Gunung Djati Conference Series, Volume 3 (2021)
Mini Seminar Kelas *Data Mining* 2020
ISSN: 2774-6585
Website: https://conferences.uinsgd.ac.id/gdcs

Implementasi Algoritma Prefixspan untuk Rekomendasi Penayangan Film di Bioskop

(*Implementation of the Prefixspan Algorithm for Recommended Cinema Showings*)

Adellia Rahmasari[1], Dimas Adi Putra Pratama[2], Nisvy Sya`bana Nugraha [3], Risnandy Maulana[4], Dinda Meysya Rochma[5]

[1]Teknik Informatika, UIN Sunan Gunung Djati Bandung, 1177050004@student.uinsgd.ac.id
[2]Teknik Informatika, UIN Sunan Gunung Djati Bandung, 1177050030@student.uinsgd.ac.id
[3]Teknik Informatika, UIN Sunan Gunung Djati Bandung, 1177050080@student.uinsgd.ac.id
[4]Teknik Informatika, UIN Sunan Gunung Djati Bandung, 1177050101@student.uinsgd.ac.id
[5]Teknik Informatika, UIN Sunan Gunung Djati Bandung, 1177050032@student.uinsgd.ac.id

Abstrak

Perkembangan teknologi yang semakin maju menyebabkan jumlah informasi pun semakin berlimpah. Setidaknya saat ini terdapat 16.3 ZB data atau setara dengan satu triliun GB. Tujuan dari penelitian ini adalah untuk mengetahui bagaimana penerapan algoritma Prefixspan untuk Sequential Pattern Mining (metode untuk mendapatkan pola yang terurut). Data yang digunakan yaitu Movie Meta Data yang mana di dalam data tersebut tedapat beberapa column di antaranya yaitu genre, rating, judul, film dan aktor. Dengan data tersebut kami bertujuan untuk menampilkan rekomendasi film untuk ditayangkan di bioskop berdasarkan antusias penonton dari data yang sudah diproses.

Kata kunci: *data mining*, *pattern mining*, prefixspan, sistem rekomendasi

Abstract

Along with technological developments that cause the amount of information available is also getting closer, at this time there are approximately 16.3 ZB data or the equivalent of one trillion GB. The purpose of this research is to see how the application of the Prefixspan algorithm for Sequential Pattern Mining (a method for obtaining an ordered pattern). The data used is Movie Meta Data in which there are several columns including genre, director, film etc. With this data, we purpose viewers to present recommendation films to be shown in theaters based on audience enthusiasm from the processed data.

Keywords: *data mining, pattern mining, prefixspan, recommendation system*

1 Pendahuluan

Menonton film merupakan sarana hiburan yang dilakukan oleh semua orang, dari anak kecil hingga orang dewasa gemar menonton film baik untuk menghilangkan stress maupun untuk mengisi waktu santai bersama keluarga. Genre film diperlukan untuk menentukan layak atau tidak film yang akan ditonton untuk berbagai usia. Dalam membuat rekomendasi menonton film ini, kami menggunakan Movie Meta Data dan melakukan proses Data Mining. Data mining bertujuan untuk menemukan, menggali atau mengumpulkan pengetahuan yang didapat dari data atau informasi yang sudah ada. Metode yang kami gunakan yaitu Algoritma Prefixspan untuk proses pattern mining agar data yang sudah ada bisa terurut dengan benar sesuai dengan rekomendasi film yang akan ditonton selanjutnya. Algoritma ini akan menampilkan pola berdasarkan kemunculan data dan urutannya.

Dalam artikel ini terdapat beberapa bagian yaitu metode penelitian, hasil, pembahasan dan kesimpulan. Juga dijelaskan bagaimana implementasi algoritma prefixspan dalam memproses movie meta data yang memiliki output rekomendasi film yang akan ditayangkan selanjutnya. Tujuan dari penelitian ini yaitu kami membuat rekomendasi film yang akan ditayangkan di bioskop. Karena dengan membuat rekomendasi ini, bioskop dapat memprediksi minat penonton berdasarkan dari genre, aktor ke aktor dan rating. Maka bioskop akan mendapatkan pasar yang lebih besar. Terdapat beberapa penelitian terkait penelitian ini, antara lain: (1) analisis frekuensi kemunculan dalam penjualan produk dengan menggunakan algoritma PrefixSpan [1], [2]; (2) analisis pola pergerakan saham menggunakan algoritma PrefixSpan [3]; dan sistem rekomendasi peminjaman buku menggunakan algoritma PrefixSpan [4];

2 Metodologi

Sebelum memproses data dilakukan data cleaning dan data selection. Data selection adalah data hasil seleksi yang akan digunakan untuk proses data mining. Dilakukan pemilihan data-data seperti apa saja yang dibutuhkan untuk proses lebih lanjut. Data yang diperoleh, baik dari database suatu perusahaan maupun eksperimen, memiliki isian-isian yang tidak sempurna seperti data yang hilang, data yang tidak valid atau juga hanya sekedar salah ketik. Data cleaning atau data juga mempengaruhi performasi dari sistem data mining karena data yang ditangani akan berkurang jumlah dan kompleksitasnya.

Algoritma PrefixSpan

PrefixSpan adalah salah satu algoritma untuk mendapatkan frequent sequential patterns (pola yang terurut dan frekuen). Algoritma ini akan mengenerate pola berdasarkan frekuensi kemunculan (hanya pola yang sering) dan ada urutan kemunculan yang juga diperhitungkan. Berikut dijelaskan tahap-tahap PrefixSpan [5], [6]. Sebuah sequence database S terdiri dari 4 sequence s yaitu:

1. < a (abc) (ac) d (cf) >
2. < (ad) c (bc) (ae) >
3. < (ef) (ab) (df) c b >
4. < e g (af) c b c >

Tahapan-tahapan algoritma PrefixSpan, antara lain [7]:

1. Mendapatkan sequential patterns *length-1*. Menghasilkan *sequential pattern Length-1* dengan struktur keluaran yang terdiri dari "<pattern>" : count" sehingga dihasilkankan <a> : 4, <b> : 4, <c> : 4, <d> : 3, <e> : 3, dan <f>i : 3. Perhatikan perhitungan frekuensi dari kumpulan *sequence*, item yang muncul beberapa kali dalam satu *sequence* dihitung 1 saja.
2. Mempartisi ruang pencarian. Pencarian secara dibagi berdasarkan 6 prefix di atas.
3. Mendapatkan subset dari sequential pattern. Setiap prefix yang telah ditemukan dalam langkah sebelumnya di*mine* untuk mendapatkan *subsets* dari *sequential patterns* melalui 2 langkah yaitu :
    a. Membentuk projected databases untuk setiap prefix. Misal pada prefix a, projected databasenya adalah <(abc) (ac) d (cf) >, < (_d) c (bc) (ae) >, < (_b) (df) c b >, < (_f) c b c >. Tanda _ menandakan bahwa a berada dalam satu event/elemen dengan item sebelahnya.
    b. Identifikasi frekuensi setiap item secara lokal dalam projected database yang telah dibentuk dan tentukan item yang frequent sesuai threshold sehingga menghasilkan a : 2, b: 4, _b : 2, c:4, d : 2, dan f : 2.
    c. Bentuk sequential pattern sesuai frequent item yang telah ditemukan sebelumnya membentuk <aa> : 2, <ab>: 4, <(ab)>: 2, <ac>:4, <ad> : 2, dan <af> : 2.

Gunung Djati Converence Series
ISSN 2774-6585
Volume 3 (2021)

4. Dari sequential pattern length-2 yang telah ditemukan diatas lakukan secara rekursif tiga step a,b,c untuk menemukan sequential pattern lainnya. Berikut langkahnya :
    a. Membentuk projected databases untuk setiap prefix. Misal pada prefix <aa>, projected databasenya adalah< (_bc) (ac) d (cf) >, < (_e) >
    b. Identifikasi frekuensi setiap item secara lokal dalam projected database yang telah dibentuk dan tentukan item-item yang frequent sesuai threshold, namun tidak ada yang memenuhi syarat frequent.
    c. Tidak ada sequential pattern yang dapat dibentuk.

3 Hasil dan Pembahasan

Data yang digunakan dalam penelitian ini adalah kumpulan data film dari mulai genre, sutradara, judul film dll. Selanjutnya data tersebut di preprocessing yang terdiri atas pengumpulan data untuk menghasilkan data mentah (raw data) yang dibutuhkan oleh data mining, penelitian ini menggunakan algoritma Prefixspan pada Sequential Pattern Mining (metode untuk mendapatkan pola yang terurut). Penelitian ini dilakukan untuk mengetahui rekomendasi film selanjutnya untuk ditonton. Sebelum memproses data dilakukan data cleaning dan data selection. Data cleaning untuk menghapus data yang tidak perlu dan *cell* yang kosong seperti durasi film. Data selection yaitu memilih data (*field*) untuk dijadikan fokus pembahasan dalam metode ini seperti genre, judul, aktor, rating dan film. Penerapan Prefixspan pada data melalui tahap mulai dari pemahaman bisnis hingga evaluasi. Detail setiap tahap dijelaskan berikut ini.

3.1 Pemahaman Bisnis

Tujuan dari menemukan pola adalah untuk melihat bagaimana antusias penonton terhadap film dengan mengacu pada daya tarik genre, aktor, rating, judul dan film. Diharapkan pihak bioskop dapat memahami pola agar menjadi bahan pertimbangan untuk menyediakan tontonan ke pasar yang lebih luas dengan cara memahami data dari kebiasaan penonton. Penawaran diharapkan akan menjadi lebih efektif karena tepat sasaran dengan adanya bantuan informasi pola kebiasaan penonton.

Pola kebiasaan penonton dapat dilihat dalam waktu 6 bulan. Pola mana yang dipakai disesuaikan dengan ritme bisnis terkait. Jumlah penonton berulang juga menjadi pertimbangan dalam menentukan jangka waktu pola.

Penonton yang akan diekstrak polanya yaitu sering (frequent) menonton di bioskop. Frequen mengacu pada banyak faktor yang disesuaikan dengan masing-masing kasus bisnis yang tidak identik. Misalkan dalam penelitian ini saya mengambil pola 3 bulanan, artinya satu tahun dibagi dalam 4 periode. Jika threshold 50% (50% dari 4 adalah 2) maka penonton dianggap sering jika dalam setahun memesan setidaknya di 2 periode atau lebih. Penonton yang sering ini yang akan menjadi referensi dalam menentukan pola pemesanan.

3.2 Pemahaman Data

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes	actor_2_name	actor_1_facebook_likes	gross
0	Color	James Cameron	723.0	178.0	0.0	855.0	Joel David Moore	1000.0	7605058
1	Color	Gore Verbinski	302.0	169.0	563.0	1000.0	Orlando Bloom	40000.0	3094041
2	Color	Sam Mendes	602.0	148.0	0.0	161.0	Rory Kinnear	11000.0	2000741
3	Color	Christopher Nolan	813.0	164.0	22000.0	23000.0	Christian Bale	27000.0	4481306
4	0	Doug Walker	0.0	0.0	131.0	0.0	Rob Walker	131.0	0.0

Gambar 1 Contoh Data

Data film yang terdapat pada Gambar 1 umumnya berupa tabel yang terdiri dari atribut/kolom yang berisi aktor, genre, tahun, rating, durasi, judul. Data film di dalam tabel berurutan setiap baris sesuai tahun film ditayangkan. Untuk penelitian ini, data yang diperlukan untuk mendapatkan pola urutan adalah data judul, genre, aktor, rating, dan film. Tabel asli data film dan attributnya lebih kompleks namun, kita mengambil data-data (atribut) yang dibutuhkan saja. Berikut contoh data yang akan ditranformasi.

3.3 Persiapan Data

Persiapan data dilakukan dengan beberapa tahapan yaitu mengambil data yang sesuai dari data film yaitu data film yang terdiri dari genre, judul, rating, film dan aktor. Data kemudian dibersihkan dari karakter-karakter khusus seperti tanda kutip, petik, spasi yang tidak perlu. Setelah itu dilakukan proses *distinct* untuk membuag data yang berulang, misalnya penonton yang sama menonton film yang berbeda dianggap menjadi satu rangkaian. Karena dalam penelitian ini memperhatikan genre dan urutan.

```
In [14]: M dataset.head()
```

Out[14]:

	director_name	actor_2_name	genres	actor_1_name	movie_title	actor_3_name	imdb_score
0	James Cameron	Joel David Moore	Action|Adventure|Fantasy|Sci-Fi	CCH Pounder	Avatar	Wes Studi	7.9
1	Gore Verbinski	Orlando Bloom	Action|Adventure|Fantasy	Johnny Depp	Pirates of the Caribbean: At World's End	Jack Davenport	7.1
2	Sam Mendes	Rory Kinnear	Action|Adventure|Thriller	Christoph Waltz	Spectre	Stephanie Sigman	6.8
3	Christopher Nolan	Christian Bale	Action|Thriller	Tom Hardy	The Dark Knight Rises	Joseph Gordon-Levitt	8.5
4	Doug Walker	Rob Walker	Documentary	Doug Walker	Star Wars: Episode VII - The Force Awakens	0	7.1

Gambar 2 Persiapan Data

3.4 Pemodelan dan Hasil

Tahap awal proses PrefixSpan adalah mendapatkan item yang frekuen (sering) yaitu penonton yang dianggap sering menonton film. Misal batasan yang dituntukan 50% dari jumlah range bulan data yang diujikan. Jika data yang diujikan dari bulan januari hingga april, hal itu berarti range 4 bulan. Penonton dianggap frekuen jika memesan 50% dari 4 bulan yaitu 2 bulan yang berarti memesan minimal di 2 bulan dalam range yang telah ditentukan. Data yang sudah ditransformasi dimasukan ke mesin PrefixSpan. Mesin membutuhkan minimum pemesan (threshold) misalkan 50 %. Kemudian pembentukan pola urutan yang frequen terbentuk dari hasil memasangkan penonton yang satu dengan

Gunung Djati Converence Series
ISSN 2774-6585
Volume 3 (2021)

yang lainnya dengan teknik prefixspan dan dilengkapi dengan berapa kali munculnya pola. Urutan film yang memenuhi minimum (minimal support) jumlah kemunculan adalah pola yang diambil. Penelitian memberikan solusi untuk merekomendasikan film yang akan ditayangkan di bioskop. Gambar 3 menunjukkan contoh pola yang dihasilkan dengan Algoritma PrefixSpan.

```
In [30]: print(ps.frequent(0.003, generator=True))
[(30, ['Blackhat']), (1, ['Blackhat', 'Rise of the Guardians', 'The Soloist', 'London Has Fallen', 'The Legend of Zorro',
'Saving Private Ryan', 'The Lord of the Rings: The Return of the King', 'Starship Troopers', 'The Hunger Games', 'X-Men:
Apocalypse', 'Seabiscuit', 'The Amazing Spider-Man 2', 'Avatar', 'Dreamcatcher', 'Mr. & Mrs. Smith', 'Harry Potter and th
e Half-Blood Prince', 'World Trade Center', 'Gods of Egypt', 'The Last Castle', 'Zodiac', 'Arthur and the Invisibles', 'T
he Great Gatsby', 'The Spanish Prisoner', 'Six Days Seven Nights', "Madagascar 3: Europe's Most Wanted", 'World Trade Cen
ter', 'The Majestic', 'Casino Royale', 'Gods and Generals', 'Catch Me If You Can']), (2, ['Blackhat', 'Rise of the Guardi
ans', 'Avatar']), (2, ['Blackhat', 'Saving Private Ryan']), (2, ['Blackhat', 'The Lord of the Rings: The Return of the Ki
ng']), (2, ['Blackhat', 'X-Men: Apocalypse', 'Avatar']), (2, ['Blackhat', 'The Amazing Spider-Man 2', 'Avatar']), (2, ['B
lackhat', 'The Amazing Spider-Man 2', 'Avatar', 'The Majestic']), (4, ['Blackhat', 'Avatar']), (2, ['Blackhat', 'The Last
Castle', 'World Trade Center']), (3, ['Blackhat', 'Zodiac']), (2, ['Blackhat', 'Zodiac', 'The Majestic']), (2, ['Blackha
t', 'Zodiac', 'Ice Age']), (2, ['Blackhat', 'The Great Gatsby']), (2, ['Blackhat', 'Six Days Seven Nights', 'The Majesti
c']), (2, ['Blackhat', "Madagascar 3: Europe's Most Wanted"]), (3, ['Blackhat', 'The Majestic']), (2, ['Blackhat', 'Missi
on: Impossible - Ghost Protocol']), (2, ['Blackhat', 'A Civil Action']), (2, ['Blackhat', 'Dr. Dolittle 2']), (2, ['Black
hat', 'Hercules']), (1, ['Blackhat', 'Superman Returns', '50 First Dates', 'Lincoln', 'Jack and Jill', 'Stealth', "Dadd
y's Home", 'Resident Evil: Afterlife', 'Bicentennial Man', 'Son of the Mask', 'Due Date', 'Eraser', 'The Fifth Element',
'Scooby-Doo', 'Seven Pounds', 'The Hobbit: The Desolation of Smaug', 'State of Play', 'Firewall', 'Black Hawk Down', 'The
Lion King', 'Cold Mountain', 'The Beach', 'Sinbad: Legend of the Seven Seas', 'The Bourne Ultimatum', 'Constantine', 'Sem
i-Pro', 'Eyes Wide Shut', 'I, Robot', 'Resident Evil: Afterlife', 'The Lego Movie']), (2, ['Blackhat', 'Lincoln']), (2,
['Blackhat', 'Jack and Jill']), (2, ['Blackhat', 'Resident Evil: Afterlife', 'Eyes Wide Shut']), (2, ['Blackhat', 'Erase
```

Gambar 3 Contoh Hasil Pola menggunakan PrefixSpan

4 Simpulan

Proses mining untuk mendapatkan pola mengikuti siklus atau metodologi yang terdiri dari tahap-tahap pemahaman bisnis dan data, mempersiapkan data, pemodelan, hingga evaluasi. Dalam kasus ini kami mengimplementasikan . Sequential pattern mining yang mana digunakan untuk mencari data yang memiliki urutan, data tersebut bisa merupakan urutan transaksi. Adapun metode yang digunakan yaitu sequence pattern mining menggunakan Prefixspan untuk menemukan pola dari sekelompok data. Apabila data tersebut berhasil di olah maka akan di jadikan sebagai pola referensi pemesanan.

Referensi

[1] M. F. Zamroni and A. Wibisono, “Analisis Frekuensi Kemunculan Fase Dalam Sales Process Pada Penjualan Perlengkapan Masak Berbasis Mobile Chat (Studi Kasus: Forbento.Com),” *J. Tek. ITS*, vol. 6, no. 2, 2017, doi: 10.12962/j23373539.v6i2.23178.

[2] P. N. Sabrina, “Penerapan Sequential Pattern Mining pada Data Pemesanan untuk Strategi Penawaran dan Pemasaran Produk Dengan Pendekatan Metode PrefixSpan,” in *Annual Research Seminar (ARS)*, 2017, vol. 2, no. 1, pp. 449–455.

[3] M. N. Wulandari, “Penggalian Pola Sekuensial Interval Waktu Fuzzy Pada Pergerakan Harga Saham Di Indonesia Menggunakan Algoritma Fp-Growth-Prefixspan.” Institut Teknologi Sepuluh Nopember, 2016.

[4] L. Lenny, “Rancang Bangun Fitur Rekomendasi Buku Menggunakan Algoritma PrefixSpan pada Sistem Peminjaman Buku Berbasis Web di Perpustakaan Universitas Ciputra,” *J. Ilm. Teknol. Inf. dan Multimed.*, vol. 3, no. 1, pp. 1–16, 2017.

[5] H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2006.

[6] D. S. Maylawati, H. Aulawi, and M. A. Ramdhani, “The concept of sequential pattern mining for text,” in *IOP Conference Series: Materials Science and Engineering*, 2018, doi: 10.1088/1757-899X/434/1/012042.

[7] J. Pei *et al.*, “Mining sequential patterns by pattern-growth: The prefixspan approach,” *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 11, pp. 1424–1440, 2004, doi: 10.1109/TKDE.2004.77.