dibuatnya proposal ini, akan dilakukan beberapa langkah kerja sebagai berikut:

1. Eksplorasi dan Stuli Literatur
   Tahap ini dilakukan dengan cara mempelajari cara kerja PrefixSpan dan PrefixMDSpan dari literatur-literatur seperti buku (textbook), paper dan sumber lain seperti website, artikel dokumen teks yang berhubungan.
2. Analisis
   Dari analisis literatur, dilakukan analisis tentang pembentukan data sekuensial multidimensional. Analisis juga meliputi pemahaman mendetail tentang sekuential pattern mining pada bentuk data sekuensial multidimensional yang diajukan, dan aspek-aspek yang meliputinya.
3. Perancangan Sistem
   Dilakukan perancangan yang mampu mengakomodasi penerapan teknik sequential pattern mining pada data sekuensial multidimensional dalam sebuah aplikasi perangkat lunak.
4. Implementasi

Implementasi dilakukan untuk mewujudkan perangkat lunak yang telah dirancang untuk penerapan teknik sequential pattern mining pada data sekuensial multidimensional.

5. Pengujian
   Pengujian dilakukan untuk memperoleh kesimpulan dari analisis dan sebagai bahan evaluasi terhadap penerapan teknik sequential pattern mining yang dilakukan.

1.5 Sistematika Pembahasan

Sistematika laporan tugas akhir ini adalah sebagai berikut :

BAB 1 Bab pertama membahas latar belakang, rumusan masalah, tujuan, batasan, dan metodologi pelaksanaan tugas akhir.

BAB 2 Bab kedua membahas dasar teori untuk sequence patter mining termasuk definisi multidimensional sequence pattern.

BAB 3 Bab ketiga adalah analisis tentang PrefixSpan, analisis tentang multidimensional sequence pattern mining dan termasuk perancangan sistem yang akan di implementasikan.

BAB 4 Bab keempat fokus pada implementasi dan pengujian, termasuk strategi pengembangan untuk menerapkan algoritma PrefixMDSpan dalam sebuah perangkat lunak dan dilakukan evaluasi terhadap hasil pengujian dan implementasi.

BAB 5 Bab kelima akan menutup laporan tugas akhir ini dengan kesimpulan dan saran.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berikut beberapa kesimpulan yang dapat diambil dari proses studi dan implementasi teknik *data mining* PrefixMDSpan pada basis data multidimensional:

1. Waktu eksekusi algoritma dipengaruhi beberapa karakteristik data yaitu *jumlah data, panjang rata-rata sequence,* dan *jumlah item yang terlibat*. Perbedaan jumlah dimensi *tidak* memberikan dampak signifikan, sedangkan *minimum support* digunakan untuk membatasi ketertarikan terhadap sebuah *pattern* sehingga akan mempersempit jumlah *sequential pattern* yang dihasilkan.
2. *Minimum support* yang menghasilkan sequence pattern output dengan jumlah item terbanyak adalah terjadi pada saat *minimum support*nya terkecil dan batas *minimum support* tertinggi yang dapat ditentukan adalah tidak melebihi *total record* yang ada pada data tersebut.
3. Semakin kecil *minimum support* yang ditentukan, semakin banyak pula *sequence pattern* yang dihasilkan, dengan konsekuensi waktu proses pun akan lebih lama dibandingkan dengan *minimum support* yang lebih tinggi.
4. Persentase perulangan pola yang dihasilkan pada tahun 1997 dibandingkan dengan hasil sequence pattern pada tahun 1998 mendekati sama dan hasil *sequence patt*ern *output* program sama dengan hasil yang didapatkan melalui perhitungan manual sehingga dapat dikatakan bahwa performasi program ini dengan menggunakan algoritma PrefixMDSpan cukup baik untuk digunakan pada kasus data *sequential multidimensional.*
5. *Item* barang yang paling banyak dibeli oleh pelanggan untuk kasus data yang digunakan menunjukkan bahwa *item* "0" paling banyak dibeli oleh sejumlah pelanggan pada waktu yang bersamaan.
6. Pada tahun 1997-1998 ada sekitar lebih dari 80% pelanggan yang membeli item "0".
7. Dengan aplikasi ini dapat diketahui jenis barang apa saja yang sering dibeli oleh konsumen yang nantinya informasi ini dapat memberikan pertimbangan tambahan bagi manajer sebuah perusahaan dalam pengambilan keputusan guna penyediaan jumlah barang yang harus disediakan dan penentuan harga yang harus diberikan pada periode waktu berikutnya atau yang akan datang.

5.2 Saran

Berikut beberapa saran yang perlu diperhatikan sebagai langkah lebih lanjut dalam studi teknik *sequential pattern mining* pada data multidimensional :

1. Penggunaancontoh kasus sesungguhnya mungkin bisa memberikan hasil implementasi yang lebih baik.
2. Usaha peningkatan kinerja teknik ini dapat menjadi satu bahan kajian tersendiri karena pada saat ini kinerja teknik masih belum optimal. Dengan data berukuran besar, diperlukan waktu yang cukup lama untuk menyelesaikan proses *mining*.
3. Selain itu, visualisasi *sequence* dalam bentuk grafik dan memiliki banyak pilihan menu tampilan yang lebih mudah dimengerti juga dapat menjadi sebuah kajian tersendiri. Saat ini masih digunakan representasi visual terbatas dalam bentuk teks.

Daftar pustaka

[1] Agrawal, Rakesh and Ramakhrisnan Srikant. (1995). *Mining Sequential patterns.* Proc. 1995 Int‟l Conf. Data. Eng. (ICDE „95).

[2] Agrawal, Rakesh and R. Srikant. (1995). *Fast algorithms for mining association rules. VLDB'94.*

[3] Han,Jiawei, J. Pei, and Y. Yin. (2001). *Mining frequent patterns without candidate generation. SIGMOD'00.*

[4] Han, Jiawei, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M.-C. (2001). Hsu. *FreeSpan: Frequent pattern-projected sequential pattern mining. KDD'00.*

[5] Han, Jiawei and Kamber, Micheline. (2001). *Data mining Concepts and Techniques*. Morgan Kaufmann.

[6] Han, Jiawei, G. Dong, and Y. Yin. (2001) Efficient mining of partial periodic patterns in time series database. ICDE'99.

[7] Pei, J., Jiawei Han, et all. (2001). *PrefixSpan: Mining Sequential patterns Efficiently by Prefix-Projected Pattern Growth*. Proc 17th int‟l Conf. Data. Eng.

[8] Pinto, Helen, Jiawei Han, Jian Pei, Ke Wang, Qiming Chen, Uweshwar Dayal. *Multi-dimensional Sequential Pattern* .ppt .

[9] Tan, Pang-Ning, Michael Steinbach, Vipin Kumar.(2006). *Introduction to Data Mining*. Adison Wesley.

[10] Yu,Chung-Ching,Yen-Liang Chen. (2005). *Mining Sequential pattern from Multidimensional Sequence Data*. IEEE Transactions on Knowledge and Data Engineering vol. 17, no. 1.