Fatwa Sifaun Nahar

2008101073

Riset Informatika A

METODE PENELITIAN

**Judul**: Perbandingan Penggunaan Jarak Manhattan, Jarak Euclid, dan Jarak Minkowski

Pada tahap awal penelitian kami menggunakan dataset yang telah ada di internet yaitu Movielens 100k, dataset tersebut berisi beberapa review film dari user yang berisi sekitar 100000 rating film. Selanjutnya pre-processing akan dilakukan pada tahap A dan pada tahap B , User-Based filter akan diimplementasikan untuk membersihkan dataset.

1. Pengambilan Data dan Pre-Processing

Pada bagian ini, dilakukan langkah pengambilan data dari dataset MovieLens yang diberi label ml-latest-small yang mencakup 100.836 penilaian dari 610 pengguna terhadap 9.742 film. Untuk menerapkan teknik sistem rekomendasi, atribut yang tidak digunakan, seperti timestamp rating, telah dihilangkan.

1. Data Processing dan Evaluasi

Nearest Neighbor akan digunakan dalam teknik filtering kolaboratif user-based, di mana tiga persamaan untuk menghitung jarak kesamaan akan dibandingkan yaitu: Euclidean, Manhattan, dan Minkowski. Untuk Minkowski sendiri parameter p dibatasi pada p = 1 dan p = 2.

Dari perhitungan bobot kesamaan pengguna, proses inverse perkalian dilakukan sehingga semakin kecil jarak antara pengguna (nilai kesamaan yang lebih kecil) dapat menghasilkan nilai bobot tertinggi, sebaliknya, nilai jarak yang lebih jauh akan menghasilkan nilai bobot yang kecil. Berikut adalah persamaannya:

Namun, fungsi diatas aka nada kemungkinan eror dimana ketika similarity mempunyai nilai 0 sehingga akan menghasilkan bobot similarity tak terbatas/infinite. Karena itulah normalisasi data diperlukan dimana bobot user akan diubah menjadi seperti rumus dibawah:

Sehingga bobot dengan nilai tak terbatas akan bisa dihindari dengan normalisasi diatas. Nilai maksimum akan menjadi 1 apabila similarity bernilai 0, nilai minimum akan mendekati 0.

Setelah membuat pembobotan user, selanjutnya yaitu melakukan pemilihan user dengan nilai k yang signifikan. Nilai k yang digunakan antara lain 3, 5, 7, 10, 20, dan 30. User yang terpilih akan dijadikan model untuk memprediksi rating film. Predeksi rating juga bisa diuraikan dalam rumus dibawah ini:

Selanjutnya untuk mengevaluasi performa dari masing masing persamaan jarak akan digunakan 5-fold cross-validation dengan Root Mean Square Eeviation (RMSE) yang diacu sebagai metriks, dimana banyak RMSE banyak digunakan untuk model prediksi. Semakin rendah hasi RMSE maka semakin bagus pula akurasinya. Untuk flowchartnya sendiri ada dibawah ini:

