ANALISIS DAN IMPLEMENTASI UTILITY MINING MENGGUNAKAN TWO-PHASE ALGORITHM UNTUK MARKET BASKET ANALYSIS

Siddhiq Amarullah Ramadhani[1], Arie Ardiyanti Suryani[2], Imelda Ataina[3]

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

Abstrak

Ada kalanya frequent itemset yang dihasilkan oleh Traditional Association Rule Mining hanyalah barang-barang yang sering terjual bersamaan saja, dan tidak menghasilkan keuntungan yang cukup besar bagi pihak retailer. Oleh karena itu, Traditional Association Rule Mining tersebut dapat dikembangkan lebih jauh lagi menjadi Utility Mining, yang dapat menggali Itemset yang menghasilkan keuntungan tinggi. Pada penelitian ini, Two-Phase Algorithm dijadikan algoritma untuk mengimplementasikan Utility Mining. Pada Utility Mining, setiap item diberikan 2 buah jenis bobot, yaitu jumlah terjualnya item tersebut pada sebuah transaksi, dan profit yang dimiliki oleh item tersebut. Kemudian Two-Phase Algorithm mengkalkulasi 2 buah jenis bobot tersebut dalam 2 tahap, untuk menghasilkan knowledge berupa sejumlah itemset yang memiliki nilai profit dan asosiasi yang tinggi Pada penelitian ini, terdapat dua buah variabel yang mempengaruhi performansi, yaitu minimum utility threshold dan minimum confidence threshold. Minimum utility threshold mempengaruhi jumlah knowledge yang dihasilkan juga waktu pemrosesan. Semakin kecil minimum utility threshold maka jumlah knowledge yang dihasilkan semakin banyak, dan waktu pemrosesan semakin lama, serta akurasi meningkat. Sedangkan minimum confidence threshold mempengaruhi jumlah knowledge yang dihasilkan dan akurasi. Jumlah knowledge semakin meningkat seiring dengan diturunkannya nilai minimum confidence threshold.

Kata Kunci : Knowledge, Market Basket Analysis, Utility Mining, Two-Phase Algorithm

Abstract

There are times when frequent itemset which is generated by the Traditional Association Rule Mining are the items that are only often sold together, and do not generate substantial profits for the retailer. Therefore, the Traditional Association Rule Mining can be further developed into Utility Mining, which can digs high profitability itemset. In this final project, Two-Phase Algorithm is used to implement the Utility Mining. On Utility Mining, each item is given two kind of weights, ie the item sold count in a transaction, and a profit for each item. Then the Two-Phase Algorithm 2 calculates the weight these 2 weights in 2 stages, to produce knowledge, which is itemsets which have high profitability and association value In this study, there are two variables that affect the performance, the minimum utility threshold and minimum confidence threshold. Minimum threshold affects the number of number of knowledge generated and processing time. The smaller the minimum utility threshold, the number of knowledge generated more and more, and the longer the processing time, with the increase of accuracy. While the minimum confidence threshold also affects the number of knowledge generated and the accuracy. The number of knowledge are increasing with lowered minimum confidence threshold

Keywords : Knowledge, Market Basket Analysis, Utility Mining, Two-Phase Algorithm

Bab I
Pendahuluan

1.1 Latar Belakang

Ada perkataan yang berbunyi "*we live in age of information*"[8]. Perkataan tersebut juga ternyata menimbulkan pertanyaan baru: "*where does information come from?*". Jawaban dari pertanyaan tersebut adalah data. Informasi bisa dibilang sebagai jembatan untuk menghubungkan 2 hal, yaitu data dan *knowledge*. Informasi merupakan data yang telah diolah, sedangkan informasi bisa dikembangkan lagi menjadi *knowledge.* Transformasi untuk menciptakan informasi yang berharga dari data yang melimpah ruah, juga mengolah informasi menjadi *knowledge* yang terorganisir, merupakan alasan utama mengapa bidang ilmu *Data Mining* muncul[1].

Oleh karena itu, sangat banyak perusahaan yang memanfaatkan teknologi yang sudah mengimplementasikan *Data Mining* untuk meningkatkan profit bagi perusahaannya. Bila ruang lingkup perusahaan dipersempit lagi menjadi *retailer*, yang bergerak pada bidang transaksi barang, cara untuk meningkatkan profit adalah mengetahui jumlah terjualnya sebuah barang, profit dari barang tersebut, beserta barang-barang yang sering terjual bersamaan. Problem tersebut menjadi inti kajian utama pada *Domain Market Basket Analysis*, yang juga merupakan subset dari salah satu bidang ilmu dalam *Data Mining*, yaitu *Association rule Mining*[7].

Problem yang sudah muncul bertahun-tahun sejak berdirinya perusahaan *retailer*, tentunya sudah menghasilkan sejumlah solusi yang diciptakan agar bisa menanggulangi masalah. Namun ternyata solusi-solusi tersebut juga menghasilkan problem baru. Algoritma untuk *traditional rule mining*, seperti Apriori yang diperkenalkan oleh Agrawal et.al, hanya dapat mencari *frequent itemset* saja[9]. Terdapat kemungkinan bahwa *Frequent Itemset* yang berhasil diidentifikasi oleh *traditional association rule mining* hanya memberikan profit

dengan porsi kecil pada *retailer*, sedangkan *non-frequent Itemset* mungkin saja bisa memberikan profit besar pada *retailer*[15].

Contohnya, sebuah *item* A dengan *profit* sebesar 2000 rupiah per *item* terjual sebanyak 10 kali pada data transaksi. Sedangkan *item* lainnya, sebut saja *item* B, memiliki *profit* sebesar 15.000 rupiah dan terjual sebanyak 2 kali. Bila diukur berdasarkan keuntungan, tentu saja *item* B yang lebih unggul. Namun *item* B bisa saja tidak diklasifikasikan sebagai *frequent Itemset* pada *traditional association rule mining*, karena frekuensi kemunculannya pada data transaksi yang sedikit. Oleh karena itu, diperlukan sebuah algoritma yang memberikan pembobotan berupa jumlah sebuah *item* yang dibeli dalam sebuah transaksi dan juga *profit* dari *item* tersebut. Algoritma tersebut diharapkan dapat menghasilkan *highly profitable Itemset* pada data transaksi dalam jumlah besar dengan kecepatan yang lebih baik dibandingkan dengan algoritma pendahulunya serta akurasi yang tinggi[2][4]

1.2 Perumusan masalah

Sejumlah permasalahan yang harus diselesaikan dalam mengimplementasikan *Two-Phase Algorithm* pada *Market Basket Analysis* adalah sebagai berikut:

1. Bagaimana cara untuk mengetahui sejumlah *item* yang memiliki nilai profit dan asosiasi yang tinggi dari keseluruhan transaksi yang ada?
2. Bagaimana cara untuk mengukur dan mengetahui performansi dari *Two-Phase Algorithm* yang akan diimplementasikan?

1.3 Tujuan

Tujuan yang akan dicapai dari Tugas Akhir ini adalah sebagai berikut:

1. Mengimplementasikan *Two-Phase Algorithm* untuk menghasilkan *knowledge*, yaitu *itemset* yang memiliki nilai *utility* dan *confidence* diatas *threshold* masing-masing

2. Menganalisis performansi dari *Two-Phase Algorithm* berdasarkan banyaknya *knowledge*, waktu pemrosesan, dan akurasi *knowledge* dengan *dataset* yang bermacam-macam

1.4 Ruang Lingkup dan Batasan Masalah

1. Metode *Weighted Association Rule Mining* yang akan diimplementasikan adalah *Two-Phase Algorithm*
2. *Software/Programming Language* yang akan digunakan adalah *Matlab R2011a*
3. *Database* yang digunakan adalah *Spreadsheet* yang dihasilkan oleh *Microsoft Excel 2010*
4. *Dataset* untuk data transaksi yang akan di-*mining* didapatkan melalui kerjasama dengan *retailer XYZ*, dari kurun waktu bulan Juli-Desember 2012, dengan jumlah total transaksi sebanyak 29.000 dan jumlah *item* sebanyak 2448
5. *DIKW Pyramid* yang dianalisis hanya hingga tahap *Knowledge* saja
6. *Data Cleaning* tidak digunakan pada tahap *Preprocessing*

1.5 Metodologi penyelesaian masalah

Beberapa metodologi untuk pengerjaan Tugas Akhir ini adalah sebagai berikut:

1. Identifikasi masalah

   Masalah yang akan dipecahkan pertama-tama diidentifikasi terlebih dahulu, agar bisa mengetahui dengan lebih detail inti dari masalah yang akan diselesaikan juga bagaimana proses serta metode untuk menyelesaikan masalah tersebut

2. Studi literatur dan wawancara

   Studi literatur dilakukan dengan cara mengunjungi *repository* Tugas Akhir, membaca jurnal ilmiah, juga buku-buku teknis yang didapat dari Internet ataupun perpustakaan. Metode wawancara juga dilakukan kepada dosen

pembimbing, dosen lain dengan kompetensi yang sama, dan mahasiswa lain yang dianggap memiliki pengetahuan yang memadai

3. Merancang kebutuhan sistem

   Penulis merancang kebutuhan sistem berupa *dataset* yang akan diinputkan, parameter apa saja yang bisa diubah, beserta hasil yang diharapkan. Perancangan ini akan dilakukan dengan matang agar proses selanjutnya, yaitu implementasi, bisa dikerjakan dengan lebih mudah karena sudah mendapatkan gambaran tentang bagaimana sistem bekerja

4. Mengimplementasikan sistem dan menganalisis hasilnya

   Berikut adalah tahap-tahap implementasi serta analisisnya:

   1. Memberikan bobot berupa *item sold count* beserta profit untuk mencari nilai *transaction–weighted utilization* dan *utility* pada *item*set
   2. Menggunakan parameter *min_utility* untuk memisahkan *high transaction-weighted utilization itemsets* dengan *weak transaction-weighted utilization itemsets*
   3. Menggunakan parameter *min_utility* untuk mengidentifikasi *high utility itemsets* dari keseluruhan *high transaction-weighted utilitzation itemsets*
   4. Menggunakan parameter *min_confidence* untuk memisahkan *weak association rule* dengan *strong association rule*
   5. Mengukur akurasi *knowledge* yang telah di-*generate*

5. Menyimpulkan hasil dari penelitian ini beserta saran untuk penelitian kedepannya

   Setelah hasil dianalisis, terdapat sebuah kesimpulan yang bisa diambil dan diharapkan berguna bagi kemajuan bidang yang penulis teliti. Penulis juga menyadari bahwa masih ada sejumlah sub-permasalahan yang bisa digali dari penelitian ini yang akan penulis sarankan untuk diteliti kembali di masa depan

6. Pembuatan buku TA

Setelah keempat proses metodologi tersebut selesai, barulah buku TA bisa dibuat dan diselesaikan secara lengkap yang dimulai dari alasan penelitian ini diadakan hingga kesimpulan yang bisa ditarik dari penelitian ini

1.6 Sistematika Penyajian

Buku Tugas Akhir ini disajikan dengan sistematika sebagai berikut:

- Bab I Pendahuluan: Berisi latar belakang, rumusan masalah, tujuan, ruang lingkup, dan metodologi penelitian
- Bab II Landasan Teori: Berisi keseluruhan teori yang diperlukan untuk memahami, mengimplementasikan, dan menganalisis hasil dari penelitian ini
- Bab III Perancangan Sistem: Berisi penjelasan tentang bagaimana sistem untuk penelitian ini dirancang, dimulai dari deskripsi dan kebutuhan yang harus dimiliki sistem hingga alur kerja dari sistem
- Bab IV Pengujian & Analisis: Berisi penjelasan data untuk pengujian beserta hasil analisis dari penelitian ini, berdasarkan dari sistem yang telah diimplementasikan dan diujikan
- Bab V Penutup: Berisi kesimpulan yang berhasil dipetik dari penelitian ini juga untuk penelitian selanjutnya

Bab V

Penutup

5.1 Kesimpulan

Beberapa kesimpulan yang didapat melalui penelitian ini adalah sebagai berikut:

1. Sebuah *itemset* bisa digolongkan sebagai *itemset* yang memiliki profit tinggi jika nilai *utility*-nya diatas *minimum utility threshold*
2. Sebuah *itemset* bisa disebut memiliki asosiasi yang bagus jika nilai *confidence*-nya diatas *minimum confidence threshold*
3. Nilai *minimum utility threshold* dan *minimum confidence threshold* sangat mempengaruhi jumlah *knowledge* yang dihasilkan. Semakin tinggi nilai *minimum utility threshold* dan nilai *minimum confidence threshold*, maka jumlah *knowledge* yang ditampilkan pun akan semakin sedikit pula. Begitu pula sebaliknya. Jumlah *knowledge* akan meningkat seiring dengan diturunkannya kedua *threshold* tersebut.
4. Berbeda dengan jumlah *knowledge*, untuk *processing time*, *independent variable* yang berpengaruh hanyalah nilai *minimum utility threshold*. Waktu pemrosesan yang dilakukan sistem berbanding terbalik dengan nilai *minimum utility threshold* yang ditetapkan. *Processing time* akan menurun jika nilai *minimum utility threshold* ditingkatkan. Semakin kecil nilai *minimum utility threshold*, maka nilai *processing time* pun akan membesar / bertambah lama
5. Sama seperti jumlah *knowledge*, akurasi yang dimiliki oleh sistem juga dipengaruhi oleh nilai *minimum utility threshold* dan nilai *minimum confidence threshold*. Hal tersebut dikarenakan, makin sedikit jumlah *knowledge* yang dihasilkan, maka akurasi pun akan bertambah baik. Sedangkan ketika nilai *minimum confidence threshold* diset pada nilai 100%, besar kemungkinannya bahwa akurasi sistem akan mencapai angka sempurna, yakni 100%

6. Berdasarkan sudut pandang jumlah *knowledge*, nilai *minimum utility threshold* yang optimal tergantung berdasarkan *dataset*-nya, namun berkisar antara 1% hingga 1.5%
7. Berdasarkan sudut pandang *processing time*, nilai *minimum utility threshold* yang terbaik berada pada nilai 3%, karena *processing time* akan semakin cepat seiring dengan ditingkatkannya nilai *minimum utility threshold*
8. Berdasarkan sudut pandang akurasi sistem, nilai akurasi terbaik dapat diraih ketika nilai *minimum confidence threshold* ditetapkan pada nilai 100% atau ketika nilai *minimum utility threshold* ditetapkan pada nilai 2.5%

5.2 Saran

Beberapa saran yang dapat dihasilkan dari penelitian ini adalah sebagai berikut:

1. Mengimplementasikan *Two-Phase Algorithm* pada *Programming Language* lainnya selain *Matlab*, seperti *C* ataupun *Java*. Untuk mengetahui perbedaan performansi, terutama dari aspek *processing time*
2. *Menggunakan Database Management System* (DBMS) seperti *MySQL* ataupun *Oracle Database* untuk menyimpan data transaksi
3. Memperhitungkan aspek kemunculan jumlah item pada keseluruhan transaksi dengan menambahkan bobot berupa *support* disamping nilai profit dan *item sold count* pada *Utility Mining*. Tentunya dengan menggunakan algoritma yang berbeda dengan *Two-Phase Algorithm*
4. Menambah proses untuk mengeliminasi *association rule* dimana *antecedent* dan *consequent itemset*-nya bersifat *redundant* / mirip antara satu sama lain, baik secara manual ataupun mengimplementasikan algoritma
5. Menggunakan metode *online learning* dan mengintegrasikan sistem dengan sistem kasir *retailer*. Agar lebih cepat dalam hal menseleksi dan memproses data secara *real time*

Daftar Pustaka

[1] H. Jiawei, M. Kamber, J. Pei (2011). *Data Mining: Concepts and Techniques, 3rd Edtion*. Burlington: Morgan Kauffman Publishers

[2] H.Yao,H.J.Hamilton ,*Mining Itemset Utilities From Transacation Databases*, in Data and Knowledge Enineering 59(2006) pp.603-626

[3] Hong Yao, Howard J. Hamilton, and Cory J Butz. *A Foundational Approach to Mining Itemset Utilities from Databases*. SDM (2004)

[4] Jiawei Han, Jian Pei, and Yiwen Yin. 2000. *Mining Frequent Pattern without Candidate Generation*. In ACM 2000 1-58113-218-2/00/05

[5] Jianying Hu, Aleksandra Mojsilovic, "*High-utility pattern mining: A method for discovery of high-utility item sets*", *Pattern Recognition*, Elsevier Science Inc, Volume 40 , Issue 11, pp. 3317-3324, (2007).

[6] Jyothi Pillai. *User Centric Approach To Itemset Utility Mining In Market Basket Analysis*. In *International Journal on Computer Science and Engineering (IJCSE)* Vol.3 No.1 Jan. 2011 pp. 393 – 400

[7] Kotsiantis, S., Kanellopoulus, D. *Association rules mining: A recent overview*. *International Transactions on Computer Science and Engineering Journal*, 2006. 32, 1, pp. 71-82.

[8] Lombardi, O.. *What is information?*. In *Foundation of Science*, 9, 105–134, 2004.

[9] R. Agrawal and R. Srikant. 1994. *Fast Algorithms For Mining Association rules*. In VLDB'94, pp. 487-499.

[10] Raorane A.A., Kulkarni R.V. and Jitkar B.D., *Association rule–Extracting Knowledge Using Market Basket Analysis*. In *Research Journal of Recent Sciences*, 1(2), 19-27 (2012)

[11] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, Shalom Tsur, *Dynamic Itemset Counting and Implication Rules for Market Basket Data* in *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, volume 26(2) of SIGMOD Record, pp. 255–264. ACM Press

[12] Trewartha, D., *Investigating Data Mining in MATLAB*, Bachelor (Honours) of Science Thesis of Rhodes University, 2006.

[13] V. Podpecan , N. Lavrac, I. Kononenkom, *A Fast Algorithm For Mining Utility-Frequent Itemsets*,in workshop on Constraint-Based Mining and Learning at ECML/PKDD ,2007, pp. 9-20.

[14] W. Wang, J. Yang, P.S. Yu, *Efficient mining of weighted association rules (WAR)*, in: Proceedings of SIGKDD 2000, 2000.

[15] Y. Liu, W.-K. Liao, A. Choudhary, *A Two-Phase Algorithm For Fast Discovery Of High Utility Itemsets*. In: *Proceedings of the Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2005, pp. 689–694.

[16] Zaki, M. "*Generating non-redundant association rules," KDD-2000*, 2000.