PROPOSAL PROYEK

Movie Recommendation System Using Content Based

Disusun oleh:

1. 12S17004 – Fivin Sadesla Tambunan
2. 12S17026 – Mika Lestari Valentina Manurung
3. 12S17037 – Nita Sophia Winandi Sirait

PROGRAM STUDI SARJANA SISTEM INFORMASI

FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO

INSTITUT TEKNOLOGI DEL

20 November 2020

DAFTAR ISI

DAFTAR TABEL

Bab 1. *Business Understanding*

Tahap pertama dalam CRISP-DM (*Cross Industry Standard Process for Data Mining*) adalah *business understanding* atau dapat juga disebut sebagai tahap pemahaman penelitian. Pada tahap ini diperlukan pemahaman mengenai substansi dari kegiatan *data mining* yang akan dilakukan serta kebutuhan (*requirement*) dari perspektif bisnis [1]. Pemahaman bisnis mengacu pada sistem rekomendasi film. Pada tahap ini diperlukan pemahaman tentang latar belakang dan tujuan pada sistem rekomendasi.

1.1 *Determine Business Objectives*

Tahap menentukan tujuan bisnis dan menentukan faktor-faktor penting yang terlibat dalam penelitian yang direncanakan dan untuk memastikan bahwa penelitian tidak menghasilkan jawaban yang benar untuk pertanyaan yang salah. Tujuan bisnis bukan merupakan tujuan langsung penelitian, namun lebih sebagai tujuan jangka panjang dalam dunia nyata. Hasil *data mining* mendeskripsikan sebuah pengetahuan yang akan menjadi tolak ukur institusi terkait dalam membuat kebijakan di masa akan datang.

Semakin berkembangnya teknologi maka akan semakin banyaknya informasi yang tersedia. Informasi ini dapat diakses melalui pemanfaatan dari teknologi multifungsi yang dikaji agar lebih efisien dan optimal melalui internet. Saat ini dengan adanya teknologi, semua informasi yang dibutuhkan sudah tersedia di internet dengan berbagai versi. Pada era sekarang, banyak masyarakat yang membutuhkan hiburan untuk menghilangkan rasa penat dari kesibukan sehari-hari, hiburan tersebut dapat diperoleh melalui film. Menurut *British Film Institute* (BFI) film *box office* yang diproduksi terus meningkat setiap tahunnya mulai tahun 2009 hingga tahun 2015 (*British Film Institute,* 2016). Pada tahun 2009 terdapat 503 film yang diproduksi pada tahun tersebut dan sebanyak 759 film diproduksi pada tahun 2015 (*British Film Institute*, 2016). Berdasarkan hal tersebut maka dapat disimpulkan bahwa banyaknya film yang tersedia di internet. Oleh karena itu, dibutuhkan sebuah sistem yang dapat membantu memberikan informasi sesuai dengan keinginan pengguna. Sistem tersebut sering disebut dengan sistem rekomendasi.

Tujuan bisnis pada sistem rekomendasi film adalah membantu pengguna mencari pilihan film mereka di antara semua jenis film yang berbeda dan juga pengguna tidak perlu menghabiskan banyak waktu untuk mencari film favorit mereka [2]. Faktor yang mempengaruhi pembuatan sistem rekomendasi film adalah informasi yang didapat dari pengguna tersebut. *Item* tersebut

atau dalam hal ini film dapat disarankan berdasarkan *rating* film yang diberikan pengguna atau berdasarkan pengguna lain yang memiliki kebiasaan yang mirip.

1.2 *Access the Situation*

Tugas ini melibatkan pencarian fakta yang lebih rinci tentang semua sumber daya (*sources*) seperti *Hardware sources* (sumber daya perangkat keras), *data sources* (sumber daya data), dan *personel sources* (sumber daya personal)

- *Hardware sources* yang digunakan pada penelitian ini adalah laptop IdeaPad Lenovo 4 GB RAM, Processor Intel Core i5-7200U Dual Core 2,5 GHZ TurboBoost 3,1 GHZ, CD/DVD ROM Drive
- *Data sources* yang digunakan pada penelitian ini adalah dataset MovieLens yang diperoleh secara *online* melalui url http://grouplens.org/datasets/movielens/latest.
- *Personel sources* pada penelitian ini terdiri dari 3 orang mahasiswa yang berperan pada pengerjaan sistem rekomendasi film menggunakan *content based* mulai dari tahap *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment.*

1.3 *Determine Data Mining Goals*

Tahap mengubah pengetahuan pada domain bisnis menjadi sebuah definisi masalah *data mining* dan menentukan tujuan *data mining* (penelitian). Tujuan *data mining* atau tujuan penelitian ini adalah menggali pengetahuan (*discovering knowledge*) tentang pola (*pattern*) *item* berdasarkan riwayat pencarian pengguna sehingga dapat diketahui dan diprediksi *item* (film) yang mungkin diminati pengguna.

1.4 *Produce Project Plan*

Menjelaskan rencana yang ditujukan untuk mencapai tujuan data mining dan dengan demikian dapat mencapai tujuan bisnis, kemudian menentukan teknik dan tools yang akan dipergunakan.

Project plan :

Perencanaan proyek agar dapat mencapai tujuan data mining dan mencapai tujuan bisnis adalah sebagai berikut :

Tabel 1.1. Perencanaan Proyek

Tahap	Waktu	Kegiatan
Business Understanding	5 hari	Menentukan tujuan bisnis, menilai situasi, menentukan

Tahap	Waktu	Kegiatan
		tujuan *data mining*, dan menyusun rencana proyek.
Data Understanding	1 minggu	Mengumpulkan data, memahami data, melakukan eksplorasi data, dan mengidentifikasi kualitas data.
Data Preparation	2 minggu	Memilih dataset akhir dari data mentah, membersihkan data, membangun data, mengintegrasikan data, dan melakukan format data.
Modeling	4 minggu	Memilih teknik pemodelan dan beberapa parameter yang disesuaikan untuk mendapatkan nilai yang optimal
Evaluation	2 minggu	Melakukan evaluasi hasil, *review* proses, dan menentukan langkah selanjutnya.
Deployment	3 minggu	Menyusun rencana penerapan, menyusun rencana pemantauan (*monitoring*) dan pemelihataan (*maintenance*), membuat laporan akhir

Teknik/metode :

Metode yang akan digunakan pada penelitian ini ada *content based*. Metode *content based* memiliki deskripsi *item* yang didasarkan pada kecenderungan pengguna. Algoritma ini akan

mencoba untuk merekomendasikan item yang sama dengan apa yang disukai pengguna pada masa lampau melalui metode *cosine similarity.*

Cosine similarity berarti menghitung nilai kemiripan antar kalimat dengan melakukan fungsi pencarian berdasarkan setiap indeks yang dimasukkan. Hasil dari pencarian dan pengujian yang dilakukan metode ini, akan memberikan rekomendasi anime pilihan *user* dengan perhitungan yang akurat.

Tools :

Tools yang digunakan untuk penelitian ini adalah python

Bab 2. *Data Understanding*

Data Understanding atau pemahaman data adalah tahapan mengumpulkan data awal dan mempelajari data tersebut untuk bisa mengenal dan memahami apa saja yang bisa dilakukan pada data tersebut. Pemahaman data mengacu pada *dataset* MovieLens. Terdapat tahap memahami format data secara permukaan (format *form* dan *report*) dan secara lebih mendalam (bentuk fisik data).

2.1 *Collect Initial Data*

Pada tahap ini dilakukan pengumpulan data. Dilakukan pemilihan data relevan yang di dapat dari dataset yang dilakukan sebelum tahap penggalian informasi dalam *Knowledge Discovery and Data Mining* (KDD) dimulai. Dalam tahapan ini, akan dipilih data seperti apa saja yang dibutuhkan untuk diproses lebih lanjut. Dataset yang digunakan pada penelitian ini berasal dari MovieLens yang dikumpulkan dari periode 9 Januari 1995 sampai 26 September 2018 dan tersedia untuk umum di url http://grouplens.org/datasets/movielens/latest. Dataset yang digunakan dapat berubah seiring dengan waktu.

2.2 *Describe Data*

Tugas ini melibatkan deskripsi terhadap isi data. Data berasal dari lebih dari satu file dengan ekstensi .csv selanjutnya akan dilakukan proses integrasi data atau penggabungan data. Data tersedia dalam 6 *file* yaitu genome-score.csv, genome-tags.csv, links.csv, movies.csv, rating.csv dan tags.csv. Data berisi 27753444 *ratings* dan 1108997 *tag applications* pada 58098 *movies*. Data ini dibuat oleh 283228 pengguna. Pengguna dipilih secara acak dan setidaknya memberikan *rating* untuk 1 *movie*. Setiap pengguna diwakili oleh sebuah id (userId) dan *movie* diwakili oleh sebuah id (movieId).

Pada penelitian ini hanya menggunakan file movies,csv dan rating.csv untuk membangun sistem rekomendasi.

2.3 *Explore Data*

Tugas ini membahas mengenai *data mining* dengan menggunakan teknik kueri, visualisasi, dan pelaporan. Eksplorasi data juga termasuk distribusi atribut kunci (misalnya, atribut target dari tugas prediksi) hubungan antara pasangan atau sejumlah kecil atribut, hasil agregasi sederhana, properti sub-populasi yang signifikan, dan analisis statistik sederhana. Analisis ini dapat secara langsung membahas tujuan *data mining*; mereka juga dapat berkontribusi atau menyempurnakan deskripsi data dan laporan kualitas, dan dimasukkan ke dalam transformasi dan langkah-langkah persiapan data lainnya yang diperlukan untuk analisis lebih lanjut.

2.4 *Verify Data Quality*

Tahap ini berisi evaluasi kualitas data dan kelengkapan data. Nilai-nilai yang hilang sering terjadi, terutama jika data yang dikumpulkan di jangka waktu yang lama. Memeriksa atribut yang hilang atau kosong. Menilai apakah semua nilai masuk akal, ejaan nilai-nilai, dan apakah atribut dengan nilai yang berbeda memiliki arti yang sama. *Data cleaning* dibutuhkan untuk menghilangkan noise dan data yang tidak konsisten atau data tidak relevan. Pembersihan data akan mempengaruhi performansi dari sistem *data mining* karena data yang ditangani akan berkurang jumlah dan kompleksitasnya. Hasil penelusuran menemukan :

1. *Dataset* di tulis sebagai *file* dengan nilai yang dipisahkan koma dengan satu baris header.
2. File dikodekan sebagai UTF-8

Daftar Pustaka

[1] p. Chapman, . J. Clinton , . R. Kerber , T. Khabaza , T. Reinartz, C. Shearer dan R. Wirth, "CRISP-DM 1.0," 2000.

[2] S. Agrawal dan P. Jain, "An Improved Approach for Movie Recommendation System," *International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, p. 336, 2017.