PEMBANGKIT DATA OTOMATIS BERBASIS POLA DISTRIBUSI POISSON UNTUK KEBUTUHAN PENGETESAN PERANGKAT LUNAK DATA MINING DALAM PENCARIAN POLA ASOSIASI DAN POLA SEKUENSIAL

Arif Djunaidy, Rully Soelaiman dan Adhita Pratiwi

Fakultas Teknologi Informasi - Institut Teknologi Sepuluh Nopember
Kampus ITS, Jl.Raya ITS – Sukolilo, Surabaya 60111, Indonesia
Telp. (031) 5939214, Fax. (031) 5939363
Email: arif@its-sby.edu

ABSTRAK

Data transaksi tiruan yang menyerupai transaksi nyata pada lingkungan ritel dibutuhkan dalam pengetesan teknik data mining untuk pencarian pola asosiasi dan pola sekuensial dari basis data berskala besar. Dalam dunia nyata, terdapat kecenderungan bahwa pembeli melakukan pembelian beberapa item secara bersamaan dengan ukuran transaksi terkelompok di sekitar nilai rerata banyaknya item yang dibeli dan membentuk pola distribusi Poisson..

Makalah ini membahas pengembangan pembangkit data otomatis yang mengikuti pola distribusi Poisson untuk kebutuhan pengetesan perangkat lunak data mining dalam pencarian pola asosiasi dan pola sekuensial. Dalam proses pembangkitan data, perangkat lunak ini menggunakan beberapa parameter, seperti jumlah item, ukuran rerata itemset, ukuran maksimum large itemset, jumlah large itemset, ukuran rerata transaksi, ukuran maksimum transaksi, dan jumlah transaksi. Sedang tahapan pembuatan transaksi tiruan meliputi pembentukan item yang akan dimasukkan ke dalam transaksi, pembuatan large itemset dari kumpulan item, dan pembuatan transaksi. Ukuran masing-masing itemset dan transaksi didasarkan pada pola distribusi Poisson dengan rerata sama dengan ukuran rerata large itemset/transaksi.

Uji coba perangkat lunak yang dilakukan terhadap berbagai nilai parameter membuktikan bahwa pembangkit data otomatis mampu menghasilkan data transaksi tiruan dalam jumlah besar dengan waktu komputasi yang relatif singkat. Hasil uji coba menunjukkan bahwa (a) semakin besar ukuran rerata transaksi, semakin besar pula jumlah record, waktu pembuatan dataset, ukuran basis data, dan jumlah frequent itemset yang ditemukan, (b) semakin besar jumlah transaksi, semakin besar pula jumlah record, waktu pembuatan dataset, dan ukuran basis data yang dihasilkan, dan (c) semakin besar jumlah itemset yang dibuat, semakin sedikit jumlah pola yang ditemukan.

Kata kunci : pembangkit data otomatis, pola distribusi poisson, data mining, pola asosiasi, pola sekuensial.

1. PENDAHULUAN

Perkembangan teknologi dan sistem informasi telah memungkinkan berbagai perusahaan ritel untuk mengumpulkan data transaksi jual-beli dalam jumlah besar dan menyimpannya dalam bentuk basis data. Sebuah record pada basis data tersebut biasanya terdiri dari tanggal transaksi, kode transaksi dan item–item yang dibeli pada transaksi itu. Pada mulanya, basis data transaksi dalam jumlah besar itu disimpan hanya untuk keperluan pengarsipan. Namun seiring dengan perkembangan manajemen basis data yang sangat pesat dewasa ini, muncullah kebutuhan untuk memanfaatkan basis data transaksi dalam jumlah besar itu sebagai bahan analisa untuk memperoleh informasi yang bermanfaat bagi proses pengambilan keputusan organisasi [4].

Data mining muncul sebagai jawaban atas permasalahan tersebut. Data mining mempelajari proses ekstraksi informasi yang bermanfaat dan potensial dari sekumpulan data yang secara implisit terdapat dalam suatu basis data. Dalam kaitan ini, data transaksi yang harus dianalisa biasanya melibatkan basis data dalam ukuran yang sangat besar.

Dua diantara sejumlah teknik *data mining* yang berkembang saat ini adalah pencarian pola-pola asosiasi (*mining association rules*) dan pencarian pola-pola sekuensial (*mining sequential patterns*) dari basis data berskala besar. Pencarian pola asosiasi bertujuan untuk menemukan hubungan–hubungan antar objek dalam basis data yang dievaluasi, misalnya 95% orang yang membeli kopi

dan gula ternyata juga membeli susu atau 90% orang yang membeli roti juga membeli selai [12]Sedangkan pencarian pola-pola sekuensial bertujuan untuk menemukan pola-pola hubungan antar objek yang berurutan, misalnya sebagian besar orang cenderung akan membeli buku cerita "Star Wars", yang kemudian diikuti dengan pembleian buku "Empire Strikes Back" dan baru membeli buku "Return of the Jedi" [3]. Dari hasil analisa tersebut, manajemen perusahaan akan memperoleh informasi tentang pola transaksi yang dilakukan pelanggan sehingga perusahaan dapat menerapkan strategi manajemen yang tepat, seperti penambahan stok barang, pengaturan tata letak barang berdasarkan pola beli konsumen, dan pembuatan katalog produk serta segmentasi pasar.

Oleh karena basis data yang terlibat dalam data mining ini berukuran sangat besar, maka teknik data mining yang digunakan haruslah melibatkan algoritma yang dapat memberikan kinerja yang efisien dan realistik untuk data transaksi yang berukuran sangat besar. Untuk itu pengembang data mining memerlukan data transaksi tiruan dalam jumlah besar untuk mensimulasikan teknik data mining yang dikembangkannya.

Data transaksi tiruan tersebut haruslah meniru tipe transaksi pada lingkungan ritel, dimana para pembeli cenderung untuk membeli sekumpulan item bersamaan [2] dimana banyaknya item pada setiap transaksi bervariasi dan terklaster di sekitar nilai rerata. Pola distribusi Poisson merupakan distribusi peluang yang sesuai untuk memodelkan banyaknya kejadian yang terjadi pada suatu interval waktu[9][12]. Dengan menggunakan distribusi Poisson dapat ditentukan banyaknya item pada suatu transaksi.

Dalam pembangkit data otomatis yang pernah dibuat sebelumnya, ukuran transaksi dan jumlah item pada data transaksi yang dihasilkan ditentukan secara pasti oleh pengguna, sehingga hasilnya tidak terdistribusi dengan baik dan bersifat uniform [7]. Pembangkit data otomatis ini tidak dapat menghasilkan data tiruan yang menyerupai data transaksi sebenarnya, karena data yang dihasilkan masih bersifat uniform dan tidak mengikuti pola distribusi yang sesuai, walaupun pengguna dapat memasukkan probabilitas kemunculan suatu item. Atas dasar inilah, maka dikembangkan perangkat lunak untuk membuat data transaksi yang mampu meniru pola data transaksi pada lingkungan *ritel* untuk keperluan simulasi *data mining*.

Pembahasan selanjutnya dari makalah ini disusun seperti berikut. Bab 2 membahas mengenai karakteristik dasar dari pola asosiasi dan pola sekuensial. Sifat statisitik dari transaksi dijelaskan dalam bab 3. Bab 4 dan 5 berturut-turut menjelaskan mengenai pembuatan data tiruan dan detil hasil uji coba dari pembangkit data otomatis yang berhasil dikembangkan. Akhirnya, kesimpulan yang dapat ditarik dari pembuatan pembangkit data otomatis ini dibahas dalam bab 6.

2. POLA ASOSIASI DAN SEKUENSIAL

Data mining merupakan sebuah proses ekstraksi informasi yang potensial, implisit, dan tidak diketahui sebelumnya –seperti pola-pola, batasan–batasan, dan regularitas– yang diperoleh dari sekumpulan data pada basis data besar [5]. Dalam kaitan ini, terdapat beberapa teknik data mining yang berkembang saat ini, seperti pencarian pola asosiasi, pencarian pola sekuensial, klasifikasi data dan klasterisasi data. Dalam sub-bab selanjutnya diuraikan mengenai pencarian pola asosiasi dan pola sekuensial pada basis data besar. Teknik yang lain tidak dibahas karena data transaksi tiruan yang dibuat dikhususkan untuk keperluan pencarian pola asosiasi dan pola sekuensial.

2.1 Pola Asosiasi

Pola asosiasi mendeskripsikan hubungan asosiasi antar item dalam suatu basisdata transaksi, dimana jika beberapa item dibeli di sebuah transaksi maka item-item lain juga ikut dibeli [1]. Untuk menemukan pola asosiasi, pertama-tama harus dicari sejumlah itemset yang cukup sering terjadi dalam basis data transaksi. Itemset adalah sekumpulan item yang dibeli secara bersamaan. Setelah sejumlah itemset ditemukan, pola asosiasi dapat diperoleh dengan cara seperti berikut. Jika itemset yang ditemukan adalah Y $\{Y = I_1, I_2, ..., I_k \mid k \geq 2\}$, maka pola asosiasi item-item dari himpunan $\{I_1, I_2, ..., I_k\}$ dapat dicari. Anteseden dari pola ini adalah X, yang merupakan subset dari Y, dan konsekuennya adalah Y-X. Suatu pola asosiasi $X \rightarrow Y - X$ dimiliki oleh basis data transaksi D dengan *confidence factor* c jika terdapat paling tidak c% transaksi di D yang mengandung X juga mengandung Y-X. Salah satu contoh dari pola asosiasi tersebut adalah: 95% dari transaksi dimana kopi dan gula dibeli, susu juga ikut terbeli. Bentuk dari pola asosiasi ini adalah "kopi, gula → susu" . Anteseden dari pola ini adalah kopi dan gula, sedangkan konsekuennya terdiri dari susu. Prosentase sebesar 95% merupakan *confidence factor* dari pola tersebut.

Sebuah transaksi dikatakan mendukung (sebagai *support*) sebuah itemset Z, jika Z terdapat pada

transaksi tersebut [1]. Nilai *support* untuk sebuah itemset didefinisikan sebagai rasio dari jumlah transaksi yang memiliki itemset ini terhadap jumlah seluruh transaksi dalam D. Dengan demikian permasalahan utama dalam pencarian pola asosiasi adalah untuk menemukan semua itemset yang memenuhi *minimum support* yang ditentukan oleh pengguna. Setiap itemset semacam ini diistilahkan sebagai *large itemset*.

2.1 Pola Sekuensial

Pencarian pola-pola sekuensial bertujuan untuk menemukan pola-pola hubungan antar objek yang berurutan (sekuensial) [3]. Contoh dari sebuah pola sekuensial adalah seorang pelanggan yang membeli buku tentang "Konsep Dasar Komputer" dan kemudian tentang "Bahasa Pemrograman" dan kemudian "Sistem Pemrograman" secara berurutan.

Penemuan pola sekuensial dapat dijelaskan seperti berikut [3]. Sebuah *sequence* (urutan) ialah sebuah *ordered list itemset*. Sebuah *sequence s* dinyatakan sebagai $<s_1, s_2, ..., s_n>$, dimana s_j adalah sebuah itemset. Item-item pada s_j menyatakan bahwa item-item tersebut dibeli secara bersamaan. Pada sekumpulan *sequence*, sebuah *sequence s* dikatakan maksimal jika *s* tidak terdapat pada *sequence* lain. Semua transaksi yang dilakukan pembeli, yang diurutkan berdasarkan waktu transaksi disebut sebagai sebuah *customer-sequence*.

Seorang pelanggan dikatakan mendukung (sebagai suppoort) suatu *sequence s*, jika *s* terdapat pada *customer-sequence* untuk pelanggan tersebut. Nilai *support* untuk sebuah *sequence* didefinisikan sebagai rasio dari pelanggan yang mendukung *sequence* tersebut terhadap jumlah seluruh *customer-sequence* yang ada. Setiap *sequence* yang memenuhi minimum *support* diistilahkan sebagai *large sequence*.

3. SIFAT STATISTIK DARI TRANSAKSI

Dalam dunia nyata, seseorang cenderung membeli beberapa item sekaligus pada satu kali transaksi [2]. Untuk memodelkan kecenderungan ini, digunakan distribusi peluang kontinyu (distribusi Poisson) untuk menentukan banyaknya item yang dibeli oleh seorang pelanggan pada satu kali transaksi [12].

Percobaan Poisson merupakan percobaan yang menghasilkan variabel acak X yang bernilai numerik yang banyaknya hasil selama selang waktu tertentu atau dalam daerah tertentu. Panjang selang waktu tersebut dapat terjadi dalam satu menit, satu hari, satu minggu, satu bulan atau bahkan satu tahun. Jadi percobaan Poisson dapat menghasilkan pengamatan untuk variabel acak X yang menyatakan banyaknya item pada sebuah large itemset. Percobaan Poisson ini berasal dari proses Poisson dan memiliki sifat seperti berikut :

- Banyaknya hasil yang terjadi dalam suatu selang waktu atau daerah tertentu tidak terpengaruh oleh (bebas dari) apa yang terjadi pada selang waktu atau daerah lain yang terpisah. Dalam hubungan ini proses Poisson dikatakan tak punya ingatan.
- Peluang terjadinya suatu hasil (tunggal) dalam selang waktu yang amat pendek atau dalam daerah yang kecil sebanding dengan panjang selang waktu atau besarnya daerah dan tidak tergantung pada banyaknya hasil yang terjadi di luar selang waktu atau daerah tersebut.
- Peluang terjadinya lebih dari satu hasil dalam selang waktu yang pendek atau daerah yang sempit tersebut dapat diabaikan.

Banyaknya hasil X dalam suatu percobaan Poisson disebut variabel acak Poisson dan distribusi peluangnya disebut distribusi Poisson. Distribusi peluang variabel acak Poisson X, yang menyatakan banyaknya sukses yang terjadi dalam suatu interval atau jangkauan tertentu dinyatakan dengan t, dapat dinyatakan sebagai persamaan berikut [6]:

$$p(x;\lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}, x = 0, 1, 2, 3, ... \qquad (1)$$

Sebagai satu ilustrasi sederhana, perhatikan hasil pengamatan seorang manajer pada sebuah supermarket, dimana diasumsikan rerata sebuah item dari 100 item akan dibeli oleh seorang pelanggan. Jika dilakukan pengamatan terhadap 300 transaksi yang dilakukan oleh pelanggan, berapakah probabilitas dibelinya item sebanyak 1, 2, 3, 4, 5 pada masing-masing transaksi yang dilakukan ?

Pada ilustrasi tersebut, n (jumlah percobaan) = 300, p (peluang) = 1/100, $np=\lambda=300/100=3$, maka berdasarkan persamaan (1) akan diperoleh distribusi Poisson seperti berikut :

$$\begin{cases} p(1;3) = \dfrac{e^{-3}(3)^1}{1!} = 0{,}14936121 \\ p(3;3) = \dfrac{e^{-3}(3)2}{2!} = 0{,}2240418 \\ p(3;3) = \dfrac{e^{-3}(3)^3}{3!} = 0{,}2240418 \\ p(4;3) = \dfrac{e^{-3}(3)^4}{4!} = 0{,}16803135 \\ p(5;3) = \dfrac{e^{-3}(3)^5}{5!} = 0{,}10081881 \end{cases} \quad (2)$$

4. PEMBUATAN DATA TIRUAN

Dalam bagian ini dibahas konsep pembuatan data tiruan yang didasarkan pada pola distribusi poisson. Metode yang digunakan dalam pembuatan data tiruan didasarkan pada pendekatan yang digunakan dalam [12].

4.1 Pembuatan Data untuk Pola Asosiasi

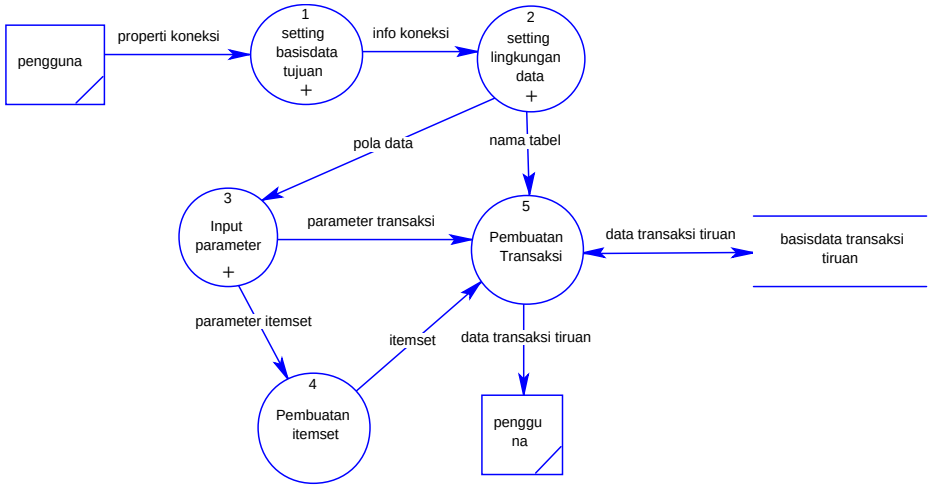
Data transaksi tiruan dibuat menyerupai transaksi pada lingkungan ritel, dimana pada dunia nyata orang cenderung membeli sekumpulan item secara bersamaan [2]. Item-item seperti itu secara potensial membentuk kumpulan item yang disebut *large itemset*. Sebuah contoh dari kumpulan item semacam itu adalah sprei, sarung bantal, dan selimut. Bagaimanapun juga, beberapa orang mungkin hanya membeli beberapa item dari sebuah kumpulan item. Sebagai contoh, beberapa orang mungkin hanya membeli sprei dan sarung bantal, sementara beberapa orang hanya membeli sprei. Sebuah transaksi dapat terdiri dari lebih dari satu large itemset. Sebagai contoh, seseorang mungkin membeli baju dan jaket disamping membeli sprei dan selimut, dimana baju dan jaket itu sendiri membentuk sebuah large itemset. Ukuran transaksi berkisar di sekitar nilai rerata ukuran transaksi dengan beberapa transaksi memiliki banyak item. Ukuran *large itemset* juga berkisar di sekitar nilai rerata ukuran *large itemset*, dengan beberapa *large itemset* terdiri dari sejumlah besar item.

Parameter–parameter yang digunakan pada pembuatan data transaksi tiruan ini dapat dilihat dalam Tabel 1. Sedangkan metode pembuatan transaksi tiruan ini dapat dijelaskan seperti berikut. Pertama–tama dibuat himpunan large itemset potensial L. Kemudian dipilih sebuah large itemset dari L untuk ditempatkan pada sebuah transaksi.

Tabel 1. Parameter Pembuatan Data untuk Pola Asosiasi

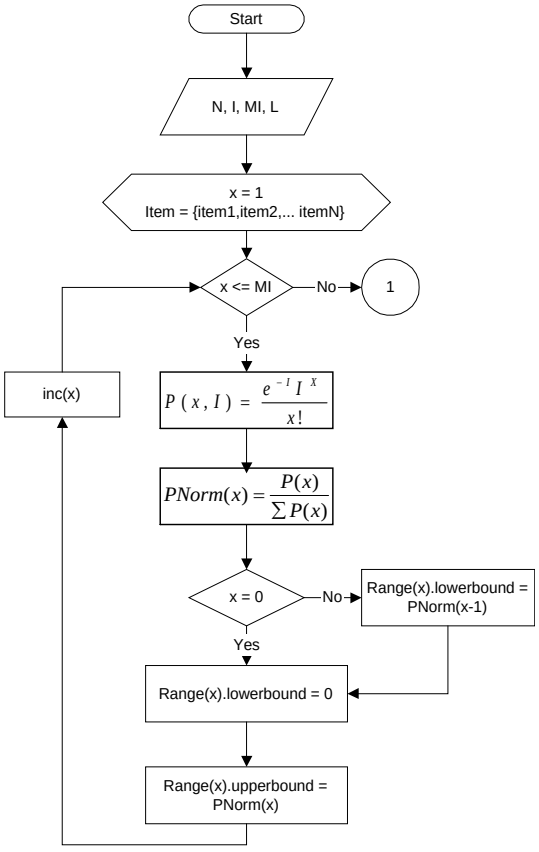
Notasi	Keterangan
|D|	jumlah transaksi
|I|	ukuran rata – rata large itemset potensial
|MI|	ukuran maksimum large itemset potensial
|L|	jumlah large itemset
|T|	ukuran rata – rata transaksi
|MT|	ukuran maksimum transaksi
N	jumlah item

Ukuran setiap large itemset potensial berkisar antara 1 dan |MI|, dimana peluang munculnya large



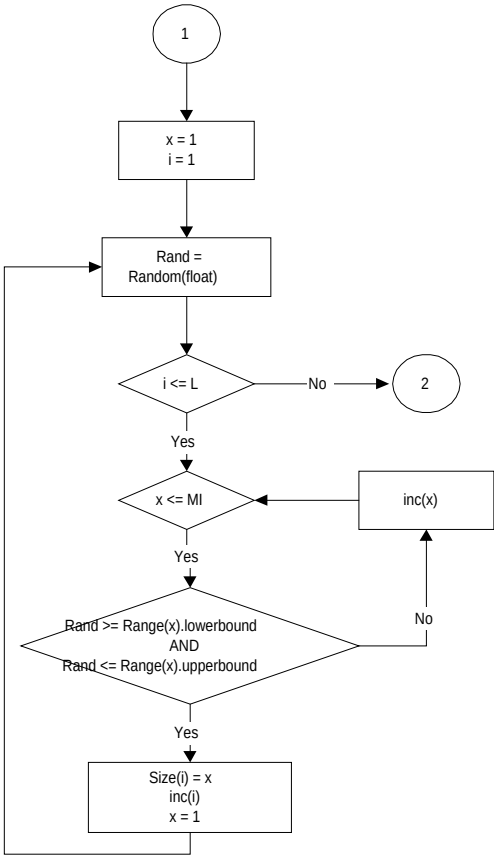
Gambar 1. DAD Tingkat-1 dari Perangkat Lunak

itemset dengan ukuran 1, 2, …, dan |MI| diperoleh dari distribusi Poisson dengan *mean* sama dengan |I|. Peluang–peluang tersebut kemudian dinormalkan sehingga jumlah peluang–peluang tersebut sama dengan 1. Sebagai contoh, misalkan ukuran rerata |I| dari large itemset adalah 3 dan ukuran maksimum |MI| dari large itemset adalah 5, maka menurut distribusi Poisson dengan mean |I|, peluang untuk ukuran 1, 2, 3, 4, dan 5 berturut-turut adalah 0.17, 0.26, 0.26, 0.19 dan 0.12, setelah proses normalisasi. Peluang–peluang tersebut kemudian diakumulasi sehingga setiap ukuran berada pada suatu interval tertentu seperti yang ditunjukkan pada Tabel 2. Untuk setiap large itemset potensial, dihasilkan sebuah angka acak real antara 0 dan 1 untuk menentukan ukuran large itemset potensial.



Gambar 2. Algoritma Perhitungan Distribusi Poisson untuk Large Itemset

Selanjutnya jumlah large itemset potensial dalam L diset sejumlah |L|. Item–item pada large itemset yang pertama dipilih secara acak. Beberapa item pada large itemset selanjutnya dipilih dari large itemset yang telah dibuat sebelumnya.



Gambar 3. Algoritma Penentuan Ukuran Masing-masing Large Itemset

Tabel 2. Probabilitas Ukuran-ukuran Itemset

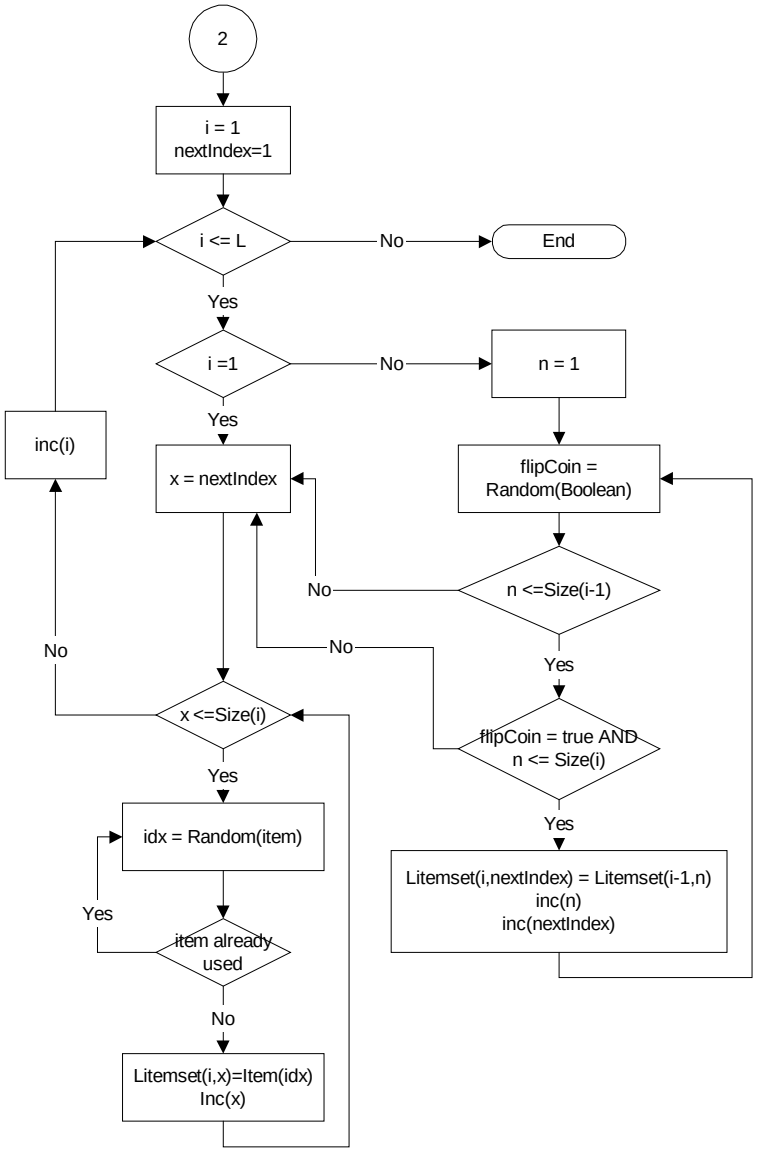
Ukuran	Jangkauan
1	0 ~ 0,17
2	0,18 ~ 0,43
3	0,44 ~ 0,69
4	0,70 ~ 0,88
5	0,89 ~ 1

Cara pemilihan item-item tersebut adalah sebagai berikut: untuk setiap item pada large itemset sebelumnya, dilemparkan sebuah mata uang untuk menentukan apakah item tersebut tetap dipakai pada large itemset yang sedang dibuat atau tidak. Jika jumlah item pada large item yang sedang dibuat belum memenuhi ukuran large itemset itu, maka sisa item pada large itemset tersebut diambil secara acak dari kumpulan item. Setelah himpunan large itemset

L selesai dibuat, langkah selanjutnya adalah membuat transaksi pada basis data. Ukuran masing–masing transaksi diambil dari distribusi Poisson dengan mean sama dengan |T| dimana ukurannya berkisar antara 1 dan |MT|. Metode untuk menentukan ukuran transaksi sama dengan metode untuk menentukan ukuran sebuah large itemset. Untuk sebuah transaksi, dipilih secara acak sebuah large itemset dari L dengan ukuran yang tidak melebihi ukuran transaksi tersebut dan menempatkannya pada transaksi tersebut. Sisa item pada transaksi yang pertama dipilih secara acak. Beberapa item pada transaksi berikutnya diambil dari transaksi yang telah dibuat sebelumnya. Untuk setiap item pada transaksi sebelumnya, dilemparkan sebuah koin untuk menentukan apakah item tersebut tetap ada pada transaksi tersebut atau tidak. Setelah mengambil item–item dari sebuah large itemset dari transaksi sebelumnya, sisa item pada transaksi tersebut ditentukan secara acak. Perlu diketahui bahwa setiap transaksi disimpan pada sebuah sistem berkas dengan format `<id transaksi, jumlah item, item>`.

4.2 Pembuatan Data untuk Pola Sekuensial

Pembuatan data transaksi untuk keperluan



Gambar 4. Diagram Alir Pemilihan Item untuk Large Itemset

pencarian pola sekuensial ini ditujukan untuk mensimulasikan keadaan pada dunia nyata dimana orang membeli kumpulan item (itemset) secara berurutan. Setiap urutan itemset semacam itu berpotensi membentuk sebuah large-sequence yang maksimal. Sebuah contoh dari urutan seperti itu adalah pembelian sprei dan sarung bantal, diikuti dengan selimut, kemudian diikuti dengan syal. Namun beberapa orang mungkin hanya membeli beberapa item dari urutan (*sequence*) tersebut. Sebagai contoh, beberapa orang mungkin hanya membeli sprei dan sarung bantal diikuti selimut dan mungkin beberapa orang hanya membeli selimut. Sebuah *customer-sequence* dapat terdiri dari lebih dari satu urutan. Sebagai contoh, seorang pembeli mungkin juga membeli baju dan jaket ketika membeli sprei dan sarung bantal, dimana baju dan jaket itu membentuk urutan lain. Pada umumnya ukuran customer-sequence berkisar di sekitar niilai rerata dan beberapa pembeli mungkin memiliki banyak transaksi. Ukuran transaksi juga biasanya berkisar di sekitar nilai rerata dan beberapa transaksi mungkin memiliki banyak item [3].

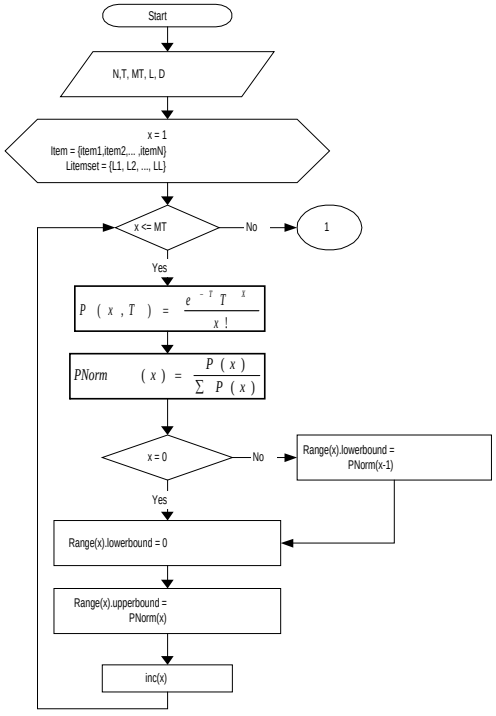
Metode untuk membuat data transaksi tiruan yang digunakan untuk keperluan pencarian pola sekuensial pada data mining hampir sama dengan pembentukan data transaksi untuk pencarian pola asosiasi seperti dijelaskan sebelumnya. Setiap transaksi disimpan dalam sebuah basisdata dengan format `<TID,itemsets>`. Perbedaannya, jika item-item pada data transaksi tiruan untuk aturan asosiasi diurutkan secara leksikal maka pada pembuatan data untuk pola sekuensial, item-item tersebut tersusun secara acak [12] dan jika TID pada data asosiasi berarti penanda transaksi maka pada data sekuensial TID berarti penanda customer-sequence. Parameter yang digunakan sama seperti pada Tabel 1 dengan beberapa modifikasi. Istilah transaksi diubah menjadi *customer-sequence*, istilah *itemset* diubah menjadi *sequence*, item diubah menjadi itemset, itemset diubah menjadi sequence dan notasi L, T, MT diubah menjadi LS, C, MC. Daftar parameter pembuatan data untuk pencarian pola sekuensial secara lengkap ditunjukkan pada Tabel 3.

5. DESAIN PERANGKAT LUNAK

Desain perangkat lunak pembangkit data otomatis didasarkan pada metode fungsional yang visualisasinya digambarkan dengan menggunakan Diagram Alir Data (DAD). Secara garis besar, DAD tingkat-1 dari perangkat lunak dapat dilihat pada gambar 1.

Tabel 3. Parameter Pembuatan Data untuk Pola Sekuensial

Notasi	Keterangan
|D|	jumlah *customer-sequence*
|I|	ukuran rata – rata *large sequence* potensial
|MI|	ukuran maksimum *large sequence* potensial
|LS|	jumlah *large sequence*
|C|	ukuran rata – rata *customer-sequence*
|MC|	ukuran maksimum *customer-sequence*
N	jumlah itemset

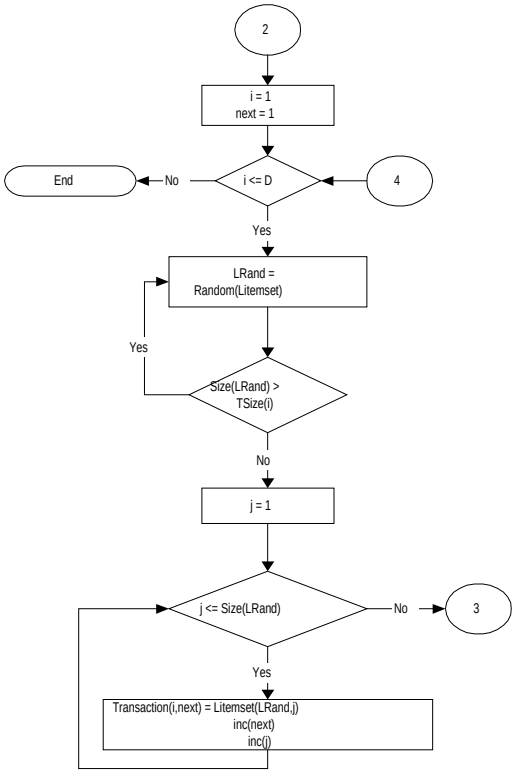


Gambar 5. Algoritma Perhitungan Distribusi Poisson

Seperti ditunjukkan dalam gambar ini, secara keseluruhan perangkat lunak pembangkit data otomatis terdiri dari 5 proses utama. Proses setting basis data tujuan terdiri dari 3 sub-proses, yaitu pemilihan jenis basis data, setting properti koneksi, dan pengetesan koneksi. Sedang proses setting lingkungan data terdiri dari 2 sub-proses, yaitu penentuan tabel tujuan dan penetuan pola data yang akan dibuat (pola asosiasi atau pola sekuensial).

Selebihnya merupakan proses-proses primitih yang tidak dipecah menjadi sub-proses yang lebih kecil, yaitu proses input parameter (memungkinkan pengguna untuk memasukkan nilai-nilai parameter dari data yang akan dibangkitkan), proses pembuatan itemset, dan proses pembuatan transaksi berupa sebuah tabel dalam basis data tujuan.

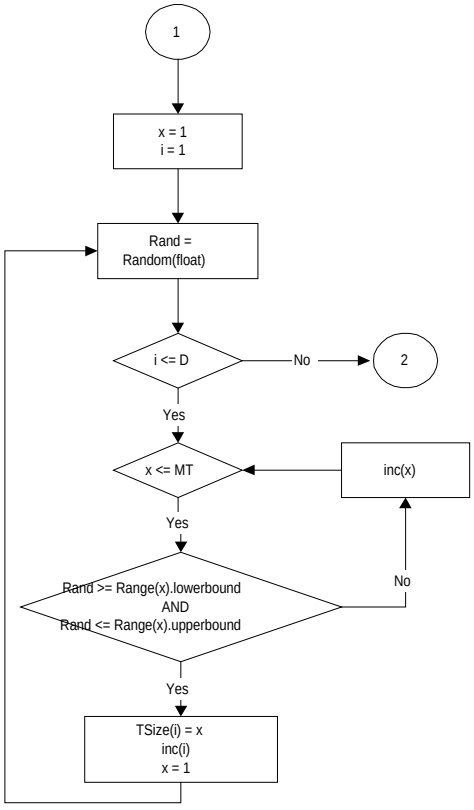
Dalam bagian ini, beberapa algoritma utama (dalam bentuk diagram alir/flowchart) dari proses atau sub-proses yang dilibatkan dalam implementasi perangkat lunak dibahas secara ringkas, seperti algoritma pembuatan itemset, algoritma pemilihan item untuk large itemset, dan algoritma perhitungan distribusi Poisson untuk pembuatan transaksi.



Gambar 6. Algoritma Penentuan Ukuran Transaksi

Proses Pembuatan Itemset

Proses pembuatan itemset ini bertujuan untuk membentuk kumpulan itemset. Hal ini perlu dilakukan untuk memodelkan fenomena bahwa orang cenderung melakukan pembelian beberapa item secara bersamaan. Seperti yang telah dijelaskan sebelumnya, perangkat lunak ini menerima masukan dari pengguna berupa parameter pembuatan data. Setelah nilai-nilai dari parameter tersebut diinisialisasikan, proses pembuatan large itemset dapat dimulai. Proses pembuatan large itemset ini terdiri beberapa tahapan, yaitu perhitungan peluang distribusi Poisson untuk masing-masing ukuran large itemset, penentuan ukuran masing-masing large itemset, dan yang terakhir pemilihan item dari kumpulan item (item-item sebanyak N) untuk dimasukkan pada large itemset. Diagram alur untuk proses-proses tersebut disajikan pada Gambar 2 sampai dengan Gambar 4.

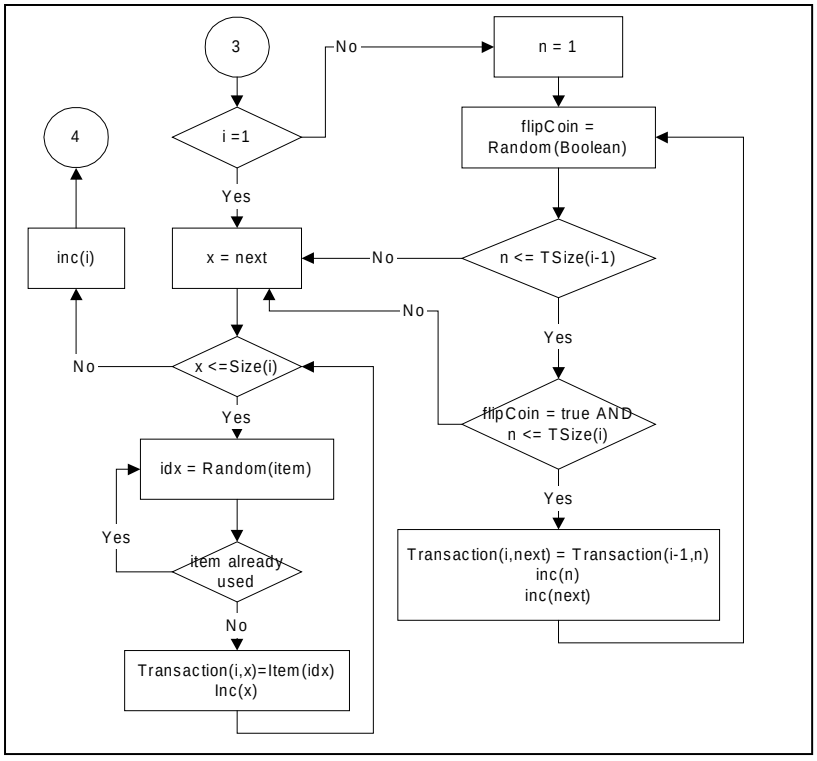


Gambar 7. Algoritma Pemilihan Item dari Large Itemset

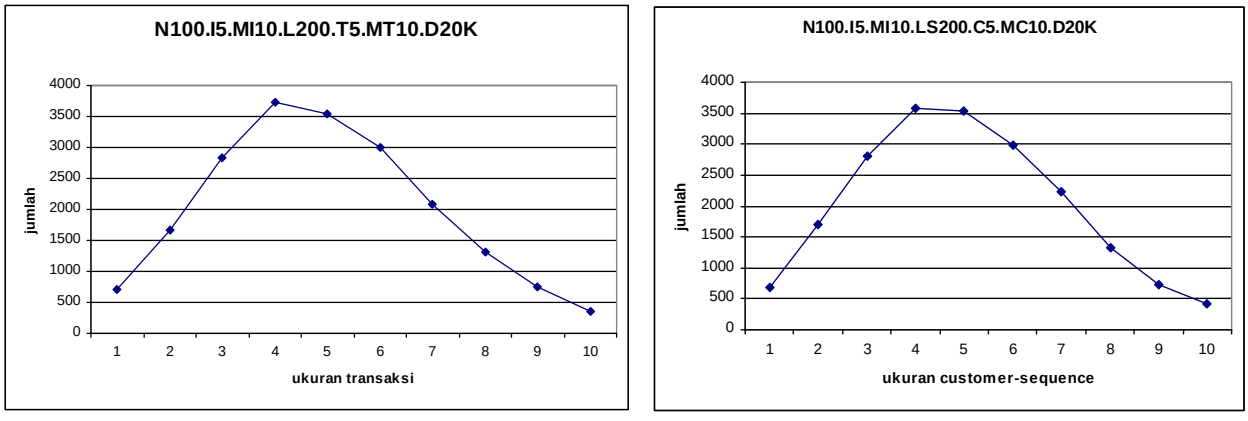
Seperti ditunjukkan dalam Gambar 1, setelah dibuat himpunan item yang anggotanya sebanyak N, langkah selanjutnya dilakukan perhitungan peluang distribusi Poisson untuk masing-masing ukuran large itemset (yaitu jumlah item pada sebuah large itemset). Ukuran masing-masing large itemset tersebut berkisar antara 1 hingga |MI| (ukuran maksimum large itemset). Peluang munculnya large itemset dengan item sebanyak 1, 2,…, |MI| diperoleh dari distribusi Poisson dengan *mean* sama dengan I. Peluang-peluang tersebut kemudian dinormalkan sehingga jumlah peluang-peluang tersebut sama dengan 1.

Proses selanjutnya adalah menentukan jumlah item untuk large itemset sebanyak L yang akan dibentuk. Untuk setiaplarge itemset, dibuat angka acak real antara 0 dan 1. Kemudian dilihat pada interval mana angka acak tersebut berada. Diagram alur proses penentuan ukuran large itemset ini disajikan pada gambar 3.

Setelah menentukan ukuran untuk setiap large itemset, maka langkah selanjutnya adalah memilih item untuk dimasukkan ke large itemset tersebut. Item-item pada large itemset yang pertama dipilih secara acak dari kumpulan item. Item-item pada large itemset berikutnya sebagian diambil dari large itemset sebelumnya dimana untuk setiap item pada



Gambar 8. Algoritma Pemilihan Sisa Item pada Transaksi.



(a) Distribusi Ukuran Transaksi (b) Distribusi Ukuran Customer-Sequence

Gambar 9. Contoh Distribusi Data Tiruan

large itemset sebelumnya dilemparkan sebuah mata uang untuk menentukan apakah item tersebut akan tetap digunakan pada large itemset yang berikutnya atau tidak. Jika jumlah item pada large itemset berikutnya masih kurang dari ukuran large itemset tersebut setelah item-item dari large itemset sebelumnya dimasukkan ke large itemset tersebut, maka sisa item untuk memenuhi kekurangan itu diambil dari kumpulan item secara random. Diagram alir untuk proses pemilihan item untuk large itemset ini disajikan dalam Gambar 4.

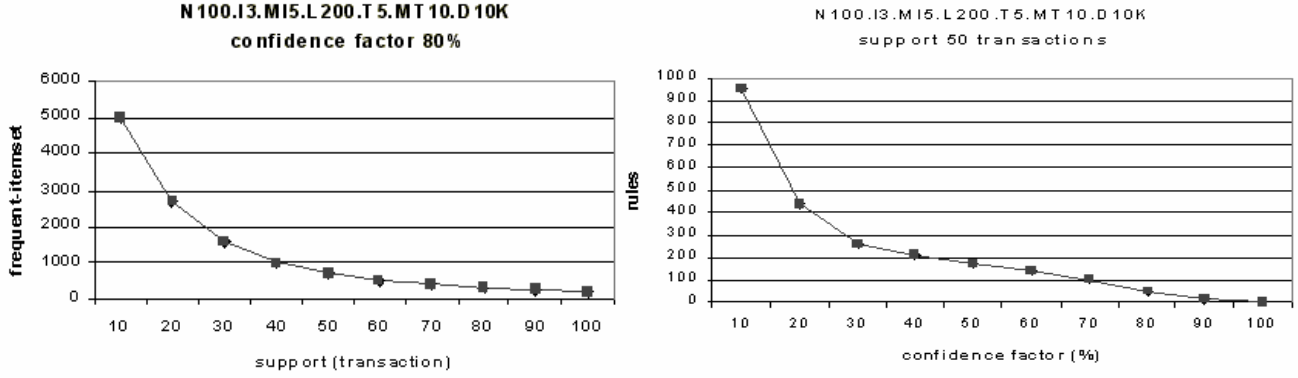
Proses Pembuatan Transaksi

Proses ini bertujuan untuk membuat transaksi tiruan yang akan disimpan pada sebuah tabel dalam basis data tujuan. Proses pembuatan transaksi ini juga terdiri dari tiga bagian, yaitu penghitungan peluang munculnya transaksi dengan ukuran 1 sampai dengan MT menggunakan distribusi Poisson (Gambar 5), kemudian penentuan ukuran transaksi berdasarkan interval peluang yang terbentuk (Gambar 6), dan pemilihan item untuk transaksi (Gambar 7 dan 8). Perbedaan pokok penghitungan peluang Poisson untuk proses pembuatan transaksi dibandingkan dengan perhitungan untuk pembuatan itemset adalah bahwasanya pada pembuatan transaksi ini mean diset sama dengan T.

Proses selanjutnya berkaitan penentuan ukuran transaksi yang disajikan dengan diagram alur seperti pada gambar 6. Setelah diperoleh ukuran untuk masing-masing transaksi maka langkah selanjutnya adalah pemilihan item-item untuk transaksi. Pada dasarnya proses ini hampir sama dengan proses pembuatan large itemset. Pada proses pembuatan large itemset, item-item untuk sebuah large itemset dipilih dari kumpulan item sedangkan pada proses pembuatan transaksi ini, pertama-tama item-item pada sebuah transaksi diambil dari kumpulan large itemset yang telah dibuat sebelumnya. Diagram alir untuk proses terakhir ini ditunjukkan pada Gambar 7 dan 8.

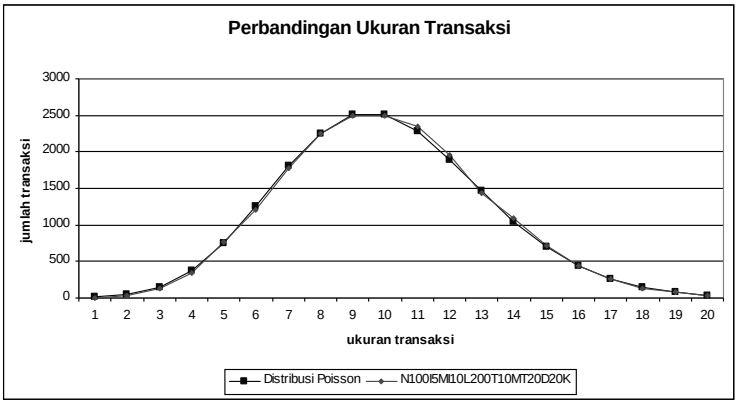
6. HASIL UJI COBA

Data yang dihasilkan oleh pembangkit data ini



(a) Uji Coba untuk berbagai Nilai Support (b) Uji Coba untuk Berbagai Nilai Confidence

Gambar 10. Uji Coba pada Aplikasi Data Mining untuk Pencarian Pola Asosiasi



Gambar 11. Perbandingan Distribusi Ukuran Transaksi

akan digunakan sebagai bahan analisa untuk menguji aplikasi data mining. Oleh karena itu data transaksi tiruan yang dihasilkan harus benar-benar memiliki karakteristik seperti pada data transaksi *retail* pada dunia nyata. Selain itu data transaksi tiruan yang dihasilkan harus dapat diaplikasikan pada algoritma data mining dengan berbagai nilai *support* dan *confidence*. Pada bab ini dibahas berbagai uji coba yang dilakukan untuk membuktikan bahwa data yang dihasilkan benar-benar sesuai dengan spesifikasi yang diharapkan. Uji coba pembuatan data dilakukan menggunakan komputer dengan prosesor AMD Athlon 1 GHz dengan memori 256 Mb. Sedangkan basis data yang digunakan adalah Personal Oracle 8.1.7.

6.1 Pembuatan Data

Untuk membuat data transaksi tiruan yang akan digunakan untuk melakukan pengetesan ekstraksi pola asosiasi, nilai pada parameter-parameter pembuatan data diset sebagai berikut: N=100, I=5, MI=10, L=200, T=5, MT=10, dan D=20.000 transaksi. Dengan parameter data seperti ini akan dihasilkan 20000 transaksi (98574 record). Ukuran masing-masing transaksi berkisar antara 1 sampai dengan 10 dengan pengelompokan di sekitar nilai reratanya, seperti diperlihatkan dalam Gambar 9(a). Hal ini telah sesuai dengan pola distribusi Poisson dimana banyak item pada transaksi terkelompok di sekitar nilai reratanya.

Pada dasarnya data transaksi tiruan yang dihasilkan untuk pencarian pola sekuensial adalah sama dengan data transaksi tiruan yang dihasilkan untuk pencarian pola asosiasi. Perbedaannya jika pada data asosiasi item-item disusun secara leksikal, maka pada data sekuensial item-item disusun secara acak. Sebagai contoh, dibuat data transaksi tiruan dengan parameter N=100, I=5, MI=10, LS=200, C=5, MC=10, D=20.000. Dengan parameter tersebut dihasilkan data transaksi tiruan sebesar 99.285 record dengan distribusi urutan seperti terlihat pada Gambar 9(b).

6.2 Uji Coba pada Aplikasi Data Mining

Dari data transaksi tiruan yang telah dibuat digunakan sebagai bahan masukan bagi aplikasi data mining. Data tiruan yang dihasilkan diujicobakan terhadap dua aplikasi data mining untuk menggali pola-pola asosiasi dengan menggunakan metode pelacakan Buttom-Up [8] dan metode Hybrid [10]. Tabel data tiruan yang digunakan adalah tabel N100I3MI5L200T5MT10D10K, yang dibuat dibuat dengan menggunakan parameter N=100, I=3, MI=5, L=200, T=5, MT=10, D=10000. Tabel tersebut memiliki record sebanyak 49.392. Penggalian pola dilakukan terhadap data tersebut dengan berbagai nilai *support* dan *confidence factor*. Uji coba yang pertama dilakukan dengan *confidence factor* 80% dan nilai *support* mulai dari 10 transaksi hingga 100 transaksi. Hasil uji coba dengan berbagai nilai *support* ini disajikan pada gambar 10(a). Uji coba kedua dilakukan terhadap tabel yang sama dengan nilai *support* tetap, yaitu 50 transaksi, yang hasilnya disajikan pada gambar 10(b).

Ternyata baik algoritma Bottom-Up maupun algoritma Hybrid menghasilkan jumlah itemset yang memenuhi *support* (frequent-itemset) yang sama pada tiap-tiap *support* yang berbeda. Hal ini dibuktikan dengan grafik Bottom-Up dan Hybrid yang berimpit. Hal lain yang dapat disimpulkan dari grafik tersebut adalah pada saat *support* semakin besar, jumlah frequent-itemset yang terjadi semakin sedikit.

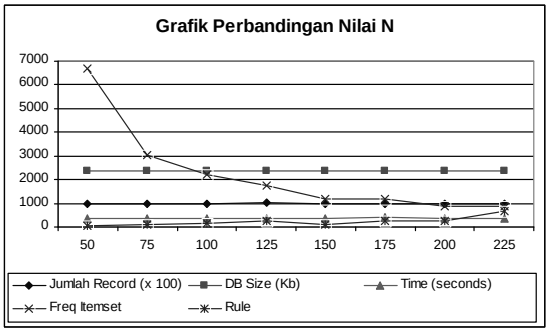
6.3 Distribusi Data

Banyaknya item yang dihasilkan untuk setiap transaksi pada data transaksi tiruan bervariasi antara 1 sampai dengan MT dengan pengelompokan pada nilai rerata jumlah item. Ukuran masing-masing transaksi tersebut dihitung dengan distribusi Peluang Poisson dengan mean sama dengan T. Untuk kebutuhan perbandingan mengenai pola distribusi data yang dihasilkan, digunakan data dengan parameter N100.MI5.MI10.L200.T10.MT20.D20K (mean = T = 10). Pola distribusi yang dihasilkan oleh pembangkit data otoamtis dibandingkan dengan pola distribusi Poisson yang diperoleh berdsarkan perhitungan manual [11] untuk masing-masing ukuran transaksi (dari 1 hingga MT). Pola distribusi yang dihasilkan diperlihatkan pada Gambar 11. Dari gambar ini terlihat bahwa pola distribusi yang dihasilkan oleh pembangkit data otomatis sangat mirip dengan pola distrobusi perhitungan manual.

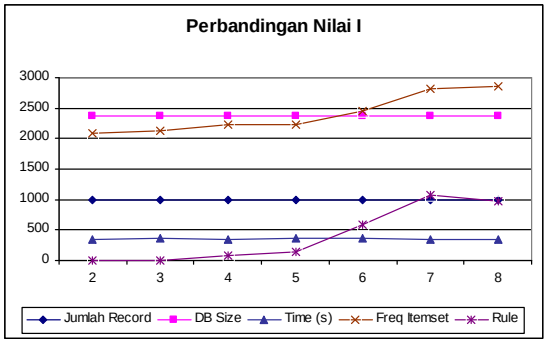
6.4 Analisis terhadap Parameter-parameter N, I, MI, L, T, MT, dan D

Analisis terhadap berbagai parameter untuk pembangkitan data (N, I, MI, L, T, MT, dan D) dilakukan dengan membuat beberapa dataset dengan nilai parameter yang berbeda. Kemudian dilakukan analisis terhadap jumlah record yang dihasilkan,

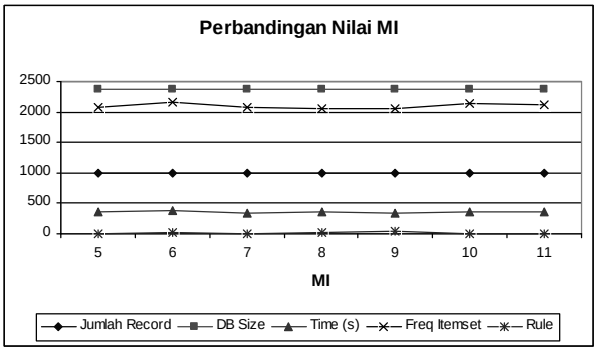
besar basis data, waktu komputasi serta frequent itemset dan pola (*rule*) yang berhasil ditemukan. Untuk mengetahui jumlah frequent itemset dan jumlah pola yang dihasilkan, dilakukan penggalian data terhadap dataset tersebut dengan nilai minimum support 1% dan minimum confidence 75%. Hasil analisa terhadap dataset yang dihasilkan untuk parameter-paramater tersebut, berturut-turut ditunjukkan dalam gambar 12 sampai dengan gambar 18.



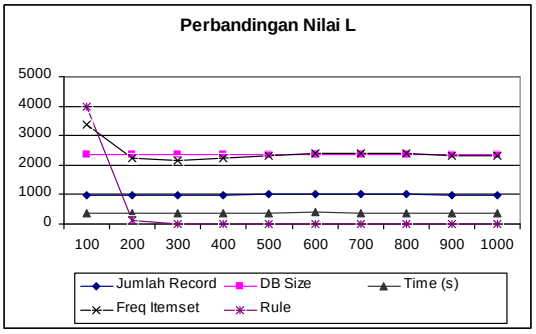
Gambar 12. Grafik Analisis untuk Berbagai Nilai N



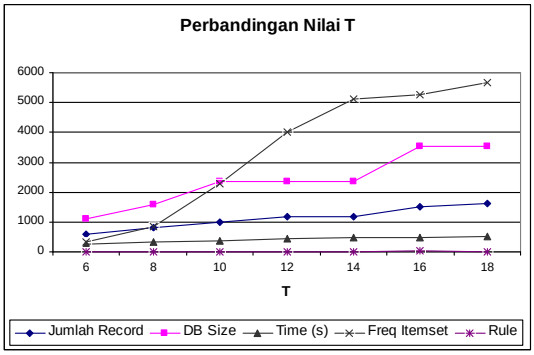
Gambar 13. Grafik Analisis untuk Berbagai Nilai I



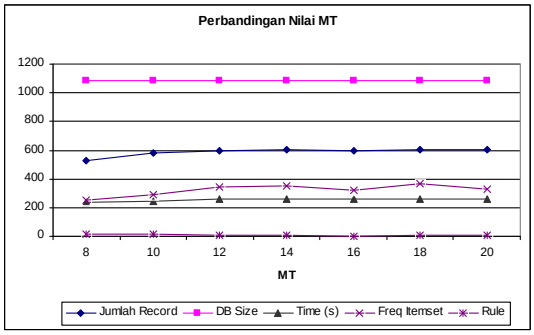
Gambar 14. Grafik Analisis untuk Berbagai Nilai MI



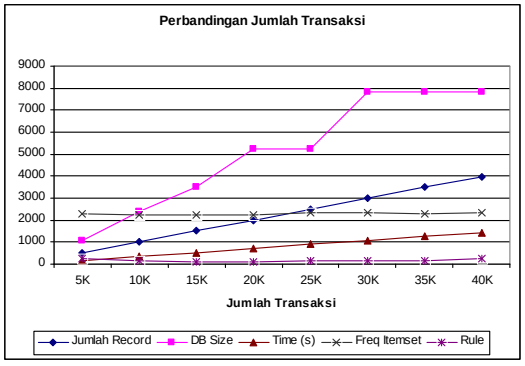
Gambar 15. Grafik Analisis untuk Berbagai Nilai L



Gambar 16. Grafik Analisis untuk Berbagai Nilai T



Gambar 17. Grafik Analisis untuk Berbagai Nilai MT



Gambar 18. Grafik Analisis untuk Berbagai Nilai D

7. KESIMPULAN

Berdasarkan hasil uji coba perangkat lunak pembangkit data otomatis yang telah berhasil didesaind an diimplementasikan, dapat ditarik disimpulkan seperti berikut:

- Perangkat lunak pembangkit data otomatis yang dibuat dapat menghasilkan data transaksi dalam jumlah besar dengan masing-masing transaksi terdiri dari beberapa item dengan peluang kemunculan ukuran transaksi didasarkan pada distribusi Poisson. Dengan demikian perangkat lunak ini mempunyai kelebihan dibandingkan pembangkit data yang pernah dibuat sebelumnya oleh penulis, karena pada pembangkit data tersebut ukuran transaksi ditentukan secara pasti oleh pengguna sehingga hasilnya tidak terdistribusi dengan baik dan bersifat uniform.
- Tahapan pembuatan data transaksi tiruan meliputi pembuatan himpunan item, dilanjutkan dengan pembuatan itemset dan diakhiri dengan pembuatan transaksi. Parameter-parameter yang digunakan untuk membentuk data transaksi tiruan ini adalah jumlah item, ukuran rata-rata itemset, ukuran maksimum itemset, jumlah itemset, ukuran rata-rata transaksi, ukuran maksimum transaksi dan jumlah transaksi. Untuk ini, pengaruh parameter terhadap data yang dihasilkan memberikan kkecenderungan sperti berikut: (a) semakin besar ukuran rerata transaksi, maka semakin besar pula jumlah record, waktu pembuatan dataset, ukuran basisdata, dan jumlah *frequent itemset* yang ditemukan, (b) semakin besar jumlah transaksi, maka semakin besar pula jumlah record, waktu pembuatan dataset, dan ukuran basisdata yang dihasilkan, dan (c) semakin besar jumlah itemset yang dibuat, maka semakin sedikit jumlah aturan yang ditemukan

DAFTAR ACUAN

[1] Agrawal R., Tomasz Imielinski, and Arun Swami, "*Mining Association Rules Between Sets of Items in Large Databases"*, *Proceeding of the 1993 ACM SIGMOD Conference*, May, 1993.

[2] Agrawal R., and R. Srikant, "*Fast Algorithm for Mining Association Rules"*, *Proceedings of the 20th VLDB Conference*, 1994.

[3] Agrawal R., and R. Srikant, "Mining Sequential Patterns", *Proceedings of International Conference on Data Engineering*, 1995.

[4] Berry M.J. and G. Linoff, *Data Mining Techniques for Marketing, Sales, and Customer Support*, John Wiley and Sons, 1997.

[5] Chen M., Jiawei Han, and Phillip S. Yu, "Data Mining: An Overview from a Database Perspective", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8 No. 6, December 1996.

[6] Dajan A., *Pengantar Metode Statistik*, Jilid II, PT. Pustaka LP3S Indonesia, 1996.

[7] Djunaidy A., R. Soelaiman and D.H. Prasetyo, "Development of a Scenario-Based Data Generator for Data Warehousing and Data Mining Applications", *Proceedings of the Int'l Conference on Eletrical, Electronics, Communication, and Information (CECI'2001)*, Jakarta, March 2001.

[8] Evianti Hera H. dan A. Djunaidy (Supervisor) "Perancangan dan Pembuatan Perangkat Lunak Data Mining untuk Pencarian Kaidah Asosiasi Dengan Metode Bottom-Up", *Tugas Akhir*, Jurusan Teknik Informatika, Fakultas Teknologi Informasi - ITS, 2002.

[9] Matsumoto M. and Takuji Nishimura, "Mersenne Twister: A 623-dimensionally Equidistributed Uniform Pseudoacak Number Generator", *ACM Transaction on Modeling and Computer Simulation*, 1998.

[10] Tyaspamadya D. dan A. Djunaidy (Supervisor), "Perancangan dan Pembuatan Perangkat Lunak Data Mining untuk Penggalian Kaidah Asosiasi Menggunakan Metode Hybrid", *Tugas Akhir*, Jurusan Teknik Informatika, Fakultas Teknologi Informasi - ITS, 2002

[11] Walpole R.E. and Raymond H. Myers, "*Ilmu Peluang dan Statistika untuk Insinyur dan Ilmuwan*," Edisi Ke-4, Penerbit ITB, 1995.

[12] Yen S-J. and Arbee L.P. Chen, "An Efficient Approach to Discovering Knowledge from Large Databases,", *Proc. IEEE/ACM International Conference on Parallel and Distributed Information Systems (PDIS)*, 1996.