FILTRIRANJE TOKOVA INFORMACIJA NA INTERNETU

Krešimir Pripužić, Marin Vuković

Fakultet elektrotehnike i računarstva

{kresimir.pripuzic, marin.vukovic}@fer.hr

+385 1 6129 745

Sažetak

Na današnjem Internetu postoji ogroman broj izvorišta informacija koja velikim intenzitetom proizvode informacije u obliku toka informacija. Primjeri takvih izvorišta su mreže senzora, RSS kanali, burze dionica, aukcijska web sjedišta, računalne i telekomunikacijske mreže, itd. Sustavi za filtriranje informacija prepoznaju i uklanjaju one informacije iz ulaznog toka na koje korisnik nije pretplaćen (tj. koje nisu od interesa za korisnika), a od preostalih informacija stvaraju izlazni tok. Inherentna karakteristika postojećih sustava za filtriranje tokova informacija je neizvjesnost intenziteta izlaznog toka informacija koja je ekvivalentna spamu zato što rezultira isporukom posve nebitnih informacija korisniku. U radu je predstavljen novi model filtriranja toka informacija – "najboljih k u vremenskom prozoru" – koji neovisno o karakteristikama ulaznog toka informacija garantira konstantni intenzitet izlaznog toka.

1. Uvod i motivacija

Digitalni svemir – skup informacija koje su proizvedene, pohranjene ili kopirane u digitalnom obliku –trenutno se nalazi u fazi eksponencijalnog rasta [Gan08]. Prema istraživanju [Gan08], veličina digitalnog svemira je bila 161 EB[1] 2006. godine, 281 EB 2007. godine , a prema procjenama će 2011. biti oko 1800 EB. Drugim riječima, ukupna godišnje stopa rasta digitalnog svemira je preko 60%.

Važno je napomenuti da veličina digitalnog svemira ne bi bila problem ako bi se cjelokupni digitalni svemir mogao pohraniti na digitalne medije za pohranu. Međutim, njihov raspoloživi kapacitet je u ovom trenutku manji od ukupne veličine digitalnog svemira, a već sljedeće godine će preko 50% digitalnog svemira ostati nepohranjeno [Gan08]. Zbog toga je za svaki oblik proizvedene informacije bitno što prije utvrditi ima li ona zapravo uopće vrijednost da bude pohranjena za budućnost.

Na ovom tragu nalaze se sustavi za filtriranje informacijskih tokova koji omogućavaju filtriranje velike količine informacija u stvarnom vremenu, ali ne daju nikakve garancije za broj isporučenih informacijskih objekata u izlaznom toku tj. za intenzitet izlaznog toka. Kako je razdioba informacijskih objekata u ulaznom toku većinom unaprijed nepoznata, pretplaćivanje podsjeća na pokušaj pogađanja s posve nepoznatim ishodom. Pri tome postoje dva granična slučaja: preopćenita i prespecificirana pretplata koje u prvom slučaju rezultiraju

[1] EB (*exabyte*) = 10^{18} okteta.

prevelikim, a u drugom premalim brojem isporučenih informacijskih objekata u izlaznom toku.

2. Model obrade toka podataka "najboljih k u vremenskom prozoru"

U radu se predlaže novi model filtriranja toka podataka koji korisnicima omogućava da po svakoj svojoj pretplati kontroliraju broj informacijskih objekata koje žele primiti u odabranom vremenskom intervalu. U ovom modelu filtriranja, pretplate definiraju funkciju za rangiranje korisnosti informacijskih objekata za pretplatnika, parametar *k* te veličinu vremenskog prozora *w*. U bilo kojem trenutku *t*, parametar *k* ograničava broj informacijskih objekata dodanih u izlazni tok na *k* najbolje rangiranih od onih koji su se pojavili u ulaznom toku u periodu između *t*-*w* i *t*.

U radu su predstavljeni rezultati komparativne eksperimentalne evaluacije postojećeg i novog modela filtriranja informacijskih tokova. Ovi rezultati pokazuju da se novi model, za razliku od postojećeg modela, izvrsno prilagođava kako razdiobi, tako i intenzitetu ulaznog toka informacija.

U radu su predstavljeni rezultati eksperimentalne evaluacije centralizirane i raspodijeljene implementacije predloženog modela filtriranja informacijskih tokova. Raspodijeljeno rješenje se temelji na prekrivajućoj mreži ravnopravnih čvorova (*peer-to-peer*) CAN [Rat01] te u potpunosti nasljeđuje njene dobre karakteristike kao što su samoorganiziranost, skalabilnost i otpornost na ispade.

3. Zaključak i smjernice daljnjeg rada

Predloženi model filtriranja tokova informacija se izvrsno prilagođava intenzitetu i razdiobi informacijskih objekata u ulaznom toku te stoga predstavlja značajno poboljšanje u području obrade tokova informacija.

Postojeća implementacija sustava omogućava filtriranje tokova sa strukturiranim informacijama kao što su mreže senzora, burze dionica te aukcijska web sjedišta. U daljnjem radu se planira implementacija sustava za filtriranje tokova informacija s nestrukturiranim informacijama (tekst). Pri tome će najveći problem biti razvoj i implementacija algoritama za indeksiranje dinamičke kolekcije tekstualnih dokumenata.

4. Literatura

[Gan08] J. F. Gantz: "The diverse and exploding digital universe: An updated forecast of worldwide information growth through 2011", 2008., http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf.

[Rat01] S. Ratnasamy, P. Francis, M. Handley, R. Karp i S. Schenker: „A scalable content-addressable network", 2001., http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.116.7700&rep=rep1&type=pdf