Dubinska analiza tokova podataka i njena primjena – pregled metoda i istraživanja

Sabina Šišović
Odjel za informatiku, Sveučilište u Rijeci
Radmile Matejčić 2, 51000 Rijeka
ssisovic@inf.uniri.hr

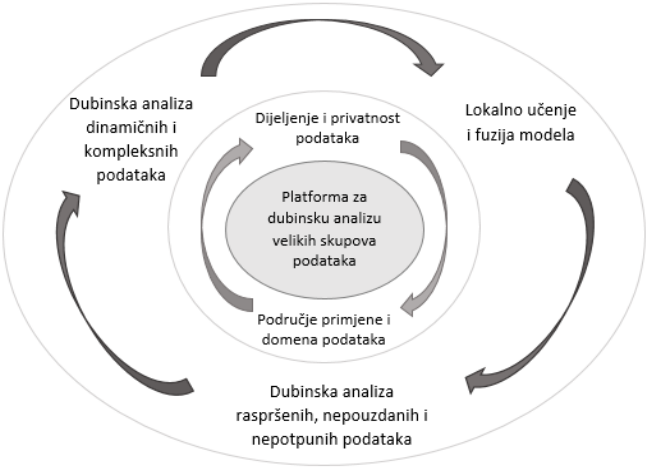
Sažetak - Iz dubinske analize velikih skupova podataka se, s vremenom, specijalizirala grana koja znanje crpi iz tokova podataka. Tokovi podataka kontinuirano pristižu iz nekog izvora, a samim time i uzorci kroz vrijeme evoluiraju. Javlja se potreba za algoritmima dubinske analize koji su prilagođeni neprekidnim promjenama. Ovaj rad služi kao uvod u područje dubinske analize tokova podataka. U uvodnom dijelu opisani su veliki skupovi podataka i navedena njihova obilježja. O dubinskoj analizi podataka riječ je u drugom poglavlju, dok je u trećem poglavlju dan uvod u karakteristike tokova podataka i tehnike za njihovo procesiranje. Metode dubinske analize tokova podataka navode se u četvrtom poglavlju, a slijedi ga poglavlje u kojem su dani primjeri primjene dubinske analize tokova podataka. U zaključku su navedeni neki od izazova u području rada s tokovima podataka.

Ključne riječi: dubinska analiza tokova podataka, klasifikacija, grupiranje, otkrivanje znanja

I. UVOD

Količina dostupnih podataka, gledano desetljećima unazad, eksponencijalno je rasla u svim promatranim područjima aktivnosti ljudi. Potreba za informacijama pokretala je nove načine za generiranje podataka i bilježenje gotovo svega što se bilježiti da, kako bi se povećale šanse da u moru podataka budu otkrivene upravo najkorisnije informacije koje će biti temelj za, prvenstveno, stvaranje novih profita. Između generiranja podataka i samog njihovog tumačenja i pretvaranja u informaciju nastao je raskorak koji raste proporcionalno s količinom podataka, a savladavanje tog raskoraka provodi se kroz metode otkrivanja znanja u podacima. Osnova otkrivanja znanja u podacima počiva na algoritmima iz područja strojnog učenja. Metode koje pri otkrivanju znanja koriste algoritme koji traže ponavljajuće uzorke među podacima i koriste ih za otkrivanje složenih utjecaja između varijabli promatranog sustava, ali i predviđanje budućih događaja zovemo dubinskom analizom podataka (engl. *Data Mining*). Witten i Frank su u [1] ponudili pregled alata i tehnika strojnog učenja u službi dubinske analize podataka pri pronalaženju i tumačenju ponavljajućih uzoraka u podacima.

Dio svakodnevice u različitim znanstvenim i inženjerskim područjima su veliki, kompleksni, rastući setovi podataka generirani od strane različitih izvora. Takvi setovi podataka poznati su pod nazivom „Big Data“ Vrlo su poznate njihove karakteristike u obliku tri „V“: volumen, raznolikost i brzina (engl. *Volume, Variety i Velocity*) [2], ali su se s vremenom izdvajale i nove karakteristike koje su pridodane ovoj skupini. Najprije su prepoznate još dvije karakteristike: varijabilnost i vrijednost (engl. *Variability i Value*) [3]. Iz [4] saznajemo da navedenih karakteristika trenutno ima 7, a dvije posljednje su istinitost i vizualizacija (engl. Veracity i Visualization). Korištenje takvih podataka i dobivanje informacija iz njih donosi izazove koji su proporcionalni obujmu takvih skupova te se bave rješavanjem problema vezanih uz osobine velikih skupova podataka. Te osobine su Wu i drugi objasnili HACE teoremom [5], prema kojem Big Data polazi iz izvora koji su golemi, heterogeni i autonomni (engl. *Heterogeneous, Autonomous*) s distribuiranom i decentraliziranom kontrolom, a nastaje kako bi se istražili kompleksnost i evoluirajuće veze (engl. *Complex, Evolving*) među podacima Sukladno teoremu, autori su procesiranje Big Data prikazali okvirom koji izazove istraživanja stavlja na tri razine koji zaokružuju platformu za dubinsku analizu Big Data tj. velikih skupova podataka (Slika 1).



Slika 1. Okvir za procesiranje velikih skupova podataka. (prilagođeno iz [5])

U sljedećem poglavlju riječ je o dubinskoj analizi velikih skupova podataka, s naglaskom na tri navedene razine. Treće poglavlje uvod je u tokove podataka, tj. obuhvaća njihovu strukturu, tehnike procesiranja i sustave za upravljanje. U četvrtom poglavlju navedene su i opisane metode dubinske analize tokova podataka, dok su u petom poglavlju opisana okruženja u kojima nastaju podaci tokovne prirode i primjena dubinske analize na njih. Peto poglavlje donosi zaključak s mogućim smjernicama za daljnje istraživanje.

II. DUBINSKA ANALIZA VELIKIH SKUPOVA PODATAKA

Na prvoj razini okvira za procesiranje velikih skupova podataka riječ je o izazovima vezanim uz raspolaganje podacima i računalne postupke nad njima. S obzirom da se podaci skladište na različitim lokacijama, učinkovita platforma za rad nailazi na tehničke barijere ili pak na intenzivne troškove potrebne da bi se podaci preselili na druge lokacije (intenzivna mrežna komunikacija), čak i kad je na raspolaganju glavna memorija velikog obujma. Nadalje, svaki izvor podatke skladišti prema vlastitoj zamisli raspoređivanja, a aplikacije za pohranu također za te podatke koriste i različite načine prikaza.

Druga razina obuhvaća problematiku semantike i primjene znanja za raznolike velike skupove podataka, što se u najvećoj mjeri odnosi na probleme pri dijeljenju i privatnosti podataka, te je posebno vezano uz pozadinsko znanje eksperta iz područja primjene i njegova očekivanja vezano uz otkrivanje znanja.

Na trećoj razini obuhvaćeni su izazovi koncentrirani na dizajn algoritama pri rješavanju teškoća nastalih zbog obujma podataka, kompleksnosti i dinamičnosti podataka. Iz slike 1 vidljivo je da je da ova razina obuhvaća algoritme i tehnike iz triju faza, koje se cirkularno odvijaju. Najprije se koriste tehnike pretprocesiranja i fuzije podataka da bi se integrirali podaci koji su raspršeni, heterogeni, nepouzdani, nepotpuni i potječu iz više izvora. Na njima se provodi prva faza dubinske analiza podataka. Potom slijedi faza koja obuhvaća algoritme dubinske analize kompleksnih i dinamičkih podataka. Zatim se traže korelacije i kreiraju modeli putem kojih bi se od uzoraka, dobivenih na lokalnoj razini (iz konkretnih podataka) i njihovim kombiniranjem s obzirom na izvore, oblikovali uzorci na globalnoj razini. Naposljetku, stečeno znanje iz prethodnih faza ulazi u početnu fazu gdje ponovo prolazi proces dubinske analize podataka kako bi se modeli i svi parametri prilagodili novom stanju.

Metode dubinske analize podataka prilagodile su se, s vremenom, osobinama podataka spomenutim u trećoj razini. Kompleksni i dinamički podaci generiraju se, primjerice, u okruženjima kao što su komunikacijske mreže, društvene mreže, Internet (stranice s hipervezama), itd. Radi se o mrežama gdje svi sudionici mogu potencijalno imati koristi od otkrivanja znanja iz podataka koji se generiraju, a kompleksnost leži u strukturi takve mreže i njenoj ekspanziji, što čini podatke iz takvih izvora jednim od ključnih izazova aplikacijama za dubinsku analizu podataka. Tako su razvijene metode za otkrivanje znanja iz skupova podataka u obliku kompleksnih mreža i s dinamičnim promjenama njihovog obujma. Primjerice, metode detektiranja zajednica unutar mreža i evolucije veza između njih ključne su za razumijevanje kompleksnih sustava poput društvene mreže [6], dok su pak metode otkrivanja anomalija unutar mreže, čiji pregled su ponudili Bindu i Thilagam [7], sredstvo za detektiranje *spammera* i osiguravanje društvenih aktivnosti preko mreža.

Prilagodba na nove osobine podataka je zamjetna i u metodama koje se ne tiču kompleksnih mreža, iako se i dalje radi o podacima pristiglih iz više izvora, primjerice metoda paralelnih algoritama za podatke iz različitih medija [8]. U [9] se navodi metoda klasifikacije unutar sustava baza podataka koja funkcionira na način da klasificira više baza podataka sa sličnim karakteristikama u jednu klasu. Unutar nekoliko takvih klasa, izdvojila bi se uvijek jedna kao najrelevantnija te bi naposljetku smanjeni broj baza iziskivao trošenje manje resursa na pretraživanje. Autori u [10] su predstavili logički okvir dizajniran da identificira pouzdana znanja koja crpe iz različitih izvora te eliminira ona nepouzdana. Istaknuli su da su se dotadašnja istraživanja temeljila na pretpostavci da se ulazni podaci sastoje od dobro definiranih podataka, koji ne sadrže nedosljedne ili netočne vrijednosti te u njima ne nedostaju vrijednosti.

Podaci iz stvarnog života prispijevaju i bilježe se najčešće kao tokovi podataka pa su metode morale razviti i prilagodbu za otkrivanje znanja iz takve, dinamičke strukture podataka. Statičke metode dubinske analize podataka ne mogu se prilagoditi karakteristikama tokova podataka kao što su kontinuitet, varijabilnost, brzina ili neograničenost, te bi mogle rezultirati i gubitkom korisnog znanja. Izvori koji generiraju takve, kontinuirane, podatke su sve brojniji, a osim već spomenutih izvora s Interneta (društvene mreže, mreže komunikacije), sve se više spominju sustavi Interneta stvari - IoT (engl. *Internet of Things*) objekata, mreže senzora, M2M komunikacije (engl. *machine to machine*) [11].

III. PRIKAZI, PROCESIRANJE I UPRAVLJANJE TOKOVIMA PODATAKA

Podaci koji se sastoje od objekata koji se bilježe u stvarnom vremenu nazivaju se tokovima podataka. Prema [11], tokovi podataka bilježe se s uređenim parom *(S, T)*, gdje je *S* niz n-torki, a *T* niz pozitivnih vremenskih intervala. Niz podataka u toku je potencijalno neograničen, što znači da se tok podataka može nastaviti generirati. Važno je spomenuti da se nizovi tokova podataka mogu pročitati samo jednom ili jako mali broj puta.

Prema [11], pri procesiranju tokova podataka podrazumijevaju se dva načina dobivanja informacija: rad s upitima nad tokovima podataka [12] i dubinska analiza tokova podataka.

A. *Struktura tokova podataka*

Osnovna struktura tokova podataka prethodno je objašnjena. Slika 2. predstavlja isječak iz tablice tokova podataka, gdje su periodički bilježene IP sesije u nekom sustavu. Tok podataka je strukturiran tako da je svaki zapis obilježen vremenskom oznakom koji može biti fizička (datum) ili logička (jednoznačna brojčana oznaka zapisa) [13], čiji je primjer i Slika 2. Nadalje, jedan se zapis definira uređenim parom (t_i, m_i) gdje je t_i spomenuta vremenska oznaka a m_i n-torka zabilježenih vrijednosti. U slučaju kad je vremenska oznaka logičkog tipa, t_i se izjednačava s i.

Timestamp	Source	Destination	Duration	Bytes	Protocol
…	…	…	…	…	…
12342	10.1.0.2	16.2.3.7	12	20K	http
12343	18.6.7.1	12.4.0.3	16	24K	http
12344	12.4.3.8	14.8.7.4	26	58K	http
12345	19.7.1.2	16.5.5.8	18	80K	ftp
…	…	…	…	…	…

Slika 2. Tablični prikaz zapisa IP sesija koje su se bilježili periodički

B. *Tehnike procesiranja tokova podataka*

Sukladno karakteristikama toka podataka, razvijaju se i prilagođavaju postojeći algoritmi za povećavanje efikasnosti upita nad tokovima podataka te njihove dubinske analize. U [14] je dan pregled dotadašnjih rješenja koja su bila temelj radu s tokovima podataka s glavnom podjelom na dvije grupe tehnika: **Tehnike temeljene na podacima** (engl. *data-based*), kao što su uzorkovanje (engl. *sampling*), rasterećenje (engl. *load shedding*) i skiciranje (engl. *sketching*), vođene su idejom sažimanja cijelog seta podataka te idejom odabira dijela dolazećeg toka podataka za analizu npr. sažete strukture podataka(engl. *synopsis data structures*). **Tehnike temeljene na zadacima** (engl. *task-based*) prilagođavaju postojeće tehnike da bi se postigla što veća efikasnost, a radi se o tehnikama poput algoritama aproksimacije (engl. *approximation algorithms*) ili kliznih prozora (engl. *sliding window*).

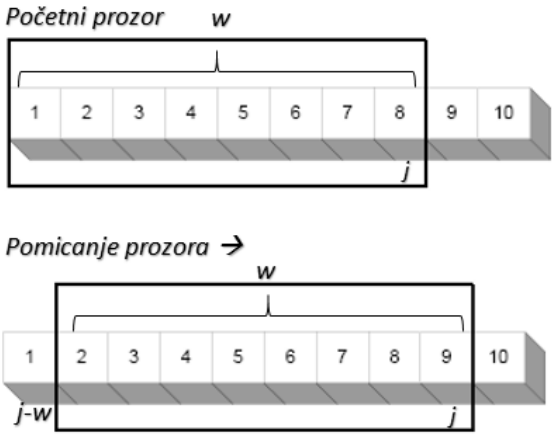
Tehnika uzorkovanja je odavno poznata u statistici, a bazira se na izdvajanju zapisa pri ulazu toka podataka. Cilj je kreirati uzorak takav da je rezultat upita nad uzorkom približan rezultatu upita nad cijelim tokom. Pri određivanju ulazi li određeni zapis u uzorak može se koristiti *hash* funkcija [15] ili neke druge metode za koje nije potrebno poznavati broj zapisa u skupu podataka, npr. tehnika uzorkovanja s rezervoarom (engl. *reservoir sampling technique*). Sličan pristup ima i **tehnika rasterećenja** kojom se odbacuju cijeli nizovi podataka na ulazu kad ulazni podaci prelaze kapacitet sustava [16]. U [17] je tehnika rasterećenja objašnjena te su prikazane dvije varijante odbacivanja niza podataka: odbacivanje slučajno odabranih nizova i odbacivanje nizova podataka prema važnosti. Uzorkovanje ili rasterećenje mogu postati i uzrokom smanjene točnosti rezultata upita nad podacima ili buduće dubinske analize podataka [16]. Naime, odbacivanjem zapisa moguća eliminacija nekih anomalija iz podataka, tj. stršećih podataka (engl. *outliera*). Kod rasterećenja pak, zbog odbacivanja cijelih serija zapisa, moguća je eliminacija podataka potencijalno korisnih za generiranje modela. Tehnika **sažete strukture podataka** koristi metode histograma, malih valova (engl. *wavelet*), momenti frekvencija (engl. *frequency moments*), a rezultati se kasnije koriste za daljnje upite.

Algoritmi **aproksimacije i randomizacije** se koriste za potrebe mjerenja entropije sustava, dubinske analize podataka korištenjem asocijacijskih pravila, grupiranja korištenjem *k-means* algoritma [16]. Ovi algoritmi imaju potencijal za efikasno rješavanje problema dubinske analize tokova podataka, ali ne rješavaju problem ograničenosti resursa pa se podrazumijeva paralelno korištenje i drugih alata u svrhu prilagodbe dostupnih resursa [14].

Prema [13], nudi se nekoliko mogućnosti za dobivanje informacija iz tokova podataka. Primjerice, ako je tražena informacija odgovor na pitanje koliko je bajtova preneseno u svakom intervalu vremena (između vremenskih oznaka), onda se koristi **model vremenskih nizova** (engl. *time-series model*). Ukoliko je tražena informacija poput ukupnog broja bajtova prenesenih od početka postojanja toka koristi se **model blagajne** *(*engl. *cash register model)* koja inkrementalno povećava vrijednost pri prolaženju kroz zapise kako bi se dobila nova, tražena, vrijednost. Model u kojem se vrijednost smanjuje ima naziv **model okretišta** (engl. *turnstile model).*

Također, autori navode da je moguće izdvojiti samo neke zapise u svrhu dobivanja informacije i u tom slučaju sva tri prethodno navedena modela mogu biti korištena. U većini slučajeva, krajnji korisnik ne traži rezultat analize cijelog toka podataka nego samo nekog dijela koji se može ograničiti prozorom, a najčešće su predmet zanimanja nedavni intervali vremena. Prozor može biti **fiksan** (engl. *fixed window*), ukoliko je definiran fiksiranim granicama (npr. 01/01/2015 do 31/01/2015), **djelomično fiksiran** (engl. landmark window) gdje je jedna granica fiksirana na određenu točku u vremenu pa veličina prozora raste s pristizanjem novih podataka (npr. od 01/01/2015 do danas) ili pak može biti **klizni** (engl. *sliding window*), ukoliko se obje granice pomiču (npr. zanima nas informacija za zadnjih 10 minuta). Korisnik, osim prozora mora odrediti i stopu osvježavanja (engl. *refreshment rate)* rezultata, odnosno ponavljanja upita. Klizni prozori funkcioniraju po principu FIFO strukture (*first-in-first-out*), gdje se na navedeni način izmjenjuju podaci. Pri čitanju i ulasku zapisa j u prozor, element j-w, gdje je w veličina prozora, biva zaboravljen [16] (Slika 3).

Prema [18], prozori se mogu definirati na dva načina: fizički ili logički. Veličina prozora se fizički ili temeljeno na vremenskoj oznaci (engl. *time-stamp based*) određuje prema traženom vremenu trajanja te prozor sadrži sve zapise iz traženog vremenskog intervala.



Slika 3. Klizni prozor

Logički ili temeljeno na nizu podataka (engl. *sequence based*) podrazumijeva određivanje veličine prozora prema broju promatranih zapisa. U potonje ubrajamo klizne prozore, djelomično fiksirane te prozore u kojima se zapisima dodaje težina koja se smanjuje s vremenom (engl. *damped window*) [19], [20].

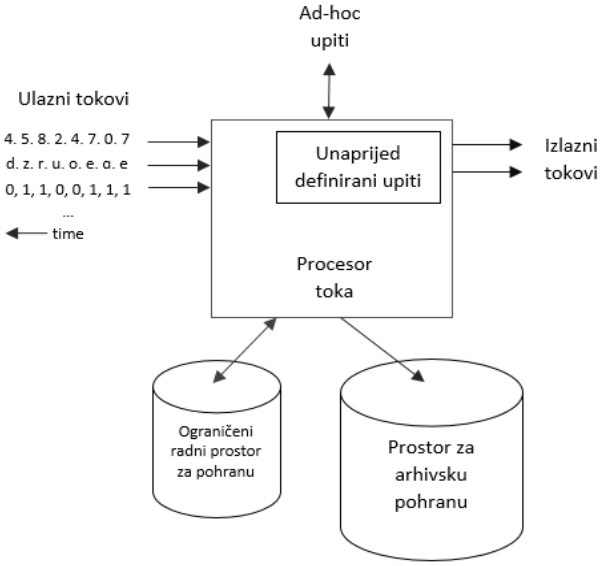
C. *Sustavi za upravljanje tokovima podataka*

Hebrail [13] sustave za upravljanje tokovima podataka (engl. *DSMS - Data Stream Management Systems*) smatra ekstenzijom sustava za upravljanje bazama podataka (engl. *DBMS - Data Base Management Systems*) s podrškom za tokove podataka. Struktura toka podataka definirana je na isti način kao i relacijska baza podataka, ali nema prostornog ograničenja za pohranu. Tok podataka iz izvora pristiže kontinuirano i ukoliko DSMS sustavi nisu u mogućnosti čitati sve podatke koje pristižu, nepročitani podaci nepovratno su izgubljeni.

Nad podacima se provode upiti kakvi se koriste i za relacijske baze u definiranim intervalima vremena. Rezultati mogu kreirati nove tokove podataka i pritom se rezultat ne pohranjuje nego procesira u nastavku ili mogu biti u obliku trajnih tablica koje se ažuriraju po potrebi. Međutim, postoje i slučajevi u kojima je potrebno pohraniti cijeli novi tok podataka. Sukladno navedenom, DSMS sustavi pružaju mogućnosti rada i s trajnim tablicama i tokovima podataka.

DSMS sustavi su prikazani slikom 4. Prema [15], u sustav se istovremeno može učitavati više tokova podataka i svaki od njih može imati vlastitu dinamiku ulaska u sustav. DSMS sustav nema kontrolu nad brzinom ulaska podataka, za razliku od DBMS sustava koji kontrolira brzinu ulaska podataka i time je osiguran od gubitka podataka za vrijeme pokretanja upita.

Tokovi se pohranjuju u bazu za arhivsku pohranu koja je velikih dimenzija, ali se upiti nad njom u pravilu ne izvršavaju jer su mogući samo u specijalnim okolnostima i kroz dugotrajni proces. Dijelovi toka koji se obrađuju smještaju se u ograničeni radni prostor za pohranu nad kojim se izvršavaju upiti.



Slika 4. Primjer sustava za upravljanje tokovima podataka (prilagođeno iz [15])

Upiti se u sustavu pojavljuju u dva oblika: unaprijed definirani (fiksni) upiti su ugrađeni u sustav i pokreću se konstantno (npr. produciraju upozorenje kad neki tok generira neku vrijednost) dok se *ad-hoc* upiti postavljaju jednom i tiču se trenutnog stanja toka ili tokova. Točnost odgovora na upite ovisi o karakteristikama sustava i ugrađenim tehnikama za rad s tokovima, npr. kliznom prozoru.

IV. METODE DUBINSKE ANALIZE TOKOVA PODATAKA

Domingos i Hulten su se 2000. godine u [21] osvrnuli na dotad prisutne metode u dubinskoj analizi podataka uzimajući u obzir potrebu za analizom tokova podataka. Istaknuli su da su resursi trenutnih metoda (vrijeme, memorija i veličina uzorka) ograničeni te da su te metode pogodne za statičke podatke s fiksnom veličinom uzorka. Ograničenja DSMS sustava su prepoznata i u dubinskoj analizi tokova podataka: prolazi se jednom kroz podatke, a resursi su ograničeni (memorija, procesor).

Odabir algoritma dubinske analize podataka ovisi o odabranom tipu prozora. Ukoliko se radi o cijelom toku podataka, idealni algoritmi su neuralne mreže ili stabla odluke. Ukoliko se bira klizni prozor, obično se odabiru algoritmi s mogućnošću da „zaboravljaju“ prošlost toka (npr. PCA metoda). Ako se analizira neki dio prošlog vremena koji nije unaprijed određen, algoritam bi trebao moći čuvati neke rezultate u privremenoj memoriji (npr. mikroklastering). U nastavku su prikazani neki od pristupa u dubinskoj analizi podataka, primijenjeni na tokove podataka.

A. *Klasifikacija*

Klasifikacija je nadzirani pristup dubinske analize podataka, koji na temelju skupa podataka za učenje oblikuje model za predviđanje kvalitativne vrijednosti atributa klase. U [22] se ističe kako je tehnika klasifikacije nastajala pod pretpostavkom da skupovi za učenje stanu u raspoloživu memoriju. S vremenom su

skupovi podataka s puno većim obujmom postali uobičajeni te autori navode klasifikatore razvijene kako bi mogli vršiti dubinsku analizu podataka nad podacima koji ne stanu u memoriju na način se vrši sekvencijalno skeniranje trenutnih podataka u memoriji, a primjeri tih klasifikatora su SLIQ [23], SPRINT [24] i CLOUDS [25]. Navedeni algoritmi temeljeni su na stablima odluke.

Metoda Hoeffdingovih stabala, prikazana u [21], temelji se na rastućem stablu odluke koje omogućava učenje iz *online* tokova podataka čiji volumen eksponencijalno raste. Hoeffdingova stabla polaze od činjenice da mali uzorak podataka može biti dovoljan za odabir optimalnog atributa za dijeljenje. Hoeffdingova granica je broj koji kvantificira broj opažanja (zapisa u bazi podataka) potrebnih za procjenu neke statistike sa zahtijevanom preciznošću, a dobiva se na sljedeći način: Neka je *r* realna varijabla slučajne vrijednosti iz ranga R i neka postoji n zapisa u bazi podataka o vrijednosti *r*, a njihova izračunata srednja vrijednost je $\bar{r}$. S vjerojatnošću *1 - δ*, srednja vrijednost varijable je najmanje $\bar{r} - \mathcal{E}$, gdje je

$$\mathcal{E} = \sqrt{\frac{R^2 \ln(\frac{1}{\delta})}{2n}}. \qquad (1)$$

Autori Domingos i Hulten su zatim predstavili VFDT (engl. *Very Fast Decision Tree Learner*) kao sustav za dubinsku analizu podataka visokih performansi koji se temelji na Hoeffdingovim stablima. VFDT algoritam se temelji na korištenju malog seta zapisa za izradu testa za odvajanje u nekom čvoru. Potrebno je u određenom čvoru dobiti zadovoljavajući rezultat na statističkom testu (Hoeffdingovu granicu) pa VFDT ispituje nove zapise dok ne postigne rezultat za odluku. Stablo nastaje rekurzivnim smještanjem čvorova odluke na mjesto listova, a svaki čvor pohranjuje statističke vrijednosti za dostupne atribute. Pri učitavanju novog zapisa, algoritam prolazi stablom odluke od korijena prema listovima, pritom evaluirajući vrijednosti atributa u svakom čvoru, što mu određuje smjer kretanja kroz grane. U trenutku dolaska u čvor lista izvršava se, ranije spomenuti, statistički test prema kojem se bira atribut za odluku, a na čvor će se nadovezati onoliko novih listova, koliko vrijednosti atribut može poprimiti [16].

U [26] autori predlažu algoritam baziran na CART stablu odluke koji pri odabiru atributa za odluku u čvoru lista koristi Gaussovu aproksimaciju. Kroz istraživanje su dokazali da je atribut koji se izdvoji pri primjeni modificiranog CART algoritma, nazvanim dsCART, upravo atribut koji se izdvaja iz cijelog toka podataka, što potvrđuje točnost algoritma.

Aggarwal [27] navodi još neke metode klasifikacije kao što su klasifikacija na zahtjev (engl. *on demand classification*), klasifikacija bazirana na cjelini (engl. *ensemble-based classification*) i metode bazirane na sažimanju (engl. *compression-based methods*).

B. *Grupiranje*

Grupiranje je nenadzirana tehnika strojnog učenja koja se provodi nad objektima na način da se oni grupiraju u klastere prema zajedničkim osobinama (sličnim vrijednostima). Analogno tome, tehnike grupiranja za tokove podataka neprekidno grupiraju objekte uz potencijalna vremenska ograničenja [11]. Prema [16], proces grupiranja dijeli se u dva sloja: prvi sloj generira lokalne modele (mikroklastere), dok drugi sloj generira globalni model iz lokalnog.

Domingos i Hulten su, paralelno s VFDT metodom, unutar pristupa VFML (engl. *Very Fast Machine Learning*) razvili i VFKM varijantu [28], koja također koristi Hoeffdingovu granicu i to na način da se njome određuje broj zapisa u svakom koraku K-means algoritma. Ta metoda izvršava niz K-means pokretanja te pri svakom pokretanju koristi sve više zapisa, dokle god nije zadovoljena Hoeffdingova granica.

U počecima razvoja metoda grupiranja tokova podataka razvijene su još neke metode poput poboljšanja k-means algoritma koje vrši grupiranje nad binarnim tokovima podataka [29] ili kombinacije metoda STREAM i LOCALSEARCH [30].

Metoda BIRCH spada u hijerarhijske metode grupiranja [31], a njena karakteristika je gradnja CF (engl. *clustering feature*) stabla, u kojem svaki čvor sadrži informacije o svojoj djeci. Navedeni algoritam je primjer kombinacije klasifikacije i grupiranja. Na tom algoritmu počiva sustav *CluStream*, predstavljen u [32], kao tadašnja novost u grupiranju tokova podataka. U sustavu se može provoditi grupiranje u bilo kojem intervalu vremena, ukoliko su podaci numerički, a sastoji se od dvije komponente: *online* statističkog prikupljanja podataka i *offline* analitičke komponente. Dva su koncepta bila ključna pri dizajnu metode: mikroklasteri, koji su omogućili registraciju detaljnijih podataka unutar klastera u odnosu na dotad korištenu k-means metodu, a time i poboljšanu točnost, i piramidalni vremenski okviri koji predstavljaju pohranu mikroklastera, kao snimaka podataka u vremenu, u piramidalnoj shemi. Takva shema omogućava ravnotežu između uvjeta skladištenja i mogućnosti da se dobije rezultat analize podataka iz vremenskih razdoblja po želji. Još jedna hijerarhijska metoda je CURE (engl. *Clustering Using REpresentatives*).

C. *Otkrivanje čestih uzoraka*

Otkrivanje čestih uzoraka (engl. *frequent pattern mining*) se temelji na izdvajanju skupova zapisa ili vrijednosti koje se često pojavljuju zajedno te tako tvore uzorak u ponašanju toka. Pri otkrivanju uzoraka može se koristiti tehnika prozora ili promatrati cijeli skup podataka [27]. Za korištenje tehnike kliznog prozora karakteristična je pretpostavka da česti uzorci nisu brojni te da će biti moguće sve uzorke zadržati u kliznim prozorima u glavnoj memoriji. Uz klizne prozore, koristiti se mogu i prozori u kojima se zapisima dodaje težina. Ukoliko odabrani pristup uključuje promatranje

cijelog skupa podataka, uzorke je potrebno pronaći pri prvom i jedinom prolasku kroz tok pa se koriste tehnike prilagođene takvom načinu rada, npr. skice (engl. *sketches*).

D. Otkrivanje promjene

Obzirom da se uzorci unutar podataka mogu razvijati kako vrijeme prolazi i pristižu novi zapisi, preporučljivo je evidentirati njihove promjene te vršiti analizu nad njima. Za otkrivanje promjene (engl. *change detection*) koristi se nekoliko metoda, među kojima su i neki algoritmi grupiranja i klasifikacije, a izdvaja se metoda procjene brzine i gustoće (engl. *velocity density estimation*). U toj se metodi računa stopa promjene gustoće podataka u različitim vremenskim točkama.

Pri otkrivanju promjene i otkrivanju čestih uzoraka koriste se i tehnike asocijacijskih pravila (*Apriori, FP-tree, FR-stream algorithm*).

E. Otkrivanje stršećih vrijednosti i anomalija

Stršeće vrijednosti pojavljuju se u raznim tipovima podataka pa tako i u tokovima podataka. Zapise koji se u većoj mjeri razlikuju od ostalih potrebno je izdvojiti. Takvi se zapisi detektiraju praćenjem podataka kroz vrijeme korištenjem metoda koje se temelje na konstantnom ažuriranju trendova podataka sukladno s pristizanjem novih zapisa u sustav [33]. Algoritmi za otkrivanje stršećih vrijednosti (engl. *outlier detection*) često se temelje na ispitivanju razlike između određenih parova zapisa te se traže oni s najvećom razlikom [11].

V. PRIMJENA DUBINSKE ANALIZE TOKOVA PODATAKA

Podaci tokovne prirode produkt su evidentiranja događaja u raznim djelatnostima i aktivnostima pa se tako i dubinska analiza tokova podataka provodi prema potrebi nad raznim podacima.

Edukacija zahtijeva korištenje raznih digitalnih pomagala, pa je tako u svakoj visokoškolskoj ustanovi u upotrebi neki od sustava za e-učenje. Za takve je sustave tipično da generiraju podatke o korištenju sustava, a ukoliko se radi o *Moodle* sustavu, onda je to pregledno evidentirano i vremenskom oznakom te predstavlja tok podataka. U [34], Romero, Ventura i Garcia ponudili su široki pregled korištenja dubinske analize podataka u *Moodle* sustavima za učenje. Primijenili su dubinsku analizu podataka u edukaciji (engl. *Educational Data Mining – EDM*) i sugerirali predavačima, nastavnicima i voditeljima sustava za e-učenje kako da provedu svoja istraživanja nad podacima koji su im na raspolaganju.

U [35], zapisi iz *Moodle* sustava *Mudri* evidentirani na jednom od uvodnih kolegija programiranja, uspoređeni su s podacima koji se tiču prethodnog znanja studenata, koristeći metode klasifikacije C4.5, JRip i PART. Zaključeno je da je aktivnost u sustavu za učenje imala veći utjecaj na završni uspjeh, odnosno prolaz kolegija.

Pronalaženje asocijacijskih pravila unutar seta podataka provedeno je u [36], kako bi se analizirale specifične aktivnosti na uvodnom kolegiju programiranja. Apriori metodom dobivena su značajna pravila i zatim su izglasavana od strane eksperata te su pokazala da video-lekcije koje su dostupne u *Moodle* sustavu imaju pozitivan utjecaj na usvajanje znanja na kolegiju.

U [37] analizirana su studentska izvješća kako bi se ekstrahiralo korisno znanje i istražila mogućnost predviđanja finalnog uspjeha pomoću njih na kolegiju Algoritmi i strukture podataka. Na navedenim tekstovima provedena je dubinska analiza teksta (engl. *text mining*) te je pronađena veza između kvalitete tekstualnih izvješća i finalnih rezultata na kolegiju.

Podaci skupljeni do polovice semestra pokazali su, u [38], potencijal pri predviđanju studenata koji nisu uspješno okončali semestar. Provedeno je grupiranje nad zapisima iz Mudri sustava, točnije primijenjen je K-means algoritam, koji je izdvojio rizičnu skupinu u zaseban klaster. Takvi su studenti kandidati za slanje ranog upozorenja (engl. *early alert*) u nadolazećim akademskim godinama. Također, nad podacima je provedena tehnika SMOTE, kako bi se stvorio podjednak broj studenata među uspješnima i neuspješnima te je to povećalo točnost i smanjilo broj pogrešno grupiranih studenata.

Vremenska oznaka u navedenim istraživanjima nije uzimana u obzir , tako da ni sami podaci nisu tretirani kao tokovi.

Učinkovitost metoda za rad s tokovima podataka ne ovisi o izvoru podataka. Priroda podataka koje generiraju senzori dovela je do toga da se brojne metode razvijaju upravo na senzorskim podacima [27]. Vojne aplikacije prikupljaju velike količine podataka kroz više vrsta izvora, npr. monitoring aktivnosti u svrhu detekcije prijetnji [39]. Velike količine podataka generiraju, također, aplikacije koje se koriste u radu svemirskih postaja ili promatračnica. Senzori su ugrađeni i u mobilne uređaje te se podaci mogu kontinuirano isporučivati i koristiti u svrhu integriranja podataka, najčešće o kretanju i potrošačkim navikama [40]. Slično mobilnim uređajima, osmišljeni su uređaji koji se ugrađuju u odjeću ili dodatno koriste u zdravstvene ili sportske svrhe [41]. U vozilima postoje razni sustavi koji generiraju tokove podataka koji se mogu koristiti u svrhu dijagnostike, navigacije, upozoravanja na opasnosti ili nepredviđene situacije, a integracija tokova podataka iz različitih izvora unutar istog sustava (vozila) temelj je za razvoj pametnih aplikacija koje bi unaprijedile korištenje vozila. Ideje nekoliko aplikacija, čija obilježja nalazimo u nekim već postojećim aplikacijama, a temelje se na gradnji prediktivnih modela nad tokovima podataka, prezentirane su u [42]. Spomenute aplikacije nude rješenje za pomoć pri traženju mjesta za parkiranje, automatski poziv za pomoć na cesti pri sudaru, povezivanje podataka o brzini vozila s detekcijom promjena u ritmu otkucaja srca ili umora kao čestih uzročnika prometnih nesreća te za dijagnostiku potencijalnih kvarova na vozilu s promptnim ili naknadnim alarmiranjem, ovisno o hitnoći.

Upravljanje hladnim lancima u transportu hrane se vrši u svrhu očuvanja svježine i sigurnosti proizvoda [43]. Bežični senzori su sastavni dio lanaca u kojima se prati prisutnost nekih parametara koje utječu na kvalitetu hrane. Bežična mreža senzora (WSN – engl. *wireless sensor network*) sastoji se od čvorova koji skupljaju podatke, vrše komunikaciju, a ovisno o vrsti čvora, i računanja [44]. Dubinska analiza podataka, u ovakvim slučajevima kao i u drugim slučajevima gdje se javlja struktura Interneta Stvari, podrazumijeva rad u realnom vremenu s višestrukim izvorima [19].

Osim senzorskih podataka, prirodu toka imaju i podaci generirani primjerice na Internetu, bilo pri širenju podataka mrežnom infrastrukturom ili pak na webu, bilježenjem kretanja web stranicama, klikova, pretraga i aktivnosti na društvenim mrežama. Slike iz izvora kao što su sateliti ili video-nadzor se automatizirano pohranjuju u određenim intervalima te tako tvore tokove za analizu [15]. Takav tip podataka zahtijeva prvenstveno tehnologiju za obradu slika. Burzovni podaci su, također, izvor velikih skupova tokova podataka koji se analiziraju u realnom vremenu kao podrška odlučivanju. Primjer korištenja dubinske analize podataka nad burzovnim zapisima je prikazan u [45], a korištene su metode grupiranja, Bayesovih mreža i stabala odluke, prilagođene strukturi toka podataka. Navedene metode i druge metode dubinske analize tokova podataka koriste se i pri praćenju kretanja ljudi, stvaranju društvene slike, praćenju kupovanja na akcijama, klimatskih promjena i kretanja temperature [27]. U zadnja dva spomenuta područja, ponovo su od velike važnosti senzori, a koriste se razni parametri iz okoline za predviđanje promjena u temperaturi, dolaska oluja i vjetrova velikog intenziteta, globalnog zatopljenja, itd. U [46] objašnjene su metode za predviđanja vremenskih i klimatskih uvjeta, za što je od prethodno navedenih metoda korištena detekcija promjene. U tom području, tokovi podataka se sastoje od numeričkih vrijednosti, ali dolaze i u obliku satelitskih snimaka.

VI. Zaključak

U ovom radu dan je uvod u procesiranje velikih skupova podataka s naglaskom na tokove podataka. Tokovi podataka predstavljaju izazov za algoritme dubinske analize podataka jer su kod potonjih prisutna ograničenja u odnosu na veliki obujam podataka i tempo pristizanja novih. U području dubinske analize tokova podataka proveden je značajan broj istraživanja, ali su neka područja unutar domene još uvijek neistražena ili pak neotkrivena.

Među tehnikama za procesiranje tokova podataka u ovom radu izdvojeni su sažeci, prozori, uzorci, itd., a prije svega je važno da rezultirajući skupovi podataka preslikavaju cijeli tok podataka i time osiguravaju točniji rezultat nakon primjene algoritama dubinske analize podataka. U radu su prikazane neke od metoda dubinske analize toka podataka. Opisana su i područja u kojima nastaju tokovi podataka te primjena dubinske analize tokova podataka na određene slučajeve iz gospodarstva, zdravlja, tehnologije, itd.

Većina istraživanja polazi od činjenice da postoji ograničenje od jednog, ili vrlo malog broja prolaska kroz podatke, što često uzrokuje gubitke informacija. U svrhu povećanja učinkovitosti i uz pretpostavku da će obujam i kompleksnost tokova podataka u budućnosti rasti, otvara se prostor za istraživanja i razvoj novih metoda koje nisu temeljene na jednom prolasku kroz podatke.

Kako je veći dio tokova podataka okarakteriziran kao senzorski, određene manjkavosti senzora mogu se preslikati i na kvalitetu podataka pa često dolazi do šumova među podacima a modeli koji se na njima grade imaju smanjenu točnost. U rad senzora mogu se uključiti dodatni algoritmi, koji bi prema modelima „čistili“ podatke koji pristižu u neki sustav. Dinamika pristizanja podataka iz senzora u neki sustav i stvaranje modela u realnom vremenu nameću potrebu za rješenjem koje će nadvladati navedene izazove.

Literatura

[1] I. H. Witten i E. Frank, “Data Mining: Practical Machine Learning Tools and Techniques,” Elsevier, San Francisco, vol. 2 edition. 2005.

[2] L. Douglas, “3d data management: Controlling data volume, velocity and variety,” Application Delivery Strategies, vol. 6, p. 4, 2001.

[3] W. Fan i A. Bifet, “Mining Big Data : Current Status , and Forecast to the Future,” ACM SIGKDD Explorations Newsletter, vol. 14, no. 2, pp. 1–5, 2013.

[4] M. van Rijmenam, “Why The 3V’s Are Not Sufficient To Describe Big Data,” Big Data startup, 2015. [Online]. Dostupno: https://datafloq.com/read/3vs-sufficient-describe-big-data/166. [Pristupano: 09-Nov-2016].

[5] X. Wu, X. Zhu, G.-Q. Wu, i W. Ding, “Data Mining with Big Data,” Knowledge Data Engineering IEEE Transactions, vol. 26, no. 1, pp. 97–107, 2014.

[6] D. Centola, “The spread of behavior in an online social network experiment,” Science (80-.)., vol. 329, no. 5996, pp. 1194–1197, 2010.

[7] P. V. Bindu i P. S. Thilagam, “Mining Social Networks for Anomalies: Methods and Challenges,” Journal of Network and Computer Applications, vol. 68, pp. 213–229, 2016.

[8] E. Y. Chang, B. Hongjie, i Z. Kaihua, “Parallel Algorithms for Mining Large-Scale Data,” u Proceedings of the 17th ACM international conference on Multimedia, 2009, pp. 917–918.

[9] X. Wu, C. Zhang, i S. Zhang, “Database classification for multi-database mining,” Information Systems., vol. 30, no. 1, pp. 71–88, 2005.

[10] K. Su, H. Huang, X. Wu, i S. Zhang, “A logical framework for identifying quality knowledge from different data sources,” Decision Support Systems, vol. 42, no. 3, pp. 1673–1683, 2006.

[11] D. Namiot, “On Big Data Stream Processing,” International Journal of Open Information Technologies, vol. 3, no. 8, pp. 48–51, 2015.

[12] C. C. Aggarwal i J. Wang, “Data Streams: Models and Algorithms,” Data Streams, vol. 31, pp. 9–38, 2007.

[13] G. Hebrail, “Data stream management and mining,” Mining Massive Data Sets for Security, pp. 89–102, 2008.

[14] M. M. Gaber, A. Zaslavsky, i S. Krishnaswamy, “Mining data streams: A Review,” ACM Sigmod Records, vol. 34, no. 2, pp. 18–26, 2005.

[15] J. Leskovec, A. Rajaraman, i J. D. Ullman, “Mining Data Streams,” u Mining of massive datasets, 2014, pp. 131–162.

[16] J. Gama, “A survey on learning from data streams: current and future trends,” Progress in Artificial Intelligence, vol. 1, pp. 45–55, 2012.

[17] N. Tatbul, U. Çetintemel, S. Zdonik, M. Cherniack, i M. Stonebraker, “Load Shedding in a Data Stream Manager,” 29th International Conference VLDB, vol. 54, pp. 309–320, 2003.

[18] B. Babcock, M. Datar, i R. Motwani, “Sampling From a Moving Window Over Streaming Data,” u Proceedings of 13th Annual ACMSIAM Symposium on Discrete Algorithms, 2002, vol. 102, no. 2001–33, pp. 633–634.

[19] P. Juric, M. Brkic Bakaric, X. Wang, X. Zhang, i M. Matetic, “Mining Data Streams for the Analysis of Parameter Fluctuations in IoT-Aided Fruit Cold-Chain,” odabrano za objavu u Proceedings of the 27th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2016.

[20] N. Jiang i L. Gruenwald, “Research Issues in Data Stream Association Rule Mining,” SIGMOD Record., vol. 35, no. 1, 2006.

[21] P. Domingos i G. Hulten, “Mining High-Speed Data Streams,” u Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 71–80, 2000.

[22] M. J. Zaki i R. Agrawal, “Parallel classification for data mining on shared-memory multiprocessors,” u Proceedings 15th International Conference on Data Engineering, pp. 198–205, 1999.

[23] M. Mehta, R. Agrawal, i J. Rissanen, “SLIQ: A fast scalable classifier for data mining,” Advances in Database Technology — EDBT '96 SE - 2, vol. 1057, pp. 18–32, 1996.

[24] J. Shafer, R. R. Agrawal, i M. Mehta, “SPRINT: A scalable parallel classfier for data mining,” u Proceedings of 1996 International Conference of Very Large Data Bases, pp. 544–555, 1996.

[25] K. Alsabti, S. Ranka, i V. Singh, “CLOUDS : Tree Classifier for Large Datasets,” Proceedings of the Fourth Knowledge Discovery and Data Mining Conference, pp. 2–8, 1998.

[26] M. Jaworski et al., “The CART decision tree for mining data streams,” Information Sciences. (Ny)., 2014.

[27] C. C. Aggarwal, “Mining sensor data streams,” u Managing and Mining Sensor Data, vol. 9781461463, Springer US, 2013, pp. 143–171.

[28] P. Domingos i G. Hulten, “A general method for scaling up machine learning algorithms and its application to clustering,” u Machine Learning-International Workshop Then Conference, pp. 106–113, 2001.

[29] C. Ordonez, “Clustering binary data streams with K-means,” u Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, pp. 12–19, 2003.

[30] L. O’Callaghan, N. Mishra, A. Meyerson, S. Guha, i R. Motwani, “Streaming-data algorithms for high-quality clustering,” u Data Engineering, 2002. Proceedings. 18th International Conference on, pp. 685–694, 2002.

[31] J. Gama i M. M. Gaber, Learning from Data Streams: Processing Techniques in Sensor Networks. 2007.

[32] C. C. Aggarwal, T. J. Watson, R. Ctr, J. Han, J. Wang, i P. S. Yu, “A Framework for Clustering Evolving Data Streams,” u Proceedings of the 29th international conference on Very large data bases, pp. 81–92, 2003.

[33] M. Gupta, J. Gao, C. Aggarwal, i J. Han, „Outlier Detection for Temporal Data“, Synthesis Lectures on Data Mining and Knowledge Discovery, vol. 5, no. 1. 2014.

[34] C. Romero, S. Ventura, i E. García, “Data mining in course management systems: Moodle case study and tutorial,” Computers and Education, vol. 51, no. 1, pp. 368–384, 2008.

[35] S. Sisovic, M. Matetic, i M. B. Bakaric, “Mining Student Data to Assess the Impact of Moodle Activities and Prior Knowledge on Programming Course Success,” u Proceedings of the 16th International Conference on Computer Systems and Technologies, pp. 366–373, 2015.

[36] M. Matetic, M. B. Bakaric, i S. Sisovic, “Association Rule Mining and Visualization of Introductory Programming Course Activities,” u Proceedings of the 16th International Conference on Computer Systems and Technologies, 2015, pp. 374–381.

[37] M. B. Bakaric, M. Matetic, i S. Sisovic, “Text Mining Student Reports,” u Proceedings of the 16th International Conference on Computer Systems and Technologies, pp. 382–389, 2015.

[38] S. Sisovic, M. Matetic, i M. B. Bakaric, “Clustering of imbalanced moodle data for early alert of student failure,” u SAMI 2016 - IEEE 14th International Symposium on Applied Machine Intelligence and Informatics - Proceedings, pp. 165–170, 2016.

[39] J. Yin, Q. Yang, i J. J. Pan, “Sensor-based abnormal human-activity detection,” IEEE Transactions on Knowledge and Data Engineering., vol. 20, no. 8, pp. 1082–1090, 2008.

[40] A. Krause, E. Horvitz, A. Kansal, i F. Zhao, “Towards community sensing,” u Proceedings - 2008 International Conference on Information Processing in Sensor Networks, IPSN 2008, pp. 481-492, 2008

[41] H. Ghayvat, J. Liu, S. C. Mukhopadhyay, i X. Gui, “Wellness Sensor Networks: A Proposal and Implementation for Smart Home for Assisted Living,” IEEE Sensors Journal, vol. 15, no. 12, pp. 7341–7348, 2015.

[42] M. Swan, “Connected Car: Quantified Self becomes Quantified Car,” Journal of. Sensor and Actuator Networks, vol. 4, no. 1, pp. 2–29, 2015.

[43] C.-W. Shih i C.-H. Wang, “Integrating wireless sensor networks with statistical quality control to develop a cold chain system in food industries,” Computer Standards & Interfaces, vol. 45, no. February, pp. 62–78, 2016.

[44] A. Z. Abbasi, N. Islam, i Z. A. Shaikh, “A review of wireless sensors and networks’ applications in agriculture,” Computer Standards & Interfaces vol. 36, no. 2, pp. 263–270, 2014.

[45] H. Kargupta, B. Park, S. Pittie, L. Liu, D. Kushraj, i K. Sarkar, “MobiMine: Monitoring the stock market from a PDA,” Newsletter ACM SIGKDD Explorations Newsletter, vol. 3, no. 2, pp. 37–46, 2002.

[46] A. Garg, V. Mithal, C. Vijay, M. Dunham, K. Vikrant et al. “Gopher: Global observation of Planetary Health and Ecosystem Resources,” u International Geoscience and Remote Sensing Symposium (IGARSS), 2011, pp. 1449–1452.