Računalništvo in informatika

Pridobivanje podatkov iz omrežja DHT

Analiza pretočnega prometa skozi vozlišča protokola BitTorrent in prenos metapodatkov

Anton Luka Šijanec <anton@šijanec.eu>

4. letnik

□

Mentor: Andrej Šuštaršič, univ. dipl. ing. elektr.

2023
Gimnazija Bežigrad

Kazalo

Povzetek

Porazdeljene razpršilne tabele (angl. distributed hash table) so razpršilne tabele, ki podatke, ponavadi so to dokumenti, strukturirani kot vrednost in njen pripadajoč ključ, hranijo distribuirano na več vozliščih, kjer se podatki shranjujejo. V računalniških sistemih se DHT uporablja za hrambo podatkov v omrežjih P2P (angl. peer to peer), kjer se podatki vseh uporabnikov enakomerno porazdelijo med vozlišča in so tako decentralizirani in preprosto dostopni članom omrežja. Ker se podatki izmenjujejo znotraj omrežja na vozliščih, ki z izvorom in destinacijo podatkov niso povezani, jih lahko vozlišča v velikih količinah shranjujejo.

V raziskovalni nalogi je preverjena praktična zmožnost pridobivanja velike količine podatkov v omrežju BitTorrent za P2P izmenjavo datotek, pridobljeni podatki pa so analizirani. Vsaka poizvedba po seznamu imetnikov datotek vsebuje ključ podatka v DHT in se prenese preko okoli $\log_2 n$ vozlišč, kjer je n število vseh uporabnikov v omrežju. Ker vsaka poizvedba obišče tako veliko število vozlišč, lahko eno vozlišče prejme veliko obstoječih ključev v omrežju, s katerimi si lahko prenese metapodatke v omrežju BitTorrent.

Naloga se osredotoči na pridobivanje metapodatkov v omrežju BitTorrent, glede prenosa datotek, ki jih ponujajo računalniki, pa se vsled njihove velikosti ne opredeli. Metapodatki konceptualno sicer niso shranjeni v DHT (namesto metapodatkov o datotekah so v omrežju shranjeni seznami računalnikov, od katerih si metapodatke lahko prenesemo), vendar odkrivanje njihovega obstoja omogoči DHT.

Ključne besede porazdeljena razpršilna tabela, distribuirani sistemi, P2P omrežje, podatkovno rudarjenje, BitTorrent

1 Uvod

1.1 Peer-to-peer omrežja za distribucijo datotek

Izmenjava in distribucija velikih datotek na internetnih omrežjih veliki količini odjemalcev predstavlja težavo, saj je v osnovi TCP/IP sklada protokolov isto datoteko poslati tolikokrat, kolikor odjemalcev imamo. Distributorji večjih količin podatkov na internetu se morajo zaradi centraliziranega modela infrastrukture strežnikov, kjer centraliziran strežnik posreduje identične informacije večkrat večim odjemalcem, ki med seboj ne komunicirajo, posluževati dragih metod kolokacije strežnikov.

Koncept P2P (angl. peer-to-peer) predstavlja alternativen način distribucije identičnih datotek večim odjemalcem. Namesto enega strežnika, ki iste podatke pošlje vsakič znova odjemalcem, v omrežjih P2P za distribucijo datotek ni razlike med strežnikom in odjemalcem. Vsak odjemalec podatke tako prejema kot tudi pošilja. Takoj ko odjemalec prejme vsebino od drugega odjemalca, jo bo tudi sam začel deliti naprej drugim odjemalcem, ki to vsebino tudi sami iščejo. S svojim sodelovanjem v distribuciji vsebine razbremeni ostale odjemalce, ki datoteke distribuirajo prosilcem, saj so P2P omrežja izdelana tako, da lahko odjemalci vsebino prejemajo od večih odjemalcev hkrati. Čim več odjemalcev razpolaga z neko vsebino, tem manj podatkov mora poslati posamezen odjemalec novemu odjemalcu, ki si to vsebino želi prenesti. Tako se zmanjša obremenitev omrežja, saj je količina prenesenih podatkov po omrežni topologiji čedalje bolj razporejena.

Sistem pa ni povsem brezhiben, saj je še vedno treba na nek zunanji način med seboj povezati odjemalce, ki so zainteresirani za določeno temo (recimo za določeno datoteko). Druga očitna slabost pa je, da je možno ugotoviti, kdo prenaša kakšno vsebino, ker odjemalci (neke datoteke) vedo za internetne naslove drugih odjemalcev, saj lahko le tako neposredno čim bolj učinkovito komunicirajo z njimi.

Koncept P2P ni namenjen le distribuciji datotek, temveč se zaradi svoje prednosti razbremenitve strežnikov dandanes vse pogosteje uporablja, na primer pri spletnih videokonferencah, anonimizacijskih omrežjih, kriptovalutah, internetu stvari in drugje.

1.2 Protokol BitTorrent

Za distribucijo datotek morajo odjemalci za medsebojno komunikacijo uporabljati standardiziran protokol za signalizacijo prenosov. Eden izmed najbolj razvitih (ci-

tiraj) in uporabljenih protokolov je BitTorrent. Prvo implementacijo je idejni avtor protokola izdelal leta 2001(citiraj), od leta 2008 pa lahko z objavo dodatkov pri razvoju skupaj sodeluje širša javnost(citiraj). Zaradi razširljive zasnove je protokol namreč moč dopolnjevati in mu s tem dodajati nove funkcije. Sprva je na primer protokol omogočal le pospešeno distribucijo datotek iz enega strežnika k več odjemalcem (citiraj), saj so si odjemalci koščke vsebine delili med seboj, vendar je še vedno temeljil na centralnih strežnikih, ki stalno gostijo datoteke in koordinirajo skupek odjemalcev, danes pa omogoča (citiraj) od centraliziranih strežnikov povsem neodvisno delovanje, prav z uporabo protokola DHT.

Za nadaljnji opis je potrebno poznavanje pojmov, ki jih uvede BitTorrent:

Pojem	Izvirno angleško ime	Razlaga
soležnik (citraj)	peer	odjemni program na računalniku ali računalnik, za povezavo nanj potrebujemo njegov IP naslov in vrata
roj (citiraj)	swarm	več soležnikov, ki prenašajo datoteke torrenta
torrent/metainfo (ni ustaljenega prevoda, neposredni prevod bi bil *hudournik*)	torrent ali metainfo	strukturirana datoteka v obliki bencoding, ki vsebuje metapodatke o datotekah, torej imena datotek, njihove velikosti, razporeditev po imenikih, zgoščene vrednosti za preverjanje istovetnosti ob prenosu in drugo
sledilnik (citiraj)	tracker	centraliziran strežnik, ki hrani podatke o tem, kateri soležniki so v roju določenega torrenta
košček (citiraj)	piece	del vsebine torrenta konstantne dolžine
infohash (ni ustaljenega prevoda)	infohash	zgoščena vrednost serializiranih podatkov pod ključem info v torrentu, ki unikatno opišejo ključne metapodatke o torrentu
announce (ni ustaljenega prevoda, neposredni prevod bi bil *obvestilo*)	announce ali ~ment	obvestilo ali obveščanje o obstoju soležnika za torrent, ki ga pošlje soležnik bodisi sledilniku bodisi v DHT in s tem zagotovi, da bodo ostali soležniki izvedeli za njegov obstoj in se potencialno povezali nanj

Tabela 1.1: Nepopoln seznam pojmov BitTorrenta, potrebnih za razumevanje naloge

BitTorrent protokol ne omogoča iskanja po datotekah, ki se prenašajo po omrežju. Za prenos datoteke je najprej treba poznati metapodatke o obstoječih datotekah. Ti metapodatki so shranjeni v t. i. obliki torrent, strojno berljivi datoteki, serializirani

s preprosto serializacijsko metodo bencoding. Vsebujejo imena in poti datotek ter njihove zgoščene vrednosti, ime torrenta, lastnosti prenosa: velikost koščka, ime, zasebnost (angl. private torrent).

V nalogi se ne osredotočam na klasičen način iskanja soležnikov s sledilniki, prav tako ne govorim o prenosu datotek od soležnikov ter o signalizaciji za omejevanje pasovne širine prenosa (choking), temveč samo o prenosu metapodatkov.

1.3 Protokol BitTorrent DHT

DHT je kot koncept definiran zelo splošno, za BitTorrent je uporabljen sistem DHT, imenovan Kademila. Uporablja se odpravo odvisnosti od sledilnika, saj lahko v njej hranimo seznam soležnikov v roju.

Pojem	Izvirno angleško ime	Razlaga
vozlišče (citiraj)	node	odjemni program na računalniku ali računalnik
usmerjevalna tabela (citiraj)	routing table	seznam vozlišč, ki ga hrani posamezno vozlišče
ID vozlišča	node ID	160 bitov dolga naključno generirana številka, ki pripada vsakemu vozlišču
merilo za razdaljo	distance metric	funkcija (XOR), ki izrazi konceptualno razdaljo kot 160 bitov dolgo številko med dvema vozliščema
koš (citiraj)	bucket	na posamezno vozlišče relativna množica drugih vozlišč, ki so si glede na merilo za razdaljo blizu, shranjena v usmerjevalni tabeli

Tabela 1.2: Nepopoln seznam pojmov Kademile, potrebnih za razumevanje naloge. Za noben pojem nisem našel ustaljenih slovenskih prevodov.(citiraj)

Kademilo, kot se uporablja v BitTorrentu, si lahko za začetek predstavljamo kot abstraktno razpršilno tabelo, ki je shranjena porazdeljeno na velikem omrežju vozlišč/računalnikov in podpira naslednji operaciji (citiraj):

Pridobi soležnike Vrne seznam soležnikov (IP naslov in vrata) za torrent, opisan z njegovim infohashom.

Announce V seznam soležnikov za torrent, opisan z njegovim infohashom, vstavi IP naslov in vrata pošiljatelja zahteve.

Cilj raziskovalne naloge je s sodelovanjem v DHT omrežju pridobiti čim več obstoječih ključev v razpršilni tabeli, da lahko z operacijo **pridobi soležnike** pridobimo sezname soležnikov, na katere se lahko povežemo in od njih prenesemo metapodatke o torrentih, da lahko te podatke kot izvleček celotnega omrežja kasneje uporabimo za analiziranje.

1.4 Obstoječe implementacije

Da je to pridobivanje mogoče, se ve že od vpeljave protokola DHT, saj obstaja mnogo implementacij koncepta pridobivanja podatkov iz omrežja DHT za prenos metapodatkov torrentov:

- Spletna stran in istoimenski program **Btdigg** (citiraj)
- Spletna stran v kitajščini pod več imeni: **clzhizhu**.com, cilizhizhu, clzz1020.buzz, clzz1025.buzz, clzz1026.buzz idr. Za obstoj te strani sem ugotovil med implementacijo programa, saj je njeno iskanje invazivno in moti obstoječe delovanje DHT. Več o tem v sledečih poglavjih.
- Spletna stran **I know what you download** (citiraj), ki hrani najdene podatke o rojih in s tem razkrije identiteto prenašalcev.

2 Teoretični del

2.1 Protokol BitTorrent

2.2 Protokol BitTorrent DHT

3 Eksperimentalni del

3.1 Program travnik

3.2 Prestrezanje podatkov

3.3 Obdelava podatkov

4 Rezultati

4.1 Analiza podatkov

5 Razprava

5.1 Uporabna vrednost korpusa prenesenih podatkov

5.2 Invazivnost v omrežje

6 Zaključek

6.1 Načrti za prihodnost

Zahvala

Za pomoč pri obdelavi podatkov se zahvaljujem Oliverju Wagnerju (oliwerix.com.) in Adrianu Sebastianu Šiški (ass.si.).

Literatura