A cost optimization model and its heuristic algorithm for a content distribution network

Junfang Guo*

Wenzhou University, Zhejiang, 325035, China

Received 1 October 2014, www.cmnt.lv

Abstract

Content Distribution Network (CDN) is an effective approach to improving the quality of Internet service. In a CDN, it is an important problem how to minimize the connection cost over all input requests. This paper firstly reviews some state-of-the-art research works related to the classical facility location problem. Secondly, based on existing Facility Location Models, the this paper proposes a mathematical programming formulation for optimizing the connection cost in a CDN and then presents a corresponding heuristic algorithm for solving this optimization problem. Finally, the theoretical analysis demonstrates that our proposed algorithm performs better than previous algorithms in terms of effectiveness and efficiency.

Keywords: facility location, content distribution network, approximation algorithm, modeling

1 Introduction

Content Distribution Network (CDN) is an effective approach to improving the quality of Internet service. CDN replicates the content from the origin place to the replica servers scattered over the Internet, and serves a request from a replica server close to which the request originates in. Meanwhile, a CDN is also an interconnected system of computers on the Internet that provides Web content rapidly for numerous users by duplicating the content on multiple servers and then directing the content to users based on proximity. Recently, with the emergence of CDN, CDNs are targeted for speeding up the delivery of normal Web content, and reducing the load on the origin servers and the network. In a CDN, a content provider is allowable to sign up for the service and place its content on the content servers. In a CDN, the on-demand users replicate its content when requesting it, or replicating beforehand, by pushing the content on the content servers. CDNs can be used by Internet service providers to deliver static or dynamic Web pages, but the technology is more suitable for streaming audio, video, and Internet television program.

Generally speaking, the CDN model can be as described follows [1]. Suppose that an autonomous system (AS) has n servers, and each server can serve a certain number of requests per unit time. A request enters the system through ingress Provider Edge (PE) routers. The cost of serving one particular request depends upon the distance between the server and the PE within the AS. As a result, the goal of CDN providers is to optimize the connection cost over all input requests.

The above problem can extend to the classical facility location problem. The facility location problem [2] is one of the fundamental combination optimization problems in the branch of optimization or operations research in mathematics. The problem of locating facilities is not a new direction in the operations research community. It has been mentioned in [2] that the challenge of the facility location problem is that where the best site facilities located has inspired a rich, colorful and ever growing body of literature. To cope with a multitude of applications encountered in the business world and the public sector, some expanding models have emerged [3-5]. Specifically, location-allocation models cover many formulations which range in complexity from simple linear, single-stage, single-product, incapacitated deterministic models to non-linear probabilistic models, wherein, the corresponding algorithms are generally developed based on local search mathematical programming and so on. Thus, first of all, this paper will review some state-of-the-art related works on the facility location problem. The focus is on the fundamental assumptions, mathematical models and specific references to the solutions.

The main contributions of this paper are twofold. First, we propose a mathematical programming formulation for the connection cost optimization problem in a content distribution network. Second, we develop an algorithm for the mathematical programming problem. The remainder of this paper is organized as follows. In section 2, we review the facility location problem. In Section 3, based on the facility location problem, we develop a mathematical programming formulation for the connection cost optimization in a content distribution network. Section 4 provides a heuristic solution algorithm used to solve the mathematical programming formulation. Finally, we conclude our work in Section 5.

---

* *Corresponding author's* e-mail: gjf@wzu.edu.cn

2 Related work

In a CDN, multiple versions of the same content exist on strategically dispersed servers; that is, content replication. A large CDN may have thousands of servers, consequently, making it possible to provide the same content simultaneously for many users efficiently and reliably, even at the time of maximum Internet traffic or during sudden demand "spikes." When a specific page, file, or program is requested by a user, the server closest to the user (i.e., the server between which and the user there is the minimum amount of nodes) is dynamically determined, so as to optimize the speed for delivering the content to the user. The use of CDN technology has obvious economic advantages to the enterprises, which expect or experience large numbers of hits on their Web sites from locations all over the world. For example, if dozens or hundreds of users select the same Web page or content simultaneously, the CDN sends the content to each of the users without any delay or time-out. Consequently minimizing the problems with excessive latency, as well as large variations in latency from moment to moment (which can lead to annoying "jitter" in streaming audio and video) namely maximizing he bandwidth that each user "sees". Another advantage of CDN technology is its content redundancy that provides a fail-safe feature and allows for graceful degradation in the event of damage or malfunction of a part of the Internet. Even during a large-scale attack that disables many servers, the content on a CDN will remain available to some users. Besides, the CDN technology can also enhance the backup, archiving, and storage capacity of data. This can benefit individuals and enterprises that heavily rely on online data backup services. CDN distribution scheme is shown in Figure 1.
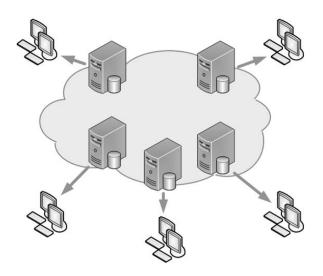

FIGURE 1 CDN distribution scheme

A CDN contains three elements: Request, Distribution and Delivery. The "request" element of a CDN deals with the requests of users or systems for specific content, regardless of a file or a video. When a request occurs at the user terminal, the protocols used in a CDN would intercept and redirect thee request to the hardware components or content engines closest to the user. Once a request is posted by a user, the content engine can decide whether it can answer the request. The "distribution" element of a CDN determines what decision (answer or proxy) is choose. All content has to come from somewhere of the architecture (i.e., origin servers), and based on the pattern of using and requesting, the content can be appropriately distributed by the administrators of a CDN. Thus, the choice of distribution directly affects the details of a request. The "delivery" element is responsible for delivering the content to the correct locations in the architecture. This element heavily relies on the routing and switching infrastructure so as to complete the reliable and efficient delivery.

The first approximate algorithm that has a constant approximation ratio less than 3.16 for the facility location problem, was developed by Shmoys and Tardos in [6], and then through being improved by Lin and Vitter in [7], a randomized filtering and rounding technique was proposed which has a better performance. The approximation algorithm for the Incapacitated Facility Location problem is based on LP rounding so as to solve the relaxation of the integer program for Incapacitated Facility Location. The algorithm of Shmoys, Tardos and Aardal first partitions the cities into clusters and then for each cluster opens exactly one facility so as to supply services for all the cities in the cluster.

The main drawback of this approach is that there is an imbalance between bounding facility and service costs in most cases. Using a different rounding approach, that is the randomized rounding technique, Chudak and Shmoys [8] presented an algorithm with ratio $1+2/e \approx 1.736$, which is a significant improvement on the previously approximation guarantees. Their techniques use the property of optimal dual solution to the linear program and complementary slackness, randomized rounding that can be easily randomized using the method of conditional expectations, as well as a generalization of decomposition techniques. This was further improved to 1.582 by Sviridenko [8] by rounding an optimal fractional solution to a linear programming relaxation. Recently, Byrka [9] used the idea of scaling-up fractional solution and techniques from [10] to obtain another LP-rounding algorithm with currently best known ration of 1.5. Guha and Khuller [11] have shown that the metric incapacitated facility location problem is Max SNP-hard, they have also shown that the existence of a ρ-approximation algorithm for ρ <1.463. The best known ration of 1.5 is very proximal to the result of 1.463. All these algorithms are based on solving linear programming relaxations of the facility location problem and rounding the solution obtained. The drawback of these algorithms based on LP-rounding is that they need to solve large linear programs, and so have prohibitive running times for most applications.

3 Classical facility location models

3.1 INCAPACITATED FACILITY LOCATION MODEL

Incapacitated Facility Location [2] is one of the classical problems that have been widely studied in operations research。 It can be defined as follows. We are given several set as follows: a finite set of potential facilities F ; a finite set of cities or demands C ; a non-negative number f_i as the opening cost of facility i ; a connection cost c_{ij} between facility i and city j for every facility-city pair (i, j) , $i \in F$, $j \in C$. The objective of Incapacitated Facility Location is to open a subset of the facilities in F , and connect each city to open facilities, such that the total cost of opening facilities and connecting cities to open facilities is minimized. We will consider the metric version of this problem, i.e., the connection distances are symmetric and satisfy the triangle inequality, or equivalently satisfy $c_{ij} + c_{i'j} + c_{i'j'} \geq c_{ij'}$ for all $i, i' \in F$ and $j, j' \in C$. This is natural assumption, since connection costs are often associated with the distance between points in Euclidean space representing facilities and cities. Therefore we speak of metric connection costs if the above condition is satisfied. We make this assumption in the following sections. The Incapacitated Facility Location problem can be formulated by the following integer linear program.

$$\text{Minimize} \quad \sum_{i \in F} f_i y_i + \sum_{i \in F} \sum_{j \in C} c_{ij} x_{ij}$$

Subjected to

$$\forall j \in C : \sum_{i \in F} x_{ij} \geq 1 \tag{1}$$

$$\forall i \in F, j \in C : y_i \geq x_{ij} \tag{2}$$

$$\forall i \in F, j \in C : x_{ij} \in \{0,1\} \tag{3}$$

$$\forall i \in F : y_i \in \{0,1\} \tag{4}$$

In the above formulation, y_i is an indicator variable denoting whether facility i is open, and x_{ij} is an indicator variable denoting whether city j is connected to facility i . The first constraint ensures that each city $j \in C$ is connected to at least one facility, whereas the second ensures that cities are connected or served only an by open facility, that is whenever a city j is assigned to facility i , then a facility must have been opened (and paid for).

In addition, an extended version is considered in [10], in which each facility i can be assigned to serve at most u_i cities, that indicates the capacity of this facility, i.e., the number of cities it can serve, where u_i is a positive integer, if facility i is opened y_i times, it can serve at most $u_i y_i$ cities. That is a capacitated version of Incapacitated Facility Location. First of all, there are two variants of the problem, depending on whether each city's demand must be assigned to only one facility, or the demand may be fractionally split among more than one (completely) open facility. We focus on the latter case. We allow the demand of cities to be split, i.e. assigned to multiple open facilities. While the total demand assigned to a facility must not exceed its capacity.

In real application, the k -median problem is considered in [14, 15]. Similar to the facility location problem, there are two respects different from the facility location problem in the following. There is no cost for opening facilities, and there is an upper bound k on the number of facilities, where k is an input parameter that is not fixed. The objective is to select a set of k facilities (or centers) so as to minimize the sum of the assignment costs for the city.

3.2 The metric fault tolerant facility location model

In Section 3.1, we have introduced the Incapacitated Facility Location problem. However, in Incapacitated Facility Location, only one connection is required for each city which is not the case in many application scenarios. Therefore, researchers extend the problem by proposing Fault-Tolerant Facility Location. In this subsection, we will introduce the Fault-Tolerant Facility Location problem [3]. The Fault-Tolerant Facility Location problem is a generalization of the Incapacitated Facility Location problem where each client has independent fault-tolerance requirement on connectivity. In many setting it is essential to provide safeguards against failures by designing fault-tolerant solutions. For example, in a distributed network, we want to place caches and assign data requests to caches so as to be resistant against caches becoming unavailable due to node or link failures. A common solution is to replicate data items across caches and build some resilience in the network. This motivates the Fault-Tolerant Facility Location problem, wherein a city $j \in C$ has a requirement r_j and has to be assigned to r_j distinct open facilities instead of just one. The facilities other than the closest one are 'backup' facilities for the city, and will be used only if the closer facility (or the link to it) fails. In this problem, we are given a set of facilities F and s set of cities C , and each city j has a requirement $r_j \geq 1$. Each facility i has an opening cost f_i , the cost c_{ij} is to establish a connection between facility i and city $j \in C$. The FTFL problem can be formulated by the following integer linear program.

$$\text{Minimize} \quad \sum_{i \in F} f_i y_i + \sum_{i \in F} \sum_{j \in C} c_{ij} x_{ij}$$

Subjected to

$$\forall j \in C : \sum_{i \in F} x_{ij} \geq r_j \tag{5}$$

$$\forall i \in F, j \in C : y_i \geq x_{ij} \qquad (6)$$

$$\forall i \in F, j \in C : x_{ij} \in \{0,1\} \qquad (7)$$

$$\forall i \in F : y_i \in \{0,1\} \qquad (8)$$

In the above formulation, variable y_i indicates whether facility i is open, and x_{ij} indicates whether client j is assigned to facility i. The connection cost or assignment cost incurred for j is the sum of the distances from j to these r_j facilities. The objective of the problem is to open proper number of the facilities and assign each city j to r_j distinct open facilities so as to minimize the total facility opening and client assignment costs. Note that the first constraint ensures the number of connections required by each city and the second constraint reflects that a city can only be assigned to an open facility.

4 A novel facility location model in content distribution networks

4.1 PROBLEM FORMULATION

Content distribution networks problems involve strategic decisions which influence tactical and operational decisions [16-19]. In particular, they involve facility location, transportation [12] and inventory decisions, which affect the cost of the distribution system and the quality of the customer service level. So, they are core problems for each company. With incorporation of CPU computing capacity of each server as new parameter, old model of facility location problem in CDN gets mapped to multi-index quadratic assignment problem. In our new model, we consider CPU computing capacities while finding out optimal connection cost for serving input requests.

In this section, a relaxation of the facility location problem called facility location in content distribution networks problem is studied. Given a set of CDN providers, a set of clients, and corresponding server operating cost at each site as well as connection cost for each site-client pair. The facility location in content distribution networks problem requires to allocate a proper number of servers, and further such that clients can fetch objects from the nearest CDN server that containing the requested object. In this problem, we are given a set S of CDN servers and a set C of n_c clients. At each server $i \in S$, an unlimited number of servers with a non-negative cost s_i as costs can be operated. There is also a connection cost c_{ij} between each client $j \in C$ and all servers at site i: Note that each client requires a pre-specified number of connections for fault tolerance purpose. Let the connectivity requirement of client $j \in C$ be r_j. The objective is to optimally allocate a certain number servers

from each i to serve every client j with r_j requests while minimizing the total combined cost for server operating and service access incurred. The facility location in content distribution networks problem can be formulated by the following integer linear program.

Minimize $\sum_{i \in S} f_i y_i + \sum_{i \in S} \sum_{j \in C} c_{ij} x_{ij}$

Subjected to

$$\forall j \in C : \sum_{i \in S} x_{ij} \geq r_j \qquad (9)$$

$$\forall i \in S, j \in C : y_i \geq x_{ij} \qquad (10)$$

$$\forall i \in S, j \in C : x_{ij} \in \{0,1,...,N\} \qquad (11)$$

$$\forall i \in S : y_i \in \{0,1,...,N\} \qquad (12)$$

In the above formulation, non-negative integer y_i indicates how many servers are opened at i and x_{ij} indicates how many connections are established between i and j. The first constraint is to ensure that each client $j \in C$ has fault tolerance of connections, that is to say, it is assigned totally r_j connections but it can tolerate up to $r_j - 1$ connection failures. The second constraint insures that each client $j \in C$ has fault tolerance of servers that is assigned distant servers for any two connections and is able to tolerate up to $r_j - 1$ server. In this section, we consider the metric version of the problem where the given connection costs satisfy triangle inequality.

The differences between the facility location in Content Distribution Networks problem and two Facility Location problems mentioned in the mentioned above sub-sections are as follows: the facility location in Content Distribution Networks problem is quite close to the well studied Fault-Tolerant Facility Location problem which has the same objective function and constraints as the facility location in Content Distribution Networks problem, except for this condition, the range of variants is different. For x_{ij} and y_i are non-negative integers in the facility location in Content Distribution Networks but binary integers (0 or 1) in Fault-Tolerant Facility Location problem. The Fault-Tolerant Facility Location problem restricts at most one server, without the restriction on the maximum number of servers at each site, the facility location in Content Distribution Networks problem is less constrained and therefore incurs less cost which is desirable in real applications. The facility location in Content Distribution Networks problem can be applied in the deployment of ATMs, server farms applications, of which multiple servers can be deployed if necessary and service the clients.

4.2 A FAST ALGORITHM FOR THE FACILITY LOCATION IN CONTENT DISTRIBUTION NETWORKS

A CDN consists of several individual components, or infrastructures, whose purpose is to provide a service to the user community. An example of an infrastructure component could be consolidated file servers, which act as repositories for user files and application data. These servers sit atop another component of the architecture: the routing and switching infrastructure. The foundation for network architecture lies in the routing and switching infrastructure, which provides transport for all other infrastructure components and their various forms of data.

Some researchers have proposed a heuristic known as "heuristics concentration", and applied this to the p-median problem. Heuristic concentration (in the context of the p-median problem) is based on running an efficient heuristic numerous times as to identify a subset of the potential facility sites that may warrant further investigation. After repeatedly running a randomized greedy heuristic, the information about which nodes are in the "p" set is collected, and a final subset of nodes is used when solving the subset problem to optimality by using integer-linear programming. Thus, using a subset of the potential facility nodes results in an optimal solution if all nodes that are in the optimal "p" set are in the subset. Also, this subset is limited, and the time to find the optimal solution to this sub-problem is only a fraction of that of optimally solving the problem using the full set of potential facility sites. It is our intent to propose a heuristic for solving problem Refurb that takes advantage of the above observations. Then some authors compared heuristic concentration to tabu search and concluded that heuristic concentration is superior to tabu search for the p-median problem. We then propose several replica placement locations for CDN servers that can determine the optimal number of replicas we should select from a given set of potential location such that clients can fetch objects from the nearest CDN server that containing the requested object.

Algorithm 1. A Fast Algorithm for the Facility Location in Content Distribution Networks.
Begin
Step 1. Give the facility location in Content Distribution Networks like a rooted binary tree, with servers and clients at leaves.
Step 2. Initialize DP for clients:
If there is a client with bandwidth demand N and CPU demand C at a leaf v
For a demand, let for any server x, dynamic programming value associated with (v, 0, 0, 0, 0, NONE, x) be N * C * distance between x and v; all other values corresponding to this leaf are set to infinity;
If server i is located at this leaf,
Let (v, 0, 0, 0, 0, NONE, NONE) gets value 0;
Put value infinity for any other DP cell corresponding to this sub tree;
Step 3. Update the dynamic programming memoriza-

tion table values bottom-up. If the node u has two children u_1 and u_2, a cell X =(u, F-N, F-C, N, C, x, y) is updated (if this gives a better value) from X1=(u1, F-N1, F-C1, N1, C1, x1, y1) and X2=(u2, F-N2, F-C2, N2, C2, x2, y2) to cost(X1)+cost(X2), if C is consistent with X1 and X2.
Step 4. Build the answer recursively from DP(r, 0, 0, 0, 0, NONE, NONE), with r being the root of the tree.
End

4.3 THE COST OF FACILITY LOCATION IN CONTENT DISTRIBUTION NETWORKS

For a given server location, we can calculate c_{ij}, the distance for client j to its nearest replica, and c_i, the distance from replica i to the origin server. To make a theoretical analysis of the total cost, we make an assumption about the relationship between the average distance from clients to origin servers and the average distance from clients to their nearest replicas. Based on it, we derive a model for the total cost of replica placement changing as the number of replicas increase.

Assume that the average distance from clients to origin servers (over all clients, origin server pairs) is r. As we put more replicas, the average distance from clients to their nearest replicas will decrease. It can be modeled as a function of r and the number of replicas k, denoted as $f(r,k)$. We make a hypothesis about the relationship by

defining $f(r,k) = \dfrac{r-1}{\beta^{k-1}} + 1$, where β is a decay factor

and $\beta > 1$. For a network with a few hundreds of nodes, β will be large, because with the increasing of the number of replicas, the distance from a client to its nearest replica will be reduced greatly. For a network of the Internet size, β tends to be very close to 1, but it is still an exponential decrease when we increase the number of replicas.

Based on the above model, we analyze the total cost of facility location in content distribution networks when the number of replicas changes. We illustrate the relationship between three distances: c_{ij} the distance from a client to its nearest replica, c_i the distance from the replica to the origin server, and co_j the distance from the client to the origin server. We know co_j can be modeled by the average distance r and c_{ij} can be modeled by the function $f(r,k)$. The other term used in calculating the total cost is $c_{ij} + c_i$. Usually replicas are located at access points of clients to the network and therefore we have $c_i \leq co_j$. Combining it with the triangle relation, we have $co_j \leq c_{ij} + c_i \leq co_j + c_{ij}$. As we put more replicas, the replica is more likely to be located on the path from the

origin server to the client and therefore $c_{ij} + c_i \approx co_j$. With this simplification we can use r to model the cost $c_{ij} + c_i$.

5 Conclusions

In this paper, the novel facility location model for Content Distribution Networks (CDN) is proposed. We formulate this model as a combinatorial optimization problem. As each server in a CDN is treated as a node with finite storage capacity, the goal of this optimization problem is to minimize the total CDN server operating cost as well as

the connection cost for each client-server pair such that clients fetch objects from the nearest CDN server containing the requested object, Then, a corresponding heuristic algorithm for this optimization problem is given. In addition, by theoretical analysis, we also studied the total cost of this model in a CDN.

References

[1] Jin S, Bestavros A 2000 Greedy dual web caching algorithm: exploiting the two sources of temporal locality in web request streams *Proceeding of the 5th Web Caching and Content Distribution Workshop* .22-4
[2] Cornuejols G, Nemhauser G L, Wolsey L A 1990 The incapacitated facility location problem. Discrete Location Theory 119-71
[3] Srinivasan B A, Swamy C 2010 Fault-tolerant facility location: a randomized dependent LP-rounding algorithm *Proceeding of the 14th conference on Integer Programming and Combinatorial Optimization* 244-57
[4] Swamy C, Shmoys D 2008 Fault-tolerant facility location *ACM Transaction on Algorithms* **4**(4) 1-27
[5] Guha S, Kamesh Munagala A M 2003 A constant factor approximation algorithm for the fault-tolerant facility location problem *Journal of Algorithms* **48**(2) 429-40
[6] Shmoys D B, Tardos E, Aardal K 1997 Approximation algorithms for facility location problems *Proceedings of the 29th Annual ACM Symposium on Theory of Computing* 265-74
[7] Lin J H, Vitter J S 2006 ε-approximation with minimum packing constraint violation *Proceedings of the 34th ACM STOC* 771-82
[8] Mahdian M, Ye Y, Zhang J 2006 Approximation algorithms for metric facility location problems *SIAM Journal on Computing* **36**(11) 411-32
[9] Sviridenko M 2009 An improved approximation algorithm for the metric incapacitated facility location problems *Proceedings of 16th International Conference on Integer Programming and Combinatorial Optimization* 240-57

[10] Byrka J, Aardal K I 2010 An optimal bi-factor approximation algorithm for the metric incapacitated facility location problem *SIAM Journal on Computing* **39**(10) 2212-31
[11] Guha S, Khuller S 1998 Greedy strikes back: Improved facility location algorithms *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms* 127-32
[12] Crainic T G, Laporte G 1997 Planning models for freight transportation *European Journal of Operational Research* .**97**(5) 409-38
[13] Xu D, Zhang S 2008 Approximation algorithm for facility location with service installation costs *Operations Research Letters* **36**(1) 46-50
[14] Thorup M 2005 Quick k-Median, k-Center, and Facility Location for Sparse Graphs *SIAM Journal on Computing* **34**(2) 405-32
[15] Fotakis D 2006 Incremental Algorithms for Facility Location and k-Median *Theoretical Computer Science* **361**(2-3) 275-313
[16] Qiu L, Padmanabhan V N, Voelker G M 1999 On the placement of web server replicas *Proceedings of IEEE Infocom* 21-5
[17] Cidon I, Kutten S, Soffer R 2001 Optimal allocation of electronic content *Proceedings of IEEE Infocom* 322-6
[18] Chen H, Luo X, Li Y, Fu H, Zheng X, Yuan X 2012 Ill-conditioned Load Flow Study Based on Asynchronous Parallel Computing in Rural Power Distribution Networks *JCIT: Journal of Convergence Information Technology* **7**(19) 261-8
[19] Yu C 2011 A TVC based Clock Fault Detection Circuit for Modern Clock Distribution Network *AISS: Advances in Information Sciences and Service Sciences* **3**(3) 75-80

Author

Guo Junfang, 30th July, 1972, Wenzhou Zhejiang, China.

Current position, grades: senior experimental division in Wenzhou University Oujiang college.
University studies: Wuhan University.
Scientific interest: computer science and application.
Publications: 10 papers.