RESUME VISUALISASI DATA

Mirzan Yusuf Rabbani

2024-08-28

Visualisasi data, yang mengubah data abstrak menjadi visi fisik (misalnya, panjang, posisi, bentuk, warna, dan sebagainya), adalah cara yang ampuh untuk menyajikan cerita yang menarik dari yang menarik dari data kepada manusia yang lebih berorientasi pada visual.

Perkembangan Visualisasi Data Tidak diragukan lagi, visualisasi data telah membuat langkah besar di banyak bidang, yang dikontribusikan oleh banyak komunitas. Komunitas grafik komputer telah secara signifikan memajukan teknologi rendering visualisasi yang indah namun dapat ditafsirkan sendiri. Komunitas visualisasi memudahkan pengguna untuk menentukan dan berinteraksi dengan visualisasi. Komunitas basis data telah secara signifikan meningkatkan pengalaman pengguna dalam melihat dan berinteraksi dengan visualisasi data secara real time, bahkan untuk data yang besar.

Jalur Visualisasi Data:

1. Impor data adalah mengambil data yang diperlukan dari sumber data yang diinginkan.
2. Persiapan data adalah mempersiapkan data yang diimpor untuk visualisasi, misalnya dengan menormalkan nilai, mengoreksi entri yang salah, dan menginterpolasi nilai yang hilang.
3. Manipulasi data adalah memilih data yang akan divisualisasikan dan mungkin dengan operasi umum lainnya seperti penggabungan dan pengelompokan.
4. Pemetaan adalah memetakan data yang

diperoleh dari proses di atas ke primitif geometris bersama dengan atributnya.

5. Rendering adalah mengubah data geometris di atas menjadi representasi visual.

1. Spesifikasi Visualisasi menyediakan berbagai cara agar pengguna dapat menentukan apa yang mereka inginkan.
2. Pendekatan yang Efisien untuk Visualisasi Data agar dapat secara efektif melibatkan pengguna dalam pipeline yang berulang, proses pembuatan visualisasi data harus efisien dan terukur, terutama untuk dua komponen, “Manipulasi Data” dan “Pemetaan”.
3. Rekomendasi Visualisasi Data menentukan visualisasi dengan tepat adalah hal yang sulit, bahkan untuk para ahli, hanya karena pemahaman tentang data apa yang harus divisualisasikan, cerita apa yang harus diceritakan, dan bagaimana memvisualisasikannya adalah latihan coba-coba.

Survei Terkait Sebagian besar survei yang ada tentang visualisasi berfokus pada topik tertentu, seperti visualisasi grafik, visualisasi data terkait, visualisasi ontologi, visualisasi data berdimensi tinggi, visualisasi data temporal.Untuk spesifikasi visualisasi, berikan survei tentang klasifikasi, sumber data, media presentasi, dll., dari bahasa visualisasi. Untuk pendekatan yang efisien untuk visualisasi data, pertimbangkan bagaimana mengintegrasikan database, visualisasi data, dan analisis data sehingga pengguna dapat dengan mudah bekerja dalam satu sistem, tetapi tanpa diskusi untuk efisiensi. Untuk rekomendasi visualisasi data, meskipun sudah banyak karya tentang sistem rekomendasi dan karya-karya tentang rekomendasi untuk tugas yang berbeda.

2 Spesifikasi visualisasi

2.1 Spesifikasi visualisasi data

Secara umum, bahasa visualisasi data terdiri dari tiga bagian:

- Data
- Catatan: data yang perlu divisualisasikan.
- Transformasi: operasi-seperti group, bin, filter, dan sorting-digunakan untuk mengubah catatan data.
- Tanda (atau petunjuk visual)
- Jenis: representasi visual untuk catatan data, seperti batang, garis, atau titik.
- Ukuran: lebar, tinggi visualisasi.
- Legenda: informasi legenda.
- Lain-lain: properti lain, seperti lebar dan warna batang.
- Pemetaan: memetakan data ke tanda yang sesuai. Operasi visual berbasis GUI biasanya diterjemahkan ke dalam bahasa visualisasi data.

2.2 Kategorisasi bahasa visualisasi data

Strategi yang umum digunakan untuk mengkategorikan bahasa visualisasi data didasarkan pada ekspresifitasnya, seperti yang ditunjukkan pada sisi kiri Gambar.

Bahasa Tingkat Rendah adalah bahasa yang dibutuhkan pengguna untuk menentukan semua elemen pemetaan.

Bahasa Tingkat Tinggi merangkum detail konstruksi visualisasi, seperti fungsi pemetaan, serta beberapa properti untuk tanda seperti ukuran kanvas, legenda, dan properti lainnya.

2.3 Operasi visual berbasis GUI

Dibandingkan dengan menggunakan bahasa visualisasi deklaratif untuk menentukan visualisasi, cara yang lebih ramah pengguna dalam memberikan spesifikasi adalah dengan mengikuti “prinsip manipulasi langsung”, sebuah konsep yang digunakan secara luas dalam aspek interaksi manusia-komputer.

Visualisasi Data Interaktif Rasionalitas, di balik visualisasi data interaktif adalah bahwa dalam banyak kasus, visualisasi data adalah proses eksplorasi, di mana pengguna perlu terus menyempurnakan spesifikasi dari visualisasi yang sedang dieksplorasi hingga mendapatkan visualisasi yang diinginkan dalam proses eksplorasi.

1. Stepwise Query Refinement Polaris dan Tableau menyediakan templat bagan untuk menunjukkan visualisasi multidimensi.
2. Navigasi Segi Banyak DeepEye mendukung navigasi segi banyak untuk membantu pengguna menjelajahi ruang desain visualisasi.

Catatan Meskipun alat interaktif berbasis GUI menyediakan antarmuka sederhana untuk membangun visualisasi umum dengan cepat.

2.4 Spesifikasi yang kurang spesifik

Visualisasi tidak ada artinya jika tidak dapat memberikan wawasan tentang data. Untuk spesifikasi yang kurang spesifik, pengguna hanya memberikan beberapa “petunjuk”, dan merupakan tugas sistem visualisasi untuk menafsirkan input yang tidak dispesifikasi, dengan (mungkin) cara yang berbeda.

Jenis petunjuk pertama adalah “berbasis referensi”, di mana pengguna memberikan referensi visualisasi sebagai seed dan sistem menyarankan visualisasi berdasarkan referensi tersebut.

Jenis petunjuk kedua adalah “berbasis kata kunci”, dengan gaya Google. APT [88] menerima tujuan melihat data pengguna dari kolom yang diinginkan. Jenis petunjuk ketiga adalah “berbasis bahasa alami”, yang mempertimbangkan konteks input pengguna dan status sistem dalam siklus eksplorasi data, bukan hanya sekali tembak pada petunjuk “berbasis kata kunci”.

3 Pendekatan yang efisien untuk visualisasi data

3.1 Visualisasi data yang tepat

Penerjemahan Kueri Cara alami untuk menggunakan kembali banyak sistem (DBMS) yang sudah matang adalah dengan menerjemahkan kueri visualisasi ke kueri yang diterima oleh sistem tersebut.

Mengintegrasikan Sistem Visualisasi dengan DBMS Meskipun menggunakan penerjemahan kueri adalah hal yang alami, ada beberapa kelemahan.

Penyimpanan Kolom Dalam manajemen data, faktor kinerja utama adalah tata letak data, misalnya, tata letak berbasis baris dan berbasis kolom, yang mungkin memiliki perbedaan kinerja yang sangat besar untuk aplikasi OLAP.

Indeks banyak digunakan untuk meningkatkan kinerja pencarian dengan cara mengurangi jumlah record/baris dalam tabel yang perlu diperiksa.

Komputasi Paralel telah banyak digunakan untuk pemrosesan kueri dalam sistem visualisasi data.

Prediksi dan Prapengambilan Salah satu langkah penting dalam visualisasi data adalah eksplorasi data-pengguna secara terus menerus menelusuri visualisasi yang mereka minati untuk mendapatkan gambaran tentang apa yang akan divisualisasikan.

1. Visualisasi yang Sedang Dieksplorasi. XmdvTool mengelompokkan tupel dalam granularitas yang berbeda untuk mendukung navigasi hirarkis pengguna.
2. Data Historis. Selanjutnya, kita akan membahas teknik yang memanfaatkan lintasan historis untuk melakukan prefetching, mengusulkan tiga strategi untuk melakukan prefetching data berdasarkan data historis:

- arah: pilih arah yang paling mungkin berdasarkan pelacakan lintasan pengguna sebelumnya,
- fokus: pilih arah dengan daerah yang panas, dan
- vektor: pilih arah berdasarkan vektor lintasan pergerakan pengguna, dalam bentuk <posisi awal, lebar, level>, di mana posisi awal adalah lokasi awal dan

orientasi pergerakan, lebar adalah jarak pergerakan, level adalah agregasi hirarki data yang dieksplorasi dalam gerakan.

Ada dua tahap prediksi data:

- Memprediksi fase analisis: memprediksi fase eksplorasi pengguna dengan model Support Vector Machine (SVM).
- Memprediksi ubin data: menggunakan strategi yang sesuai untuk merekomendasikan data yang telah diambil sebelumnya.

Studi Kasus menggunakan Kyrix dan Tableau

Kyrix [108] adalah sistem visualisasi data interaktif yang dapat diskalakan.

1. Spesifikasi Visualisasi di Front end. Ada dua abstraksi dalam bahasa spesifikasi visualisasi dari Kyrix: kanvas dan lompatan.

2. Pendekatan yang Efisien untuk Visualisasi Data di Back-end. Ada dua peningkatan penting dalam Kyrix: mengambil granularitas dan pengindeksan.

TDE adalah mesin data yang disesuaikan untuk visualisasi di Tableau 6.0. TDE mengoptimalkan mesin data terutama dalam perspektif berikut.

1. Penyimpanan dan Kompresi Berorientasi Kolom. Karena biaya I/O yang tinggi dari database Tableau sebelumnya, Firebird dan data visualisasi yang biasanya disimpan dalam kolom yang berbeda, teknik penyimpanan dan kompresi berorientasi kolom telah dirancang untuk mengatasi masalah ini di TDE.
2. Pengurutan Ulang Operator. Operator pilihan dan operator dengan kolom tunggal yang dikompresi didorong ke bawah dalam pohon rencana kueri SQL.
3. Pengurangan Kardinalitas. Untuk kolom dengan kolom kardinalitas tinggi, TDE secara otomatis mengubah kolom-kolom ini ke hirarki yang lebih tinggi.
4. Dukungan Visualisasi Lainnya. TDE menyediakan informasi domain (misalnya, kardinalitas, nilai maksimum dan minimum domain) dari kolom.

3.2 Perkiraan visualisasi data

Ketika volume data tumbuh secara eksponensial, modul pemrosesan data tradisional tidak dapat memberikan hasil pemrosesan interaktif yang cepat.

Berbasis AQP Cara yang mudah untuk menghasilkan visualisasi perkiraan dalam waktu interaktif adalah dengan memanfaatkan teknik AQP.

Berbasis Pengambilan Sampel Tambahan Beberapa karya mencoba menghubungkan teknik kueri data tambahan dengan visualisasi data.

Berbasis Persepsi Manusia Terkadang, meningkatkan ukuran sampel tidak selalu meningkatkan kualitas visualisasi.

3.3 Visualisasi data progresif

Banyak karya dalam visualisasi data perkiraan menghasilkan hasil visualisasi yang progresif kepada pengguna.

Range-Based Binning imMens menyediakan visualisasi resolusi yang berbeda dengan mengubah ukuran bin.

Binning Berbasis Rentang dan Konten Karya ini menyediakan dua struktur pohon untuk eksplorasi hirarki: HETree-R (HETree berbasis rentang) dan HETree-C (HETree berbasis konten).

4 Rekomendasi visualisasi

4.1.1 Spesifikasi yang tidak lengkap

Proses visualisasi data bersifat berulang, dan masalah utama para praktisi adalah mereka harus terlibat dalam setiap langkah untuk membuat beberapa modifikasi.

Ikhtisar Solusi Secara umum, untuk menyelesaikan semua masalah di atas, sistem rekomendasi visualisasi harus terlebih dahulu menghitung semua visualisasi yang mungkin dan kemudian merekomendasikan visualisasi dengan peringkat teratas.

Memangkas Visualisasi yang Tidak Berarti Untungnya, ada banyak sinyal (atau batasan) baik dari pengguna atau dari kebijaksanaan tradisional yang dapat digunakan untuk memangkas visualisasi yang “buruk”.

– Batasan yang ditentukan pengguna. Pengguna dapat menentukan elemen visualisasi yang diminati seperti kolom atau catatan data. – Batasan yang diberikan oleh pakar. Beberapa kombinasi variabel, transformasi, dan penyandian visual mungkin tidak menghasilkan visualisasi yang valid.

4.1 Rekomendasi berdasarkan spesifikasi

4.1.1 Spesifikasi yang tidak lengkap

Sistem rekomendasi visualisasi dengan spesifikasi kosong tidak memerlukan masukan dari pengguna, sedangkan sistem rekomendasi dengan spesifikasi parsial menerima masukan spesifikasi sebagian elemen visualisasi pengguna untuk visualisasi yang diinginkan.

Pemeringkatan visualisasi berbasis aturan

Sistem rekomendasi berbasis aturan mengurutkan kandidat visualisasi berdasarkan aturan yang telah ditetapkan. Aturan Statistik Kerangka kerja peringkat berdasarkan fitur adalah sistem rekomendasi berbasis aturan statistik. Aturan Perseptual Kerangka kerja peringkat berdasarkan fitur hanya dapat membedakan antara jenis visualisasi yang sama (misalnya, histogram, plot kotak) dengan metrik statistik tunggal, sementara Voyager memeringkat jenis visualisasi yang berbeda dengan skor efektivitas perseptual dengan mempertimbangkan jenis data, kardinalitas, preferensi visual manusia, dan sebagainya.

Pemeringkatan visualisasi berbasis pembelajaran mesin Dengan pesatnya perkembangan machine learning dan deep learning, semakin banyak sistem yang berfokus pada rekomendasi visualisasi berbasis machine learning. Pembelajaran dengan Kendala Lunak Draco mengekspresikan preferensi dengan kendala lunak, dan kendala lunak (misalnya, lebih baik untuk nilai temporal menggunakan tipe pengkodean: Sumbu X) . ditentukan oleh persepsi manusia. Belajar dengan Contoh Batasan dalam Draco ditentukan sebelumnya ke sistem oleh pengguna atau pengembang, bukan dipelajari oleh mesin.

4.1.2 Spesifikasi berbasis referensi

Berbasis deviasi SeeDB merekomendasikan visualisasi berdasarkan deviasi dengan beberapa visualisasi referensi.

Berbasis anomali Profiler merekomendasikan visualisasi yang dapat membedakan anomali dalam visualisasi utama.

Berbasis kesamaan/jarak Zenvisage mencoba menemukan visualisasi menarik lainnya ketika pengguna memberikan tren, pola, atau wawasan yang diinginkan.

4.2 Rekomendasi berbasis perilaku Sistem rekomendasi berbasis perilaku menangkap perilaku pengguna sebagai input, kemudian menyimpulkan tugas yang diinginkan pengguna dan merekomendasikan visualisasi yang berguna berdasarkan tugas mereka.

4.3 Rekomendasi yang dipersonalisasi Sistem rekomendasi yang dipersonalisasi menangkap perilaku historis pengguna sebagai masukan untuk merekomendasikan visualisasi yang sesuai dengan minat pengguna. Linear Model VizDeck memberikan hasil rekomendasi visualisasi yang dipersonalisasi dengan melatih model linear untuk setiap pengguna menggunakan perilaku historis mereka.

Penyaringan Kolaboratif Selain melatih model untuk setiap pengguna, ada banyak teknik lain dalam sistem rekomendasi yang dipersonalisasi. Mengusulkan tiga metode untuk rekomendasi visualisasi yang dipersonalisasi.

1. Pemfilteran Kolaboratif. VizRec membangun sebuah matriks A berukuran m × n.
2. Pemfilteran berbasis konten. Untuk pengguna yang baru mengenal sistem, rekomendasi berbasis CF tidak berlaku.
3. Pemfilteran Hibrida. Metode hibrida dari dua metode di atas akan membawa sejumlah manfaat.

5 Arah penelitian lainnya

5.1 Persiapan data untuk visualisasi data

Data kehidupan nyata biasanya kotor, dan memvisualisasikan data yang kotor dapat menyesatkan pengguna.

– Analisis Bagaimana-Jika untuk Pencilan: Scorpion memungkinkan pengguna untuk secara manual menentukan outlier dari hasil kueri agregasi. –Mengevaluasi Visualisasi dengan Data yang Hilang: melakukan studi crowdsourced untuk mengukur faktor-faktor yang memengaruhi akurasi respons, kualitas data, dan kepercayaan diri dalam interpretasi untuk data deret waktu dengan nilai yang hilang. – Mendeteksi visualisasi yang bias. Visualisasi yang tampaknya bagus mungkin sebenarnya bias; oleh karena itu, diperlukan untuk mendeteksi visualisasi tersebut secara otomatis. – Pembersihan data yang sadar tugas. Secara intuitif, akan lebih mudah untuk membersihkan kumpulan data jika tugas penargetan diketahui.

5.2 Tolak ukur visualisasi data

Tolak ukur harus sesuai dengan tugas analisis visual, menyediakan jejak dan data yang dapat digunakan kembali, dan dalam hal rekomendasi, memiliki cakupan dan kualitas label yang tinggi.

– Sebuah karya penelitian VizNet telah menyajikan sebuah korpus berskala besar dengan lebih dari 31 juta set data yang dikumpulkan dari repositori data terbuka dan galeri visualisasi online. – Kategorisasi visualisasi. Untuk

ImageNet, mudah untuk menetapkan kategori, seperti “balon” atau “stroberi”, karena tugas klasifikasi lebih mudah. Tidak jelas bagaimana mendefinisikan kategori yang serupa untuk visualisasi dalam tingkat konseptual, seperti “tren” atau “distribusi”. – Data pelatihan. Dengan asumsi kategori dapat disediakan, masih ada tugas yang menakutkan untuk memberi label pada visualisasi, dan setiap visualisasi mungkin memiliki beberapa label.

5.3 Visualisasi data untuk aplikasi yang berhubungan dengan basis data aplikasi

Dengan perkembangan teknik visualisasi yang cepat, ada lebih banyak peluang tentang penggunaan visualisasi data untuk aplikasi yang berhubungan dengan database.

– Visualisasi data untuk penemuan data. Penemuan data, masalah menemukan set data yang menarik untuk aplikasi tertentu dari data lake dengan ribuan atau jutaan silo data, tetap menjadi masalah yang sulit dipecahkan. – Visualisasi data untuk debugging data. Salah satu masalah yang baru-baru ini diangkat oleh sistem Data Civilizer adalah data debugging, di mana output dari sebuah alur kerja data analytics salah bukan karena adanya bug pada program.