# DATA MINING : MASA LALU, SEKARANG DAN MASA MENDATANG

Ronsen Purba

STMIK Mikroskil
Jl. Thamrin No. 122, 124, 140 Medan
ronsen@mikroskil.ac.id

**Abstrak**

*Data mining* telah menjadi disiplin ilmu yang dibangun dalam domain kecerdasan buatan (AI), dan rekayasa pengetahuan (KE). *Data mining* berakar pada *machine learning* dan statistika, tetapi merambah bidang lain dalam ilmu komputer dan ilmu lainnya seperti biologi, lingkungan, finansial, jaringan dan sebagainya. *Data mining* telah mendapatkan begitu besar perhatian pada dekade terakhir sehubungan dengan perkembangan *hardware* yang menyediakan kemampuan komputasi luar biasa yang memungkinkan pengolahan data besar. Tidak seperti kajian lain dalam AI dan KE, *data mining* dapat diperdebatkan sebagai sebuah aplikasi dibandingkan dengan sebuah teknologi, dengan demikian diharapkan akan menjadi topik yang hangat dibahas di masa mendatang, mengingat pertumbuhan data yang bersifat eksponensial. Paper ini memberikan kilas balik perjalanan sejarah *data mining*, keadaan saat ini dan beberapa pandangan dan perkembangan ke depan.

**Kata kunci :** ***kecerdasan buatan, rekayasa pengetahuan, data mining, machine learning***

## 1. Pendahuluan

Asal usul *data mining* dapat dilihat kembali ke akhir tahun 1980-an pada saat istilah tersebut mulai digunakan, paling tidak dalam kalangan komunitas riset. Pada awalnya terdapat sedikit perdebatan tentang makna dan cakupan dari istilah tersebut dan sampai saat ini pertentangan tersebut masih terjadi. Dalam arti luas *data mining* dapat didefenisikan sebagai sekumpulan mekanisme dan teknik yang direalisasikan dalam perangkat lunak untuk mengekstrak informasi tersembunyi dari kumpulan data. Pengertian tersembunyi dalam defenisi di atas sangat penting; *query* yang dilakukan dalam SQL meskipun sangat rumit bukanlah *data mining*. Juga istilah informasi harus diinterpretasikan dalam arti luas. Sebelum tahun 1990-an *data mining* umumnya dikenal sebagai sub proses dalam lingkup lebih besar yang disebut *Knowledge Discovery in Databases* (KDD). Meskipun dalam konteks modern dari *data mining* KDD akan lebih sesuai, karena sumber pengetahuan bukan lagi terbatas pada *database*. Defenisi yang lebih umum dari KDD adalah apa yang dikemukakan oleh Fayyad et al. yakni: "*the nontrivial process of identifying valid, novel, potentially useful, ultimately understandable patterns in data*" [6, 7]. Dengan anggapan sebagai sub proses dalam cakupan KDD, *data mining* terkait dengan penemuan akan "informasi tersembunyi". Sub proses lainnya yang juga merupakan bagian dari KDD di antaranya persiapan data (*warehousing*, *data cleaning*, *pre-process*, dan lain-lain) dan analisis serta visualisasi dari hasil. Untuk beberapa tujuan praktis KDD dan *data mining* sering dianggap sinonim, tetapi secara teknis ternyata yang satu merupakan sub proses dari yang lain [7, 31].

Data yang digunakan dalam proses *data mining* pada awalnya hanya untuk data dalam bentuk tabel (relasional) mengingat keterbatasan kemampuan komputasi saat itu. Dengan peningkatan kemampuan komputasi, maka waktu komputasi (meskipun tetap menjadi isu penting) tidak lagi menjadi persoalan utama dan digantikan dengan tujuan lain yakni akurasi dan keinginan untuk menambang data yang jauh lebih besar. Saat ini, dalam konteks data

bentuk tabel telah diperlakukan secara khusus dengan teknik yang tersendiri. Hal ini dapat dilihat dari banyaknya sistem seperti SPSS yang melakukan penambangan data dari dalam tabel. Akan tetapi, jenis data yang tersedia secara digital semakin banyak dan berasal dari berbagai sumber berbeda. Jenis data saat ini sangat beragam seperti gambar, teks, video, multimedia, graf dan jaringan. Hal ini membuat kajian *data mining* terus berkembang dan terus menghadirkan tantangan baru [31].

Popularitas *data mining* terus meningkat secara signifikan pada tahun 1990-an, terutama dengan pelaksanaan sejumlah konferensi yang dikhususkan pada kajian tersebut, seperti: ACM SIGKDD *annual conference* tahun 1995, European PKDD and Pacific/Asia (PAKDD) pada tahun 1997. IEEE ICDM mulai diadakan pada tahun 2001 sebagai konferensi SIAM yang pertama [8, 20]. Hal ini meningkatkan popularitas *data mining* yang secara bersamaan didukung oleh kemajuan teknologi, kemampuan CPU dan media yang menyimpan data dalam jumlah besar dan mengolahnya dalam waktu yang lebih cepat. Menjadi hal biasa untuk perusahaan komersial untuk memelihara data dalam berbagai bentuk yang dapat dibaca komputer, dan dalam kebanyakan kasus digunakan untuk mendukung aktivitas bisnis dan pemikiran bahwa data tersebut dapat "ditambang" sering masih dianggap sebagai prioritas kedua. Era tahun 1990-an juga ditandai dengan penggunaan kartu kesetiaan pelanggan yang memungkinkan perusahaan untuk mencatat belanjaan dari setiap pelanggan. Data yang tersimpan dalam volume besar tersebut dapat "ditambang" untuk mengetahui pola/perilaku pembelian oleh masing-masing pelanggan. Popularitas *data mining* terus berkembang sampai hari ini terlebih dengan kemungkinan menambang data dari data yang tidak standar (non tabular) [21, 31].

## 2. Mekanisme dan Teknik dalam *Data Mining*

Mekanisme dan teknik dalam cakupan *data mining* dapat dijelaskan sebagai gabungan dari pendekatan dalam *machine learning* dan statistika; dan dari perspektif ini dapat dikatakan bahwa *data mining* "tumbuh" keluar dari disiplin *machine learning* dan statistika. Sebagai bukti bahwa komunitas *data mining* dipenuhi oleh mereka yang berasal dari ilmu komputer dan statistika, European Conference on Machine Learning (ECML) dan European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD) lahir bersamaan pada tahun 2001 dan tetap bersatu sampai sekarang. Akan tetapi, terdapat perbedaan antara keduanya. *Data mining* fokus pada data (dapat dianggap sebagai domain aplikasi); sementara *machine learning,* setidaknya dalam bentuk tradisional, fokus pada mekanisme untuk membuat komputer dapat "belajar". *Machine learning* dapat dianggap sebagai teknologi, sementara *data mining* dan dengan perluasan KDD dianggap sebagai aplikasi. Teknik dalam *data mining* dapat dikategorikan sebagai (i) ekstraksi pola, (ii) *data clustering* dan (iii) *classification/categorization*. Dari literatur tentang *data mining* dapat ditemukan beberapa teknik lain yang diadopsi dari bidang statistika dan matematika seperti regressi dan *principle component analysis* (PCA) [6, 21].

### 2.1 Ekstraksi Pola

Sepanjang sejarahnya, *data mining* telah memfokuskan kajian untuk menemukan pola dalam data. Pola-pola tersebut dapat dalam berbagai bentuk, seperti pola belanja pelanggan; pola alternatif dapat berupa trend dalam temporal atau longitudinal, subgraf yang sering muncul dalam graf dan lain-lain. Sebuah pola sering muncul sebagai kombinasi dari sejumlah entitas, kejadian, objek dan sebagainya. Salah satu teknik dalam penambangan pola adalah *Association Rule Mining* (ARM) dengan algoritma Apriori yang dikemukakan pertama kali oleh Agrawal et al. dalam konteks *super market basket analysis* [2]. Tujuannya adalah mengidentifikasikan pola yang sering muncul dalam data dan kemudian dari pola tersebut dilakukan ektraksi untuk mendapatkan *Association Rules* (ARs). Sebuah AR adalah aturan

probabilistik yang menyatakan bahwa jika kumpulan atribut data tertentu muncul maka sesuatu yang lain juga kemungkinan terjadi. Tantangan mendasar dari ARM adalah jika kumpulan data dengan *N* atribut, maka akan terdapat $2^N$ kandidat pola yang dapat diperoleh [6, 19]. Sejumlah perluasan (pengembangan) telah dilakukan seperti pembobotan dan pemanfaatan ARM, *incremental* ARM, *fuzzy* ARM dan lain-lain. Riset terbaru terkait dengan pola yang sering muncul telah diarahkan pada sistem rekomendasi. Pola yang sering muncul terpopuler saat ini adalah algoritma *Frequent Pattern* (FP) *Growth* [9-10, 22, 31].

### 2.2 *Clustering*

*Clustering* berkenan dengan pengelompokan data ke dalam kategori tertentu. Hal ini dimaksudkan dalam konteks data besar untuk aktivitas eksplorasi. Untuk berbagai persoalan, *clustering* dilengkapi dengan jumlah *cluster* (seperti penerapan k-means) [25] atau tergantung pada nilai *threshold* tertentu (seperti penerapan kNN) [12, 14, 19]. Pendekatan alternatif adalah mengadopsi bentuk tertentu dari *clustering* hirarkis dimana data secara iteratif dipartisi untuk membentuk himpunan *clusters*. Algoritma *clustering* hirarkis yang paling umum adalah BIRCH [31,33]. Kelebihan dari konfigurasi *cluster* biasanya diukur dalam hal *intra-cluster cohesion* dan *inter-cluster separation*. Isu terkait dengan algoritma *clustering*, seperti pada k-means dan kNN, adalah bahwa *clusters* yang terbentuk dinyatakan sebagai *hyper-spheres* dimana bentuk ini belum tentu ideal. Isu lainnya adalah dimensi yang tinggi dari data dan penanganan terhadap *noise* (*outliers*) serta data dalam bentuk kategori. Satu hal menarik adalah tidak ada algoritma *clustering* terbaik yang dapat diaplikasikan terhadap semua bentuk data [1, 5, 14].

### 2.3 Klasifikasi

Klasifikasi berkenan dengan pembentukan "*classifier*" yang dapat digunakan terhadap data tersembunyi sehingga data dapat diklasifikasikan ke dalam grup (kelas). Klasifikasi demikian dapat dilakukan dengan *clustering*. Perbedaannya, adalah bahwa klasifikasi membutuhkan data *training* yang sudah diberi label dari mana *classifier* dapat membangun kelas. Klasifikasi seperti ini sering disebut dengan *supervised learning*, sementara *clustering* dianggap sebagai *unsupervised learning* [5]. *Classifier* dapat berbentuk pohon keputusan, *support vector machine* (SVM), atau aturan dan lain-lain [31]. Algoritma paling berpengaruh dalam pembentukan pohon keputusan adalah C4.5 yang telah berkembang menjadi See5/C5.0 [24, 31]. Dalam konteks *classifier* berbasis aturan, klasifikasi dapat dianggap sebagai kasus khusus seperti AR dengan teknik ARM yang dapat menghasilkan aturan yang diperlukan. Algoritma yang paling sering dijadikan referensi tentang ARM adalah Algoritma CBA [25, 27, 31]. Teknik klasifikasi lainnya termasuk untuk regressi adalah algoritma CART, dan Naive Bayes [6, 8-9, 31]. *Classifiers* dapat berupa: (i) *binary classifiers*;(ii) *multi*-class *classifers*; atau (iii) *multi-labelled*. Kualitas kelas yang dihasilkan biasanya diukur dalam hal akurasi, sensivitas, dan spesifikasi. Untuk memperluas klasifikasi dimungkinkan untuk mencari kemiripan antar kelas dan *reasoning* berbasis kasus. Klasifikasi terus mendapat perhatian dari komunitas *data mining*. Salah satu bentuk pengembangan dari klasifikasi adalah konsep tentang *ordinal classifiers* dimana kelas-kelas yang diperoleh dapat diurutkan sedemikian rupa. Juga berkembang kajian terhadap klasifikasi dinamis, seperti pengklasifikasian aliran video [8].

## 3. Aplikasi *Data Mining*

Berikut ini diberikan sejumlah aplikasi *data mining* di luar dari data tabular.

### 3.1 *Text Mining*

Langkah alami berikutnya setelah data tabular tradisional adalah *text mining*. Aplikasi ini membangun sebuah *classifier* untuk mengelompokkan kumpulan dokumen yang banyak (seperti kumpulan artikel, berita atau laman web). Aplikasi lainnya adalah penambangan opini atau *questionare* untuk mendapatkan informasi penting dan berguna terkait dengan satu isu atau produk. Aplikasi lebih lanjut adalah membuat ringkasan yang dimulai dari kondisi kabur sampai penemuan informasi. Dalam konteks klasifikasi teks, SVM bekerja dengan baik (tetapi tidak memberikan penjelasan tentang klasifikasi yang dihasilkan) [15, 18]. Secara umum, isu dalam *text mining* adalah bagaimana cara terbaik untuk merepresentasikan data tekstual sehingga memungkinkan penggunaan teknik *data mining* yang lebih mudah. Yang menjadi pertanyaan adalah kata apa yang dimasukkan sebagai kata kunci. Hal ini dapat ditentukan oleh pakar, atau diekstrak pada saat penerapan teknik *data mining* atau teknik *Natural Language Processing* (NLP). Alternatif representasi pengelompokan kata dapat dilakukan dengan menggunakan ungkapan (*phrase*). Akan tetapi, dalam kedua kasus, urutan kata/ungkapan akan hilang. Teknik alternatif mencoba untuk mempertahankan pengetahuan tentang urutan, meskipun akan menambah kompleksitas komputasi. *Text mining*, dalam segala bentuk dan keinginan, akan tetap populer dalam aktivitas *data mining* di masa mendatang [8, 31]

### 3.2 *Image Mining*

Terdapat banyak koleksi data citra ukuran besar yang telah dihasilkan dalam berbagai jenis aplikasi. Sama seperti pada *text mining*, *image mining* berkenan dengan representasi citra (2D atau 3D) sedemikian hingga teknik penambangan dapat diterapkan. Teknik dasar adalah menghasilkan histogram, pohon atau graf. Alternatif lain, dapat merepresentasikan citra dalam bentuk kumpulan objek dan diidentifikasi dengan teknik segmentasi dan registrasi. Teknik segmentasi citra mempunyai sejumlah keterbatasan yang tergantung pada sifat alami dari citra dan merupakan subjek riset lanjutan dalam komunitas bidang pengolahan citra. Analisis citra tetap menjadi topik riset yang menantang. Dalam bidang tertentu, seperti bidang medis, dimana masalah dapat dipersempit *image mining* telah mencapai sukses tersendiri. Misalnya klasifikasi data citra retina mata untuk identifikasi manusia dan pemindaian data dengan *Magnetic Resonance Imaging* (MRI) untuk mengindentifikasi ketidakteraturan (kerusakan) yang terjadi pada tubuh pasien. Bidang populer lainnya adalah aplikasi dalam *image mining* berbasis satelit, seperti GoogleMap atau prediksi cuaca dalam pengenalan pola. Riset terkini dalam *image mining* tetap akan menjadi fokus pada seberapa baik representasi citra sehingga teknik *data mining* dapat digunakan. Dalam hal ini, penting untuk melakukan observasi bahwa untuk aplikasi teknik *data mining* kita tidak harus merepresentasikan citra yang mudah diinterpretasikan oleh manusia, sepanjang *data mining* bekerja dengan baik [4, 17, 19].

### 3.3 *Graph Mining*

*Graph mining* secara prinsip merupakan perluasan dari *frequent pattern mining*, misalnya apa yang menarik dalam subgraf yang sering muncul. Praktisi *graph mining* mengatakan bahwa segala sesuatu dapat direpresentasikan dengan graf. Pada kenyataannya dapat dilihat bagaimana dokumen, email dan citra dapat direpresentasikan dengan graf (seperti pada *search engine*). Pada level yang lebih tinggi kita dapat mengidentifikasikan dua bentuk masalah yakni: (i) *frequent sub-graphs* yang sering terjadi pada koleksi graf dan (ii) *frequent sub-graphs* yang terjadi dalam satu graf besar. Kita dapat membedakan antara *graph mining* dan *tree mining*; *tree mining* lebih mudah menelusurinya karena sifat alami dari pohon (seperti tidak ada siklus dan factor percabangan yang teratur dan lain-lain). *Graph/tree mining* membutuhkan bentuk kanonik tertentu dengan mana kita merepresentasikan graf [29]. Isu utama saat ini terkait *graph mining* antara lain pembentukan kandidat subgraf dan pengujian isomorfisme subgraf. Algoritma *mining* yang paling berpengaruh dalam analisis *frequent sub-*

*graph* adalah gSpan [21, 23, 29, 32]. Pengembangan populer dari *graph mining* adalah *social network mining* (*social network analysis*). Motivasi untuk hal ini adalah kepopuleran dari situs jejaring sosial saat ini seperti *facebook* dan *twitter*. Akan tetapi, terdapat sejumlah bentuk jejaring sosial seperti jaringan transportasi atau *co-authoring* dimana teknik *social network mining* dapat digunakan [31].

## 4. Trend Masa Depan

4.1 Menambang Objek Kompleks untuk Tipe Tertentu

Metode otomatis modern untuk pengukuran, koleksi, dan analisis data dalam berbagai bidang telah menghasilkan data dalam ukuran masif dan struktur yang kompleks. Kompleksitas tersebut muncul karena kebutuhan akan deskripsi yang lebih presisi dan lebih kaya dari objek dunia nyata. Di sisi lain, perbaikan yang begitu cepat dalam teknik pengukuran dan analisis memungkinkan eksplorasi terhadap objek dengan multi tujuan. Untuk dapat mengelola volume data yang sangat besar dari data yang begitu kompleks, digunakanlah sistem basis data. Secara tradisional, basis data relasional digunakan untuk menyimpan informasi tersebut dalam bentuk kumpulan atribut. Basis data relasional-objek bahkan memungkinkan seseorang untuk mendefenisikan tipe data untuk memodelkan objek tertentu. Menyadari bahwa analisis data manual dengan volume besar dan tingkat kompleksitas yang tinggi secara praktis adalah mustahil, sehingga muncul kebutuhan akan teknik *data mining* yang dapat menemukan pengetahuan baru dan menarik dari kumpulan data tersebut. Beberapa metode untuk data yang kompleks telah dikemukakan selama beberapa tahun terakhir, seperti penambangan *multi-instance objects* [1, 3, 10, 14, 20, 21, 30] atau penambangan *multi-represented objects* dalam bentuk terawasi [4, 17, 22, 33], tak terawasi [5, 14, 18, 20] atau *graph mining* [23, 29, 32].

Akan tetapi, pendekatan *data mining* biasanya hanya menangani sub tipe data. Sebagai contoh, *itemset mining* dikhususkan pada data string atau daftar nilai yang mungkin, beberapa pendekatan klasifikasi dan *clustering* membutuhkan data numerik, sementara yang lain memungkinkan dengan kategori. Kemungkinan lain adalah menggabungkan pendekatan berbeda untuk memberikan hasil yang lebih sesuai. Contohnya, kelas-kelas tertentu dari *string kernel* untuk mengukur *frequent substrings* dari sebuah barisan atau teks dan pada dasarnya dilakukan dengan menghitung frekuensinya dan akhirnya membentuk fitur numerik [20, 31].

Memodelkan dunia nyata akan menciptakan representasi yang sangat sederhana dengan keterbatasan tampilan. Konsep "*object-oriented modeling*" dimaksudkan untuk mendeskripsikan objek yang rumit dalam bentuk sederhana melalui cara yang formal. Dalam hal ini, atribut dari sebuah objek mungkin saja tipe primitif atau objek itu sendiri. *Object-oriented* dan *object-relational databases* dapat merepresentasikan koleksi objek yang demikian. Hal ini diharapkan mampu secara langsung menambang pada objek-objek tersebut bukan hanya dari sebagian objek. Beberapa tahun terakhir, sejumlah langkah dilakukan untuk menambang data yang dimodelkan sebagai graf, atau multi representasi, multi relasional dan multi bentuk data. Dalam aspek tertentu, pendekatan ini adalah penggabungan atau generalisasi dari pendekatan sebelumnya untuk data yang tak terstruktur. Di sisi lain, pendekatan yang hampir sama dapat dipahami sebagai penyesuaian terhadap jenis representasi yang lebih umum, tetapi bukan representasi yang universal. Akan tetapi, semua model tersebut mengasumsikan bahwa data bersifat statis. Gambaran "object-oriented modeling" juga menyangkut sifat objek yang disebut dengan "methods", yakni sifat dinamis. Lebih jauh, model *sequence diagrams* atau *activity diagrams* dapat memodelkan kronologi dari sifat-sifat pola. Dalam kenyataannya, perilaku program merupakan tugas *data mining* secara umum [15, 27, 31]. Beberapa tahap dilakukan untuk secara langsung menambang sistem *object-oriented* misalnya dalam Kanellopoulos et al. [15].

Merepresentasikan objek kompleks dengan cara sederhana seperti fitur vektor numerik dapat dipahami sebagai satu cara untuk menggabungkan domain pengetahuan ke dalam proses *data mining*. Domain pakar mencari cara untuk menggunakan fitur penting dari sebuah objek ke klasifikasi objek baru dari tipe sama yang pada akhirnya dengan menggunakan fungsi kompleks untuk mentransformasikan atribut dari berbagai tipe ke tipe lain [16].

Lebih jauh, pengetahuan spesifik untuk domain tertentu meningkat dalam jumlah dan kompleksitas itu sendiri. Dengan demikian, biasanya tidak dapat disurvei dengan kepakaran manusia semata – komunitas akhirnya mulai memasok pengetahuan mereka dalam basis data atau basis pengetahuan. Sehingga di masa mendatang, algoritma *data mining* harus mampu secara otomatis mengambil domain pengetahuan terpercaya untuk meningkatkan efektifitasnya.

Untuk memroses objek kompleks, *data mining* terdistribusi menjadi semakin penting [31]. Beberapa domain aplikasi mempertimbangkan objek kompleks yang sama sesuai dengan karakteristik sama pada lokasi berbeda dan/atau pada waktu berbeda (misalnya seorang pasien dapat melakukan konsultasi dengan dokter berbeda atau observasi berkelanjutan atau observasi sebuah bintang dengan melibatkan teleskop di seluruh dunia). Pada sisi lain, *data mining* pada objek kompleks membutuhkan kemampuan komputasi yang lebih besar dibandingkan dengan proses *mining* pada vektor fitur. Akhirnya, tidak semua partisipan dalam aktivitas bersama dari *data mining* yang ingin berbagi semua data yang mereka kumpulkan, mungkin untuk memproteksi kerahasiaan pelanggan mereka, sehingga muncul kebutuhan akan algoritma *data mining* terdistribusi dan mempertahankan kerahasiaan data untuk data kompleks [31, 33].

4.2 Aspek Temporal: Dinamis dan Hubungan

Pengetahuan tentang perilaku dari objek merupakan bagian integral dari pemahaman hubungan yang rumit dalam sistem dunia nyata dan aplikasi. Semakin banyak metode baru untuk observasi dan pembentukan data yang sesuai untuk menangkap hubungan rumit tersebut. Namun, beberapa arah penelitian dalam *data mining* lebih fokus pada deskripsi statis dari objek atau tidak dimaksudkan untuk menangani data yang bersifat dinamis dalam hal perilaku dan hubungan.

Sebagai tambahan terhadap model data yang lebih kompleks, hubungan antar objek juga ditentukan dengan aspek temporal yang tersembunyi dalam data. Terdapat dua tantangan yang muncul terkait dengan aspek temporal ini. Pertama, data dapat menjelaskan perkembangan dari waktu ke waktu atau mekanisme temporal, seperti aliran data atau data dalam bentuk *time-series*. Meskipun menambang dari tipe data temporal telah mendapat perhatian besar dalam beberapa tahun terakhir [9, 25], masih banyak tantangan yang harus dihadapi. Misalnya, algoritma *data mining* di masa mendatang harus mampu menemukan korelasi berbeda dalam *time-series* dimensi tinggi atau harus mampu mengeksplorasi tipe baru dalam model kemiripan untuk data temporal untuk mengatasi permasalahan praktis yang berbeda-beda. Jadi, mengikuti pendekatan yang lebih berorientasi pada aplikasi dapat memberika tantangan baru terhadao komunitas *data mining*.

Tantangan kedua adalah bahwa pola yang diobservasi juga mempunyai sifat temporal artinya berubah dari waktu ke waktu. Tantangan penting adalah mempertahankan pola untuk tetap terbarukan tanpa melakukan perhitungan ulang dari nol. Secara umum, hal ini diperlukan dalam menambang data *stream*. Akan tetapi, juga dalam lingkungan data base yang dinamis dimana input, penghapusan dan edit sering terjadi membuat pola terus berubah dan menjadi tantangan berat di masa mendatang [1, 11, 31]. Selanjutnya, juga sangat menarik dalam berbagai aplikasi untuk memonitor evolusi pola yang terjadi dan menurunkan pengetahuan terkait dengan perubahan tersebut atau perilaku pola yang dinamis dan lengkap. Menemukan "perubahan pola dalam pola" merupakan tantangan penting yang masih kurang

mendapat perhatian dari para peneliti *data mining*. Jadi pendekatan untuk aspek temporal ini diproyeksikan akan menjadi bidang kajian yang akan banyak mendapat perhatian, karena hal ini akan memainkan peranan penting dalam proses pemahaman hubungan kompleks dan perilaku objek atau sistem [11, 31].

4.3 Pre-processing Data: Cepat, Transparan, dan Terstruktur

Penemuan pengetahuan dalam *data mining* untuk dunia nyata diakui tidak hanya sekedar pengenalan pola. Proses yang dilakukan bukan hanya sekedar analisis data, tetapi bagaimana membawa data tersebut ke dalam sebuah bentuk yang memungkinkan analisis data dilakukan. Menurut estimasi bahwa penambangan data yang aktual hanya sekitar 10% dari waktu yang dibutuhkan untuk proses penemuan pengetahuan secara keseluruhan. Dalam kenyataan bahwa persiapan data agar dapat ditambang ternyata tidak kalah penting dibanding dengan proses *data mining* itu sendiri yang sangat mempengaruhi hasil pada tahap berikutnya [11, 19, 21].

Sebagai tambahan terhadap format dan kelengkapandata, algoritma *data mining* secara umum membutuhkan data yang berasal dari satu sumber. Akan tetapi, entitas dalam basis data berbeda, mungkin mempunyai skala berbeda dan mungkin saja dihasilkan oleh teknik eksperimen dan *software* berbeda. Sebelum memulai analisis data, perbedaan-perbedaan tersebut harus diseimbangkan melalui proses integrasi data. Jika tidak, akan muncul resiko menemukan pola dalam data yang diakibatkan oleh perbedaan sumber, dan bukan oleh fenomena dalam domain aplikasi yang ingin dipelajari. Disamping masalah integrasi ini, upaya lebih jauh masih perlu dilakukan untuk mengatasi format dan integrasi semantik, yang membentuk tantangan jangka panjang untuk komunitas *data mining* [17, 31]. Jadi, integrasi data adalah isu sentral lain yang harus ditempuh sebelum proses penemuan pengetahuan. Dimensi data yang tinggi akan mengarah pada masalah skalabilitas untuk algoritma *pre-processing*, dan urusan data yang hilang dan integrasi pada struktur data seperti string dan graf justru menjadi tantangan besar secara teori. Khususnya dalam komunitas *data mining* statistika akan menghadapi tantangan untuk mendesain uji statistika dan mendapatkan algoritma yang secara efisien dan skala terkendali melakukan *pre-processing* data yang multi dimensi dan struktur yang kompleks.

Di masa mendatang kita akan melihat bahwa *pre-processing* akan lebih bagus, lebih cepat dan lebih transparan dari yang ada saat ini. Untuk aplikasi *data mining* yang cepat dan *user-friendly*, *pre-processing* data akan secara otomatis dikerjakan oleh sistem, dan semua langkah yang diambil akan dilaporkan ke pengguna atau bahkan mungkin secara interaktif dapat dikendalikan oleh pengguna (seperti yang ada pada microsoft surface). Representasi data yang umum dan bahasa deskripsi untuk *pre-processing* data akan memudahkan komputer dan pengguna untuk memutuskan dan mempelajari pengetahuan yang ada dan cara ini sudah dan akan terus digunakan. Sistem lanjutan akan secara otomatis melakukan *pre-processing* dengan cara berbeda-beda dan memberikan laporan hasil dalam berbagai bentuk. Uji statistika yang baru dan algoritma *pre-processing* yang lebih baik akan terus bermunculan untuk dapat menangani data multi dimensi dan multivariat [13, 16, 27].

4.4 Penggunaan yang Terus Meningkat

Trend terakhir menunjukkan bahwa penggunaan *data mining* meningkat secara terus menerus. Namun yang tidak kalah penting adalah kebutuhan akan panduan terhadap pengguna mengingat semakin banyaknya permasalahan yang pada akhirnya didekati dengan teknik *data mining*. Sistem saat ini mungkin sudah memberikan “help” tetapi belum ada penjelasan mengapa digunakan sebuah metode untuk mengerjakan pekerjaan tertentu. Sebagai contoh, pembobotan Euclidean mungkin sering digunakan tetapi tidak ada penjelasan apa kelebihan dan kekurangan rumus tersebut dan bagaimana pemilihan parameter yang sesuai agar memberikan hasil yang baik. Pemilihan metode yang sesuai dan menemukan parameter

yang masuk akal sering menjadi kendala bagi pengguna sistem. Keinginan pengguna yang mengharapkan interaksi yang tidak terlalu banyak juga mungkin menjadi salah persyaratan ke depan. Aspek lain yang terkait dengan penggunaannya adalah *intuitiveness* pada saat melakukan penyesuaian parameter dan sensitifitas dari parameter. Jika hasil tidak tergantung pada perbedaan kecil pada variasi parameter, maka menyesuaikan algoritma sedikit lebih mudah. Untuk memenuhi persyaratan ini, kita harus terlebih dahulu membedakan dua jenis parameter dan setelah itu kita lihat apa yang bisa dilakukan di masa mendatang [21, 31].

Parameter tipe pertama adalah pemilihan algoritma *data mining* untuk menghasilkan pola penting. Sebagai contoh nilai *k* pada *classifier k-NN* secara langsung akan mempengaruhi akurasi klasifikasi yang dihasilkan atau kualitas klasifikasi. Parameter tipe yang kedua terkait dengan penjelasan semantik dari objek yang ada, misalnya matriks biaya yang digunakan dalam perubahan jarak telah menjadi domain pengetahuan dan bervariasi antar aplikasi berbeda [4]. Aspek penting lainnya adalah bahwa parameter yang digunakan untuk memodelkan konstrain dari dunia nyata. Berdasarkan pertimbangan ini, beberapa hal yang perlu diperhatikan di masa mendatang dapat dirangkum sebagai berikut:

1. Menghindarkan parameter tipe pertama jika mungkin untuk membangun sebuah algoritma
2. Jika parameter tipe pertama memang wajib, maka perlu menemukan *setting* otomatis yang optimal.
3. Dari pada mencari pola untuk satu nilai yang mungkin untuk parameter tipe kedua, dapat dilakukan secara simultan menurunkan pola untuk setiap parameter *settings* dan menyimpan untuk *postprocessing*.
4. Mengembangkan metode yang ramah terhadap pengguna untuk mengintegrasikan domain pengetahuan dimana diperlukan.

Aspek penting yang kedua terkait dengan penggunaan yang semakin luas adalah proses atau cara menurunkan pola itu sendiri. Saat ini, kebanyakan algoritma *data mining* menghasilkan pola yang dapat didefenisikan dalam format atau bentuk matematis. Akan tetapi, makna dari pola yang ditemukan masih sangat sedikit. Dengan kompleksitas objek yang semakin tinggi, masalah ini akan mendapat perhatian lebih dari para penggiat *data mining* di masa mendatang. Meskipun dimungkinkan untuk menginterpretasikan makna dari permukaan dalam sebuah ruang vektor, pola yang didapat dari objek yang lebih kompleks mungkin tidak mudah diinterpretasikan bahkan oleh pakar sendiri. Jadi, bukan hanya data input untuk *data mining* yang semakin kompleks, tetapi juga pola yang diperoleh akan jauh lebih kompleks. Dalam beberapa aplikasi, pola yang paling umum yang diturunkan oleh metode standar belum menghasilkan solusi yang memuaskan untuk tugas tertentu. Untuk mengatasi masalah ini, pola yang didapat perlu untuk mengisi sekumpulan kendala (*constraints*) yang membuatnya semakin menarik untuk aplikasi tersebut. Contoh untuk tipe pola ini adalah *correlation clusters* [3] dan *constrained association rules* [27]. Untuk itu, kita dapat mendaftarkan beberapa tantangan lainnya berikut ini:

1. Pola yang dijelaskan oleh algoritma *data mining* masih terlalu abstrak untuk dipahami. Akan tetapi, sebuah pola yang disalahtafsirkan dapat mengandung resiko yang sangat besar. Misalnya kebanyakan algoritma *data mining* tidak membedakan antara *causality* dan *co-occurrence*.
2. Seperti telah disebutkan di atas, algoritma yang ada saat ini lebih fokus pada pola standar. Akan tetapi, menurunkan pola ini sering tidak menghasilkan solusi yang langsung dan komplit untuk kebanyakan kasus. Lebih jauh, dengan peningkatan kompleksitas dari data yang dianalisis, kemungkinan bahwa pola yang diturunkan juga akan semakin sulit. Jadi trend masa depan salah satunya adalah menemukan pola yang lebih kaya.

3. Tugas terakhir terkait dengan pola di masa mendatang adalah peningkatan jumlah pola yang valid, yang mungkin diperoleh pada data besar dari objek yang sangat kompleks yang akan terlalu besar untuk ditangani oleh manusia tanpa bantuan sistem dalam mengorganisirnya.

## 5. Kesimpulan

*Data mining* telah menjadi terkenal dalam dua dekade terakhir sebagai sebuah disiplin ilmu atau kajian yang menawarkan sejumlah keuntungan dengan keterkaitannya dengan berbagai domain ilmu pengetahuan baik secara komersial dan akademis. Secara luas *data mining* dapat dilihat sebagai domain aplikasi, bukan teknologi. Peningkatan kemampuan institusi mengumpulkan dan menyimpan data digital, difasilitasi oleh perkembangan teknologi prosesor dan media penyimpanan data serta kemajuan teknologi web, membuat keinginan untuk "menambang" data menjadi lebih meluas. Komunitas *data mining* mempunyai sejumlah teknik yang telah maju dan tersedia yang dapat digunakan pada variasi data yang lebih besar. Secara umum proses aktual dari *data mining* tanpa disadari telah terjadi dalam keseharian terutama pada saat kita mengakses internet, seperti pada kebanyakan sistem cerdas saat ini. Isu aktual yang lebih menantang dari sekedar mengolah data adalah teknik yang dapat digunakan setelah pengolahan seperti visualisasi, pemberian alasan dan lain-lain. Jadi, meskipun telah diperoleh kemajuan dalam pemrosesan mendapatkan pengetahuan dari pekerjaan "*mining*", proses "*end-to-end*" dari *data mining* masih membutuhkan riset lanjutan yang mampu memberikan perbaikan yang signifikan. Dalam visualisasi sebagai contoh, Microsoft telah menghasilkan *microsoft surface* yang memungkinkan pengguna dengan interaktif melakukan perubahan dalam visualisasi data. Bentuk lain yang masih juga perlu riset lanjutan adalah semakin besarnya ukuran dan variasi data yang ingin diproses yang mengharapkan teknik dengan kecepatan yang bagus dan akurasi tinggi. Isu penting lainnya adalah masalah keamanan yang mencakup kerahasiaan, integritas data serta kepemilikan data. Juga tidak kalah menarik adalah *distributed data mining* dan *mining multi-agent system*, serta penerapan *data mining* dalam bio-informatika dan lingkungan.

## Referensi


[1] Achtert E., Böhm C., Kriegel H.P., Kröger P., 2005, *Online hierarchical clustering in a data warehouse environment*, Proc. of the 5th international conference on data mining (ICDM), Houston, TX, 10–17

[2] Agrawal, R, Imielinski, T. and Swami, A., 1993, *Mining association rules between sets of items in large databases*. Proc. ACM SIGMOD International Conference on Management of Data, ACM Press, 207-216

[3] Böhm C, Kailing K, Kröger P, Zimek A., 2004, *Computing clusters of correlation connected objects*, Proc. of the SIGMOD conference, Paris, France, 455–466

[4] Cronea S.F., Lessmann S., Stahlbock R., 2005, *The impact of preprocessing on data mining*: *an evaluation of classifier sensitivity in direct marketing*, Eur J Oper Res, Vol 173, No 3, 781–800

[5] Domeniconi C., Gunopulos D., 2001, *Incremental support vector machine construction*, Proc. of the 1st international conference on data mining (ICDM), San Jose, CA, 589–592

[6] Fayyad, U., Piatetsky-Shapiro, H., Smyth, P., 1996, *The KDD process for extracting useful knowledge from volumes of data*, Comm. of the ACM, Vol 39, No 11, 27 - 34

[7] Fayyad U., Piatetsky-Shapiro G., Smyth P., 1996, *Knowledge discovery and data mining*: *Towards a unifying framework*, Proc. of the 2nd, ACM international conference on knowledge discovery and data mining (KDD), Portland, OR, 82–88

[8] Gaber M.M., Zaslavsky A, Krishnaswamy S., 2005, *Mining data streams: a review*,

SIGMOD Records Vol 34 No.2

[9] Han, J., Pei, J., Yin, Y., 2000, *Mining frequent patterns without candidate generation*, Proc. ACM SIGMOD Conference on Management of Data (SIGMOD). ACM Press, 1-12

[10] Han J, Kamber M., 2001, *Data mining: concepts and techniques*, Academic Press, San Diego

[11] Hand, D.J., Yu, K., 2001, *Idiot's Bayes: Not So Stupid After All*? Internat. Statist. Rev. Vol 69, 385-398

[12] Hastie, T., Tibshirani, R., 1996, *Discriminant Adaptive Nearest Neighbor Classifcation*, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol 18, No. 6, 607-616

[13] Jörnsten R., Wang H.Y., Welsh W.J., Ouyang M., 2005, *DNA microarray data imputation and significance analysis of differential expression*, Bioinformatics Vol 21, No. 22, 4155–4161

[14] Kailing K., Kriegel H.P., Pryakhin A., Schubert M., 2004, *Clustering multi-represented objects with noise*, Proc. of the 8th pacific-asia conference on knowledge discovery and data mining (PAKDD), Sydney, Australia, 394–403

[15] Kanellopoulos Y., Dimopulos T., Tjortjis C., Makris C., 2006, *Mining source code elements for comprehending object-oriented systems and evaluating their maintainability*, SIGKDD Explorations Vol 8, No. 1, 33–40

[16] Keogh E, Kasetty S., 2002, *On the need for time series data mining benchmarks: A survey and empirical demonstration*, Proc. of the 8th ACM international conference on knowledge discovery and data mining (SIGKDD), Edmonton, Alberta, 102–111

[17] Kittler J., Hatef M., Duin R., Matas J., 1998, *On combining classifiers*, IEEE Trans Pattern Analysis and Machine Intelligence, Vol 20, No. 3, 226–239

[18] Kriegel H-P, Kröger P, Pryakhin A, Schubert M., 2004, *Using support vector machines for classifying large sets of multi-represented objects*, Proc. of the 4th SIAM international conference on data mining (SDM), Orlando, FL, 102–113

[19] Kriegel H.P., Pryakhin A., Schubert M., 2005, *Multi-represented kNN-classification for large class sets*, Proc. of the 10th international conference on database systems for advanced applications (DASFAA), Beijing, China, 511–522

[20] Kriegel H-P, Pryakhin A, Schubert M., 2006, *An EM-approach for clustering multi-instance objects*. Proc. of the 10th pacific-asia conference on knowledge discovery and data mining (PAKDD), Singapore, 139–148

[21] Kriegel**,** H-P., et al., 2007, *Future trends in data mining*, Data Min Knowl Disc. DOI 10.1007/s10618-007-0067-9, No. 5, 87–97

[22] Liu, B., Hsu, W., Ma, Y. M., 1998, *Integrating classifcation and association rule mining,* Proc. KDD-98, ACM press, 80-86

[23] Liu C, Yan X, Yu H, Han J, Yu PS., 2005, *Mining behaviour graphs for "backtrace" of noncrashing bugs*, Proc. of the 5th SIAM international conference on data mining (SDM), Newport Beach, CA, 286–297

[24] Quinlan, J. R., 1993, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers Inc.

[25] MacQueen, J. B., 1967, *Some methods for classifcation and analysis of multivariate observations*, Proc. 5*th* Berkeley Symp. Mathematical Statistics and Probability, University of California Press, 281-297

[26] Ramon J., Bruynooghe M., 2001, *A polynomial time computable metric between points sets*, Acta Informatica Vol 37, 765–780

[27] Srikant R, Vu Q, Agrawal R., 1997, *Mining association rules with item constraints*. Proc. of the 3rd ACM international conference on knowledge discovery and data mining (KDD), Newport Beach, CA, 67–73

[28] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB, 2001, *Missing value estimation methods for DNA microarrays*, Bioinformatics, Vol 17, No. 6, 520–525

[29] Washio T, Motoda H., 2003, *State of the art of graph-based data mining*. SIGKDD Explorations Newsletter Vol 5, No. 1, 59–68

[30] Weidmann N., Frank E., Pfahringer B., 2003, *A two-level learning method for generalized multi-instance problems,* Proc. of the 14th european conference on machine learning (ECML), Cavtat-Dubrovnik, Croatia, 468–479

[31] Wu X., Kumar V., 2009, *The top ten algorithm in data mining*, Chapman & Hall/CRC, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

[32] Yan, X., Han, J., 2002, *gSpan: Graph-Based Substructure Pattern Mining*, Proc. IEEE International Conference on Data Mining (ICDM '02), IEEE, 721-724

[33] Zhang, T., Ramakrishnan, R., Livny, M., 1996, *BIRCH: an efficient data clustering method for very large databases*, Proc. ACM SIGMOD international Conference on Management of Data, ACM Press, 103-114