Pencarian Asosiasi Produk Pada Dataset Transaksi Retail Menggunakan Algoritma Apriori

Mukhlis Saputro
Riset Manajemen dan Keamanan Informasi
Institut Teknologi Bandung
Bandung, Jawa Barat
23223063@mahasiswa.itb.ac.id

***Abstract*—Apriori merupakan algoritma yang umum digunakan untuk menemukan pola keterhubungan antar produk yang sering dibeli dalam suatu data transaksi. Algoritma ini dapat membantu seorang manajer dalam menentukan tata letak barang, menyusun strategi promosi, waktu penawaran dan lain sebagainya berdasarkan pola pembelian yang ditemukan. Apriori membantu meningkatkan profit, efisiensi, sekaligus meningkatkan pengalaman belanja pelanggan.**

***Index Terms*—apriori, dataset, market basket analysis**

I. Pendahuluan

Penjualan memiliki peranan krusial dalam keberhasilan usaha. Strategi penjualan yang baik sangat penting, karena dapat berdampak positif pada laba dan pendapatan usaha. Strategi ini bisa didapat dengan mengumpulkan data transaksi dengan jumlah yang sangat besar [1]. Selanjutnya, data tersebut diolah dan dianalisis untuk menghasilkan informasi menggunakan metode khusus. Secara umum, pengolahan data menjadi informasi ini sering disebut sebagai data mining.

Aturan asosiasi adalah salah satu metode dalam data mining yang digunakan untuk menemukan pengetahuan dari sejumlah besar data dalam basis data [2]. Metode ini bertujuan untuk menemukan informasi dari produk yang sering muncul dalam transaksi menggunakan minimum nilai dukung (*support*), serta mencari keterhubungan dari produk-produk dengan tingkat keyakinan tertentu (*confidence*). Umumnya metode yang digunakan dalam mencari asosiasi produk ini yaitu apriori. Informasi yang akan diekstrak memainkan peran penting dalam mendukung pengambilan keputusan perusahaan, organisasi bisnis, dan proses bisnis.

Pada pendalaman algoritma apriori ini, data penjualan yang akan digunakan merupakan data buatan yang di *generate* menggunakan python. Data penjualan terdiri dari 30.000 transaksi yang berisi nama kostumer, barang, kuantiti, jumlah uang yang dibayar dan lain sebagainya dengan format csv [3]. Data ini akan membantu dalam memberikan contoh untuk menemukan wawasan tentang perilaku pembelian pelanggan dan operasi toko menggunakan algoritma apriori.

II. Pengenalan Aturan dan Data

A. Definisi Aturan

Definisi yang digunakan dalam metode apriori akan dijelaskan di bawah ini. Definisi ini akan membantu dalam memahami penerapan apriori dalam mengelola data tahap demi tahap.

1) Itemset
   Itemset merupakan kumpulan dari item atau produk yang muncul bersamaan dalam satu transaksi. Contoh itemset p = `{Kopi, Gula, Roti}`
2) Transaksi
   Transaksi merupakan data yang terdiri dari satu kostumer, satu atau banyak produk, kuantiti, total pembelian yang terdaftar pada satu kunci utama. Contoh transaksi t = `{Johan, (itemset p), 8, 25.000}`
3) Aturan Asosiasi
   Setiap aturan asosiasi menjelaskan keterhubungan suatu produk dengan produk lainnya. Contohnya jika Johan membeli kopi, maka dia juga membeli gula. Aturan asosiasi ini dapat disimbolkan menjadi `{A}` to `{B}`
4) *Support*
   Support digunakan untuk mencari kemungkinan hubungan antara transaksi A terhadap seluruh total transaksi. *Support* dapat dirumuskan sebagai berikut:
   $$\bar{Support}(A) = \frac{A}{N}$$
5) *Confidence*
   Confidence digunakan untuk mencari seberapa besar kemungkinan asosiasi antara item A dan B ditemukan dalam transaksi A. *Confidence* dapat dirumuskan sebagai berikut:
   $$\bar{Confidence}(A \rightarrow B) = \frac{AUB}{A}$$
6) *Lift*
   Lift digunakan untuk mengetahui kekuatan aturan asosiasi yang telah ditemukan dari nilai *support* dan *confidence*. Nilai *lift* digunakan sebagai penentu apakah aturan asosiasi valid atau tidak valid. *Lift* dapat dirumuskan sebagai berikut:
   $$Lift(\bar{A} \rightarrow B) = \frac{AUB}{Support(A) x Support(B)}$$

B. Data Cleaning

Sebelum data diolah, dilakukan penghapusan beberapa karakter atau simbol spesial yang ada di dalam dataset. Penghapusan ini dimaksudkan untuk menghindari *error* saat

pemanggilan atau pemrosesan data menggunakan kode pemrograman. Karakter spesial yang dihapus dalam dataset seperti tanda petik (”) dan buka tutup kurung ([]]).

C. Pengolahan Data

Data yang terkumpul dan akan diolah bersifat semi struktured. Sehingga sebelum informasi yang dibutuhkan dapat diekstraksi, diperlukan lebih banyak tahapan pra pemrosesan dan parsing. Tahapan tersebut antara lain sebagai berikut:

1) Perubahan format data
   Apriori yang dibangun akan menggunakan bahasa pemrograman PHP dengan memanfaatkan database mysql. Maka dari itu, perubahan format csv menjadi sql diperlukan, sehingga dataset dapat dimasukkan ke dalam database mysql. Perubahan ini menggunakan *tools* yang beredar secara online dan gratis.
2) Penghapusan transaksi dengan satu jenis produk
   Apriori bertujuan untuk menemukan keterkaitan minimal 2 produk. Sehingga transaksi yang hanya memiliki satu jenis produk dapat dihapus atau diabaikan. Hasilnya, dari 30.000 transaksi yang dimiliki, tersisa 24.077 yang memenuhi syarat apriori.
3) Penghapusan transaksi dengan satu kuantiti
   Dataset yang di *generate* memiliki kandungan transaksi yang bernilai lebih dari satu jenis produk, namun hanya memiliki total kuantiti sebanyak 1. Hal ini mungkin terjadi bilamana dalam satu transaksi, terdapat 1 produk yang terjual, dan pembeli mendapatkan bonus berupa produk lain secara gratis. Data transaksi dengan jenis seperti ini dihapus atau diabaikan karena dapat membuat hasil asosiasi tidak relevan. Walaupun jenis produk yang terdaftar cukup banyak, namun produk tersebut tidak memiliki profit dari penjualan.
4) Pengabaian transaksi dengan jumlah kuantiti kurang dari jenis produk yang terjual
   Dataset yang di *generate* juga terindikasi memiliki transaksi yang menyimpan lebih dari satu jenis produk, namun memiliki total kuantiti kurang dari jenis produk tersebut. Hal ini mungkin terjadi bilamana dalam satu transaksi, terdapat beberapa produk yang terjual, dan dalam pembelian tersebut, salah satu atau beberapa produknya memiliki bonus berupa produk lain secara gratis. Sama seperti pada tahap 2, transaksi dengan jenis seperti ini diabaikan karena dapat membuat hasil asosiasi tidak relevan. Hasilnya, tersisa 21.616 yang memenuhi untuk dilanjutkan dalam proses apriori. Keseluruhan hasil dapat dilihat pada Fig 1.

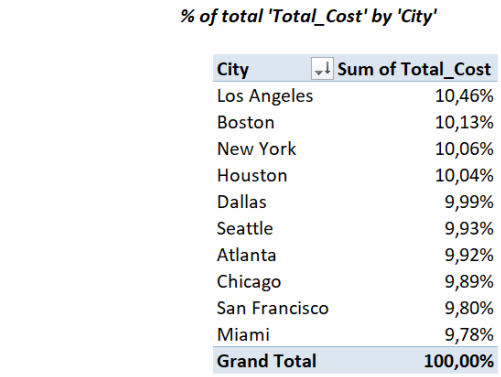
Table	Action							Rows	Type
datacleansing	Browse	Structure	Search	Insert	Empty	Drop		21,636	InnoDB
transaction	Browse	Structure	Search	Insert	Empty	Drop		30,000	InnoDB
2 tables	Sum							51,636	InnoDB

Fig. 1. Dataset Sebelum dan Sesudah Pra Pemrosesan Data

III. Perancangan Apriori

A. Analisis Data Model

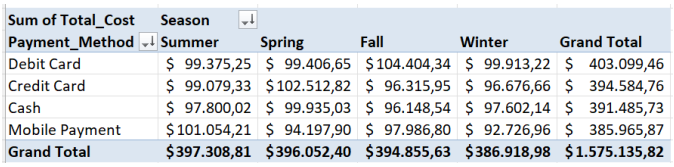
Proses untuk menemukan asosiasi, terlebih dahulu diawali dengan menemukan model data yang ingin diolah. Model data begitu penting karena mempengaruhi dapat tidaknya aturan asosiasi yang baik. Model data yang dimaksud seperti apakah jenis transaksi yang dilakukan hanya menggunakan debit, apakah kostumer hanya berkategori professional, apakah kota tertentu saja yang akan diproses atau apakah data merupakan gabungan dari berbagai kondisi dan lain sebagainya. Proses ini menggunakan analisis yang ada pada program Microsoft Excel. Contoh analisis yang dilakukan seperti mencari kota yang memiliki total transaksi terbanyak (Fig 2), total transaksi dengan menggunakan jenis pembayaran (fig 3), maupun produk favorit yang dibeli di setiap kota (fig 4).



% of total 'Total_Cost' by 'City'

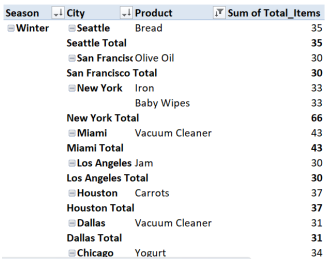
City	Sum of Total_Cost
Los Angeles	10,46%
Boston	10,13%
New York	10,06%
Houston	10,04%
Dallas	9,99%
Seattle	9,93%
Atlanta	9,92%
Chicago	9,89%
San Francisco	9,80%
Miami	9,78%
Grand Total	**100,00%**

Fig. 2. Jumlah Kota Dengan Transaksi Terbanyak



Sum of Total_Cost	Season				
Payment_Method	Summer	Spring	Fall	Winter	Grand Total
Debit Card	$ 99.375,25	$ 99.406,65	$ 104.404,34	$ 99.913,22	$ 403.099,46
Credit Card	$ 99.079,33	$ 102.512,82	$ 96.315,95	$ 96.676,66	$ 394.584,76
Cash	$ 97.800,02	$ 99.935,03	$ 96.148,54	$ 97.602,14	$ 391.485,73
Mobile Payment	$ 101.054,21	$ 94.197,90	$ 97.986,80	$ 92.726,96	$ 385.965,87
Grand Total	**$ 397.308,81**	**$ 396.052,40**	**$ 394.855,63**	**$ 386.918,98**	**$ 1.575.135,82**

Fig. 3. Total Transaksi Setiap Jenis Pembayaran



Season	City	Product	Sum of Total_Items
Winter	Seattle	Bread	35
	Seattle Total		35
	San Francisc	Olive Oil	30
	San Francisco Total		30
	New York	Iron	33
		Baby Wipes	33
	New York Total		66
	Miami	Vacuum Cleaner	43
	Miami Total		43
	Los Angeles	Jam	30
	Los Angeles Total		30
	Houston	Carrots	37
	Houston Total		37
	Dallas	Vacuum Cleaner	31
	Dallas Total		31
	Chicago	Yogurt	34

Fig. 4. Favorit Produk Disetiap Kota dan Musimnya

B. Penentuan Data Model

Model data yang digunakan pada uji coba ini dicontohkan dengan mencari hubungan produk yang tidak memiliki diskon, serta saat pembayaran tidak menggunakan kode promo di kota

Los Angeles. Dari keterhubungan ini akan dianalisis asosiasinya dan ditentukan strategi penjualan yang bisa diterapkan.

C. Pembuatan Tampilan

Selanjutnya, dilakukan pembuatan program apriori menggunakan kode pemrograman PHP dan database MYSQL. Tampilan dibuat sesederhana mungkin berupa tabel dan nilai-nilai dari setiap produk yang diolah menggunakan rumus. Diharapkan dengan tampilan ini memudahkan pembacaan tahap demi tahap suatu aturan ditemukan.

IV. Hasil

Analisis menggunakan model yang sudah dibangun menghasilkan nilai tingkat asosiasi yang beragam. Terutama ketika jumlah set yang diatur bertambah, nilai tingkat asosiasi yang ditemukan semakin tinggi. Contohnya ketika dataset yang ditentukan minimal itemset dengan 2 jenis produk, dengan nilai support, confidence dan lift diatur minimal 0,2, ditemukan 15 transaksi yang memenuhi. Salah satu produk yang memenuhi yaitu buah apel dengan penjualan sebanyak 3 buah, bernilai *support* yaitu 3 / 15 = 0.2, nilai *confidence* jika membeli Apel maka akan membeli *tootbrush* 0.67, dan *Lift* 0.167.

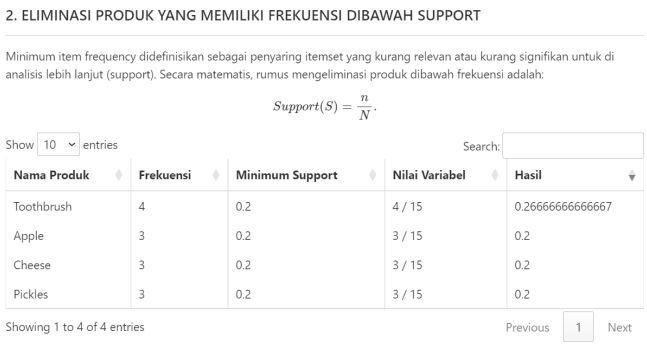


Fig. 5. Contoh Tampilan Program

Dicontohkan lain ketika yang ditentukan minimal itemset dengan 3 jenis produk, dengan nilai support, confidence dan lift diatur minimal 0,3, ditemukan 11 transaksi yang memenuhi. Produk yang memenuhi yaitu *tootbrush* dengan penjualan sebanyak 4 buah, bernilai *support* yaitu 4 / 15 = 0.36, nilai *confidence* jika membeli *tootbrush* dan apel maka akan membeli *yogurt* 0.5, dan *Lift* 0.5. Sehingga bisa disimpulkan, strategi yang mungkin dapat diaplikasikan yaitu pada apel dengan memberi promosi. Ketika seorang pelanggan tertarik membeli apel, ada kemungkinan pelanggan akan membeli *tootbrush* dan juga yogurt.

Hasil masing-masing nilai maksimal yang bisa ditemukan dengan berbagai rentang dataset dapat dilihat pada tabel di bawah ini.

Sayangnya, meskipun jumlah *support*, *confidence* dan *lift* semakin tinggi ketika nilai itemset dinaikkan, transaksi yang masuk ke dalam kategorinya juga semakin sedikit. Jumlah transaksi ini membuat bahan pertimbangan dirasa terlalu kecil dan kurang akurat dalam analisis. Sebaliknya, jika transaksi yang digunakan tidak dibatasi, terjadi penurunan *support* yang dapat mengakibatkan penurunan *confidence* dan *lift*, karena rumus *confidence* dan *lift* melibatkan perbandingan dengan nilai *support*. Hal ini bisa diakibatkan karena dataset yang digenerate cukup besar [4], itemset yang sering muncul jarang terjadi dan hampir rata pada setiap kota, jenis pembayaran, kategori kostumer, promosi dan lainnya.

TABLE I
Tabel Asosiasi

Jumlah Dataset	Minimal Nilai Asosiasi				
	Transaksi	Nilai S, C, L	*S*	*C*	*L*
2	15	0.2	0.26	0.6	0.7
3	11	0.3	0.36	1	1
4	11	0.4	0.36	1	1

[a]S = Support — C = Confidence — L = Lift

V. Kesimpulan

Algoritma apriori dapat membantu dalam menemukan keterhubungan antar satu produk dengan yang lainnya. Algoritma ini mudah diaplikasikan dan memiliki tingkat kepercayaan yang bisa dijanjikan. Kesulitan dalam mengaplikasikan metode ini adalah menemukan model dataset yang sesuai, ataupun mencari kondisi-kondisi dari dataset tersebut. Selain dari kesulitan pencarian model, beberapa hal yang bisa dijadikan perkembangan dalam apriori yaitu proses perhitungan yang dilakukan membutuhkan waktu dan sumber daya cukup besar untuk *scanning* dataset di dalam *database*, terutama jika dataset yang digunakan cukup besar.

References

[1] H. K. D. Sarma and S. Mishra, "Mining Time Series Data with Apriori Tid Algorithm", 2016 International Conference on Information Technology (ICIT), pp. 160-164, 2016.

[2] Patil, P. (2023, Oktober). Retail Transactions Dataset. Retrieved December 14, 2023 from https://www.kaggle.com/datasets/prasad22/retail-transactions-dataset/data.

[3] P. R P. and K. Prasad Rao, "A Comparative Study On Apriori And Reverse Apriori In Generation Of Frequent Item Set," 2019 1st International Conference on Advances in Information Technology, India, 2019, pp. 337-341, doi: 10.1109/ICAIT47043.2019.8987430.

[4] Y. Cong, "Research on Data Association Rules Mining Method Based on Improved Apriori Algorithm," 2020 International Conference on Big Data, Artificial Intelligence, Software Engineering (ICBASE), Bangkok, Thailand, 2020, pp. 373-376, doi: 10.1109/ICBASE51474.2020.00085.