Proposition de recherche doctorale

Joint Optimization of Content Caching and Recommendation for Mobile Edge Systems

Mots clés :
● Directeur de thèse : christian BONNET
● Co-encadrant(s) :
● Unité de recherche : Laboratoire de recherche d'EURECOM
● Ecole doctorale : École Doctorale Informatique, Télécommunications, Électronique de Paris
● Domaine scientifique principal: Divers

Résumé du projet de recherche (Langue 1)
Summary Caching popular content at the edge of future mobile networks (e.g. small cells or user devices) has been widely considered in order to alleviate the backhaul bottleneck resulting from very dense, small cell networks, that are deemed necessary to keep up with the data tsunami experienced. While a number of interesting techniques have been proposed that target the wireless environment explicitly, such as femto-caching and coded caching, the majority of these approaches suffer from the rather limited storage capacity of edge caches. Even if one considers the additional “global” caching gains these methods promise, the envisioned gains in practical settings are rather pessimistic, due to the tremendous and rapidly increasing size of the Internet content catalog. In this project, we propose to depart from the assumption of hard cache misses, common in most of these works, and consider soft cache misses, where alternative content, related to the original one not found locally, can be recommended possibly leading to complete or partial user satisfaction, without the need to retrieve the original content over the expensive backhaul and core network. This thesis proposes a first-of-its-kind joint consideration of caching and recommendation systems, towards facilitating mobile edge caching performance. Such a convergence is quite timely due to popular content providers envisioning becoming virtual operators, facilitated by RAN Sharing and Mobile Edge Computing architectures considered in 5G.

Context The “network side” of caching: While caching had been widely studied in peer-to-peer systems and content distribution networks (CDNs), the seminal work of femto-caching was among the first to propose caching content on a set of inexpensive edge nodes (called “helper”) nodes, which can be, for example, femto-cells with limited backhaul capacity. In this context, the network topology and the content popularity distribution (both assumed to be known) are used to solve the optimal content placement problem: which contents to store in which cache. A number of follow-up works within the “femto-caching” framework have appeared, considering aspects such as storage in user devices, constraints on transmission capacity for helper leading to a placement and routing problem, multi-layer video streaming where video quality can be traded off with hit rate, and multicast through multiple helper nodes, using LTE’s eMBMS framework. A number of other aspects have also been considered in this framework, such as local popularity patterns, using social relations of users to improve prefetching, as well as cache replacement policies. These latter works use stochastic geometry methods to model the random placement of users and helper nodes and resulting topology, rather than assuming a given one. The common denominators between these works can be summarized as follows: (i) The main bottleneck is the backhaul link (ii) the transmission phase is ignored or simplified, (iii) global caching gains stem from coverage overlaps between nearby cells. The “communications side” of caching: When multiple nearby users are requesting content at the same time, the content delivery side of the problem in a wireless setup becomes just as important as the placement problem. Recent work by Maddah-Ali and Niesen revealed quite interesting findings about the fundamental gains achievable by jointly considering caching and coded transmissions. In a simple setup with a catalog of N files, K receiving users each with a request for any one of the N files, storage at the user devices with a cache of size of M (< N) files, the authors show that the amount of time T can be reduced not only due to the local caching gain (M/N) of traditional caching schemes, but also due to a global caching gain that scales with the total size of caching memory, KM. While the above results were derived in a rather stylized setting, a number of follow-up studies have tried to address various shortcomings of the original work, towards distributed implementations, non-uniform popularity distributions, replacement algorithms, hierarchical topologies, etc. Beyond these works, caching has also been considered in the context of advanced cooperative transmission techniques on the physical layer, such as CoMP. CoMP can significantly improve performance, but requires all BS involved to exchange contents, to create a virtual MIMO channel, a large overhead on the already taxed backhaul. However, if every BS involved already stores the contents requested by every user involved, then a collaborative transmission can be performed by just knowing the channel matrix. Some authors argue that the caching and transmission algorithms at each involved BS must be jointly designed in order to facilitate such CoMP opportunities. Finally, in a very recent work of ideas from coded caching are also used to derive fundamental performance bounds on the impact of caching for a simple K-user interference channel. Beyond state-of-the-art – the “content side” of caching: In the content placement problem, the global gain comes from K > 1 small cells being in range of a user, thus increasing the effective cache size by up to K. Even with high densification K is expected to be less than 10 in most cases. In coded caching, the global gain comes from K nearby nodes overhearing a single broadcast transmission, and again ripping gains proportional to K. In practice, small cells cannot have more than 10-16 users associated. Hence, if we consider that each cache can store about 0.1-0.01% of the entire content catalogue (an already highly optimistic value), one would require orders of magnitude larger K values to obtain real gains. The main novelty of this thesis, in terms of approach, is to depart from the assumption of hard cache misses and consider soft cache misses. A request for content X leading to a cache miss is not immediately served from the core network, but alternative contents “similar” to X and locally stored can be proposed to the user. The user might then be fully satisfied with the proposed content (in the best case), fully dissatisfied (in the worst case) leading to a miss, or partially satisfied (a new scenario). Hence, (soft) cache hits are no more binary, but may also take non-integer values (?1). Within this framework, the main technical novelty of this project will be a first-of-its-kind joint consideration of wireless edge caching algorithms (subject to such a soft cache hit/miss) and recommendation systems that aim to facilitate wireless edge caching systems, in case of soft cache misses.

Résumé du projet de recherche (Langue 2)
Scientific Content and Methodology

In this project, we propose a joint treatment of caching and recommendation systems for storing content on the edge of future networks. The proposed work consists of 3 main tasks: Task 1 – Recommendation-aided Optimization of Edge Caching: Consider the basic framework of femto-caching [1], where content must be allocated in K small cells (SCs), each equipped with a cache of size M contents, during off-peak hours. There are S total users, and each might request content n from a catalog of N total contents with a probability p_n. Each user i is within range of a number of SCs. The relation between users and SCs is a bipartite graph G = {V,E}, where G(i) = {j: (i,j) ? E} denotes the set of SCs within range of user i. The problem variables are x_nj where x_nj=1, if SC j stores content n, and 0 otherwise. The basic caching placement problem can be summarized as max?(x_nj )????(n=1)^N??(1-x_nj)?) subject to ?_(n=1)^N??x_nj?M,?j ?? ? (1) The term in the parenthesis implies that if content n is requested by node i, and none of the SCs within range of i store that content (i.e. x_nj=0,?j?G(i)), then there is a cache miss. Assume now that if there is a miss, the system can recommend instead to user i an alternative content among the ones that are locally available. The relation between contents could be captured with a N×N utility matrix U={u_ij}, where u_ij denotes the utility a given user gets if she originally asks for content i but instead receives content j, where u_ii=1 and u_ij?[0,1]. This matrix is expected to be sparse, and in practice, it could be inferred from the related content graph (e.g. related video graph on YouTube). A very simple example would be for example to mark the L (e.g. top 10) most related videos for video i, with u_ij=1. The optimal caching problem can be rewritten as max?(x_nj )??_(i=1)^S??(1-?_(j?G(i))??(1-x_nj^(u_nm )))? subject to ?_(n=1)^N??x_nj?M,?j ?? ? (2) Hence, a cache hit for content n is now achieved if n is stored in SC j?G(i) (i.e., if x_nj=1 as before) but also if x_mj=1, i.e. some other content m is available locally, for which u_nm=1 (i.e., content m is related to n, and could equally satisfy the request). Within this task, we will investigate problems of this type, e.g., whether the objective remains submodular and thus amenable to greedy approximation algorithms, as well as continuous relaxations. We will further consider more realistic scenarios where u_ij can take any values in [0,1]. Task 2 – Modeling and Impact of Content Relation Matrix U: The goal of this task would be to analytically understand the properties of related content matrix U and their impact on performance. Example 1: Each content i has on average L related contents, chosen uniformly and independently for each i. Given that L?N, one can conjecture that caching gains from alternative contents, in case of a cache miss, will be rather modest. Example 2: The L elements of each row i for which u_ij=1 are again chosen randomly, but now proportionally to their popularity. A content request resulting in a miss, will now have a higher chance to be satisfied by one of its related contents. However, the exact benefits depend both on the content popularity distribution and how this relates to the u_ij entries. First, we will study real datasets for Internet content and its relations (e.g., publicly available ones for YouTube file relations) to understand both the qualitative and quantitative properties of matrix U. Second, we will study analytically what specific properties of matrix U most facilitate caching algorithms. One promising direction is the Spectral Graph properties of U, which can nicely and compactly characterize macroscopic relations between many contents, which we expect to have an impact on cache gains. Task 3 – Caching-aware Biasing for Recommendation Systems: There exists an abundance of recommendation systems and algorithms, and the topic is a fertile field for ongoing research. The investigation of novel recommendation systems per se is beyond the scope of this project. Instead, our goal in this last task is how one could modify the basic idea(s) of existing recommendation algorithms to facilitate caching gains, with a minimum footprint on the recommendation algorithm’s intended semantics and goals. For example, we plan to investigate how a YouTube-like recommendation algorithm could be within the context of future cellular (e.g. 5G) and edge (e.g. MEC) architectures, so that the contents it shows as recommended don’t only depend on the pure recommendation results, but are “biased” in favor of the caching algorithm. This problem could be formulated as maximizing the objective of the caching problem, by appropriate knobs in the recommendation algorithm, while minimizing the impact on the recommendation objective (e.g. captured as a constraint on user QoE).