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
CPSY2018-109,DC2018-91（2019-03）

広域データ分析システムにおける サービス品質を考慮した類似データ除去手法

一角　健人† 　五木田　駿† 　久保田　真† 　福山　訓行†

† (株) 富士通研究所　〒211-8588 川崎市中原区上小田中 4-1-1
E-mail: †{ikkaku.kento,gokita.shun,kubota.makoto,noriyuki}@fujitsu.com

あらまし　近年，車のドライブレコーダーに記録された映像データなど地理的に分散蓄積されたエッジデータを，クラウド上で実行される各種サービスからオンデマンドに活用可能にしたいとのニーズが高まりつつある．しかし，要求に合致するすべてのデータをクラウドに収集するというアプローチは，ネットワーク帯域が狭いためデータ転送に時間がかかること，エッジノードの通信帯域にばらつきがあるためデータが揃うまでに時間がかかることがあるために対応できない．本研究では，複数のエッジノードにサービスにとって同一の意味を持つデータである類似データが重複して保存されている可能性があることを利用し，サービスにとって必要なデータの収集完了時間が最小になるようにエッジノード間で類似データを除去する手法を提案する．

キーワード　広域データ分析，IoT，類似データ除去

QoS Aware Similar Data Reduction Method in On-demand Wide-Area Data Analysis System

Kento IKKAKU†, Shun GOKITA†, Makoto KUBOTA†, and Noriyuki FUKUYAMA†

† FUJITSU LABORATORIES LTD. Kamikodanaka 4-1-1, Nakahara-ku, KAWASAKI, 211-8588 Japan
E-mail: †{ikkaku.kento,gokita.shun,kubota.makoto,noriyuki}@fujitsu.com

Abstract　Recently, needs for making it possible to utilize geo-distributed and accumulated data, such as video data recorded in a car drive recorder, from various services on demand are increasing. The approach of collecting all the data that meets an user request in the cloud greatly impairs responsiveness because the network bandwidth is low and it can not deal with dynamically changing the communication bandwidth and computer load in each edge node. In this paper, we use the fact that each edge node stores similar data having the same meaning as viewed from the service in duplicate, and propose a method to efficiently remove similar data adjusting to current bandwidth between edge nodes, which minimize the data collection completion time.

Key words　Widely data analysis, IoT, similar data removal

1. はじめに

近年，IoT デバイスの増加に伴い，従来は無かったデータが大量に生成されている．例えば，今日の自動車は，後で故障分析や事故原因分析を行うためにエンジン回転数などの計測用センサーやドライブレコーダを装備し，現在の走行状態や周辺状況を常に計測/記録している．計測用センサーから取得される CAN(Controller Area Network) データのデータ量は最大 1 GB/day，ドライブレコーダーに記録された映像データのデータ量は車 1 台当たり数 GB/h，にも上る [1]．今後は，都市計画や街の安全などを目的としたサービス事業者が，地域や時間帯を指定し，その地域・時間に走行した複数車両のドライブレコーダーの映像データを用いて歩行者の人流分析 [2] をするなど，複数車両に跨って任意のデータをオンデマンドに活用したいとのニーズが高まりつつある．

現場で大量に発生するデータをオンデマンドに活用可能とするシステムでは，利用者から使われるかどうかわからないため，発生するすべてのデータをクラウドに集約することは無駄である．[3] では，オンラインサービスプロバイダである Microsoft のクラウド Cosmos 上に蓄積されたデータの 80% は，5 日間でアクセスが 2 回以下であり，クラウドに収集されたデータの大部分がほとんどアクセスされないことを示している．実際，歩道の人流分析において必要なデータは，データ利用者が指定した時間帯/地域を走行する車のデータのみであり，限定的であ

る．そこで，我々の研究グループでは [4] 及び [5] において，映像データなどのデータを車載器などのコンピューティング/ストレージ/ネットワーク機能を持った機器 (エッジ) に蓄積し，データの保存場所をクラウドで管理することで，データ利用者はオンデマンドにデータにアクセス可能とする技術を提案している．一方，エッジからクラウドにデータを収集する際，エッジ-基地局間の移動通信として一般に LTE 通信が用いられるが，その帯域はクラウド内のネットワーク帯域より狭いため，複数のエッジに蓄積された大量のデータを収集するには長時間を要することがある．

データの収集時間を短縮する従来手法として，2 つのアプローチがある．一つ目のデータ配置技術 [6] は，利用者から分析要求が到着する前に，データ転送開始から完了までに最も時間がかかる (ボトルネック) と推測されるエッジからボトルネックではない他のエッジに，利用される可能性が高いデータのコピーを移動させておくことで，分析要求が到着した時にボトルネックのエッジから収集されるデータの転送量を削減しているが，本来エッジからクラウドへの転送に使いたい LTE 帯域をエッジ間の事前データ移動に使用されてしまい，効率が悪い．2 つ目のデータ集約技術 [7] は，近くを走行する車が保持するデータはサービスにとって同一の意味を持つ可能性があること (類似性) を利用し，近くの車同士で組んだクラスタ中の 1 台が車車間通信を用いてクラスタ内の全ての車から集約したデータを圧縮してクラウドに転送することでデータ転送量を削減する．この技術は，事前のデータ移動は LTE 回線を使わず車車間通信を使うため，選択された 1 台は LTE 通信を全てクラウドへのデータ転送に利用でき，かつ，類似した情報を圧縮して転送するため，圧縮せずに集めた場合と比べて同様の品質のサービスや分析を提供できる．しかし，車は移動するため，過去データを持つ車の間で，車車間通信はできない．また，クラウドにデータを転送する 1 台を選択するアルゴリズムは，各エッジに割り当てられた LTE のネットワーク帯域を考慮できていないため，選択された車に割り当てられる帯域が他のエッジより狭い場合にデータ転送量を削減できても収集時間が長くなる．

本研究では，データの品質を維持しながら収集完了時間を削減することを目的とし，データ集約技術をベースに，収集時間が最小となるように類似性のあるデータ (類似データ) をタイムスロットごとに分割した上で，タイムスロットごとに収集先エッジを変える手法を提案する．類似データの分割では，ボトルネックのエッジに割り当てられる類似データのデータ量は，他のエッジよりも少なくなるように割り当てる．車から映像データを収集するシナリオを模擬したシミュレーションで，類似データを持つエッジの中から最も帯域が広い 1 台を選択する方式と比較してデータ収集時間において，68 %以上の削減効果が得られることを示す．

以降，2 章では，我々が想定するシステムと関連技術を説明する．3 章では，提案手法を説明し，4 章では，提案手法のシミュレーション評価の結果について示す．5 章ではまとめと今後の課題について述べる．

2. 想定するシステムと関連技術

本章では，想定するシステム，および既存のデータ収集時間削減技術について述べる．

2.1 想定するシステム

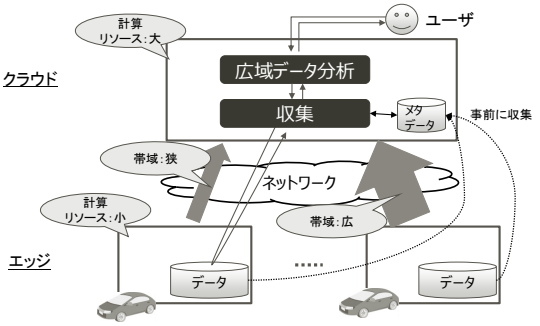

図 1　想定システム

我々が想定するシステムの概要を図 1 に示す．本システムは，広域に散在するエッジとクラウドとそれらをつなぐネットワークとで構成される．クラウドでは，利用者から要求があった時に該当するデータを読み出せるようにデータの属性情報 (メタデータ) を保持している．メタデータは，データ位置としてデータが保存されているエッジの場所 (アドレス/ホスト名/URI)，データの発生時刻，データサイズ，およびアクセス方法を示すプロトコルなどを含む [4]．広域データ分析を実行する一連の流れとして，クラウドが利用者から分析要求を受付けると，要求に合致するデータをメタデータから検索してエッジに対してデータ読出し要求を送信する．読出し要求を受け取ったエッジは，該当するデータをストレージから読み出してクラウドに転送する．クラウドは，エッジから転送されてきたデータを用いて利用者指定の分析処理を実行する．分析処理の例として，車のドラレコ映像を用いた人流解析などが挙げられる．この例では，利用者が地域と時間帯を指定して読み出された映像データを用いて画像処理することにより，街頭を歩く人の方向と人数を可視化することができる [2]．

本システムでは，エッジは車載器などを想定し，CPU やメモリなどの計算リソースとデータを蓄積するストレージを備えている．計算リソースはクラウドと比較して乏しいことを前提としている．ネットワーク帯域も同様，クラウド内と比較して，エッジごとに割り当てられる帯域は狭い．また，LTE などの移動体通信ではネットワーク帯域はエッジごとでばらつきがある．帯域がばらつく状況の例として，[8] では，ドルトムント周辺を走る車から LTE 基地局を介してクラウドにデータを収集するシナリオを想定し，各車に割り当てられる帯域に差があることをトラヒックシミュレータ SUMO とネットワークシミュレータ Opnet を使って再現している．そこでは，車が渋滞しているところとそうでないところで車一台あたりに割り当てられる帯域に 7 倍の差があることを示している．原因として，各エッジに割り当て可能な帯域リソースの最小単位であるリソースブロックの数が LTE 基地局のセル内の車の数により異なる点などが挙げられている．以上のように，ネットワークの帯域が狭く，エッジの計算リソースが乏しいことに加え，エッジごとに

帯域にばらつきがあることから，エッジからデータ収集時，複数のエッジの中で最も収集に時間のかかるエッジがボトルネックとなり，データ収集完了時間が膨大となる．

2.2 関連技術

本節では，1 章で記述した従来技術を詳しく説明する．

2.2.1 データ配置技術

地理的に分散蓄積されたデータを一箇所に集めずに分析処理フレームワーク (Apache Spark, MapReduce など) を実行する研究がある．[6] では，分析要求に対する応答時間を短縮するため，インターネットなどの広域なネットワークを跨いだデータ転送の完了時間が最も遅い場所 (ボトルネック) にあるエッジのデータ量が少なくなるように事前にボトルネックとならない他の場所にあるエッジにデータのコピーを移動する手段と，分析要求到着時に事前移動後のデータを用いてタスク配置を決定する手段を持っている．事前のデータ移動では，現在ボトルネックの場所にいるエッジから，利用頻度，データアクセス間隔，移動データ量により算出されたスコアが最も高いデータセットを移動する．この事前データ移動を次の分析要求が到着するまで繰り返す．この方法により，分析要求到着時にボトルネックの場所からのデータ転送量が減るため，広域ネットワークを跨いだデータ転送完了時間が短縮され，応答時間が短くなる．

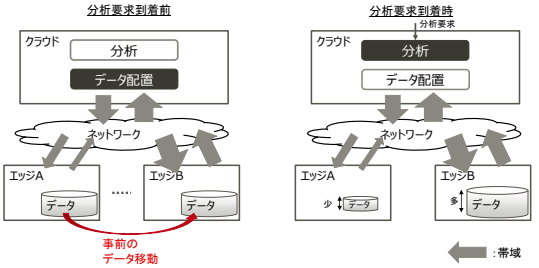

図 2　データ配置技術の例

2.2.2 データ集約技術

エッジ間でセンシング対象が類似している場合，エッジが保持するデータも類似している可能性が高い．例えば，近くを縦走する 2 台の車のドライブレコーダーが記録した歩道の映像データにはほぼ同じ人物が映っている．この特性に着目し，エッジ間で類似データを圧縮することでデータの品質を維持しつつデータ量を削減する手法がある．データ品質を維持するとは，類似データを圧縮しても収集しても，圧縮せずに集めたときと同様のサービスや分析を提供できることをいう．[7] では，現在走行中の車両からデータを収集するシーンにおいて，車車間の相互通信によりデータが類似している可能性が高い車同士でクラスタを組み，クラスタの中から 1 台の車を選出してその車のみデータをクラウドに転送する (図 3)．クラスタの組み方は，まず各車両は無線通信方式 802.11p を用いて車車間通信で到達できる車両を把握し，その中から自身の速度と進行方向との差がしきい値以下の車両を探索する．該当する車両の中で，過去にクラスタヘッドになった回数と各々が発見できた条件に該当する車両数の和が最も大きい車両をクラスタヘッドとして選択する．クラスタヘッドは，クラスタ内からデータを集約し，圧縮したデータをクラウドに転送する．このとき，クラスタ内の車両が保持するデータは類似しているために圧縮率が高い．そのため，クラスタ内で発生するデータを実質 1 台のデータと見立てて収集されるため，データ量が大幅に削減される．

クラウドで各車両の位置情報と速度情報と方向情報を保持していれば，車車間通信を用いなくてもクラウドでセンシング対象が類似している車両を発見することは可能である．圧縮率が高いことから，車車間でデータを集約する工程を省略し，1 台の車両からデータを転送することでデータの品質を維持しつつデータ量を削減できる．

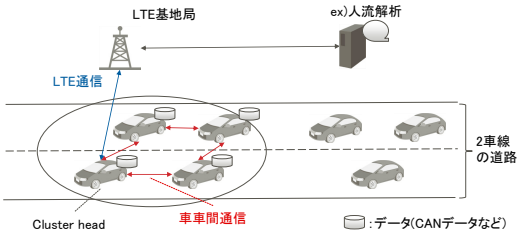

図 3　データ集約技術の例

2.3 課　　題

データ移動技術はデータセンタ間の広域データ分析を想定しており，使用可能な帯域リソースが十分にある環境であるが，我々が想定する LTE 環境では，各エッジに割り当てられる帯域が狭く，エッジ間で繰り返しデータ移動することは本来クラウドへの転送に使用したい帯域を消費してしまうため，非効率である．一方，データ集約技術では，LTE 回線を使ったエッジ間のデータ移動がないため，クラウドへの転送に使用できる帯域の消費を抑えられる．しかし，エッジ間でのデータ集約および圧縮後のクラウドへのデータ転送において，クラウドへ転送する 1 台のエッジを選択するときに帯域を考慮していないため，帯域の狭いエッジが選択された場合，このエッジがデータ転送完了時間のボトルネックとなり収集完了時間が長くなる．

ネットワーク帯域を考慮しつつ類似データを除去をする必要性を動機付け，そして従来の解決策が不十分であることを図 4 の試算結果を用いて示す．3 台のエッジ i に蓄積されたデータ I_i に対して前処理をして中間データ収集する場合，収集する中間データ $S_i = \alpha \cdot I_i$ と表される．ここで，前処理はデータ量を削減する処理と想定し，前処理によるデータ量削減率 α は $(0 < \alpha \leq 1)$ であると仮定する．次に，ボトルネックのエッジ $(i = 0)$ に割り当てられる帯域 b_i は 1Mbps，それ以外のエッジは 10Mbps とする．表 1 は，その他詳細なパラメータを示す．これらの設定パラメータの下で，データ収集にかかる収集完了時間 $T_i^{Collect} = S_i/b_i$ を試算する．図 4 の左は，3 台の中から 1 台選んで，その 1 台から全ての類似データを収集したときの結果を示す．もしボトルネックのエッジが選択された場合は，ボトルネックのエッジから 150MB 収集されることになり，収集完了時間は 1200s であり，もしボトルネック以外のエッジが選択された場合は，収集完了時間は 40s である．(後者のボトルネック以外のエッジが選択された場合は，4 章の評価において比較手法のエッジ選択手法として用いる．) [7] では，ボトルネックのエッジが選択されることがあり，非効率である．以降は，非効率な従来手法に変わる我々のアプローチの例である．図 4 の真ん中は，左と同様に 3 台のエッジから類似データを 1

つ収集するが，3 台で 1 つの類似データとなるように類似データを均等に分割し，その分割されたデータをそれぞれのエッジから収集した場合の結果を示す．(4 章の評価において比較手法の均等分割手法として用いる．) この場合，各エッジが収集する類似データは 3 分割されるので，全て 50MB である．このとき，収集完了時間はボトルネックのエッジの 400s となる．1 台のエッジを選択して収集する手法でボトルネックのエッジを選択する場合と比べると，収集完了時間は短くなる．最後に，図 4 の右は，真ん中と同様に類似データを 3 台で分割して収集するが，ボトルネックのエッジから収集するデータ量を少なくして収集した場合の結果を示す．(3 章では，このボトルネックのエッジから収集する類似データのデータ量を最適化する．これを 4 章の評価において提案手法として用いる．) この例では，各エッジから収集するデータ量は，ID の若い順に，10MB, 70MB, 70MB とした．この場合，収集完了時間はボトルネックのエッジから収集する時間の 80s で，どの手法よりも短く収集できていることがわかる．

表 1　設定パラメータ

パラメータ	値
エッジ数	3
ボトルネックのエッジ数	1
類似するデータのデータ量	300[MB]
前処理によるデータ量削減率	50[%]
中間データのデータ量	150[MB]
帯域	1，10[Mbps]

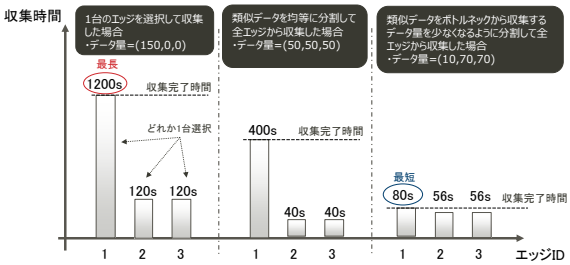

図 4　試　　算

3.　提 案 手 法

前述のとおり，我々の想定するシステムは帯域が狭いため，データ配置技術の適用は難しい．本章では，データ集約技術をベースとして，データの品質を維持しつつデータ収集完了時間を短縮する類似データ除去手法について説明する．

3.1　概　　要

提案手法の概要を図 5 を用いて説明する．図 5 は図 1 の収集部に類似データ探索部，類似データ除去部を追加拡張したものである．まず，類似データ探索部において，利用者が所望する収集対象データの中からエッジ間で類似するデータの探索を行い，以下に述べる方法でエッジ間で類似するデータである類似データと自身しか持ち合わせていないデータである希少データに切り分ける．類似データ探索部は，エッジを車としたシーンを想定した場合，近隣の車両とそれらの車両が保持するデータに類似性がある特性を用いる．エッジに蓄積された過去のデータに対して類似データ探索するため，クラウドにおいて 2.2.2 項で示したエッジの状態情報 (位置，方向，速度) を保持しておく．分析要求が到着すると，指定する時間帯と場所に当てはまるエッジの状態情報を用いて，指定された時間帯でエッジ間の時空間相関をとり，そのエッジ同士の相関が強い状態を維持し続けている時間帯のデータを類似データとする．相関が強いエッジが見つからない時間帯のデータは希少データとする．

次に，類似データ除去部では，データの品質を維持しつつ収集完了時間が最小化するように，各エッジから収集すべきデータ決定する．具体的には，複数のエッジから類似データの一部を収集すれば 1 つの類似データが完成するように，類似データを分割する．これは，類似データは 1 つ集めればデータの品質を損なわないことを前提としている．分割するとき，類似データのデータ量情報と希少データのデータ量情報，各エッジの現在の帯域情報を用いて，収集時間が最小になるようにエッジ間で分割する割合を決定する．分割が完了すると，エッジごとに類似データから割り当てられた部分以外のデータを除去し，希少データと合わせて収集すべきデータとする．収集部から収集すべきデータを指定してエッジに要求を送信する．次節では，類似データの分割の詳細について説明する．

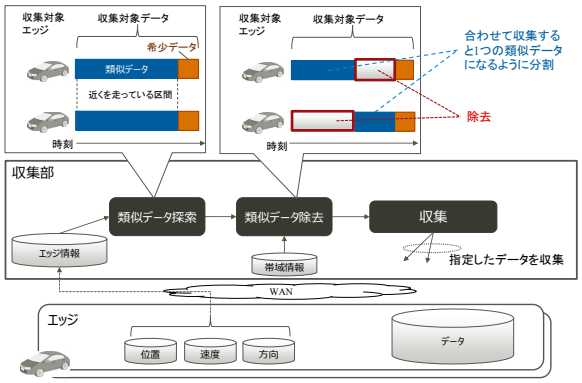

図 5　提案手法の概要

3.2　類似データの分割

入力となる類似データのデータ量が既知の場合に，収集時間が最小になるようにエッジ間で類似データを分割する方法を提案する．エッジ間で類似データの分割割合を調整することにより収集時間を最小化する問題を定式化し，線形計画法を用いてこの問題を解く．(以降，同じ類似データを持つエッジの集団を類似グループと呼ぶ．) 類似グループ内で各エッジに割り当てられる帯域は，2.1 節でも示したとおりエッジの現在いる場所が渋滞しているか渋滞していないかに依存する．本手法は，各エッジが所属する類似グループが 1 つか複数かで場合分けして説明する．各エッジが 1 つの類似グループに所属する場合，類似グループ内でのみ類似データの分割割合を調整する．一方，各エッジが複数の類似グループに所属する場合，各類似グループ間で類似データを分割し合うエッジが異なる．この場合，グループ内のエッジの帯域の偏りのみを考慮して分割の割合を調整しても，グループ外のエッジを含めた全エッジで収集時間を最小化できない．そのため，全エッジの帯域の偏りを考慮し，類似グループ内の類似データの分割割合を調整できるような要素を付け加える．これにより，帯域が狭いエリアにいるエッジ

は，所属する全ての類似グループでより多めにデータ量を減らせるように調整できる．

まず，各エッジが 1 つの類似グループに所属している場合の定式化を行う．ある類似グループ内のエッジ数を M とし，各エッジ $i(0 \leq i \leq M)$ から収集する前処理後の中間データのデータ量が S_i である場合において，エッジ間でデータ収集時間を最小化することを考える．S_i は，入力データのデータ量 I_i に前処理によるデータ削減率 α $(0 < \alpha \leqq 1)$ を掛けて算出されたものであり，また，類似データのデータ量 D_i と希少データのデータ量 R_i の和である．本手法では，エッジ間で最長の収集時間 y を最小化するために各エッジに割り当てる類似データの分割割合 ri を決定する．この問題を定式化するために，類似データは分割可能と仮定する．r_i の決定に関わる重要な要素は各エッジに割り当てられた帯域 b_i である．この帯域下で各エッジから収集するデータのデータ量は $S_i = D_i \cdot r_i + R_i$ となる．このとき各エッジのデータ収集時間は，$T_i^{Collect}(r_i) = (D_i \cdot r_i + R_i)/bi$ となる．よって，類似データの分割割合 r_i を決定する問題は以下のように定式化できる．

$$
\begin{aligned}
&min && y \\
&s.t. && \sum_{i=0}^{M} r_i = 1 \\
& && \forall i : (D_i \cdot r_i + R_i)/bi \leq y
\end{aligned}
$$

次に，各エッジが複数の類似グループに所属している場合の定式化を行う．ここでは，全エッジ $j(0 \leq j \leq N)$ 間で収集時間 z が最小になるように各類似グループ内で類似データの分割割合の調整を可能にするための重み $w_{\mathrm{j}}\,(0 < w_j \leq 1)$ を決定する．w_i は，以下のようにして求める．

$$
\begin{aligned}
&min && z \\
&s.t. && \sum_{j=0}^{N} w_j = 1 \\
& && \forall\ \mathrm{j} : w_j \cdot S_j/b_j \leq z
\end{aligned}
$$

上記で求まった w_j を用いて以下のように類似グループ内の類似データの分割割合 r_i を求める．

$$
\begin{aligned}
&min && y \\
&s.t. && \sum_{i=0}^{M} w_i \cdot r_i = 1 \\
& && \forall i : (D_i \cdot w_i \cdot r_i + R_i)/bi \leq y
\end{aligned}
$$

以上で定式化した問題を線形計画法で解くことにより，データの品質を維持しつつ収集時間を最小化するように，各エッジから類似データの分割して除去することが可能となる．

4. 性 能 評 価

4.1 シミュレーションモデル

提案手法の有効性を評価するため，想定するシナリオから作成した簡易なモデルにおいてシミュレーションを行った．本評価では，過去に都会を走行していた車のドラレコから映像データをクラウドに収集して人流解析等の分析を行う場合を想定する．シミュレーションでは，利用者が指定する時間帯に近くを走行していた車 5 台から 10 分間の映像データを収集するシーンを想定する．各車から収集する映像データは 500MB で，これらの映像にはほぼ同じ人物が映っていることを想定し，人流解析に用いるうえでは類似データと見なせる．利用者の要求が到着した時の各車の位置は，過去に近くを走っていた場所からそれぞれ散らばって走行しているとする．このとき，LTE 環境だと，各車に割りあてられる帯域は，渋滞エリアと渋滞していないエリアで帯域にばらつきが生じ，渋滞エリアにいる車の方が帯域が細くなる [8]．本シナリオも同様，帯域が狭いエリアにいるエッジと帯域が広いエリアにいるエッジの 2 種類にわけ，エッジ間で平均で 10Mbps になるように，①帯域が狭いエリアのエッジと帯域が広いエリアのエッジの帯域の差の絶対値 (帯域差) と②帯域が狭いエリアにいる車の割合とを変化させ，各エッジに割り当てる帯域を決定した．図 6 に，帯域差と，帯域が狭いエリアにいるエッジ数に対する，帯域が狭いエリアにいるエッジと帯域が広いエリアにいるエッジの帯域分布を示す．上側 (赤) の分布が帯域が広いエリアにいるエッジの帯域分布を示し，下側 (青) の分布が帯域が狭いエリアにいるエッジの帯域分布を示す．帯域が広いエリアにいるエッジは，帯域差と帯域が狭いエリアにいる割合が小さいほど，割り当てられる帯域が小さくなる．一方，帯域が狭いエリアのエッジは，帯域差が大きく，帯域が狭いエリアのエッジ数が少ないほど，割り当てられる帯域が小さくなる．その他，詳細なパラメータを表 2 に示す．評価では，この帯域分布を用い，類似データを持つエッジの中から最も帯域が広いエリアにあるエッジを 1 つ選択する方式 (エッジ選択方式)，類似データを持つエッジ間で類似データを均等に分割して割り当て，エッジごとに割り当てられなかった部分を除去する方式 (均等除去方式) との比較により，提案手

表 2　シミュレーションパラメータ

パラメータ	値
エッジ数	5
各エッジから収集するデータ量	500[MB]
前処理によるデータ量削減率	50[%]
平均帯域	10[Mbps]
類似データ数	1
類似グループ数	1
類似データのデータ量	500[MB]

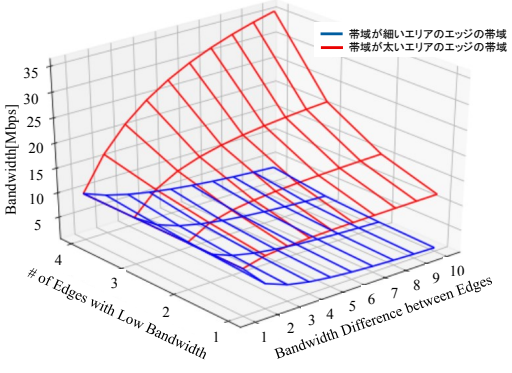

図 6　帯 域 分 布

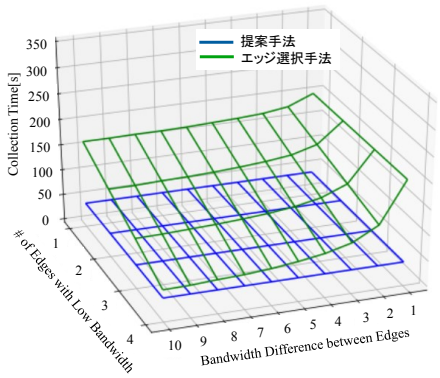

図 7　収集時間 (エッジ選択方式との比較)

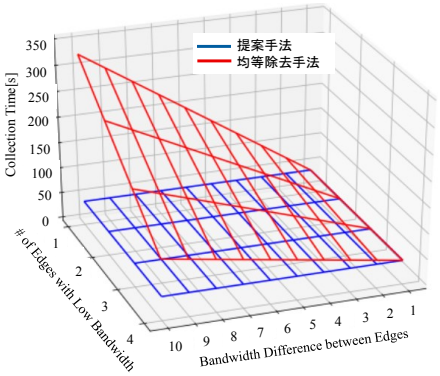

図 8　収集時間 (均等除去方式との比較)

法の収集時間平準化の有効性を示す．

4.2　評 価 結 果

図 7, 8 は，帯域差と，帯域が狭いエリアにいる割合に対する収集時間について，提案手法とエッジ選択方式，提案手法と均等除去方式を比較した結果を示す．グラフは，縦軸の数値を見やすくするため，図 6 の横軸を 180% 回転している．図 7，8 より，提案手法は，帯域分布の偏りに応じて類似データを分割して除去しているため，帯域差，帯域が狭いエッジの変化に対して収集時間が 40 s と低く安定している．それに対して，エッジ選択方式は，帯域が広いエッジ 1 台から全ての類似データを収集するため，収集時間はそのエッジの帯域に依存し，最大で 200 s となる．均等除去方式は，帯域分布の偏りに依存せずに類似データを均等に分割して除去するため，収集時間は帯域が狭いエッジがボトルネックとなり，最大で 328s かかる．以上から，我々が想定するシナリオでは，収集対象となるエッジの場所は制御不能であり，そのエッジの場所による帯域分布の偏りがある中で，提案手法は他の手法と比べて収集時間を安定して短く維持できることを示した．

図 9 は，帯域差に対する収集時間削減率を示している．収集時間削減率は，提案手法のデータ収集時間をそれぞれの比較方式のデータ収集時間で割った値である．帯域が狭いエリアにいるエッジ数は 3 台で，車が渋滞エリアに存在する割合が大きい状況における結果である．図 9 から，提案手法は，エッジ選択方式と比較して最低でも 68%，均等除去方式と比較して最低でも 37% 削減できていることが分かる．

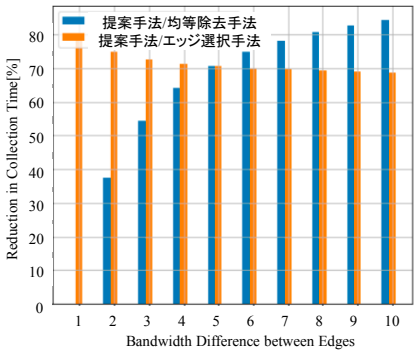

図 9　収集時間削減率

5. ま と め

利用者によりアドホックに指定された分析対象データの条件（場所・時間帯）に従い，該当するデータを持つエッジからデータ収集して分析を可能とするシステムにおいて，データの品質を落とさずに収集時間を最小化するように，事前にエッジ間で類似するデータを除去して収集可能にする技術を開発した．その結果，簡易なモデルによるシミュレーション評価では，類似するデータを持つエッジの中で最も帯域が広いエッジ 1 台を選択して収集する方式と比較して最低でも 68 %削減できることを確認した．

今後は，現実的な帯域情報を使ったシミュレーションの評価および事前計算にかかる時間のオーバヘッドの評価を行う予定である．また，分析アプリをつかって分析品質が維持できているかの確認も行う予定である．

文　　献

[1] Pillmann, J., C. Wietfeld, A. Zarcula, T. Raugust, and D. C. Alonso, “Novel Common Vehicle Information Model (CVIM) for Future Automotive Vehicle Big Data Marketplaces, ” In Intelligent Vehicles Symposium (IV), 2017 IEEE (pp. 1910-1915). IEEE.

[2] 原 佑輔, 他, “ドライブレコーダー映像を用いた頭部検出に基づく人流推定法の提案, ” 分散協調とモバイルシンポジウム (DICOM0), July 2016

[3] G. Ananthanarayanan, S. Agarwal, S. Kandula, A. Greenberg, I. Stoica, D. Harlan, and E. Harris. “Scarlett: Coping with skewed content popularity in mapreduce clusters, ” In Proceedings of the Sixth Conference on Computer Systems (EuroSys), 2011.

[4] 山崎公敬, 他, “広域分散データアクセスにおいてトラフィック集中を回避するメタデータ管理手法, ” IEICE ICM 研究会, Mar. 2017.

[5] 松田一仁, 他, “データの地理的分散管理のための分散 KVS 冗長化手法, ” IEICE ICM 研究会, July 2017.

[6] Q. Pu, G. Ananthanarayanan, P. Bodik, S. Kandula, A. Akella, P. Bahl, and I. Stoica, “Low Latency Geo-distributed Data Analytics, ” In ACM SIGCOMM, 2015.

[7] C. Ide, F. Kurtz, and C. Wietfeld, “Cluster-Based Vehicular Data Collection for Efficient LTE Machine-Type Communication, ” in Vehicular Technology Conference (VTC Fall), 2013 IEEE 78th, Sept 2013, pp.1-5.

[8] J. Pillmann, B. Sliwa, J. Schmutzler, C. Ide, and C. Wietfeld, “Car-To-Cloud Communication Traffic Analysis Based on the Common Vehicle Information Model, ” Proc. IEEE Vehicular Technology Conference (VTC-Spring) Workshop on Wireless Access Technologies and Architectures for Internet of Things (IoT) Applications, pp.1-5, June 2017.