PERBANDINGAN PENCARIAN FREQUENT ITEMSET MENGGUNAKAN ALGORITMA CUT BOTH WAYS DAN ALGORITMA APRIORI COMPARISON OF FREQUENT ITEMSET GENERATION USING CUT BOTH WAYS ALGORITHM AND APRIORI ALGORITHM

Oyo Sukarya[1], -[2]

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

Abstrak

Penggalian kaidah asosiasi (mining association rules) merupakan salah satu proses data mining untuk menemukan pola dan aturan (rule) dari sekumpulan data yang besar. Pola-pola ini merupakan kumpulan item (itemset) yang sering muncul secara bersamaan (frequent itemset) dalam transaksi pada basis data. Proses pencarian frequent itemset membutuhkan waktu yang sangat lama, oleh karena itu diperlukan suatu algoritma yang bisa mengefisiensi waktu yang dibutuhkan.

Algoritma yang paling populer saat ini adalah algoritma apriori yang menggunakan support base pruning (membuang ruang pencarian dengan batasan nilai support). Algoritma ini memiliki kelemahan ketika kardinalitas pada longest frequent itemset berupa k, membutuhkan sebanyak k pembacaan basis data dan memiliki sifat computation-intensive dalam membangkitkan kandidat itemset dan penghitungan nilai support, khususnya untuk aplikasi yang memiliki nilai support yang sangat rendah dan atau item yang sangat banyak.

Algoritma Cut Both Ways (CBW) menggunakan gabungan beberapa teknik dan menggunakan cutting level (α) untuk membagi ruang pencarian menjadi dua bagian. Strategi top-down untuk menemukan frequent itemset yang berada dibawah cutting level dikombinasikan dengan strategi pencarian breadth first search dan horizontal counting untuk penghitungan nilai support. Sedangkan bottom-up untuk menemukan frequent itemset yang berada diatas cutting level dikombinasikan dengan depth first search dan vertical intersection. Nilai cutting level merupakan nilai rata-rata dari kardinalitas frequent itemset.

Pada tugas akhir ini akan mengimplementasikan proses pencarian frequent itemset dengan menggunakan algoritma Apriori dan CBW. Kemudian membandingkan kinerjanya dengan menggunakan beberapa parameter nilai support.

Kata Kunci : mining association rules, itemset, frequent itemset, support, support base pruning, longest frequent itemset, computation-intensive, cutting level, top-down, bottom-up, breadth first search, dept first search, vertical intersection.

Abstract

Mining association rules is a data mining process to find rule and pattern from a large database. The pattern can be frequent itemset from the transaction of databases. Frequent itemset generation is most time-consuming process, so we need an algorithm that can be eficient a time consuming.

A most popular algorithm is Arpriori which use support base pruning to prune a vast amount of non-candidate itemsets. This algorithm has disadvantages when the cardinality of longest itemset is k, apriori needs k passes of database scan, and it has. In addition, the apriori algorithm is computation-intensive in generating the candidate itemsets and counting the support values, especially for applications with very low support treshold and/or a vast amount of items.

Cut Both Ways (CBW) combine a various technic and use cutting level (α) to divide a search space into two different part. Top-down strategy combined with breadth first search and horizontal counting, are used to find frequent itemset at below of the cutting level. In the other hand, bottom-up strategy combined with depth first search and vertical intersection, are used to find frequent itemset at upper of the cutting level. Cutting level is an average cardinality of frequent itemsets, expecting that most of the frequent itemsets will apear in this level.

In this final project will implement frequent itemset generation using Apriori and CBW algorithm. Then, compare its performance by using different parameter of minimum support.

Keywords : mining association rules, itemset, frequent itemset, support, support base pruning, longest frequent itemset, computation-intensive, cutting level, top-down, bottom-up, breadth first search, dept first search, vertical intersection.

BAB I
PENDAHULUAN

1.1 Latar Belakang

Penggalian kaidah asosiasi (*mining association rules*) merupakan salah satu proses *data mining* yang mampu menemukan pola dan aturan (*rule*) dari sekumpulan data yang besar. Pada manajemen bisnis proses ini digunakan untuk menemukan pola-pola dan perilaku pelanggan dalam melakukan suatu transaksi dimana pola-pola ini merupakan kumpulan item (*itemset*) yang sering dibeli oleh pelanggan. Dengan mengetahui pola tersebut, proses manajemen bisnis menjadi lebih mudah.

Proses pencarian item-item yang sering dibeli atau sering muncul dalam transaksi (*frequent itemset*) pada data yang besar membutuhkan waktu yang sangat lama, dan hal ini merupakan syarat untuk membentuk kaidah asosiasi. Dari kondisi tersebut, diperlukan suatu algoritma yang bisa meminimalisasi pembacaan basis data, sehingga bisa mengoptimasi waktu yang dibutuhkan.

Algoritma yang paling populer saat ini adalah algoritma Apriori yang menggunakan prinsip Apriori, yaitu jika suatu *itemset* merupakan *frequent* (yang sering muncul), maka semua *subset*-nya akan berupa *frequent*. Algoritma ini merupakan pionir dalam menggunakan *support-base pruning* untuk mengontrol pertumbuhan eksponensial dari kandidat *itemset* secara sistematis. Proses pada algoritma ini membangkitkan *frequent itemset* per level, dimulai dari level 1-*itemset* sampai ke *longest itemset*, kandidat level yang baru dibentuk dari *frequent itemset* yang ditemukan di level sebelumnya lalu menentukan nilai *support*-nya.

Kelemahan yang paling kritis pada Apriori adalah bahwa ketika kardinalitas pada *longest frequent itemset* berupa k, Apriori membutuhkan sebanyak k pembacaan basis data. Hal lainnya adalah algoritma Apriori memiliki sifat *computation-intensive* dalam membangkitkan kandidat *itemset* dan

penghitungan nilai *support*, khususnya untuk aplikasi yang memiliki nilai *support* yang sangat rendah dan atau item yang sangat banyak.

Ada banyak macam algoritma yang sudah diajukan untuk memperbaiki efisiensi, termasuk DHP, Partition, DIC, Eclat, Top-down, FP-growth, dan yang lainnya. Walaupun semua macam algoritma ini mengadopsi teknik yang berbeda dan sudut pandang yang berbeda, tapi dapat dikelompokan menjadi tiga aspek yang berbeda, yaitu :

1. Strategi arah pencarian: *top-down* dan *bottom-up*
2. Strategi pencarian: *breadth first search (BFS)* dan *Dept first search (DFS)*
3. Strategi penghitungan : *vertical intersection* dan *horizontal counting*

Algoritma *Cut Both Ways* (*CBW*) menggunakan gabungan dari beberapa teknik ini. Kebanyakan algoritma yang mirip dengan algoritma Apriori, lebih menitikberatkan pada pembatasan dengan nilai *support* minimal untuk membuang sejumlah besar *infrequent itemset*, hal inilah yang mejadi motivasi dibentuknya algoritma CBW. Oleh karena itu algoritma ini menggunakan *cutting level* (α) untuk membagi ruang pencarian menjadi dua bagian. Nilai *cutting level* merupakan nilai rata-rata dari kardinalitas *frequent itemset*, yang diharapkan pada level ini akan ditemukan banyak *frequent itemset*.

Setelah menentukan nilai *cutting level*, algoritma ini melakukan pencarian ke bagian bawah untuk menemukan *frequent itemset* yang nilai kardinalitasnya kurang dari dan sama dengan nilai *cutting level* dan menentukan nilai *support*-nya. Selanjutnya melakukan pencarian ke bagian atas untuk menentukan semua *frequent itemset* dengan kardinalitas lebih dari *α* dengan mengikuti paradigma Apriori (membangkitkan *frequent itemset* per level) dan menggunakan *vertical intersection* untuk mengurangi penghitungan nilai support yang tidak perlu.

Permasalahan efisiensi dalam pencarian *frequent itemset* adalah proses minimalisasi pembacaan basis data dan penghitungan nilai *support*. Algoritma CBW menggunakan nilai *cutting level* untuk membagi ruang pencarian dan menggunakan *vertical intersection* untuk menentukan nilai *support* pada proses pencarian *bottom-up*.

1.2 Perumusan Masalah

Pencarian *frequent itemset* merupakan syarat dalam penggalian kaidah asosiasi dan memerlukan banyak waktu. Pada beberapa kasus, algoritma Apriori kurang baik digunakan untuk pencarian *frequent itemset* terutama untuk basis data yang besar, item yang sangat banyak dan nilai minimum *support* yang kecil. Algoritma CBW menggabungkan beberapa teknik dan membagi ruang pencarian menjadi dua bagian dengan nilai *cutting level,* sehingga memungkinkan algoritma ini lebih baik dibanding algoritma Apriori walau dengan kondisi yang sebelumnya disebutkan, tetapi mungkin saja algoritma Apriori lebih baik digunakan untuk keadaan yang sebaliknya.

1.3 Tujuan

Tujuan dari tugas akhir ini adalah untuk mengimplementasikan algoritma Apriori dan algoritma CBW untuk proses pencarian *frequent itemset.* Selanjutnya membandingkan kinerjanya untuk menentukan kelebihan dan kekurangannya terhadap kondisi data, sekaligus membuktikan efisiensi yang dihasilkan dari penggabungan beberapa teknik yang digunakan dan penggunaan nilai *cutting level*.

1.4 Batasan Masalah

Batasan permasalahan dalam tugas akhir ini antara lain :

1. Proses yang dilakukan sampai ditemukan *frequent itemset* maksimal, tidak dilajutkan dengan membentuk kaidah asosiasi.
2. Jenis data yang digunakan adalah data transaksi yang telah melalui proses *data cleaning*.
3. Analisis kompleksitas algoritma hanya dilakukan untuk kompleksitas waktu.

1.5 Metodologi

Metodologi yang digunakan untuk menyelesaikan tugas akhir ini adalah sebagai berikut :

1. Studi Pustaka

Pada tahap ini yang dilakukan adalah mencari, mengumpulkan dan mempelajari segala macam informasi yang berhubungan dengan *data mining*, *assocaition rules,* algoritma Apriori, teknik-teknik pencarian *frequent itemset* dan algoritma CBW.

2. Perancangan Perangkat Lunak

   Pada tahap ini dilakukan perancangan perangkat lunak, dimana ada 3 perancangan yang dibuat yaitu perancangan data, perancangan proses, dan perancangan antarmuka.

3. Pembuatan Program

   Pada pembuatan program dilakukan implementasi dari tahap sebelumnya. Pembuatan perangkat lunak untuk tugas akhir ini menggunakan bahasa pemrograman C#.Net .Sistem operasi yang digunakan adalah Windows XP. Basis data yang digunakan adalah SQL Server 2000.

4. Uji Coba dan Evaluasi

   Pada tahap ini dilakukan uji coba dari program yang telah dibuat dengan menggunakan data yang telah dipersiapkan sebelumnya. Selanjutnya, hasil dari pengujian program akan dievaluasi untuk menentukan kebenaran dan kehandalan.

5. Penyusunan Buku Tugas Akhir

   Pada tahap ini dilakukan penyusunan laporan dari mulai tahap studi literatur sampai tahap pengujian dari evaluasi.

1.6 Sistematika Penulisan

Sistematika penulisan dibagi menjadi :

BAB I PENDAHULUAN

Pada bab ini diuraikan latar belakang, perumusan masalah, tujuan, batasan masalah, metodologi serta sistematika penulisan.

BAB II DASAR TEORI

Pada bab ini dibahas teori umum yang meliputi definisi dan uraian singkat tentang *Data mining* dan *Association rule,* uraian

singkat dari algoritma Apriori, algoritma CBW dan teknik-teknik pencarian *frequent itemset* yang digunakan.

BAB III ANALISIS DAN PERANCANGAN SISTEM

Bab ini berisi analisis algoritma Apriori dan algoritma CBW dari segi teori dan kompleksitas, analisis dan perancangan perangkat lunak sebagai alat bantu dalam proses analisa.

BAB IV IMPLEMENTASI DAN HASIL PENGUJIAN

Bab ini memuat tentang analisis terhadap performansi algoritma Apriori dan algoritma CBW, membandingkan kinerja kedua algoritma berdasarkan parameter yang ada.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan yang diambil dari pembahasan bab-bab sebelumnya serta saran-saran untuk pengembangan selanjutnya.

BAB V
PENUTUP

5.1 Kesimpulan

1. Algoritma CBW mampu menambang *frequent itemset* lebih efisien dibanding algoritma Apriori, karena proses pembacaan basis data yang lebih sedikit. Hal ini terbukti dengan perbandingan waktu yang digunakan algoritma Apriori dua kali lipat dari algoritma CBW dengan beberapa parameter minimal *support*.
2. Perbedaan jumlah pembacaan basis data tidak berbanding langsung dengan waktu yang digunakan, karena ada perbedaan proses pembentukan kandidat *itemset.*
3. Performansi kedua algoritma berbanding terbalik dengan nilai minimal *support*, semakin kecil nilai minimal *support* semakin besar waktu yang digunakan.
4. Pertambahan jumlah transaksi untuk kedua algoritma cenderung linier dengan waktu yang digunakan. Hal ini disebabkan karena semakin banyaknya pembacaan basis data.
5. Performansi algoritma CBW pada saat minimal *support* semakin kecil lebih baik dibanding algoritma Apriori. Sedangkan pada saat minimal *support* semakin besar performansi algoritma Apriori mendekati performansi algoritma CBW, tetapi tidak bisa lebih baik dan hanya bisa sama untuk yang tidak ditemukan *frequent itemset*.

5.2 Saran

1. Prosedur *DwnSearch* pada algoritma CBW bisa lebih dioptimalkan lagi karena proses yang ada masih kurang baik dalam pembentukan kandidat *itemset* dan penghitungan nilai *support*-nya.
2. Cara mendapatkan nilai *cutting level* pada algoritma CBW masih terlalu sederhana sehingga nilainya tidak sesuai dengan harapan untuk data yang memiliki rata-rata *frequent item* yang tinggi.
3. Fasilitas yang disediakan oleh bahasa pemrograman bisa digunakan untuk membentuk algoritma baru yang lebih cepat prosesnya, misalkan penggunaan objek dataset pada C#.Net bisa dioptimalkan sehingga bisa mendapatkan algoritma yang efisien.

DAFTAR PUSTAKA

[1] Agrawal R., Srikant R., 1994, "Fast Algorithm for Mining Association Rules", Proceeding of the 20th VLDB Conference, Santiago, Chile.

[2] Dana Sulistyo, 2003, "Data Mining dengan Algoritma Apriori pada RDBMS Oracle", Tugas Akhir Jurusan Informatika STT Telkom Bandung.

[3] Craig Lamran, 1998, "Applying UML and Patterns An Introduction to Object-Oriented Analysis and Design", Prentice Hall, New Jersey.

[4] Dunham M. H., Xiao Y., Gruenwald L., Hossain Z., "A Survey of Association Rules", [12 Juli 2004] http://www.cs.uh.edu/~celck/6340/grue-assoc.pdf

[5] Fahry Ady Yamin, 2004, "Implementasi Data Mining untuk Penggalian Kaidah Asosiasi Menggunakan Metode Bottom-Up Algoritma Eclat", Tugas Akhir Jurusan Informatika STTTelkom Bandung.

[6] Han Jiawei, Kamber Micheline, 2001, "Data Mining Concepts and Techniques".

[7] Hwung Su J., Yang Lin W. 2004, "CBW: An Efficient Algoritma for Frequent Itemset Mining". Proceeding of the 37th Hawaii International Conference on System Sciences. http://csdl.computer.org/comp/proceedings/hicss/2004/2056/03/205630064c.pdf.

[8] J. Hipp, U. Guntzer, G. Nakhaeizadeh, "Mining Association Rules: Deriving a Superior Algorithm by Analyzing Today's Approaches", in Proceedings of 4th European Symposium on Principles of Data Mining and Knowledge Discovery(PKDD'00), 2000, pp. 159-168

[9] Sucahyo Y. G. 2003, Data Mining: Menggali Informasi yang Terpendam, Artikel Populer Ilmu Komputer.Com Copyright 2003. [23 september 2004]

[10] Widodo S. 2004. Perancangan Data Mining dengan Metode Association Rule untuk Analisis Cross-Market (Studi Kasus Toko Sinar Bogor), Skripsi S1 Departemen Ilmu Komputer, FMIPA Institut Pertanian Bogor.

[11] Zaki M. J. 2000, Scalable Algorithm for Association Mining, IEEE Transaction on Knowledge and Data Engineering, Vol. 12 No. 2, pp. 372-390