A Decentralized PageRank Based
Content Dissemination Model in the Edge of Network

Xin Zhang

UCAS

April 11, 2019

Outline

Outline

1. Introduction

In order to improve the quality of content services, content is usually placed at the edge of the network which is close to users, such as CDN. In recent years, with the enhancement of edge device capabilities, it is hoped that more and more content will be placed on the wider edge including user devices and operator devices, such as set-top boxes, APs. Nevertheless, the traditional distribution methods are mostly based on centralized management. As the scale of edge devices growing larger, it will have a negative impact on the dissemination efficiency. Therefore, how to disseminate content in the edge of the network in a decentralized manner is an important issue.

Outline

2. Related Work

- **Diffusion Models**
  - Linear Threshold (LT) Model
  - Independent Cascade (IC) Model

- **Existing Content Distribution Solutions**
  - CDN
  - Peer-to-peer
  - Hybrid-P2P

2. Related Work
Diffusion Models

- Diffusion models were originally used in social networks to model the spread of influence in a network.
- The Susceptible-Infected-Recovered (SIR) model is the most classic information diffusion model.
- In these models each node is either active or inactive. Over iterations an inactive node becomes active as more of its neighbors become active. The two most popular diffusion models are the linear threshold model and the independent cascade model.
- The two most popular diffusion models are:
  - Linear threshold (LT) model
  - Independent cascade (IC) model

2. Related Work

Diffusion Models

LT Model Each edge has an activation probability and the influence is propagated by activated nodes independently activating their inactive neighbors based on the edge activation probabilities.

IC Model Each edge has a weight, each vertex has threshold chosen uniformly at random, and a vertex becomes activated if the weighted sum of its active neighbors exceeds its threshold.

Neither of the above two solutions can be used

- Application scenarios of the above two diffusion models are different from the self-diffusion scenario of the home router cluster of the community network in this paper.
  - It is hoped that the content will be distributed evenly in the network.
  - However, in the above two models, only after enough neighbor nodes have cached the content, the current node is willing to cache the content. This can lead to uneven distribution of the content.

2. Related Work

Existing Content Distribution Solutions - CDN

CDN

A CDN is a geographically distributed network of proxy servers and their data centers. Three mechanisms are as follows:

Pull-based
: Dispatching user requests to appropriate edge nodes. If the current node does not have the content required by the user, the layer-by-layer query is performed upwards.

Push-based
: Replicating objects on certain servers in advance, namely deciding how to pre-deploy the content based on the analysis on the user behavior.

Hybrid
: First half of which is a pre-deployment via a push-based mechanism, and the latter half is based on a pull-based mechanism. In fact, YouTube, as the largest online video-sharing service, keeps employing CDNs to deliver its most popular videos to its users.

2. Related Work

Existing Content Distribution Solutions - Peer-to-peer

Peer-to-peer

Peer-to-peer (P2P) computing or networking is a distributed application architecture that partitions tasks or workloads between peers, without the need for central coordination by servers or stable hosts.P2P media streaming aims at utilizing the uploading bandwidth of the clients. Two mainstream architectures have been devised:

Tree-based push The tree-based system acts as a natural extension of CDNs. Instead of only having two layers of the client-server structure, it has many such layers by allowing every client to become a potential server to some other clients.

Mesh-based pull Unlike the tree-based push system, the mesh-based system requires peers to share the information about their media repository, which guides a peer to pull its desired media chunk from others.

2. Related Work

Existing Content Distribution Solutions - Hybrid-p2p

Hybrid-p2p

Hybrid P2P systems are a combination of peer-to-peer and client-server models.

- Servers in CDNs are organized in a tree structure, which consists of, from top to bottom, the source server owned by content providers, core service nodes, and edge ones. Each edge service node serves end users that are usually in the same ISP.

- Clients form a tree-mesh combined overlay to forward the received streams.

2. Related Work
What's New in Our Model

Different from the existing methods, a **self-diffusion** process is added into the home router cluster in the edge of network to better assist the video distribution and offline downloading.

- The application scenario of this paper is how to make edge devices such as home routers, instead of operator servers in the traditional sense, work together to pre-deploy the content to enhance the content service performance.

- In our scenario, the framework is similar to the existing CDN-P2P model.

Outline

3. Content Dissemination Model

- **Statement**
- **Decentralized Content Dissemination Algorithm (DCDM)**
- **Decentralized Weighted PageRank-style Node Selection Algorithm**
    - PageRank
    - Weighted PageRank
    - Parameters Approximation

3. Content Dissemination Model

Statement

The application scenario in this paper is the community network, where each node is embodied as a home router. Every node is aware of the following information:

$$< nodeId, peerList, fileState, curCov, needCov, bandwidth >$$

- *nodeId* is the ID of the node.
- *fileState* records the file status on the current node.
- *curCov* is the current coverage rate of the specific file.
- *needCov* represents the expected coverage rate.
- *bandwidth* represents the rate of data transfer.

3. Content Dissemination Model

Decentralized Content Dissemination Algorithm (DCDM)

Each node has at least a certain number of neighbors with good link connectivity. Five steps:
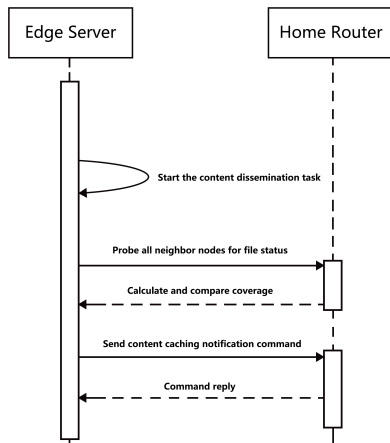
1. Once the content dissemination task starts, the edge server probes all neighbor nodes for file status.

2. then calculates the current coverage and compares it with the requirement.

3. If not reached, the node selection algorithm is used to rank neighbors and select the appropriate ones with a certain coverage rate.

3. Content Dissemination Model
Decentralized Content Dissemination Algorithm (DCDM)
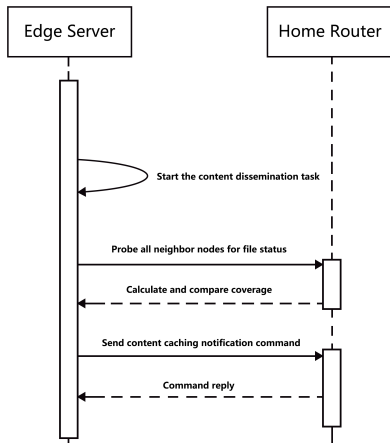
4. A content caching notification command will be sent to neighbors selected by the node selection algorithm.

5. When the ratio of neighbor nodes owning content reaches or exceeds the default coverage requirement, or there are no more quotas, the current node has already completed the dissemination step.

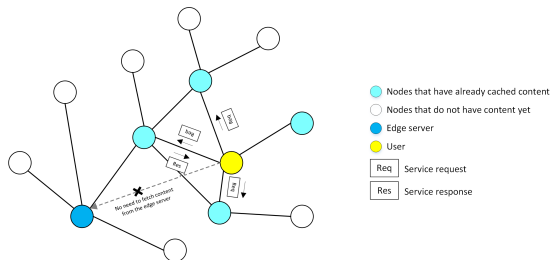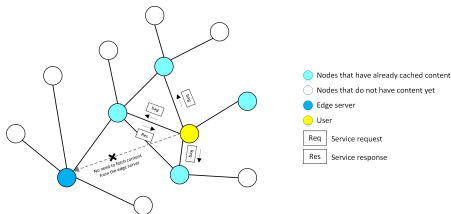# 3. Content Dissemination Model

Decentralized Content Dissemination Algorithm (DCDM)

Figure: Network Diagram

3. Content Dissemination Model
Decentralized Content Dissemination Algorithm (DCDM)

- Each node has a certain number of neighbors, some of which have already cached content.
- Twenty percent of nodes are randomly selected as users.
- User requests obey Poisson distribution.
- When there is a user request, the user can directly obtain data from neighbor nodes without obtaining from a remote edge server.

3. Content Dissemination Model
Decentralized Weighted PageRank-style Node Selection Algorithm

Because the node selection algorithm is the most important part of the entire model, a decentralized weighted PageRank-style node selection algorithm is proposed, which is based on the following thoughts. The content should be cached on nodes as follows:

- Firstly, nodes with a central position.
- Secondly, nodes with high bandwidth and low-latency network connections, which are able to support QoS requirements.
- Thirdly, nodes that are close to users, from the perspective of hops, etc.

3. Content Dissemination Model

Decentralized Weighted PageRank-style Node Selection Algorithm - PageRank

PageRank is a link algorithm assessing the importance of web pages. The fundamental idea is that the importance of a node is judged by the number of nodes linking to it as well as their importance.

$$PR(P_i) = \frac{1-\alpha}{N} + \alpha \sum_{P_j \in B_i} \frac{PR(P_j)}{l_j}$$

where B_i denotes a collection of nodes pointing to P_i, l_j denotes the number of outgoing links of P_j, N denotes the total number of nodes, and α denotes the damping factor, with $0 < \alpha < 1$ and usually set to a value around 0.85. With probability , a surfer chooses the next node at random.

3. Content Dissemination Model

Decentralized Weighted PageRank-style Node Selection Algorithm - Weighted PageRank

In a directed weighted complex network with n nodes, for node v, links to node v are denoted by $v_1, v_2, \ldots, v_l$, the corresponding weights are denoted by $\omega(v_i, v)$ where $1 \leq i \leq l$. Then the calculation formula of weighted PageRank is derived as follows:

$$
\begin{aligned}
PR(v) &= \frac{1 - \alpha}{n} + \alpha \left(\frac{\omega(v_1, v)}{S_{out}(v_1)} \right) PR(v_1) + \cdots + \frac{\omega(v_l, v)}{S_{out}(v_l)} PR(v_l) \\
&= \frac{1 - \alpha}{n} + \alpha \sum_{i=1}^{l} \frac{\omega(v_i, v)}{\sum_{j=1}^{m_i} \omega(v_i, z_j)} PR(v_i)
\end{aligned}
\tag{1}
$$

New Weighting Scheme

This new weighting scheme does not only evaluate the topological position of a node, but also its network and hardware resources.

3. Content Dissemination Model

Decentralized Weighted PageRank-style Node Selection Algorithm - Parameters Approximation
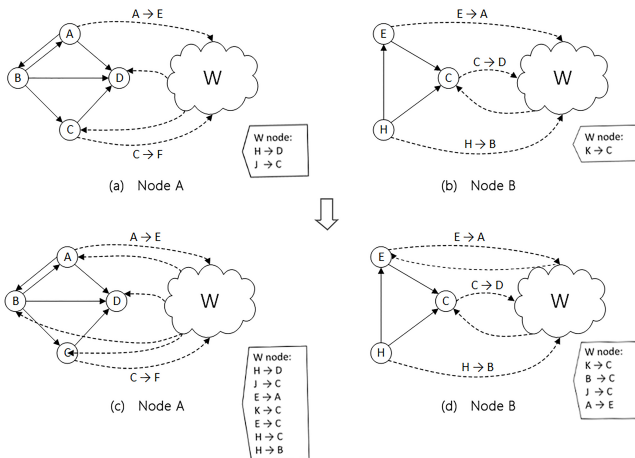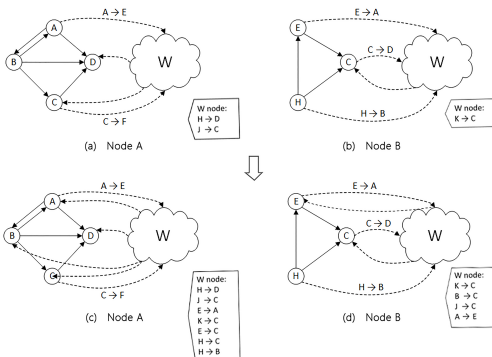

Figure: Demonstration of the light-weight merging phase
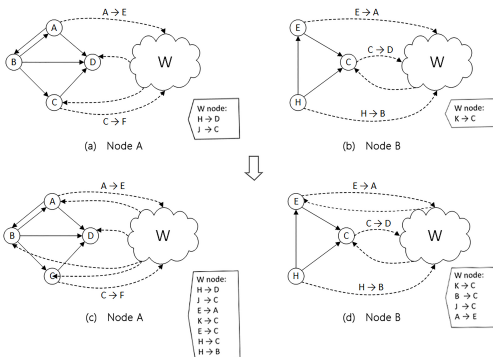
3. Content Dissemination Model
Decentralized Weighted PageRank-style Node Selection Algorithm - Parameters
Approximation

(a) Node A

(b) Node B

(c) Node A

(d) Node B

- Starting with the local graph G of a node, the node first extends G by adding a special node W, called world node, which represents all nodes in the network that do not belong to G.

- Since it is not sufficient to estimate the global score by only knowing the local information, nodes improve their knowledge by meeting other nodes and exchanging information, namely the extended local graph and the score list.

3. Content Dissemination Model
Decentralized Weighted PageRank-style Node Selection Algorithm - Parameters Approximation

(a) Node A

(b) Node B

(c) Node A

(d) Node B

- The graphs in figure (a) (b) represent the original local graphs of nodes A and B before their encounter. The graphs in figure (c) (d) represent the updated ones after the encounter.

- A dashed arrow indicates that either start point or end point of the link is world node. Conversely, the meaning of a solid arrow is that both ends of the link are local nodes.

3. Content Dissemination Model

Decentralized Weighted PageRank-style Node Selection Algorithm - Parameters Approximation
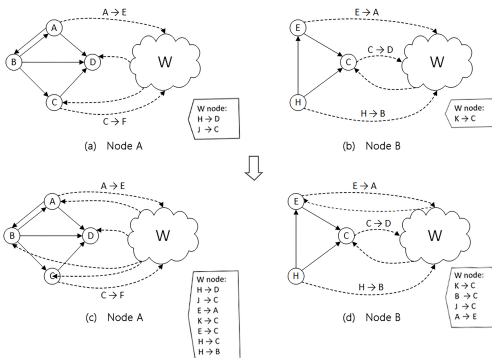
- At a meeting, instead of merging the graphs and world nodes, only relevant information received from the other node into the local world node will be added.

- The only changes are due to insertion of links from the world node to local nodes, whereas links from local nodes to the world node are invariant during all iterations.

3. Content Dissemination Model

Decentralized Weighted PageRank-style Node Selection Algorithm - Parameters Approximation

(a) Node A

(b) Node B
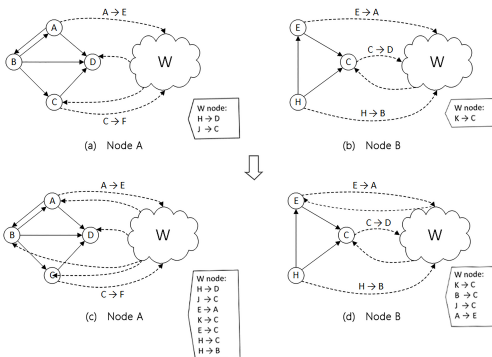
(c) Node A

(d) Node B

- Firstly, updating node A. Obviously, we need to find links in node B that points to the vertex in the vertex set of node A:

$$C \rightarrow D, E \rightarrow A, K \rightarrow C,$$
$$E \rightarrow C, H \rightarrow C, H \rightarrow B$$

- Since the link from C to D already exists in node A, it can be removed from the set. Then these links are added to the local graph of node A, as shown in figure (c).

3. Content Dissemination Model

Decentralized Weighted PageRank-style Node Selection Algorithm - Parameters
Approximation

(a) Node A

(b) Node B
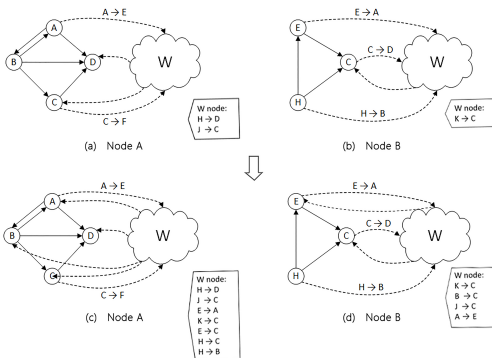
(c) Node A

(d) Node B

- Secondly, updating node B. Similarly, we need to find links in node A that points to the vertex in the vertex set of node B:

$$B \to C, J \to C, A \to E$$

Then these links are added to the local graph of node B. The updated node B is shown in figure 3(d).

Outline

4. Evaluation

- **Experimental Setup**
- **Baseline Algorithms for Comparison**
    - Maximum degree method
    - EasyRank method
- **Metrics**
    - The number of nodes that meet the required coverage requirement
    - The distribution makespan
    - The mean of coverage
    - The average latency
- **Result Analysis**
- **Discussion**

4. Evaluation

Experimental Setup

Symbol	Hyper-parameters	Value
Q	The number of quotas	Ten percent of the total number of nodes
F	File size	1MB
C	Coverage requirement	0.1
B	Bandwidth allocation	640kbps: 40%, 1Mbps: 40%, 2Mbps: 10%, 4Mbps: 10%
m	The number of interactions	Ten percent of the total number of nodes
E	The number of edge servers	5

Table: Simulation Hyper-parameters

4. Evaluation

Baseline Algorithms for Comparison

In our decentralized PageRank-based content dissemination model, the node selection algorithm is the core of the model. Therefore, different node selection algorithms are compared here.

- **Maximum degree method**
  - Generally, nodes with more connections are considered more important.
  - However, to some extent, the degree can only reflect the direct neighbor connections. It is inadequate for discovering the overall topology, such as how neighbors are connected with others in the network.

- **EasyRank method**
  - The EasyRank (*Khan, M. A., Debnath, H., & Borcea, C., 2016*) centrality is a function of the number of friends of its 1-hop neighbors.
  - It assigns relatively higher scores to peers whose neighbors have few friends in order to reduce the skewness in storage availability and improve replication fairness/success.

4. Evaluation

Metrics

The number of nodes that meet the required coverage requirement

As the name suggests, this is a quantitative measurement.

The distribution makespan

It refers to the time spent by the entire model to complete the content dissemination process.

The mean of coverage

It means the average coverage of all nodes when the dissemination process finishes.

The average latency

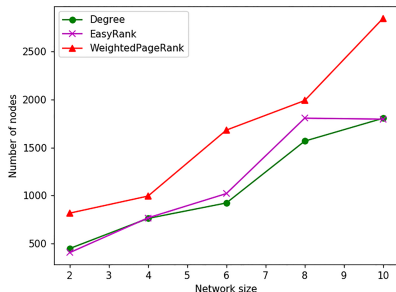It means the average latency of users accessing content.

4. Evaluation

Result Analysis

Figure: The percentage of nodes that meet the required coverage requirement

Since the decentralized PageRank approximation introduces a certain degree of randomness, the ranking of neighbor nodes may be slightly different from that of the global PageRank. Therefore, not only nodes selected by the other two methods, but also the remaining nodes, are likely to be selected. While in degree and EasyRank method, replicas are just placed on nodes with more neighbors, which causes uneven distribution of the content.

4. Evaluation

Result Analysis
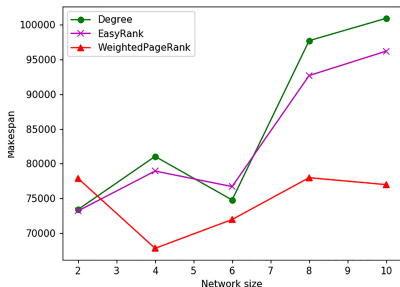
Figure: The dissemination makespan

It can be seen that when the network size is 2000, the makespan of weighted PageRank is slightly higher than the other two methods, but as the network size increases, the makespan of weighted PageRank becomes significantly less than the maximum degree and the EasyRank method. This is because the weighted PageRank method incorporates the bandwidth as a significant part into the link weights between node links, while both the maximum degree and the EasyRank methods are not optimized for the bandwidth.

4. Evaluation

Result Analysis
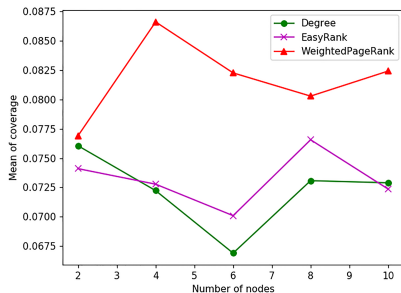
Figure: The mean of coverage

Figure on the left shows a comparison of the mean of coverage under the condition of a fixed number of quotas. As the number of quotas is limited, the higher the mean of coverage rate, the better. The reason is that when some of these neighbors fail to provide services normally, nodes with higher average coverage rate have higher fault tolerance. In figure 6, the weighted PageRank approach achieves the best performance regarding the mean of coverage, as there are more nodes reaching the coverage rate requirement.

4. Evaluation

Result Analysis
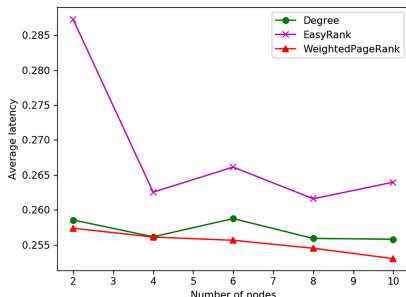
Figure: The average latency

From the end-uses point of view, the service latency of the weighted PageRank method is the smallest compared to the other two methods. There are two main reasons for this. Firstly, the introduction of the bandwidth weights makes the selected node more bandwidth-intensive. Secondly, the weighted PageRank method has higher average coverage rate, so the service fault tolerance is better.

4. Evaluation

Discussion

Generally, there are tradeoffs between different measurements. The main goal is to enhance the service performance, alternatively, increase the number of nodes that meet the coverage requirements, while keeping other measurements improved. 20% of nodes are randomly selected as users, whose requests follow the Poisson distribution. Through the simulation, the service rejection rate decreased by an average of 5.2% in the case of high concurrent requests. Similar to the discussion in the mean of average section, nodes with higher average coverage rate have higher fault tolerance. Therefore, the service rejection rate will be reduced.

Outline

5. Future Work

Future work will mainly focus on two aspects:

- **Edge network churn** brings about changes in network topology and content distribution.
- How to utilize the **edge intelligence** to enhance multiple services, for instance, content distribution, content-based computing, high-performance processing, is also an interesting subject we will focus on.

Outline

6. Conclusion

- This paper proposes a decentralized PageRank-based content dissemination model in the edge of network. Our model solves the problem of how to effectively spread the content in the edge of network.

- The core of this model is a PageRank-based node selection algorithm in a decentralized manner to quantify the service capability of nodes for efficient caching.

- Through the simulation, he service rejection rate decreased by an average of **5.2%** in the case of high concurrent requests.

Bonus

Figure: IJWSR