一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SITE2022-65,IA2022-88(2023-03)

# 多段キャッシュネットワークにおける 類似性キャッシングの特性解析に関する検討

中村　遼[†]　上山　憲昭[††]

† 福岡大学 工学部
〒 814–0180 福岡県福岡市城南区七隈 8–19–1
†† 立命館大学 情報理工学部
〒 525–8577 滋賀県草津市野路東 1–1–1
E-mail: †r-nakamura@fukuoka-u.ac.jp, ††kamiaki@fc.ritsumei.ac.jp

**あらまし**　近年，コンテンツの大規模化および多様化に伴い，ユーザのコンテンツ要求を柔軟に解決する方法として，類似性キャッシングが注目を浴びている．類似性キャッシングは，キャッシュにコンテンツ要求が到着した時に，要求されたコンテンツに類似したコンテンツを返送することを可能とする．類似性キャッシングによって，類似したコンテンツを近隣のキャッシュから取得できる可能性が高まるため，ユーザの通信に対する満足度を高めることが期待できる．ただし，類似性キャッシングは，ユーザが観測する通信品質を優先するのか，もしくは，ユーザが要求したコンテンツと実際に配送されたコンテンツとの類似度を優先するのか，というトレードオフに関する問題を生じさせる．本稿では，類似性キャッシングのためのユーザ指向な性能指標として，類似コンテンツ配送遅延 (ユーザが望む類似度を満たすコンテンツをはじめて取得するまでの時間) を定式化し，一般的なキャッシュネットワークにおける類似コンテンツ配送遅延を導出する．さらに，いくつかの数値例を通して，多段キャッシュネットワークにおける類似性キャッシングの特性を分析する．その結果として，コンテンツの要求者であるユーザのネットワーク上における位置が，ユーザにとって許容可能な類似度に影響を与えることなどを示す．
**キーワード**　類似性キャッシング，キャッシュネットワーク，コンテンツ配送遅延，数学的解析

# A Study on Characteristic Analysis of Similarity Caching on Multi-Stage Cache Network

Ryo NAKAMURA[†] and Noriaki KAMIYAMA[††]

† Faculty of Engineering, Fukuoka University
8–19–1 Nanakuma, Jonan, Fukuoka, Fukuoka 814–0180, Japan
†† College of Information Science and Engineering, Ritsumeikan University
1–1–1 Noji-higashi, Kusatsu, Shiga 525–8577, Japan
E-mail: †r-nakamura@fukuoka-u.ac.jp, ††kamiaki@fc.ritsumei.ac.jp

**Abstract**　In recent years, with the rapid growth of content space, *similarity caching* has attracted attention as a way to flexibly resolve content requests from users. The similarity caching enables in-network caching nodes, e.g., a cache server, to return any similar contents, i.e., contents which is similar to the requested one, as well as an exact content, i.e., content exactly-matching the requested one. Through such a mechanism, users can obtain similar-contents from neighboring caches; the similarity caching is expected be to provide us more comfortable communication experience. However, the similarity caching might cause a problem regarding on the trade-off: which should we prioritize, the communication quality observed by users or the similarity between the requesting content and the actually-returned content? In this paper, to deeply understand the similarity caching on a general cache network, we formulate *similar-content delivery delay* as a user-centric performance metric for the similarity caching, and then, through several numerical examples, investigate characteristics of the similarity caching on a multi-stage cache network. As a consequence, we reveal, for instance, that the tendency of the similar-content delivery delay is dominated by the position of user, i.e., content requester, as well as the content similarity.
**Key words**　Similarity Caching, Cache Network, Content Delivery Delay, Mathematical Analysis



This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.
Copyright ©2023 by IEICE

## 1 はじめに

動画像に代表されるコンテンツの継続的な増加に伴い，大規模なコンテンツ空間からユーザのコンテンツ要求を柔軟に解決する方法として，類似性に基づくコンテンツ探索 (注1) [1] が検討されている．類似性に基づくコンテンツ探索は，一定の基準に従って定義されるコンテンツ間の類似性に着目し，ユーザが要求したコンテンツに対して類似したものを発見する，というものである．その有効性は，ユーザは，要求したコンテンツそのものではなくても，ある程度類似したコンテンツを取得することができれば，一定の満足度を得られる，ということに由来する．

同時に，コンテンツキャッシングが，コンテンツをユーザに効率的に配信するための基盤的な技術であることから，ネットワーク内におけるキャッシュを活用したネットワーク (キャッシュネットワーク) の設計が活発に議論されている．コンテンツキャッシングの恩恵 (例: コンテンツ配送遅延の削減およびトラヒック量の削減) は，現在広く普及している CDN (Content Delivery Network) [2] が我々の社会生活を支える基盤になっている，ということからも明らかである．今後もコンテンツキャッシングの活用が有望視されていることから，ネットワークのエッジに位置するノード (例: アクセスポイント) におけるキャッシュを活用した MEC (Mobile Edge Computing) が活発に研究されている [3]．さらに，次世代のインターネットアーキテクチャとして，キャッシュを有するルータから構成されるネットワークである，CCN (Content-Centric Networking) [4]/NDN (Named Data Networking) [5] に対する検討も進められている [6]．

このような進展に伴い，近年では，コンテンツ間における類似性の概念をキャッシュの制御に応用した，類似性キャッシング [7] が注目を浴びている．従来のコンテンツキャッシングでは，キャッシュを保有するノード (キャッシュノード) は，ユーザからの要求が到着した時には以下の動作を行う; (i) ユーザが要求するコンテンツを保有しているかどうかを確認する; (ii) 当該コンテンツをキャッシュしていれば即座に返送する，一方，そうでない場合には，キャッシュミスとみなし，他ノード (例えば，コンテンツのオリジナルが格納されているノード) に向けて要求を転送する．一方，類似性キャッシングでは，キャッシュノードは，ユーザが要求するコンテンツに対して完全に一致するコンテンツを保有していない時に，要求されたコンテンツに対して類似したものをキャッシュしていれば，類似コンテンツをユーザに返送する．(注2) 類似性キャッシングは，例えば，ユーザがビデオを要求した時に，キャッシュノードは，ユーザが真に要求するビットレートのビデオを保有していないとしても，類似コンテンツとしてビットレートの異なるビデオを返送できるため，ユーザは速やかにビデオを視聴できる [8]，ということを可能とする．

類似性キャッシングは，ユーザに対して類似したコンテンツをキャッシュから配信することを可能とするが，ユーザが観測する通信品質を優先するのか，もしくは，要求したコンテンツと配信されたコンテンツとの類似性を優先するのか，というトレードオフに関する問題を生じさせる．通信品質を優先するユーザに対しては，ユーザが要求したコンテンツとは異なるものであったとしても，近隣のキャッシュから類似コンテンツが迅速に配送される方が望ましい．一方，自身が要求したコンテンツそのものが配信されることを望むユーザに対しては，通信品質は劣化したとしても，近隣のキャッシュではなく，ユーザが真に望むコンテンツを保有するノードから，当該コンテンツが返送される方が望ましい．

本研究は，類似性キャッシングが機能する，複数のキャッシュノードから構成される一般的なキャッシュネットワークを対象とし，通信品質とコンテンツ間の類似度との関係を定量的に明らかにすることを目指す．単一のキャッシュではなく，複数のキャッシュから構成される一般的なキャッシュネットワークでは，ネットワーク内に多様なコンテンツが散在する．このような状況下では，ネットワーク内におけるどのキャッシュがコンテンツを返送するか，かつ，どのような類似コンテンツが返送されるかが，ユーザが観測する通信品質およびコンテンツ間の類似度に影響を与える．この影響は容易に導かれるものではなく，一般的なキャッシュネットワークにおける類似性キャッシングの特性を理解することが求められる．

上述した目的を達成するために，本稿では，類似性キャッシングのための，ユーザ指向な指標である類似コンテンツ配送遅延を導入し解析的に導出する．類似コンテンツ配送遅延は，ユーザが許容するコンテンツの類似度に対する閾値が与えられた時に，ユーザがコンテンツの要求をネットワーク中に送出してから，その閾値を上回るコンテンツを初回に取得するまでに要した時間を表す．本稿では，キャッシュを有するノードから構成される一般的なキャッシュネットワークにおける，類似コンテンツ配送遅延を導出する．

さらに，いくつかの数値例によって，多段キャッシュネットワークにおける類似性キャッシングの特性を，類似コンテンツ配送遅延の観点で分析する．特に，小規模なタンデム型トポロジを対象とした数値例を通して，多段キャッシュネットワークにおける類似性キャッシングに関する問いに答える (4 節を参照されたい)．

本稿の貢献を以下に要約する．

[C1] 一般的なキャッシュネットワークにおいて類似性キャッシングを導入した時の，ユーザが観測する指標として，類似コンテンツ配送遅延を定式化するとともに，代数的に導出する．

[C2] キャッシュの多段化が類似性キャッシングに与える影響を数値例によって調査した結果として，(i) 類似コンテンツ配送遅延は類似度に応じて大きく異なること，特に，類似度が極めて高いコンテンツに対する遅延は大幅に増大すること，(ii) ユーザが許容可能な類似度は，そのユーザとコンテンツのオリジンノード (つまり，サーバ) との距離に依存することなどを示す．

本稿の構成は以下の通りである．2 節では，類似キャッシン

---

(注1)：従来研究では *similarity searching* と呼ばれる．

(注2)：要求されたコンテンツと完全に一致するコンテンツがキャッシュから返送される，従来のコンテンツキャッシングは，*exact caching* と呼ばれる．一方，本稿で対象とする類似性キャッシングは *similarity caching* と呼ばれる．

グに関する研究動向を紹介し，従来研究に対する本研究の位置付けを説明する．3 節では，本稿で対象とする解析モデルを説明し，類似コンテンツ配送遅延の導出を説明する．4 節では，小規模なタンデム型トポロジを対象としたいくつかの数値例を通して，多段キャッシュネットワークにおける類似性キャッシングの特性を分析する．最後に 5 節では，本稿のまとめと今後の課題を述べる．

## 2 関 連 研 究

本節では，まず，類似性キャッシングに関する従来研究を紹介し，その後に，従来研究に対する本研究の位置付けを説明する．

### 2.1 単一キャッシュを対象とした類似性キャッシング

我々が知る限り，類似性キャッシングの領域における先駆けとなる研究は文献 [9, 10] であり，これらの研究は，類似性キャッシングの概念を提案し，類似性キャッシングのためのキャッシュ置き換え方式を検討している．文献 [9] は，クエリによるコンテンツ検索システムにおいて，類似性キャッシングの概念を提案している．また，文献 [9] は，$k$ 近傍法を基礎とするキャッシュ制御法を提案しており，大規模な写真データから構成されるデータセットを利用した実験によって，提案手法の有効性を示している．文献 [10] は，ページに掲載されている，ユーザのプロファイルおよびコンテンツをもとに，与えられた広告集合から，最も適切な広告を提示する，というコンテンツマッチシステムに，類似性キャッシングを導入している．また，文献 [10] は，従来のコンテンツキャッシングにおける代表的なキャッシュ置き換え方式である LRU (Least-Recently Used) や LFU (Least-Frequently Used) を拡張した，類似性キャッシングのためのキャッシュ置き換え方式を提案している．シミュレーションによって，提案されたキャッシュ置き換え方式は，キャッシュヒット率を改善できることを示している．これらの研究は，対象とするシステムおよび設計した制御方式は異なるが，コンテンツ間の類似性をキャッシュの制御に応用するという点が共通する．

近年では，文献 [7] において，類似性キャッシングが再び着目され，類似性キャッシングの数理的な特性が解析されている．文献 [7] は，まず，キャッシュに到着する要求に対して，キャッシュヒット時に発生するコスト，および，キャッシュミス時に発生するコストによって定式化される目的関数を最小化するようなキャッシュへのコンテンツ配置を求めることは NP 困難である，ということを証明し，キャッシュに到着する要求パターンに応じたアルゴリズムを提案している．特に，文献 [7] は，$q$ LRU-$\Delta C$ および DUEL と呼ばれる，類似性キャッシングのためのキャッシュ置き換え方式をそれぞれ提案しており，その有効性を数理的および実験的に示している．

文献 [7] を発端として，単一のキャッシュを対象とした，類似性キャッシングのためのキャッシュ置き換え方式の検討が行われている [11, 12]．文献 [11] において提案されているキャッシュ置き換え方式の主要なアイデアは，可能な限り多くの異なる要求を解決できるようなコンテンツをキャッシュするように，キャッシュの状態を逐次的に更新する，というものである．一方，文献 [12] において提案されているものは，文献 [7] において定式化されているコストを最小化するように，オンラインアルゴリズムに従って，キャッシュの状態を更新する，というものである．これらの方式は，従来に提案されている，類似性キャッシングのためのキャッシュ置き換え方式 [9, 10] より良好な性能を示している．

### 2.2 一般的なキャッシュネットワークを対象とした類似性キャッシング

さらに，類似性キャッシングを一般的なキャッシュネットワークに拡張し，ネットワーク内の各ノードにおけるキャッシュへの最適なコンテンツ割り当てを求める，というコンテンツ配置問題に対する解法が検討されている [13, 14]．文献 [13] は，コンテンツ配送遅延とコンテンツの非類似性との重み付け和を最小化するようなコンテンツ配置を求めるためのアルゴリズムを提案している．ただし，文献 [13] で提案されたアルゴリズムは大規模なコンテンツ空間に適用できないため，文献 [14] は，離散的なコンテンツ空間だけでなく連続的なコンテンツ空間を想定し，木構造を有する規則的なネットワークにおけるコンテンツ配置問題に対する最適解を求めている．さらに，文献 [14] は，一般的なネットワークを想定した解法として，ヒューリスティックなキャッシュ置き換え方式を提案し，人工的なデータセットおよび現実的なデータセットを用いた実験によって，その有効性を示している．

### 2.3 本研究の位置付け

本研究は，類似性キャッシングを一般的なキャッシュネットワークに適用した時の，ユーザが観測する通信品質とコンテンツ間の類似度とのトレードオフを明らかにするものであり，本研究と上述した従来研究との違いは以下のようにまとめられる．

- 本研究は，単一のキャッシュではなく，複数のキャッシュから構成されるキャッシュネットワークを対象とする．

  近年では，一般的なキャッシュネットワークに対する類似性キャッシングの応用は検討されているが，従来研究は，単一のキャッシュを対象としたキャッシュ制御方式の設計が中心であり，一般的なキャッシュネットワークにおける類似性キャッシングの特性は十分に理解されていない．1 節で述べたように，キャッシュを活用したネットワークが将来的に普及することを期待できるため，キャッシュネットワークにおける類似性キャッシングの特性を明らかにすることは不可欠である．我々は，類似性キャッシングは，CCN/NDN のようなキャッシュネットワークとの親和性を有していると考える．類似性キャッシングを，コンテンツに付与されている識別子の完全一致によって要求を解決する CCN/ NDN にそのまま適用することはできないが，コンテンツに付与されている属性の類似度に基づき要求を解決するネットワークアーキテクチャ [15] には適用できる．このような研究動向を鑑みた上で，本研究は，一般的なキャッシュネットワークにおける類似性キャッシングを考える．

- 本研究は，コンテンツの類似度に対する許容度をユーザが有していると想定し，この想定を表現した性能指標である類

似コンテンツ配送遅延の観点から，類似性キャッシングの特性を分析する．

従来研究の大半は，性能を測るための指標として，キャッシュノードにおけるキャッシュヒット率をもとに，類似性キャッシングの良さを評価している．キャッシュネットワークを設計する上では，キャッシュヒット率は重要な指標ではあるが，通信ネットワークにおいて真に重要な指標は，ユーザの QoS (Quality of Service) および QoE (Quality of Experience) に直結する指標 (例: スループットまたはコンテンツ配送遅延) である．この見地に立てば，文献 [13] は，性能指標として，コンテンツ間の類似性を考慮したコンテンツ配送に要する時間を用いており，本研究と最も近いと言える．ただし，本研究では，ユーザは，得られたコンテンツに対して，そのコンテンツが自身の望むものであるかどうか，つまり，コンテンツの類似度が自身の基準を満たしているかどうか，を評価する，ということを想定する．そして，このような想定を表現した性能指標である類似コンテンツ配送遅延を定義し，類似コンテンツ配送遅延の観点で多段キャッシュネットワークにおける類似性キャッシングの特性を定量的に分析する．

## 3 解　　析

本節では，本稿で対象とする，一般的なキャッシュネットワークにおける類似性キャッシングを記述した解析モデルを紹介し，その後に，類似コンテンツ配送遅延の定義およびその導出を述べる．

### 3.1 解析モデル

本研究では，CCN/NDN で採用されている要求-応答型の通信モデルを利用したコンテンツ配送を考える．要求-応答型の通信モデルでは，ユーザはコンテンツ要求をネットワーク中に送出し，対応するコンテンツがユーザに返送されることで，コンテンツ配送が実現される．

有限のキャッシュを有するノード (キャッシュノード)，および，コンテンツのオリジナルを保有するノード (オリジンノード) から構成されるキャッシュネットワークを対象とする．キャッシュネットワークを無向グラフ $G = (\mathcal{V}, \mathcal{E})$ と表記し，キャッシュノードの集合およびオリジンノードの集合を $\mathcal{V}_c \subset \mathcal{V}$ および $\mathcal{V}_o \subset \mathcal{V}$ とそれぞれ表記する．ただし，キャッシュノードとオリジンノードが重複することはないものとする (つまり，$\mathcal{V}_c \cap \mathcal{V}_o = \phi$ である)．ノード間のリンクにおける伝搬遅延を $\tau$ と表記する．リンクにおける通信帯域は十分に大きいものと仮定し，ネットワーク内において要求やコンテンツは廃棄されないものとする．後述するように，キャッシュノードにはクライアントが接続されていることを仮定するが，クライアントとキャッシュノードとのリンクにおける伝搬遅延は無視できるくらいに小さいものとする．

ネットワーク中に存在する $C$ 種類のコンテンツ集合を $\mathcal{C} = \{1, \ldots, C\}$ と表記する．任意の二つのコンテンツ間の類似度を $s(c, c')\,(0 \leqq s(c, c') \leqq 1)$ と表記する．定義より，$s(c, c) = 1$ である．

各キャッシュノードは，大きさを $B$ とする有限のキャッシュを保有しており，自身が中継するコンテンツをキャッシュすることができる．この時に，キャッシュノードは，自身のキャッシュが他のコンテンツによって完全に占有されている場合には，キャッシュ置き換えアルゴリズムに従っていずれかのコンテンツを破棄した上で，当該コンテンツをキャッシュする．

コンテンツ要求がキャッシュノードに到着した時に，キャッシュノードは，確率 $\alpha\,(0 \leqq \alpha \leqq 1)$ でキャッシュからコンテンツを返送するかどうかを決定する．キャッシュノードがコンテンツを返送する場合には，要求されたコンテンツに対して，キャッシュ内に存在する，類似度が最も高いコンテンツを返送する．形式的に述べると，要求されたコンテンツおよびキャッシュに存在するコンテンツ集合を $c$ および $\mathcal{B} \subseteqq \mathcal{C}$ とそれぞれ表記すれば，$\arg\max_{c' \in \mathcal{B}} s(c, c')$ を満たすコンテンツ $c'$ をキャッシュノードは返送する．一方，キャッシュノードがコンテンツを返送しない場合には，キャッシュノードはコンテンツ要求を次ノードに転送する．

キャッシュノードには単一のクライアントが接続されており，クライアントは，コンテンツ要求割合 $\lambda_c$ に従って，コンテンツを要求する．ここで，クライアントは，ある類似度を満たすコンテンツを得られるまで，コンテンツ要求を繰り返し送出するものとする．

従来の CCN/NDN では，クライアントは，識別子を記述したコンテンツ要求を送出すると，その識別子に対して完全に一致するコンテンツが (ネットワーク中に存在すれば) 必ず返送される．一方，コンテンツ間の類似性を考慮する場合には，クライアントからのコンテンツ要求に対して類似したコンテンツが返送されるため，クライアントが望むコンテンツが必ずしも返送されるとは限らない．そこで，本稿では，クライアントが許容できる類似度の下限値が与えられるものとし，クライアントはその値を満たすまでコンテンツ要求を一定の間隔で繰り返し送出する．クライアントがコンテンツを取得した後に，コンテンツ要求を再送するまでの時間間隔を $\delta$ と表記する．

ノードにおける経路制御表はルーティングプロトコルに従って適切に設定されているものとし，クライアントから送出されるコンテンツ要求は当該コンテンツを保有するオリジンノードまで転送されるものとする．ノード $v$ に接続されているクライアントが送出するコンテンツ $c$ に対する要求が転送される経路を $\mathbf{P}_v^c = (v, \ldots, v_c)$ と表記する．ここで，$v_c\,(\in \mathcal{V}_o)$ はコンテンツ $c$ のオリジンノードである．

### 3.2 類似コンテンツ配送遅延の定義

本稿で対象とする類似コンテンツ配送遅延の定義を説明する．ここでは，あるキャッシュノード $v$ に接続されているクライアント (以降では，単にクライアント $v$ と表記) がコンテンツ $c$ を要求した時の類似コンテンツ配送遅延 $D_v^c$ に着目する．

ある類似度 $s$ に対するコンテンツ配送遅延 $D_v^c(s)$ を，クライアントが要求を送出してから，与えられた類似度 $s$ 以上のコンテンツを初回に取得するまでの時間と定義する．以降では，これを単に $s$-コンテンツ配送遅延と呼ぶ．上述したように，クラ

イアントは類似度 $s$ を満たすコンテンツを取得するまでコンテンツ要求を繰り返すため，$s$-コンテンツ配送遅延は，単純なコンテンツ配送遅延 (クライアントが要求を送出してから，その要求に呼応して返送されたコンテンツを受け取るまでの時間)，および，コンテンツ要求を送出する時間間隔によって構成される．

類似コンテンツ配送遅延 $D_v^c$ は，$s$-コンテンツ配送遅延 $D_v^c(s)$ と，クライアントが許容できる類似度の度合い表すパラメータ $\beta_v(s)$ との荷重平均により定式化できる．

$$D_v^c = \sum_s \beta_v(s)\, D_v^c(s) \tag{1}$$

$\beta_v(s)$ は，任意の確率質量関数によって与えられるものとする．このため，規格化条件から，$\sum_s \beta_v(s) = 1$ である．例えば，$\beta_v(1) = 1$ の単純な二値関数で与えた時の類似コンテンツ配送遅延は，従来の CCN/NDN と同様に，要求コンテンツと返送されたコンテンツとが完全に一致する時のコンテンツ配送遅延に相当する．

### 3.3 類似コンテンツ配送遅延の導出

まず，準備として，コンテンツ $c \in \mathcal{C}$ に対する要求がキャッシュに到着した時に，類似度が $s$ であるコンテンツが存在する確率 $\phi_c(s)$ を求める．コンテンツ $c$ が要求された場合に，類似度が $s$ であるコンテンツがキャッシュから返送されるのは，類似度が $s$ であるコンテンツ $c$ がキャッシュ空間に存在する，かつ，キャッシュの残余空間が類似度が $s$ 以下であるコンテンツのみによって占有されている場合である．これは，コンテンツ $c$ との類似度が $s$ 以下であるコンテンツのみによって構成されるキャッシュの組み合わせと，類似度が $s$ 未満であるコンテンツのみによって構成されるキャッシュの組み合わせとの差から表現できる．コンテンツ $c$ との類似度が $s$ 以下であるコンテンツ数およびコンテンツ $c$ との類似度が $s$ 未満であるコンテンツ数を $C^c_{\leq s}$ および $C^c_{<s}$ と表記し，次式を得る．

$$\phi_c(s) = \frac{\binom{C^c_{\leq s}}{B} - \binom{C^c_{<s}}{B}}{\binom{C}{B}} \tag{2}$$

ただし，上式は，コンテンツに対する要求頻度 (コンテンツの人気度) とコンテンツの類似性とは無相関である，という仮定の上で成立することに注意されたい．

コンテンツ $c$ に対する要求に対して，類似度が $s$ 以上であるいずれかのコンテンツが返送される確率 $p_c(s)$ は，類似度が $s$ 以上であるコンテンツが返送される確率の総和であることから，類似度が $s$ 以上である類似度の集合 $\mathcal{S}^c_{\geq s} = \{s(c,c') \mid s(c,c') \geqq s, \forall c' \in \mathcal{C}\}$ を用いて，次式で与えられる．

$$p_c(s) = \sum_{s \in \mathcal{S}^c_{\geq s}} \phi_c(s) \tag{3}$$

以降では，キャッシュノード $v$ に接続されているクライアントがコンテンツ $c$ を要求した時の，$s$-コンテンツ配送遅延の導出を説明する．特に断りのない限り，ノードを表す記号である $v$，および，コンテンツを表す記号である $c$ を省略する．例えば，$D_v^c(s)$ を単に $D(s)$ と表記する．

クライアントが初回のコンテンツ要求を送出してから $n$ 回目の要求でコンテンツを取得するまでに要した時間，および，$n$ 回目の要求で類似度が $s$ 以上であるコンテンツをクライアントがはじめて取得できる確率を，$\Delta_n$ および $R_n$ とそれぞれ表記する．$\Delta_n$ および $R_n$ を用いて，$s$-コンテンツ配送遅延 $D(s)$ は次式で与えられる．

$$D(s) = \sum_{n=1}^{\infty} \Delta_n(s)\, R_n(s) \tag{4}$$

本稿では，返送されるコンテンツの類似度に応じた 2 種類のコンテンツ配送遅延 (あるコンテンツ要求に対してコンテンツが返送されるまでの時間) に着目する．類似度が $s$ 以上であるコンテンツに対する配送遅延を $d(s)$ と表記する．また，類似度が $s$ 未満であるコンテンツに対する配送遅延を $\tilde{d}(s)$ と表記する．

$\Delta_n(s)$ は，初回のコンテンツ要求から $n$ 回目のコンテンツ要求までに要した時間 (つまり，類似度が $s$ 未満であるコンテンツの配送遅延 $\tilde{d}(s)$ と要求を再送する間隔 $\delta$ との総和) と，$n$ 回目のコンテンツ要求で類似度が $s$ 以上であるコンテンツの配送遅延 $d(s)$ との和で与えられる．

$$\Delta_n(s) = (n-1)\,(\tilde{d}(s) + \delta) + d(s) \tag{5}$$

経路 $\mathbf{P}$ における $i$ 番目のノードから類似度が $s$ 以上であるコンテンツが返送される確率を $\eta_i(s)$ と表記すれば，類似度が $s$ 以上であるコンテンツの配送遅延 $d(s)$ は，$\eta_i(s)$ と，クライアントが接続されているノード $v$ から $i$ 番目のノードまでの往復伝搬遅延 $2(i-1)\tau$ とを用いて次式で与えられる．

$$d(s) = \sum_{i=1}^{|\mathbf{P}|} \Big( 2(i-1)\tau \times \eta_i(s) \Big) \tag{6}$$

ここで，$\mathbf{P}$ は，クライアントが接続されているノード $v$ からコンテンツ $c$ のオリジンノード $v_c$ までの経路 $(v, \ldots, v_c)$ であり，$|\mathbf{P}|$ はその経路の要素数 (つまり，経路上に存在するノードの数) である．

$q_{v'}(s)$ を，ノード $v'$ から類似度が $s$ 以上であるコンテンツが返送される確率と表記する．ノードから類似度が $s$ 以上であるコンテンツが返送されるのは，キャッシュノードが確率 $\alpha$ でコンテンツを返送する，かつ，類似度が $s$ 以上であるいずれかのコンテンツがキャッシュされている，という場合である．ただし，オリジンノードは，オリジナルコンテンツ (類似度が 1 であるコンテンツ) を必ず返送することに注意すれば，$q(s)$ は次式で与えられる．

$$q_{v'}(s) = \begin{cases} \alpha\, p(s) & \text{if } v' \in \mathcal{V}_c \\ 1 & \text{otherwise} \end{cases} \tag{7}$$

ここで，$p(s)$ は，式 (3) における $p_c(s)$ であることに注意されたい．よって，$\eta_i(s)$ は，経路上の 1 番目から $i-1$ 番目のノードから当該コンテンツが返送されずに，$i$ 番目のノードから類似度が $s$ 以上であるコンテンツが返送される確率であるから，

次式で与えられる．

$$\eta_i(s) = \prod_{j=1}^{i-1}\Big(1 - q_{\mathbf{P}[j]}(s)\Big)\, q_{\mathbf{P}[i]}(s) \qquad (8)$$

類似度が $s$ 未満のコンテンツに対する配送遅延 $\tilde{d}(s)$ は，式 (3) における $\mathcal{S}^c_{\geq s}$ を集合 $\{s(c,c') \mid s(c,c') < s, \forall c' \in \mathcal{C}\}$ に置き換えれば，$d(s)$ と同様の手順によって導出できる．

最後に，クライアントが，$n$ 回目のコンテンツ要求で，類似度 $s$ 以上のコンテンツを取得できる確率 $R_n(s)$ を求める．クライアントからオリジンノードまでの経路上のいずれかのノードから当該コンテンツが返送される確率を $\rho(s)$ と表記すれば，$R_n(s)$ は次式で与えられる．

$$R_n(s) = (1 - \rho(s))^{n-1}\, \rho(s) \qquad (9)$$

$\rho(s)$ は，経路上のいずれかのノードから類似度が $s$ 以上であるコンテンツが返送される確率であるから，

$$\rho(s) = \sum_{i=1}^{|\mathbf{P}|}\Big((1-\alpha)^{i-1}\, q(s)\Big) \qquad (10)$$

である．

## 4 数 値 例

本節では，いくつかの数値例により，多段キャッシュネットワーク上の類似性キャッシングがコンテンツ配送に与える影響を，類似コンテンツ配送遅延の観点で調査する．特に，本数値例を通して，以下のリサーチクエスチョン (RQ; Research Question) に答えることを試みる．

[RQ1] キャッシュの多段化は，類似コンテンツの配送遅延にどのような影響を与えるのか?

[RQ2] ユーザにとって許容可能な類似度は，ユーザの位置に応じてどのような影響を受けるのか?

### 4.1 パラメータ設定

ネットワークトポロジとして，5 台のキャッシュノード $v_1, \ldots, v_5$ および 1 台のオリジンノード $v_6$ が直列に接続されており，各キャッシュノードに単一のクライアントが合流する，というタンデム型のトポロジを用いた．キャッシュネットワークを対象とした評価では，木構造によって表現される階層的なトポロジ，および，ネットワーク生成モデルによって人工的に生成された複雑なトポロジを用いることが典型ではある [16] が，本稿は，キャッシュの多段化による効果を正確に解釈できるように，意図的に単純なネットワークトポロジを採用した．ノード $v_i$ に接続されているクライアントが送出する要求が転送される経路は，コンテンツに依らず，一意な経路 $(v_i, v_{i+1}, \ldots, v_6)$ を設定した．また，ノード間の伝搬遅延 $\tau$ を 1 [ms] とした．

コンテンツ間の類似度は，その距離(注3)を正規化することによって与えた．文献 [13] と同様に，任意の二つのコンテンツ $a, b \in \mathcal{C}$ 間における距離 $d(a,b)$ を $d(a,b) = |a-b|^\gamma$ として与えた．ここで，$\gamma$ はコンテンツ間の距離を調整するパラメータであり，本数値例では，$\gamma = 1$ とした．コンテンツ間の距離の最小値および最大値を $d_{\min}$ および $d_{\max}$ とし，コンテンツ間の類似度を $s(a,b) = (d_{\max} - d(a,b))/(d_{\max} - d_{\min})$ と正規化した．図 1 に，コンテンツ数 $C = 100$ および $\gamma = 1$ である時の，コンテンツ間の類似度を示す．

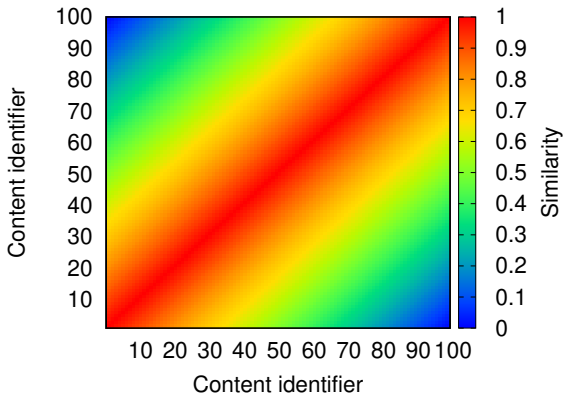


図 1: コンテンツ間の類似度

特に断りのない限り，以下のパラメータを用いた; コンテンツ数 $C = 100$，キャッシュノードにおけるキャッシュサイズ $B = 10$，コンテンツ要求割合 $\lambda_c = 1/C$，クライアントによる要求の再送間隔 $\delta = 10$ [ms].

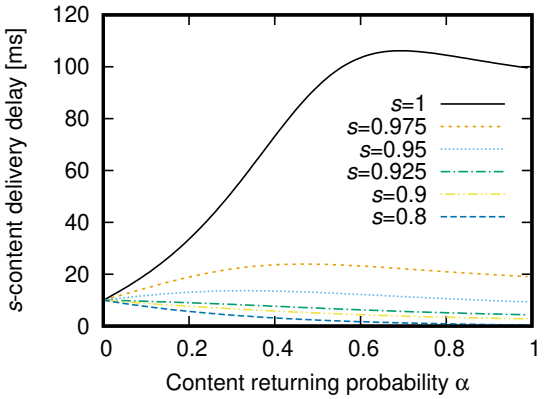
### 4.2 結　　果

以降の数値例では，多段キャッシュネットワークにおける類似性キャッシングによる影響を詳細に分析するため，あるコンテンツに対するクライアント毎の $s$-コンテンツ配送遅延 $D^c_v(s)$ (式 (4)) を示す．また，ノード $v_i$ に接続されているクライアントを単にクライアント $i$ と呼ぶ．

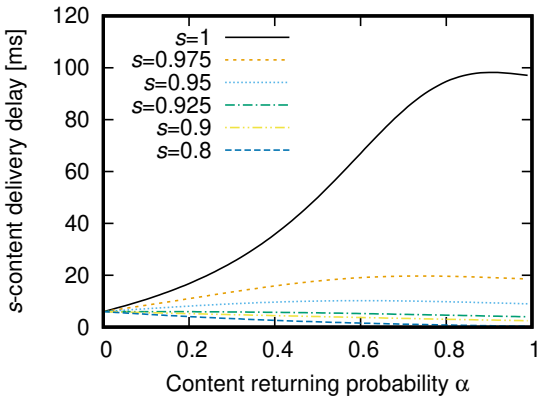
図 2 に，$\alpha$ (キャッシュノードが類似コンテンツを返送するかどうかを決定するパラメータ) を変化させた時の $s$-コンテンツ配送遅延を示す．図中には，コンテンツ間の異なる類似度 $s$ に対する結果を示している．ただし，$s$ が 0.8 未満である場合の結果に対する傾向はほぼ同様であったため，$s$ が 0.8 以上である場合の結果のみを示している．

リサーチクエスチョンに答える前に，図 2 の結果から，多段キャッシュネットワークにおいて類似キャッシングを用いた時の基本的な傾向を確認する．まず，類似度がそこまで高くないコンテンツに対する配送遅延は，$\alpha$ の増加に伴い，減少していることがわかる．キャッシュノードに，完全に一致したコンテンツがキャッシュされている可能性は低いとしても，類似したコンテンツはキャッシュされる可能性が高い．このため，キャッシュノードが積極的に類似コンテンツを返送することによって，クライアントにコンテンツが配送されるまでの時間を削減できる．一方，類似度が極めて高いコンテンツに対する配送遅延 (例: 0.975-コンテンツ配送遅延および 1-コンテンツ配送遅延) は，$\alpha$ の増加に比例して増加していることがわかる．これは，キャッシュノードは $\alpha$ に従って何らかの類似コンテンツをクライアントに返送するが，クライアントは，返送されたコンテンツが自身の望む類似度を満たさず，コンテンツに対する

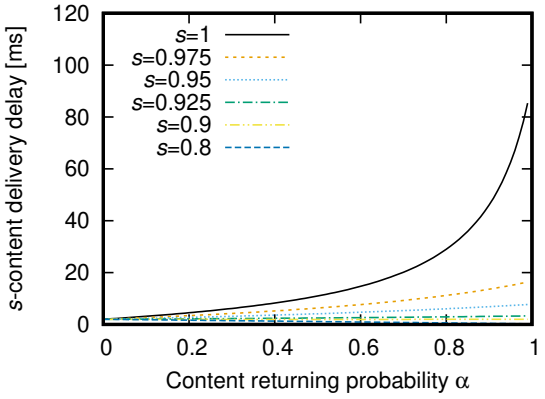
(注3)：距離は，*dissimilarity* または *cost* とも呼ばれる．




(a) クライアント 1




(b) クライアント 3




(c) クライアント 5

図 2: タンデム型トポロジにおける $s$-コンテンツ配送遅延

再送要求を繰り返すためである．その結果として，全体的なコンテンツ配送遅延の増加につながる．

図 2 の結果から，キャッシュの多段化が $s$-コンテンツ配送遅延にどのような影響を与えるか [RQ1] を確認する．クライアント 1 の結果 (図 2(a)) とクライアント 5 の結果 (図 2(c)) とを比較すると，クライアントからの要求が経由するキャッシュの台数が変化することによって明らかに傾向が異なることがわかる．オリジンノードから最も近くに位置するクライアント 5 の結果 (図 2(c))，つまり，単一キャッシュを用いた場合に相当する結果に着目すると，類似度が比較的高いコンテンツに対する配送遅延は，$\alpha$ の増加に伴い，単調に増加する．特に，この傾向

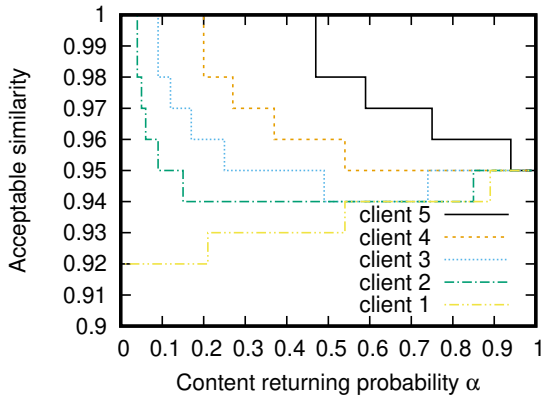



図 3: クライアント毎における許容可能な類似度

は，$s = 1$ の時には顕著である．一方で，オリジンノードから最も遠くに位置するクライアント 1 の結果 (図 2(a)) に着目すれば，$\alpha$ の増加に対するコンテンツ配送遅延の増大が抑えられていることがわかる．これは，経路上の複数のキャッシュを活用することによって，類似コンテンツが返送される可能性が高まるためである．ただし，多段キャッシュを活用することで，遅延の増大が抑えられるという傾向を示すものの，遅延が大幅に削減されるような恩恵は得られていないことに注意されたい．

RQ2 に答えるための準備として，$s$-コンテンツ配送遅延の観点でクライアントが許容可能な類似度を導入する．本稿では，許容可能な類似度は，与えられた $\alpha$ に対して，$s$-コンテンツ配送遅延が基準となるコンテンツ配送遅延以下である，ということを満たす類似度における最大値と定義する．形式的には，基準となるコンテンツ配送遅延を $D^*$ と表記すれば，許容可能な類似度は $\sup(\{s \mid D(s) \leqq D^*, \forall s\})$ で与えられる．本稿では，基準となるコンテンツ配送遅延は，クライントからオリジンノードまでの往復伝搬遅延とした (例えば，クライアント 1 の場合は，基準となるコンテンツ配送遅延は，10 (= 2 × 5) [ms] である)．図 3 に，上記の定義を用いた時の，クライアント毎における許容可能な類似度を示す．図中には，クライアント 1 ~ 5 の結果を示している．

図 3 は，RQ2 の答えとして，$\alpha$ が極端に大きい場合を除いて，クライアントにとって許容可能な類似度が，クライアントの位置によって大きく異なる，ということを定量的に示している．オリジンノードから近くに位置するクライアントに着目すると，キャッシュから類似コンテンツが返送されなかったとしても，近くに位置するオリジンノードからコンテンツが短い遅延で返送されるため，許容可能な類似度は 1 に近付くことがわかる．ただし，キャッシュノードが積極的にコンテンツを返送することは，クライアントに配送されるコンテンツの類似度が低下することにつながるので，許容可能な類似度は減少する傾向にある．一方，オリジンノードから遠くに位置するクライアント (例: クライアント 1) にとっては，近くに位置するキャッシュノードからコンテンツが返送される方が望ましいため，$\alpha$ の増加に対して許容可能な類似度は増加する傾向にある．

## 5 まとめと今後の課題

本稿では，多段キャッシュネットワークにおける類似性キャッシングに着目し，ユーザが観測する通信品質とコンテンツ間の類似性とのトレードオフを数理的に分析した．具体的には，まず，類似性キャッシングのためのユーザ指向な指標である，類似コンテンツ配送遅延を定式化し導出した．さらに，小規模なタンデム型トポロジを対象とした数値例により，多段キャッシュネットワークにおける類似性キャッシングの特性を分析した．その結果として，類似コンテンツ配送遅延は類似度に応じて大きく異なること，ユーザが許容可能な類似度はネットワーク上におけるユーザの位置に依存することなどを示した．

今後の課題として，まず，本稿の数値例ではネットワークトポロジに関する議論が限定的であるため，ネットワークトポロジの構造が類似性キャッシングに与える影響を解明することが挙げられる．さらに，人工的に生成したコンテンツの類似度を用いるのではなく，実際の通信を観測することによって得られる情報をもとに構築したデータセットを用いた評価を行うことも重要な課題である．

## 謝　　辞

本研究の一部は JSPS 科研費 21H03436 の助成を受けたものである．

## 文　　献

[1] E. Chávez, G. Navarro, R. Baeza-Yates, and J. L. Marroquín, “Searching in metric spaces,” *ACM Computing Surveys*, vol. 33, pp. 273–321, Sept. 2001.

[2] A. Vakali and G. Pallis, “Content delivery networks: Status and trends,” *IEEE Internet Computing*, vol. 7, pp. 68–74, Dec. 2003.

[3] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, “A survey on mobile edge computing: The communications perspective,” *IEEE Communications Surveys & Tutorials*, vol. 19, pp. 2322–2358, Aug. 2017.

[4] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, “Networking named content,” in *Proceedings of the 5th International Conference on emerging Networking EXperiments and Technologies (CoNEXT 2009)*, pp. 1–12, Dec. 2009.

[5] L. Zhang *et al.*, “Named data networking,” *ACM SIGCOMM Computer Communication Review*, vol. 44, pp. 66–73, July 2014.

[6] G. Xylomenos *et al.*, “A survey of information-centric networking research,” *IEEE Communications Surveys & Tutorials*, vol. 16, no. 2, pp. 1024–1049, 2014.

[7] G. Neglia, M. Garetto, and E. Leonardi, “Similarity caching: Theory and algorithms,” *IEEE/ACM Transactions on Networking*, vol. 30, pp. 475–486, Dec. 2021.

[8] A. Araldo, F. Martignon, and D. Rossi, “Representation selection problem: Optimizing video delivery through caching,” in *Proceedings of the 15th IFIP Networking Conference and Workshops*, pp. 1–9, May 2016.

[9] F. Falchi, C. Lucchese, S. Orlando, R. Perego, and F. Rabitti, “A metric cache for similarity search,” in *Proceedings of ACM Workshop on Large-Scale Distributed Systems for Information Retrieval (LSDS-IR 2008)*, pp. 43–50, Oct. 2008.

[10] S. Pandey, A. Broder, F. Chierichetti, V. Josifovski, R. Kumar, and S. Vassilvitskii, “Nearest-neighbor caching for content-match applications,” in *Proceedings of the 18th International Conference on World Wide Web (WWW 2009)*, pp. 441–450, Apr. 2009.

[11] A. Sabnis, T. S. Salem, G. Neglia, M. Garetto, E. Leonardi, and R. K. Sitaraman, “GRADES: gradient descent for similarity caching,” in *Proceedings of IEEE Conference on Computer Communications (INFOCOM 2021)*, pp. 1–10, May 2021.

[12] T. S. Salem, G. Neglia, and D. Carra, “AÇAI: Ascent similarity caching with approximate indexes,” in *Proceedings of the 33rd International Teletraffic Congress (ITC 2021)*, pp. 1–9, Aug. 2021.

[13] J. Zhou, O. Simeone, X. Zhang, and W. Wang, “Adaptive offline and online similarity-based caching,” *IEEE Networking Letters*, vol. 2, pp. 175–179, Dec. 2020.

[14] M. Garetto, E. Leonardi, and G. Neglia, “Content placement in networks of similarity caches,” *Computer Networks*, vol. 201, pp. 1–13, Dec. 2021.

[15] M. Papalini, *TagNet: A Scalable Tag-Based Information-Centric Network*. PhD thesis, Faculty of Informatics of the Università della Svizzera Italiana, Oct. 2015.

[16] A. Ioannou and S. Weber, “A survey of caching policies and forwarding mechanisms in information-centric networking,” *IEEE Communications Surveys & Tutorials*, vol. 18, pp. 2847–2886, May 2016.