Priority Control Based on Website Categories in Edge Computing

Noriaki Kamiyama1,2 Yuusuke Nakano1,2 Go Hasegawa2 Hideo Miyahara2
1NTT Network Technology Laboratories 2Osaka University

2016.4.11

Copyright©2016 NTT corp. All Rights Reserved.

Increase of Web Response Time

- Longer than 5 seconds in 50% webpages, and longer than 10 seconds in 10% webpages
- Amazon increased revenue 1% for every 0.1 second reduction in web response time.

Need to reducing web response time

*Web response time: waiting time after clicking hyperlink until entire part of webpage is shown

Complementary cumulative distribution (CCD) of web response time of most popular 1,000 websites when accessing from Tokyo, Japan, in June 2015

CDN: Platform Delivering Web Objects

- 74% of 1,000 most popular websites use CDN*, and CDN is most common technique for reducing HTTP response time.

Reducing HTTP response time by delivering objects from cache server close to user terminal

Obtain objects from origin server and store them

CDN provider, e.g., Akamai, provides many cache servers at network edge.

Edge Computing

- Many objects are dynamically generated in modern webpages.
- Edge computing* is effective to deliver dynamic objects efficiently.

*J. Ott, et al., Content Delivery and the Natural Evolution of DNS, ACM IMC 2012

Complexity of Web Traffic Pattern

One website consists of multiple objects which are delivered from various object servers using HTTP sessions.

Effect of Edge Computing

Geographical deployment pattern may differ among website categories, e.g., Sports and News, and effect of edge computing will depend on website categories.

Yahoo Answers, McAfee SiteAdvisor, ... Society

Yelp, Groupon, ... Home

Identical content from North America

Unique content at each region

Contribution of this work

1. Propose to differentiate caching priority among website categories
2. Roughly analyze effect of category-based priority control in edge computing using active measurement data from 12 locations in world
**Measurement Procedure**

1. Selected 12 PlanetLab hosts as measurement terminals accessing various websites.
2. Measured various properties, e.g., object count obtained and RTT, by executing program at each PlanetLab host to access various websites sequentially.
3. Collected measurement results at collector terminal.

**URL List of Measurement Target**

- Selected 300 most popular websites in each of 16 categories based on public information of Alexa.
- Totally Selected 927 websites from which measurement data were successfully obtained at all 12 measurement locations.

**Classifying Objects Based on CDN Use**

- Classified objects into CDN objects delivered using CDN or non-CDN objects delivered without using CDN.
- Identified CDN objects by comparing second-level domain obtained by dig command with entries of generated list.

**Clustering Analysis of Webpages based on RTT**

- Geographical pattern of original objects, i.e., non-CDN objects, and CDN caches delivering CDN objects will differ among access locations even when accessing same website.
- Analyzed geographical tendencies by clustering websites based on average RTT at 12 access locations.

**Geographical Distribution of Original Objects**

- Cluster 1: RTT was small only in North America. Geographical locality is weak, and identical content are viewed from various regions.
- Cluster 3: RTT was small in all areas except Africa. Geographical locality is strong, and unique content are viewed in each region.

**Platform Measuring Deployment of Web Objects**

- Sets caching priority of each web category at edge servers by continuously measuring geographical deployment of web objects from edge servers in world.
Priority Control among Categories in Edge Computing

Each edge server autonomously makes caching judgement based on web categories according to policy set by controller.

- Controller obtains URL list of website categories using Alexa Web Information Service API® and sends it to edge servers.
- When cache miss, edge server obtains application code of dynamic object from origin server and judges whether to store code based on URL list.

Roughly Estimating Web Response Time (1)

To investigate potential of differentiating caching policy among web categories in edge computing, roughly evaluate reduction effect of web response time

- Number of parallel sessions established with one web server is limited below P
  - P = 2 (suggested in the HTTP/1.1 specification)
  - P = 6 (Explore 8, Firefox 3)

Roughly Estimating Web Response Time (2)

Assumption

- Starts obtaining objects on all TCP connections with all servers
- Fairly obtains objects over all TCP connections with each server
- Continuously receives objects on each TCP connection
- Obtains each object from edge server with probability H with zero RTT

Flow sequence on TCP connection

Applying measured value

Average Reduction in Response Time of Four Categories

Confirm difference of E(D) between Universal websites (adult, society) and Localized websites (home, shopping)

Effect of Web Category Differentiation in Edge Computing

E(G), average reduction ratio of web response time:

Without priority differentiation: delivering 50% of objects of each category
Prioritizing universal group: delivering all objects of Adult and Society webpages
Prioritizing localized group: delivering all objects of Home and Shopping webpages

Conclusion

- Actively measured RTT and object count of most popular 1,000 webpages from 12 locations in world using PlanetLab
  - Confirmed difference of geographical tendencies of object deployment among website categories
  - Universal websites: Adult and Society
  - Localized websites: Home and Shopping
- Proposed to differentiate caching priority among web categories in edge computing
  - Roughly estimated reduction effect of web response time by edge computing
  - Numerical confirmed effect of differentiating caching priority among web categories in edge computing
**Measurement Program**

- Generated URL list and sent it to each PlanetLab host.
- Starting from 0:00 (midnight) or 12:00 (noon), each PlanetLab host executed the following procedures:
  1. Accessed websites according to URL list and obtained HAR (HTTP Archive) files.
  2. Extracted information of HTTP response time from obtained HAR files.
  3. Measured RTT to each object server by sending a ping.
  4. Obtained domain name of each object server using the `dig` command.
  5. Sent measurement results to collector terminal.

**Obtaining HAR Files**

- Obtained HTML file initially, and obtained each object embedded in HTML file.
- HAR (HTTP Archive) file: outputs various properties of each object in JSON (JavaScript Object Notation) format.

**Example of HAR File**


**クラスタリング手法**

- **k-means法**: 非階層型クラスタリング手法の一つで、クラスタの重心を用いて、各要素をk個のクラスターに分類する。
  - 各要素を重心の距離が最も近いクラスターに分類する処理をクラスタが収束するまで反復する。
- **k-means++法**: 距離の離れた要素を初期クラスターの重心に設定することで、分類精度を向上する。
  - ランダムに一つの要素を選び、クラスタ重心に設定する。
  - 各要素xに関する、その最近傍重心との距離D(x)を計算する。
  - D(x)に比例する確率に従い、新しいクラスター重心としてランダムに一つ要素を選択する。
  - k個のクラスター重心が選択されるまで上記処理を反復する。
  - 以後はk-means法を用いてクラスターを生成する。

**Basic Properties**

- エンターテイメントウェブサイト、たとえば、Arts、Shopping、およびSportは、多くのオブジェクトと大きなデータサイズをもつ傾向がある。
- 情報ウェブサイト、たとえば、Business、Computers、Health、およびReferenceは、少ないオブジェクトと小さなデータサイズ傾向が見られる。
各サイトの応答時間削減率のCCD

キャッシュヒット率をH=1とした場合の、各Webページの応答時間削減率GのCCDを4つのカテゴリごとにプロット

UniversalとLocalizedで、応答時間削減率に明確な差異を確認

アクセス拠点による傾向の差異

各地点・4ジャンルの応答時間削減量下限値Dの平均値(上図)と応答時間削減率下限値Eの平均値(下図)をプロット

ジャンルごとのEで見たエッジ配信の効果の順位は、どの地点でもほとんど同一

品質が良好な北米も含めて、全地域で、UniversalはLocalizedよりもエッジ配信の効果が高く、効果にジャンルグループ間の差異が見られることを確認