UNIVERZITA KOMENSKÉHO V BRATISLAVE
Fakulta matematiky, fyziky a informatiky

Vizuálna analýza dát

Bakalárska práca

Eva Paľovičová

Študijný odbor: 9.2.1 Informatika

Vedúci bakalárskej práce:
Mgr. Matej Novotný

Bratislava 2009

Čestné prehlásenie

Vyhlasujem, že som bakalársku prácu vypracovala samostatne s použitím uvedenej odbornej literatúry.

Bratislava, máj 2009 .....

Abstrakt

Práca poskytuje teoretický prehľad z oblasti vizualizácie informácií. Popisuje spôsoby vizuálnej reprezentácie dát, najčastejšie zobrazované dátové typy, spôsoby interakcie užívateľa s vizualizačnými systémami a princíp hľadania informácie. Tiež podrobne rozoberá niektoré problémy pri návrhu moderných vizualizačných nástrojov a formuluje odporúčania pre ich prekonanie.

Kľúčové slová: vizualizácia, informácia, vizualizačný systém

Obsah

1 Vizualizácia

Vývoj zaznamenaný v oblasti hardware umožňuje dnešným počítačom uchovávať obrovské množstvo dát. Výskumníci z Univerzity v Berkeley odhadujú, že každý rok je vygenerovaný 1 exabyte (1 milión terabytov) dát, z čoho je veľká časť dostupná v digitálnej forme. Dáta sú zaznamenávané rôznymi senzormi a monitorovacími systémami. I informácie o jednoduchých transakciách každodenného života, ako platba kartou, či telefonovanie, sú poväčšine ukladané v počítačoch. Ľudia zbierajú tieto dáta, lebo veria že môžu byť potenciálnym zdrojom hodnotných informácií, avšak ich nájdenie v takom množstve údajov je náročná úloha. Dnešné systémy pre správu dát sú schopné zobraziť len malé časti z nich. Bez možnosti ich primerane preskúmať sa tieto údaje stávajú nepotrebnými, a databázy sú tak len smetiskom dát. Vzniká teda potreba informácie vhodne vizualizovať a poskytnúť tak človeku možnosť získať náhľad o uložených dátach.
Vizualizáciu teda môžme definovať ako kognitívnu aktivitu, ktorá prebieha v mysli človeka a jej výsledkom je sformovanie interného modelu, označovaný aj ako kognitívna mapa, ktorý umožňuje pochopiť význam zobrazených dát a vzťahy medzi nimi.

1.1 Úlohy vizualizácie

Cieľom vizualizácie býva niektorá z nasledujúcich úloh:

- **Preskúmanie dát** - v tomto prípade užívateľ nevie presne, čo hľadá. Prezerá si zobrazené dáta, a snaží sa nájsť určitý trend, pochopiť vzťahy medzi údajmi, vytvoriť si hypotézu o zobrazených dátach. Vhodné je poskytnúť užívateľovi viacero zobrazení prostredníctvom viacerých vizualizačných techník, čo môže pomôcť odhaliť zložité vzťahy medzi premennými. V tejto časti je taktiež veľmi dôležitá interakcia.

- **Potvrdenie hypotézy** - užívateľ má už vytvorenú hypotézu, ktorú potrebuje potvrdiť alebo vyvrátiť. Uplatnenie tu nachádzajú analytické nástroje, ktoré za použitia rôznych matematických vzťahov a štatistických výpočtov umožňujú overiť určité tvrdenie.

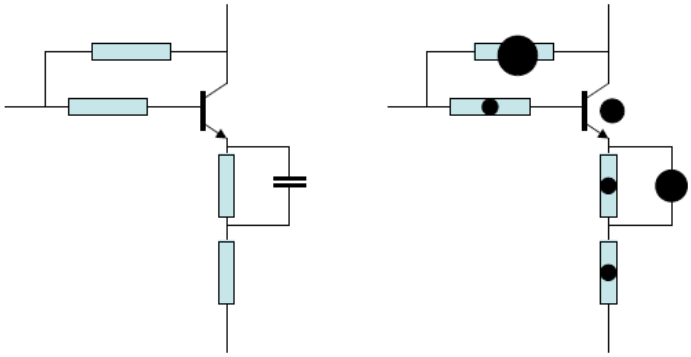
- **Zobrazenie informácie** - užívateľ má potvrdenú hypotézu a presne vie, akú informáciu chce podať ďalej. Potrebné je vybrať vhodnú vizualizačnú techniku vzhľadom k zvolenému typu dát, rovnako ako aj ďalšie parametre prezentácie.

1.2 Prostriedky vizualizácie

V nasledujúcom texte si predstavíme niekoľko prostriedkov, ktorými je možné reprezentovať informácie.

1.2.1 Veľkosť

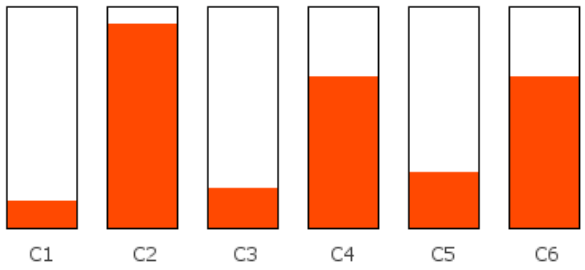
V prípadoch, kedy nepotrebujeme poznať presnú hodnotu premennej, stačí nám na jej zobrazenie použiť vhodnú veľkosť určitého grafického prvku. Na ilustráciu popíšeme príklad zobrazenia elektrického okruhu, kde je nad každým prvkom záujmu zobrazený kruh, ktorého veľkosť indikuje vplyv daného prvku na vlastnosti okruhu. Takéto symbolické zobrazenie číselných dát je vhodné v štádiu návrhu okruhu, kedy dizajnér väčšinou nepotrebuje vedieť, či citlivosť má hodnotu 5.2 alebo 5.3, ale potrebuje len zistiť, či je malá, stredná, alebo veľká. Takúto informáciu je možné nájsť rýchlejšie v spomenutom zobrazení, ako v číselnej reprezentácii.



1.2.2 Dĺžka a výška

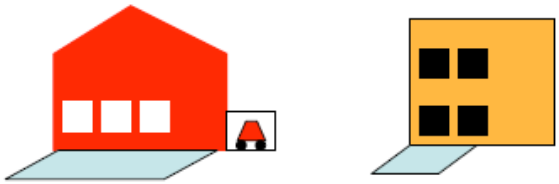
Kódovanie číselných dát pomocou dĺžky a výšky sa používa keď je potrebné vytvoriť kvalitatívny obraz dát, a môže byť užitočné pri robení rozhodnutí. Ako príklad môžeme uviesť systém, v ktorom niekoľko elektrických komponentov rýchlo (niekoľko krát za sekundu) a automaticky mení svoje hodnoty. Užívateľ, ktorý tento proces monitoruje, tak musí pozorovať niekoľko hodnôt naraz. Za týchto okolností by číselné zobrazenie bolo nevhodné, oveľa viac sa hodia stĺpcové výšky (bar heights). Dĺžka každého stĺpca znázorňuje aktuálnu hodnotu pre vybraný komponent, a rámček okolo každého stĺpca vyjadruje jeho maximálnu výšku, reprezentujúc tak najväčšiu možnú hodnotu, ktorú môže komponent nadobudnúť. Ako tento systém beží a počítač automaticky upravuje hodnoty vybraných komponentov, výšky stĺpcov sa rýchlo zväčšujú

a zmenšujú, čo umožňuje pozorovateľovi získať dobrý prehľad o celkovom fungovaní systému.



1.2.3 Ikony

Ikony sa často používajú na zobrazenie viacrozmernej informácie. Obrázok (ikona) sa skladá z niekoľkých prvkov, kde zobrazenie každého z nich závisí od hodnoty premennej, ktorú tento prvok zobrazuje. Príkladom použitia môže byť vybratie domu spĺňajúceho určité kritériá. Ak na zobrazenie domov použijeme ikony, kde farba reprezentuje cenu (červená farba - cena nad 400 000 USD, oranžová farba - cena medzi 300 a 400 000 USD, atď.), tvar reprezentuje typ domu (dom, chata, houseboat, ...), počet izieb je zobrazený počtom okien v ikone, veľkosť záhrady obdĺžnikom pred domom, a pod., tak dokážeme vybrať dom podľa určitých kritérií približne o polovicu skôr, ako keby sme ho hľadali v tradičnej, textovej reprezentácii.



Známym príkladom použitia ikon sú tiež tzv. **Chernoffove tváre**, kde sú na zobrazenie hodnôt premenných použité črty tváre, ako napríklad veľkosť očí, výška obočí nad očami, tvar úst, a pod. Keďže ľudia sú vnímaví k širokej škále rôznych výrazov a vzhľadov tváre, je pomocou tejto techniky možné pomerne ľahko identifikovať určité skupiny vzoriek.

1.2.4 Farba

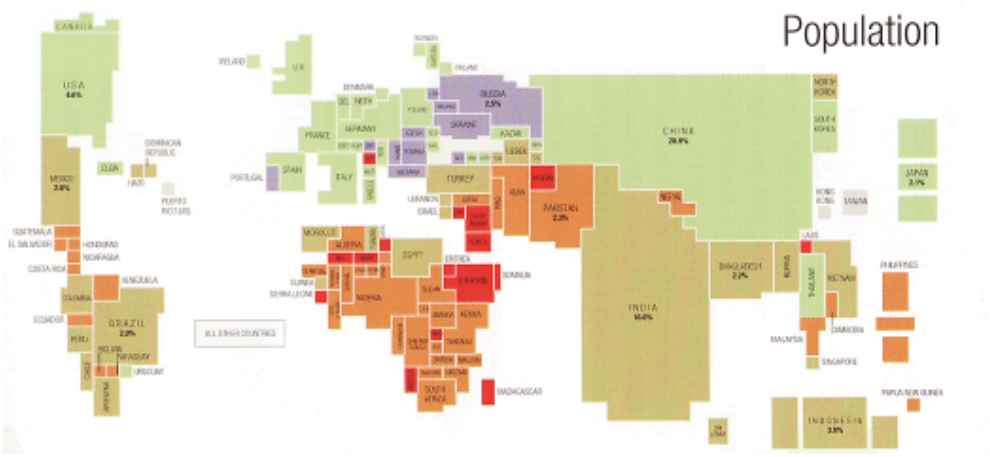
Asi najčastejším spôsobom reprezentácie numerických dát je farba. Často používanou voľbou je farebná reprezentácia veľkosti - červená farba pre veľké, a modrá pre malé elementy. Ďalší príklad použitia farby je už spomenutý príklad s ikonami, kde bola farbou zobrazená cena domu - škála červená-oranžová-žltá-biela tu znázorňovala klesajúcu hodnotu. Zaujímavé zistenie je, že pri testoch spojených s týmto zobrazením, žiadny z opýtaných subjektov nepotreboval danú schému vysvetliť, teda vedeli ju pochopiť bez väčšieho úsilia. Takéto techniky, ktoré sú užívateľmi rozoznané veľmi rýchlo a jednoducho nazývame **preatentívne**.
Rôzne odtiene šedej bývajú tiež často používané pre zobrazenie informácie, je však treba vhodne zvoliť odtiene tak, aby ich užívateľ vedel od seba bez problémov odlíšiť.

1.2.5 Priestorovosť

Možným spôsobom odovzdania informácie je tiež využitie dobrej priestorovej pamäte človeka. Napríklad problém usporiadania veľkej zbierky kníh o amerických štátoch by väčšina ľudí riešila abecedným usporiadaním. Ako dobré riešenie, ktoré si ale vyžaduje výbornú znalosť geografickej polohy týchto štátov, sa ukázalo aj rozmiestnenie kníh do políc podľa pozície štátu, teda knihy o Floride boli v ľavom dolnom rohu, knihy o Kalifornii v pravej časti, v strednej výške, a pod.

1.2.6 Zväčšenie

Je možné zapojiť ľudskú, dobre vyvinutú pamäť tradičného atlasu sveta a použiť tak zväčšenie ako techniku kódovania geografických dát. Teda ak by počet bicyklov na obyvateľa na Novom Zélande bol 10krát väčší ako v Austrálii, môžme tento fakt zobraziť na mape, kde Nový Zélalnd bude relatívne zväčšený oproti Austrálii. Táto metóda bola efektívne využitá napríklad v *The New State of the World Atlas*, kde si na mape zobrazujúcej počet obyvateľov, ihneď všimneme malú veľkosť Kanady a Austrálie, a veľkú rozlohu Indie. Hoci je táto technika efektívna v poskytnutí okamžitého dojmu, príliš sa spolieha na zapamätanie veľkostí.



1.2.7 Zvuk

Aj keď 'vizualizácia' obsahuje slovo vizuálny, môže byť ku podaniu informácie použitý tiež sluchový vnem. Keďže ľudia sú schopní rozlišovať rôzne tóny, akordy, tempo, či rytmus, zdá sa byť použitie zvuku vhodným nástrojom pre kódovanie informácie. Výhodou oproti klasickej grafickej vizualizácii je, že určitý zvuk, či tón, vieme zachytiť aj keď danej aktivite práve nevenujeme pozornosť.

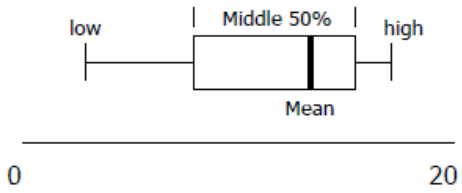
1.3 Typy dát

V rámci vizualizácie informácií sa stretávame s dátami, ktoré obvykle pozostávajú z veľkého množstva záznamov, každý obsahujúc niekoľko premenných, či atribútov. Každý záznam zodpovedá nejakému pozorovaniu, meraniu, transakcii a podobne. Počet atribútov sa môže medzi jednotlivými množinami dát líšiť: určité pozorovanie môže byť opísané piatimi premennými, pričom ďalšie potrebuje aj niekoľko stoviek premenných. Počet týchto premenných nazývame rozmernosť - dimenzionalita dát. Dáta teda môžu byť jednorozmerné, dvojrozmerné, viacrozmerné, či komplexné dátové typy ako text/hypertext, hierarchie/grafy.

1.3.1 Jednorozmerné dáta

Jednorozmerné dáta sú najjednoduchší dátový typ, ktorý má len jeden rozmer. Typickým príkladom sú časové dáta, kde s každým časovým bodom je asociovaná nejaká hodnota. Na zobrazenie takýchto dát je možné použiť jednoduchú

tabuľku, efektívnejšie zobrazenie je vo forme bodov na určitej škále, histogram, alebo Tukey Box Plots.



Počet zobrazených bodov môže byť - a zvyčajne býva - dosť veľký. V takom prípade je určite prínosná možnosť priblížiť zvolený interval vo väčšom detaile. Jednoduché približovanie umožňuje užívateľovi len vidieť zväčšený bod, oveľa efektívnejšie je použitie logického priblíženia, pri ktorom sa s narastajúcim stupňom priblíženia objavujú stále nové a nové dáta. Táto aktivita sa tiež nazýva **sémantické približovanie** (semantic zooming).

1.3.2 Dvojrozmerné dáta

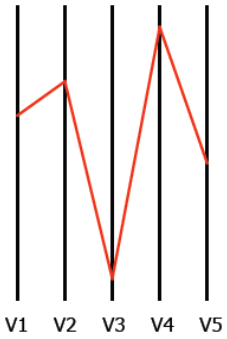
Dvojrozmerné dáta majú dve rozličné dimenzie. Tradičný prístup k vizualizácii a interpretácii dvojrozmerných dát je predstavovaný dvojrozmerným diagramom dvoch premenných. V príklade hľadania vhodného domu, kde na jednej osi je zobrazená cena, a na druhej počet izieb, umožňuje táto technika zobrazenia ľahko identifikovať všeobecné trendy (cena sa zvyšuje s narastajúcim počtom izieb), ako aj okrajové hodnoty, ktoré môžu odhaliť hodnotné informácie (lacný dom s veľa izbami).

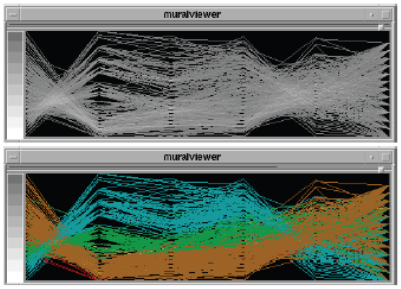
Ďalším príkladom sú geografické dáta, kde tieto dve dimenzie sú šírka a dĺžka, a mapa je len špeciálnym typom x-y diagramov pre dvojrozmerné geografické dáta.
Aj keď sa vizualizácia týchto dát zdá jednoduchá, v prípade veľkého množstva

zobrazovaných záznamov sa môže mapa príliš preplniť a nemusí tak vôbec pomôcť k pochopeniu zobrazovaných dát.

1.3.3 Viacrozmerné dáta

Mnohé množiny dát pozostávajú z viac ako troch atribútov a tak neumožňujú takú jednoduchú vizualizáciu ako 2 či 3 rozmerné diagramy. Príkladom takýchto dát sú záznamy z relačných databáz, ktoré často obsahujú desiatky až stovky stĺpcov (atribútov). Keďže nemáme žiadne jednoduché mapovanie toľkých atribútov do dvoch rozmerov obrazovky, sú potrebné zložitejšie vizualizačné techniky.
Jednou z takých je technika paralelných súradníc, kde každý atribút (rozmer) je znázornený vertikálnou osou. Pre každý záznam je potom na každej osi vyznačená hodnota príslušného atribútu, a tieto hodnoty prislúchajúce jednému záznamu sú medzi každými susednými osami spojené horizontálnymi čiarami. Isté vzťahy medzi premennými je možné vyčítať zo vzhľadu diagramu. Ako však bude diagram vyzerať vo veľkej miere závisí od usporiadania jednotlivých paralelných súradníc, je preto dobré vedieť určiť ich správne poradie.





1.3.4 Text a hypertext

Nie všetky dátové typy je možné opísať pomocou dimenzionality. V dobe World Wide Web sa ďalsím dôležitým dátovým typom stáva text a hypertext, ako aj multimediálny obsah stránok. Tieto dáta sa od predchádzajúcich odlišujú tým, že ich nieje možné jednoducho opísať číslami a teda na ich zobrazenie sa väčšinou nedajú použiť tradičné vizualizačné techniky. Najprv teda musí byť aplikovaná určitá transformácia dát do opisných vektorov, ktoré sú potom určitým spôsobom zobrazené.

1.3.5 Hierarchie a grafy

Dátové záznamy sú častokrát v nejakom vzťahu s inými časťami informácie. Grafy sú široko rozšírenou reprezentáciou práve takýchto vzťahov. Príkladmi sú e-mailové vzájomné vzťahy medzi ľuďmi, ich nakupovacie návyky, štruktúra súborov na hard disku alebo hyperlinky vo www. Vizualizácia grafu je veľmi jednoduchá - uzly sú zobrazené ako body, navzájom pospájané čiarami. Ako dodatočné techniky pre kódovanie informácie môžu byť použité rôzne farby a hrúbky spájajúcich čiar.

1.3.6 Algoritmy a softvér

Ďalšou triedou dát sú algoritmy a softvér. Cieľom ich vizualizácie je podpora vývoja softvéru umožnením pochopenia algoritmov, napríklad zobrazením toku informácií v programe, pochopenie písaného kódu, a pod. Napríklad môžme chcieť zobraziť adresár obsahujúci 20 súborov zdrojových kódov, ktoré dokopy obsahujú niekoľko desiatok tisíc riadkov. Súbory môžme reprezentovať ako stĺpce, a jednotlivé riadky kódu ako tenké riadky v týchto stĺpcoch. Farba riadku znázorňuje jeho vek (najnovšie riadky sú červené, najstaršie modré) a odsadenie riadkov od ľavého okraja zobrazuje štruktúru kódu - funkcie, vetvenia, cykly. V takomto zobrazení vieme potom napríklad rýchlo identifikovať, ktoré časti kódu v ktorých súboroch sú rovnako staré, alebo ktoré boli nedávno zmenené, a pod.

1.4 Interakcia

Pre efektívne skúmanie dát je nevyhnutné použitie techník interakcie a pretvorenia (interaction and distortion techniques). Techniky interakcie umožňujú analytikovi priamo interagovať so zobrazením a dynamicky ho meniť podľa úlohy skúmania, a tiež umožňujú spojiť a skombinovať niekoľko nezávislých vizualizácií. Techniky pretvorenia zas poskytujú možnosti zamerať sa na určité detaily, pričom zachovávajú celkový pohľad na dáta. Rozlišujeme pojmy dynamicky a interaktívne podľa toho, či zmeny vo vizualizácii sú robené automaticky alebo manuálne.

1.4.1 Dynamická projekcia

Základnou ideou dynamických projekcií je dynamicky meniť zobrazenia za účelom preskúmania viacerých rozmerov multirozmerných dát. Typickým príkladom je projekcia všetkých dvojrozmerných zobrazení viacrozmernej

množiny dát ako série x-y diagramov (scatterplots). Počet všetkých zobrazení je však exponenciálny k dimenzii dát, je teda ťažko použiteľný pre vysokú rozmernosť údajov. Postupnosť zobrazení môže byť náhodná, manuálna, prednastavená, alebo riadená dátami.

1.4.2 Interaktívne filtrovanie

Pri skúmaní rozsiahlych skupín dát je dôležité môcť interaktívne rozdeliť celú skupinu dát do segmentov, a zamerať sa na určitú podskupinu, ktorú považujeme za zaujímavú. Toto môžme dosiahnuť buď priamym výberom želanej podmnožiny (prehliadanie - browsing) alebo špecifikovaním vlastností podmnožiny (dotazovanie - querying). Prehliadanie je ťažko realizovateľné pre veľké súbory dát, dotazovanie zas nemusí priniesť želané výsledky, preto boli vyvinuté viaceré interakčné techniky, ktoré umožnili ľahšie filtrovanie údajov. Príkladom je napríklad technika Magic Lenses, ktorej základnou ideou je použitie zväčšovacieho skla priamo vo vizualizácii dát. Dáta nachádzajúce sa pod zväčšovacím sklom sú spracované filtrom, a výsledná množina je zobrazená odlišne od zvyšných údajov. Táto technika poskytuje upravený pohľad na vybraný región, pričom zvyšok zobrazenia ostáva nezmenený.

1.4.3 Interaktívne približovanie

Približovanie (zooming) je známa technika široko používaná mnohými aplikáciami. Keď sa zaoberáme veľkým množstvom dát, je dôležité prezentovať dáta vo vysoko komprimovanej podobe, ktoré poskytnú celkový prehľad, zároveň však poskytnúť zobrazenia dát v rôznych rozlíšeniach. Približovanie neznamená len zobrazenie väčších objektov reprezentujúcich dáta, ale taktiež to znamená automatickú zmenu reprezentácie dát tak, aby poskytla viac informácií na vyššej úrovni priblíženia. Objekty môžu byť napríklad zobrazené ako jednotlivé pixely pri nízkej úrovni priblíženia, ako ikony pri strednej úrovni a ako popísané objekty na vysokej úrovni priblíženia.
Zaujímavým príkladom je aplikácia priblíženia vo vizualizačnej technike TableLens. Získať prehľad o rozsiahlych tabuľkových dátach môže byť zložité, ak sú zobrazené v textovej podobe. TableLens reprezentuje numerickú hodnotu malým stĺpcom, všetky stĺpce majú výšku 1 pixel, a dĺžku vypočítanú podľa zobrazovanej hodnoty. To znamená, že počet zobrazených riadkov je takmer tak vysoký, ako vertikálne rozlíšenie, a počet stĺpcov závisí od maximálnej šírky stĺpca pre každý atribút. Počiatočný pohľad umožňuje užívateľovi hľadať vzory, vzájomné vzťahy, či okrajové hodnoty v danej množine dát. Za účelom prehliadnutia určitej oblasti si ju užívateľ priblíži, pričom zahrnuté

riadky sú zobrazené vo väčšom detaile - ak je to možné, tak v textovej forme.

1.4.4 Interaktívne pretvorenie (distortion)

Techniky interaktívneho pretvorenia podporujú proces skúmania dát zachovaním celkového pohľadu na dáta, pokým sú vykonávané operácie na nižšej úrovni. Základná idea je zobraziť časti dát s vyššou úrovňou detailu, pokiaľ zvyšok dát je zobrazený s nižšou úrovňou detailu. Populárnymi technikami sú hyperbolické a guľové pretvorenia (distortions), ktoré sú často používané na hierarchiách a grafoch, ale môzu byť tiež aplikované aj na ostatné vizualizačné techniky.

1.4.5 Interaktívne spájanie a vykresľovanie (linking and brushing)

Existuje veľa možností, ako zobrazovať viacrozmerné dáta a každá z nich má svoje výhody aj nevýhody. Ideou interaktívneho spájania a vykresľovania je skombinovať rôzne vizualizačné metódy tak, aby sa obišli slabé stránky jednotlivých techník. Rôzne projekcie x-y diagramov môžu byť skombinované s ofarbením a spájaním množín bodov vo všetkých projekciách. Podobným spôsobom môže byť spájanie a vykresľovanie aplikované na všetky vizualizačné techniky. Výsledkom je, že ofarbené body sú zvýraznené vo všetkých zobrazeniach, čo umožňuje odhaliť rôzne závislosti a korelácie. Interaktívne zmeny spravené v jednom zobrazení, sú automaticky premietnuté aj do ostatných vizualizácií. Spojenie viacerých zobrazení touto technikou poskytuje viac informácií, ako keby sme jednotlivé zobrazenia posudzovali nezávisle od seba.

1.5 Visual Information Seeking Mantra

Existuje mnoho odporúčaní pre vizuálny návrh, pričom ich základný princíp môžme sformulovať ako mantru hľadania informácie (Information Seeking Mantra) a to - prehľad, priblíženie a filtrovanie, a potom detaily na požiadanie (Overview first, zoom and filter, then details-on-demand). Túto skupinu úloh môžme rozšíriť ešte o ďalšie tri - zobrazenie vzťahov (relate), história (history), výber (extract), ktoré sú tiež dôležité pri vizualizácii dát.

1.5.1 Prehľad - Overview

Prehľad poskytuje všeobecné súvislosti potrebné pre pochopenie daného súboru dát; vykresľuje obraz dát ako celku, ktorý reprezentuje daná vizualizácia. Užitočné vzory a trendy zobrazovaných dát sú častokrát viditeľné len z pohľadu, ktorý pozostáva z celkového zobrazenia dát. Z tejto perspektívy sú viditeľné hlavné komponenty a ich vzťah jeden k druhému. Význačné črty môžu byť potom vybrané pre ďalšie skúmanie. Ich objavenie môže pomôcť užívateľovi v odfiltrovaní vedľajších informácií a tak môže efektívnejšie vykonať svoju úlohu.

1.5.2 Priblíženie - Zoom

Keďže maximálne množstvo zobrazenej informácie môže byť limitované rozlíšením, či farebnou hĺbkou displeja, je približovanie dôležitou technikou, ktorá prekonáva toto obmedzenie. Užívatelia sa zvyčajne zaujímajú o určitú časť zobrazených dát, je teda potrebné poskytnúť nástroje umožňujúce kontrolovať cieľ a faktor priblíženia.
Plynulé priblíženie je dôležité pri zachovaní dojmu o celkovom kontexte. Predstavme si, že si užívateľ prezeral nejaký objekt A a chcel by prejsť k skúmaniu objektu B, ktorý sa však nenachádza v práve zobrazenom náhľade. Vhodné je nevymeniť len jednoducho obraz objektu A za obraz objektu B, ale zobraziť oddialený pohľad, v ktorom sú oba objekty zájmu zobrazené naraz, a potom plynulo prejsť k detailu objektu B. Takýto prístup pomáha zachovať interný model dát v mysli užívateľa, ktorý potom vie lepšie pochopiť vzťah vybraných objektov k celkovej množine dát.

1.5.3 Filtrovanie - Filter

Filtrovanie umožňuje užívateľovi odfiltrovať dáta, o ktoré sa nezaujíma a zamerať sa tak len na určitú podmnožinu záujmu. Často používanou je technika dynamických dotazov (dynamic queries), kde pomocou posuvníkov užívateľ určí rozsah hodnôt jednotlivých premenných, a zobrazené sú potom len dáta spĺňajúce tieto filtrovacie kritéria. Dôležité je dosiahnuť rýchly update zobrazenia pri zmene kritérií (pod 100 milisekúnd), aj v prípade, že ide o tisícky zobrazených položiek.

1.5.4 Detaily na požiadanie - Details-on-demand

Umožňuje vybrať objekt alebo skupinu objektov a zobraziť detaily ak je to potrebné. Keď máme dáta, ktoré boli zúžené na niekoľko desiatok položiek, je

jednoduché prehliadať detaily o skupine, či jednotlivých položkách. Tradičný prístup je jednoducho kliknúť na vybraný objekt a zobraziť prekrývacie okno s detailami.

1.5.5 Zobrazenie vzťahov - Relate

Zobrazenie vzťahov medzi jednotlivými položkami môže byť užitočné, avšak navrhnutie akcií užívateľského rozhrania a výber vzťahov, ktoré majú byť objasnené, nieje jednoduchá úloha.
Príkladom je systém pre vyhľadávanie filmov, kde si užívatelia môžu vybrať atribút, napríklad režisér filmu, a potom v okne obsahujúcom detaily na požiadanie posuvníkom vybrať meno, čím sa následne zobrazia len filmy vybraného režiséra.

1.5.6 História - History

Skúmanie informácií je proces pozostávajúci z viacerých krokov, a je veľmi zriedkavé aby užívateľ po jednom kroku dospel k želanému výsledku. Je preto dôležité udržovať históriu akcií a poskytnúť tak užívateľovi možnosť vrátiť krok späť (undo), či znovuvykonať vrátenú akciu (redo).

1.5.7 Výber - Extract

Častokrát je vhodné umožniť užívateľovi nielen definovať a zobraziť želanú podmnožinu dát spĺňajúcu určité kritériá, ale aj vybrať a uložiť tieto údaje do súboru vo formáte, ktorý je možné ďalej používať - či už vložiť do prezentácie, poslať e-mailom, alebo vytlačiť. Alternatívou je uloženie nastavení, ktoré viedli k zobrazeniu danej množiny dát.

2 Analytické rozdiely - Analytic Gaps

Moderné systémy pre vizualizáciu informácií poskytujú rozsiahle prehľady údajov, podporujú výber a overenie individuálnych dát a umožňujú vykonávať dynamické dotazy. Napriek tomu môžme povedať, že tieto úlohy sa sústreďujú hlavne na podanie reprezentácie dát a existujú stále pochybnosti o schopnosti týchto systémov podporovať vyššie-úrovňové analytické úlohy akými sú učenie, či robenie rozhodnutí. Najčastejšími nedostatkami súčasných systémov je:

- **Nedostatočná funkčnosť** - operácie poskytované mnohými vizualizačnými systémami sú ekvivalentné jednoduchým databázovým dotazom ako zoraďovanie, filtrovanie, zobrazenie dvojrozmerných vzťahov. Hoci tieto operácie sú užitočné pri počiatočnom skúmaní dát, pri rozhodovaní potrebujeme častokrát brať do úvahy iné, štatistické vlastnosti.

- **Preddefinované reprezentácie** - reprezentácie používané bežnými vizualizáciami nebývajú dostatočne prispôsobivé, podporujúc vytvorenie niekoľkých statických modelov z elementárnych dotazov. Ak vizualizačný softvér podporuje x-y diagramy (scatterplots) a užívateľ by potreboval zobraziť vrstevnicovú mapu, zvyčajne musí použiť iný balíček. Hoci existujú vizualizácie, ktoré sú vhodné pre určitú špecifickú doménu, či problematiku a ich použitie môže byť veľmi efektívne, vynára sa otázka, či každá nová doména vyžaduje novú techniku vizualizácie.

- **Pokles determinizmu v rozhodovaní** - dôležitým faktorom je tiež skutočnosť, že dnešnému svetu nedominujú len informácie, ale tiež nepresnosti. Mnoho infovis systémov sa nezaoberá pojmom nepresnosti v dátach a týkajúcimi sa príčinami a následkami príliš dobre.

Na základe týchto nedostatkov vznikajú rozdiely medzi súčasnými vizualizačnými systémami a viac analytickými systémami, preto tieto rozdiely nazývame **analytické rozdiely**. Dajú sa zatriediť do dvoch kategórií - worldview gap a rationale gap.

2.1 Worldview gap

Worldview gap definujeme ako rozdiel medzi tým, čo je zobrazené, a tým, čo by v skutočnosti malo byť zobrazené, aby bolo možné vyvodiť jasný záver pre spravenie rozhodnutia. Základné prvky pre dosiahnutie tohto sú: zobrazenie vhodných dát, použitie vhodných zobrazení pre znázornenie dát a ukázanie jasných vzťahov.
Systémy, ktoré prekonávajú worldview gap nielen, že zobrazujú užitočné vzťahy medzi dátami, ale taktiež ukazujú užitočné reprezentácie a ich obmedzenia.

2.2 Worldview-based precepts

Uvedieme tri odporúčania, ktoré podporujú formuláciu stratégie pre prehliadanie vizualizácie tým, že odporúčajú užívateľovi aké dáta by mali byť prezreté na ujasnenie vzťahov, či umožňujú otestovanie hypotézy.

2.2.1 Stanovenie parametrov domény

Atribúty dát vo vizualizácii, a tým aj parametre, podľa ktorých sú dáta vo vizualizácii organizované, vyjadrujú ako pravidlá merania v rámci súboru dát, tak aj kľúčové parametre pre pochopenie domény. Napríklad, fakt, že zbierka výsledkov amerického baseballu obsahuje počet úderov, home run-ov a iných atribútov znamená, že sú tieto parametre považované za dôležité a môžu vyžadovať ďalšie objasnenie.
Systém teda môže pomôcť prekonať worldview gap poskytnutím možnosti pre vytvorenie, nadobudnutie a prenos poznatkov a metadát o dôležitých parametroch domény v rámci daného súboru dát.

2.2.2 Poskytnutie viacerých vysvetlení

Väčšina vizualizačných systémov umožňuje pochopiť vzťahy medzi dvoma - troma premennými. Avšak, niektoré vzťahy zahŕňajú viac ako tri vysvetľujúce premenné, či jednoduché transformácie samostatných premenných s použitím logaritmov alebo polynomiálnych vzťahov. Takéto korelácie často niesú spravované bežnými vizualizačnými nástrojmi. V takýchto prípadoch správna interpretácia zvyčajne vyžaduje určité navádzanie od užívateľa. Vo všeobecnosti, štatistika ponúka metódy, ktoré automaticky dokážu vybrať vhodný reprezentačný model, no bezhlavé používanie takýchto nástrojov nemusí viesť k správnym výsledkom. Kombinovanie daných metód s užívateľským riadením však môže priniesť veľmi užitočné prostriedky pre analýzu dát.
Sila tohto pravidla spočíva v objavení užitočných vysvetľujúcich premenných, či už automaticky alebo manuálne, čo prispieva k prekonaniu worldview gap.

2.2.3 Umožnenie otestovania hypotézy

Užívatelia potrebujú možnosť otestovať správnosť svojich dedukcií o danom súbore dát. Nástroje preto musia pomôcť užívateľovi definovať hypotézu, simulovať možné výsledky a overiť tak pravdivosť takej hypotézy, k čomu bývajú použité rôzne štatistické metódy. Ak je nájdený určitý zaujímavý výsledok, potom overenie hypotézy môže spočívať tiež v tom, ako jednoducho dokáže užívateľ s týmto výsledkom pracovať.
Tento analytický proces je zložité podporovať všeobecne pri návrhu interface a reprezentácií, môže však byť užitočný pri rozhodovaní o určitých črtách dizajnu.

2.3 Rationale gap

Rationale gap definujeme ako rozdiel medzi vnímaním vzťahu a schopnosťou skutočne vysvetliť užitočnosť daného vzťahu. Prvky tohto zahrňujú: presvedčenie o dátach, zvýraznenie nejasností v dátach, a pochopenie dôsledkov zmeny. Niektoré systémy síce pomáhajú vo vnímaní vzťahov medzi dátami, častokrát však zlyhávajú pri vysvetlení ich silných stránok. Spôsoby, ako prekonať rationale gap je nielen poskytnúť presné, jasné odpovede, ale tiež poskytnúť užívateľom logický základ o rozhodnutiach, ktoré môžu byť s ich použitím urobené.

2.4 Rationale-based precepts

Užívatelia potrebujú byť schopní spájať dáta do určitých oblastí, v rámci ktorých môžu byť robené rozhodnutia. Napríklad analýza chemických dát môže ukázať novú základnú zlúčeninu pre tvorbu lieku, a správna vizualizácia môže pomôcť pri rozhodovaní, ako upraviť existujúcú zlúčeninu aby mohla byť vytvorená nová.

2.4.1 Odhalenie nejasností

Určité nejasnosti obsahuje každý súbor dát. Je daný súbor dát dostatočne veľký, aby nedošlo k výberovej chybe? Existujú v dátach čísla obsahujúce nejasnosti, ako napríklad populačné odhady so spojenými odchýlkami? Pochopenie, kde sú hodnoty nepresné a ako táto nepresnosť ovplyvňuje schopnosť dát byť spoľahlivým zdrojom výsledkov, je preto veľmi dôležité.
Napríklad, keď uvažujeme niekoľko predajcov jednej súčiastky, ktorej šírka musí byť presne v určitom rozsahu, potrebujeme poznať nielen šírku priemernej vyrobenej súčiastky, ale takisto aj jej štandardnú odchýlku.

2.4.2 Konkrektizovanie vzťahov

Ďalšie pravidlo pre prekonanie rationale gap sa zameriava na schopnosť racionálne odôvodniť rozhodnutia a výsledky založené na vnímaných vzťahoch. Je podobné predchádzajúcemu pravidlu, s tým rozdielom, že sa viac zameriava na koncoužívateľskú prezentáciu, ako objavovanie. Zhrnúť to teda môžme tak, že systém by mal jasne prezentovať, čo zahŕňa reprezentácia určitého vzťahu, a tiež prezentovať konkrétne výsledky kde je to vhodné.

2.4.3 Odhalenie príčiny a následku

Pri skúmaní dát zvyčajne existuje príčina, prečo sa dané dáta dostali práve do daného súboru dát. Pre utvorenie kognitívneho modelu u užívateľa je preto dôležité vysvetliť za akých predpokladov daný súbor dát vznikol, a teda čím boli výsledky ovplyvnené.

2.5 Použitie odporúčaní

Analytické rozdiely a odporúčania, ktoré boli popísané vyššie, tvoria základ pre návrh a ohodnotenie systému. V podstate, všetko, čo treba urobiť je aplikovať dané pravidlá v tej ktorej situácii.

2.5.1 Použitie pravidiel v návrhu

Keď navrhujeme vizualizáciu pre novú doménu či scenár, môžme pravidlá aplikovať nasledovne:

1. Vytvoríme zoznam podúloh, ktoré by vizualizácia mala podporovať, či vykonávať
2. Identifikujeme možné nedostatky v reprezentácii alebo dátach
3. Určíme možné vzťahy, ktoré by mali byť zvýraznené, alebo použité ako základ vizualizácie

Hlavnou myšlienkou je aplikovať dané pravidlá v každom scenári tak, ako by to urobil používateľ. Napríklad, "Čo je nepresné v daných dátach a ako to ovplyvní výsledky, ktoré uvidím?", alebo Äko zobrazím konkrétne výsledky tohto procesu?"

2.5.2 Použitie pravidiel vo vyhodnocovaní

Popísané pravidlá môžu slúžiť taktiež pri vyhodnocovaní kvality vizualizačného systému a to jednoduchým vyhodnotením ako daný systém podporuje tie ktoré odporúčania, napríklad kladením otázok ako sú aktuálne vzťahy a výsledky zobrazené užívateľovi, alebo aké presvedčenie by užívateľ mal získať o daných dátach.

3 Štruktúra operátorov vo vizualizačných systémoch

Predstavme si vizualizačnú aplikáciu (povedzme HomeFinder - vyhľadávač nehnuteľností podľa určitých kritérií) s dvoma pohľadmi na rovnaké dáta. V jednom je použitý x-y diagram (scatterplot) s posuvníkmi pre dynamické dotazovanie, a v druhom je použitá usporiadaná číselná tabuľka. Posunutím posuvníkov odfiltrujeme niektoré dáta a scatterplot sa podľa toho upraví. Avšak, otázka sémantiky sa vynára pre zobrazenie tabuľky. Jedna možná interpretácia je, že tabuľka je nezávislé zobrazenie pôvodných nefiltrovaných dát a teda nieje potrebné ju meniť. Druhá možná interpretácia je, že pôvodný súbor dát bol touto interakciou upravený a je potrebné zmeniť aj pohľad na tabuľku. Ktorá z týchto možností je teda správna? Predpokladajme, že užívateľ potrebuje vybrať nehnuteľnosti spadajúce do určitého cenového rozsahu. Ak sa však zaujíma iba o to, ako sa diagram zmení so zmenou cenových intervalov, v takom prípade tabuľku nieje treba upravovať, pretože sémantika tejto úlohy si nevyžaduje upravenie pôvodných dát. Ak však užívateľ chce vytvoriť novú množinu domov, ktoré spadajú do danej cenovej relácie, v takom prípade budú pôvodné dáta modifikované a teda aj zobrazenie tabuľky by sa malo príslušne upraviť. Obidve interpretácie sú teda možné.

Problém pre koncového užívateľa Tento príklad ukázal, že užívatelia môžu mať problém s interakciou s niektorými vizualizačnými systémami, pretože existuje rozdiel medzi zámerom a možnou vykonateľnou akciou. Niekedy býva sémantika operácií nepresná a užívateľ častokrát nemá možnosť predpovedať, aký bude výsledok jeho akcií. Týmto interakčný model môže brániť analytickému procesu, pretože nespĺňa potreby analýzy.
Problém pre návrhárov Návrhári vizualizačných systémov mávajú tri najčastejšie problémy: znovupoužitie operátorov, oddelenie hodnoty/zobrazenia a zameranie na operand (operand focus).
Existujú teda dôvody, prečo je potrebné vytvoriť štruktúru operátorov - model, ktorý umožní jasne klasifikovať a organizovať vizualizačné operácie a pomôcť tak koncovým užívateľom ako aj návrhárom lepšie pochopiť situácie, v ktorých môžu byť operátory použité a akým spôsobom.

3.1 Vlastnosti operátorov

U operátorov existujú dve dôležité vlastnosti - tá prvá, či je daný operátor operátor zobrazenia, alebo hodnoty - teda či modifiuje základné dáta alebo nie. Tá druhá je stupeň funkčnej podobnosti s inými operátormi.

3.1.1 Funkčná versus operatívna podobnosť

Niektoré operátory sú medzi aplikáciami **operatívne podobné** - operácie, ktorých základné implementácie sú úplne rovnaké od jednej aplikácie ku druhej, ako napríklad otočenie, zmena rozmerov, posunutie, manipulácia s pozíciou kamery, manipulácia s geometrickými objektami a osvetlenie. Celá trieda geometrických a scénických operátorov sú operatívne podobné, pretože predpokladáme, že keď už raz dostaneme nejaký pohľad, tak ten pozostáva z určitých grafických primitív ako čiary a mnohouholníky, s ktorými vieme pracovať vždy rovnakým spôsobom.
Ďalšia trieda operátorov sú **funkčne podobné** operátory - sémanticky sú podobné medzi jednotlivými aplikáciami, no ich základná implementácia je rozdielna pre rôzne dátové typy. Napríklad filtrovanie množiny dát je veľmi častá a užitočná operácia, no rozdielne aplikácie používajú rôzne spôsoby filtrovania. Ďalším príkladom sú operácie súčtu a rozdielu. Spôsob, akým spájame dva zoznamy vlastností nehnuteľnosti nieje rovnaký, ako napríklad spojenie štruktúr webových odkazov z dvoch rozdielnych prelezení webu.
Poslednou skupinou operátorov sú **operátory závislé od konkrétnej úlohy** - operácie, ktoré sú špeciálne navrhnuté pre špecifickú úlohu v rámci určitej domény. Príkladom je napríklad operácia parsovania HTML dokumentu za účelom výpočtu podobnosti dokumentov.

3.1.2 Zobrazenie versus hodnota

Ďalší rozmer operátorov je, či sú orientované na zobrazenie, alebo na hodnotu. Hodnotou rozumieme neupravené dáta, zatiaľ čo zobrazenie je vizualizácia koncového produktu.
Operátor hodnoty mení zdroj dát operáciami ako pridávanie alebo mazanie podmnožín dát, či filtrovanie alebo modifikácia základných dát. Takýto operátor zvyčajne generuje novú množinu dát.
Operátor zobrazenia na druhej strane mení len obsah vizualizácie. Príklady takých operátorov zahŕňajú 3D otočenie, posunutie, priblíženie, horizontálne alebo vertikálne prevrátenie obrázku, zmena priehľadnosti povrchu. Operátory zobrazenia podstatne nemenia základnú množinu dát.
Rozdiel medzi operátormi zobrazenia a operátormi hodnoty nie je vždy celkom

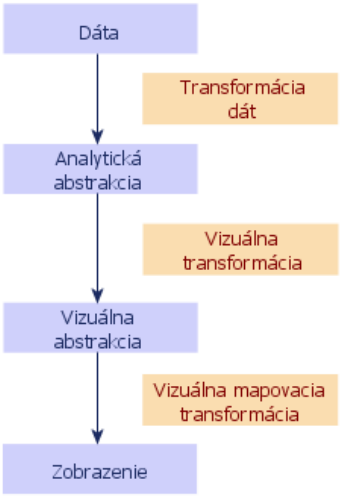
jasný. Napríklad, modifikácia farebnej mapy znamená zmenu hodnoty jednotlivých pixelov pre vybraný obrázok, a teda táto operácia môže byť klasifikovaná ako operátor hodnoty. Avšak, pri úprave tepelnej mapy 3D povrchu sa zmena teplotnej škály javí ako operácia zobrazenia, keďže nemení základné hodnoty povrchu.

3.2 Vizualizácia stavu operátorov

Pojmy funkčnej a operatívnej podobnosti sa týkajú pojmov operátorov zobrazenia a hodnoty. Operátory zobrazenia sú viac operatívne podobné, operátory hodnoty sú zas funkčne podobné, ale implementované rôzne pre rôzne dáta.

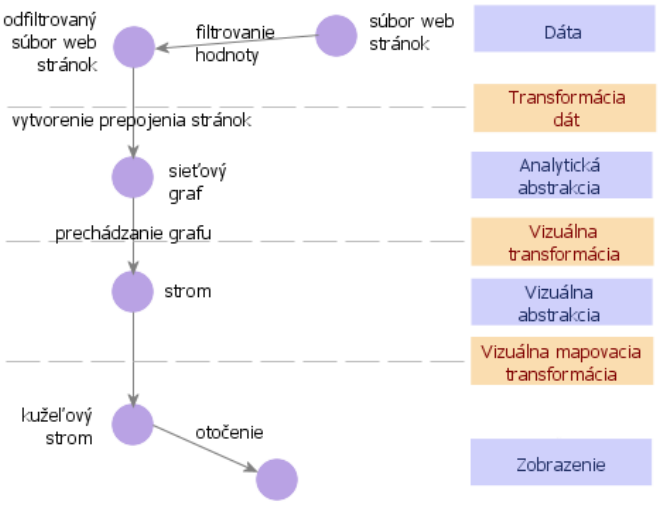
3.2.1 Vizualizačné zreťazenie (visualization pipeline)

Na jednom konci zreťazenia sú dáta (hodnoty), pričom na druhom konci je vizualizácia (zobrazenie). Základnou klasifikáciou pre určenie pozície operátora v tomto zreťazení, je rozdelenie operátorov podľa toho či ovplyvňujú dáta alebo zobrazenie. Na jednom konci sú plne zobrazovacie operátory, na druhom operátory hodnôt, a medzi nimi sú všetky tie, ktoré nepatria úplne ani do jednej kategórie.
Neupravené dáta sú zvyčajne najprv spracované do určitej formy analytickej abstrakcie pomocou procesu transformácie dát. Táto analytická abstrakcia je častokrát ďalej redukovaná použitím vizuálnej transformácie do formy vizuálnej abstrakcie, ktorá je už zobraziteľným obsahom. Tento proces zvyčajne zahŕňa redukciu dimenzie, keďže údaje sú zvyčajne komplexné a viacrozmerné. Z vizuálnej abstrakcie vedie ďalší krok k vizuálnej mapovacej transformácii, ktorej výsledkom je pohľad zobraziteľný na monitore užívateľa.



3.2.2 Stavový model

Stavový model je odvodený od vizualizačného zreťazenia s dvoma modifikáciami: Zreťazenie nebralo do úvahy viacero vstupov - ak dva rozdielne dátové súbory mali vytvoriť jednu vizualizáciu, zreťazenie sa nedalo použiť. Stavový model je rozšírený na sieť a umožňuje toľko hodnôt a toľko pohľadov, koľko je potrebné. Po druhé, stavový model používa uzly na zobrazenie stavu dát a hrany na zobrazenie operátora, ktorý transformuje dáta z jedného stavu do iného. Stavový model je teda užitočný pre niektoré vizualizačné úlohy, pretože umožňuje užívateľovi vidieť medzivýsledky pri plánovaní ďalších operácií.
Ako príklad aplikujeme túto štruktúru vo vizualizácii štruktúry web stránok. Dátami je v tomto prípade súbor podstránok vygenerovaný preliezaním webu. Najprv vykonáme filtrovanie podľa hodnoty (operátor hodnoty), kde vyhľadáme dokumenty obsahujúce určité slovo. Potom použijeme tento výber stránok na vygenerovanie sieťového grafu (analytická abstrakcia) z odkazov na stránkach (operácia transformácie dát). Na grafe môžme vybrať len určitú podskupinu uzlov, napríklad tie so vzdialenosťou nie väčšou ako tri od koreňového uzlu (operácia na úrovni analytickej abstrakcie). Z tohto potom vieme vygenerovať strom a zobraziť ho pomocou niektorej vizualizačnej techniky (operácia vizuálnej transformácie).



3.3 Zhrnutie

V tejto kapitole sme teda vytvorili klasifikáciu, ktorá ujasňuje nasledovné vlastnosti operátorov:

- **Zobrazenie vs. hodnota** - čím bližšie sa nachádza operátor na zobrazovacom konci zreťazenia, tým má viac vlastností zobrazovacieho operátora, a rovnako to platí pre operátory hodnoty.
- **Použiteľnosť operátorov** - závisí od umiestnenia operátora v zobrazovacom zreťazení (Visualization pipeline). Posúvaním sa k zobrazovaciemu koncu nás približuje viac ku všeobecným dátam, ktoré sú použiteľné vo veľkom počte rôznych dátových domén. Pri výbere implementácie operátorov je teda dobré uprednostniť operátory čo najbližšie k zobrazovaciemu koncu. Sú efektívne a jednoduchšie sa implementujú vo vizualizačnom systéme.
- **Priama manipulácia** - rozsah priamej manipulácie tiež závisí od pozície operátora vo visualization pipeline. Čím bližšie je k zobrazeniu, tým väčšia interaktivita je možná. Napríklad, geometrická pozícia a orientácia sú operácie, ktoré sú priamo upraviteľné. Čím sa posúvame po pipeline bližšie k operátoru hodnoty, tým viac narastá doménová závislosť a špecifikácia týchto operácií je zložitejšia. Takou operáciou je napríklad výber dát z rôznych formátov súborov.

4 Objavné nástroje - Spojenie vizualizácie informácie s dolovaním dát

Narastajúce používanie nástrojov pre vizualizáciu informácie a algoritmov pre dolovanie dát pochádza z dvoch rôznych smerov výskumu. Infovis výskumníci veria, že je dôležité dať užívateľom prehľad a získať náhľad do daných dát, pričom výskumníci dolovania dát veria v silu štatistických algoritmov pri hľadaní zaujímavých vzorov.
Počítače umožnili vykonávať komplexné štatistické analýzy, ktoré v minulosti nebolo možné vykonávať. Avšak narastá tiež nebezpečenstvo používania zložitých softvérových nástrojov v prípade, že im užívateľ celkom nerozumie a nevie ich kontrolovať. Preto je užitočné uvažovať o vhodnosti rôznych metód, ktoré boli použité. Tento prístup môže viesť k lepšiemu pochopeniu kedy použiť ktoré metódy, a tiež prispieť k vynájdeniu nových či zdokonaleniu existujúcich objavných nástrojov (Discovery tools). V nasledujúcom texte prediskutujeme dve otázky, ktoré ovplyvňujú návrh objavných nástrojov a to: štatistické algoritmy vs. vizuálna prezentácia dát, a overenie hypotézy vs. objavná analýza dát.

4.1 Štatistické algoritmy vs. vizuálna prezentácia dát

Prvotné snahy o zosumarizovanie dát viedli k vygenerovaniu stredných hodnôt, štandardných odchýliek a intervalov. Tieto čísla boli užitočné, lebo v porovnaní s kompletným súborom dát boli kompaktné, jasné, umožňovali porovnanie či rozhodovanie a hlavne boli objektívne. Avšak, taktiež niekedy mohli skryť zaujímavé črty, ako napríklad, či je rozloženie dát rovnomerné, nepravidelné, alebo ovplyvnené krajnými hodnotami (outliers). Riešením týchto problémov bola prezentácia dát ako vizuálneho diagramu, takže zaujímavé črty mohli byť spozorované človekom.
Vizuálna prezentácia je veľmi silná pri objavovaní trendov, zvýrazňovaní extrémnych hodnôt, zobrazovaní zoskupení a odhaľovaní rozostupov. Taktiež umožňuje užívateľom lepšie porozumieť, čo sa deje v daných dátach a odporučiť tak smery pre ďalšie štúdium. Nedostatky vizuálnej prezentácie sú v zaobchádzaní s veľkými súbormi dát, absorpcia dát, alebo ich zlá interpretácia.
Typická prezentácia výsledkov štatistického dolovania dát sú stručné sumarizujúce tabuľky, odvodené pravidlá, alebo rozhodovacie stromy. Typická vizuálna prezentácia zas zobrazuje bohaté histogramy, x-y diagramy (scatterplots), tepelné mapy, paralelné súradnice, atď. s podporou užívateľom

riadeného skúmania a dynamickými dotazmi pre filtrovanie. Porovnávacie štúdie ukázali dôležitosť oboznámenia užívateľa s každým prístupom a tiež vplyv špecifických úloh.

4.2 Overenie hypotézy vs. objavná analýza dát

Zástancovia hypotéz a kontrolovaných pokusov tvrdia, že najväčší prínos prináša to, že výskumníci musia sformulovať svoje hypotézy ešte pred zozbieraním dát, čo vedie k ostrejšiemu mysleniu, šetrnejšiemu výberu dát a presnejším meraniam. Ich cieľom je pochopiť príčinné vzťahy, vyprodukovať opakovateľné výsledky či objaviť zovšeobecniteľné náhľady. Kritici však považujú tento prístup za príliš vytrhnutý z kontextu, čo podľa nich môže viesť k ignorovaniu dôležitých premenných, ktoré ovplyvnili výsledky. Taktiež iniciálne stanovenie hypotézy môže viesť následné pozorovanie k hľadaniu faktov potvrdzujúcich danú domnienku, a prehliadnutiu zaujímavých faktov, ktoré niesú spojené s danou hypotézou.
Na druhej strane stoja zástancovia prieskumnej analýzy dát, ktorí tvrdia, že väčší prínos má zozbieranie veľkého objemu dát a následné hľadanie zaujímavých vzorov, a stanovovanie hypotéz nepovažujú za potrebné. Skeptici však tvrdia, že akýkoľvek veľký súbor dát môže byť vždy len určitým špeciálnym prípadom, a teda výsledky nieje možné vždy zovšeobecňovať. Tiež je otázne, či nájdenie vzťahov môže viesť k pochopeniu príčiny a následkov, pretože korelácia nemusí implikovať príčinnosť.
Ako príklad môžme použiť fabriku na výrobu polovodičov, v ktorej zistili vysokú mieru chybovosti. Zástancovia overovania hypotéz najprv zostavia zoznam možných príčin ako prímesy, prílišné teplo alebo príliš rýchle ochladenie, a potom hľadajú dôkazy na podporu toho ktorého tvrdenia a prípadne sa pokúsia problém znovu navodiť. Zástancovia prieskumnej analýzy dát namiesto toho zozbierajú existujúce dáta z posledného roku o produkcii za rôznych podmienok, spustia nástroje pre dolovanie dát a pokúsia sa odhaliť vzťahy medzi vysokou chybovosťou a ostatnými premennými.

Riešením tejto úvahy je vziať to najlepšie z obidvoch extrémov a vytvoriť nové objavné nástroje pre veľa rôznych užívateľov a veľa rozličných domén. Skúsení analytici často kombinujú pozorovanie v prvotných štádiách, ktoré vedie k testovaniu hypotézy. Prípadne môžu mať utvorenú presnú hypotézu, ak sú ale pozorní pozorovatelia, počas kontrolovaného experimentu môžu objaviť anomálie vedúce k novej hypotéze.

4.3 Zhrnutie a odporúčania

Výpočtové nástroje pre objavovanie dát, ako dolovanie dát a vizualizácia informácie prešli výrazným vývojom za posledných pár rokov. Nanešťastie, vývoj týchto nástrojov prebiehal poväčšinou v oddelených komunitách s rôznymi filozofiami. Výskumníci z oblasti dolovania dát veria v schopnosť štatistických metód identifikovať zaujímavé vzory bez zásahu človeka. Na druhej strane výskumníci vizualizácie informácie považujú za dôležitú kontrolu doménového experta pri vytvorení vizuálnej prezentácie, ktorá môže viesť k objaveniu neočakávaných výsledkov.

Odporúčanie 1 - integrovať dolovanie dát s vizualizáciou informácie pre vytvorenie objavovacích nástrojov. Pridaním vizualizácie k dolovaniu dát umožňíme užívateľovi hlbšie pochopiť dané dáta. Pridaním dolovania dát k vizualizácii zas užívatelia budú môcť špecifikovať, čo hľadajú. Stredná cesta umožňujúca užívateľom zostaviť ich objavnú analýzu dát aplikovaním ich doménových znalostí (ako obmedzenie dolovacích algoritmov na určitý rozsah hodnôt) môže byť zdrojom inovatívnych nástrojov.

Odporúčanie 2 - umožniť užívateľom špecifikovať, čo hľadajú a čo považujú za zaujímavé. Užívateľom je vhodné poskytnúť možnosť určiť o aké typy vzťahov sa zaujímajú, alebo aké hraničné hodnoty hľadajú. Ak užívateľ potom otestuje svoju hypotézu na daných dátach a zistí jej nepravdivosť, môže zároveň objaviť nové možnosti.
Keďže objavovanie je proces, nielen jedna udalosť, udržiavanie histórie užívateľových akcií môže byť užitočné. Užívatelia by mali mať možnosť uložiť svoj stav (dáta, nastavenia ovládacích panelov), vrátiť sa do predchádzajúceho stavu, či poslať svoju históriu ostatným.

Odporúčanie 3 - pamätať, že užívateľ patrí do sociálneho kontextu. Výskumníci zriedkakedy pracujú osamote. Potrebujú nahromadiť dáta z viacerých zdrojov, konzultovať s doménovými odborníkmi, podať ďalej čiastkové výsledky a prezentovať zistenia kolegom. Úspešné nástroje teda umožňujú vymieňať si dáta, konzultovať s ostatnými a podať im výsledky.

Odporúčanie 4 - zachovať zodpovednosť užívateľa pri návrhu objavných nástrojov. Ak sú nástroje zrozumiteľné, predpovedateľné a kontrolovateľné, užívatelia sa ich ľahko naučia dobre používať k úspešnému dokončeniu svojej práce, mali by však niesť zodpovednosť aj za prípadné svoje chyby. Keď sú nástroje príliš zložité alebo nepredvídateľné, užívatelia sa bránia ich používaniu pretože ich nedokážu kontrolovať.

Ak užívatelia dobre nerozumejú štatistickým algoritmom alebo vizuálnej prezentácii, nemôžu pracovať s ich výsledkami s istotou. Viditeľnosť týchto procesov a následkov znižuje riziko zlej interpretácie a nesprávnosti výsledkov. Pochopenie algoritmov za vizualizáciou podporuje efektívne používanie, ktoré vedie k úspešnému objavovaniu.

5 Vlastné závery a odporúčania

Na základe problematík popísaných v predchádzajúcich kapitolách môžme sformulovať niekoľko princípov pre návrh vizualizačných systémov:

- **Integrovať dolovanie dát s vizualizáciou informácie**. Vizuálna prezentácia je veľmi silná pri objavovaní trendov, zvýrazňovaní extrémnych hodnôt, zobrazovaní zoskupení a odhaľovaní rozostupov, a dolovanie dát zas využíva silu štatistických výpočtov a obchádza ľudskú subjektivitu. Oba smery majú svoje výhody aj nevýhody, preto ich správnou kombináciou môžu vzniknúť užitočné nástroje.

- Pre získanie celkového prehľadu o dátach je vhodné užívateľovi **poskytnúť viacero zobrazení** pomocou viacerých navzájom prepojených vizualizačných techník. Interaktívne zmeny spravené v jednom zobrazení, sú automaticky premietnuté aj do ostatných, čo umožňuje odhaliť rôzne závislosti a korelácie, ktoré by pri samostatnom zobrazení mohli ostať nepovšimnuté.

- Pri zobrazovaní veľkého množstva dát je dôležité prezentovať dáta vo vysoko komprimovanej podobe, ktoré poskytnú celkový prehľad, zároveň však poskytnúť zobrazenia dát v rôznych rozlíšeniach. **Použitie plynulého priblíženia** umožnuje užívateľovi zamerať sa na určitú podmnožinu záujmu a prezrieť ju vo vyššom detaile, a zároveň podporuje zachovanie interného modelu u užívateľa a tým aj dojmu o celkovom kontexte.

- Užívatelia častokrát potrebujú odfiltrovať dáta, o ktoré sa nezaujímajú a zamerať sa tak len na určitú podmnožinu záujmu. Preto je dôležité **umožniť filtrovanie** a to buď priamym výberom želanej podmnožiny (prehliadanie - browsing) alebo špecifikovaním vlastností podmnožiny (dotazovanie - querying).

- Popri vykonávaní operácií na nižšej úrovni je dôležité **zachovávať celkový pohľad na dáta** a zároveň zobraziť časti dát na vyššej úrovni detailu.
- Umožniť užívateľovi pochopiť, ktoré sú **dôležité parametre domény** v rámci súboru dát
- Umožniť užívateľovi **definovať hypotézu**, simulovať možné výsledky a overiť tak pravdivosť takej hypotézy, k čomu bývajú použité rôzne štatistické metódy.
- **Odhaliť nejasnosti**. Dôležité je pochopenie, kde sú hodnoty nepresné a ako táto nepresnosť ovplyvňuje výsledky.
- **Odhaliť príčinu a následky**. Pri skúmaní dát zvyčajne existuje príčina, prečo sa dané údaje dostali práve do daného súboru dát. Pre lepšie pochopenie vzťahov je preto dôležité vysvetliť za akých predpokladov daný súbor dát vznikol.
- **Uprednostniť operátory zobrazenia** všade tam, kde je to možné. Takéto operácie pracujú so všeobecnými dátami a preto sa dajú ľahko znovupoužiť a svojou rýchlosťou a efektívnosťou umožňujú väčšiu mieru interaktivity.
- **Pamätať, že užívateľ patrí do sociálneho kontextu**. Úspešné nástroje umožňujú importovať dáta z viacerých zdrojov, vymieňať si dáta navzájom, konzultovať ich s ostatnými a podávať im výsledky.
- **Poskytnúť históriu**. Je užitočné udržovať históriu akcií a poskytnúť tak užívateľovi možnosť vrátiť krok späť (undo), či znovuvykonať vrátenú akciu (redo).

6 Záver

Práca priniesla teoretické informácie o vizualizácii dát, podrobne rozobrala niekoľko problémov návrhu vizuálnych aplikácií a poskytla odporúčania pre ich návrh.
Prvá kapitola popísala tri základné úlohy vizualizácie, predstavila najčastejšie používané prostriedky, ktorými je možné reprezentovať informácie, ako veľkosť, dĺžka a výška, ikony, farba, priestorovosť, zväčšenie, či zvuk a niektoré z nich ilustrovala príkladmi. Ďalej rozdelila dáta na niekoľko typov:

jednorozmerné, dvojrozmerné a viacrozmerné dáta, text a hypertext, hierarchie a grafy a algoritmy a softvér, pričom zároveň odporučila vhodné techniky vizualizácie pre každý dátový typ. Okrem toho ešte poskytla prehľad o niekoľkých spôsoboch interakcie, ktoré by mali byť podporované vizualizačným systémom a v závere spomenula spôsob hľadania informácie užívateľom a popísala úlohy, ktoré toto hľadanie podporujú.
Druhá kapitola práce vysvetľuje rozdiely medzi súčasnými vizualizačnými systémami a analytickými systémami, ktoré nazývame analytické rozdiely. Prvý z nich označujeme ako worldview gap a znamená rozdiel medzi tým, čo je zobrazené, a tým, čo by v skutočnosti malo byť zobrazené, aby bolo možné vyvodiť jasný záver pre spravenie rozhodnutia. A druhý označovaný ako rationale gap sme definovali ako rozdiel medzi vnímaním vzťahu a schopnosťou skutočne vysvetliť užitočnosť daného vzťahu. Ďalej sme popísali niekoľko odporúčaní, ktoré pomáhajú eliminovať tieto rozdiely a tým zvyšujú schopnosť systémov podporovať vyššie-úrovňové analytické úlohy akými sú učenie, či robenie rozhodnutí.
V tretej kapitole sme sa zaoberali štruktúrou operátorov vo vizualizačných systémoch, popísali sme dve hlavné vlastnosti operátorov - funkčnú a operatívna podobnosť, a zameranie operátora na hodnotu alebo zobrazenie. Predstavili sme stavový model rozdeľujúci operátory podľa ich zamerania a na základe tohto modelu sme dali odporúčania pre znovupoužiteľnosť operátorov a priamu manipuláciu so zobrazením.
Obsahom štvrtej kapitoly bolo spojenie dvoch rôznych smerov týkajúcich sa analýzy dát. Prvým z nich je vizualizácia dát, ktorá zdôrazňuje dôležitosť poskytnúť užívateľom prehľad o zobrazených dátach, a druhým smerom je použitie algoritmov pre dolovanie dát, ktoré sa spoliehajú na silu štatistických výpočtov pri hľadaní zaujímavých vzorov. Ďalej sme porovnali metódy oboch smerov a to: štatistické algoritmy s vizuálnou prezentáciou dát, a overenie hypotézy s objavnou analýzou dát. Na základe týchto porovnaní sme popísali niekoľko odporúčaní, ktoré môžu pomôcť k vzniku nových, objavných nástrojov.
V poslednej piatej kapitole sme zhrnuli odporúčania pre návrh vizuálnej aplikácie, čím sme naplnili cieľ bakalárskej práce.

7 Použitá literatúra

R. Spence: *Information Visualization*, ACM Press, 2001, 206 s., ISBN 0-201-59626-1

U. Fayyad, G. G. Grinstein, A. Wierse: *Information Visualization in Data Mining and Knowledge Discovery*, Academic Press, 2002, 407 s., ISBN 1-55860-689-0

P. Isenberg, A. Tang, S. Carpendale: *An Exploratory Study of Visual Information Analysis*

R. A. Amar, J. T. Stasko: Knowledge Precepts for Design and Evaluation of Information Visualizations, *IEEE Transactions on Visualization and Computer Graphics*, 2005

B. Schneiderman: *The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization*, 1996

D. A. Keim: Information Visualization and Visual Data Mining, *IEEE Transactions on Visualization and Computer Graphics*, 2002

D. M. Russel, M. J. Stefik, P. Pirolli, S. K. Card: The Cost Structure of Sensemaking, *Proceedings of InterCHI*, 1993

E. H. Chi, J. T. Riedl: *An Operator Interaction Framework for Visualization Systems*

T. J. Jankun-Kelly, K. Ma: A Model and Framework for Visualization Exploration, *IEEE Transactions on Visualization and Computer Graphics*

S. G. Eick: Visual Discovery and Analysis, *IEEE Transactions on Visualization and Computer Graphics*, 2000

B. Schneiderman: *Inventing Discovery Tools: Combining Information Visualization with Data Mining*

M. X. Zhou, S. K. Feiner: *Visual Task Characterization for Automated Visual Discourse Synthesis*, 1998