

Penggunaan Data *Mining* untuk Penggalian Kaidah Asosiasi *Rules-Market Basket Analysis* pada *Giant MOG Malang*

Nugroho Prayitno, Novi Yuliasuti dan Anggi Nurita Dofiss

Dosen STMIK Asia Malang,

ABSTRAK

Penggalian kaidah asosiasi dipergunakan untuk mencari kaidah asosiasi antar suatu kombinasi item, mendeteksi kumpulan-kumpulan atribut yang muncul bersamaan (co-occur) dalam frekuensi yang sering, dan membentuk sejumlah kaidah dari kumpulan-kumpulan tersebut. Untuk penggalian kaidah asosiasi tunggal, pengukuran kualitas dari kaidah-kaidah yang dibangkitkan biasanya hanya berdasarkan satu kriteria evaluasi saja, yaitu confidence factor/predictive accuracy. Pada penelitian ini peneliti ingin mengetahui proses apa saja yang bisa mempengaruhi perbedaan hasil dari proses Data Mining. Apakah ada proses dari perbedaan alam, perbedaan budaya, suhu ataupun yang lainnya yang biasa disebut dengan parameter-parameter atau field-field abstrak yang tidak bisa dituliskan lebih terperinci didalam sebuah database. Di dalam sistem basis dataproduksi yang digunakan pada Giant MOG ini untuk melakukan pencatatan data dan generalisasi atau dengan nama lain OLAP (On-line Analytical Processing). Basis data yang diolah sudah menampilkan data yang spesifikasi, itulah yang akan diteliti juga. Untuk metode Asosiasi harus diselaraskan dengan metode Apriori, karena disini lebih ditekankan lagi bahwasannya Giant di MOG adalah sejenis Departement Store yang besar. Jadi harus bisa memanfaatkan segala informasi dari sekian banyak database yang menumpuk, guna untuk menjadikan informasi itu berguna untuk pengambilan keputusan selanjutnya.

Kata Kunci : *Kaidah Asosiasi, Market Basket Analysis, Multiobyektif.*

ABSTRACT

The Mining of the association rules are used to find the rules of the association between a combination of items, to detect a collection of attributes that occur together in the often frequency, and establish some rules of the multitudes. Extracting a single association rules, measuring the quality of the generated rules are usually based on a single evaluation criterion alone, that's confidence factor / predictive accuracy. In this case, the researchers want to know what the difference could affect the outcome from the process of Data Mining. Is there a process of the difference of natural, cultural, temperature or any other so-called parameters or abstract fields that can't be written in much detail for a database. On the basis of production's data system used in this Giant MOG to record data and generalizations or by any other name OLAP (On-line Analytical Processing). The database is already displaying processed data specifications, that will be research as well. Association for the method should be harmonized with the Apriori method, because it's emphasized here again that Giant MOG is a kind of a high department store. So it should be able to utilize all of database information that has accumulated, in order to make the information useful for subsequent decision making.

Keywords: *Association rules, Market Basket Analysis, Multiobjective.*

PENDAHULUAN

Pada era globalisasi ini persaingan dalam dunia industri di negara Indonesia menjadi sangat ketat, apalagi dengan adanya perusahaan pendatang baru (newcomer) yang ternyata memiliki kekuatan jauh lebih fit. Perusahaan yang sudah lama berdiri harus berani bersaing dengan perusahaan-perusahaan pendatang baru

lain lebih keras lagi, ditambah lagi dengan dicabutnya kuota perdagangan tekstil pada tahun 2005 silam. Di lain pihak, dengan adanya kelemahan-kelemahan kondisi perekonomian dunia yang mengakibatkan turunnya daya beli yang dapat mengurangi kebutuhan, sehingga akan menyebabkan terjadinya suatu tren dimana persediaan (supply) melebihi permintaan (demand) pasar.

Data mining berkembang menjadi alat bantu untuk mencari pola-pola yang berharga dalam suatu database yang sangat besar jumlahnya, sehingga tidak memungkinkan dicari secara manual. Beberapa teknik data mining dapat diklasifikasikan kedalam kategori berikut, meliputi klasifikasi, *clustering*, penggalian kaidah asosiasi, analisa pola sekuensial, prediksi, visualisasi data dan lain sebagainya. *Association mining* diusulkan pertama kali oleh Agrawal et al. (1993), yang selanjutnya berperan utama dalam mendukung penelitian, pembangunan dan aplikasi dari teknik-teknik data mining selanjutnya. Beberapa teknik dari *Association mining* ini telah dikembangkan sampai saat ini. *Association rule mining* dipergunakan untuk mencari kaidah asosiasi antara suatu kombinasi item. Mendeteksi kumpulan-kumpulan atribut yang muncul bersamaan (*co-occur*) dalam frekuensi yang sering, dan membentuk sejumlah kaidah dari kumpulan-kumpulan tersebut. Contoh, 80% dari orang yang berbelanja di suatu supermarket pada bulan Februari membeli coklat untuk acara perayaan Valentine days, ini merupakan degradasi budaya barat yang menular di Indonesia.

Kebanyakan dari permasalahan-permasalahan di dunia nyata merupakan permasalahan multi obyektif yang seharusnya secara bersama-sama dioptimalkan untuk memperoleh hasil yang terbaik dari permasalahan tersebut. Demikian halnya dengan masalah-masalah penggalian kaidah asosiasi. Menemukan sebuah solusi tunggal untuk sebuah masalah multi obyektif sulit untuk dilakukan. Sehingga merupakan hal yang umum untuk mencari sekumpulan solusi berdasarkan pada kriteria-kriteria yang tidak dominan. Sebuah pendekatan untuk menyelesaikan masalah multi obyektif disarankan oleh Vilfredo Pareto. Teknik optimasi yang berdasarkan pendekatan ini disebut dengan teknik optimasi Pareto. Pada penelitian yang dilakukan oleh Ghosh et al. (2004), digunakan tiga pengukuran meliputi *support count*, *comprehensibility* dan *interestingness*, untuk mengevaluasi sebuah kaidah sehingga dapat dipikirkan sebagai obyektif-obyektif yang berbeda dari masalah penggalian asosiasi. *Support count* adalah banyaknya *record-record* yang memenuhi semua kondisi sebelumnya dari suatu kaidah. *Comprehensibility* adalah banyaknya atribut-atribut yang terlibat dalam sebuah kaidah dan mencoba untuk menentukan kemampuan memahami dari kaidah-kaidah tersebut. Dan *interestingness* adalah seberapa pentingnya sebuah kaidah. Jika jumlah dari atribut-atribut yang terlibat dalam bagian *antecedent* lebih

sedikit, maka kaidah tersebut lebih komprehensif. Sebuah kaidah yang memiliki nilai *support count* sangat tinggi, akan diukur sebagai kurang menarik.

Pada umumnya data yang digunakan untuk penggalian kaidah asosiasi sangatlah besar dan terdapat variasi data yang sangat tinggi. Hal ini dapat mengurangi kualitas dari kaidah yang dihasilkan. Oleh karena itu, pada penelitian ini digunakan *cluster-based sampling* untuk melakukan *clustering* data, agar data yang digunakan untuk penggalian kaidah asosiasi sudah terklaster dengan baik. Kemudian dari data yang terklaster tersebut, dilakukan pengambilan sampel yang digunakan untuk penggalian kaidah asosiasi. Dengan menggunakan teknik sampling ini diharapkan dapat meningkatkan kualitas dari kaidah yang dihasilkan.

Dari uraian diatas, permasalahannya adalah membandingkan kualitas kaidah-kaidah yang dihasilkan pada proses penggalian kaidah asosiasi multi obyektif antara data yang di-*cluster* dengan data yang tidak di-*cluster*. Serta mengetahui algoritma yang lebih baik antara algoritma K-Means dan Fuzzy C Means untuk penggalian kaidah asosiasi multi obyektif. Adapun tujuan dari penelitian ini adalah penggunaan *cluster-based sampling* untuk penggalian kaidah asosiasi multi obyektif untuk membandingkan kualitas kaidah-kaidah yang dihasilkan pada proses penggalian kaidah asosiasi multi obyektif antara data yang di-*cluster* dengan data yang tidak di-*cluster*. Dalam hal ini kualitas dari kaidah yang lebih baik adalah yang mempunyai rata-rata nilai *confidence* yang lebih tinggi. Serta membandingkan algoritma K-Means dan Fuzzy C Means untuk penggalian kaidah asosiasi multi obyektif. Manfaat dari penelitian ini adalah bahwa data yang di-*cluster* akan menghasilkan kaidah-kaidah yang lebih berkualitas dibandingkan dengan data yang tidak melalui proses *clustering*. Dan mengetahui algoritma *clustering* yang lebih baik antara K-Means dan Fuzzy C Means untuk penggalian kaidah asosiasi multi obyektif. Data mining muncul disaat analisis data menjadi sangat kompleks dalam memajukan manajemen bisnis. Dimana data mining dapat membantu penggunaannya untuk mengetahui pola dan keteraturan alam himpunan data yang sifatnya tersembunyi. Data mining diartikan sebagai proses ekstraksi informasi yang berguna dan potensial dari sekumpulan data yang terdapat secara implisit dalam suatu bisnis data. Terdapat banyak istilah dari data mining yang dikenal luas seperti *Knowledge Mining From Basis Data*, *Knowledge Extraction*, *Data Archeology*, *Data Dredging* dan

lain sebagainya (Han,2000). Semakin berkembangnya kebutuhan manusia untuk mengolah basis data sehingga memicu perkembangan dari metode-metode data mining. Beberapa metode yang dikenal didalam data mining yaitu penggalian kaidah sekuensial, klasifikasi data dan korelasi data serta kaidah asosiasi.

KAJIAN TEORI

Penggalian kaidah asosiasi adalah salah satu teknik data mining untuk menemukan kaidah asosiasi antara suatu kombinasi *item*, (Agrawal et.al,1993). Sebagai contoh berdasarkan basis data penjualan dari sebuah pasar swalayan, dimana *record-record* menggambarkan transaksi pembelian yang dilakukan oleh para pelanggan dan atribut-atributnya menggambarkan barang-barang yang disediakan atau dijual. Dari kaidah asosiasi yang diperoleh dari analisa pembeliannya dapat diketahui seberapa besar kemungkinan seorang pelanggan membeli coklat dan minuman sirup. Misalnya terdapat kaidah asosiasi {coklat, sirup} \Rightarrow {coklat}, dengan nilai *support* nya 40% dan nilai *confidence* nya 50%. Artinya bahwa seorang pelanggan yang membeli coklat dan sirup mempunyai kemungkinan 50% untuk juga membeli coklat. Aturan ini cukup signifikan karena mewakili 40% dari catatan transaksi selama ini. Dengan pengetahuan tersebut pengelola pasar swalayan dapat mengatur untuk promosi pemasaran dengan menggunakan kupon diskon untuk beberapa kombinasi barang tertentu, peletakan barang dan lain-lain. Maka bisa digunakan peletakan ataupun display barang-barang secara berdampingan atau dalam bentuk paketan sekedar menarik minat konsumen dalam pembelian double. Terdapat banyak daerah aplikasi untuk teknik-teknik penggalian kaidah asosiasi, termasuk rancangan katalog, rancangan toko, pembagian pelanggan, diagnosa alarm telekomunikasi dan lain sebagainya. Dalam menentukan suatu kaidah asosiasi, terdapat suatu *interestingness measure* (ukuran kepercayaan) yang didapatkan dari hasil pengolahan data dengan perhitungan tertentu. Umumnya ada dua ukuran, yaitu :

1. *Support*: suatu ukuran yang menunjukkan seberapa besar tingkat dominasi suatu *item/itemset* dari keseluruhan transaksi. Ukuran ini akan menentukan apakah suatu *item/itemset* layak untuk dicari *confidence*-nya (misal, dari seluruh transaksi yang ada, seberapa besar tingkat dominasi yang menunjukkan bahwa *item* A dan B dibeli bersamaan) dapat juga digunakan untuk mencari tingkat dominasi *item* tunggal.

2. *Confidence* : suatu ukuran yang menunjukkan hubungan antar 2 *item* secara *conditional* (misal, seberapa sering *item* B dibeli jika orang membeli *item* A).

Sebuah kaidah asosiasi adalah sebuah implikasi $A \Rightarrow B$, dimana sekumpulan *item* A dan B tidak saling beririsan (*intersect*). Masing-masing kaidah asosiasi mempunyai dua kualitas pengukuran yaitu *support* dan *confidence* yang didefinisikan sebagai berikut:

$$\text{Support} : \text{supp}(A \Rightarrow B) = \text{prob} \{ A \Rightarrow B \} \quad (1)$$

$$\text{Confidence} : \text{conf}(A \Rightarrow B) = \frac{\text{supp}\{ A \Rightarrow B \}}{\text{supp}\{ A \}} \quad (2)$$

Kedua ukuran ini nantinya akan berguna dalam menentukan *interestingness* kaidah asosiasi, yaitu untuk dibandingkan dengan *threshold* (batasan) yang ditentukan. Batasan tersebut umumnya terdiri dari *min_support* dan *min_confidence*. Metodologi dasar penggalian asosiasi terbagi menjadi dua tahap meliputi :

1. Frequent itemset generation

Pada tahapan ini dilakukan pencarian kombinasi *item* yang memenuhi syarat minimum dari nilai *support* dalam basisdata.

2. Rule Generation

Setelah semua kaidah frekuensi tinggi ditemukan, selanjutnya mencari turunan asosiasi yang memenuhi syarat minimum *confidence* dengan menghitung *confidence* asosiasi $A \Rightarrow B$ dari *support* kaidah frekuensi tinggi A dan B dengan menggunakan rumus (2). Terdapat dua proses utama yang dilakukan pada algoritma Apriori meliputi :

1. Join (penggabungan)

Untuk menemukan L_k, C_k dibangkitkan dengan melakukan proses join L_{k-1} dengan dirinya sendiri, $C_k = L_{k-1} * L_{k-1}$, kemudian C_k diambil hanya yang terdapat dalam L_{k-1} .

Untuk menemukan L_k, C_k dibangkitkan dengan melakukan proses join L_{k-1} dengan dirinya

2. Prune (pemangkasan)

Menghilangkan anggota C_k yang memiliki *support count* lebih kecil dari *min support* supaya tidak dimasukkan ke L_k

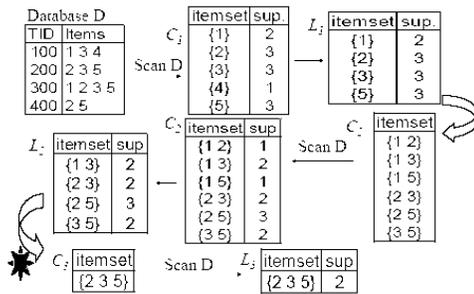
Tahapan yang dilakukan algoritma apriori untuk membangkitkan *large itemset* adalah sebagai berikut :

1. Menelusuri seluruh *record* pada basis data transaksi dan menghitung *support count* dari tiap *item*.

2. Large 1 *itemset* L_1 dibangun dengan menyaring C_1 dengan *support count* yang lebih besar atau sama dengan *min support* untuk dimasukkan kedalam L_1 .

3. Untuk membangun L_2 algoritma apriori menggunakan proses *join* untuk menghasilkan C_2 .

4. Dari C_2 , *itemset* yang memiliki *support count* lebih besar atau sama dengan *min support* akan disimpan dalam L_2 .
 5. Proses ini diulang sampai tidak ada lagi kemungkinan *k-itemset*.
- Contoh proses pembangkitan kandidat untuk dijadikan *itemset* dan *large itemset* dapat dilihat pada Gambar 1.



Gambar 1. Pembangkitan kandidat *itemset* dan *large itemset*.

Metode Data Mining yang digunakan pada perusahaan Giant department store. Sistem basis data produksi adalah generalisasi dan pencatatan data yang juga disebut dengan nama lain On-line Analytical Processing (OLAP). Generalisasi dan pencatatan ini menampilkan karakteristik umum terhadap sekumpulan data yang dispesifikasi oleh pemakai dalam basis data. Untuk kedepannya Giant department store di MOG Malang akan menggunakan SAP (System Application and Product)

Di Perusahaan Giant department store lebih menitik beratkan pada kegiatan operasionalnya, jadi lebih mengacu pada "Market Basket Analysis". Yaitu fungsi dari Asosiasi Rules seringkali disebut dengan "Market Basket Analysis" yang digunakan untuk menemukan relasi atau korelasi diantara himpunan item-item. Sedangkan pengertian dari Market Basket Analysis adalah analisa dari kebiasaan membeli customer dengan mencari asosiasi dan korelasi antara item-item berbeda dengan diletakkan customer pada keranjang belanjaan.

Fungsi ini paling banyak digunakan untuk menganalisa data dalam rangka keperluan strategi pemasaran, desain katalog, dan proses pembuatan keputusan bisnis. Tipe association rule bisa dinyatakan sebagai misal : "70% dari orang-orang yang membeli mie, juice dan saus akan membeli juga roti tawar". Aturan asosiasi mengcapture item atau kejadian dalam data berukuran besar yang berisi data transaksi. Dengan kemajuan teknologi, data penjualan dapat disimpan dalam jumlah besar yang disebut dengan "basket data." Aturan asosiasi yang didefinisikan pada basket data, digunakan untuk keperluan promosi, desain katalog,

segmentasi customer dan target pemasaran. Secara tradisional, aturan asosiasi digunakan untuk menemukan trend bisnis dengan menganalisa transaksi customer.

Dan dapat digunakan secara efektif pada bidang Web Mining yang diilustrasikan sebagai berikut : pada Web access log, kita menemukan bahwa aturan asosiasi : "A and B implies C," memiliki nilai confidence 80%, dimana A, B, dan C adalah halaman Web yang bisa diakses. Jika seorang user mengunjungi halaman A dan B, maka terdapat 80% kemungkinan dia akan mengunjungi halaman C juga pada session yang sama, sehingga halaman C perlu diberi direct link dari A atau B. Informasi ini dapat digunakan untuk membuat link secara dinamik ke halaman C dari halaman A atau B sehingga user dapat melakukan direct link ke halaman C. Informasi semacam ini digunakan untuk melakukan link ke halaman produk yang berbeda secara dinamik berdasarkan interaksi customer.

Analisis dari kebiasaan membeli customer dengan mencari asosiasi dan korelasi antara item-item berbeda yang diletakkan customer dalam keranjang belanjannya.



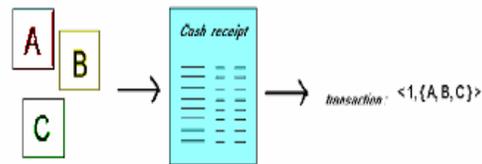
Gambar 2. Keranjang Belanja

☑Diberikan :

Suatu database transaksi customer (misal, keranjang belanja), dimana setiap transaksi adalah suatu himpunan item-item (misal produk)

☑Cari:

Grup item-item yang sering dibeli secara bersama-sama



Gambar 3. Bentuk Transaksi Keranjang Belanja

☑Mengekstraksi informasi perilaku pembelian "IF membeli bir dan sosis, THEN juga membeli mostar dengan peluang tinggi"

☑Informasi yang bisa ditindak-lanjuti:

Bisa menyarankan Tata letak toko yang baru dan campuran produk.

Bisa menyarankan Produk apa untuk diletakkan dalam promosi.

Kaidah Asosiasi yang digunakan :

1. Kaidah asosiasi penambangan Pertama kali diusulkan oleh Agrawal, Imielinski dan Swami [AIS93]
2. Diberikan:
 - a. Suatu database transaksi
 - b. Setiap transaksi adalah suatu himpunan item-item
3. Cari seluruh kaidah asosiasi yang memenuhi kendala minimum support dan minimum confidence yang diberikan user.
4. Contoh:
 - 30% dari transaksi yang memuat roti juga memuat mentega 5% dari transaksi memuat item-item berikut:
 - a. 30% : confidence dari kaidah ini
 - b. 5% : support dari kaidah ini
5. Kita berminat untuk mencari seluruh kaidah ketimbang memeriksa apakah suatu kaidah berlaku.

Dasar Kaidah Asosiasi:

1. Kaidah asosiasi penambangan:

Mencari pola yang sering muncul, asosiasi, korelasi, atau struktur sebab musabab diantara himpunan item-item atau objek-objek dalam database transaksi, database relasional, dan penyimpanan informasi lainnya
2. Kepemahaman:

Sederhana untuk dipahami
3. Kegunaan:

Menyediakan informasi yang bias ditindaklanjuti
4. Efisiensi:

Ada algoritma pencarian yang efisien
5. Aplikasi:

Analisis data keranjang pasar, pemasaran silang, rancangan katalog, analisis lossleader, clustering, klasifikasi, dsb.
6. Format penyajian kaidah asosiasi yang biasa:

Penggalan kaidah asosiasi multi obyektif

Permasalahan-permasalahan penggalan kaidah asosiasi dapat dianggap sebagai sebuah masalah yang multi obyektif, karena masalah-masalahnya kompleks. Terdapat beberapa pendekatan untuk permasalahan multi obyektif ini. Salah satunya adalah pendekatan optimasi Pareto. Dalam definisi optimasi Pareto, sebuah solusi S_1 dikatakan menjadi dominan pada solusi yang lain S_2 , jika dan hanya jika solusi S_1 tersebut adalah jelas lebih baik daripada paling tidak satu dari kriteria dan tidak lebih buruk daripada, S_1 dalam evaluasi dari keseluruhan kriteria. Sebuah solusi dinyatakan sebagai solusi tidak dominan jika solusi tersebut adalah lebih unggul dari semua solusi yang lain dalam semua kriteria dari optimasi (Freitas,2002).

Sebuah pendekatan untuk menyelesaikan masalah multi obyektif disarankan oleh Vilfredo Pareto. Teknik optimasi yang berdasarkan pendekatan ini disebut dengan teknik optimasi Pareto.

Pada penelitian yang dilakukan oleh (Ghosh et al,2004), digunakan tiga pengukuran meliputi *support count*, *comprehensibility* dan *interestingness*, untuk mengevaluasi sebuah kaidah sehingga dapat dipikirkan sebagai obyektif-obyektif yang berbeda dari masalah penggalan asosiasi. *Support count* adalah banyaknya *record-record* yang memenuhi semua kondisi sebelumnya dari suatu kaidah. *Comprehensibility* adalah banyaknya atribut-atribut yang terlibat dalam sebuah kaidah dan mencoba untuk menentukan kemampuan memahami dari kaidah-kaidah tersebut. Untuk menghitung nilai *Comprehensibility* dari sebuah rule digunakan rumus sebagai berikut :

$$\text{Comprehensibility} = \text{Log}(1+|C|/\text{log}(1+ |A \setminus C|)) \quad (3)$$

Dimana : |C| adalah jumlah atribut yang terlibat dalam bagian *consequent*.

|A \setminus C| adalah total kaidah.

Dan *interestingness* adalah seberapa pentingnya sebuah kaidah. Jika jumlah dari atribut-atribut yang terlibat dalam bagian *antecedent* lebih sedikit, maka kaidah tersebut lebih komprehensif. Sebuah kaidah yang memiliki nilai *support count* sangat tinggi, akan diukur sebagai kurang menarik. Nilai *interestingness* dari sebuah kaidah diperoleh dengan menggunakan rumus berikut ini :

Interestingness

$$= [\text{SUP}(A \setminus C) / \text{SUP}(A)] \times [\text{SUP}(A \setminus C) / \text{SUP}(C)] \times [1 - (\text{SUP}(A \setminus C) / |D|)] \quad (4)$$

Dimana : |D|= jumlah *record* dalam basisdata

Cluster-Based Sampling

Clustering merupakan salah satu metode data mining yang bersifat tanpa arahan (*unsupervised*). *Clustering* data adalah proses dari pengelompokan data berdasarkan similaritas atau kesamaan antara data. Similaritas *clustering* dapat diaplikasikan untuk beberapa bidang, misalnya di bidang penelitian pasar, *clustering* digunakan untuk membagi populasi umum dari konsumen-konsumen ke dalam segmen pasar, pembagian pasar dan menentukan sasaran pasarnya.

Terdapat beberapa pendekatan dari *clustering*, salah satunya adalah untuk database yang besar, dimana digunakan untuk *sampling* dan *compression*. Metode pengelompokan clustering adalah kumpulan obyek data dimana jika obyek data yang terletak didalam cluster memiliki kemiripan sedangkan yang tidak berada dalam satu cluster tidak memiliki kemiripan. Jika terdapat n obyek pengamatan dengan p variabel

maka sebelum dilakukan pengelompokan data atau obyek, terlebih dahulu ditentukan ukuran kedekatan sifat antar data yang ada. Ukuran kedekatan data yang biasa digunakan adalah jarak euclidian (*euclidean distance*) antara dua obyek dari p dimensi pengamatan. Jika obyek pertama yang diamati adalah $X=[x_1, x_2, x_3, \dots, x_p]$ dan $Y=[y_1, y_2, y_3, \dots, y_p]$ maka perhitungan jarak dengan menggunakan *euclidean distance* untuk satu vektor digunakan rumus (5).

$$D_{L_2}(x_2 - x_1) = \|x_1 - x_2\| = \sum_{j=1}^p |x_{2j} - x_{1j}| \quad (5)$$

Sampling adalah proses pemilihan unsur-unsur (*item-item*) yang mewakili suatu populasi (seluruh unsur/*item* yang ada) secara sistematis dengan tujuan mempelajari unsur/*item* tersebut. Pada penelitian ini digunakan *sampling* berbasis klaster, dimana teknik *clustering* yang digunakan adalah algoritma K-Means dan Fuzzy C Means.

K-Means

Algoritma K-Means merupakan metode yang umum digunakan pada teknik *clustering*. Menurut Mac Queen (1967), K-Mean adalah salah satu algoritma *unsupervised learning* yang paling sederhana yang dikenal dapat menyelesaikan permasalahan *clustering* dengan baik. Ide utamanya adalah mendefinisikan *centroid* sejumlah k , untuk masing-masing klaster. *Centroid-centroid* ini harus diletakkan dengan cara yang cerdas pada satu tempat, karena lokasi yang berbeda akan menyebabkan hasil yang berbeda pula. Maka sebaiknya meletakkan sebisa mungkin berjauhan satu dengan yang lain. Langkah berikutnya adalah mengambil masing-masing titik kepunyaan sekumpulan data tertentu dan menghubungkannya ke *centroid* yang terdekat. Ketika tidak ada lagi titik yang belum dihubungkan, maka langkah pertama terlengkapi dan satu pengelompokan awal telah dilakukan. Dalam posisi ini kita perlu menghitung kembali k *centroid-centroid* baru sebagai *barycenters* dari hasil klaster-klaster pada langkah sebelumnya. Setelah kita mempunyai *centroid* baru, satu keterikatan harus dilakukan antara titik-titik sekumpulan data yang sama dengan *centroid* yang baru. Satu pengulangan telah dilakukan, sebagai hasil dari pengulangan ini mungkin kita memperhatikan bahwa k *centroid* mengubah lokasi mereka secara bertahap sampai tidak ada lagi perubahan yang dilakukan. Dengan kata lain *centroid* tidak bergerak/berubah lagi.

Fuzzy C Means

Adalah sebuah metode *clustering* yang mengizinkan satu data menjadi milik dua atau lebih *cluster*. Metode ini sering digunakan dalam pengenalan pola (*pattern recognition*). Metode Fuzzy C Means adalah salah satu metode *clustering* yang mengalokasikan kembali data kedalam masing-masing *cluster* dengan memanfaatkan teori Fuzzy. Dalam metode Fuzzy C Means dipergunakan variabel *membership function* i_{ku} , yang merujuk pada seberapa besar kemungkinan suatu data bisa menjadi anggota ke dalam suatu *cluster* (Bezdek, 1981). Pada metode ini juga digunakan suatu variabel m yang merupakan *weighting exponent* dari *membership function*. Variabel ini dapat mengubah besaran pengaruh dari *membership function* i_{ku} dalam proses *clustering*. Variabel m mempunyai wilayah nilai $m > 1$, sampai pada saat ini tidak ada ketentuan yang jelas berapa besar nilai m yang optimal dalam melakukan optimasi suatu permasalahan *clustering*. Nilai m yang umum digunakan adalah 2. *membership function* untuk suatu data ke suatu *cluster* tertentu dihitung menggunakan rumus sebagai berikut :

$$u_{ik} = \sum_{j=1}^c \left(\frac{D(x_k, v_i)}{D(x_k, v_j)} \right)^{\frac{2}{m-1}} \quad (6)$$

Dimana : u_{ik} = *membership function* data ke- k ke *cluster* ke- i

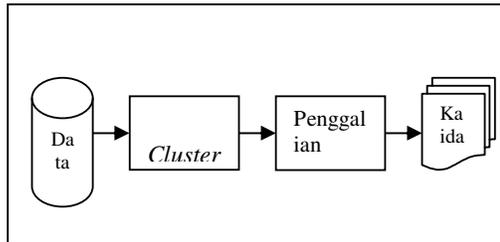
v_i = nilai *centroid cluster* ke- i

m = *weighting exponent*

Secara mendasar terdapat dua cara pengalokasian data kembali kedalam masing-masing *cluster* pada saat proses iterasi *clustering*. Yang pertama adalah pengalokasian dengan cara tegas (*hard*). Dimana data item secara tegas dinyatakan sebagai anggota *cluster* yang satu dan tidak menjadi anggota *cluster* yang lain. Yang kedua dengan menggunakan nilai Fuzzy dimana masing-masing data item diberikan nilai kemungkinan untuk bisa bergabung ke setiap *cluster* yang ada. Pada K-Means pengalokasian data kembali didasarkan pada perbandingan jarak antara data dengan *centroid* setiap *cluster* yang ada. Pada Fuzzy C Means pengalokasian kembali data kedalam masing-masing *cluster* dipergunakan variabel *membership function* u_{ik} yang merujuk pada seberapa besar kemungkinan suatu data bisa menjadi anggota dalam satu *cluster*. Selain itu juga digunakan variabel m yang merupakan *weighting exponent* dari *membership function*.

METODOLOGI PENELITIAN

Dalam hal penelitian ini dilakukan dalam wawancara manager EDP yang ada pada perusahaan Giant MOG. Wawancara yang dilakukan bertahap guna untuk mendapatkan teori dan informasi seakurat mungkin dalam penyusunan jurnal ini, dan bisa disiapkan materi lebih lanjut untuk kejelasan dalam implementasi metode yang digunakan didalam perusahaan. Langkah-langkah yang dilakukan didalam penelitian ini meliputi : Desain algoritma, Implementasi algoritma, Uji coba, Evaluasi. Adapun *block diagram* dari sistem ditunjukkan pada Gambar 2. Terdapat dua proses utama dalam sistem yaitu proses *clustering* dan proses penggalian kaidah asosiasi. Data transaksi yang berasal dari data *warehouse* akan melalui proses klasterisasi terlebih dahulu sebelum dilakukan proses penggalian kaidah asosiasinya.



Gambar 4. Block Diagram Sistem

Desain Algoritma

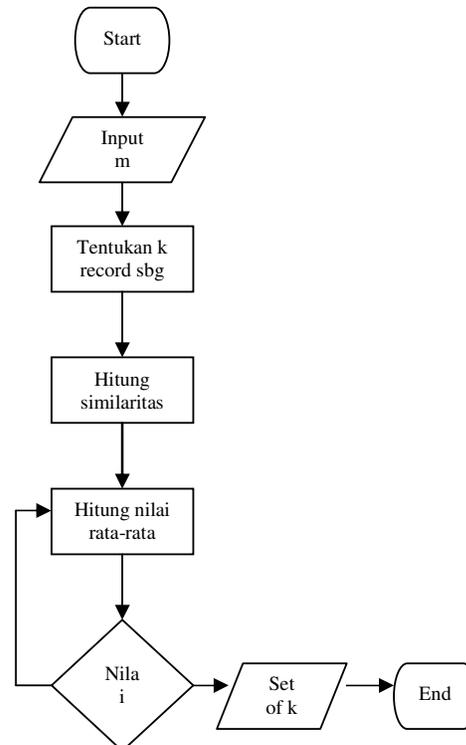
Desain algoritmanya sebagai berikut, terdapat dua proses utama yaitu proses klasterisasi data input dan proses penggalian kaidah asosiasi multi obyektif. Dari proses klasterisasi akan diperoleh data sampel yang sudah dikelompok-kelompokkan, yang selanjutnya digunakan sebagai data input untuk proses penggalian kaidah asosiasi. Didalam penelitian ini digunakan dua metode *clustering* yaitu K-Means dan Fuzzy C Means, *flowchart* dari masing-masing metode tersebut berturut-turut ditunjukkan pada Gambar 3 dan Gambar 4. Tahapan yang dilakukan dalam melakukan *clustering* dengan metode K-Means adalah pertama dipilih k record secara random sebagai *centroid* atau pusat *cluster* untuk masing-masing k *cluster*. Selanjutnya menghitung jarak dari masing-masing *record* dengan *record* yang lain dengan menggunakan rumus (5). Dengan menggunakan jarak tersebut, selanjutnya adalah menentukan *record* r_i ke sebuah *cluster*, sedemikian sehingga jarak r_i dan *centroid cluster* adalah terkecil diantara *cluster-cluster* yang lain. Berikutnya adalah menghitung kembali nilai rata-rata dari *centroid* untuk masing-masing *cluster* berdasarkan *record-record* yang menjadi anggota dari *cluster* tersebut. Adapun untuk

menghitung nilai rata-rata *centroid* digunakan rumus (7). Proses tersebut diulangi sampai nilai rata-rata dari masing-masing *centroid* tidak berubah atau mendekati tetap.

$$v_{i,j} = \frac{\sum_{k=1}^{N_i} x_{kj}}{N_i}$$

(7)

Dimana : N_i adalah jumlah data yang menjadi anggota *cluster* i .

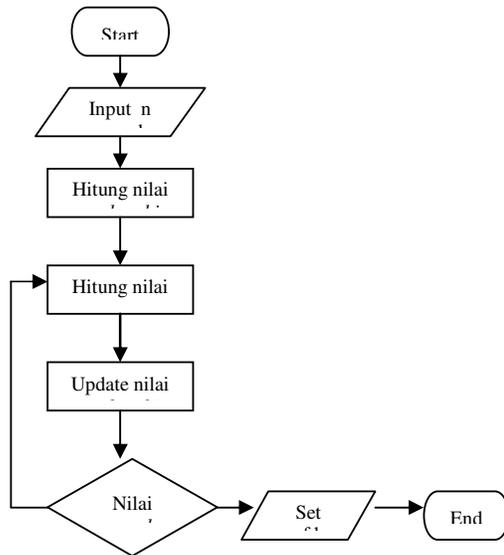


Gambar 5. Flowchart K-Means

Adapun tahapan yang dilakukan untuk melakukan proses *clustering* dengan menggunakan metode Fuzzy C Means adalah pertama dimulai dengan menentukan jumlah *cluster* yang diinginkan, menentukan nilai m (*weighting exponent*) dan menentukan nilai ϵ (*threshold*). Selanjutnya menghitung nilai *membership* masing-masing titik dari data dengan menggunakan rumus (6). Dilanjutkan dengan menghitung nilai pusat *centroid*, nilai ini diperoleh dengan menggunakan rumus (7). Nilai dari semua titik diperbaiki berdasarkan dari nilai pusat *centroid* yang baru. Penghitungan nilai pusat *centroid* dengan nilai *membership* dilakukan sampai nilai *membership* lebih kecil dari nilai ϵ (*threshold*) yang telah ditentukan diawal.

Penggalian kaidah asosiasi dilakukan dengan menggunakan algoritma Apriori, yang merupakan algoritma yang umum digunakan

dalam penggalian kaidah asosiasi. Untuk penggalian kaidah asosiasi obyektif tunggal hanya digunakan pengukuran *support count* saja. Sedangkan untuk penggalian kaidah asosiasi multi obyektif digunakan tiga pengukuran yaitu nilai *support count*, *comprehensibility* dan *interestingness*.



Gambar 6. Flowchart Fuzzy C Means

PEMBAHASAN

Dataset yang digunakan sebagai bahan uji coba adalah dataset retail dari informasi Manager EDP yang ada pada perusahaan Giant MOG Malang itu sendiri. Sebagai contohnya penelitian ini akan mengambil contoh salah satu barang, yaitu 'coklat'. Membahas penjualan, persediaan yang berhubungan dengan data mining dan semua parameter yang berhubungan. Parameter yang berhubungan ini berguna untuk mengetahui coklat tersebut mengalami fastmoved pada bulan apa saja, dan mengalami slowmoved pada bulan apa saja. Penataan barang atau displai apakah juga mempengaruhi, terutama penempatan dalam bentuk paket.

Bisa diambil contoh ketika pada bulan februari, tepatnya pada perayaan valentine, hal ini dipengaruhi dari degradasi budaya barat yang masuk ke budaya timur. Yaitu budaya indonesia, mulai mengikuti tren ini. Jadi coklat menjadi icon utama, dan otomatis penjualan coklat akan menjadi fastmoved. Selain penjualan coklat, juga bisa dibuat paket. Karena biasanya, valentine juga dirayakan dalam bentuk kebersamaan. Bisa dipaketkan dengan penjualan minuman bersoda. Dengan adanya hal-hal ini yang menimbulkan parameter baru yang tidak ada

pada database perusahaan Giant MOG Malang. Padahal hal ini sangat penting sekali untuk kemajuan dalam keberhasilan dalam penjualan barang-barangnya. Maka dalam sistem data mining jika akan memberitahukan barang apa saja yang mengalami fastmoving dan memberikan alternatif untuk melakukan inovasi-inovasi dalam melakukan penjualan. Terutama dalam bidang display, yang tidak ada di field-field tabel database perusahaan. Sebagai contoh lain, penjualan dilakukan dalam bentuk parcel. Guna penjualan paket dalam bentuk parcel coklat ini karena telah diperoleh informasi dari pelaksanaan informasi dari sistem data mining, barang apa saja, atau kombinasi apa saja yang dilakukan pembeli ketika membeli coklat. Biasanya membeli roti, minuman bersoda, atau hal lainnya. Pemberian diskon untuk pembelian diatas rata-rata juga memancing minat para pembeli untuk melakukan pembelian lebih banyak, sehingga terjadi pembelian potensial yang juga akan membuat perusahaan menjadi sukses melakukan strategin dan pemanfaatan informasi. Pihak dari manejer bagian operasional harus sudah mengantisipasi, karena sistem data mining yang ada hanya berusaha mengoptimalkan dari penumpukan data yang ada pada perusahaan, untuk melakukan penjualan secara potensial maka pihak bagian operasional harus mengetahui dan menjalankan metode setelah mengambil laporan yang turun dari pihak kantor pusat. Kaidah yang dipakai untuk mendapatkan suatu informasi, diperoleh dari adanya kaidah asosiasi. Yaitu memperoleh informasi dari atribut-atribut apa saja yang saling berhubungan. Hal ini sangat penting untuk penempatan atau display barang yang nantinya juga akan memperbaiki barang slowmoved menjadi barang yang fastmoved.

Manfaat Data Mining

Pada awal penggunaan Data Mining yang ada pada Giant MOG Malang department store ini dimulai dan dibuka. Data Mining ini merupakan cabang dari Giant yang sebelumnya. Karena data mining ini akan berfungsi maksimal ketika adanya data yang menumpuk, semakin banyak data yang digali, maka semakin kompeten informasi yang didapat. Kevalitannya menjadi semakin kuat untuk memanfaatkan informasi tersebut. Data mining mengoptimalkan informasi yang bertumpuk, yang awalnya hanya memakan tempat di database, diolah sehingga bisa mendapatkan informasi dan strategi-setrategi baru untuk kemajuan perusahaan terutamanya Giant MOG Malang. Jadi telah dilakukan riset dan smpling terlebih dahulu untuk pemanfaatannya di Giant MOG. Hal ini dirasa cukup menghemat biaya

untuk pembelian sistem lagi, jadi dilakukan kerjasama dengan pihak terdahulu sebelum cabang Giant di MOG Malang ini dibuka.

Manfaat dari Data Mining yang digunakan pada Giant departement store :

1. Untuk mengetahui stok ideal perbulan dari setiap item.
2. Mengamati pergerakan barang fast moving maupun barang slow moving.
3. Meminimalisir kerusakan barang, karena belum tentu semua barang yang rusak bisa di retur atau diganti.
4. Mengetahui variabel apa saja yang terjadi di lapangan, yang sangat mempengaruhi tidak lakunya suatu item-item di dalam operasional Giant MOG Malang.
5. Yang paling menonjol dari pemanfaatan dengan adanya Data Mining ini maka, item-item yang ada pada perusahaan Giant MOG Malang bisa lebih terkontrol dan terkendali.

Kelemahan dari digunakannya Data Mining :

1. Membayar dari sistem Data Mining yang dibuat.
2. Perawatan atau maintenance yang dilakukan pada pusat.
3. Adanya variabel lain yang tidak tertera dalam database harus bisa dirasakan para manager operasional itu sendiri. Karena hal ini sangat berpengaruh sekali dalam arus transaksi pada operasional itu sendiri.

PENUTUP

Dari hasil uji coba yang sudah dilakukan dapat diambil simpulan sebagai berikut :

Saran yang bisa diberikan berkaitan dengan penelitian ini untuk pengembangan selanjutnya adalah :

1. Pada penelitian ini algoritma, hanya digunakan untuk penggalan kaidah asosiasi multi obyektif dimensi tunggal, sehingga bisa dikembangkan untuk menyelesaikan permasalahan penggalan kaidah asosiasi multi obyektif dimensi banyak.
2. Menggunakan teknik sampling yang lain, misalnya *random sampling* atau *regression-based sampling*.

DAFTAR PUSTAKA

1. Agrawal, R., Imielinski, T., & Swami, T., 1993. Mining association rules between sets of items in large databases. In Proceedings of ACM SIGMOD international conference on management of data (SIGMOD'93) (pp. 207–216).

2. Ghosh, Ashish Bhabesh Nath, 2004. "Multi-Objective Rule Mining using Genetic Algorithms", Information Sciences 163 (2004), 123-133.
3. Han, J., Kamber M., 2000. Data Mining : Concept and Techniques. Morgan Kaufman Publishers.
4. Freitas,A. 2002. "Data Mining and Knowledge Discovery with Evolutionary Algorithms"Springer-Verlag, New York, 2002.
5. MacQueen, J.B., 1967. Some Methods For Classification And Analysis Of Multivariate Observations, Proc. Of 5th Berkeley Symposium On Mathematical Statistic And Probability, Berkeley, University Of California Press, 1:281-297.
6. Bezdek, J.C 1981. Pattern Recognition With Fuzzy Objective Function Algorithm, Plenum Press, New York