Homework 1

Lian Di 1801212881

1. Big data problem

The Internet has become an indispensable part of people's life. Internet applications must be able to run in a complex network environment. However, the network distance of different regions is far away, and the geographical location is scattered, which brings great challenges for Internet companies to provide high-quality services. As a kind of Internet service, video service has more strict requirements for transmission speed. How to make the network users in different regions get the same high-quality service experience and reduce the cost to the greatest extent has become an important research direction of major Internet companies.

At present, in order to improve the efficiency of large-scale distribution, two dominant technologies are CDN (content delivery network) and P2P (peer-to-peer) networks. Among them, the Internet mainly uses CDN network technology. In CDN network, a large number of servers are deployed in the "edge" of the Internet. These edge servers are used to store the content of the source server strategically. The resource request sent by the user is directed by the CDN routing system to the edge server closest to the user. Therefore, data transmission delay is significantly reduced, users can get better experience; at the same time, it can reduce the load of the source server, reduce network congestion, and improve the availability of the system. However, CDN network also has the following defects: (1) it is limited by the processing ability of the client / server; (2) the cost of deployment and maintenance is high; (3) it is still a client / server structure in essence, with the increase of user books requesting services, the performance of the edge server will decline significantly. Therefore, how to better arrange the location of the content distributor, reduce the delay from the server to the client, reduce the transmission cost of traffic and the cost of server deployment and maintenance is the premise of promoting and using CDN network technology.

The main problem of content distribution network is how to select some nodes as content distributor nodes when a set of consumers and a set of network nodes are given, and different nodes are connected with each other, with different bandwidth size and cost of bandwidth use, so that all consumers can meet their own requirements of content transmission speed, and at the same time, make the total The cost is minimal.

2. Big data properties

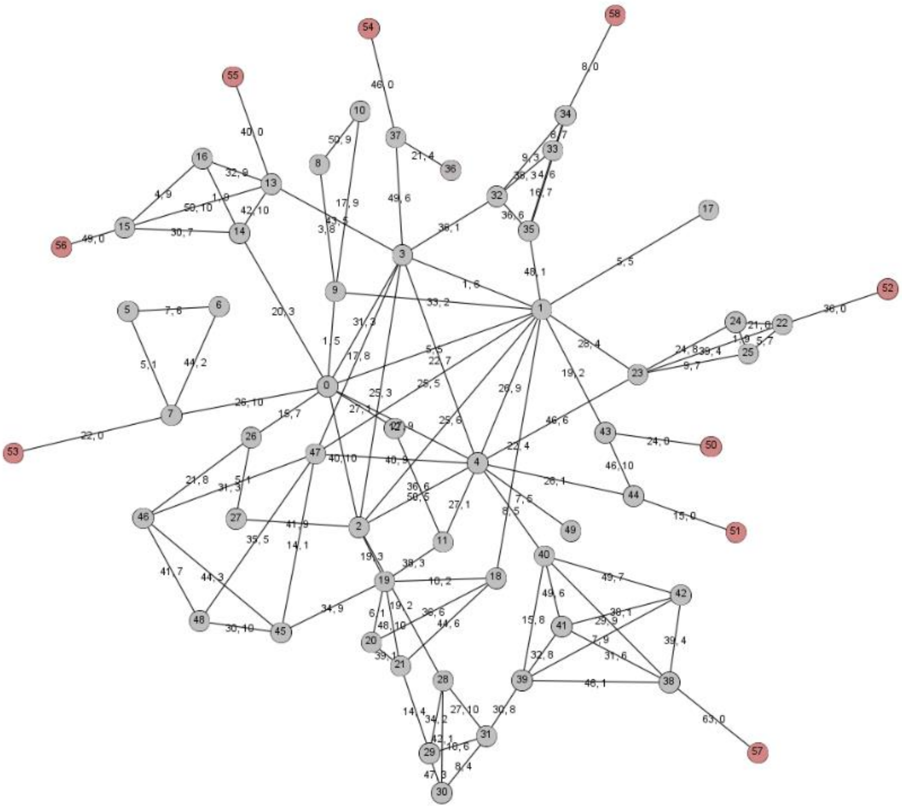
Given a undirected graph of network structure composed of several network nodes (such as router and switch), each node is connected with at least another node through network link (network link refers to the network path directly connected between two network nodes, and there is no other network node in the middle, which is equivalent to one edge in the undirected graph). One node can communicate the received data Over the network link to another connected node. The total network bandwidth of each link is different (for example, the total bandwidth of a link is 10Gbps). The video transmission carried by each link needs to charge the corresponding network rental fee according to the amount of bandwidth occupied. The unit rental fee of each link is different (for example, the rental fee of a link is 1000 yuan / Gbps, i.e. 1K / Gbps). The total bandwidth occupied on a link must not exceed the total bandwidth of the link.

Consumer node: in a given network structure, there are some network nodes directly connected to the network of residential area. Each residential area network is presented as a consumer node in the given network structure diagram, and the video bandwidth consumption needs of different consumer nodes are different.

Video content server: the video content server stores the video content (such as movie, TV series, etc.), the video data flow of the video content server can flow to the consumer node through the network path composed of network nodes and links, the output capacity of the video content server has no upper limit, it can serve multiple consumer nodes, a consumer node can also simultaneously from multiple TV stations The video content server obtains the video stream. The cost of deploying a video content server (for example, 300000 yuan / set, 300K / set) is the same for all servers.

Example: As shown in the figure below, nodes numbered 0-49 (white node) in the network are network nodes, nodes numbered 50-58 (red node), that is, terminal nodes are consumer nodes, each edge is labeled with an edge attribute, the first number represents traffic capacity, and the second number represents the cost of renting 1Gbps bandwidth of the link.

This simple example has hundreds of links and millions of potential solutions.

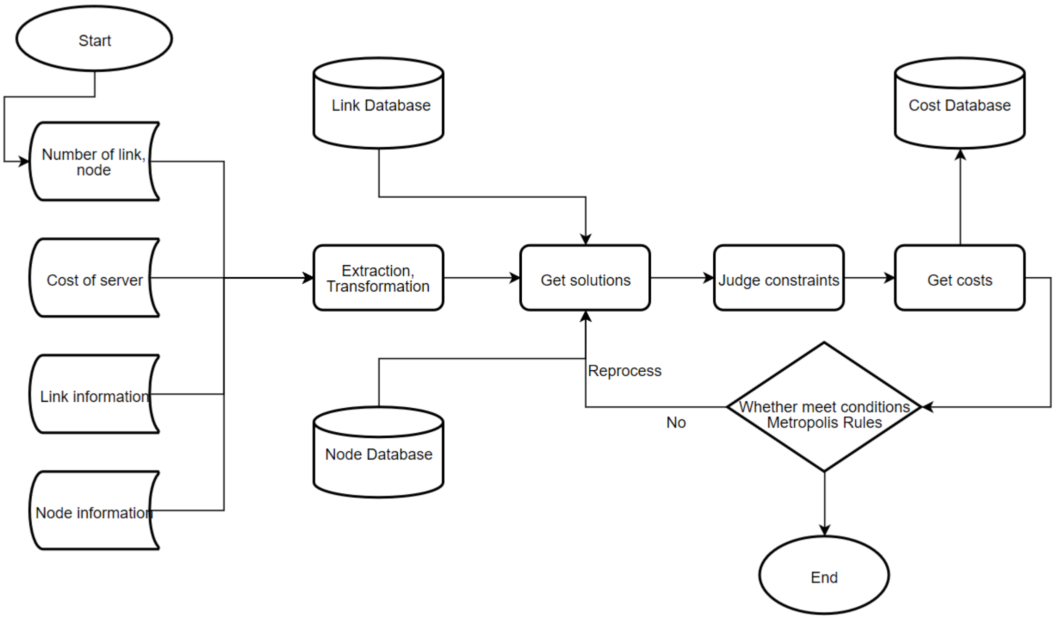


3. Workflow to solve the problem

In order to combine the network topology with heuristic algorithm, we should consider the condition of given server selection point, judge whether it can meet the needs of all consumers, and the condition of given server selection point, and choose the least server as far as possible, the algorithm with the lowest link cost is an important problem to be solved in the content distribution network design model.

A key problem to solve the content distribution network is the selection of server nodes. When the number of network nodes is small, because the number of nodes to be selected is not large, we can use exhaustive method to traverse all the node selection schemes, and select the optimal scheme with the lowest cost to meet the conditions in all the schemes. However, when the number of nodes is large and the network topology becomes complex, the exhaustive method often needs less realistic time to complete the calculation, and cannot be full To meet the needs of current network design, it is necessary to use the combination of heuristic search and network flow calculation.

Another key problem to solve the content distribution network deployment is to find the minimum cost flow to meet the current user needs after the selection of server nodes. The common algorithms to solve this problem are the negative cost loop elimination algorithm and the minimum cost path algorithm.



4. Database

For each link, link start node information, link end node information, total bandwidth size and network rental fee need to be stored

For consumer nodes, consumer node information, connected network node information and video bandwidth consumption demand information need to be stored

For the solutions, all the network nodes in the solution and the size of the occupied bandwidth need to be stored

These are obviously graph data, we should use **graph database** which is a database that uses graph structures for semantic queries with nodes, edges and properties to represent and store data.