Sveučilište u Zagrebu

Fakultet elektrotehnike i računarstva

Zavod za telekomunikacije

FILTRIRANJE TOKOVA INFORMACIJA NA INTERNETU

Marin Vuković marin.vukovic@fer.hr

Krešimir Pripužić kresimir.pripuzic@fer.hr

Konferencija CUC 2010,

Split, Hrvatska,

15.-17. 11. 2010.

Sadržaj

- Uvod
- Motivacija
- Model usporedbe top-k/w (“k najboljih u vremenskom prozoru”)
- Centralizirana i raspodijeljena obrada “top-k/w”
- Eksperimentalna evaluacija
- Zaključak i smjernice daljnjeg rada

Motivacija

- Eksponencijalni rast digitalnog svemira
  - Informacije koje su stvorene, pohranjene ili replicirane u digitalnom obliku
  - Studije korporacije IDC [92, 93] procijenile su veličinu digitalnog svemira na:
    - 161 egzaokteta (10^{18} okteta) u 2006. godini
    - 281 egzaokteta (10^{18} okteta) u 2007. godini
    - 1800 egzaokteta (10^{18} okteta) u 2011. godini
  - Godišnja stopa rasta digitalnog svemira je 60%
- Cjelokupni digitalni svemir ne možemo pohraniti
  - Po studiji istoj studiji, u 2011. godini gotovo 50% digitalno svemira neće biti pohranjeno na prijenosne medije
  - Moramo odlučiti koje informacije vrijedi pohraniti za budućnost

Model obrade podataka Store-then-Query

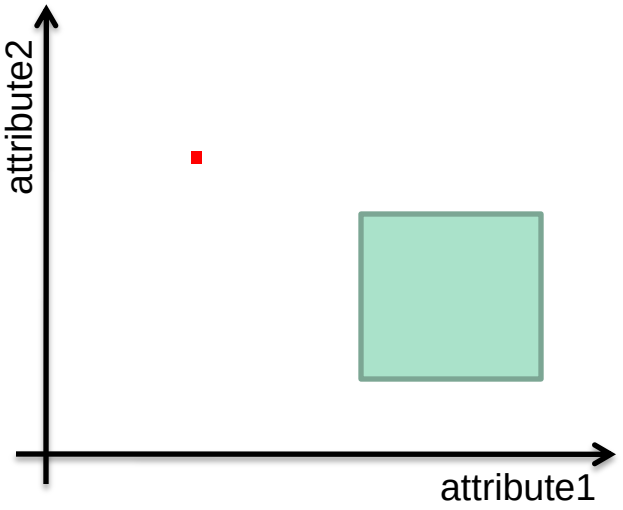
- Tradicionalni model obrade podataka
- Jednokratni upiti
- Aplikacije: sustavi za upravljanje bazama podataka i web tražilice
  - Velika frekvencija upita i mala frekvencija promjene podataka
- Nedostaci
  - Često je nemoguće pohraniti sve nastale podatke
  - Većini podataka se nakon pohrane neće više nikad pristupiti
  - Presporo za obradu u stvarnom vremenu

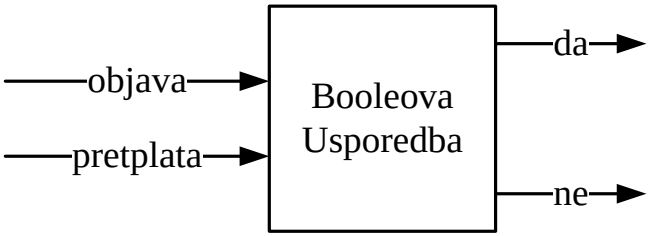
Model obrade podataka Query-then-Store

- Novi model obrade podataka
- Kontinuirani upiti (**pretplate**)
- Aplikacije: sustavi za procesiranje toka podataka (data stream processing systems) i sustavi objavi-pretplati (publish/subscribe systems)
  - Velika frekvencija dolazećih podataka (**objave**) i mala frekvencija promjene upita
  - mreže senzora, nadzor računalnih i telekomunikacijskih mreža, nadzor ponašanja korisnika (click-streams), aukcijska web sjedišta, aplikacije za brokere, tokovi vijesti (news feeds)
- Nedostaci
  - Kompliciranа je promjena aktivnog upita

Booleov model usporedbe (pretplata i objava)

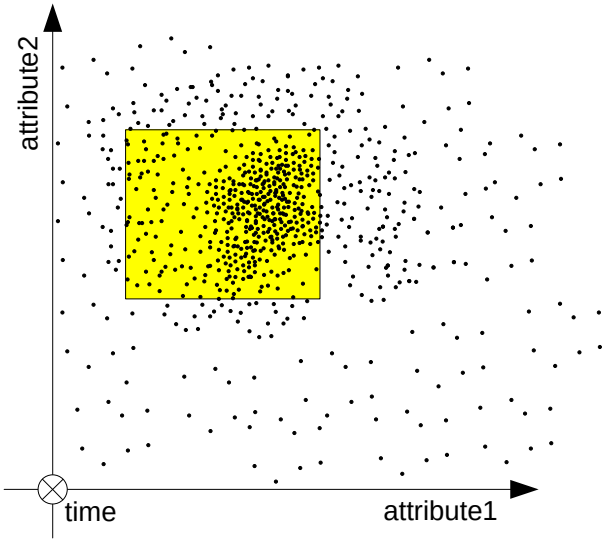
- Objava je najčešće točka u višedimenzionalnom prostoru
  - Očitanje senzora
  - Cijena dionice
  - Oglas
  - Vijest
- Pretplata je potprostor višedimenzionalnog prostora
  - Pretplata je Booleova funkcija
  - Statični potprostor pretplate – nema unutarnjeg stanja
  - Rezultat usporedbe ovisi samo o sadržaju pretplate i objave

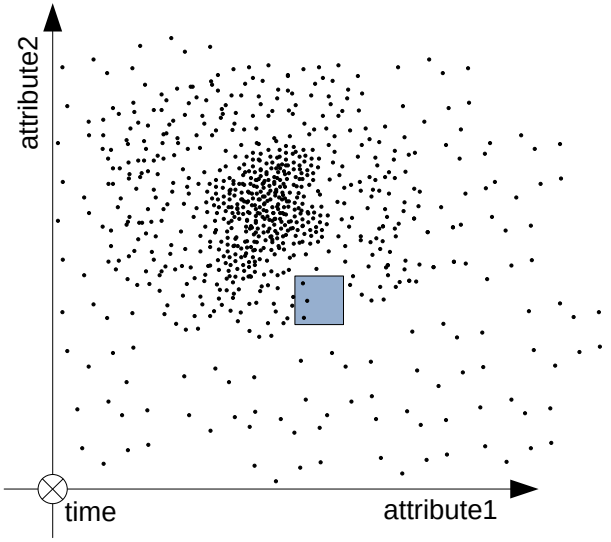




Nedostaci Booleovog modela usporedbe

- Koliko objava će odgovarati pretplati prilikom usporedbe?
  - Ovisi o sadržaju objava
  - Uglavnom unaprijed nepoznat
- Pretplaćivanje sliči nagađanju – nepoznat ishod
- Dva granična slučaja
  - Preopćenita pretplata
  - Pretplata je prespecificirana
- Rangiranje objava nije podržano
  - Sve objave koje su jednako relevantne za pretplatu

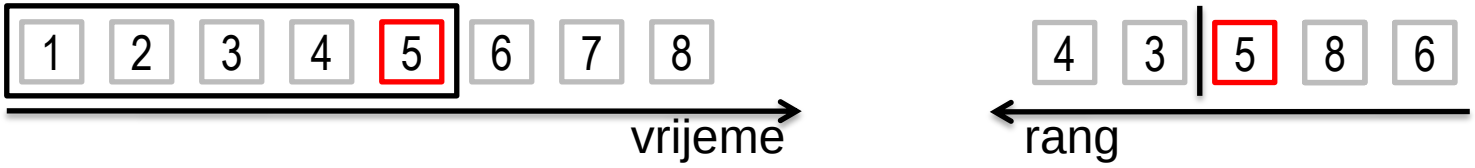




Model usporedbe “k najboljih” u vremenskom prozoru

- Naša ideja
  - Rangirati objave u svakoj poziciji prozora
  - Dostaviti *k* najbolje rangiranih objava pretplatniku

- Kada objava može postati jedna od “najboljih *k*” u prozoru pretplate?
  - Neposredno nakon objave
  - Kasnije ukoliko starije i bolje rangirane objave ispadnu iz prozora, a naknadno objavljene budu lošije rangirane
    - Primjer pretplate koja isporučuje 2 najbolje od 5 zadnje objavljenih objava (top-2/5)



- Potrebno je u memoriji pohraniti potencijalne kandidate za “k najboljih”

Implementacija

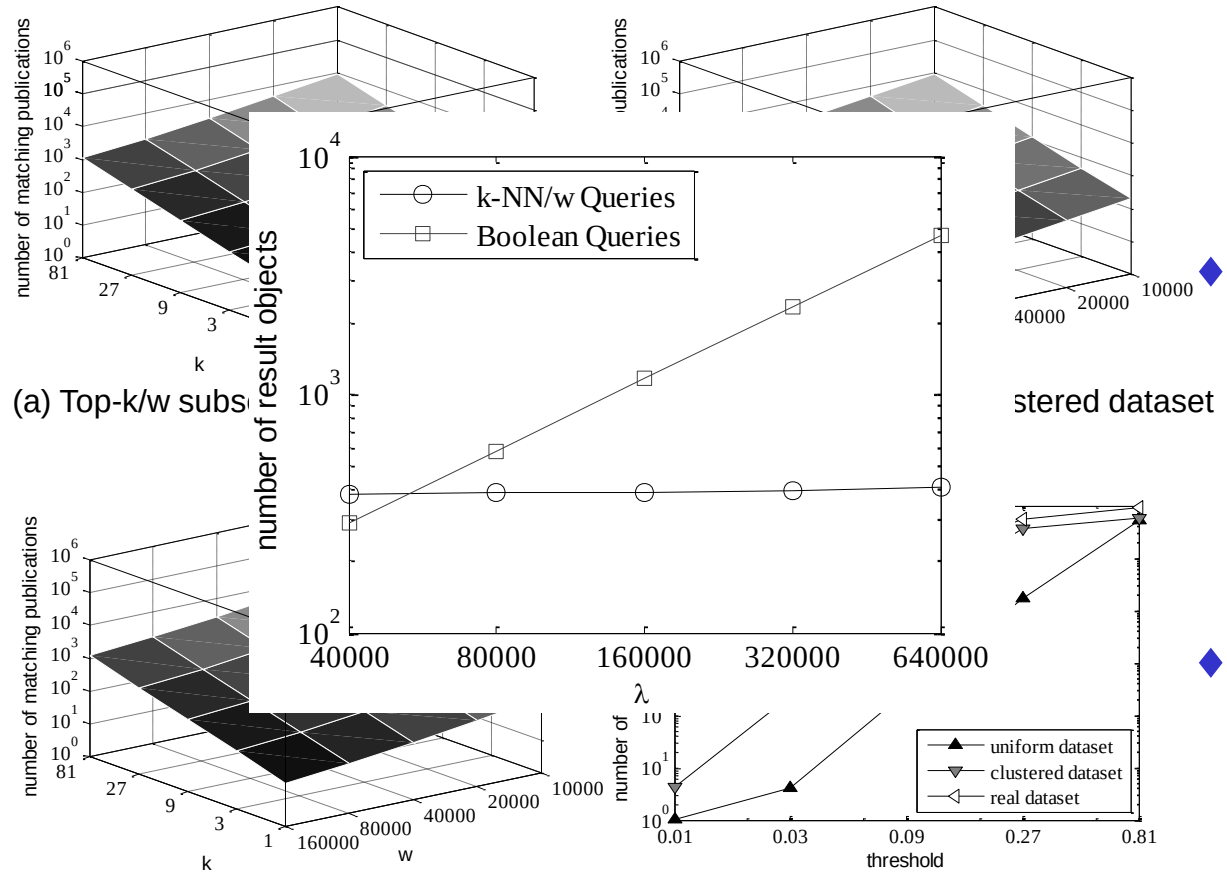
- Problemi pri implementaciji modela usporedbe *top-k/w*
  - Ograničena količina memorije
  - Ograničena procesorska snaga
  - Ograničena propusnost mreže
- Centralizirano rješenje
  - Ukoliko je frekvencija objavljivanja velika sve objave se fizički ne mogu dostaviti centraliziranom procesoru
  - Centralizirani procesor predstavlja jedinstvenu točku ispada
- Raspodijeljeno rješenje
  - Nema nedostataka
  - Problematična implementacija

Parametri eksperimentalne evaluacija

- Objave i pretplate su višedimenzionalne točke
- Euklidska udaljenost između točaka je uzeta kao koeficijent odgovaranja
- Skupovi podataka
  - Stvarni skup podataka
    - *LUCE deployment data* (podatke je skupila senzorska mreža velikih dimenzija za nadgledanje okoliša - projekt SensorScope http://sensorscope.epfl.ch/)
    - Nasumično smo uzorkovali točke iz ovog skupa podataka
  - Sintetizirani skupovi podataka
    - Uniformno generirani podaci
    - Podaci klasterirani po Gaussovoj razdiobi

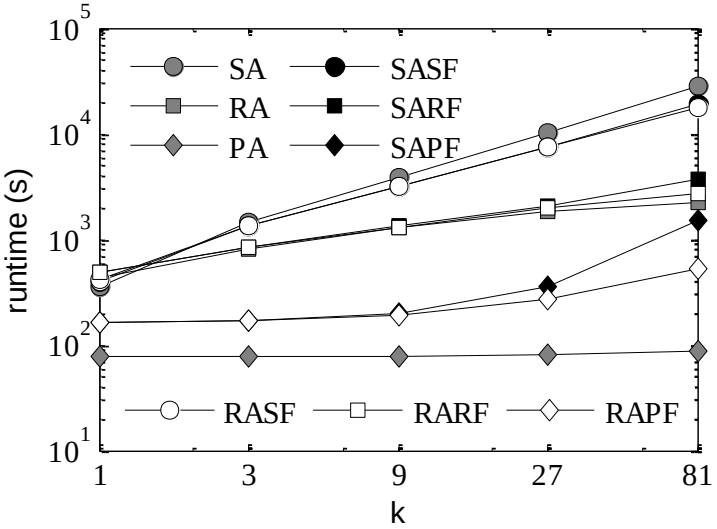
Parametar	Vrijednost
Broj objava	1.000.000
Broj pretplata	400
Parametar *k*	9
Veličina prozora pretplate	40.000
Dimenzionalnost podataka	4
Razlučivost strukture za indeksiranje	10 do 12
RA: koeficijent čišćenja	0,2
PA: vjerojatnost pogreške	1.000
Broj čvorova u mreži	256

Evaluacija modela *top-k/w*

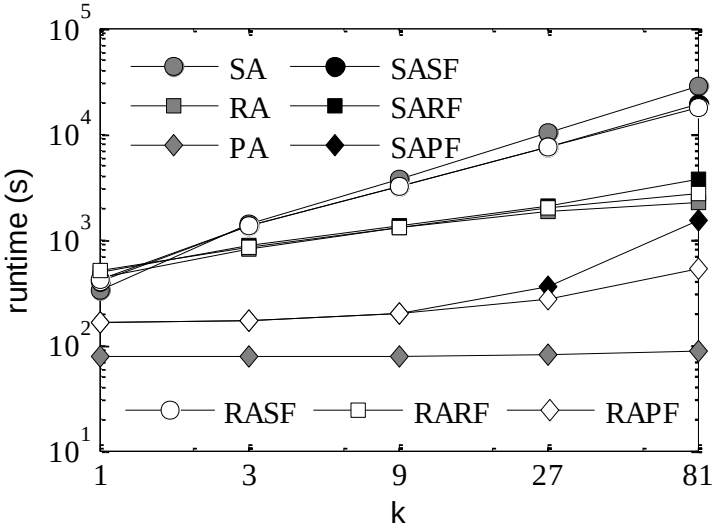


- Model top-k/w se vrlo dobro prilagođava podacima različitih razdioba
- Prag Booleove pretplate (*threshold*) predstavlja polumjer višedimenzionalne sfere
- U Booleovom modelu broj objava koje odgovaraju pretplati značajno ovisi o:
  - Razdiobi podataka
  - Intenzitetu objavljivanja

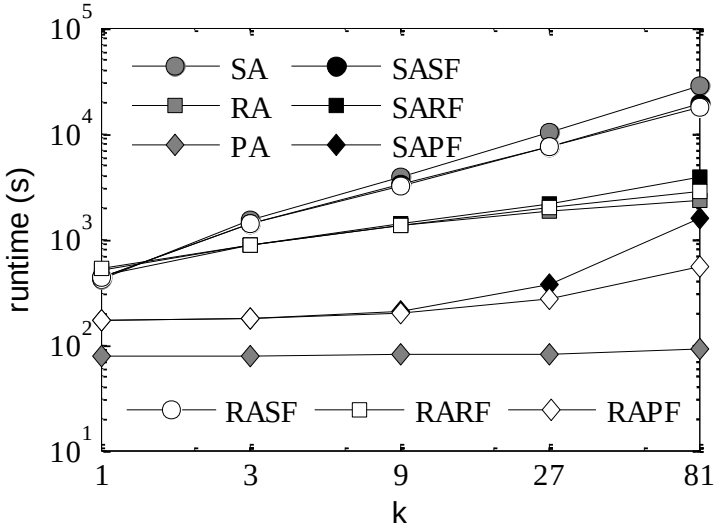
Centralizirana implementacija modela *top-k/w*



(a) Uniform dataset

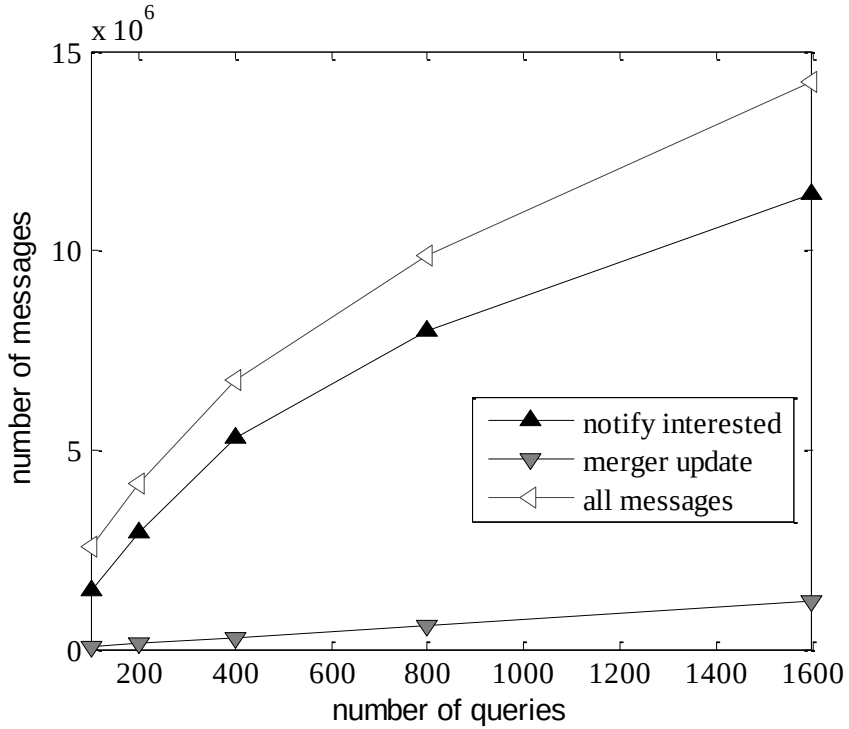


(b) Clustered dataset

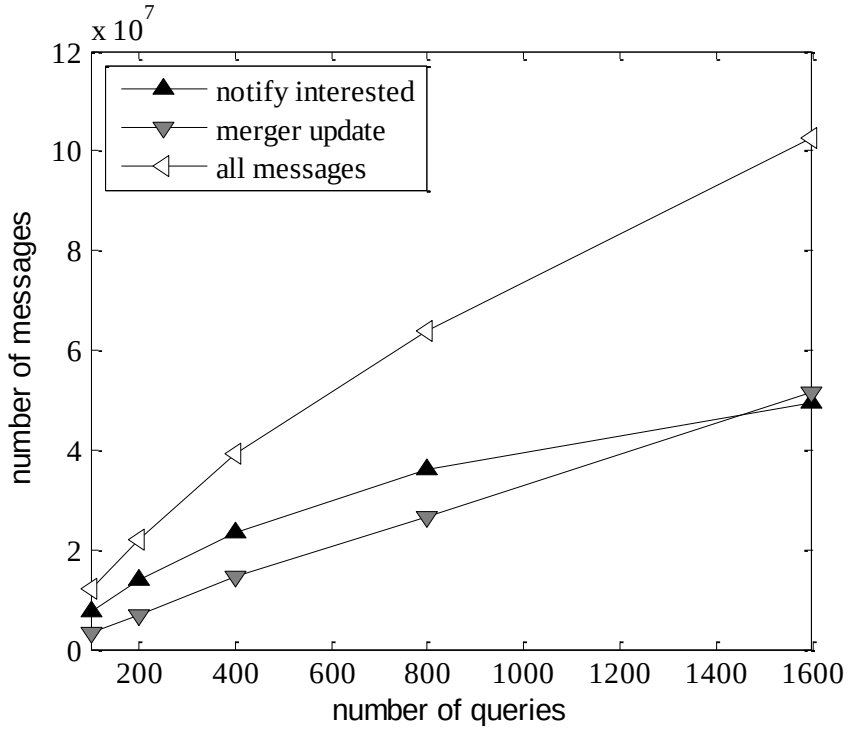


(c) Real dataset

Skalabilnost pretplata u raspodijeljenoj implementaciji modela *top-k/w*

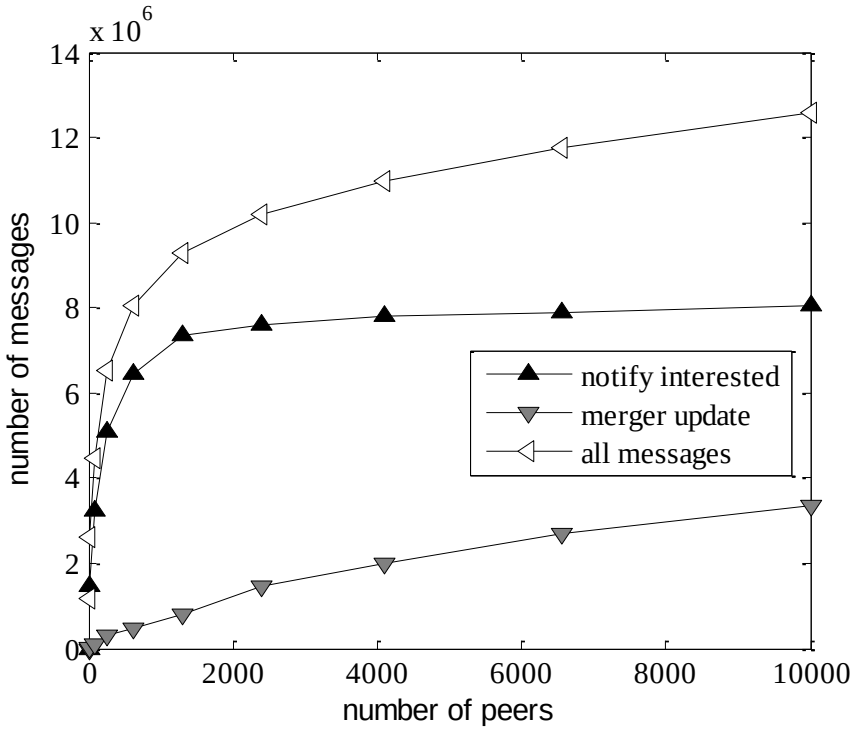


(a) PA-based subscription

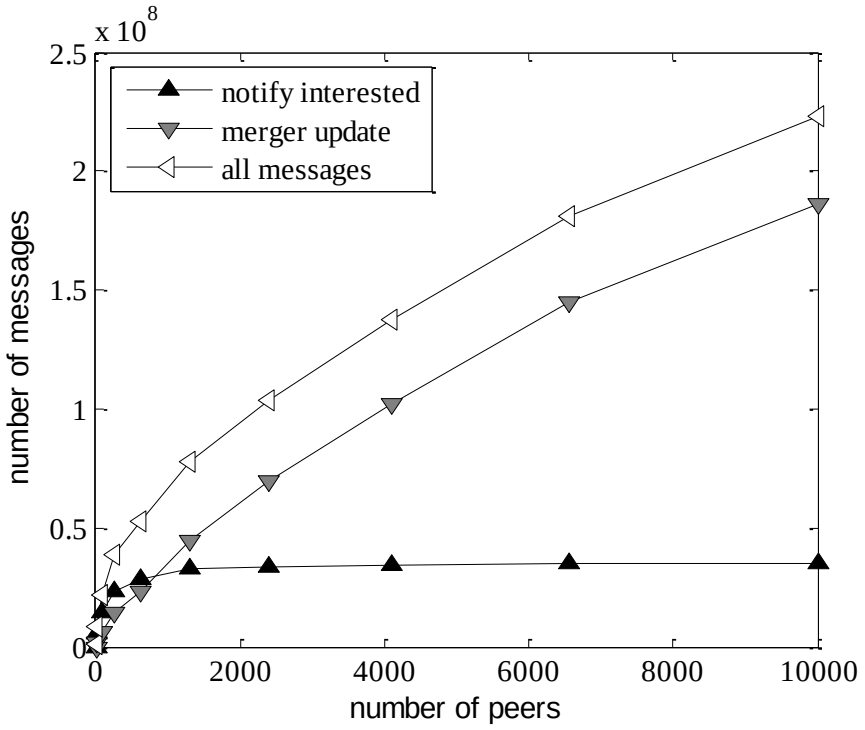


(b) RAPF-based subscription

Skalabilnost čvorova u raspodijeljenoj implementaciji modela *top-k/w*



(a) PA-based subscription



(b) RAPF-based subscription

Zaključak i smjernice daljnjeg rada

- Model top-k/w se može koristiti za učinkovito filtriranje informacija
- Razvijeni su učinkoviti algoritmi za obradu podataka
- Drugačiji pristup rangiranju objava
  - Top-k/w definiraju oštru granicu u vremenu – objave se trenutačno izbacuju iz prozora pretplate
  - Postupno smanjivanje koeficijenta odgovaranja u vremenu
- Indeksiranje pretplata u vektorskom prostoru
  - Omogućilo bi upotrebu modela *top-k/w* za tekstualne podatke
  - Kolekcija podataka (tj. objava) je vrlo dinamična u ovom slučaju što je potpuno neistraženo područje