Tehnike Text Mining-a i njihova realizacija primenom objektno-orijentisane analize

Muzafer Saračević, Sead Mašović, Hamza Kamberović

***Sadržaj* — U ovom radu su predstavljene ključne tehnike u procesu inteligentne obrade teksta (*Text Mining*) i navedeno je objektno-orijentisano modelovanje za proces klasterovanja. Odrađen je model za *K-means cluster* algoritam i njegova praktična primena. Cilj je da se prikaže nova tehnika vizuelnog planiranja i postupak dobijanja rešenja na osnovu objektno-orijentisane analize i modelovanja.**

***Ključne reči* — Java programiranje, *K-means* algoritam, Klasterovanje, *Text Mining*, UML modelovanje.**

I. Uvod

Cilj ovog rada je da se predstavi sredstvo modeliranja koje je upotrebljivo i za čoveka i za mašinu i da se uspostavi eksplicitna veza između koncepata i izvršnog koda. Analiza i rešavanje na ovakav način ima mnoge prednosti.

Modeluje se algoritam koji se primenjuje u inteligentnoj obradi teksta, odnosno u procesu izvlačenja termina i informacija i pronalaženja veza unutar teksta. Konkretno dati algoritam formira grupe na osnovu učestalosti termina u nekom dokumentu. Praktična primena ove metode modelovanja je pre svega u tome da se na osnovu vizuelnog modela može bolje razumeti problem i može lakše razraditi programski kod za aplikaciju čiji je zadatak zajedničko povezivanje izvučenih informacija da bi se stvorile nove činjenice ili hipoteze koje će kasnije biti dublje ispitivane putem konvencionalnijih sredstava za istraživanje. Ovakvo rešenje nalazi široku primenu u elektronskom poslovanju, jer se suočava sa ogromnim količinama podataka i informacija koje povezuje i grupiše, a kao što je poznato, e-poslovanje se zasniva na ogromnim bazama podataka.

Prilikom rešavanja ovakvih vrsta problema imamo mogućnost vizualizacije u više dimenzija i nivoa detalja i model je prikladan za nadogradnju i dalje modelovanje.

Inteligentna obrada podataka (*Data mining*) se na osnovu vrste podataka, koji se koriste kao ulaz, mogu okvirno podeliti na: *Text mining*, *Web mining, Data mining* multimedije i dr.

Muzafer Saračević, Prirodno-Matematički fakultet, Univerzitet u Nišu, Višegradska 33, 18000 Niš, Srbija (telefon: 381-60-4979797 , e-mail: muzafers@gmail.com)
Sead Mašović, Fakultet Organizacionih nauka, Univerzitet u Beogradu, Jove Ilića 154, 11000 Beograd, Srbija (telefon: 381-60-6660646 , e-mail: sekinp@gmail.com)
Hamza Kamberović, Departman za prirodno-tehničke nauke, Univerzitet u Novom Pazaru, Dimitrija Tucovića bb, 36300 Novi Pazar, Srbija (telefon: 381-63-8550255 , e-mail: hamzanp@hotmail.com)

Text Mining je kompjutersko otkriće nove, prethodno nepoznate informacije putem automatskog izvlačenja termina i informacija iz različitih pisanih izvora. Ključni element jeste zajedničko povezivanje izvučenih informacija da bi se stvorile nove činjenice ili hipoteze.

II. Pojam i tehnike text mining-a

Text Mining je definisan kao analiza teksta na prirodnom jeziku u cilju izvlačenja termina, entiteta i odnosa između tih termina i entiteta[1]. Kako tražnja za uspešnijim tehnikama poslovne inteligencije raste, analitičari smatraju da moraju da prošire opseg svojih podataka kako bi uključili i nestrukturirani tekst. Da bi iskoristili ove informacione resurse, neophodne su tehnike kao što je inteligentna obrada teksta. *Text Mining* je deo šireg polja otkrivanja znanja koje, počinjući od grupe podataka relevantnih za rešenje specifičnog problema, traži zanimljive i prethodno nepoznate šablone. U *Text Mining*-u, cilj je da se dođe do prethodno nepoznatih informacija za koje niko ne zna tako da one nisu mogle ranije biti zapisane. Text Mining kao varijacija polja Data Mining nastoji da pronađe interesantne modele iz većih baza podataka[2].

Ključne tehnike Text Mining-a su: izvlačenje termina i karakteristika, izvlačenje informacija, analiza veza i klasterovanje.

A. Izvlačenje termina i karakteristika

Izvlačenje termina je osnovni oblik Text Mining-a. Kao i sve druge tehnike *Text Mining*-a, ova tehnika prenosi informaciju iz nestrukturiranog podatka u strukturirani oblik. Najjednostavnija struktura podataka u *Text Mining*-u je vektor karakteristika, procenjena lista reči koje se pojavljuju u tekstu i pruža odgovarajući opis teksta. Da bi identifikovali ključne termine, sistemi za *Text Mining* izvršavaju nekoliko operacija.

Prvo, često korišćene reči, poznate kao „stop reči“ (npr. i, ili, ostali) se uklanjaju. Drugo, reči su skraćene na njihove korene. Na primer, reči „telefonirati“ i „telefonirano“ su skraćene na „telefon“. Time je omogućena analiza učestalosti korena reči koji daju smisao bez sintaksnih variranja[3].

Poslednji korak je izračunavanje vrednosti svakog preostalog pojma u dokumentu. Postoje mnoge metode za izračunavanje ovih vrednosti, ali najčešće korišćeni algoritmi upotrebljavaju merenje učestalosti termina u dokumentu (učestalost termina - UT faktor) i učestalost reči u celom skupu dokumenata (inverzna učestalost dokumenata - IUD faktor). Veliki UT faktor povećava

vrednost termina, dok veliki IUD faktor smanjuje tu vrednost.

Opšta pretpostavka celog ovog metoda je da termini koji se često pojavljuju u dokumentima (visok UT faktor) razlikuju jedne dokumente od drugih osim ako se ti termini često pojavljuju u svim tekstovima u skupu dokumenata (visok IUD faktor). Vektori karakteristika se primenjuju u nekoliko slučajeva u sistemima za *Text Mining*. Koriste se za merenje sličnosti između dokumenata. Ako posmatramo taj vektor kao liniju u višedimenzionalnom prostoru, ugao između dva vektora pokazuje sličnost između dokumenata. Pošto vektori karakteristika sadrže najvažnije termine, oni mogu da upravljaju izborom najvažnijih rečenica u dokumentu za potrebe kreiranja sažetka. Konačno, svi vektori pružaju osnovu za klasifikovanje i grupisanje dokumenata.

Izvlačenje podataka se može primeniti na niz tekstualnih izvora i, za razliku od ostalih tehnika *Text Mining*-a, ne smanjuje se kvalitet funkcionisanja kod negramatičkih tekstova. Izvlačenje pojmova je dovoljno u mnogim situacijama, ali kategorisanje i drugi viši nivoi operacija u *Text Mining*-u se poboljšavaju upotrebom izvlačenja karakteristika.

Izvlačenje karakteristika je slično izvlačenju termina samo što umesto korišćenja leksičkih sredstava kao što je razmak ili interpunkcija za identifikovanje termina, izvlačenje karakteristika koristi sintetička svojstva da identifikuje entitete, kao što su imena preduzeća, vladinih agencija, datumi, novčani iznosi i lokacije. Tehnika izvlačenja karakteristika pruža više detalja u vezi sa semantičkim atributima nego izvlačenje termina, ali joj nedostaje važna informacija o odnosima između termina. Zbog toga se na sledećem nivou analize moraju koristiti tehnike izvlačenja informacija.

B. Izvlačenje informacija

Sledeći nivo složenosti *Text Mining*-a je izvlačenje informacija. Za razliku od izvlačenja termina koji se koncentriše na pojmove, izvlačenje informacija se koncentriše na skup činjenica koje čine događaj ili stanje. Primenom sofisticiranijih metoda dodavanja mogu se identifikovati glagolske fraze, koje određuju uloge entiteta u rečenici. Sada, pored sintaksnih atributa (imenovani entitet, datum, novčani iznos), analizirani tekst je obeležen semantičkim atributima kao što su npr. kompanija kupac i iznos kupovine[4].

C. Analiza veza

Analiza veza je skup tehnika za sticanje uvida u odnose između višestrukih entiteta, sa složenim vezama ili koracima. Telekomunikacije su tipičan primer sistema na koje se primenjuje analiza veza. Pozivi telefonom i mrežni paketi počinju od određene tačke i kreću se preko složenih veza na svom putu ka krajnjem odredištu. Analiza veza počinje sa velikom grupom podataka o učestalosti zajedničkog pojavljivanja činjenica ili termina i entiteta.

Na taj način gradi se osnovna struktura veza. U slučaju zajedničkog pojavljivanja termina i entiteta, merimo koliko često se dva termina pojavljuju zajedno, što nam omogućava da identifikujemo potencijalno zanimljive odnose[5].

Osnovni koraci u analizi veza su:

1. određivanje izvora sadržaja,
2. predobrada sadržaja i dodavanje osnovnih sintaksnih i semantičkih dodataka,
3. u slučaju izvlačenja informacija, izvlačenje činjenica, uključuje agente i akcije,
4. u slučaju zajedničkog pojavljivanja termina ili osobina (karakteristika), izračunava se koliko se često termini pojavljuju zajedno.

Primena pogodna za analizu veza se može lako preneti na strukturu čvorova i veza usmerenih grafikona. Kvalitet analize veza dosta zavisi od kvaliteta analize teksta i sposobnosti za pravilno identifikovanje odnosa.

D. Klasterovanje

Klasterovanje je često korišćena aktivnost u okviru istraživanja podataka i koristi za grupisanje podataka. Međutim, grupe nisu unapred definisane, već se grupisanje vrši na osnovu pronađenih sličnosti među podacima. Na osnovu tri prethodno opisane tehnike *Text Mining*-a možemo grupisati podatke u okviru ovog postupka. Tako da možemo formirati grupe koje se nazivaju klasteri.

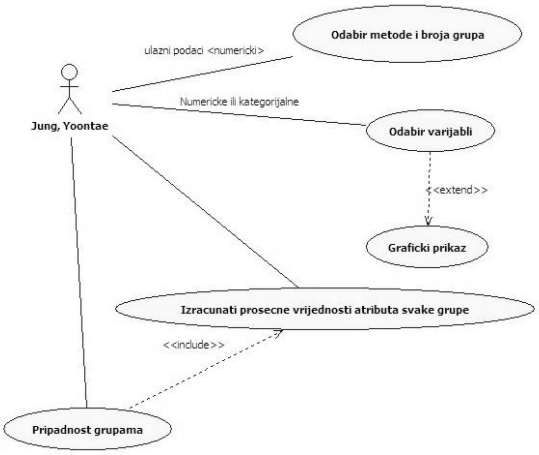
Algoritam K-srednjih vrednosti *(K-means cluster analysis*) je algoritam koji ima kao ulaznu vrednost prethodno definisan broj grupa (k). Ovaj algoritam najbolje radi kada su ulazni podaci uglavnom numerički, kada ga čine kvantitativne varijable[6].

Prema vrednostima koje pojedine varijable poprimaju, svaki zapis se smešta u multidimenzionalnom prostoru zapisa, a svaka varijabla predstavlja određenu dimenziju. Unutar tog prostora jedinke često stvaraju prirodne grupacije (segmente, odnosno klastere). Segmente karakterizuje mala udaljenost među pripadnicima jednog segmenta i veća udaljenost među pripadnicima različitih segmenata. Kada je reč o udaljenosti, onda se najčešće koristi Euklidska udaljenost.

Algoritam K-srednjih vrednosti je iterativna procedura u kojoj centralnu ulogu igra pojam centroida. Centroid je reprezentuje srednju ili prosečnu lokaciju određene grupe primera. Koordinate ove tačke izračunavaju se kao prosečne vrednosti koordinata svih primera koji pripadaju toj grupi. Obično ova iterativna procedura redefinisanja centroida i raspoređivanja primera u odgovarajuće grupe zahteva samo nekoliko iteracija do zadovoljavajuće konvergencije[7].

Za ovaj postupak je odrađeno modelovanje kako bi se što lakše došlo do rešenja, odnosno do konkretnog programskog koda.

Navedeni dijagram (Sl.1.) je dijagram slučajeva korišćenja (*use case*) koji služi za predstavljanje funkcionalnih zahteva koje sistem treba da ispuni. Sastavljen je od jednog aktera (korisnika) i slučajeva korišćenja sistema koji proizilaze iz algoritma. Jedan slučaj korišćenja predstavlja jedan slučaj iskorišćavanja funkcionalnosti sistema.

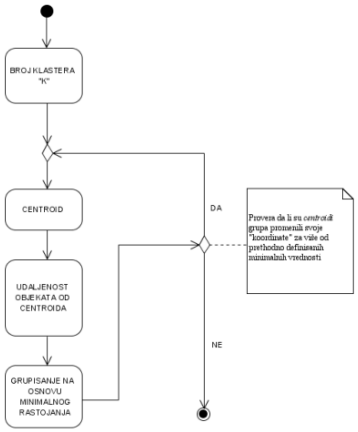


Sl. 1. Dijagram slučajeva korišćenja

Faze *K-means* algoritma su[8]:

1. odabere se *k* početnih tačaka - *centroidi ,*
2. dodeljuje se svaki podatak *centroidu* kojem je najbliži (formiraju se tzv. ekskluzivne grupe),
3. izračunavaju se novi *centroidi* novoformiranih grupa,
4. proverava se da li su *centroidi* grupa promenili svoje "koordinate" za više od prethodno definisanih minimalnih vrednosti,
5. ako jesu, kreće se ponovo od tačke 2. Ako ne, određivanje grupa je završeno.

Ovaj postupak je prikazan na dijagramu aktivnosti (Sl.2).



Sl. 2. Dijagram aktivnosti

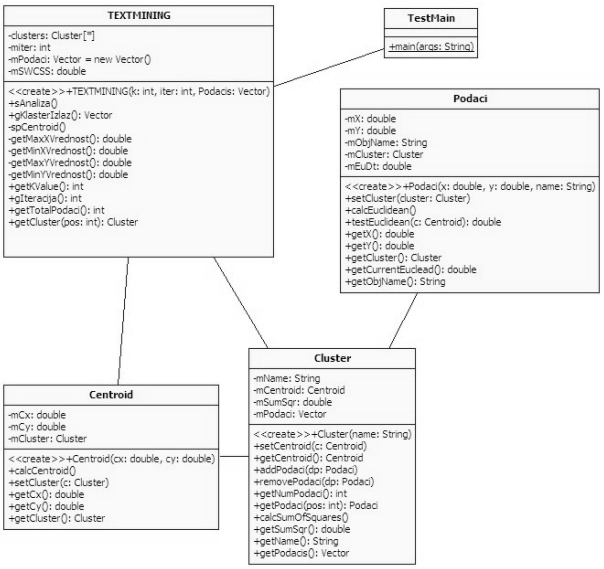
Dijagram aktivnosti *(activity diagram)* predstavlja akcije koje se izvode (Sl.2). Nakon izvršenja radnje u jednoj akciji dolazi do automatskog prelaska u sledeću akciju.

III. OBJEKTNO ORIJENTISANA ANALIZA I DIZAJN

Objektno orijentisana analiza i dizajn omogućava svim učesnicima u razvoju aplikacije da na jednostavan i sveobuhvatan način steknu uvid u analizu i implementaciju konkretnog problema. Shodno tome, javila se potreba za univerzalnim jezikom namenjenim za objektno orijentisano modelovanje. UML (Unified Modeling Language) je jezik za modelovanje koji služi za specifikaciju, vizuelizaciju, konstrukciju i dokumentaciju razvoja sistema. Koristi se u različitim fazama razvoja, od specifikacije zahteva do testiranja završenih, gotovih sistema. Za izradu UML dijagrama u ovom radu korišćeni su ***STAR UML*** i ***Visual Paradigm for UML.***

UML standard koji se primenjuje kod objektno-orijentisanog pristupa, predviđa odgovarajuće poglede na sistem, s tim što se u svakom pogledu sistem može opisati sa statičkog (strukturnog) i dinamičkog aspekta. Sasvim je sigurno, da ćemo ovakvim postupkom pojednostaviti postupak dobijanja rešenja **.**U postupku modelovanja treba najpre odraditi slučajeve korišćenja, od koga dalje potiču svi sledeći dijagrami. Zatim sledi projektni pogled, kod koga se statički aspekt sistema prikazuje preko dijagrama klasa i dijagrama objekata, a dinamički aspekt preko dijagrama interakcije, sekvenci, dijagrami promene stanja i dijagrami aktivnosti.

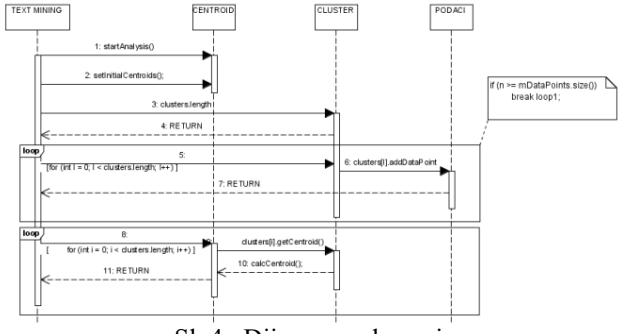
U objektno-orijentisanom pristupu razvoja sistema osnovni elementi pomoću kojih se opisuje sistem su klase i veze između njih, te objekti posmatranih klasa, njihove veze i poruke koje objekti međusobno razmenjuju u cilju izvršavanja određenih aktivnosti. Dijagram klasa (sl.3.) prikazuje skup klasa, interfejsa i saradnji, kao i njihovih veza. Klase su opisane imenom, atributima i operacijama. Veze obezbeđuju komunikaciju između klasa.



Sl. 3. Dijagram klasa

Početni korak je da dobijemo klase i metode unutar klasa. Za dalju analizu potrebno je detaljno razraditi dijagrame ponašanja.

U dijagramu sekvenci su predstavljene međusobne interakcije objekata, koje predstavljaju niz razmene poruka između klasa, pri čemu je redosled i vremenski tok slanja i primanja poruka jasno naznačen (sl.4.).



Sl. 4. Dijagram sekvenci

Dijagram sekvenci je pogodan za razradu funkcionalnosti metoda. Prikazani dijagram prikazuje osnovne korake u metodi *sAnaliza()* na osnovu kojeg je kasnije moguće lakše dobiti *izvorni kod* u nekom objektno-orijentisanom programskom jeziku. U dijagramu klasa možemo jasno videti koje metode pripadaju kojoj klasi, tako da ako želimo da odradimo kompetno modeliranje, do detalja, potrebno je odraditi sekvencijalne dijagrame za svaku navedenu metodu.

Krajnji rezultat ovakve tehnike je predstaviti metod klasterovanja koji se primenjuje u inteligentnoj obradi teksta, odnosno predstavljen je proces kvalitetnog izvlačenja informacija iz teksta na osnovu formiranja grupa i izvlačenja termina kod kojih je moguće uočiti odgovarajuće veze unutar dokumenta. Način predstavljanja je odrađen posredstvom UML modelovanja koji nam služi da odradimo vizuelni plan na osnovu kojeg možemo programirati i dalje razraditi problem i dobiti rešenje za dati algoritam. UML je metod modelovanja koji definiše nekoliko pravila koja omogućavaju da svi koji su uključeni u projekat govore istim jezikom i koriste iste činjenice.

IV. IMPLEMENTACIJA

Dat je konkretan primer gde je jedan od zadataka ovog algoritma može biti sledeći: pronaći mogu li se marke automobila grupisati prema odabranim karakteristikama.

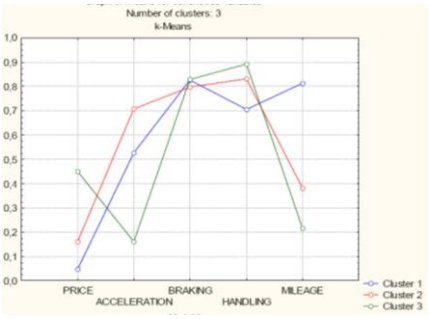
	1 PRICE	2 ACCELERATION	3 BRAKING	4 HANDLING	5 MILEAGE
Acura	-0,521	0,477	-0,007	0,382	2,079
Audi	0,866	0,208	0,319	-0,091	-0,677
BMW	0,496	-0,802	0,192	-0,091	-0,154
Buick	-0,614	1,689	0,933	-0,210	-0,154
Corvette	1,235	-1,811	-0,494	0,973	-0,677
Chrysler	-0,614	0,073	0,427	-0,210	-0,154
Dodge	-0,706	-0,196	0,481	0,145	-0,154
Eagle	-0,614	1,218	-4,199	-0,210	-0,677
Ford	-0,706	-1,542	0,987	0,145	-1,724
Honda	-0,429	0,410	-0,007	0,027	0,369
Isuzu	-0,798	0,410	-0,061	-4,230	1,067
Mazda	0,126	0,679	-0,133	0,500	-1,724
Mercedes	1,051	0,006	0,120	-0,091	-0,154
Mitsub.	-0,614	-1,003	0,084	0,382	0,718
Nissan	-0,429	0,073	-0,007	0,263	0,997
Olds	-0,614	-0,734	0,409	0,382	2,114
Pontiac	-0,614	0,679	0,536	0,145	0,195
Porsche	3,454	-2,215	-0,296	0,618	-1,026

Sl. 4. Deo baze "automobili"

Centroids for k-means clustering (Cars.sta)
Number of clusters: 3
Total number of training cases: 22

Cluster	PRICE	ACCELERATION	BRAKING	HANDLING	MILEAGE	Number of cases	Percentage(%)
1	-0,595031	-0,15541	0,003701	-0,564366	1,395032	5	22,72727
2	-0,115722	0,54973	-0,062172	0,090547	-0,261140	13	59,09091
3	1,119885	-1,59237	0,097334	0,411181	-0,895085	4	18,18182

Sl.5. Grupisanje navedenih podataka u tri klastera



Sl.6. Grafički prikaz (5 karakteristika i 3 klastera)

V. ZAKLJUČAK

U radu je naveden model za dati algoritam koji se primenjuje u formiranju grupa na osnovu pronađenih karakteristika ili termina. Algoritam se primenjuje u inteligentnoj obradi teksta, prilikom pronalaženja korisnih informacija po nekom zadatom kriterijumu. Znači postupak se može definisati kao pronalaženja veza u tekstu na osnovu nekih pravila unutar samog dokumenta ili baze. Ovim radom dajemo jedan od načina dolaženja do rešenja. Znači, naši modeli će poslužiti daljoj nadogradnji i analizi rešenja.

UML je alat koji nudi mnoge mogućnosti, između ostalih i mogućnost generisanja koda u nekom od objektno-orijentisanih jezika. Tako da nam modeli služe kao šeme (planovi) koji nam omogućavaju dalju implementaciju. Pored toga, dajemo jedan nespecifičan način dolaženja do rešenja, tako što koristimo metod postepenog rešavanja kroz objektnu analizu, koja ima opciju da ponudi mogućnost uključenja lica koja nisu bas dobra u oblasti programiranja ali zato dobro znaju da razrade kompletno problem i da ga vizuelno prikažu.

Ovakav način rešavanja je dobar zato što se UML-om detaljno opisuje problem, modeli pomažu da se sistem vizualizuje onakav kakav je ili kakav bi trebalo da bude i omogućuju da se odredi struktura ili ponašanje sistema. Modeli dokumentuju odluke koje smo donosili i daju uzorke koji nas vode prilikom konstrukcije sistema, odnosno konkretne aplikacije.

LITERATURA

[1] R.Feldman,J. Sanger, "The Text Mining Handbook", *Cambridge University Press,* ISBN 9780521836579.

[2] S.S. Weng, Y.J. Lin and F. Jen, ''A study on searching for similar documents based on multiple concepts and distribution of concepts", *Expert Systems with Applications* 25 (2003) 355–368.

[3] Mani and M.T. Maybury, "Advances in Automatic Text Summarization", *MIT Press*, 1999.

[4] V.Ramos, J.J. Merelo, "Self-Organized Stigmergic Document Maps: Environment as a Mechanism for Context Learning", *AEB'2002*, pp. 284-293, Centro Univ. de MErida, Spain, 6-8 Feb. 2002.

[5] Dan W. Patterson: "Introduction to Artificial Intelligence and Expert Systems", *Prentice Hall*, 1990.

[6] K.Alsabti, S.Ranka, V.Singh, "An Efficient k-means Clustering Algorithm", *Proc. First Workshop High Performance Data Mining*, Mar. 1998.

[7] D.Pelleg and A.Moore, " Accelerating Exact k-means Algorithms with Geometric Reasoning", *Proc. ACM SIGKDD Int'l Conf.Knowledge Discovery and Data Mining*, pp. 277-281, Aug. 1999.

[8] S.Z.Selim, M.A.Ismail, " K-means-type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 6, pp. 81-87, 1984.

ABSTRACT

This paper presents the key techniques in the intelligent processing of text (*Text Mining*) and the stated object-oriented modeling to the process of clustering. Done is a model for K-means cluster algorithm. The aim is to present new visual techniques and procedures for obtaining planning solutions based on the previous analysis and modeling.

TEXT MINING TECHNIQUES AND THEIR IMPLEMENTATION USING OBJECT-ORIENTED ANALYSIS

Muzafer Saračević , Sead Mašović , Hamza Kamberović