1. Perkenalan

1.1 Motivasi

Dalam beberapa tahun terakhir, jumlah data yang dihasilkan oleh manusia dan mesin telah meningkat pesat
ditingkatkan. Kata kunci seperti Big Data, Internet of Things, dan Machine-to-Machine
Komunikasi umumnya terdengar di media arus utama dan menunjukkan betapa lazimnya topik tersebut
adalah. Manfaat potensial dari sejumlah besar data adalah pengetahuan yang lebih luas dapat diperoleh
menganalisis data. Jenis analisis ini merupakan pendekatan bottom-up dan banyak organisasi yang menerapkannya
menerapkan pendekatan ini. Pendekatan top-down dimulai dari prinsip-prinsip umum dan cara kerja
turun untuk mengembangkan model suatu proses. Tesis ini menyelidiki arsitektur yang menggabungkan
pendekatan bottom-up dengan pendekatan top-down dan meninjau perangkat lunak yang dapat mewujudkan hal tersebut
arsitektur gabungan.

1.2 Kerangka

Kerangka kerja yang diusulkan dari metodologi gabungan telah disediakan, yang akan
dibahas secara rinci dalam bab 3. Kerangka kerja ini terdiri dari modul top-down dan bottom-up
modul beserta koneksi antara keduanya dan blok lainnya. Kerangkanya akan menjadi
dianalisis untuk menentukan bagian mana dari kerangka yang diusulkan yang dapat diterapkan dan mana yang dapat diterapkan
tidak, serta bagian mana yang mampu diotomatisasi. Kerangka kerja yang dihasilkan kemudian akan menjadi
digunakan untuk merancang arsitektur perangkat lunak yang dapat digunakan untuk membangun kerangka kerja.

1.3 Arsitektur dan Alat Perangkat Lunak

Setelah dilakukan analisis terhadap kerangka top-down bottom-up, maka akan dihasilkan kerangka kerja tersebut
digunakan untuk merancang arsitektur perangkat lunak. Alat penambangan data dan dinamika sistem yang ada akan digunakan
dimanfaatkan untuk mengusulkan implantasi perangkat lunak dari arsitektur perangkat lunak. Kumpulan fitur dan
kemampuan otomatisasi penambangan data dan alat dinamika sistem akan dianalisis untuk menentukan
alat mana yang dapat diterapkan pada implementasi perangkat lunak.

1.4 Dinamika Sistem dan Data Mining

Dinamika sistem dan penambangan data merupakan implementasi pendekatan top-down dan bottom-up masing-masing. Keduanya banyak digunakan dalam bisnis. Salah satu contoh data mining dalam bisnis adalah menentukan subkelompok calon pelanggan mana yang akan diiklankan. Sebuah perusahaan dapat menganalisisnya database pelanggan untuk menentukan tipe orang mana yang paling umum. Mengetahui hal ini perusahaan dapat menargetkan orang-orang seperti itu untuk iklan daripada mencakup semua tipe.

Dinamika sistem sering digunakan untuk memodelkan kebijakan suatu perusahaan. Contoh sederhananya adalah memodifikasi kebijakan persediaan suatu perusahaan. Berbagai kebijakan inventaris dapat disimulasikan lihat bagaimana perubahan tersebut akan mempengaruhi inventaris dan keseluruhan rantai pasokan selama periode waktu tertentu. A Model dinamika sistem dapat dikemas sebagai “simulator penerbangan” untuk memungkinkan manajer bereksperimen dengan menyesuaikan parameter dan kebijakan serta melihat bagaimana sistem berperilaku.

Metode operasional kedua sistem berbeda. Penambangan data digunakan dalam pengaturan langsung di mana data baru diproses secara terus menerus. Biasanya juga sangat otomatis jika ada sedikit atau tidak ada interaksi manusia yang diperlukan untuk mengoperasikan sistem penambangan data. Kasus penggunaan utama untuk dinamika sistem di sisi lain, adalah untuk lingkungan pengujian simulasi interaktif. Seorang pengguna bisa mengatur berbagai parameter model dan kemudian menjalankan simulasi untuk menghasilkan deret waktu keluaran. Kerangka kerja gabungan dan arsitektur perangkat lunak yang dihasilkan akan menjadi kombinasi dari keduanya. Kerangka kerja ini akan beroperasi sebagai sistem otomatis, melakukan simulasi, dan memproduksi keluaran deret waktu pada interval waktu yang telah ditentukan.

1.5 Tujuan

Sementara penambangan data dan dinamika sistem digunakan dalam bisnis, kerangka kerja gabungan sebagai dijelaskan di sini, tidak akan digunakan untuk keperluan bisnis. Sebaliknya kasus penggunaannya adalah untuk memantau dan meramalkan berbagai peristiwa yang terjadi di seluruh dunia. Kasus penggunaan lainnya adalah menganalisis sejarah peristiwa untuk membantu memahami faktor-faktor penting dari peristiwa tersebut. Kerusuhan adalah salah satu contoh peristiwa. Bencana ini sering terjadi di seluruh dunia dan menyebabkan kerusakan besar pada kota seperti Kerusuhan Inggris 2011.

Tujuan dari contoh ini adalah untuk melihat apakah kerangka kerja gabungan dapat meramalkan terjadinya kerusuhan. Ini akan memungkinkan pihak berwenang untuk mengalokasikan sumber daya dan mengambil tindakan untuk membantu mencegah kerusuhan atau mempersiapkan kerusuhan.

Penelitian bertahun-tahun diperlukan untuk menguji teori ini. Tesis ini akan memberikan kerangka dan arsitektur perangkat lunak untuk memungkinkan dimulainya penelitian.

1.6 Ringkasan Bab

Bab 2 akan memberikan gambaran umum tentang data mining dan dinamika sistem. Ini akan mencakup mereka kekuatan dan keterbatasan serta proses penerapan kedua metode tersebut.

Bab 3 akan menganalisis kerangka kerja yang menggabungkan metode top-down dan bottom-up. Itu akan membahas berbagai bagian kerangka dan modifikasi apa pun yang dibuat.

Bab 4 akan mengeksplorasi perangkat lunak yang tersedia dalam penambangan data dan dinamika sistem. A sejumlah alat komersial dan sumber terbuka akan dianalisis untuk set fiturnya untuk digunakan dalam arsitektur perangkat lunak.

Bab 5 akan membahas arsitektur perangkat lunak dan data mining serta alat dinamika sistem itu dapat digunakan untuk konstruksi arsitektur perangkat lunak.

Bab 6 akan memberikan ringkasan tesis.

2 Ikhtisar dari atas ke bawah dari bawah ke atas

2.1 Dari bawah ke atas

2.1.1 Ikhtisar

Analisis bottom-up (BU) terdiri dari analisis berbagai bentuk data, seperti angka, teks, gambar, video, suara, dll untuk menemukan hubungan dan pola untuk memperoleh pengetahuan darinya data. Analisis bottom-up telah mengalami pertumbuhan yang luar biasa selama bertahun-tahun. Jenis ini analisis telah menyebar ke sejumlah sektor termasuk keuangan, bisnis, penegakan hukum, dan pertahanan adalah beberapa di antaranya. Penambangan data, pembelajaran mesin, dan data besar adalah hal yang umum

perwakilan analisis bottom-up. Tesis ini akan fokus pada penggunaan data mining ketika
mengacu pada analisis bottom-up.

2.1.2 Penambangan Data, Pembelajaran Mesin

“Data mining adalah proses eksplorasi dan analisis, dengan cara otomatis atau semi-otomatis,
sejumlah besar data untuk menemukan pola dan aturan yang bermakna.”[1]

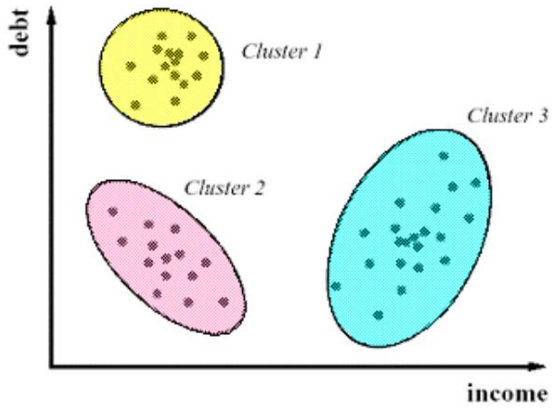
Kutipan di atas memberikan penjelasan sederhana tentang data mining. Ini adalah sistem tempat data berada
dikumpulkan, disimpan, dan kemudian dianalisis dengan metode otomatis.

Salah satu contoh kasus bisnis data mining adalah menentukan apakah seseorang akan mengajukan permohonan kartu kredit
jika diberikan iklan untuk kartu. Untuk memahami tipe orang seperti apa yang mungkin melamar
untuk kartu kredit, perusahaan kartu kredit menyimpan atribut setiap orang yang bergabung. Itu
atribut dapat mencakup usia, jenis kelamin, pekerjaan, pendapatan, status perkawinan, alamat rumah, dll. A
sejumlah metode analisis dapat digunakan untuk menentukan kombinasi atribut apa yang a
orang tersebut kemungkinan besar akan mengajukan permohonan kartu kreditnya. Bagian analisis ini adalah tempat pembelajaran mesin
diimplementasikan. Data masa lalu digunakan untuk membantu membuat algoritme yang mempelajari kombinasi apa
memberikan probabilitas tertinggi seseorang akan mengajukan permohonan kartu kredit. Algoritma yang dipelajari adalah
disebut algoritma fit. Setelah algoritma fit dikembangkan, algoritma ini digunakan dalam penambangan data
sistem untuk operasi langsung. Kembali ke contoh, alih-alih mengirimkan jutaan secara acak
aplikasi kartu kredit orang, sistem penambangan data dapat mengurai melalui database nonkredit
pemegang kartu dan menguji atribut setiap orang terhadap algoritma kecocokan. Itu adalah
bertekad untuk kemungkinan mengajukan permohonan kartu kredit dapat menjadi penerima permohonan kartu kredit.

Pembelajaran mesin adalah metode pembelajaran dari data secara otomatis. Tugas dari
menentukan apakah suatu email adalah spam atau bukan akan digunakan sebagai contoh untuk menjelaskan pembelajaran mesin
lebih jauh. Sebuah program atau model, yang berisi sejumlah parameter, dapat mengetahui apakah sebuah email adalah
spam atau tidak melalui paparan berulang terhadap email spam dan email nonspam. Setiap paparan akan
menyesuaikan parameter untuk meningkatkan kinerja. Ini adalah proses otomatisasi di mana model itu sendiri berada
menyesuaikan parameter untuk meningkatkan kinerja. Setelah kinerja berada pada tingkat yang dapat diterima
model dianggap cocok. Tesis ini akan menyebut model fit ini sebagai model data mining.

Algoritma utama yang digunakan untuk pembelajaran mesin adalah klasifikasi, pengelompokan, regresi atau prediksi, dan aturan asosiasi. Untuk klasifikasi, tujuannya adalah untuk mengklasifikasikan sesuatu menjadi a kumpulan kategori yang telah ditentukan. Misalnya jika seseorang diberikan iklan kredit kartu, apakah orang tersebut akan mendaftar atau tidak. Hanya ada dua kemungkinan yang ada untuk kasus ini.

Clustering akan mengelompokkan data ke dalam kategori serupa dimana jumlah kategori tersebut belum ada telah ditentukan sebelumnya. Titik data, sering disebut sebagai catatan, yang serupa dikelompokkan menjadi satu. A Contoh bisnis clustering adalah membandingkan jumlah pendapatan dan hutang suatu orang. Titik-titik yang berdekatan akan dikelompokkan. Gambar di bawah menunjukkan plot titik-titiknya dan tiga cluster yang dihasilkan.



Gambar 1 Contoh Cluster [2]

Regresi akan memprediksi suatu nilai, seperti harga rumah bergantung pada atribut seperti itu umur rumah, jumlah kamar, lingkungan sekitar, dll. Dengan menganalisis harga rumah dengan atributnya yang telah terjual selama bertahun-tahun sebuah model dapat diciptakan. Modelnya kemudian bisa digunakan untuk memprediksi berapa harga jual suatu rumah berdasarkan atribut rumah tersebut.

Terakhir, aturan asosiasi menentukan objek apa yang biasanya dikaitkan satu sama lain. Super pasar tertarik pada jenis data ini. Mereka tertarik untuk mengetahui barang apa saja yang lainnya biasanya dibeli dengan hotdog.

Keempat kategori ini terbagi dalam dua kelompok umum pembelajaran yang diawasi dan pembelajaran tanpa pengawasan.
Pembelajaran yang diawasi, yang mencakup algoritma klasifikasi dan regresi memberikan umpan balik. Jika
klasifikasi suatu catatan benar atau salah umpan balik dapat digunakan untuk pembelajaran.
Pembelajaran tanpa pengawasan, yang mencakup pengelompokan dan aturan asosiasi tidak menyediakan apa pun
masukan. Oleh karena itu pembelajaran tidak dapat diperoleh dengan mengolah data sejarah. Misalnya,
clustering akan mengelompokkan titik data yang serupa tetapi karena jumlahnya tidak ditentukan sebelumnya
klasifikasi yang termasuk di dalamnya tidak ada umpan balik untuk mengetahui apakah cluster tersebut benar atau tidak.

Masing-masing dari empat kategori dapat diimplementasikan melalui sejumlah algoritma. Misalnya
klasifikasi dimungkinkan melalui pohon keputusan, jaringan saraf, pengklasifikasi Bayesian, dan Dukungan
Mesin Vektor adalah beberapa di antaranya. Saat model pembelajaran mesin klasifikasi sedang dibuat
membangun sejumlah algoritma akan diuji untuk melihat mana yang berkinerja terbaik. Sama
proses dilakukan dengan kategori lainnya juga.

2.1.3 Alur Penambangan Data

Alur untuk penambangan data dijelaskan di bawah ini.

Langkah 1: Kembangkan pemahaman tentang tujuan proyek penambangan data. Apakah tujuannya satu
upaya waktu atau akankah itu berurusan dengan eksekusi berkali-kali.

Langkah 2: Dapatkan kumpulan data yang akan digunakan dalam analisis. Jika jumlah datanya sangat besar maka
mungkin cukup untuk mengambil sampel sebagian data secara acak. Seribu catatan biasanya cukup
untuk membuat model [3]. Data mungkin perlu ditanyakan dari beberapa database secara internal dan
secara eksternal.

Langkah 3: Jelajahi, bersihkan, dan proses awal data. Data mungkin banyak namun seringkali tidak bersih.
Catatan yang hilang dari kumpulan data adalah hal biasa. Keputusan harus dibuat tentang bagaimana menangani kehilangan
data. Itu dapat diabaikan atau dirata-ratakan di antara catatan-catatan di sekitarnya. Data yang salah juga
umum. Untuk kasus ini jelas data yang salah dapat diperiksa. Misalnya jika yang diharapkan
nilai suatu catatan berada di antara dua nilai dan catatan tersebut berada di luar rentang maka catatan ini berlaku
data yang salah dan dapat diabaikan.

Langkah 4: Kurangi dan pisahkan variabel. Tidak semua variabel mungkin diperlukan. Pada langkah ini variabel yang tidak diperlukan dikeluarkan dari analisis. Semakin banyak variabel yang disertakan, semakin banyak CPU waktu akan diperlukan untuk pemrosesan. Oleh karena itu, yang ideal adalah menjaga jumlah variabel serendah mungkin mungkin. Misalnya, sebuah rumah mempunyai 20 atribut. Jika semua 20 atribut digunakan membuat model fit maka model fit harus menggunakan seluruh 20 atribut untuk setiap rekaman yang diprosesnya. Jika model kecocokan yang sama atau sedikit kurang akurat dapat dibuat hanya dengan 5 atribut, maka CPU beban akan jauh lebih sedikit.

Selain itu, beberapa variabel perlu dimodifikasi atau diubah. Misalnya jika suatu variabel adalah usia seseorang resolusinya mungkin terlalu bagus. Mungkin lebih mudah untuk menganalisis jika sejumlah usia rentang digunakan sebagai gantinya.

Terakhir, ketika pelatihan yang diawasi akan digunakan, data harus dibagi menjadi tiga kelompok pelatihan, validasi, dan tes. Set pelatihan digunakan untuk melatih model. Setelah dilatih validasi dan set pengujian akan digunakan untuk melihat kinerjanya dengan kumpulan data yang berbeda.

Langkah 5: Pilih tugas penambangan data (regresi, pengelompokan).

Langkah 6: Gunakan algoritma untuk melakukan tugas. Ini biasanya memerlukan banyak upaya. Bermacam-macam kombinasi variabel serta beberapa varian dari algoritma yang sama akan diuji.
Algoritme yang menjanjikan dapat diuji dengan kumpulan data validasi untuk melihat kinerjanya terhadap a kumpulan data baru.

Langkah 7: Interpretasikan hasil algoritma. Algoritma dari salah satu dari sekian banyak algoritma yang diuji pada langkah 6 perlu dipilih. Algoritme yang dipilih juga harus diuji terhadap kumpulan data pengujian untuk melihatnya bagaimana kinerjanya dengan kumpulan data baru lainnya. Pada titik ini algoritma telah disesuaikan tugas yang ada.

Langkah 8: Algoritma fit diintegrasikan ke sistem untuk digunakan dengan data nyata. Sistem akan melakukannya mengeksekusi algoritma fit terhadap record baru untuk membuat penentuan seperti apa