Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/16BE34Radojcic

VIZUELNO PRETRAŽIVANJE PODATAKA

VISUAL DATA MINING

Tanja Radojčić, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *Ovaj rad daje uvid u podatke i tehnike za vizuelizaciju podataka u cilju lakšeg korišćenja velikih količina podataka. Primeri pokazuju koliko vizuelni prikaz može olakšati rad čoveka u svakodnevnom životu. Kombinuju se čovek i algoritam za najbolji mogući ishod.*

Ključne reči: *vizuelizacija, pretraživanje podataka, velika količina podataka, tehnike vizuelizacije*

Abstract – *This paper provides insight into data and data visualization techniques to facilitate the use of large amounts of data. Examples show how much visual representation can facilitate a person's work in everyday life. Man and algorithm are combined for the best possible outcome.*

Keywords: *visualization, data search, big data, visualization techniques*

1. UVOD

Napretkom tehnologije u 21. veku došli smo do toga da današnji kompjuteri mogu da skladište ogromne količine podataka. Od početka razvoja kompjuterske tehnologije, nije bilo predviđeno da će godišnji unos novih informacija prelaziti čak 1 milion terabajta. Kako svakim danom tehnologija napreduje, kao i potreba za većim pohranjivanjem informacija, dolazimo do procene da će u naredne tri godine biti generisano više podataka nego što je ukupno od postojanja čovečanstva.

Prvo poglavlje je uvod u vizuelno pretraživanje podataka i sam rad. Drugo poglavlje je bazirano na osnovnim informacijama vezanim za vizuelno pretraživanje podataka. Treće poglavlje predstavlja opis svih vrsta podataka koji se koriste za vizuelizaciju. Četvrto poglavlje je posvećeno tehnikama vizuelizacije i njihovim reprezentacijama.

Peto poglavlje objašnjava tehnike interakcije i izobličenja, koje uz tehnike vizuelizacije služe za njihov detaljniji prikaz. Šesto poglavlje se osvrće na praktičnu primenu i opisani su primeri iz života sa dostupnim bazama podataka. Zaključak je pregled uvida stečen tokom pisanja rada.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Dragan Ivetić, red. prof.

2. TEORIJSKE OSNOVE

Vizuelno pretraživanje podataka podrazumeva prikazivanje podataka vizuelno kroz grafike, tabele i slično, pomoću kojih čovek dolazi do određenih zaključaka i povezuje date informacije. Postoje razne tehnike koje su se pokazale krajnje efikasno u ovim pretragama, te algoritmi pretraživanja zasnovani na njima mogu da pretraže ogromne količine podataka.

Kako je čovek uključen u sam proces, zaključci i ciljevi se lako menjaju ukoliko je to potrebno. Određene zaključke i veze između podataka može doneti softver, koje čovek može lako da previdi.

2.1. Proces vizuelnog pretraživanja

Prvo, korisnici dobijaju širok pregled svih unesenih podataka. U pregledu podataka, korisnici pronalaze zanimljive obrasce i fokusiraju se na neke od njih. Za njhovu analizu, moraju da zalaze duboko u podatke kako bi dobili detaljniji prikaz. Drugo, zumiranje različitih grupa podataka unutar vizualizovanih podataka će omogućiti da se filtriraju različite vrednosti i kreiraju obrasce pre nego što se dođe do zaključka zbog kog je proces i započet.

Tehnologije vizuelizacije se mogu koristiti za sva tri koraka, tako da je dobro koristiti jednu tehniku za prvi korak i prikazivanje podataka od interesa, dok se neka druga tehnika koristi za fokusiranje na deo tih podataka, detaljnije i opširnije. Vizuelizacijom možemo prikazati i veze između ta tri koraka, ne samo konkretno njih.

2.2. Klasifikacija tehnika

Za tri dimenzije klasifikacije - tip podataka koji se vizualizuje, tehnika vizualizacije i tehnika interakcije i izobličenja - može se pretpostaviti da su ortogonalne. Ortogonalnost znači da se bilo koja od tehnika vizualizacije može koristiti zajedno sa bilo kojom od tehnika interakcije, kao i bilo kojom od tehnika izobličenja za bilo koji tip podataka.

Određeni sistem može biti dizajniran da podržava različite tipove podataka i da može koristiti kombinaciju više tehnika vizualizacije i interakcije.

3. TIPOVI PODATAKA

U vizualizaciji informacija, podaci se obično sastoje od velikog broja zapisa od kojih se svaki sastoji od niza promenljivih ili dimenzija. Svaki zapis odgovara opservaciji, merenju, ili transakciji.

3.1. Jednodimenzionalni podaci

Jednodimenzionalni podaci obično imaju jednu gustu dimenziju. Tipičan primer jednodimenzionalnih podataka je vremenski podatak, vreme ili novac. Imajte na umu da se sa svakom tačkom vremena može povezati jedna ili više vrednosti podataka.

3.2. Dvodimenzionalni podaci

Dvodimenzionalni podaci imaju dve različite dimenzije. Tipičan primer su geografski podaci gde su dve različite dimenzije geografska dužina i širina. X-Y-grafikoni su tipična metoda za prikazivanje dvodimenzionalnih podataka, a mape su posebna vrsta X-Y -grafikona za prikazivanje dvodimenzionalnih geografskih podataka.

3.3. Višedimenzionalni podaci

Mnogi skupovi podataka sastoje se od više od tri atributa i stoga ne dopuštaju jednostavnu vizualizaciju kao dvodimenzionalni ili trodimenzionalni grafikoni. Primeri višedimenzionalnih (ili viševarijantnih) podataka su tabele iz relacionih baza podataka, koje često imaju desetine do stotine kolona (ili atributa). Pošto ne postoji jednostavno mapiranje atributa u dve dimenzije ekrana, potrebne su sofisticiranije tehnike vizualizacije. Primer tehnike koja omogućava vizualizaciju višedimenzionalnih podataka je tehnika paralelnih koordinata. Paralelne koordinate prikazuju svaku višedimenzionalnu stavku podataka kao poligonalna linija koja preseca horizontalne ose dimenzija na položaju koji odgovara vrednosti podataka za odgovarajuću dimenziju.

3.4. Tekst i hipertekst

Ne mogu se svi tipovi podataka opisati u smislu dimenzionalnosti. U doba svetske mreže, jedan važan tip podataka je tekst i hipertekst, kao i sadržaj multimedijalnih veb stranica. Ovi tipovi podataka se razlikuju po tome što se ne mogu lako opisati brojevima, pa se većina standardnih tehnika vizualizacije ne može primeniti. U većini slučajeva, prvo je potrebna transformacija podataka u vektore opisa pre nego što se mogu koristiti tehnike vizualizacije. Primer za jednostavnu transformaciju je brojanje reči koje se često kombinuje sa analizom glavnih komponenti ili višedimenzionalnim skaliranjem.

3.5. Algoritmi i softver

Zapisi podataka često imaju neku vezu sa drugim podacima, oslanjaju se na sadržaj objavljen na vebu. Grafikoni se široko koriste za predstavljanje takvih međuzavisnosti. Grafikon se sastoji od skupa objekata koji se nazivaju čvorovi i veza između ovih objekata koji se nazivaju ivice. Primeri su međusobni odnosi e-pošte među ljudima, njihovo ponašanje pri kupovini, struktura datoteka na čvrstom disku ili hiperveze na svetskoj mreži. Postoji niz specifičnih tehnika vizualizacije koje se bave hijerarhijskim i grafičkim podacima.

4. TEHNIKE VIZUELIZACIJE

Postoji veliki broj tehnika vizualizacije koje se mogu koristiti za vizualizaciju podataka. Pored standardnih 2D/3D tehnika, kao što su X-Y (x-y-z) grafikoni, trakasti grafikoni, linijski grafikoni itd., Postoji niz sofisticiranijih tehnika vizualizacije. Časovi odgovaraju osnovnim principima vizualizacije koji se mogu kombinovati radi implementacije određenog sistema vizualizacije [2].

4.1. Standardni 2D/3D prikaz

Najčešće korišćen prikaz. Uglavnom predstavlja krafikone, kockaste prikaze grafikona u 3 dimenzije.

4.2. Geometrijski transformisan prikaz

Geometrijski transformisane tehnike prikaza imaju za cilj pronalaženje „zanimljivih“ transformacija višedimenzionalnih skupova podataka. Klasa tehnika geometrijskog prikaza uključuje tehnike iz istraživačke statistike kao što su matrice raspršenog grafikona i tehnike koje se mogu podvesti pod izraz „traženje projekcije“. Druge tehnike geometrijskog projektovanja uključuju tužilačke poglede (eng. Prosection Views), Hyperslice i dobro poznatu tehniku vizualizacije paralelnih koordinata. Tehnika paralelnih koordinata preslikava k-dimenzionalni prostor u dve dimenzije ekrana korišćenjem k ekvivalentnih osa koje su paralelne sa jednom od osa prikaza. Osi odgovaraju dimenzijama i linearno su skalirane od minimalne do maksimalne vrednosti odgovarajuće dimenzije. Svaka stavka podataka predstavljena je kao poligonalna linija, koja seče svaku osu u onoj tački koja odgovara vrednosti razmatranih dimenzija.

4.3. Prikaz zasnovan na ikonama

Druga klasa tehnika istraživanja vizuelnih podataka su ikonične tehnike prikaza. Ideja je mapiranje vrednosti atributa višedimenzionalne stavke podataka sa karakteristikama ikone. Ikone se mogu proizvoljno definisati: To mogu biti mala lica, ikone zvezda, ikone u obliku štapića, ikone u boji i TileBars (Slika 7). Vizualizacija se generiše preslikavanjem vrednosti atributa svakog zapisa podataka na karakteristike ikona. U slučaju tehnike štapne figure, na primer, dve dimenzije se mapiraju u dimenzije ekrana, a preostale dimenzije u uglove i/ili dužinu ekstremiteta ikone figure štapa. Ako su stavke podataka relativno guste u odnosu na dve dimenzije ekrana, rezultujuća vizualizacija predstavlja obrasce teksture koji se razlikuju u skladu sa karakteristikama podataka i stoga se mogu otkriti predpažnjom.

4.4. Prikaz sa gustim pikselima

Osnovna ideja tehnika gustog piksela je mapiranje vrednosti svake dimenzije u obojeni piksel i grupisanje piksela koji pripadaju svakoj dimenziji u susedna područja. Pošto generalno ekrani sa gustim pikselima koriste jedan piksel po vrednosti podataka, tehnike omogućavaju vizualizaciju najveće moguće količine podataka na trenutnim ekranima (do oko 1.000.000 vrednosti podataka). Ako je svaka vrednost podataka predstavljena jednim pikselom, glavno pitanje je kako rasporediti piksele na ekranu. Tehnike gustog piksela koriste različite aranžmane u različite svrhe. Raspoređivanjem piksela na odgovarajući način, rezultirajuća vizualizacija pruža detaljne informacije o lokalnim korelacijama, zavisnostima i žarišnim tačkama.

4.5. Naslagani prikaz

Tehnike naslaganog prikaza prilagođene su za hijerarhijski prikaz podataka podeljenih na particije. U slučaju višedimenzionalnih podataka, dimenzije podataka koje će se koristiti za particioniranje podataka i izgradnju hijerarhije moraju biti odgovarajuće odabrane. Primer tehnike naslaganog prikaza je Dimenzijsko slaganje.

5. TEHNIKE INTERAKCIJE I IZOBLIČENJA

Pored tehnike vizualizacije, za efikasno istraživanje podataka potrebno je koristiti i neke tehnike interakcije i izobličenja. Tehnike interakcije omogućavaju analitičaru podataka da direktno stupi u interakciju sa vizualizacijama i dinamički menja vizualizacije u skladu sa ciljevima istraživanja, a takođe omogućavaju povezivanje i kombinovanje više nezavisnih vizualizacija. Tehnike izobličenja pomažu u procesu istraživanja podataka pružajući sredstva za fokusiranje na detalje uz očuvanje pregleda podataka. Osnovna ideja tehnika izobličenja je prikazati delove podataka sa visokim nivoom detalja, dok su drugi prikazani sa nižim nivoom detalja. Razlikujemo termine dinamički i interaktivni u zavisnosti od toga da li se promene vizualizacija vrše automatski ili ručno (direktnom interakcijom korisnika).

5.1. Dinamička projekcija

Osnovna ideja dinamičkih projekcija je dinamička promena projekcija radi istraživanja višedimenzionalnog skupa podataka. Broj mogućih projekcija je eksponencijalan u broju dimenzija, to jest nerešiv je za velike dimenzionalnosti. Prikazane sekvence projekcija mogu biti nasumične, ručne, unapred izračunate ili zasnovane na podacima

5.2. Interaktivno filtriranje

U istraživanju velikih skupova podataka važno je interaktivno podeliti skup podataka na segmente i fokusirati se na zanimljive podskupove. Ovo se može uraditi direktnim izborom željenog podskupa (pregledavanjem) ili specifikacijom svojstava željenog podskupa (upiti). Pretraživanje je veoma teško za veoma velike skupove podataka i upiti često ne daju željene rezultate. Stoga su razvijene brojne tehnike interakcije za poboljšanje interaktivnog filtriranja u istraživanju podataka.

5.3. Interaktivno zumiranje

Zumiranje je dobro poznata tehnika koja se široko koristi u brojnim aplikacijama. U radu sa velikim količinama podataka, važno je predstaviti podatke u visoko sabijenom obliku kako bi se obezbedio pregled podataka, ali istovremeno omogućio promenljiv prikaz podataka u različitim rezolucijama. Zumiranje ne znači samo da se objekti podataka prikazuju veći, već takođe znači da se prikaz podataka automatski menja kako bi predstavio više detalja o višim nivoima zumiranja. Objekti mogu, na primer, biti predstavljeni kao pojedinačni pikseli na niskom nivou zumiranja, kao ikone na srednjem nivou zumiranja i kao označeni objekti u visokoj rezoluciji. Zanimljiv primer primene ideje zumiranja na velike skupove tabelarnih podataka je pristup TableLens.

5.4. Interaktivno krivljenje

Interaktivne tehnike izobličenja podržavaju proces istraživanja podataka čuvajući pregled podataka tokom operacija bušenja. Osnovna ideja je prikazati delove podataka sa visokim nivoom detalja, dok su drugi prikazani sa nižim nivoom detalja. Popularne tehnike izobličenja su hiperbolične i sferne distorzije koje se često koriste na hijerarhijama ili grafikonima, ali se mogu primeniti i na bilo koju drugu tehniku vizualizacije.

5.5. Interaktivno povezivanje i četkanje

Postoji mnogo mogućnosti za vizualizaciju višedimenzionalnih podataka, ali svi oni imaju određenu snagu i neke slabosti. Ideja povezivanja i četkanja je kombinovanje različitih metoda vizualizacije kako bi se prevazišli nedostaci pojedinih tehnika. Na primer, tabele različitih projekcija mogu se kombinovati bojenjem i povezivanjem podskupa tačaka u svim projekcijama. Na sličan način, povezivanje i četkanje se mogu primeniti na vizualizacije generisane svim gore opisanim tehnikama vizuelizacije. Kao rezultat toga, brušene tačke su istaknute u svim vizualizacijama, što omogućava otkrivanje zavisnosti i korelacija. Interaktivne promene napravljene u jednoj vizualizaciji automatski se odražavaju u drugim vizualizacijama.

6. PRIMENA SA BIG DATA PODACIMA

Pretraživanje podataka je metoda analize velikih količina podataka u nastojanju da se otkriju odnosi, tj. Veze između podataka. Ove veze se slažu sa definicijom da moraju biti smisleni jer vode do nekoliko prednosti, najčešće finansijske. Podaci su obično kvantitativni, posebno ako se uzme u obzir eksponencijalni razvoj podataka odnosno big data. Vizuelno pretraživanje podataka je prikaz podataka u okviru grafikona, slike i slično. Oni su napravljeni kao vizuelni prikaz informacija.

7. PRAKTIČNA PRIMENA

Ova studija slučaja pretraživanja vizuelnih podataka zasnovana je na otvorenim podacima koje je objavio Centar za životnu sredinu Helsinkija.

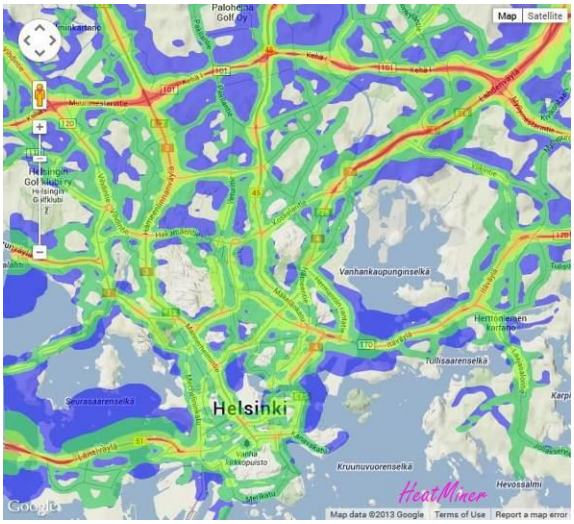
Originalni skup podataka dostupan je na zvaničnoj stranici regije Helsinki i sadrži detaljne informacije o nivoima gradske buke. Detaljan opis podataka i povezanih istraživanja mogu se pronaći iz zbirnog izveštaja koji je dostupan na veb stranicama projekta studije buke (uglavnom na finskom) [4].

Skup podataka o buci odličan je primer otvorenih podataka koji bi mogli biti vrlo vredni običnim ljudima. Na primer, porodice koje traže miran kraj u Helsinkiju verovatno bi želele da provere nivo buke u saobraćaju pre nego što kupe kuću. Nažalost, dostupne informacije su date u prilično teškom formatu za brzi pregled. Zato je napravljen ovaj primer sa toplotnim mapama kako bi na google mapu izgenerisali date podatke i njihovom vizuelizacijom pojednostavili prikaz.

Toplotna karta sa slike 1 prikazuje prosečne nivoe saobraćajne buke (LAeq) tokom dana (između 07č-22č) dok slika 2 prikazuje prosečan nivo buke tokom noći.

Boje toplotne karte ukazuju na prosečne nivoe buke, tako da su crvena područja najbučnija.

Slika 1. *Saobraćajna buka tokom dana (07č-22č)* [5]



Slika 2. *Saobraćajna buka tokom noći (22č-07č)* [5]

Nakon prikaza obe mape, preko dana i preko noći, možemo zaključiti sledeće: Saobraćaj u noćnim satima je manje bučan (što je očekivano), imamo mrežu glavnih saobraćajnica koje su preko dana veoma bučne i delove koje su čak i preko dana u najmirnijoj boji – plavoj. Ako se vratimo na primer sa početka, u kom očekujemo da bi ova mapa koristila ljudima koji biraju nekretninu u mirnijem delu, oni bez domenskog znanja mogu da prepoznaju delove grada i tako dođu do željene nekretnine, uz pomoć vizualnog pretraživanja podatka.

8. ZAKLJUČAK

Istraživanje velikih skupova podataka važan je, ali težak problem. Tehnike vizualizacije informacija mogu pomoći u rešavanju problema. Istraživanje vizuelnih podataka ima veliki potencijal i mnoge aplikacije, poput otkrivanja prevara i pretraživanja podataka, koristiće tehnologiju vizualizacije informacija za poboljšanu analizu podataka. Budući rad će uključivati usku integraciju tehnika vizualizacije sa tradicionalnim tehnikama iz disciplina kao što su statistika, mašinsko učenje, istraživanje operacija i simulacija. Integracija tehnika vizualizacije i ovih ustaljenijih metoda kombinovala bi brze algoritme automatskog rudarenja podataka sa intuitivnom snagom ljudskog uma, poboljšavajući kvalitet i brzinu procesa iskopavanja vizuelnih podataka. Viusalne tehnike rudarenja podataka takođe moraju biti čvrsto integrisane sa sistemima koji se koriste za upravljanje ogromnom količinom relacionih i polustrukturiranih informacija, uključujući upravljanje bazama podataka i sisteme skladišta podataka. Krajnji cilj je unositi moć tehnologije vizualizacije na svaku radnu površinu kako bi se omogućilo bolje, brže i intuitivnije istraživanje veoma velikih resursa podataka. Ovo neće biti samo vredno u ekonomskom smislu, već će i stimulisati i oduševiti korisnika.

9. LITERATURA

[1] D.A.Keim, „IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS“, 2002.
[2] D. A. KEIM, „Visual Data-Mining Techniques“. 2002.
[3] http://cloudnsci.fi/wiki/index.php?n=HeatMiner.RateOfIndebtedness (pristupljeno u oktobru 2021.)
[4] http://cloudnsci.fi/wiki/index.php?n=HeatMiner.TraficNoiseInHelsinki (pristupljeno u oktobru 2021.)
[5] http://cloudnsci.fi/wiki/index.php?n=HeatMiner.TraficAccidentsInHelsinki (pristupljeno u oktobru 2021.)
[6] http://cloudnsci.fi/wiki/index.php?n=HeatMiner.DeerCrashesInFinland (pristupljeno u oktobru 2021.)
[7] http://cloudnsci.fi/wiki/index.php?n=HeatMiner.VisitorLocationHeatmaps (pristupljeno u oktobru 2021.)
[8] E. Karn, „Visualization of Real Time Data Driven Systems using D3 Visualization Technique“.
[9] D. Asimov, “The grand tour: A tool for viewing multidimensional data“, 1985.
[10] S. J. Simoff1, M. H. Böhlen, and A. Mazeika, „Visual Data Mining: An Introduction and Overview“ ,2007.

Kratka biografija:

Tanja Radojčić je rođena 29.04.1996. godine u Novom Sadu. Završila je srednju školu gimnaziju „Isidora Sekulić“ u Novom Sadu 2015. godine. Fakultet tehničih nauka u Novom Sadu je upisala 2015. godine. Završila je osnovne akademske studije na Fakultetu tehničkih nauka 2019. godine. Završila je master akademske studije na Fakultetu tehničkih nauka 2021. godine, smer Primenjeno softversko inženjerstvo sa prosekom 9,20.