Μηχανική Δικτύων 2021

Βιβλιογραφική Εργασία:
Big Data Caching – Moving from Cloud to Edge

Άρτεμις Γεωργοπούλου, 3374

Στο κείμενο γίνεται αναφορά σε μια προτεινόμενη αρχιτεκτονική δικτύωσης που πηγάζει από την όλο και μεγαλύτερη εφαρμογή και χρήση των 5G δικτύων. Θα περιγραφεί το πρόβλημα με την τρέχουσα αρχιτεκτονική, έπειτα η προτεινόμενη λύση και τέλος, θα παρουσιαστεί μια έρευνα για την λειτουργικότητα της προτεινόμενης λύσης.

Αρχικά, γνωρίζουμε ότι υπάρχει μεγάλη κίνηση δεδομένων στις ασύρματες δικτυώσεις, λόγω του μεγάλου πλήθους συσκευών, της σύνδεσης οποτεδήποτε και οπουδήποτε και του κατά-απαίτηση μεγάλου όγκου περιεχομένου. Ως αποτέλεσμα έχει την χρήση υπερβολικών πόρων του backhaul (The backhaul is the link between the network serving as the backbone for other networks and other sub-networks. Also, the transportation of data or network between access points to the public is backhaul. Backhaul connects the central network to the individual networks or public networks) και αυτό το πρόβλημα θα γίνει ακόμα μεγαλύτερο με την εισαγωγή της γρήγορης 5G δικτύωσης.

Η τρέχουσα αρχιτεκτονική είναι βασισμένη μακριά από τον χρήστη (base-station centric) και λειτουργεί με reactive τρόπο (αντιδρά, αφού ο χρήστης κάνει κάποια ενέργεια). Η προτεινόμενη αρχιτεκτονική από την άλλη, θα είναι βασισμένη κοντά στον χρήστη (user centric) και θα λειτουργεί με proactive τρόπο (θα ενεργεί πριν τον χρήστη σε πολλές περιπτώσεις), καθώς θα λαμβάνει υπόψιν το περιεχόμενο που θα ζητήσει ο χρήστης (context-aware).

Για να γίνει αυτό (να προβλέπεται το περιεχόμενο που θα ζητήσει ο χρήστης) προτείνεται να γίνει χρήση Big Data. Από την μεγάλη πληθώρα δεδομένων στο ίντερνετ, αν γίνει ανάλυση, επιστρατεύοντας στατιστική και αλγορίθμους μηχανικής μάθησης, μπορεί να επιτευχθεί μια καλή κατανόηση

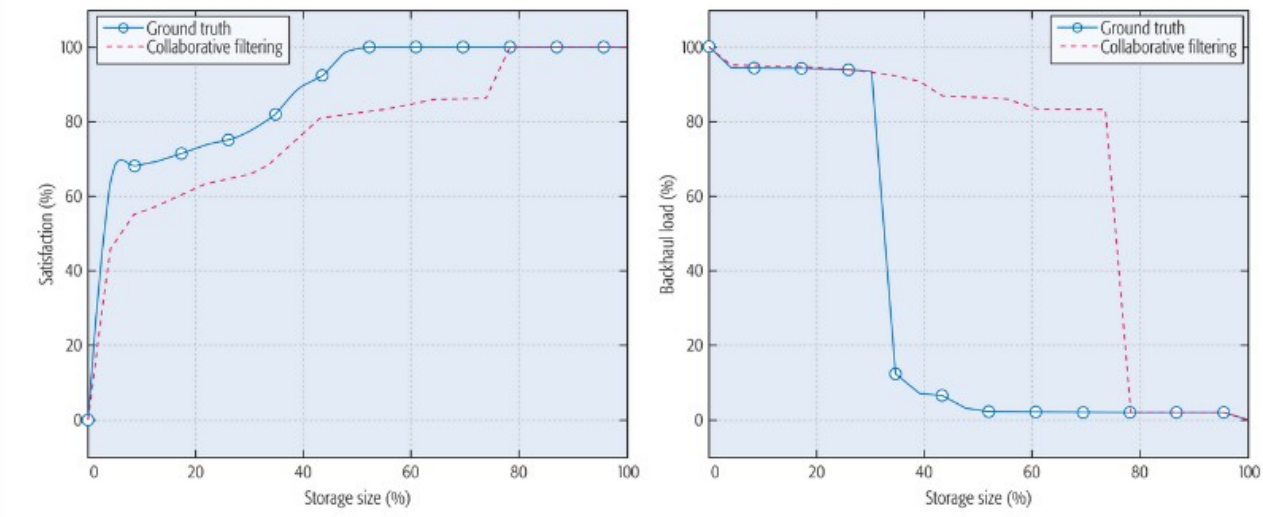
της συμπεριφοράς των χρηστών και του περιεχομένου που είναι δημοφιλές ή σχετικό. Έτσι, προβλέποντας το περιεχόμενο που θα ζητήσει ο χρήστης, δίνεται η δυνατότητα caching του περιεχομένου στον κοντινότερο σταθμό (caching at the edge). Όταν τελικά ζητήσει ο χρήστης το συγκεκριμένο περιεχόμενο θα μειωθεί η απόσταση μεταξύ χρήστη και σταθμού και θα αυξηθεί η ταχύτητα μετάδοσης. Παράλληλα μειώνεται η χρήση του backhaul.

Για την πρόβλεψη του περιεχομένου, πρέπει να γίνει ανάλυση πολλών δεδομένων του δικτύου. Πρέπει να αφαιρεθούν τα λανθασμένα ή μη λειτουργικά πακέτα, έπειτα να διαλεχθούν τα σχετικά πεδία από τα πακέτα και τέλος, να φορμαριστούν καταλλήλως,. Έπειτα, βρίσκονται συσχετισμοί μεταξύ των δεδομένων βάσει τοποθεσίας και χρονικής στιγμής και δημιουργείται μια βάση δεδομένων και ένας πίνακας δημοτικότητας (popularity matrix) από όπου γίνονται οι προβλέψεις για τον χρήστη. Η βάση επαναχρησιμοποιείται και επαυξάνεται βάσει δημοτικότητας του περιεχομένου σε κάθε τοποθεσία.

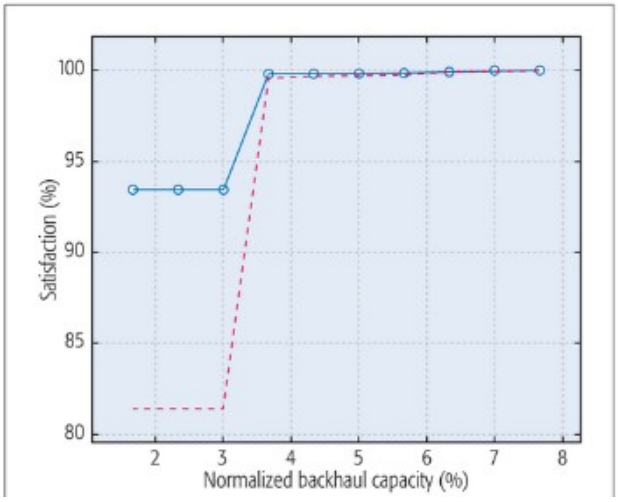
Για την έρευνα που διεξήχθη πάνω στην προτεινόμενη λύση, συλλέχθηκαν δεδομένα από πάροχο δικτύου της Τουρκίας. Συνολικά αναλύθηκαν 80 TB δεδομένων για χρήση proactive caching, χρησιμοποιώντας μια πλατφόρμα ανάλυσης δεδομένων βασισμένη σε Hadoop (Apache Hadoop is a collection of open-source software utilities that facilitates using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model). Η δημιουργία του πίνακα δημοτικότητας εξετάστηκε με δύο τρόπους: είτε χρησιμοποιώντας αποκλειστικά τα δεδομένα που συλλέχθηκαν σε ποσοστό 100% (ground truth), είτε παίρνοντας τυχαίο ποσοστό 30% από τα συλλεχθέντα δεδομένα και προβλέποντας το υπόλοιπο ποσοστό με SVD (regularized singular value decomposition από μεθόδους Collaborative filtering).

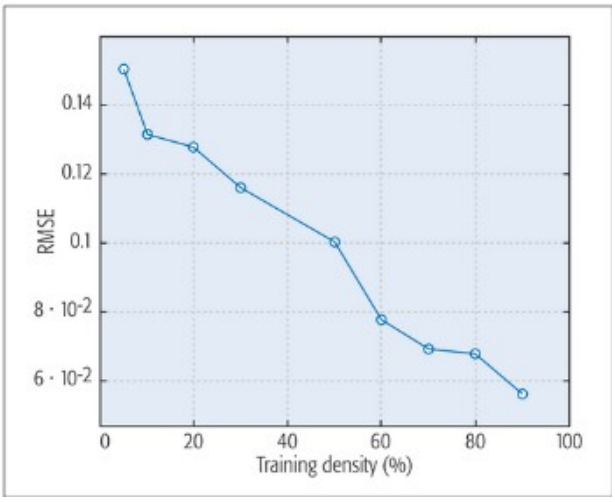
Έπειτα, υποτέθηκε ότι από τα υπάρχοντα δεδομένα που μαζεύτηκαν, ζητήθηκαν τα D (D=422.529 requests) περιεχόμενα, σε μια χρονική περίοδο περίπου 7 ωρών. Τα υποτιθέμενα αιτήματα ανατέθηκαν ψευδοτυχαία σε N (N=16) μικρά base-stations, τα οποία έχουν πανομοιότυπες ιδιότητες (wireless link/back haul link/storage capacities). Σε κάθε base-station έγιναν cache όσα δημοφιλή περιεχόμενα χωρούσαν.

Κάνοντας cache στα μικρά base-stations τα δεδομένα/περιεχόμενα σε ποσοστό από 0-100% (100%=17.7GB) του συνολικού μεγέθους της βιβλιοθήκης, παρατηρούμε ότι και με τη μέθοδο ground truth και με την μέθοδο Collaborative filtering τα ποσοστά ικανοποίησης των χρηστών αυξάνονται (request satisfaction = amount of contents delivered at a given target rate), όσο αυξάνεται και το μέγεθος των δεδομένων που γίνονται cache. Επίσης, παρατηρείται μείωση φόρτου του backhaul όσο αυξάνεται το storage size. Η μεγάλη μείωση του φόρτου με την μέθοδο Ground truth οφείλεται στο ότι με αυτή τη μέθοδο υπάρχουν όλες οι πληροφορίες για την αξιολόγηση δημοτικότητας των περιεχομένων σε σχέση με την μέθοδο CF.



Τέλος, παρατηρείται αύξηση της ικανοποίησης όσο αυξάνεται το normalized backhaul capacity (total backhaul capacity link/total wireless link capacity). Επίσης, αν και τα παραπάνω αποτελέσματα γίναν με 30% rating density σε CF, παρατηρείται ότι αυξάνοντας το ποσοστό training rating density του CF μειώνεται το estimation error και άρα αυξάνονται τα ποσοστά ικανοποίησης με CF.


Satisfaction (%)
100
95
90
85
80
2
3
4
5
6
7
8
Normalized backhaul capacity (%)


RMSE
0.14
0.12
0.1
8 · 10-2
6 · 10-2
0
20
40
60
80
100
Training density (%)