Rudarenje teksta sa društvenih mreža API pristupom

Nenad Mirkov
Informatika
OŠ „Ivan Goran Kovačić“
Subotica, Srbija
nenad.mirkov@gmail.com

Miodrag Peranović
Poslovna Informatika
Univerzitet u Istočnom Sarajevu, Ekonomski fakultet Brčko
Brčko, BiH
miodrag.peranovic.efb@gmail.com

Sadržaj— **Količina informacija koja protiče društvenim mrežama svakodnevno se povećava, i predstavlja bogat izvor podataka, koji ako se pravilno semantički obrade, može biti veoma koristan u najrazličitijim oblastima. Cilj ovog projekta je kreiranje sistema za automatizaciju procesa rudarenja teksta sa Twiter mreže, te semantičke obrade i klasterizacije, skladištenja i mogućnosti kasnije ponovne pretrage i obrade. U radu su prikazani rezultati testiranja projektovanog sistema, koji pokazuju svrsishodnost istog.**

Ključne reči – rudarenje teksta; Twiter API; društvene mreže;

I. UVOD

Razvoj Interneta doneo je velike pomake u oblasti komunikacija i razmene informacija. Sve više se pojavljuju novi zahtevi za prijemom i obradom informacija u realnom vremenu. Korišćenje Interneta kao komunikacionog kanala i fenomena u psihologiji ljudi koji se odnosi na potrebu da se određeno znanje istog momenta ili što je pre moguće podeli sa drugima, omogućilo je da društvene mreže, i servisi za mikroblogovanje postanu značajan potencijalni izvor znanja koje se kasnijom semantičkom obradom i klaster analizom mogu uskladištiti, analizirati i koristiti u procesima donošenja odluka. Ovaj fenomen je posebno primetan tokom prirodnih katastrofa ili drugih kriza koje pogađaju ljudsko društvo. Tada su i informacije koje se dobijaju sa lica mesta od običnih ljudi, korisnika servisa, smatraju izuzetno vrednim [1]. Smatra se da su podaci proizvedeni mikroblogovanjem nedvosmisleni, brzi i pristupačni [2], kao i da pomažu građanima da budu svesniji situacije tokom kriza, i da donesu kvalitetnije odluke koje će im pomoći da prevaziđu krizu [3].

Mikroblogovanje kao jedan od oblika društvenih mreža se već uveliko integrisao u svakodnevni život i postao deo šeme svakodnevne komunikacije i razmene informacija. S Obzirom na ogroman broj mobilnih uređaja koji se koristi i u svrhe mikroblogovanja, mnogi smatraju da je ovaj vid komunikacije ozbiljan činilac u sprečavanju ljudskih žrtava u momentima kriza i prirodnih katastrofa. U poslednjoj deceniji sprovedena su brojna istraživanja o upotrebi društvenih mreža i mikroblogovanja tokom kriznih situacija, kao što su upotreba Fejsbuka tokom pucnjave u Univerzitetskom kampusu u Virdžiniji 2007. godine i Univerzitetu Severnog Ilinoisa 2008. godine [1] i velikih požara koji su zahvatili Južnu Kaliforniju 2007. godine [4]. Pored informacija od značaja tokom kriznih situacija, postoji velika korist od informacija dostupnih u realnom vremenu kada je saobraćaj u pitanju. Naime, korisnici veoma često koriste društvene mreže da pošalju informaciju o saobraćajnim nezgodama, blokadama puteva, kašnjenjima u javnom transportu, što može biti od značaja za veliki broj drugih korisnika. Naravno, da bi ove informacije bile u pravo vreme prosleđene potencijalnim korisnicima, potrebno ih je prikupiti, obraditi i proslediti u realnom vremenu.

U ovom radu prikazano je istraživanje upotrebljivosti API-ja za pretragu podataka na Twitter mreži. Objašnjeni su problemi sa kojima se korisnik može susresti i dat je predlog okvira za prevazilaženje istih u vidu sistema koji se sastoji iz nekoliko modula koji su u daljem tekstu opisani i predstavljeni dijagramom. Prikupljani su podaci o stanju na putevima odnosno saobraćaju uopšte i to posebno set podataka vezan za englesko govorno područje i srpsko govorno područje. Nakon toga je analizirana relevantnost dobijenih podataka i mogućnosti njihove klasterizacije i skladištenja radi kasnije ponovne upotrebe. U konkretnom primeru vezanom za saobraćaj izvršena je klasifikacija prema kriterijumima : jak, srednji i slab (umeren) saobraćaj.

Velika konkurencija koja zahvaljujući globalizaciji dodatno izvršila pritisak na kompanije koje svoje poslovanje zasnivaju na potrošnji je uticala na promene u načinima prikupljanja informacija sa tržišta. Kompanije su prepoznale potencijal društvenih mreža kako za reklamiranje, tako i za prikupljanje mišljenja korisnika o konkretnom proizvodu ili usluzi, odnosno i samoj kompaniji. Istraživanja su pokazala da sposobnost kompanije da uče prikupljajući eksterna znanja određuje njenu inovativnost i produktivnost. Postoji veoma tesna veza između prikupljanja i upravljanja znanjem, učenja i inovativnosti. Inovativnost u kompanijama direktno zavisi od ljudske sposobnosti da stvori i podeli određeno znanje sa drugima. Newell i drugi [5] ističu da stvaranje znanja nije individualna aktivnost, već rezultat saradnje većeg broja pojedinaca da sintetizuju znanja deljenjem informacija i razmenom ideja i perspektiva.

Različiti aspekti upotrebe društvenih mreža i koliko je inovativnost zavisna, odnosno bazirana na društvenoj interakciji svojih zaposlenih, je tema koja se sve više istražuje u proteklim godinama [6]. Određeni broj istraživanja je fokusiran posebno na razmenu znanja unutar same organizacije (transfer znanja) [7]. Ova istraživanja potvrđuju neophodnost razmene i transfera znanja, kao i uloge pojedinca u tom procesu. Pokazalo se tačnim da „ono koga poznajemo“ direktno utiče i na „ono šta znamo“, s obzirom da stvaranjem veza preko društvenih omogućuje rešavanje težih problema tj. učenje kako da se određeni problem reši.

Međutim kada se traga za znanjima koja nije su zatvorena i dostupna samo određenom skupu korisnika mreže, kao što je slučaj sa Fejsbuk, Majspejs, Badoo i sličnim oblicima društvenih mreža gde je uključen apsekt privatnosti, i gde korisnici moraju prvo uzajamno izvršiti povezivanje da bi mogli deliti informacije, tada su mikroblogerski oblici društvenih mreža mnogo pogodnije tlo sa istraživanje. Najpoznatiji servis za mikrobloging je Twitter pa će u nastavku biti detaljnije objašnjena njegova arhitektura i mogućnosti za rudarenjem podataka i teksta (data mining, i text mining).

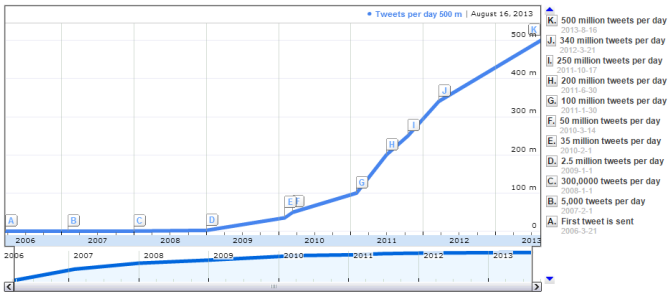
Posebno je važno istaći i svakodnevni porast i sveprisutnost mobilnih uređaja u svakodnevnom životu, koji su olakšali pristup društvenim mrežama, i pojednostavili način deljenja određene informacije jednim klikom. Upravo su informacije koje su ključne za donošenje odluka u realnom vremenu pogodne za mreže poput Twitera. Pew Research Center ističe da preko 40% korisnika mobilnih uređaja koristi društvene mreže, i od toga 28% čini to svakodnevno [8].

U nastavku ovog rada dato je objašnjenje ključnih metodologija i termina koji su neophodni za istraživanje informacija i podataka dostupnih na društvenim mrežama (pre svega Twiteru). U trećem delu je opisano sprovedeno istraživanje kao i dobijeni rezultati. Četvrto poglavlje se bavi zaključnim razmatranjima i idejama za buduća istraživanja.

II. PRIKUPLJANJE PODATAKA PUTEM TWITER API SERVISA

Kada je rudarenje podataka iz društvenih mreža u pitanju, Twiter servis je u prednosti, jer su sve informacije javno dostupne svima, onog momenta kada su objavljene. Poslednjih godina mogućnosti su dodatno proširene podacima o geolokaciji svakog tvita (naziv za pojedinačnu informaciju koju korisnik postavlja na mrežu). Ovo otvara nove pravce u istraživanjima omogućujući proučavanje kako tekstualnih tako i lokacijskih karakteristika online sadržaja. Velika količina podataka koja protiče kroz ovaj servis ima potencijal novih znanja i otvaranja novih pogleda kako akademskoj zajednici, tako i marketinškim kompanijama i drugim organizacijama zainteresovanim za monitoring određenih trendova i statusa. Poslednja istraživanja [9],[10] pokazuju da je broj aktivnih korisnika Twitera dosegao blizu 646 miliona u 2014. godini. Svake sekunde postavlja se oko 6000 novih tvitova što odgovara brojki od oko 500 miliona tvitova dnevno (Slika 1). Pored toga podaci sa Twitera su posebno interesantni jer se tvitovi pojavljuju „brzinom misli“ odnosno dostupni su za preuzimanje momentalno, jer se sve odigrava u „skoro realnom vremenu“ (near real-time).

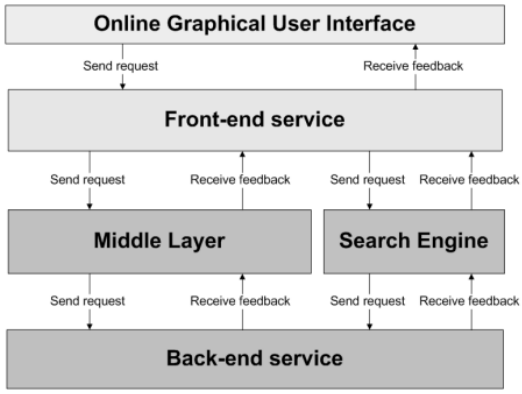
Broj upita koji se prosleđuju Twiterovim serverima svakog dana broji oko 2.1 milijarde, što je pokazatelj da se ide ka upotrebi Twitera kao servisa umesto kao veb aplikacije. Ovo je ozbiljan zahtev pa iza svega mora postojati čvrsta i sigurna serverska arhitektura kao i način za pristup i pretragu podataka.



Slika 1. Broj tvitova po danu. Preuzeto 3.2.2015, sa http://www.Internetlivestats.com/twitter-statistics/

A. *Arhitektura Twiter servisa*

Od svog nastanka 2006. godine Twiter arhitektura je pretrpela nekoliko izmena. Trenutna arhitektura obuhvata srednji sloj koji je uveden upravo zbog velikog broja zahteva u sekundi koje front-end servis nije mogao više opslužiti. (Slika 2). Sistem zadužen za pretragu (search engine) je izdvojen i njegova namena je da odgovori na ogroman broj zahteva koji pristižu preko Search API interfejsa. Od nedavno je ovaj sistem implementiran korišćenjem Apache Lucene sistemom visokih performansi i koji služi za pretragu teksta i koji koristi metodu obrnutih indeksa.



Slika 2. Trenutna arhitektura Twiter servisa [11]

Twiter API (Application Programming Interface) je sistem za pristup Twiter serverima i bazi podataka radi ekstrakcije rezultata upita. Upiti se šalju u obliku url adrese, i u sebi sadrže ključne reči pretrage i druge parametre. Naravno postoje ograničenja sa serverske strane o broju upita, odnosno zahteva koje se mogu uputiti u minuti. Trenutna verzija API-ja je v1.1

Razvijene su mnogobrojne biblioteke za različite programske jezike koje koriste Twiter API za pretragu podataka, odnosno rudarenje teksta (text mining). Za potrebe ovog istraživanja korišćena je biblioteka za C# programski jezik pod nazivom Linq2Twitter. Osnovna ideja je prikupljanje tekstualnih podataka u cilju prikupljanja, analize, geolociranja, i skladištenja podataka od interesa za korisnika, na osnovu kriterijuma pretrage.

B. *Prikupljanje podataka (text mining)*

Rudarenje teksta predstavlja proces koji koristi metodologije u razvoju i koje se baziraju na ekstrakciji informacija od značaja iz nestrukturiranih tekstualnih izvora. Rudarenje teksta je deo rudarenja podataka koji se odnosi na tekstualne izvore. Da bi se iz ogromne količine tekstualnih podataka, na brz i efikasan način izvukli korisni podaci, neophodno je koristiti automatizovane računarske tehnologije.

Dosadašnja istraživanja su pokazala da je moguće uspešno koristiti tehnike rudarenja teksta radi analize velike količine tekstualnih podataka u poslovne svrhe [13] kao i oblasti obrazovanja [14]. Fuller, Biro i Delen su metodama rudarenja podataka i rudarenja teksta uspešno detektovali laži i prevare u tekstu. Veoma značajno je istraživanje kojim se korišćenjem tehnika rudarenja teksta vrši grupisanje sadržaja i dokumenata za elektronsko učenje, na osnovu prepoznavanja sličnosti a koje pripadaju potpuno različitim oblastima [15].

C. *Rudarenje podataka sa Twitera (twittermining)*

Istraživanja vezana za prikupljanje i analizu podataka sa Twitera su do sada sprovedena, kako od strane naučnika, tako i od većih kompanije poput Microsoft-a koji su sa partnerima u projektu istraživali u ulogu različitih tipova uloga društvenih medija u širenju informacija tokom kritičnih i osetljivih pitanja u svetu [16].

Istraživanje koje su sproveli Kwak i drugi [17] imalo je za cilj da potvrdi mogućnosti Twitera kao medija za deljenje informacija. Kreirali su program koji je uspešno krstario Twiterom i prikupio 41,7 miliona korisničkih profila, 1,47 milijardi društvenih veza i 106 miliona tvitova. Za prikupljanje ovih podataka, istraživači su uspešno održavali vezu preko Twiter API-ja tokom jednog meseca.

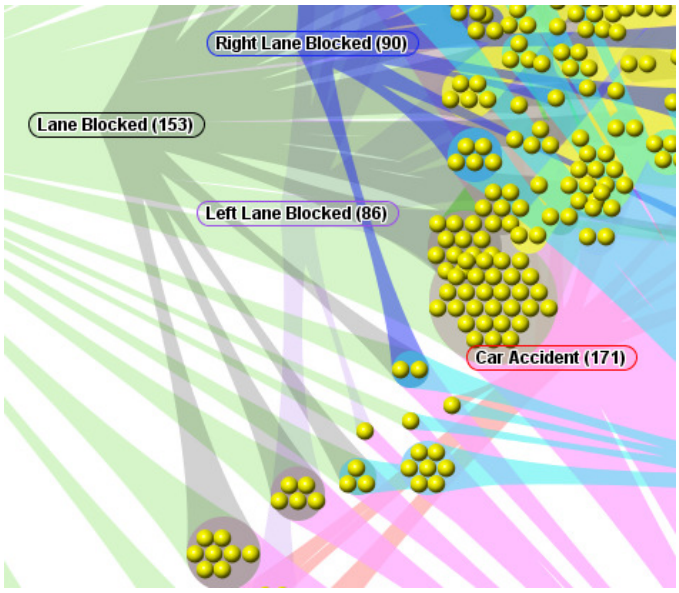
Ekstrakciju podataka o saobraćaju je istraživalo nekoliko različitih autora, pa je tako zanimljiva disertacija [18] u kojoj se opisuje projektovani sistem za monitoring saobraćajnih gužvi, prikupljanjem i obradom podataka sa Twitera u realnom vremenu.

III. ISTRAŽIVANJE I REZULTATI ISTRAŽIVANJA

Projektovani sistem je zasnovan na Search API-ju koji za razliku od Stream API-ja poseduje znatna ograničenja prilikom slanja upita i pretraživanja podataka. Naime, upiti koji pretražuju tvitove mogu da vrate maksimalno do 100 tvitova po jednom upitu. Takođe nije moguće pretraživati stare tvitove, pa smo sistem projektovali da znanja prikuplja progresivno, tokom dvonedeljnog testiranja. Prvo je sistem vršio detekciju učestalosti pojave tvitova, drugim rečima vremenski interval u kojem se pojavljuju novi tvitovi na tražene reči iz rečnika. Tek nakon toga API za prikupljanje tvitova na osnovu dobijenog vremenskog intervala vrši slanje upita i skladišti sirovi materijal koji se kasnije semantički obrađuje.

Prilikom testiranja odlučili smo se za rudarenjem tekstova koji su povezani za saobraćajem i stanjem u saobraćaju zbog pretpostavke da se pojedinci koji koriste mikroblogovanje kao način deljanja informacija najčešće odlučuju da ovakav vid informacija podele sa drugima. Jedan od zahteva istraživanja je bio da se utvrdi frekventnost informacija koje se pojavljuju, a zadovoljavaju zadate kriterijume. Zbog potrebe za brzom obradom podataka u „skoro realnom vremenu“, podaci se po prikupljanju smeštaju u sirovom obliku u xml formatu. Za klasterizaciju smo koristili projekat Carrot2 koji je prevashodno namenjen radu sa tekstualnim dokumentima. U mogućnosti je da primeni više različitih algoritama za klasterizaciju [19].

U izdvojenom skupu podataka o stanju u saobraćaju najbolje rezultate dao je STC (Suffix Tree) algoritam, koji je na ulaznom skupu od 20000 dokumenata dao rezultat od 16 klastera za 5,5 sekundi. Program nudi i vizuelizaciju rezultata od čega je jedan segment prikazan na Slici 3.



Slika 3. Vizuelizacija klaster mape rezultata (Carrot2 Workbench)

Arhitektura sistema prikazana je na Slici 4. Nakon obrade podaci se smeštaju u bazu podataka pri čemu se čuvaju relacije između entiteta korisnika i entiteta teksta. Arhitektura podataka je vršena na osnovu analize rezultata koji se dobiju slanjem online upita na Twiterovom sistemu za testiranje API-ja koji je dostupan na adresi: https://dev.twitter.com/rest/tools/console.

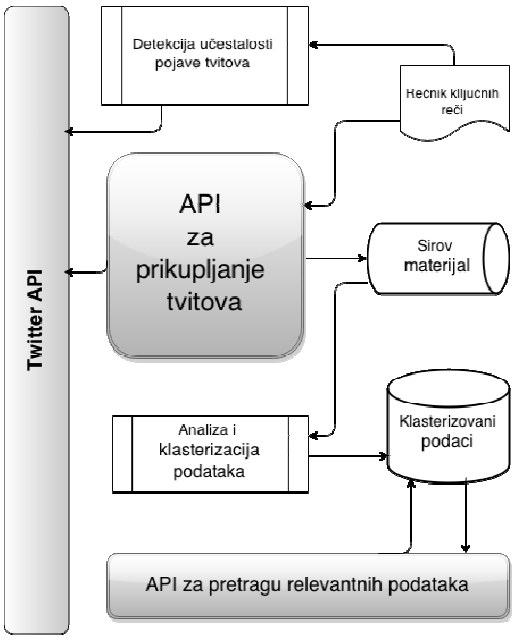
Podaci koji su ekstrahovani iz sirovih rezultata prikupljenih pomoću Twiter API-ja su prikazani u tabeli 1. Ovi podaci predstavljaju osnov za klasu podataka koja se koristi za smeštanje u bazu podataka.

TABELA I. STRUKTURA USKLADIŠTENIH PODATAKA

Naziv	Tip	Opis
CreatedAt	DateTime	Datum i vreme nastanka tvita
Lat	Double	Latituda
Lon	Double	Longituda
PlaceName	String	Opis lokacije
PlaceType	String	Tip lokacije
Retweeted	Boolean	Da li postoji odgovor na tvit
UserName	String	Korisničko ime
Source	String	Kompletan URL koji predstavlja

TweetText	String	Sadržaj tvit poruke
MaxID	String	ID poslednje poruke
FollowersCount	Int	Broj pratilaca korisnika
GeoEnabled	Boolean	Da li je korisnik uključio opciju
Name	String	Naziv korisnika

S obzirom da Twiter zahteva autorizaciju prilikom korišćenja svojih servisa, bilo je neophodno kreirati i registrovati aplikaciju na razvojnoj stranici Twitera. Postoji mogućnost i autorizacije bez aplikacije, ali su tada ograničenja veća, pa je zbog toga odlučeno da se koristi OAuth 1 sistem autorizacije koji poseduje četiri parametra za pristup: API key, API secret, Access Token i Access Token Secret. Ovi parametri su trajni kada se jednom kreiraju pa ih je moguće koristiti dok god postoji registrovana aplikacija.



Slika 4. Arhitektura projektovanog sistema za rudarenje teksta pomoću Twiter API-ja

Prilikom dvonedeljnog testiranja sistema korišćena su dva rečnika ključnih reči. Jedan je sadržao ključne reči na engleskom, a drugi na srpskom jeziku. Testiranje je vršeno odvojeno kako bi se ustanovile razlike u frekvenciji pojavljivanja novih informacija vezanih za saobraćaj i stanje na putevima.

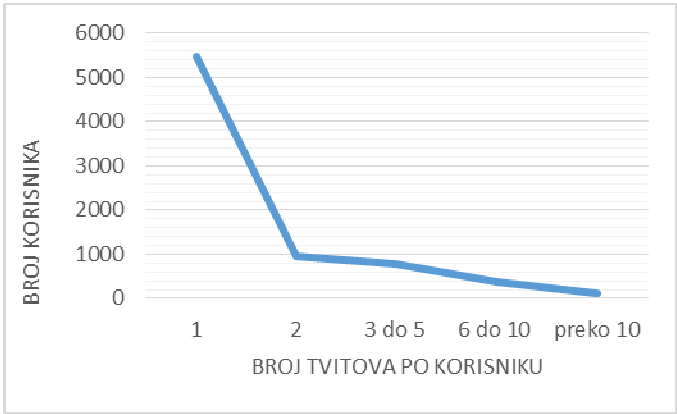
Kao što je i bilo očekivano postoji velika razlika u frekvenciji pojavljivanja tvitova koji zadovoljavaju kriterijume upita na engleskom u odnosu na srpski jezik. U test primeru na sa engleskim ključnim rečima, tvitovi su se pojavljivali u proseku svakih 6,03 sekundi, dok su se tvitovi koji sadrže ključne reči na srpskom jeziku pojavljivali svakih 898,2 sekunde. Primećeno je i potpuno odsustvo novih tvitova u periodu od ponoći do 4h ujutru u testovima sa rečima na srpskom jeziku, pa je ovo vreme izuzeto iz proseka.

U prvobitnim softverskim zahtevima nameravali smo da analiziramo i lokacijske podatke korisnika koji dele informacije. Međutim analizom je zaključeno (Tabela 2) da je i dalje veoma mali broj korisnika koji deli geolokacijske informacije prilikom tvitovanja kod nas, pa je za sada ovakva funkcionalnost projektovanog sistema izostala. Procenat prepoznatih geolokacijskih podataka (latitude i longitude) iznosi 6,19% (za test na srpskom svega 1,05%), iako je procenat korisnika koji imaju uključenu opciju deljenja geolokacije znatno veći i iznosi 37,62 %.

TABELA II. BROJ KORISNIKA SA UKLJUČENIM GEOLOCIRANJEM

Jezik ključnih reči	Broj korisnika sa uključenom geolokacijom	Ukupan broj korisnika
ENGLESKI	6015	15990
SRPSKI	1635	2865

Analiza korisnika koji objavljuju informacije pokazuje da su u prikupljenim podacima uglavnom različiti korisnici, a korisnici koji imaju preko 10 tvitova su uglavnom nalozi medijskih kuća. Na Slici 5 je prikazana dinamika broja korisnika prema broju objavljenih tvitova.



Slika 5. Prikaz broja korisnika prema broju objavljenih tvitova

Za potrebe ovog testiranja podaci su klasifikovani u tri grupe prema detektovanom stanju saobraćaja na osnovu analize teksta: slab saobraćaj, saobraćaj srednje gustine, gust saobraćaj (kolaps). Rezultati nakon semantičke analize klasifikacije su prikazani u Tabeli 3. Ono što je prosto uočljivo je da broj poruka raste što je saobraćaj teži. Najčešće se prijavljuju blokade na putevima, zastoji, i saobraćajne nezgode. Mada su pronađene i poruke koje govore u prilog prohodnosti puta.

TABELA III. PRIKAZ KLASIFIKACIJE PODATAKA

Tip saobraćaja	Broj poruka
Slab saobraćaj	660
Srednja gustina	2130
Saobraćajni kolaps	14940

IV. ZAKLJUČAK I BUDUĆA ISTRAŽIVANJA

Projektovani sistem uspešno je zadovoljio inicijalno testiranje koristeći Twiter API-je za rudarenje teksta. Većina prikazanih rezultata je bila očekivana, osim podataka o geolokacijama korisnika odnosno tvitova. Zbog veoma malog procenta prisutnih informacija o geolokaciji korisnika nismo bili u mogućnosti kreirati vezu korisnika sa geografskom lokacijom. Međutim s obzirom da ova opcija nije postajala na samom početku pojave Twiter servisa, i da sve više mobilnih uređaja poseduje module za geografsko pozicioniranje, realno je očekivati da će ovaj procenat vremenom biti veći.

Pokazalo se da je trenutni Search API moguće koristiti u cilju prikupljanja podataka, i da se predloženi model može koristiti u opštem slučaju za pretragu po različitim kriterijumima, ali kvalitet rezultata direktno zavisi od zadatih ključnih reči. Primer sa ključnim rečima vezanim za stanje u saobraćaju, pokazalo je nedostatak podataka na srpskom jeziku, pa je ideja korišćenja modela za analizu stanja na putevima u Srbiji za sada nepraktična korišćenjem predloženog modela.

Buduća istraživanja bi trebala da obuhvate i sentiment analizu (često se spominje i kao rudarenje mišljenja), koja je veoma često deo sistema za prikupljanje podataka sa društvenih mreža i koristi se u sistemima za odlučivanje i marketinškim istraživanjima.

Planirano je uvođenje algoritama za NLP koji bi trebali omogućiti bolju klasifikaciju i klasterizaciju podataka, kao i usavršavanje API poziva kako bi se projektovani sistem mogao koristiti od strane drugih programa sa korisničkim interfejsom.

LITERATURA

[1] Palen, L., Vieweg, S. , "The Emergence of Online Widescale Interaction in Unexpected Events: Assistance, Alliance & Retreat.", CSCW, ACM Press, 2008, pp. 117-126

[2] Vieweg, S., Palen, L., Liu, S. B., Hughes, A. L., and Sutton, J., "Collective intelligence in disaster: An examination of the phenomenon in the aftermath of the 2007 Virginia Tech shootings.", In Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM). 2008.

[3] Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L., "Microblogging during two natural hazards events: what twitter may contribute to situational awareness.", In Proceedings of the SIGCHI conference on human factors in computing systems, pp. 1079-1088), 2010.

[4] Sutton, J., Palen, L., & Shklovski, I., "Backchannels on the front lines: Emergent uses of social media in the 2007 southern California wildfires.", In Proceedings of the 5th International ISCRAM Conference, pp. 624-632, 2008.

[5] Newell S., Robertson, M., Scarbrough, H., & Swan, J., " Managing knowledge work and innovation.", Palgrave Macmillan, 2002.

[6] Lin, H. F., "The effects of employee motivation, social interaction, and knowledge management strategy on KM implementation level.", Knowledge Management Research & Practice, 9(3), 263-275, 2011.

[7] Pérez-Nordtvedt, L., Kedia, B. L., Datta, D. K., & Rasheed, A. A., "Effectiveness and efficiency of cross-border knowledge transfer: An empirical examination.", Journal of Management Studies, 45(4), 714-744., 2008.

[8] Pew Research Center, "Social Networking Fact Sheet", 2014. [Na mreži], dostupno na: http://www.pewInternet.org/fact-sheets/social-networking-fact-sheet. [Poslednji pristup: 15.1.2015]

[9] Internet Live Stats, "Twitter Usage Statistics", 2013. [Na mreži], dostupno na: http://www.Internetlivestats.com/twitter-statistics. [Poslednji pristup 1.2.2015]

[10] Statistic Brain, "Twitter Statistics", 2014. [Na mreži], dostupno na: http://www.statisticbrain.com/twitter-statistics/. [Poslednji pristup 1.2.2015]

[11] deBoer T., Lossek M., Janssen R., and Neppelenbroek M., "Twitter: An architectural review. Scholars paper on Software Architecture at Utrecht University.", 2011. Source: http://www. timdeboer. eu/paper_publishing/Twitter_An_Architectural_Review. pdf.

[12] Zeng, L. i drugi, "Distributed data mining: a survey". Information Technology and Management, Vol.13, pp 403-409, 2012.

[13] Ingvaldsen, J. E., and Gulla, J. A., "Industrial application of semantic process mining.", Enterprise Information Systems, Vol 6, pp 139-163, 2012.

[14] Abdous M. H., He W., and Yen C. J., "Using data mining for predicting relationships between online question theme and final grade.", Educational Technology & Society, Vol 15, pp 77-88, 2012.

[15] Tane J., Schmitz C., and Stumme G., "Semantic resource management for the web: an e-learning application.", In Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters (pp. 1-10). ACM., 2004.

[16] Lotan G., Graeff E., Ananny M., Gaffney D., Pearce I., and Boyd D., "The revolutions were tweeted: information flows during the 2011 Tunisian and Egyptian revolutions", International Journal of Communication, vol. 5, 2011.

[17] H. Kwak, C. Lee, H. Park and S. Moon., "What is Twitter, a Social Network or a News Media?", In the Proceedings of the International World Wide Web Conference, 2010.

[18] Elsafoury A.F., "Monitoring urban traffic status using Twitter messages.", MA thesis. University of Twente. 2013

[19] Osiński S., Dawid W., "Carrot2: Design of a flexible and efficient web information retrieval framework." Advances in Web Intelligence. Springer Berlin Heidelberg, pp 439-444, 2005.

ABSTRACT

Quantity of the information flow in social media, is increasing daily. Such information represents a great data source which can be reused in a wide range of areas. This study intended to create an automotive system for preprocessing data and text mining using Twitter API's. The system is designed to semantically process data, store it into a database for later reuse. The paper presents results from tests that show the system can be used to server intended purpose.

TEXT MINING FROM SOCIAL NETWORKS USING API's

Nenad Mirkov, Miodrag Peranović