INFOTEH-JAHORINA Vol. 9, Ref. E-I-2, p. 423-426, March 2010.

EKSTRAKCIJA SVOJSTAVA PODATAKA, PREDSTAVLJANJE I OTKRIVANJE ZNANJA U KONTEKSTU MULTIMEDIA DATA MINING

DATA FEATURE EXTRACTION, PRESENTATION AND DISCOVERY OF KNOWLEDGE IN THE CONTEXT OF MULTIMEDIA DATA MINING

Milorad Banjanin[1], Fakultet tehničkih nauka Novi sad
Igor Lazarević[2], Fakultet tehničkih nauka Novi Sad, postdiplomac
Anton Vrdoljak[3], Građevinski fakultet Sveučilišta u Mostaru
Tijana Gavrić-Mičić[4], Saobraćajni fakultet Doboj, postdiplomac

***Apstract**: Danas je dobro poznato da su multimedia (multimedijalne informacije) sveprisutne i često zahtevane, ako ne i neophodne, u mnogim aplikacijama. S druge strane, data mining je takođe raznolika, interdisciplinarna i transdisciplinarna istraživačka oblast, koja je originalno započela sa otkrivanjem znanja u bazama podataka, ali je danas napredovalo mnogo izvan područja baza podataka. Aktuelna literatura iz oblasti pretraživanja multimedijalnih podataka fokusira se na dve paradigme učenja ili teorije pretraživanja-teorija statističkog učenja i teorija soft računarstva, koje se mogu koristiti odvojeno ili zajedno u određenim aplikacijama pretraživanja. Sistem za pretraživanje multimedijalnih podataka funkcioniše kroz tri koraka izvođenja : 1. ekstrakcija svojstava sirovih multimedijalnih podataka kroz predstavljanje svojstava u apstraktnom prostoru (prostor svojstava) 2. predstavljanje znanja, za podršku aktivnostima očekivanog otkrivanja znanja u multimedijalnoj bazi podataka. 3. otkrivanje znanja u multimedijalnoj bazi podataka kroz stvarno pretraživanje ili teorija i/ili tehika učenja, što čini srž sistema za pretraživanje multimedijalnih podataka.*

Ključne reči: multimedia, data mining, prostor svojstava, predstavljanje znanja, otkrivanje znanja

1. UVOD

Dve posebne interdisciplinarne i multidisciplinarne oblasti u računarskoj nauci, čiji je razvoj započet tokom prethodne decenije, a poslednjih godina ubrzan sa značajnim aplikacionim zahtevima, su MULTIMEDIA i DATA MINING. Multimedija je kombinacija teksta, grafike, zvuka i slike u jedan objekat sa kojim se može ostvariti interakcija uz upotrebu računara. Dok se reč *multimedija* odnosi na kombinaciju više vrsta medija, multimedija kao istraživačka oblast se odnosi na studiju i razvoj efektivnog i efikasnog multimedijalnog sistema orijentisanog na specifičnu aplikaciju. Multimedijalni podaci, po tipu numerički, tekstualni, audio, slikovni, grafički, video, animirani, formiraju sveprisutne i neophodne informacije u mnogim aplikacijama. Zbog toga, „istraživanje u multimediji pokriva veoma širok spektar subjekata, od multimedijalnog indeksiranja i pronalaženja, multimedijalnih baza podataka, multimedijalnih mreža, multimedijalnih prezentacija, multimedijalnog kvaliteta usluga, multimedijalne upotrebe i korisničke studije, do multimedijalnih standarda“ [1].

Termin DATA MINING ili *pretraživanje podataka* se može opisati kao proces identifikacije novih, potencijalno korisnih i razumljivih uzoraka (patterns) i odnosa (relationships) među podacima u bazama podataka. Prema Faloutsos-u [2] i Fishman-u [3] , data mining podrazumeva „otkrivanje znanja, koje je originalno započeto u bazama podataka a već je napredovalo mnogo izvan tog područja“

2. NOVE PARADIGME UČENJA ILI TEORIJE SISTEMA PRETRAŽIVANJA PODATAKA

Danas se u istraživanju otkrivanja podataka zahtevaju napredni i alati i teorije, posebno iz matematike, statistike, mašinskog učenja i prepoznavanja modela, dok su eksplozija potreba skladištenja podataka i prisustvo multimedijalnih podataka skoro svuda doprineli preorijentaciji razvoja tradicionalnih baza podataka u skladišta podataka. Istovremeno se „tradicionalni struktuirani podaci“ razvijaju u nestruktuirane podatke tipa slikoviti podaci, podaci vremenskih serija, prostorni podaci, video, audio i opšti multimedijalni podaci. Ovakav trend je posebno „prepoznatljiv u oblastima kao što su umetnost, dizajn, hipermedijska i digitalna medijska produkcija, rezonovanje na osnovu slučaja (CBR) i računarsko modelovanje kreativnosti, uključujući i evolucionarno računarstvo i medicinske multimedijalne podatke“[1]. Rezultat toga je „sve veće interesovanje za nove tehnike i alate koji mogu detektovati i otkriti modele pretraživanja koji vode do novog znanja u domenu problema gde su podaci sakupljeni“.

Takođe, sve je veće interesovanje za analizu multimedijalnih podataka generisanih u različitim distributivnim aplikacijama, kao što su *kolaborativna virtuelna okruženja*, *virtuelne zajednice* i *multi-agentni sistemi*. U oblasti pretraživanja multimedijalnih podataka posebno su aktuelne dve paradigme učenja ili teorije pretraživanja-*teorija statističkog učenja* i *teorija soft računarstva*, koje se mogu koristiti odvojeno ili zajedno u određenim aplikacijama pretraživanja.

Teorija statističkog učenja fokusira statističke funkcije koje uključuju histograme, promenu koeficijenata, povezane vektore i korelograme. Histogrami kao poznati grafikoni, datiraju iz rane literature uzoraka prepoznavanja i analize slike (5). Histogram je statistički metod za pretvaranje originalne prezentacije podatka u učestalo pojavljivanje informacija, merenu za posebnu količinu u originalnom podatku: prema tome histogram je predstavljen kao jednodimenzionalan vektor, gdje je X-osovina niz specifičnih količina , a Y-osovina je učestalo ponavljanje informacija, mereno za svaku vrednost u nizu specifičnih količina. Specifična količina zavisi od različitih modaliteta podataka i takodje od različitih aplikacija i obično je definisano unapred od korisnika.

Povezani vektori prvo su predloženi u ranim danima nalaženja slika sredinom 19.veka, a koristili su se u ranoj literaturi pronalaženja slika i prvobitno su se razvili za pronalaženje slika u boji. Iako je poznato da histogrami nisu jedinstveni za predstavljanje jedinice multimedijalnog podatka, povezani vektori su predloženi da bi poboljšali jedinstvenost. Ono što je bitno za ideju o povezanim vektorima je da treba da spoji prostorne informacije u histogram. Stoga, povezan vektor je definisan na vrhu regularnog histograma koji je vektor. Dok povezani vektori objedinjuju prostorne informacije u odlike histograma obeležavanjem tačke podataka u svaki bucket histograma u dve grupe- povezanu i nepovezanu, kroz povezano traženje komponenata- korelogrami su korak dalje u objedinjavanju prostornih informacija u funkcijama histograma. Svaki zvuk je u suštini multimedijalni podatak. Kao digitalni zvuci, sve različite matematičke promene mogu im biti priložene i mogu se crtati iz njihovih prvobitnih područja u različite oblasti, koje se zovu *učestale oblasti*. Dakle, koeficijent ovih promena šifrira statističke raspodele multimedijalnog podatka u svojim digitalnim područjima kao rapsodela energije u područjima učestalosti. Stoga, koeficijenti ovih promena takođe se mogu koristiti kao funkcija za predstavljanje originalnog multimedijalnog podatka.

Sve statističke funkcije daju samo statistički opis skupa podataka iz koga nije moguće očekivati da se otkriju originalne informacije. One nisu jedinstvene za segmentaciju jedinica u identifikovane delove već su primenjene na celu jedinicu multimedijalnog podatka. Jedinica multimedijalnog podatka je tipično definisana kao specifični modalitet podatka. Npr. za audio tok (strim) jedinica je audio frejm, za kolekciju slika jedinica je slika, a za video tok (strim) jedinica je video frejm. Deo jedinice multimedijalnog podatka zove se objekat. Objekat je dobijen segmentacijom jedinica multimedijalnog podatka.

Teorija *soft računarstva* je implementirana u metodologiji kooperativnih aktivnosti razvoja novih računarskih paradigmi kao što su fazi logika, neuronske mreže, genetski algoritmi, teorija haosa i evolutivno računarstvo. Drugim rečima, soft računarstvo otvara nove pravce istraživanja načina rešavanja problema, koji su teži u odnosu na tradicionalni pristup u računarstvu (hard computing).

Sistem za pretraživanje multimedijalnih podataka funkcioniše kroz tri koraka izvođenja : 1. ekstrakcija svojstava sirovih multimedijalnih podataka kroz predstavljanje svojstava u apstraktnom prostoru (prostor svojstava) vrši se pre sprovođenja bilo koje aktivnosti pretraživanja; 2. predstavljanje znanja odgovarajućim metodama, za podršku aktivnostima očekivanog otkrivanja znanja u multimedijalnoj bazi podataka; 3. otkrivanje znanja u multimedijalnoj bazi podataka kroz stvarno pretraživanje ili teorija i/ili tehnika učenja, što čini srž sistema za pretraživanje multimedijalnih podataka.

3. EKSTRAKCIJA SVOJSTAVA SIROVIH MULTIMEDIJALNIH PODATAKA

Analiza prethodna tri koraka u pretraživanju multimedijalnih podataka podrazumeva prethodno razjašnjavanje pitanja njihovog predstavljanja. Pri tome treba istaći razliku između struktuiranih i nestruktuiranih podataka. Struktuirani podaci su svi podaci koji mogu biti predstavljeni i sačuvani u pojedinim strukturama baza podataka. Uključene su uobičajeno korišćene i povezane baze podataka kao i objektno- orijentisane strukture baze podataka. Nestruktuirani podaci su multimedijalni podaci koji ne mogu biti predstavljeni ili pokazani u utvrdjenoj strukturi baze podataka. Oni se mogu predstaviti u prostoru sa odgovarajućim dimenzijama u kojima je podatak. To znači da je neki tip multimedijalnog podatka: 0-*dimenzionalni podatak*- ovo je regularan tip *numeričkog* podatka a karakterističan primer je *tekstualni* podatak. 1-*dimenzionalni podatak* - ima jednu dimenziju prostora zadatu u njima. Tipičan primer je *audio* podatak, 2-*dimenzionalan podatak*- ovaj podatak ima dve dimenzije prostora odredjene u njima. Dva uobičajena primera za ovaj tip podatka su *slikovni* i *grafički* podaci , 3-*dimenzionalni podatak* –ovaj tip podatka ima tri dimenzije prostora odredjene u njima. Standardni primeri ove vrste podataka su *video* i *animirani* podaci.

Za multimedia data mining najbitnije su osobine podatka i predstavljanje znanja. Osobine su apstrakcije podatka u specifičnom modalitetu i definisane u količini koja se može izmeriti u specifičnom Euklidskom prostoru. Euklidov prostor dobija naziv funkcija prostora. Funkcije, takođe zvane atributi, su apstraktni opis originalne multimedia data u funkciji prostora. Tipično, postoji više nego jedna funkcija koja se koristi da bi se opisao podatak , a to su višestruke funkcije oblika funkcije vektora u prostoru funkcija. Proces identifikacije funkcije vektora iz originalnih multimedia data naziva se funkcija ekstrakcija.

Funkcija *ekstrakcije osobina* je preslikavanje iz jedinice multimedijalnog podatka ili nekog objekta u vektorsku funkciju ili funkciju prostora. Kažemo da je osobina jedinstvena ako i samo ako različite jedinice multimedijalnog podatka ili različiti objekti mapiraju različite vrednosti osobina. U ostalim slučajevima je mapiranje jedan-na-jedan. Medjutim, umesto položaja jedinice multimedijalnog podatka, ova jedinstvena definicija podatka je iznešena na nivo objekta umesto na nivo jedinice multimedijalnog podatka, različiti objekti se tumače u terminima različitih semantičkih objekata, suprotno od različitih varijacija istog objekta. (Npr, kruška i jabuka su dva semantički različita objekta, dok su različiti pogledi iste jabuke ustvari različite varijacije istog objekta, ali ne različitih semantičkih objekata).

Poznati originalni formati (kao što su JPEG, TIFF ili čak Raw Matrix predstavljanje) se smatraju kao neodgovarajuće prezentacije u sistemima multimedia data mining i prema tome su veoma retko direktno korišćeni u bilo kojoj multimedia data mining aplikaciji.

4. ANALIZA METODA PREDSTAVLJANJA ZNANJA U APLIKACIJAMA MULTIMEDIA DATA MINING

Najčešće metode predstavljanja znanja, koje imaju široku upotrebu u aplikacijama za sve multimedia data mining probleme su : logički utemeljene prezentacije , prezentacije zasnovane na semantičkim vezama, frejmovi ili okvirno bazirane prezentacije kao i prinudne prezentacije; Posebnu grupu čine prezentacijske metode zasnovane na sumnjama. U cilju što bogatijih multimedijalnih podataka, važno je da se ne koristi samo odgovarajuće predstavljanje funkcije za multimedijalne podatke, ali takođe i da je odgovarajuća podrška znanja dostupna u multimedija bazama podataka, da bi se olakšali zadaci pretraživanja.

4.1. Logičke prezentacije-Za prestavljanje znanja koje je razumljivo ljudima, najčešće se koristi prirodni jezik. U data mining sistemu često je ograničeno područje prirodnog modela jer predstavljanje mora biti razumljivo ne samo čoveku kao korisniku već i računaru kao elektronskom sistemu za informaciono procesiranje. Zato logička prezentacija predstavlja efektivniji put kojim prirodni jezik postaje razumljiv računaru. Uobičajeno korišćena logika je predikativna ili iskazna logika pr*vog reda* (FOL) gde svaka promenljiva može uzeti, za nju definisanu vrednost u skupu promenljivih. Sve funkcije u FOL se zovu predikati, odnosno Boolean predikati, jer im vrednosti mogu vratiti jednu ili dve vrednosti: 0 za lažnu i 1 za istinitu. Postoje **tri operatora** koji su definisani za sve promenljive, kao i za predikate. To su:

1. *unarni operator***,** primenjen ili na promenljivu ili na predikat, rezultirajući negacijom vrednosti operanda (Ako operand ima vrednost 1, operacija ima vrednost 0, i obrnuto)

2. *binarni operator* Λ, primenjen ili na promenljive ili na predikate operacijom množenja vrednosti dva operanda(operator vraća jedan ako i samo ako su obe vrednosti dva operanda 1, a u protivnom vraća 0).

3. *binarni operator* V**,** primenjen na promenljive ili na predikate a uzima dodatak između dve vrednosti dva operanda, (vraća 0 ako i samo ako oba operanda imaju vrednost 0, a u protivnom povrate 1).

Dva kvantifikatora definisana su samo za promenljive, ali ne i za predikate. To su *univerzalni kvantifikator* $\forall$**, (**za sve vrednosti promenljive kojoj ovaj kvantifikator pripada) i *egzistencijalni kvantifikator* $\exists$**.** (označava da tu postoji bar jedna vrednost promenljive kojoj ovaj kvantifikator pripada).

Prednost korišćenja FOL za prezentaciju znanja u multimedia data mining sistemu je laka i moćna dedukcija, a proces rasuđivanja, uz korišćenje FOL izraza je takođe veoma uspešan prema simboličkim izračunavanjima. Osnovni nedostatak je, što je teško sačuvati postojanost FOL izraza u bazama znanja za multimedia data mining sistem zbog dinamičnog i stalnog ažuriranja.

4.2. Semantičke mreže su veoma moćan alat za predstavljanje znanja u istraživanju veštačke inteligencije i aplikacijama. U multimedia data mining, semantičke mreže se koriste za predstavljanje koncepata, u suštini, prostornih koncepata i njihovovih mreža. Primer je KMeD sistem, razvijen od strane dr Hsu [4], koji koristi hijerarhijske semantičke mreže za predstavljanje znanja o medicinskim slikama u bazama podataka koje je potrebno za olakšavanje rasuđivanja, bogatstva i pretraživanja medicinskih slika u bazi podataka.

Poznata je aplikacija WorldNet (baza podataka reči engleskog jezika), kao dobar primer korišćenja semantičkih mreža za predstavljanje reči i njihovih veza. Tako se u grafičkom prikazu semantičke mreže digraf, koriste: Meronymy – A je deo B; Holonymy – B je deo A; Hiponimi – A je zavisno od B; Hipenimi – A je nadređeno B; Sinonimi – A je isto ili slično kao B; Antonimi – A je suprotno do B; Predloženi su da koriste grafičke prikaze i literature, da predstave koncepte i njihove veze, u kojima su tačke presecanja u grafičkim prikazima koncepti, a ivice u grafičkim prikazima su veze. Istorijski, oni su bili korišćeni da predstave engleske reči, kao i veze u prirodnom jeziku, razumevajući istraživanja u veštačkoj inteligenciji.

4.3. Frejmovi ili okviri-Frejm ili okvir je metoda predstavljanja znanja koja se koristi u multimedijalnim bazama podataka za opisivanje posebnog tipa objekta ili posebnog koncepta. Frejm može imati isto ime kao i druge osobine koje imaju vrednost, a zovu se **slotovi.** Koncepti povezani sa svim drugim konceptima mogu biti opisani za jedan frejm kao slot vrednost drugog frejma. Frejm se može zamisliti kao mreža čvorova i veza u kojoj su veći nivoi fiksirani i predstavljaju stvari koje su uvek istinite za pretpostavljenu situaciju. Niži nivoi imaju mnogo slotova, koji moraju biti ispunjeni posebnim primerom podataka. Svaki slot može zahtevati uslove za terminalni zadatak čiji se argumenti moraju poklopiti. Složeniji uslovi mogu odrediti odnose između stvari dodeljenih terminalima a terminalni zadaci mogu biti osoba, predmet dovoljne vrednosti ili pokazivač na subfrejmovima određenog tipa. .

5. OTKRIVANJE ZNANJA U MULTIMEDIJALNOJ BAZI PODATAKA

Multimedia data mining sistem je tipični inteligentni sistem koji sadrži komponentu za podršku znanja koja omogućava ' pretraživanje zadataka. Standardni tipovi znanja u komponentama podrške znanja uključuju *oblast znanja, znanje zdravog razuma***,** kao i *meta znanje*. Zbog toga je kao značajna postala tema istraživanja, kako prikladnije i efektivnije predstaviti ove tipove znanja u multimedia data mining sistemu. S druge strane, u okviru generalne data mining aktivnosti, standardni zadatak multimedia data mining-a je da automatski otkrije znanje u specifičnom kontekstu**,** pri čemu, takođe postoji i problem predstavljanja znanja posle otkrivanja znanja u aktivnosti pretraživanja.

6. UMESTO ZAKLJUČKA

Cilj multimedia data je da otkrije znanja predstavljena na prikladan način. DATA MINING je relativno nova interdisciplinarna oblast, ponikla, pre svega iz statistike, mašinskog učenja i teorije informacija**.** Premda je ovaj termin prvi put upotrebljen tek pre 15-tak godina, svedoci smo eksplozivnog širenja područja primene datih metoda. Sve šira primena Data Mining pristupa je podstaknuta izuzetnim rezultatima koji su ostvareni u mnogim veoma značajnim naučnim i inženjerskim istraživanjima. Originalni formati podataka uobičajeno zauzimaju više mesta nego što je potrebno. To uzrokuje dva problema- više utrošenog vremena i više angažovanog skladišnog prostora. Drugo i mnogo bitnije je da su ovi originalni formati dizajnirani za najbolje arhiviranje podataka (npr. za minimalno gubljenje integriteta podataka, dok u isto vreme služi za održavanje prostora), ali ne za ispunjavanje multimedia data mining svrhe.

Prema tome, ono što navedeni originalni formati predstavljaju je samo podatak. S druge strane, za multimedia data mining svrha je da predstavi multimedia data kao korisnu informaciju koja bi mogla olakšati različite utrošene i mining operacije. Data Mining u svetu postaje standard, načinivši metamorfozu od endemičnog postupka, dostupnog samo velikim korporacijama, do alata koji se koristi kao potpora u mnogim iole ambicioznijim naučnim projektima, koji se oslanjaju na velike količine podataka zapisane u digitalnom obliku.

LITERATURA

[1] Z. Zhang, R, Zhang,(2009) *Multimedia Data mining*, *A Systematic Introduction to Concepts and Theory*, Taylor and Francis Group, LLC

[2] C.Faloutsos, (1996) *Searching Multimedia Databases by Content,* Kluver Academic Publishes,

[3] G. Fishman, (1996) Monte Carlo Concepts, Algorithms and Applications, Spinger Verlag,

[4] C.C. Hsu, W.W, Chu, and R.K. Raira, (1996) A knowledge-based approach for retrieving images by content, *IEEE Transactions on Knowledge and Data Engineering*, 8(4);522-532, August

[5]R.O.Duda, and P.E. Hart, (2001) *Pattern Classification and Scene Analysis,* John Willey and Sons