Proses Data Mining dalam Sistem Pembelajaran Berbantuan Komputer

Mewati Ayub
Jurusan Teknik Informatika,
Fakultas Teknologi Informasi
Universitas Kristen Maranatha, Bandung
Email : mewati.ayub@eng.maranatha.edu

Abstract

Web-based educational systems and intelligent tutoring systems collect large amounts of students data, from web logs to student models. Data mining applications on those data can help discovering relevant knowledge for improving computer aided learning system. Using the knowledge, teacher can understand more about how students learn by studying a group of students in order to enhance teaching and learning. In this paper, data mining process will be separated into data collection, data transformation, and data analysis. Association rules, classification, and clustering are data mining algorithms that explored in data analysis for computer aided learning systems.

Keywords : data mining, computer aided learning system, knowledge

1. Pendahuluan

Ketersediaan data yang berlimpah yang dihasilkan dari penggunaan teknologi informasi di hampir semua bidang kehidupan, menimbulkan kebutuhan untuk dapat memanfaatkan informasi dan pengetahuan yang terkandung di dalam limpahan data tersebut, yang kemudian melahirkan *data mining*. *Data mining* merupakan proses untuk menemukan pengetahuan (*knowledge discovery*) yang ditambang dari sekumpulan data yang volumenya sangat besar. Aplikasi *data mining* pada pengelolaan bisnis, pengendalian produksi, dan analisa pasar misalnya, memungkinkan diperolehnya pola dan hubungan yang dapat dimanfaatkan untuk peningkatan penjualan, atau pengelolaan sumber daya dengan lebih baik.

Dunia pendidikan memiliki data yang berlimpah dan berkesinambungan mengenai siswa yang dibina dan alumni yang dihasilkannya. Hal ini membuka peluang diterapkannya *data mining* untuk pengelolaan pendidikan yang lebih baik [Jing, 2004] dan *data mining* dalam pelaksanaan pembelajaran berbantuan komputer yang lebih efektif [Merceron, 2005].

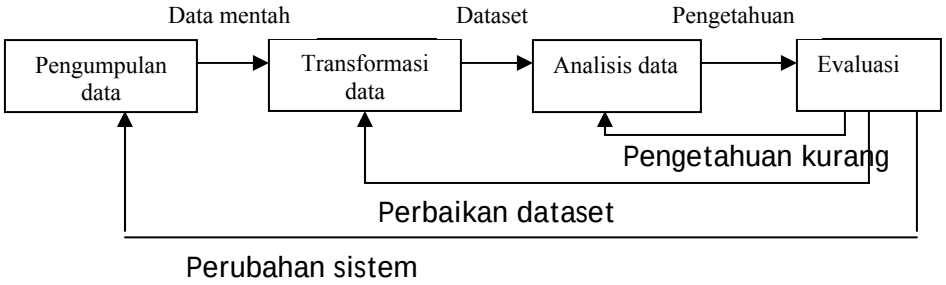
Sistem pembelajaran berbantuan komputer *(computer aided learning system)* dapat diimplementasikan sebagai sistem tutorial berbasis web (*web-based tutoring tool*) [Merceron, 2005] atau sistem tutorial cerdas (*intelligent tutoring system*) [Nilakant, 2004]. Di dalam sistem tutorial berbasis web maupun sistem tutorial cerdas, setiap interaksi siswa dengan sistem akan dicatat dalam suatu basis data dalam bentuk web log atau model siswa (*student model*). Setelah sistem tersebut digunakan dalam proses pembelajaran selama jangka waktu tertentu, maka akan terkumpul sejumlah besar data. Kumpulan data tersebut dapat diproses lebih lanjut dengan *data mining* untuk memperoleh pola baru yang dapat digunakan untuk meningkatkan efektifitas dalam proses pembelajaran.

Makalah ini akan membahas bagaimana *data mining* dapat dimanfaatkan untuk meningkatkan efektifitas dalam proses pembelajaran berbantuan komputer dari sudut pedagogi.

2. Data Mining

Data mining mengacu pada proses untuk menambang (*mining*) pengetahuan dari sekumpulan data yang sangat besar [Jiawei, 2001]. Sebenarnya *data mining* merupakan suatu langkah dalam *knowlegde discovery in databases* (KDD). *Knowledge discovery* sebagai suatu proses terdiri atas pembersihan data (*data cleaning*), integrasi data (*data integration*), pemilihan data (*data selection*), transformasi data (*data transformation*), *data mining*, evaluasi pola (*pattern evaluation*) dan penyajian pengetahuan (*knowledge presentation*).

Kerangka proses *data mining* yang akan dibahas tersusun atas tiga tahapan, yaitu pengumpulan data (*data collection*), transformasi data (*data transformation*), dan analisis data (*data analysis*) [Nilakant, 2004]. Proses tersebut diawali dengan *preprocessing* yang terdiri atas pengumpulan data untuk menghasilkan data mentah (*raw data*) yang dibutuhkan oleh *data mining*, yang kemudian dilanjutkan dengan transformasi data untuk mengubah data mentah menjadi format yang dapat diproses oleh kakas *data mining*, misalnya melalui filtrasi atau agregasi. Hasil transformasi data akan digunakan oleh analisis data untuk membangkitkan pengetahuan dengan menggunakan teknik seperti analisis statistik, *machine learning*, dan visualisasi informasi.



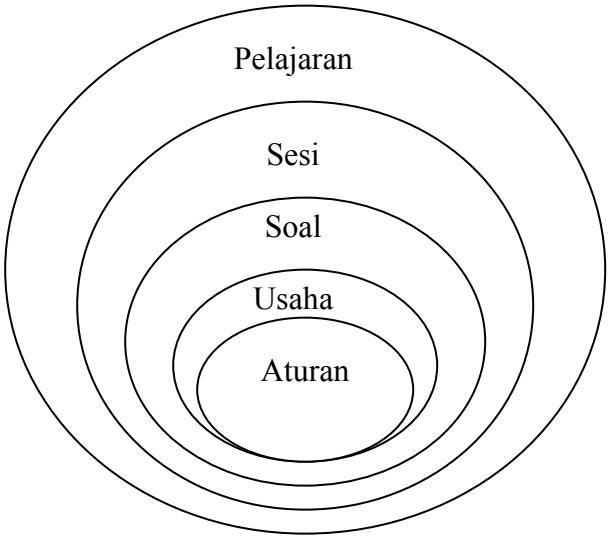
Gambar 1. Aliran informasi dalam data mining

Pada gambar 1 ditunjukkan diagram yang menggambarkan aliran informasi dalam proses *data mining* yang diadaptasi dari [Nilakant, 2004]. Proses *data mining* pada gambar tersebut ditunjukkan sebagai proses yang iteratif. Hasil evaluasi pengetahuan yang dihasilkan *data mining* dapat menimbulkan kebutuhan pengetahuan yang lebih lengkap, perbaikan kumpulan data (dataset) atau perubahan pada sistem.

3. Data mining dalam pembelajaran

Untuk menentukan variabel apa saja yang perlu dicatat dari interaksi siswa dengan sistem, perlu dikembangkan suatu model interaksi siswa-sistem. Gambar 2 menunjukkan analisis interaksi tersebut pada beberapa lapisan (layer) yang berbeda [Nilakant, 2004].

Apabila evaluasi sistem tutorial diterapkan pada lapisan terluar, maka akan dilakukan *pre-test* sebelum pelajaran dimulai dan *post-test* setelah pelajaran selesai diberikan. Perbedaan *pre-test* dan *post-test* akan menunjukkan perubahan kinerja setiap siswa dengan granularitas pada tingkat pelajaran. Jika diperlukan analisis yang lebih terinci, maka eksplorasi hasil belajar (*learning outcome*) harus dilakukan pada lapisan yang lebih dalam. Sebagai contoh, pembelajaran dalam suatu pelajaran tersusun atas beberapa sesi. Setiap sesi diawali dan diakhiri dengan tes untuk mengevaluasi perubahan kinerja setiap siswa per sesi. Selama suatu sesi, siswa akan mencoba mengerjakan sejumlah soal. Untuk mengerjakan suatu soal, siswa mungkin akan mencoba beberapa kali sebelum mendapatkan solusi yang paling tepat. Dari interaksi tersebut, beberapa informasi dapat dicatat oleh sistem, seperti misalnya apakah siswa dapat menjawab setiap soal dengan benar, berapa kali siswa mencoba sebelum akhirnya memberikan jawaban yang tepat. Pada tingkat granularitas yang paling baik, setiap usaha siswa menjawab soal dibedakan atas melanggar atau memenuhi sejumlah aturan. Informasi mengenai pelanggaran atau pemenuhan tersebut dapat dicatat, dan akan menghasilkan representasi status kognitif siswa, yang dikenal sebagai model siswa.



Gambar 2. Model interaksi siswa-sistem (lapisan granularitas)

Informasi yang dihasilkan pada berbagai lapisan tersebut di atas dapat saling melengkapi, sehingga analisis data dapat menunjukkan hubungan antara data dari berbagai lapisan. Sebagai contoh, untuk setiap pelanggaran aturan yang dicatat, informasi mengenai siswa yang melakukan pelanggaran, pelajaran yang sedang diikuti, soal yang sedang dipelajari, serta usaha yang menyebabkan pelanggaran harus disimpan juga. Pada tabel 1 ditunjukkan ringkasan data yang dapat disimpan untuk setiap usaha yang dilakukan siswa dalam menjawab soal.

Tabel 1. Taksonomi variabel dari usaha siswa menjawab soal

Kegiatan	Variabel	Keterangan
persiapan	• umpan balik yang tersedia • banyaknya soal • banyaknya usaha • tingkat kesulitan soal • konteks soal	informasi yang dipakai siswa sebelum mencoba menjawab suatu soal
pelaksanaan latihan soal	waktu yang diperlukan untuk menjawab soal	informasi mengenai usaha siswa menjawab suatu soal
evaluasi	• aturan (relevan, dipenuhi, dilanggar) • tingkat umpan balik yang diminta • permintaan melihat solusi	informasi yang berhubungan dengan hasil (outcome) dari suatu usaha.

Tahap pengumpulan data akan menyediakan data dalam volume yang cukup besar, namun analisis data tidak dapat langsung dilakukan terhadap kumpulan data tersebut, karena harus dilakukan transformasi terhadap data sehingga analisis siap dilakukan.

Data mentah yang dihasilkan dari pengumpulan data, biasanya tersimpan dalam bentuk beberapa tabel basis data. Karena analisis data umumnya dilakukan terhadap suatu tabel tunggal, maka perlu dilakukan penggabungan (join) beberapa tabel yang relevan. Hasilnya adalah suatu struktur yang disebut dengan dataset, seperti tampak pada gambar 3 [Nilakant, 2004]. Dataset dapat dikelompokkan secara vertikal sebagai kumpulan atribut dan secara horisontal sebagai kumpulan instans. Setiap atribut mempunyai tipe data, yang dapat berupa numerik, teks, atau bentuk lainnya. Jika domain nilai suatu atribut berhingga, maka disebut atribut nominal. Suatu instans adalah data yang dihasilkan dari suatu kejadian di dunia nyata, yang dicatat dalam beberapa atribut.

	atribut-1	atribut-2	. . .	atribut-n
instans-1	$x_{1,1}$	$x_{1,2}$	. . .	$x_{1,n}$
instans-2	$x_{2,1}$	$x_{2,2}$	. . .	$x_{2,n}$
. . .	. . .	. . .	. . .	. . .
instans-m	$x_{m,1}$	$x_{m,2}$	. . .	$x_{m,n}$

Gambar 3. Format Dataset

Transformasi dataset dapat dilakukan dalam beberapa cara, antara lain filtrasi dataset dan konversi atribut [Nilakant, 2004][Jiawei, 2001]. Filtrasi dataset dilakukan dengan mengurangi ukuran dataset, yaitu dengan membuang beberapa informasi yang tidak relevan. Sebagai contoh, dari analisis terhadap data mentah ditemukan bahwa beberapa soal dalam basis data cenderung menimbulkan pelanggaran terhadap aturan tertentu. Untuk eksplorasi penyebabnya, analisis harus dibatasi hanya terhadap kumpulan soal tersebut. Dengan menyaring informasi tersebut, proses analisis akan memberikan hasil yang lebih dapat diandalkan (*reliable*). Filtrasi dilakukan terhadap salinan data, sehingga data asli tidak mengalami perubahan data.

Cara berikutnya untuk transformasi data adalah konversi atribut, yaitu bekerja pada nilai atribut di setiap instans dari dataset. Tujuan dari konversi atribut adalah mengubah atribut bernilai kontinu (tidak berhingga) menjadi atribut dengan nilai nominal (berhingga), karena sebagian cara analisis dengan *machine learning* tidak dapat berfungsi pada atribut yang bernilai kontinu.

Terdapat dua cara untuk melakukan diskritisasi nilai atribut. Cara pertama dengan melakukan penelusuran (*scanning)* seluruh dataset untuk semua nilai kontinu yang muncul, kemudian menggunakan nilai tersebut sebagai domain dari atribut nominal. Teknik ini membuat domain nilai menjadi himpunan tertutup dari nilai yang mungkin muncul, sehingga dataset menjadi '*sparse*'.

Terdapat teknik lainnya, yaitu *'binning'*, yang mendefinisikan kumpulan kelas nominal untuk setiap atribut, kemudian menetapkan setiap nilai atribut ke dalam salah satu kelas. Misalnya, jika domain atribut numerik mempunyai nilai dari 0 sampai dengan 100, domain tersebut dapat dibagi menjadi empat bin (0..24, 25..49, 50..74, 75..100). Setiap nilai atribut akan dikonversi menjadi atribut nominal yang berkorespondensi dengan salah satu bin.

Terdapat tiga cara untuk mendefinisikan interval nilai, yaitu *equal-width*, *equal-frequency*, dan *customised*. *Equal-width* akan membagi interval nilai atribut menjadi n interval yang lebarnya sama. *Equal-frequency* menghitung interval dari setiap kelas sehingga setiap kelas yang dialokasikan akan mempunyai frekuensi instans dataset yang hampir sama.

4. Penerapan Teknik Data Mining

Proses analisis data dengan menerapkan teknik *data mining* dapat dilakukan melalui analisis statistik atau dengan pendekatan *machine learning*. Analisis data pembelajaran dengan pendekatan *machine learning* akan menggunakan tiga teknik, yaitu *association rules*, *clustering,* dan *classification* [Nilakant, 2004][Merceron, 2005].

Algoritma *association rule* (AR) digunakan untuk menemukan hubungan antar nilai tertentu dari atribut nominal dalam suatu dataset. Aturan yang dihasilkan dapat ditulis dalam bentuk "if-then" dengan mempertimbangkan besaran *support* dan *confidence* untuk menilai reliabilitas aturan. Bentuk umum aturan dalam *association rule* adalah :

$$(X = x_i) \rightarrow (Y = y_i) \quad [sup, conf]$$

dengan $X = \{ x_1, x_2, \ldots, x_n \}$, $Y = \{ y_1, y_2, \ldots, y_m \}$, sup = probabilitas bahwa suatu instans dalam dataset mengandung $X \cup Y$, conf = probabilitas kondisional bahwa instans yang mengandung X juga mengandung Y.

Pada gambar 4 ditunjukkan contoh atribut dataset yang dapat diturunkan dari Tabel 1 yang akan digunakan sebagai acuan untuk pembahasan dalam algoritma AR, classification, dan clustering berikut ini. Dataset tersebut dapat merupakan gabungan dari beberapa data yang diperoleh saat siswa berinteraksi dengan sistem pembelajaran.

No	Atribut	Keterangan
1	IdSiswa	Identitas siswa
2	NoSoal	Nomor soal yang dikerjakan
3	JenisSalah	Jenis kesalahan
4	NoAturan	Nomor aturan/konsep yang dipakai dalam soal
5	JmlCoba	Banyaknya usaha siswa mencoba menjawab soal
6	Tingkat	Tingkat pencapaian dalam pengerjaan soal
7	Nilai	Nilai yang diperoleh

Gambar 4. Contoh Atribut Dataset

Dalam *mining* data pembelajaran, algoritma AR dapat dimanfaatkan untuk menemukan kesalahan yang sering terjadi pada saat siswa mengerjakan latihan soal. Sebagai contoh, dari dataset pada gambar 4, diperoleh kumpulan instans mengenai siswa yang melakukan kesalahan dengan frekuensi tertentu. Diasumsikan kumpulan instans tersebut memenuhi kondisi jika siswa melakukan kesalahan A dan kesalahan B, maka mereka juga melakukan kesalahan C, misalnya dengan support 30% dan confidence 60%, akan ditulis sebagai :

A and B → C [30%,60%]

Aturan tersebut dapat dibaca sebagai berikut : dari 30% siswa yang melakukan kesalahan A dan kesalahan B (dari seluruh siswa yang mengerjakan latihan soal), 60% diantaranya melakukan kesalahan C.

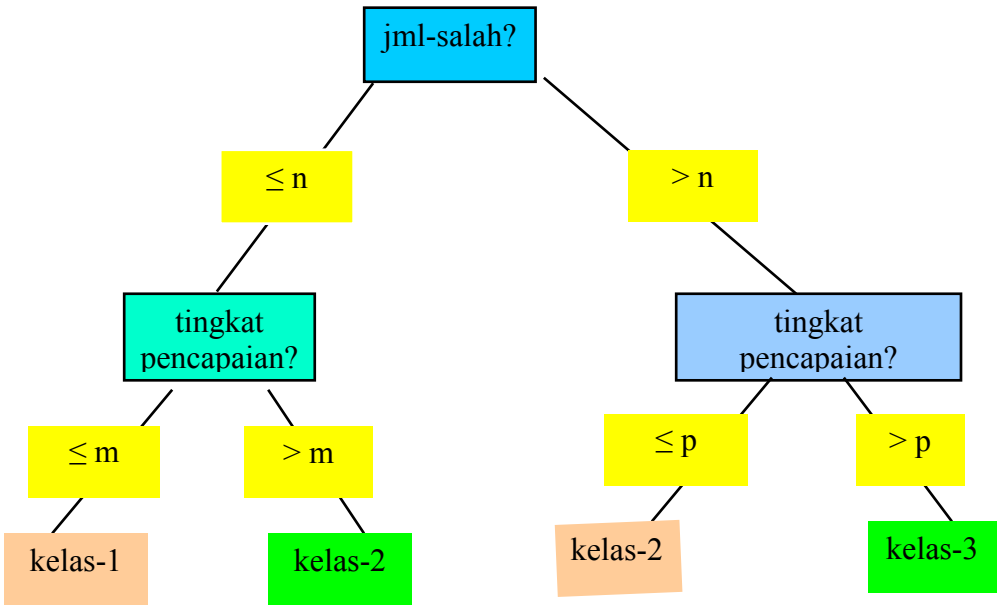
Algoritma AR juga dapat menyatakan hubungan antara beberapa atribut yang berbeda, misalnya kesalahan A pada konsep X menimbulkan kesalahan B pada konsep Y, yang ditulis sebagai

A and X → B and Y

Teknik *classification* bekerja dengan mengelompokkan data berdasarkan data training dan nilai atribut klasifikasi. Aturan pengelompokan tersebut akan digunakan untuk klasifikasi data baru ke dalam kelompok yang ada. Classification dapat direpresentasikan dalam bentuk pohon keputusan (*decision tree*). Setiap node dalam pohon keputusan menyatakan suatu tes terhadap atribut dataset, sedangkan setiap cabang menyatakan hasil dari tes tersebut. Pohon keputusan yang terbentuk dapat diterjemahkan menjadi sekumpulan aturan dalam bentuk IF condition THEN outcome.

Perbedaan utama antara aturan hasil algoritma AR dengan aturan hasil *classification* adalah *classification* hanya membuat model untuk satu atribut, yaitu atribut kelas. Pada algoritma AR, bagian konsekuen aturan (bagian kanan aturan) dapat mengandung lebih dari satu atribut, sedangkan pada *classification* hanya mengandung nilai atribut dari atribut kelas. Hal ini dapat digunakan untuk analisis secara *top-down*, yaitu mulai dengan algoritma AR untuk memperoleh hubungan antara beberapa atribut, kemudian analisis dipersempit pada atribut tertentu dengan menggunakan *classification*.

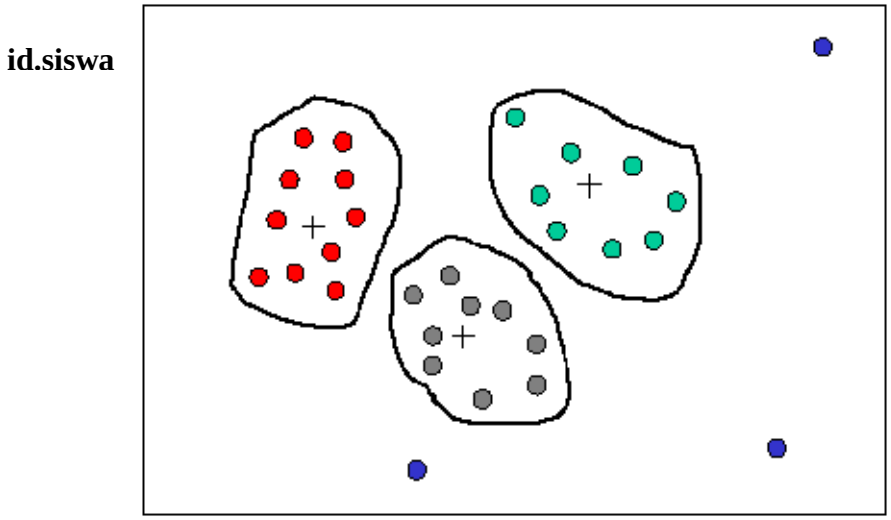
Dengan menggunakan dataset hasil belajar siswa seperti pada gambar 4, sebagai data training pada suatu tahun, dapat dibuat pohon keputusan untuk memperkirakan distribusi nilai siswa pada tahun berikutnya. Gambar 5 menunjukkan contoh pohon keputusan yang dihasilkan dari teknik *classification*. Pada gambar 5 terlihat klasifikasi siswa yang didasarkan pada jumlah kesalahan yang dilakukan siswa dan tingkat pencapaian dalam pengerjaan soal. Pemilihan atribut yang digunakan dalam pohon keputusan ditentukan secara heuristik dengan *information gain* [Jiawei, 2001].



Gambar 5. Contoh Pohon Keputusan

Teknik *clustering* bekerja dengan mencari kemiripan di antara objek dengan memperhatikan karakteristik objek, sekelompok objek yang mirip akan termasuk ke dalam satu *cluster*. Teknik yang dapat digunakan untuk melakukan *clustering* antara lain algoritma *k-means* atau algoritma *k-medoids* [Jiawei, 2001]. *Clustering* dapat diterapkan untuk mengenali karakteristik siswa yang mengalami kesulitan belajar. Misalnya kita ingin menganalisis siswa yang sudah mencoba mengerjakan latihan, namun tidak dapat menyelesaikannya sampai tuntas. Dalam hal ini, data yang dapat

digunakan adalah jumlah kesalahan yang dilakukan seorang siswa dalam mengerjakan suatu soal. Dengan demikian, siswa-siswa dengan frekuensi dan jenis kesalahan yang hampir sama (mirip), dapat dikelompokkan dalam satu cluster yang sama. Sebagai contoh, bila n < m, cluster 1 adalah kelompok siswa yang frekuensi kesalahannya lebih kecil dari n, cluster 2 adalah kelompok siswa dengan frekuensi kesalahan di antara n sampai dengan m, dan cluster 3 adalah kelompok siswa dengan frekuensi kesalahan lebih besar dari m. Contoh visualisasi clustering dapat ditunjukkan seperti pada gambar 6. Pada gambar tersebut terdapat tiga *cluster* dan beberapa *outlier*.



Gambar 6. Contoh Clustering

5. Kesimpulan

Penerapan data mining dalam sistem pembelajaran berbantuan komputer diawali dengan pengumpulan data, yang dilanjutkan dengan transformasi data, dan diakhiri dengan analisis data. Pada pengumpulan data, harus didefinisikan suatu model interaksi siswa-sistem untuk menetapkan data yang harus dicatat dari suatu proses pembelajaran. Model interaksi siswa-sistem tersebut dapat tersusun atas beberapa lapisan untuk memungkinkan penangkapan data pada tingkat granularitas yang berbeda. Proses transformasi data mengubah data mentah menjadi dataset yang siap dianalisis. Transformasi dapat dilakukan pada instans dataset melalui proses filtrasi, maupun pada atribut dari dataset melalui filtrasi ataupun konversi. Analisis data hasil pembelajaran dapat dilakukan dengan menerapkan teknik algoritma *association rules, classification*, dan *clustering* untuk menghasilkan pengetahuan yang dapat membantu guru

dalam mengelola kelasnya dengan memahami cara belajar siswa, dan memberikan umpan balik proaktif kepada siswanya.

Daftar Pustaka

Jiawei, H., Kamber, M. (2001). Data Mining Concepts and Techniques, Morgan Kaufmann Publishers.

Jing, L.(2004). Data Mining Applications in Higher Education, www.spss.com/events/e_id_1471/Data Mining in Higher Education.pdf, diakses tanggal 7 Februari 2007.

Nilakant, K. (2004). Application of Data Mining in Constraint Based Intelligent Tutoring System, www.cosc.canterbury.ac.nz/research/reports/HonsReps/2004/hons_0408.pdf, diakses tanggal 28 Februari 2007.

Merceron, A., Yacef, K. (2005). Educational Data Mining : a Case Study, http://www.it.usyd.edu.au/~kalina/publis/merceron_yacef_aied05.pdf, diakses tanggal 7 Februari 2007.

Merceron, A., Yacef, K. (2005). TADA-Ed for Educational Data Mining, imej.wfu.edu/articles/2005/1/03/printver.asp, diakses tanggal 1 November 2006.