J M S K
Jurnal Matematika, Statistika, & Komputasi
http://journal.unhas.ac.id/index.php/jmsk
Vol. 3, No.1, 1-8, Juli 2006

ISBN:1858-1382

Penggunaan Data Mining Saat Ini Dan Tantangannya Di Masa Depan

Sri Astuti Thamrin*

Abstract

This paper begins by describing the main activities involved in the data mining process and highlight the two major styles of data mining: *supervised* and *unsupervised*. It then describes two "hot" areas where data mining applications are being used successfully business database systems and the Internet. Finally, it concludes by examining the challenges and research issue data mining will face in the future.

business database systems, data mining, internet, supervised, unsupervised.

1. Pendahuluan

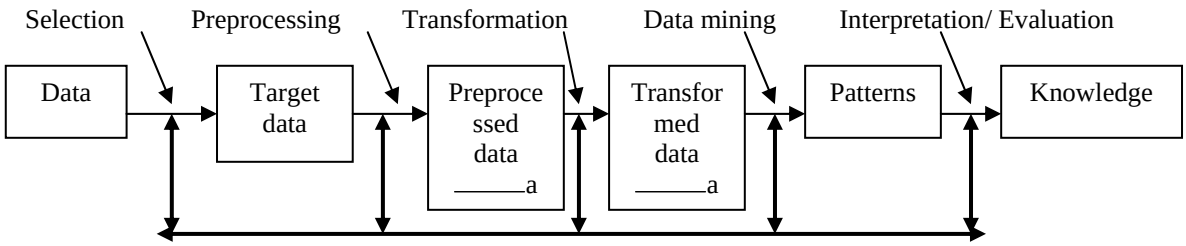
Data mining merujuk kepada proses non-trivial dalam menemukan pola data yang berguna dan menarik (Fayyad, Piatetsky & Shapiro, 1996). Jumlah data yang besar dan terus menerus berkembang secara konstan akan dikumpulkan oleh sistem informasi digital dalam aktivitas sehari-hari seperti menempatkan sebuah panggilan telepon, membeli sesuatu dari toko grosir, atau mengunjungi seorang dokter (Fayyad dkk, 1996). Pada masa yang lalu, analisis data mining harus dilakukan secara manual – sebuah cara yang sangat tidak praktis tentunya. Sekarang ini dengan kekuatan sebuah komputer yang cukup modern, data yang sama dapat diinterpretasi dan dijelaskan dalam kurun waktu yang sangat efisien dengan menggunakan sejumlah teknik data mining yang bervariasi dan terotomatisasi.

2. Proses Data Mining

Proses data mining merujuk kepada semua aktivitas yang biasanya muncul dalam proses data mining; penggalian data (mencari dan menemukan pola-pola data) hanya satu bagian dari proses tersebut. Walaupun tidak ditemukan dalam literatur sebuah gambaran proses yang sifatnya standar, namun pada umumnya literatur tersebut membagi proses data mining kedalam serangkaian aktivitas yang lebih kecil lingkupnya. Kita akan mencoba melihat proses data mining secara lebih detail sebagaimana telah digambarkan oleh Fayyad dkk (1996). Gambar 1 merupakan gambaran dari proses tersebut.

Langkah pertama dari proses data mining mencakup pembelajaran dan pemahaman akan domain atau wilayah penerapan. Sebuah kumpulan data yang telah ditetapkan untuk

* *Staf Pengajar pada Jurusan Matematika FMIPA Universitas Hasanuddin Makassar*

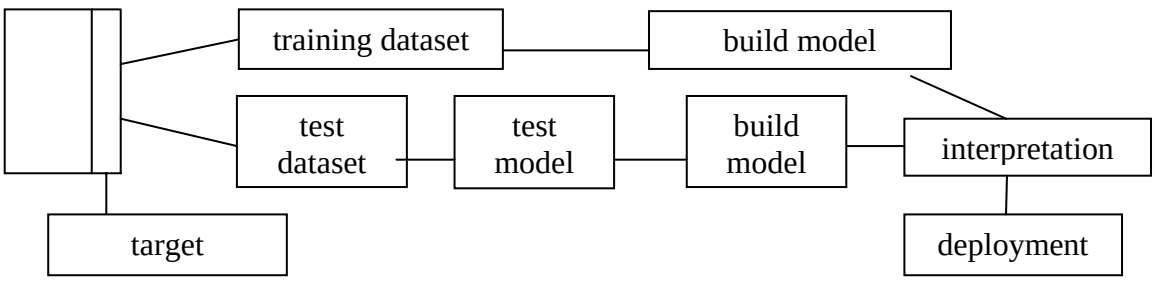


Gambar 1. Alur Kerja Proses Data Mining (Fayyad dkk, 1996)

diolah dalam aktivitas data mining haruslah terlebih dahulu diseleksi. Data tersebut kemudian harus "dibersihkan" dan diubah kedalam bentuk yang lebih layak bila memang diperlukan. Aktivitas–aktivitas sebelum data diproses antara lain memindahkan data pencilan, mengatasi data hilang dan issu–issu DBMS. Data kemudian direduksi dan diproyeksi untuk menemukan bagian–bagian yang berguna untuk menggambarkan data dengan sejumlah variabel-variabel yang telah direduksi dengan pertimbangan-pertimbangan tertentu. Tujuan dari data mining kemudian haruslah diputuskan (contohnya untuk klarifikasi atau kesimpulan) untuk memudahkan pemilihan data mining yang layak dalam mencapai tujuan yang telah ditetapkan. Penambangan data *(data mining)* yang sebenarnya (mencari pola dalam data tersebut) selanjutnya dapat dilakukan, yang kemudian hasilnya diserahkan kepada pemakai dalam bentuk yang cocok dan dapat dimengerti. Pemakai kemudian dapat melakukan interpretasi data dan memutuskan apakah pola-pola yang ditemukan itu bermanfaat dan atau menarik. Sekali kesimpulan dapat dibuat, pemakai kemudian dapat menggunakan pengetahuan yang telah ditemukan dan menjalinnya kedalam sistem untuk mempertinggi kinerja atau hanya menjadikannya dokumen sebagai bahan rujukan atau verifikasi dimasa datang. Selanjutnya perlu dikemukakan bahwa paling sedikit terdapat dua jenis teknik data mining yang biasa digunakan: yang tersupervisi *(supervised)* dan tidak tersupervisi *(unsupervised)*. Kedua jenis ini akan digambarkan pada bagian berikut ini.

2.1. Data Mining Tersupervisi (*Supervised Data Mining*)

Supervised Data Mining, selanjutnya disingkat SDM, dapat dilihat sebagai sebuah pendekatan yang sifatnya *top-down* dalam data mining. Pendekatan ini digunakan ketika hasil yang sedang dicari atau diperkirakan diketahui. Model ini dapat dianggap sebagai sebuah kotak hitam karena pentingnya akurasi peramalan, bukan bagaimana model tersebut bekerja. Proses SDM berawal dari membangun sebuah model dan mengujinya pada sebuah kumpulan dataset yang merupakan *training dataset* .

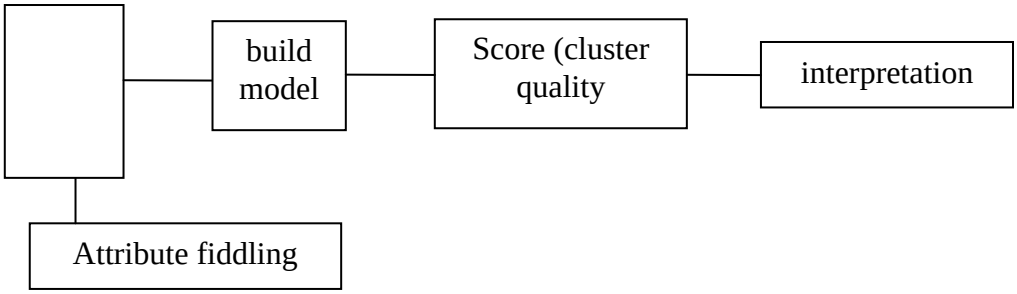


Gambar 2. Data Mining Tersupervisi (Skillicorn, 2001)

Ketika dipercaya telah akurat, model tersebut diuji pada kumpulan data yang baru yang hasilnya kemudian diskor dan dikembalikan kepada pemakai untuk dimanfaatkan dalam proses interpretasi. Jika dianggap berhasil atau baik untuk memprediksi data yang baru, maka model tersebut dapat digunakan dalam kehidupan praktis dan perbaikan atau penyempurnaan lebih lanjut dapat dilakukan terhadap model tersebut. Gambar 2 memperlihatkan komponen dari pendekatan ini.

2.1. Data Mining Tak Tersupervisi (*Unsupervised Data Mining*)

Dalam *Unsupervised Data Mining,* selanjutnya disingkat UDM, tidak ada tujuan atau pola khusus yang sedang dicari. Pola-pola yang ditemukan dalam data langsung disampaikan kepada pemakai untuk ditentukan apakah pola-pola tersebut menarik atau tidak. UDM dapat dilihat sebagai sebuah proses yang sifatnya *bottom-up*. Melalui UDM pemakai ingin mengetahui bagaimana model bekerja dan bagaimana model tersebut menghasilkan sebuah jawaban. Proses ini mungkin melibatkan beberapa tipe pemanipulasian atribut dari kumpulan data (memberikan bobot yang lebih kepada atribut ketimbang yang lain dalam algoritma data mining), kemudian mencoba membangun sebuah model yang dapat menjelaskan hasil-hasil yang baru. Sekali sebuah model terbangun, pemakai dapat melakukan interpretasi hasil-hasil yang telah dinilai untuk pola-pola yang berguna atau menarik. Jika model ini dianggap sebagai sebuah peramal (*predictor*) yang baik untuk hasil pada masa datang, maka model tersebut dapat digunakan dalam dunia praktis. Gambar 3 menggambarkan komponen dari pendekatan ini.



Gambar 3. Data Mining Tak Tersupervisi (Skillicorn, 2001)

3. Data Mining dan Sistem Database Bisnis

Dengan ketersediaan dan kemampuan dari komputer yang canggih, perusahaan dapat melakukan penganalisaan data yang sebelumnya disimpan baik untuk kegunaan transaksi atau dokumentasi atau dengan tujuan menemukan sesuatu yang menarik yang terdapat pada data sehingga dapat membantu perusahaan-perusahaan tersebut dalam meningkatkan proses usaha mereka dan selanjutnya dapat memberikan keuntungan kompetitif dalam dunia usaha (pasar) (Brachman et.al.,1996). Instrumen data mining cenderung dikategorikan ke dalam dua kategori: Instrumen statistik umum atau software yang dikembangkan untuk wilayah bisnis tertentu. Kategori ini lebih lanjut dapat dipecah lagi kedalam instrument untuk tugas yang lebih banyak (*multi task tools*). Instrumen yang umum mungkin meliputi SAS seperti software penggambaran grafik dan diagram, lembar kerja (*spreadsheets*) dan mesin penyelidikan database. Para pemakai instrumen-instrumen ini adalah para spesialis dalam teknik-teknik statistik. Dengan instrumen-instrumen ini

para pemakai dapat memahami kesulitan dari setiap alat analisis atau tool yang berbeda-beda (Brachman et.al.,1996). Alat (*tools*) yang domainnya spesifik akan mendukung penemuan-penemuan dalam domain yang tunggal dan sifatnya lebih abstrak. Pengguna dari tool ini tidak perlu harus mengetahui secara mendalam mengenai proses analisis itu sendiri (Brachman et.al., 1996). Tantangan-tantangan dunia usaha sekarang ini adalah menemukan sesorang yang mempunyai pengetahuan yang memadai baik dalam analisis data statistik maupun domain bisnis itu sendiri. (D. Skillicorn, personal communication, 2001).

Sejauh ini telah banyak ditemukan aplikasi teknologi data mining yang berhasil baik termasuk diwilayah yang terkait dengan bisnis. Kita akan menyoroti 3 wilayah tersebut: Pemasaran (*marketing*), deteksi penipuan (*fraud detection*) dan investasi keuangan (*financial investment*)

3.1 Pemasaran

Pemasaran telah memperoleh keuntungan secara mendasar dari aplikasi penelitian data mining. Contoh klasik yang sering disampaikan dalam literatur adalah “the beer and diapers” (bir dan pembalut) dimana sebuah supermarket melakukan analisis data mining terhadap pola pembelian para pelanggangnya dan menemukan bahwa terdapat korelasi antara pembelian bir dan pembalut. Pihak supermarket menemukan cara bahwa dengan menempatkan dua buah produk tersebut secara berdekatan dan menggiring pelanggan untuk berjalan diantara kedua produk, dapat meningkatkan penjualan pada kedua item tersebut. (Palace, 1996). Menempatkan makanan ringan (*chips*) kentang diantara bir dan pembalut kemungkinan besar juga dapat meningkatkan penjualan *chips* kentang tersebut. Aplikasi pemasaran lain yang populer melibatkan penelusuran pola pembelian pelanggan dan iklan untuk pengguna individual. Contohnya situs The Amazon.com akan tetap menelusuri setiap pembelian yang dilakukan oleh pelanggan. Bila pada kesempatan berikutnya seorang pelanggan masuk ke situs tersebut, pelanggan akan menemukan iklan untuk produk serupa dengan produk yang pernah dibeli sebelumnya dengan tujuan agar pelanggan membeli produk baru tersebut. Produk baru yang “dipaksakan (*pushed*)” kepada pelanggan tersebut dipilih berdasarkan reviu terhadap pola-pola pembelian pelanggan tersebut disamping pola-pola pembelian yang dilakukan pelanggan yang lain. Hal ini untuk melihat produk serupa lainnya yang dibeli berdasarkan produk yang asli tadi.

3.2 Deteksi Penggelapan (*Fraud Detection*)

Sistem Deteksi Penggelapan, selanjutnya ditulis DP, digunakan oleh perusahaan kartu kredit, bank dan telepon untuk mendeteksi penipuan atau penggelapan dalam sistem mereka. Sistem deteksi ini secara umum akan mencoba mendeteksi pola-pola yang tidak umum khususnya dalam aktivitas pelanggan. Contohnya, jika seorang pelanggan tiba-tiba merubah pola dalam pemakaian telepon dan memulai penipuan tagihan jarak jauh yang mahal dengan tujuan seluruh dunia, sistem penipuan menjaga-jaga software yang akan dapat mendeteksi pola-pola yang tidak umum dan secara otomatis berjaga-jaga dan menampilkan kegiatan telepon pada arah yang baik ke operator manusia, yang mana juga dapat menentukan jika langkah selanjutnya diperlukan (Brachman et.al.,1996).

3.3 Investasi Keuangan (*Financial Investment*)

Penerapan data mining juga telah dilakukan pada sektor keuangan. Contohnya para manajer bank dapat menggunakan sebuah model guna membantu menentukan pelanggan mana yang memiliki kemungkinan paling besar tidak mampu membayar pinjaman. Analisis pasar modal/stok memberikan perhatian yang besar terhadap penerapan data mining. Calrberg & associates telah mengembangkan sebuah model untuk memperkirakan standard dan indeks 500 untuk orang miskin dengan menggunakan nilai bunga, pendapatan, deviden, indeks dollar dan harga minyak sebagai nilai bunga dari data yang ada. Model mereka ini dianggap cukup berhasil dan menjelaskan 96% variasi indeks dari tahun 1986 hingga tahun 1995 (Brachman, 1996). Perlu diingat bahwa para investor menggunakan model ini dengan penuh kehati-hatian sebagaimana dalam jangka pendek, pasar market/stok agak sukar untuk diprediksi dan banyak kejadian lain yang mempengaruhi outcome pasar.

3.4 Data mining dan internet

Jaringan internet (*The World-Wide-Web)* menyediakan ilmu pengetahuan serta informasi yang cukup memadai buat kita. Akan tetapi, strukturnya yang dapat didefinisikan secara secara bebas dan bahasa yang terbatas dan hanya dapat dibaca oleh mesin, membuat internet sangat menantang untuk mengakses isinya dan melakukan analisis pengeksplorasian secara efektif (Etzioni, 1996). Kebutuhan akan penambangan web atau jaringan dirasakan dimana-mana, contohnya mesin pencari (*search engines*) mengandalkan teknik yang efisien dan efektif untuk kembali kepada daftar hasil yang sepadan dengan isi dari sebuah pencarian yang dilakukan oleh seorang pemakai. Web mining dapat dibagi kedalam tiga kategori penelitian: penemuan sumber daya (*resource discovery),* penyerapan informasi *(information extraction) dan* generalisasi *(generalization)(Etzioni, 1996).* Ketiga bagian tersebut akan dibahas berikut ini:

***a. Penemuan Sumber Daya** (Resource Discovery)*

Penelitian mengenai Penemuan Sumber Daya pada dasarnya terfokus pada penciptaan indeks dokumen yang dapat dicari yang digunakan pada mesin pencari web. Mesin ini dapat membuat indeks semua kata-kata yang terdapat pada HTML yang jumlahnya jutaan dan kemudian mengakses halaman tersebut ketika seorang pengguna meminta semua dokumen yang mengandung kata-kata kunci tertentu yang terdapat dalam indeks (Etzioni, 1996). Keberadaan dari banyaknya indeks yang berbeda dan disertai dengan interface yang unik memiliki sebuah tantangan yang cukup berarti: dapatkah kita menciptakan beberapa jenis interface yang terintegrasi ketika kita melakukan sebuah pencarian, interface tersebut dapat menemukan semua jenis indeks yang berbeda, pengeksplorasi indeks tersebut kemudian kembali lagi dan melakukan sintesis untuk mendapatkan hasil yang diinginkan? Pertanyaan ini telah memunculkan mesin pencari meta seperti MetaCrawler dan Google. Penelitian masa depan pada wilayah ini akan lebih intensif melibatkan pemanfaatan isi halaman web dalam mesin pencari dan perbaikan kategorisasi situs web secara otomatis (Etzioni, 1996).

b. Ekstraksi Informasi *(Information Extraction)*

Sekali sebuah halaman web ditemukan, informasi harus dapat diserap atau diekstrak. Sebagaimana halaman web berasal dari sumber yang bervariasi, tujuan yang diinginkan adalah agar kita dapat menulis fasilitas ekstraksi otomatis yang dapat bekerja baik dengan sumber data yang kita kenal maupun yang tidak dikenal (Etzioni, 1996). Contohnya penelitian terbaru dalam bidang ini mencoba memprediksi kapan perang sipil pada sebuah bangsa akan pecah berdasarkan informasi yang diserap dari berita (Adler, 2001). Sebuah persamaan yang menggambarkan situasi konflik dibangun berdasarkan anggapan apakah informasi yang didapatkan dapat dikategorikan kedalam tindakan opresif pemerintah atau protes kalangan sipil. Hasil menunjukkan bahwa sistem ini ternyata cukup akurat dan penulis mengatakan bahwa mereka mungkin dapat memperkirakan perang sipil di Algeria 6 hingga 9 bulan sebelum kejadian tersebut meletus jika seandainya mereka telah memiliki teknologi ini (Adler, 2001).

c. Generalisasi *(Generalization)*

Ketika penemuan dan penarikan informasi telah dilakukan secara otomatis, kita dapat melakukan usaha untuk membangun sistem generik/umum yang memungkinkan kita untuk mengklaisifikasikan tipe dokumen yang terdapat pada web dan menggunakannya untuk meningkatkan hasil yang didapatkan dari mesin pencarian. Sebagai contoh, penelitian yang telah mencoba untuk mengelompokkan dokumen berdasarkan kategori tertentu seperti home pages, iklan dan FAQs (Etzoni, 1996). Kemampuan untuk mengelompokkan web page dalam cara seperti ini akan menguntungkan bagi mesin pencari (search engine) dan akan menyediakan mesin tersebut informasi yang lebih semantik mengenai tipe informasi yang sedang dicari dan kemudian meningkatkan kualitas dari hasil pencarian yang diberikan kepada pengguna/pencari informasi (Etzoni, 1996).

3.5 Tantangan dan issu penelitian data mining

Penelitian data mining belum mencapai bentuk yang sempurna/matang dan banyak issu penelitiam yang harus dapat dipecahkan jika teknologi ingin dianggap benar-benar berhasil. Kita akan menyoroti beberapa tantangan penelitian data mining ini dalam bagian-bagian berikut khususnya: mengatasi dataset yang heterogen dan sifatnya massif, mengelola data dinamis dan menilai signifikansi secara statistik.

a. Dataset sifatnya besar/massif dan heterogen

Database terutama pada perusahaan tampaknya semakin besar saat ini. Database ini berisi tabel mendasar dan sifatnya penting dalam jumlah yang banyak (Fayyad, Piatetsky-Shapiro,Smyth, 1996). Tampaknya sudah umum kita dapat menemukan informasi dari database dalam terabytes dengan jutaan catatan atau transaksi. Yang lebih penting dari hal ini, data yang didaptkan mungkin berasal dari sumber yang heterogen dan mungkin saja tidak dalam bentuk yang tepat atau kemungkinan terdapat nilai yang hilang. Dataset seperti ini dapat menjadi sebuah masalah tersendiri bagi algoritma data mining ketika aktivitas penelitian tumbuh secara eksponensial

sebagaiaman berkembangnya dataset. Penelitian pada area ini difokuskan pada penciptaan algoritma yang sangat efisien, penyatuan ilmu pengetahuan yang didapat sebelumnya dan pemprosesan yang paralel guna mengatasi dataset yang jumlahnya sangat besar dan heterogen tersebut (Fayyad et. Al. 1996).

b. Mengelola Data dinamis

Data yang dinamis juga menjadi tantangan tersendiri bagi hasil yang didapatkan dari proses data mining. Data yang berubah mungkin telah mengakibatkan pola dan model lama menjadi tidak up to date lagi. Tampaknya tidak terlalu realistis lagi beranggapan bahwa semua jenis data akan tetap bersifat statis. Penelitian pada area ini (Pengelolaan data dinamis) berpusat pada pengembangan sistem pembelajaran yang dapat melakukan adaptasi dari model-model sebelumnya terhadap dataset yang dimodifikasi ketika ditemukan data yang sifatnya baru (Fayyad dkk, 1996).

c. Signifikansi secara statistik *(Statistical Significance)*

Penilaian signifikansi secara statistik akan selalu menjadi perhatian dari sebuah penelitian data mining. Hal ini disebabkan seringkali terjadi kemungkinan dimana algoritma bersifat berlebihan (*over fitting*) terhadap data. Fenomen ini dikemukakan oleh theorema Bonferroni yang mengatakan bahwa ketika terdapat terlalu banyak kesimpulan yang mungkin diambil dari data, kesimpulan yang pasti dan benar dengan alasan yang sifatnya murni statistik. Membatasi lingkup pencarian merupakan tugas yang sangat menantang khususnya dalam data mining yang tidak terawasi/tak tersupervisi dimana lingkup pencarian hampir tak berakhir. Model data mining yang diawasi/tersupervisi perlu di proses kembali atau dibatasi dengan menggunakan berbagai cara yang bervariasi, contohnya menggunkan strategi penjernihan aturan untuk mendapatkan hasil yang lebih signifikan dengan tingkat kinerja yang lebih baik.

4. Kesimpulan

Tulisan ini diawali dengan gambaran tentang beberapa aktivitas yang berbeda dalam proses data mining. Kemudian dijelaskan dua tipe utama data mining yaitu tersupervisi dan tidak tersupervisi. Kita kemudian menyoroti dua area penelitian data mining yang cukup populer yaitu aplikasi bisnis dan internet serta kontribusi yang telah diberikan oleh kedua area penelitian ini. Tulisan ini diakhiri dengan melihat beberapa tantangan dan issu yang harus di pecahkan oleh penelitian data mining guna menjamin keberhasilannya dimasa datang. Banyaknya data yang harus dianalisa, imbalan yang potensial dari penerapan data mining serta kebutuhan ekonomi dan sosial akan menjamin keberlanjutan dari pertumbuhan dan popularitas penelitian data mining (Fayyad dkk, 1996). Ladang baru bagi ilmu komputer, Komputerisasi bio medis akan memiliki masa depan yang cukup cerah bagi penerapan data mining. Sekarang ini ketika para ilmuan telah hampir selesai melakukan pemetaan *genome* manusia, analisis data mining dapat juga digunakan untuk menemukan pola serta elemen-elemen *genome* manusia. Penelitian pada area ini akan membawa pada pemahaman yang lebih mendalam tentang bagaimana manusia terbentuk dan selanjutnya menemukan cara-cara penyembuhan yang potensial terhadap berbagai kondisi dan penyakit yang mempengaruhi manusia.

Daftar Pustaka

[1] Robert Adler, 2001, "*Conflict Index Warn when a nation faces civil war*", in New Scientist [online], http://archive.newscientist.com/archive.jsp?id=23140700 (23 November 2003).

[2] R.J.Brachman, Khabaza T., Kloesgen W., Piatetsky-Shapiro G., Simoudis E, 1996, *"Mining Business Databases"*, in Communications of the ACM, Vol. 39, No. 11. (pp. 42-48).

[3] Oren Etzioni, 1996, *"The World-Wide Web: Quagmire or Gold Mine?"*, in Communications of the ACM, Vol. 39, No. 11. (pp. 65-68).

[4] U. Fayyad , Smyth P, 1996, *"The KDD Process for Extracting Useful Knowledge from Volumes of Data"*, in Communications of the ACM, Vol. 39, No. 11. (pp. 27-34).

[5] B. Palace, 1996, *"What Is Data Mining"*, Technology Note Prepared for Management 274A, http://anderson.ucla.edu/faculty/jason.frand/teacher/technologies/ palace/datamining.htm (23 November 2003).

[6] David Skillicorn, 2001, *"Course Handouts in Data Mining"*, Department of Computer and Information Science, Queen's University, at Kingston Canada.