# Aplikasi dan analisis clustering pada data akademik

Andina Budiarti, author



---

Abstrak

Seiring dengan berkembangnya teknologi basis data dan volume data yang terkumpul di dalamnya, muncul kebutuhan untuk mendapatkan informasi yang lebih dalam, yaitu dengan data mining. Penelitian ini bertujuan untuk menemukan informasi baru yang belum diketahui sebelumnya dari domain data yang tersedia (data MTI) dan mempelajari berbagai algoritma clustering yang telah ada serta menemukan algoritma yang paling cocok digunakan untuk domain tersebut. Penelitian tugas akhir ini terbatas pada analisis data dan algoritma yang sudah tersedia serta analisis hasil yang didapatkan pada masing-masing percobaan. Metode penelitian mencakup studi literatur, analisis data dan algoritma, percobaan, serta analisis hasil percobaan. Dalam melakukan data mining, digunakan panduan (CRISP-DM) [OY+07] yang terdiri dari tahapan Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation dan Deployment. Namun, tahap deployment tidak dilakukan karena berada di luar lingkup penelitian. Penyiapan dan pemurnian data dilakukan dengan standarisasi penamaan, pengubahan bentuk dan diskretisasi. Untuk memproses data dengan nilai atribut yang tidak lengkap diserahkan kepada mekanisme masing-masing algoritma. Untuk keperluan penelitian ini, 3 implementasi clustering pada WEKA akan dimanfaatkan, yaitu K-Means, EM dan COBWEB. Implementasi Apriori juga dimanfaatkan untuk menemukan association rules. Untuk mengatasi permasalahan yang mungkin timbul akibat high dimensionality dari domain data, dilakukan dekomposisi secara iteratif (5 iterasi) dengan mengambil subset dari seluruh atribut. Pada setiap percobaan, hasil clustering akan divisualisasikan dalam gambar 2-dimensi dengan bantuan program Applet Java yang dibuat oleh penulis. Visualisasi ini terbatas untuk kebutuhan pengamatan saja karena tidak menggambarkan kemampatan yang sebenarnya dari masingmasing cluster yang berdimensi tinggi. Informasi hasil dari percobaan data mining yang paling menonjol adalah mengenai kaitan antara 'Jalur lulus' dan 'Lama studi' di mana 'Proyek akhir' memungkinkan mahasiswa untuk dapat lulus lebih cepat. Tidak ada hubungan yang cukup berarti antara data latar belakang dengan IPK, menandakan siapa saja dapat berprestasi di program studi ini. Sementara itu, 'Sektor kerja' juga menjadi faktor yang cukup mempengaruhi pengelompokkan data. Algoritma yang menentukan sendiri banyak clusters yang dihasilkan lebih cocok untuk dipakai. Perubahan volume data sangat berpengaruh pada hasil clustering. Oleh sebab itu pula, algoritma tanpa input banyak cluster seperti K-Means kurang cocok dipakai sampai volume data mencapai suatu titik yang stabil. Partitioning algorithm cocok digunakan jika sudah ada dugaan atau perkiraan yang didukung hasil data mining sebelumnya mengenai banyak cluster yang dihasilkan dan seperti apa struktur clusters tersebut. Untuk kasus yang sudah diketahui sebelumnya mengenai struktur kelompok dalam data, kemungkinan clustering dengan algoritma yang memerlukan input banyak cluster lebih 'baik' daripada algoritma yang menentukan sendiri banyak cluster yang dihasilkan sehingga perlu diinterpretasi lebih jauh lagi hasilnya. COBWEB yang mewakili hierarchical algorithm menunjukkan hasil clustering yang lebih alamiah dan mudah untuk diinterpretasikan jika dibandingkan hasil dari algoritma EM maupun K-Means. Akan tetapi, tidak seperti partitional algorithm yang dari cluster yang dihasilkan dapat ditarik kesimpulan yang baru, hierarchical algorithm dalam kasus ini hanya mengelompokkan data yang 'mirip' tanpa bisa digali

informasi dari masing-masing cluster yang dihasilkan. Untuk jumlah data yang digunakan dalam percobaan kali ini, algoritma EM, K-Means yang diimplementasi WEKA dapat mengeluarkan hasil dalam waktu yang relatif cepat (di bawah 30 detik). Lain halnya dengan COBWEB yang lebih memakan waktu, misalnya pada iterasi kedua algoritma ini memerlukan 12 menit.