博士論文

エッジ AI における訓練データ検索システムに関する研究

公立はこだて未来大学大学院 システム情報科学研究科
システム情報科学専攻

森 郁海

2022 年 3 月

Doctoral Thesis

Research on System of Training Data Searching on Edge AI

要旨

エッジ AI（Artificial Intelligence）は，セキュリティの問題や応答時間へのリアルタイム要求に応えるなどの目的で，データの生成元に近いエッジ上で機械学習や深層学習などの知的処理を実行するものである．

エッジ AI が適したユースケースは，工場の検品作業や人物行動分析，自動運転などのような，画像を入力として即時的な出力が要求されるものである．さらに，これらのユースケースでは，高度なセキュリティ対策やプライバシー保護が求められるため，クラウド上に訓練用の画像データを集約することが難しい．したがって，エッジ上で収集した画像データのみを訓練に利用することになるが，しばしば訓練データの不足が問題となる．

このような訓練データ不足を解消するために，高度なセキュリティ対策やプライバシー保護を実現しながら，AI の学習に有効な訓練データを検索するシステムが必要である．この訓練データ検索システムでは，まず，訓練データの暗号化したキーワード群と，その訓練データを所有するエッジの所在地を示す，暗号化した URI（Uniform Resource Identifier）からなるカタログ情報をクラウド上に集約しておく．次に，訓練データを検索するエッジは，クラウド上のカタログ情報の暗号化されたキーワードに対して検索を行い，訓練データの候補となるデータの URI を得る．そして，訓練データを検索するエッジは，URI に示されたエッジにアクセスし，データの利用許諾を取得した後，データをダウンロードする．最後に，訓練データを検索するエッジは，ダウンロードした画像データが訓練データとして有効かどうかを判定する．

以上のような訓練データ検索システムを実現するためには，クラウドにおいてスケーラビリティとセキュリティを維持しながら暗号化されたキーワードどうしを検索する技術と，エッジにおいて訓練データとして有効な画像かどうかを判定する技術の開発が必要である．本論文では，クラウドとエッジは，担保すべきセキュリティ対策とプライバシー保護のレベルが異なり，処理も分離可能であることから，これらの技術開発を独立した 2 つの問題ととらえ，各々の問題を解決することで訓練データ検索システムが構築可能であることを示す．

1 つ目の問題である，暗号化されたキーワードどうしの検索を高速に実行する技術については，高度なセキュリティ対策とプライバシー保護を実現するために，キーワードに加え検索処理自体もクラウドから秘匿する検索可能暗号を用いる．そのうえで，スケーラビリティを確保するために，検索速度とセキュリティを考慮した検索可能暗号の高速化パラメータを自動的に決定する手法を提案する．従来は，高速化パラメータの設定値をシステム管理者が手動で決定していたが，提案手法は，検索に使用するキーワードの最小エントロピーとk-匿名性を用いて，検索性能とセキュリティがバランスするパラメータ値を自動的に求める．数十万キーワード規模での評価の結果，提案手法によって検索時間を最大 97.2%削減しつつ，データベースが 2,598-匿名性を持ち，検索速度を向上させながら高いセキュリティレ

ベルを達成していることを確認した.

2 つ目の問題である，訓練データとして有効な画像かどうかを判定する技術については，エッジの限られた計算資源で実行できるよう，特徴点マッチングをベースとした類似画像検索を用いる．そのうえで，訓練データとして有効な画特を判定できるような，特徴点マッチングの類似度指標を提案する．従来のユークリッド距離などの単純な類似度指標は，照明変動や画像に占める被写体の割合の違いなどの環境ノイズを含む画像を検索しにくい．そこで，提案方式は，このような環境ノイズに対して不変性を持たせるために，画像のヒストグラムの形状に着目して類似度を計算する．具体的には，ヒストグラムの類似度計算において，ヒストグラムの形状が平行移動したり，伸縮したり，相似形である場合でも類似度が高くなるように,類似度評価区間を極値で分割し,区間ごとに DTW（Dynamic Time Warping）距離を求め，各距離を結合することで類似度を得る．画像認識でのユースケースを想定した評価において，提案手法は，ヒストグラムの形状に平行移動や伸縮，相似形が存在する画像どうしの類似度を高く算出できることを確認した.

1 つ目の提案により，十分なスケーラビリティを持ち，高いセキュリティを維持しつつ，検索速度が向上する秘匿検索技術を確立した．2 つ目の提案により，環境ノイズに対して不変性を持つ訓練画像データを検索できるようになり，エッジが収集した,ドメインが似ている画像データを AI の学習に相互利用できる技術を確立した．これらの結果から，エッジ AI における訓練データ検索システムの実現課題である 2 つの独立した問題が解決され，システムが構築可能であることが示された.

Abstract

Edge AI (Artificial Intelligence) executes intellectual processing such as machine learning and deep learning on the edge close to the data generation source for both solving high-level security problems and responding to real-time demands.

Use cases suitable for Edge AI such as factory inspection work, human behavior analysis, and autonomous driving require immediate output when images are inputted. Furthermore, edges may not upload images for training to the cloud because these use cases require high-level security and privacy protection. Therefore, only small image data set collected by the edge is used for the training of AI, but the lack of training data is often a problem.

To solve this lack of training data problem, we propose a training data search system. This search system can find effective image data for the training of AI while maintaining high-level security and privacy protection.

In this search system, the data catalog consisting of the encrypted keywords associated with the image data and the encrypted URI (Uniform Resource Identifier) indicating the location of the edge that has the image data are uploaded to the cloud. Then, the search edge that had would like to search for image data sends encrypted search keywords to the cloud. The cloud compares the encrypted keywords in the data catalog on the cloud and the sent search keywords using special matching processing. The cloud returns URIs of the image data that is a candidate of the training data to the search edge if the keywords hit in the matching process. Next, the search edge accesses the edge indicated by the URI and downloads the image data after accepting EULA (End User License Agreement). Finally, the search edge determines whether the downloaded image data is valid as training data.

The training data search system needs two new technologies. One is to match between the encrypted keyword of the data catalog and the search keyword while maintaining large scalability and high-level security at the cloud; the other is to determine whether the image data is effective as training data at the edge. In this study, we show that the training data search system can be constructed by solving these independent problems. The reason why these problems can be independent is that the cloud and the edge have different levels of security and privacy protection, and the processing is also separable.

To take measures to cope with the first problem, we use searchable encryption to make sure high-level security and privacy protection. The searchable encryption can protect both the search process and the keywords of the image data from the malicious cloud administrator. Moreover, to ensure scalability, we propose a method to

automatically determine the acceleration parameter of the searchable encryption while maintaining high search performance and high-level security. The current method for determining the acceleration parameter requires setting the parameter manually by a system administrator. On the other hand, our proposed method for compromising between search speed and security finds an optimum parameter automatically using conditional minimum entropy of the keyword. When over 685 thousand keywords were stored in the cloud database, the search time was reduced to 97.2% compared to not using the acceleration parameter. At the same time, the database table had 2598-anonymity. This means the narrowed-down database table by the acceleration parameter includes at least 2,598 keywords. Therefore, the database table was narrowed down from over 685 thousand keywords to only 2,598 keywords using our proposed method while maintaining high-level security.

To take measures to cope with the second problem, we use a similar image search based on feature point matching so that it can be executed with limited computational resources of the edge. Moreover, we propose a new similarity metric for feature point matching that can find an effective image as training data. The conventional simple similarity metrics such as Euclidean distance is difficult to search for images that include environmental noise such as lighting fluctuations and differences in the proportion of the subject in the image. On the other hand, our new similarity metric uses the shape of the histogram of the image to give invariance against such environmental noises. Specifically, our new similarity metric evaluates data similarity using Dynamic Time Warping (DTW) with partition by peaks of the image histogram. In evaluation assuming the use case of image recognition, we confirmed that the proposed method can extract similar images from the source image set even if peak shifts and similar in peak shape exist when comparing image data.

With the first proposal, we have established a secure search technique that has large scalability maintains high-level security and high search performance. The second proposal made the edge possible to search for training image data that is invariant to environmental noises. Furthermore, the images that are found by our new similarity metric will be able to be used as good training data. These results had shown that the training data search system can be constructed by solving the two independent problems that are required to implement this search system for edge AI.

目次

図目次

表目次

第1章 序論

本章では，まず，エッジ AI（Edge Artificial Intelligence）における訓練データ検索システムの必要性を述べる．次に，エッジ AI で考慮すべき IoT（Internet of Things）でのセキュリティ対策とプライバシー保護の動向を紹介する．そして，エッジ AI がセキュリティ対策やプライバシー保護に有効であるが，訓練データが不足するおそれがあることを説明する．訓練データ不足に起因して学習モデルの品質低下が生じる問題があり，訓練データ不足を解消する手段として，訓練データ検索システムを用いた転移学習の導入が効果的であることを示す．最後に，研究目的と貢献，本論文の構成について述べる．

1.1 エッジ AI における訓練データ検索システムの必要性

この節では，エッジ AI の説明と，訓練データ検索システムの必要性，課題解決の方策について述べる．

エッジ AI[1]は，セキュリティの問題や応答時間へのリアルタイム要求に応えるなどの目的で，データの生成元に近いエッジ上で機械学習や深層学習などの知的処理を実行するものである[2][3].

エッジ AI が適したユースケースは，工場での検品作業[4][5][6][7]や人物行動分析[8][9]，自動運転[10][11][12][13]などのような，画像を入力として即時的な出力が要求されるものである．さらに，これらのユースケースでは，高度なセキュリティ対策やプライバシー保護が求められるため，クラウド上に訓練用の画像データを集約することが難しい[14][15][16][17][18]．したがって，エッジ上で収集した画像データのみを訓練に利用することになるが，特に，機密性が高いデータや生起確率が低い事象を扱う場合に，訓練データ不足に陥りやすい[19][20][21][22][23].

このような訓練データ不足を解消するために，高度なセキュリティ対策やプライバシー保護を実現しながら，AI の学習に有効な訓練データを検索するシステムが必要である．

本論文では，訓練データ検索システムをクラウドとエッジが協調する形態をとる．訓練データの検索範囲がエッジ内に限られると，訓練に有効な画像が存在しない可能性がある．そこで，広域に分散するエッジが持つ画像を検索対象とするために，画像にキーワード等のメタデータが付与されている前提で，画像のキーワードと画像の所在地を示す URI（Uniform Resource Identifier）をクラウド上にセキュアに集積しておく．訓練データを検索するエッジは，まず，クラウドを使って訓練データの候補となる画像をキーワードベースで絞り込む．次に，キーワードによって絞り込んだ画像が訓練データとして有効かどうかを判定するために，実際に画像をエッジからダウンロードし，エッジ内の訓練画像との類似性を評価する．

本論文の訓練データ検索システムは，クラウドを使用するため，画像のキーワードと URI に対するセキュリティ対策やプライバシー保護が要求される．とくに，プライバシー保護規

制に関しては，世界各国で法律の整備などが進められており，パーソナルデータの越境移転が発生する場合は注意が必要である[24][25]．日本を含む各国の個人情報保護の考え方の基礎になっている OECD8 原則[26]に従い，データの所有者からデータの利用許諾をとってから，データを利用しなければならない[27]．さらに，IoT の普及により，パーソナルデータを含む機微情報を扱う，金融や医療関係等のユースケースが増えており，データの暗号化だけでなく，データの処理過程の秘匿も求められている[28][29]．

これらのことから，以下の考え方に沿って，訓練データ検索システムを構築する．構築には，データカタログ[30]の概念を参考にした．

- データの所在と越境移転の有無を明確化するために，データ所有者とデータの利用者が直接データを授受する
- データの適正利用のために，データの利用者がデータの取得の前に，データ保有者からデータの利用許諾を取ることができる
- 悪意のあるクラウド管理者からデータを保護するために，データの暗号化とデータの処理過程の秘匿を行う
- エッジは信頼できるとみなし，画像を直接扱う処理はエッジ上で実行する

上記の考え方に沿った，訓練データ検索システムのシステム構成を図 1 に示す．訓練データ検索システムの処理フローを説明する．まず，訓練データの暗号化したキーワード群と，その訓練データを所有するエッジの所在地を示す，暗号化した URI からなるカタログ情報をクラウド上に集約しておく（図 1①）．次に，訓練データを検索するエッジは，クラウド上のカタログ情報の暗号化されたキーワードに対して検索を行い，訓練データの候補となるデータの URI を得る（図 1②）．そして，訓練データを検索するエッジは，URI に示されたエッジにアクセスし，データの利用許諾を取得した後（図 1③），データをダウンロードする（図 1④）．最後に，訓練データを検索するエッジは，ダウンロードした画像データが訓練データとして有効かどうかを判定する（図 1⑤）．

この訓練データ検索システムを実現するためには，エッジを横断する広い検索範囲への対応と品質の良い訓練データの収集の観点から，以下の新しい技術開発が必要である．

- クラウドにおいて，キーワード数に対するスケーラビリティとセキュリティを維持しながら，暗号化されたキーワードどうしを検索する技術
- エッジにおいて，訓練データとして有効な画像かどうかを判定する技術

なお，クラウドとエッジは，担保すべきセキュリティ対策とプライバシー保護のレベルが異なるため，疎結合にする方が良い．上記の技術課題は，クラウドとエッジのそれぞれで処理を分離でき，訓練データ検索システムの性能は，クラウドとエッジの直列処理の結果と等

しくなる．したがって，本論文では，上記の技術開発を独立した 2 つの問題ととらえ，各々の問題を解決することで訓練データ検索システムが構築可能であることを示すこととする．

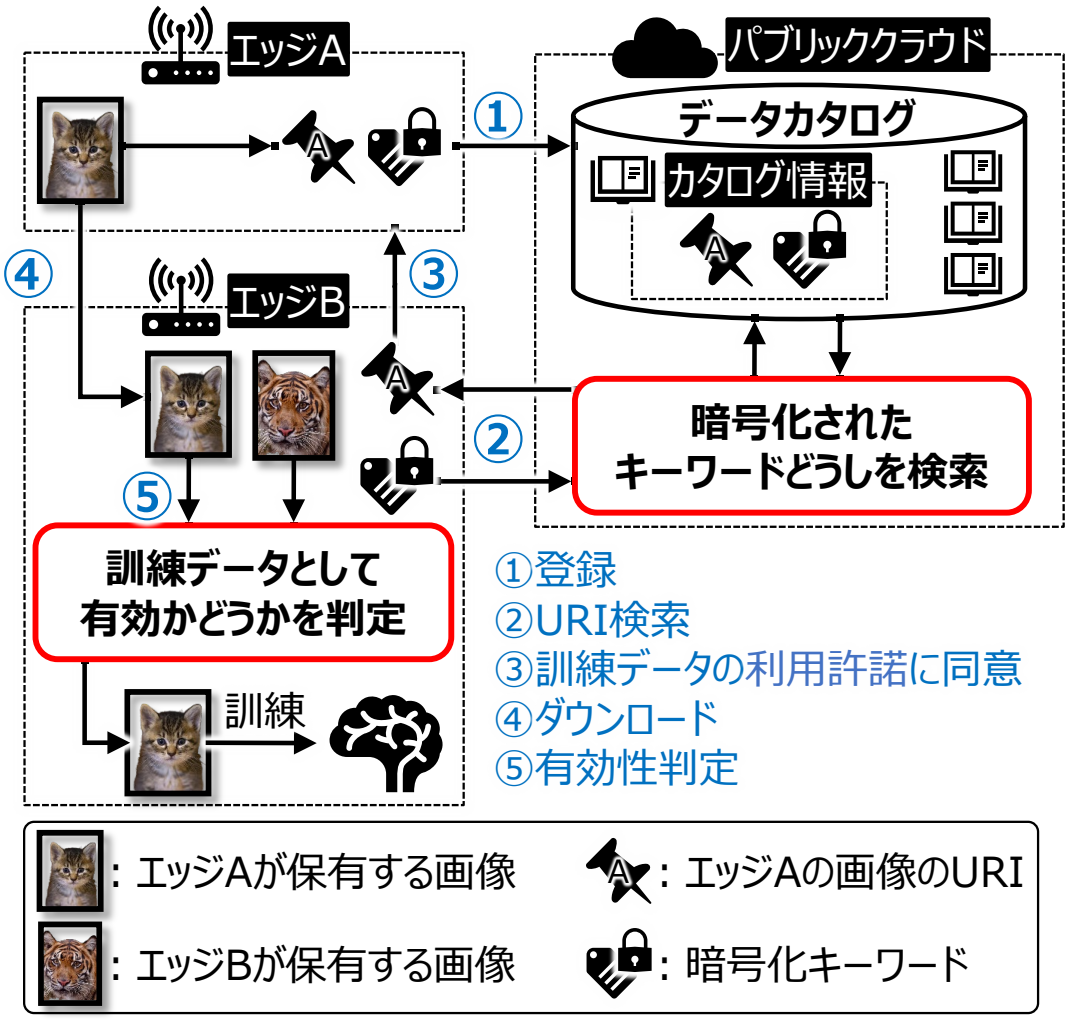


図 1 訓練データ検索システムのシステム構成

1.2 IoT でのセキュリティ対策とプライバシー保護の動向

1.1 節で述べた訓練データ検索システムでのセキュリティ対策とプライバシー保護のために，エッジやクラウド上に保管されているデータが，実際にどこに配置されているかを注意深く確認し，データを収集・利用する必要がある．さらに，訓練データは，パーソナルデータに該当しない場合でも，ビジネス上重要な資産である．そのため，訓練データやそのキーワードをエッジやクラウド上で扱う際には，各種ガイドラインや法律に従って，暗号化を施してデータを保護することや，データの所有者から利用許諾を取るなどの適切な契約を取り交わすことが必要である．この節では，IoT におけるセキュリティ対策とプライバシー

保護に関連する規格や法律について述べる．

エッジ AI は，1.1 節で述べたように，エッジデバイスに搭載された AI である．このエッジデバイスは，IoT デバイスを指すことが多い．よって，エッジ AI も IoT のセキュリティ対策とプライバシー保護に関する施策を順守する必要がある．

IoT は，様々な「モノ（物）」がインターネットに接続され，情報交換することにより相互に制御する仕組み，と定義される[31][32][33][34]．わが国では，「Society 5.0」[35]や「Connected Industries」[36]などの概念の中で，IoT の利活用方法が述べられている．

まず，内閣府は提唱する Society 5.0 は，「サイバー空間（仮想空間）とフィジカル空間（現実空間）を高度に融合させたシステムにより，経済発展と社会的課題の解決を両立する，人間中心の社会（Society）」と定義されている．具体的には，「IoT で全ての人とモノがつながり，様々な知識や情報が共有され，今までにない新たな価値を生み出すことで，これらの課題や困難を克服する」，「人工知能（AI）により，必要な情報が必要な時に提供されるようになり，ロボットや自動走行車などの技術で，少子高齢化，地方の過疎化，貧富の格差などの課題が克服される」とある．

次に，経済産業省が提唱する「Connected Industries」は，将来的に目指すべき未来社会である Society5.0 の実現のために，主に製造業を対象とした「データを介して，機械，技術，人など様々なものがつながることで，新たな付加価値創出と社会課題の解決を目指す産業のあり方」を表したものである[37]．Connected Industries の重点 5 分野とは，「自動走行・モビリティサービス」，「ものづくり・ロボティクス」，「バイオ・素材」，「プラント・インフラ保安」，「スマートライフ」である．さらに，Society5.0 が対象とする，ものづくり，自動運転，ヘルスケア，防災，農業などの分野のうち，Connected Industries 関連分野である，ものづくり，自動運転，ヘルスケアについては，データ社会推進協議会[38]において，分野横断的なセンサデータ等のデータ連携の仕組みを検討している[39]．このような分野横断的なデータ連携や，同一分野での他社間データ連携では，セキュリティ対策とプライバシー保護が必須である．

そこで，経済産業省は，Society 5.0／Connected Industries の進展に対応した「サイバー・フィジカル・セキュリティ対策フレームワーク（CPSF）」を策定している[40]．CPSF では，Society 5.0／Connected Industries におけるサプライチェーン全体のサイバーセキュリティ確保を目的として，産業に求められるセキュリティ対策の全体像を整理している．CPSF は，平時および緊急時のリスク管理・対応体制の構築と運用のほか，アクセス制御や暗号化によるデータ保護などの技術的な対策にも言及している．さらに，対策要件の一つである CPS.AC-9 において，プライバシーリスクへの対応を明記している．このように，セキュリティを維持するための組織体制の構築や，ルール作りなどのセキュリティ活動の手段を規定するものを，本論文では，「プロセスガイド」と表記する．一方，システムやコンポーネントが持つべきセキュリティ機能の要件を解説しているものを，本論文では，「対策ガイド」と表記する．したがって，CPSF は，プロセスガイドと対策ガイドの両方に該当す

る．

CPSF は主に国内を対象としたものであるが，世界的にも IoT におけるセキュリティ対策関連で様々な規格が存在する．従来の IT セキュリティで，古くからある世界的な標準として，ISO/IEC 27001[41]がある．

ISO/IEC 27001 は，組織の情報セキュリティマネジメントシステム（ISMS : Information Security Management System）を遂行するための要求事項を提供する．ISMS は，組織における情報資産のセキュリティを管理するための枠組みであり，情報の機密性，完全性及び可用性を維持し，かつ，リスクを適切に管理しているという信頼を利害関係者に与える．保護の対象となる情報には，個人情報などのパーソナルデータも含むこともあるが，保護のための施策を自主的に規定する必要がある．これらのことから，ISO/IEC 27001 は，セキュリティを維持するための組織体制の構築や，ルール作りなどのセキュリティ活動の手段を規定するため，プロセスガイドにあたる．

ISO/IEC 27001 をベースとして，業界別に規格が整備されている．たとえば，産業分野では，IEC 62443[42]，スマートグリッドでは，NISTIR 7628 Rev. 1[43]がある．

IEC 62443 では，産業用オートメーション及び制御システム（IACS：Industrial Automation and Control Systems）のセキュリティを確保するために開発されたもので，9 つの標準と，テクニカルレポート（TR），および，技術仕様（TS）が含まれている．ただし，セキュリティのみを対象としており，パーソナルデータの扱いには触れていない．セキュリティプログラムエレメント（SPE）として詳細を定義しており，具体的には，SPE 1 でセキュリティを維持するためのポリシー策定や組織体制の構築が規定され，SPE 5 にデータの保護，SPE 6 にユーザアクセス制御に関する記述がある[44]．したがって，IEC 62443 は，プロセスガイドと対策ガイドにあたる．

NISTIR 7628 は，米国国立標準技術研究所（NIST）が発行した，スマートグリッドのサイバーセキュリティに関するガイドラインである．NIST IR 7628 では，セキュリティ要件，リスク評価，スマートグリッドの設計に必要なツール，個人データを扱う際のプライバシーへの取り組みに必要なガイドなど幅広くカバーしている[45][46]．さらに，要員やポリシーの策定方法，リスクマネジメントにも言及している[47]．したがって，NISTIR 7628 は，スマートグリッド分野のプロセスガイドと対策ガイドにあたる．

クラウドに関わるセキュリティでは，産業別ではなく国別に整備されることが多く，米国では NIST[48]，欧州では ENISA[49]が発行するガイドが有名である．最近では，CPS（Cyber Physical Systems）分野のセキュリティが注目されており，例えば，2020 年に策定された NIST SP800-53[50][51]では，組織と情報システムのためのセキュリティおよびプライバシー管理策が述べられている．セキュリティに関する組織の管理体制と具体的な管理策が規定されており，プロセスガイドと対策ガイドの両方にあたる．

ENISA は，2020 年に IoT サプライチェーンの保護に関するガイドラインである，GUIDELINES FOR SECURING THE INTERNET OF THINGS Secure supply chain for

IoT[52]を公開している．このガイドラインでは，IoT サプライチェーン上のセキュリティリスクに対し，暗号化によるデータ保護や，プライバシーに関する法規制の順守などを述べている．よって，位置づけとしては，対策ガイドとなる．

これらのガイドでは，クラウド上のデータ秘匿やプライバシー保護を必須要件としている．様々な産業分野に対して IoT が適用されるにつれ，特に，金融分野やパーソナルデータを扱う医療分野においては，保管されたデータの秘匿だけではなく，データ処理の秘匿も必要とされる．

IoT のセキュリティに関しては，GSMA（Global System for Mobile Communications Association）[53]が公開している GSMA IoT Security Guidelines and Assessment[54]がある．このガイドラインでは，IoT サービスのリスク評価や，その評価結果を用いたセキュリティモデルの継続的な是正手段を定義している[55]．さらに，IoT デバイスから取得した画像の扱いにも触れており，具体的なセキュリティ対策やプライバシー保護にも言及している．

IIC（Industry IoT Consortium）[56]は，IISF（Industrial Internet Security Framework）[57]を公開している．IISF では，Industry IoT の信用度（trustworthiness）を安全性，信頼性，強靭性，セキュリティ，プライバシーの観点で評価している．IISF，セキュリティ等の機能要件を示したものであり，体制構築は対象としていない[58]．

GSMA IoT Security Guidelines and Assessment と IISF は，IoT システムやサービスにおけるリスクや信頼度の評価を対象としているため，対策ガイドとなる．

表 1 は，ここまでに述べたセキュリティ対策・プライバシー保護に関する主なガイドラインとカバー範囲をまとめたものである．ほとんどのガイドラインで，セキュリティ対策とプライバシー保護に関する記述があり，IoT において，この 2 つが重要であることが分かる．

表 1 セキュリティ対策・プライバシー保護に関する主なガイドラインとカバー範囲

ガイドライン	プロセスガイド	対策ガイド	セキュリティ対策	プライバシー保護	カバー範囲		
					システム	クラウド	デバイス
ISO/IEC 27001	✔		✔	✔	✔	✔	✔
経済産業省 CPSF	✔	✔	✔	✔	✔	✔	✔
IEC 62443	✔	✔	✔		✔		✔
NISTIR 7628	✔	✔	✔	✔	✔		
NIST SP800-53	✔	✔	✔	✔	✔	✔	✔
ENISA GUIDELINES FOR SECURING THE INTERNET OF THINGS		✔	✔	✔	✔	✔	✔
GSMA IoT Security Guidelines and Assessment		✔	✔	✔	✔	✔	✔
IIC IISF		✔	✔	✔	✔	✔	✔

プライバシー保護に関しては，その対策に法的根拠が存在する．この法的根拠は，国によって異なるため，主なものをピックアップして説明する．表 2 は，各国のプライバシー保護に関する法律をまとめたものである．法律が対象とするデータは，法律内で定義されているが，本論文では，パーソナルデータと表記する．詳細な定義は，各国の法律を確認すること．パーソナルデータの越境移転を許可していれば「○」を，許可していなければ「×」と記載している．パーソナルデータの利用許諾方法に関しては，オプトイン，オプトアウトの観点で規定があるかを記載する．オプトインとは，本人が事前許諾したパーソナルデータだけを第三者提供することである．オプトアウトとは，パーソナルデータを第三者提供するにあたって，そのパーソナルデータを持つ本人が反対をしない限り，第三者提供に同意したものとみなし，第三者提供を認めることである．

表 2 各国のプライバシー保護に関する法律

	プライバシー保護に関する法律	パーソナルデータの越境移転	パーソナルデータの利用許諾方法
日本	個人情報保護法	○ 本人の同意等が必要	原則オプトイン， オプトアウトは届け出が必要
EU	GDPR	× 原則禁止	原則オプトイン
米国	CCPA	○ 本人の同意，販売行為に対する義務あり	オプトイン（16 歳未満）， オプトアウトの権利を規定
中国	CS 法	○ データローカライゼーション，本人の同意を含む安全評価義務あり	原則オプトイン

わが国の IoT のプライバシー保護に関わる法律は，個人情報の保護に関する法律（個人情報保護法）である[59]．個人情報保護法は，時代に合わせて改正されており，IoT の普及によりプライバシー保護に対する規制が進んでいる[60]．現行の法律では，パーソナルデータの第三者提供などの目的外利用には，データ保有者の本人同意が必要である．パーソナルデータの越境移転は可能だが，前述のとおり本人同意を要求する．パーソナルデータの利用許諾方法は，原則オプトインであり，オプトアウトには，個人情報保護委員会への届け出が必要である．

EU，米国，中国にも同様の法律が存在する[25]．EU の GDPR（General Data Protection Regulation：一般データ保護規則）[61]は，パーソナルデータの取り扱いと欧州域外への越境移転を定めた法律である．パーソナルデータの越境移転は，原則禁止であり，パーソナルデータの利用許諾方法は，原則オプトインである．GDPR に違反すると巨額の罰金が科される．最も高額な罰金を科せられた事例は，2019 年 1 月にフランス当局が Google に対して 5000 万ユーロ（約 63 億円）の罰金を科したものである[62]．

米国では，CCPA（California Consumer Privacy Act：カリフォルニア消費者プライバシー法）[63]をはじめとする州法による規制が進んでいる[64]．CCPA は，カリフォルニア州の住人（消費者）のパーソナルデータが対象であり，パーソナルデータの所有者からの要求に応じて，パーソナルデータの販売を停止することを要求する．パーソナルデータの越境移転には，本人同意が必要であり，販売行為に対する義務を課せられる．パーソナルデータの利用許諾方法は，16 歳未満であればオプトイン，16 歳以上の消費者であれば，企業に対してオプトアウトする権利を規定している．現在，米国連邦データプライバシー法案提出され

ており，米国全土に効力を持つ法律が立法化されると予想されている．

中国のサイバーセキュリティ（CS）法は，セキュリティ対策，パーソナルデータの取り扱い，データ越境移転について定めた法律である[65]．GDPR に準じているが，独自の要求仕様も追加されている．とくに，中国でパーソナルデータを取り扱う場合には，現地にデータを保管する必要があり（データローカライゼーション），そのうえで海外に移転する場合には，定められた安全評価が課される．したがって，パーソナルデータの越境移転は可能であるが，データローカライゼーションと，本人同意を含む安全評価を実行する義務が発生する．パーソナルデータの利用許諾方法は，原則オプトインであるが，中国当局に強い権限が規定されているため，注意が必要である．

このように，各国の主要な法律は，パーソナルデータの越境移転を厳しく規制しており，違反時には巨額の罰金などの制裁が課される．したがって，エッジやクラウド上に保管されているデータの実際の配置を意識して，データの越境移転を最小限に抑えつつ，収集・利用しなければならない．さらに，訓練データは，ビジネス上重要な資産であるため，ここで挙げたガイドラインや法律に従って，訓練データやそのキーワードを暗号化して保護することや，データの所有者から利用許諾を取るというような適切な契約締結が必要である．とくに，新たに訓練データを取集する際や，他のエッジで収集したデータやクラウドに集積されているデータを訓練データとして利用する際は，これらの点に配慮しなければならない．

1.3 エッジ AI の重要性

この節では，一般的な AI 処理の流れと，エッジ AI がセキュリティ対策やプライバシー保護などの問題解決に有効であることを述べる．

IoT の普及により，モノの情報の可視化や，モノの制御を行うソリューションが増加[66]したことで，セキュリティの問題や応答時間へのリアルタイム要求などに応える必要が出てきている．前記のソリューションでは，AI（Artificial Intelligence）処理と呼ばれる機械学習や深層学習などの知的処理を扱うものがある．一般的な AI 処理の流れを，図 2 に示す[67][68]．

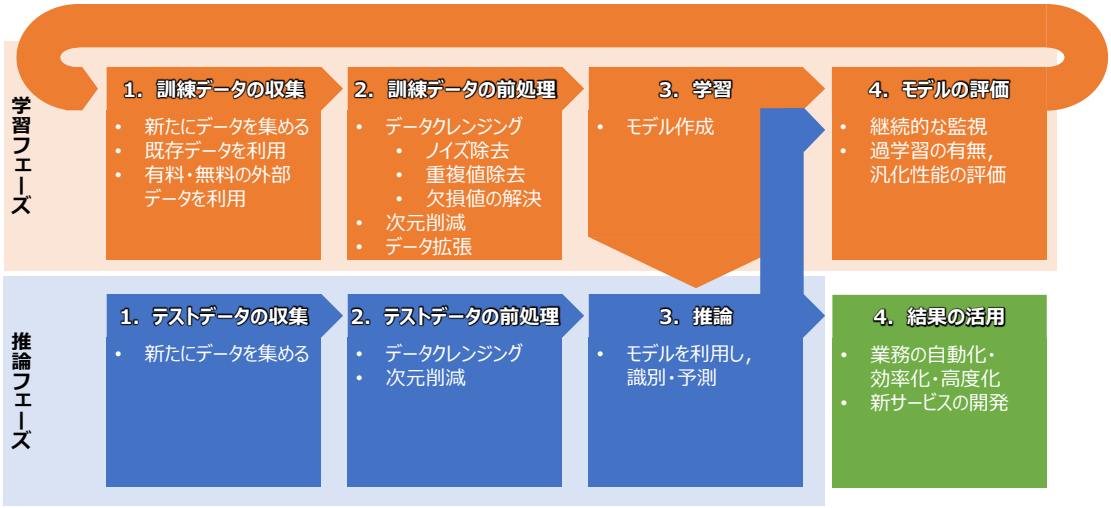


図 2 AI 処理の流れ

AI 処理は，学習フェーズと推論フェーズからなり，学習フェーズでは，訓練データの収集，訓練データの前処理，学習によるモデルの作成，モデルの評価を行う．モデルの更新の有無は，モデルの評価の結果によって決定する．モデルの更新サイクルは，モデルを利用するソリューションによるが，1 日から数か月と幅広い．とくに，高度な安全性を求める自動運転などの制御系ソリューションは，更新後のモデルを使用したときに危険な事象が発生しないかを確かめるために，モデルの評価に時間がかかることが多い．一方，推論フェーズでは，テストデータの収集，テストデータの前処理，モデルを利用した推論を行う．推論により得られた結果は，業務の自動化・効率化・高度化や，新サービスの開発などに活用される．さらに，推論の結果は，学習フェーズのモデルの評価にも用いられることがある．

訓練データの前処理では，訓練データ不足を解消するために，データ拡張を行うことがある．データ拡張とは，画像に対して，移動や回転，ノイズ付加等の人為的な処理を加えることによって，学習に用いる画像の数を増やすことである[69]．ただし，データ拡張を行ったとしても，深層学習のような大量の訓練データを必要とするアルゴリズムでは，過学習が生じるリスクがある[70][71][72]．

訓練データが不足する問題に対して，少ない訓練データでモデルを作成する方法が存在する．少ない訓練データでモデルを作成する手法として，敵対的特徴生成[73]や Few-shot learning[74]などの技術が存在する．敵対的特徴生成は，深層ネットワークの中間層で得られる特徴量を意図的に変化させることで，認識が難しい学習データを人工生成するものである．この処理は，学習内部で実行されるため，訓練データを増やす必要なしに，過学習を防ぎながら推論精度を向上させることができる．Few-shot learning は，少ない画像で過学習を抑えながら品質の良いモデルを作成する技術の総称で，特に，教師データとして 1 枚の画像サンプルのみを用いるものを One-shot learning，教師データを全く用いずに未知の画像を識別するものを Zero-shot learning と呼ぶ．しかしながら，これらの技術は，ほとん

どが深層学習向けのものであり，事前に類似ケースで学習をした学習済みモデルをベースとして用いることが多い．すなわち，セキュリティやプライバシーの問題で事前に学習済みモデルを用意できないユースケースには，適用が難しい．

AI 処理は，従来，クラウド上で実行されることが多かったが，モノの増加[66][75][a]によって AI 処理の計算負荷が高まり，一極集中のクラウドコンピューティングだけでは対応が難しくなっている．クラウドコンピューティングを用いて AI 処理を実行する形態は，クラウド AI と呼ばれる（図 3）．クラウド AI では，AI 処理における学習と推論をクラウドサーバ上で実行する．クラウドの豊富な資源を利用できるため，複雑で大規模な AI アルゴリズムを実行でき，AI による分析を使用したデータマイニングなどに適している．一方で，一極集中によりクラウドのエネルギー消費が増大する問題[76]や通信遅延に起因するデータ遅延の問題[77], 1.2 節で述べたセキュリティ対策やプライバシー保護の問題などがある．そのため，工場での検品作業や人物行動分析，自動運転などの IoT デバイスやエッジで AI の処理の結果を使用するユースケースに対してクラウド AI を使用すると，リアルタイム性を保てず，サービスを提供できないおそれがある．

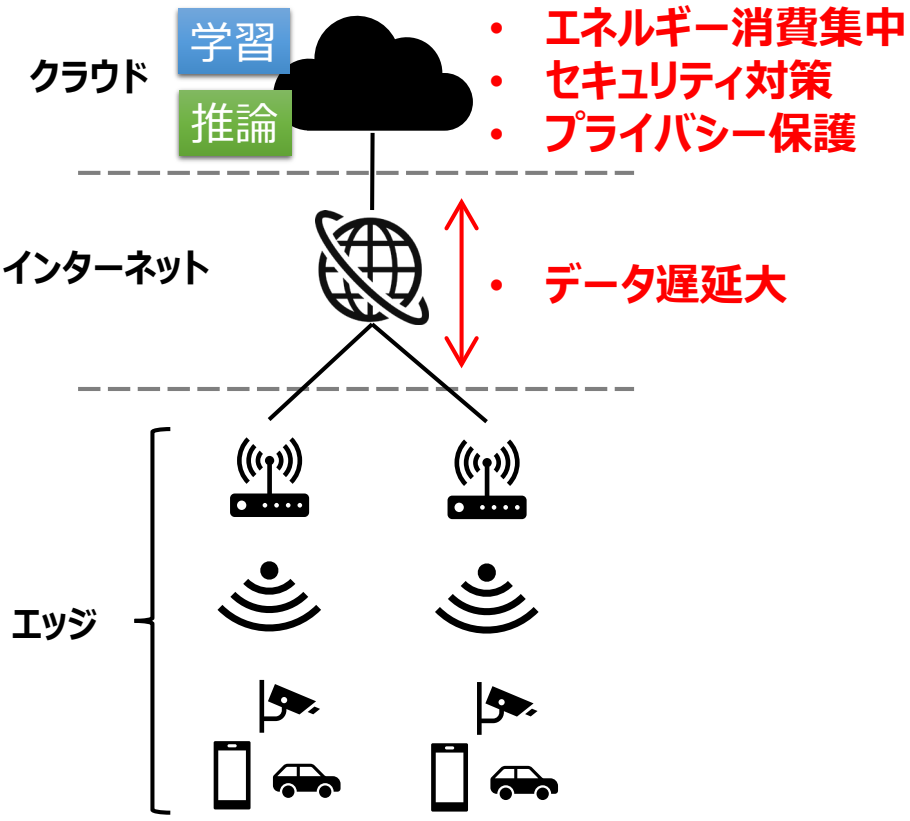


図 3 クラウド AI の実行形態と問題

a) 世界の IoT デバイス数は，2021 年時点で 348.3 億台に上り，特に，「医療」，「産業用途」，「コンシューマ」及び「自動車・宇宙航空」で高成長が見込まれている．

そこで，従来のクラウド AI では対応できないユースケースに対応するために，エッジ AI が注目されている．エッジ AI は，AI の処理結果を利用するデバイスの近傍で，AI 処理を実行する形態である．1.2 節で述べたように，訓練データの越境移転は最小限に抑える必要がある．しかし，クラウド AI では，エッジ上で取得したデータを学習に使用するか否かにかかわらず，クラウドに集約するため，データの越境移転が発生しやすい（図 4 左）．一方，学習をエッジ上で実行するタイプのエッジ AI は，訓練データがエッジに留まるため，データの越境移転を最小限に抑えることができる（図 4 右）．

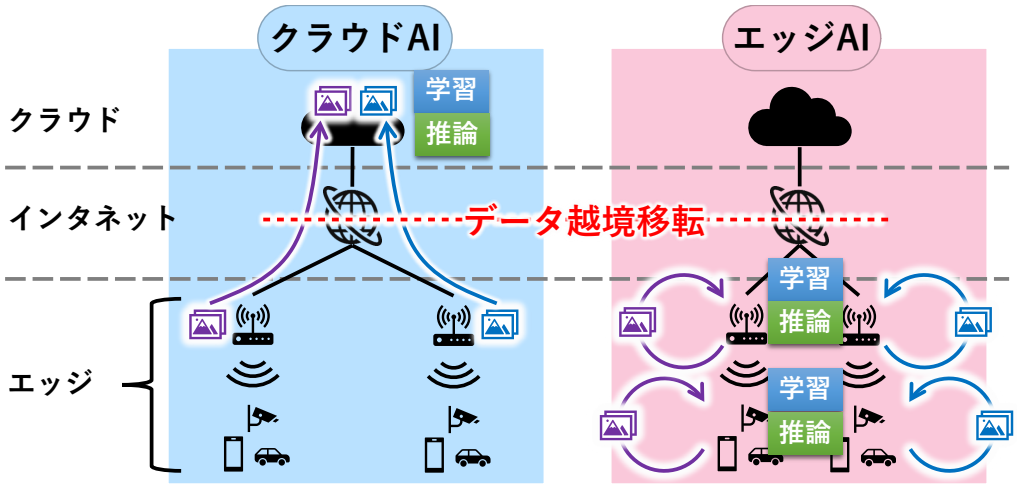


図 4 データの越境移転の発生

エッジ AI には，計算負荷の高い学習をクラウド側で実行し，クラウドで作成したモデルをエッジに送信して，エッジ上で推論を実行する形態（図 5 (A)）と，学習も推論もエッジ上で実行する形態（図 5 (B)）がある．

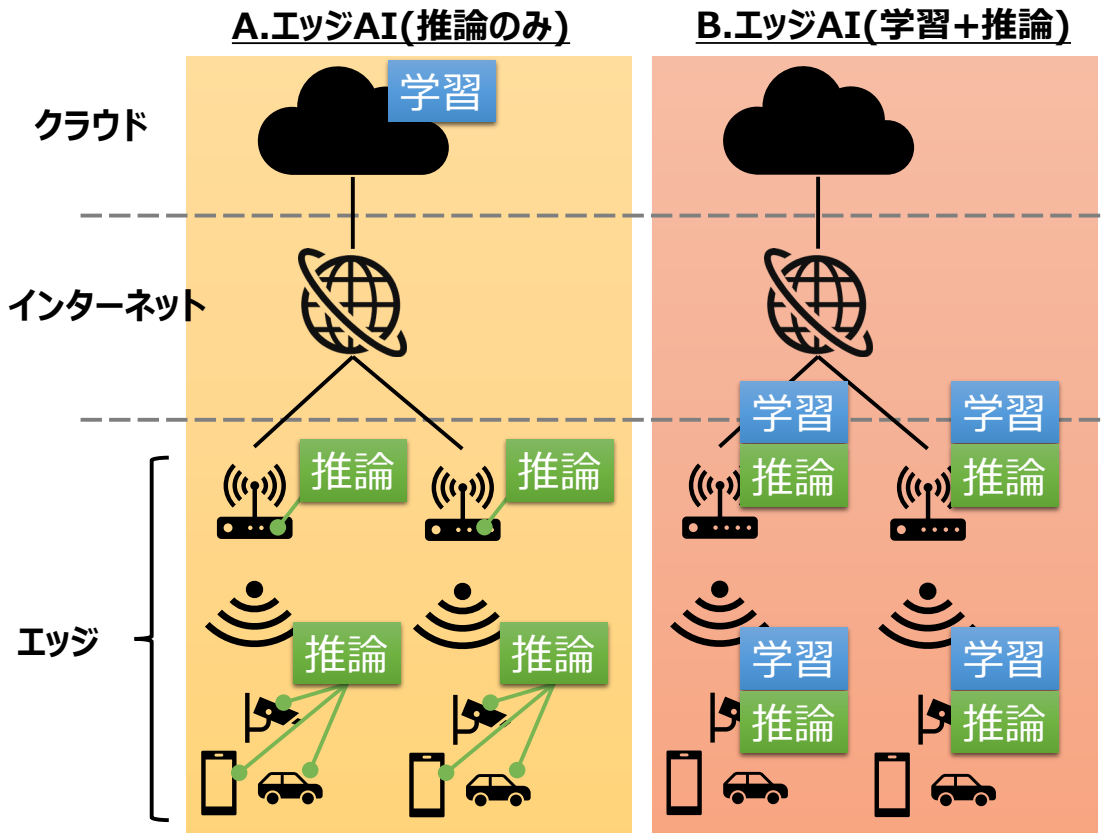


図 5 エッジ AI の実行形態

エッジ上で推論のみを実行するエッジ AI は，エッジの限られた計算資源の中で複雑な AI アルゴリズムを用いた推論結果を低遅延で利用できる．しかしながら，クラウドで学習を行うため，デバイスで取得したデータをクラウドへアップロードする必要があり，通信帯域が過剰に消費されるとともに，データの越境移転が発生する可能性がある．さらに，セキュリティの問題でクラウドへデータをアップロードできないことがある．加えて，クラウドでアップデートしたモデルが，改ざんされることなく宛先のエッジに正しく配送する仕組みも必要になる．この形態は，複雑で高度な予測が低遅延で求められる，工場内のデバイスの制御や故障予測などの製造系の AI や，株価のトレンド予測などの金融系の AI，がけ崩れの監視などの防災系の AI に適している．

エッジ上で学習と推論の両方を実行するエッジ AI は，すべての AI 処理がエッジ内部で完結するため，セキュアかつ低遅延で通信帯域の消費も少なく，モデルのアップデートの問題もない．ただし，限られた計算資源のエッジ上で学習を行うため，小規模な AI アルゴリズムが対象となる．ビデオカメラが使用される自動運転やドライバーの状態監視，行動認識や顔認証などのプライバシー保護が必要で，かつ，個別最適が要求されるユースケースに向いている．

エッジ AI には，エッジ上で推論だけ行うものと，推論に加えて学習も行うものがあると

述べた．より詳細には，エッジ AI は 6 つのレベルに分けられる[1]（図 6）．レベル 1 は，クラウドとエッジの両方で推論を実行するものである．レベル 2 は，エッジで推論を実行するものである．レベル 3 はエッジでもよりデバイスに近いところで推論を行うものである．レベル 1～3 では，クラウド上で学習が行われる．レベル 4 はクラウドとエッジの両方で学習を行うものである．たとえば，深層学習では，デバイスで共通に使用するモデルはクラウドで作成し，デバイス固有の特徴獲得は，デバイス側で行わせる．この処理をファインチューニングと呼び，これがレベル 4 に当たる．レベル 5 は，エッジで学習と推論を行う形態で，本研究の対象とするものである．最後にレベル 6 は，デバイス上で学習と推論を行うものである．レベルが上がるにつれ，ネットワークに流れるデータ量は少なくなり，通信距離も短くなる．一方，レベルが上がるにつれて計算資源が小さくなるため，AI アルゴリズムも複雑なものは使用できなくなる．

文献[1]によれば，エッジとクラウドの協調利用によって，低遅延と消費電力削減の両方を達成できる．観点(2)で触れたように大量の訓練データをクラウドへ送信することによる通信帯域の消費の問題や，問題(3)で述べたセキュリティの問題によって，訓練データをクラウドに上げられない場合は，エッジまたはデバイス上で学習を行うレベル 5 以上が対象となる．

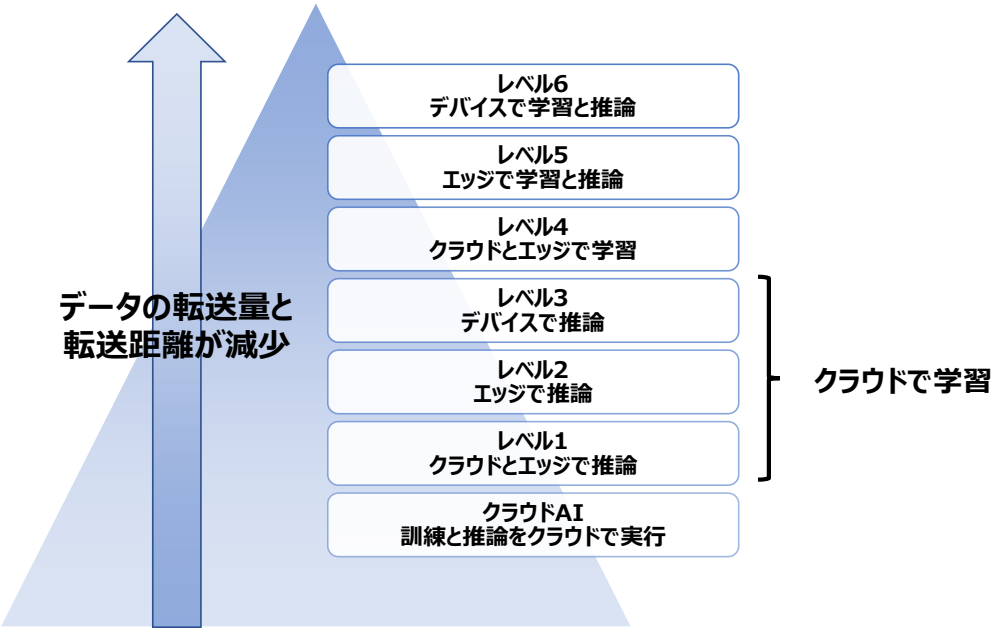


図 6 エッジ AI のレベル分け[1]

エッジで推論に加え学習も行う場合は，分散型とハイブリッド型の 2 つのアーキテクチャを使用する（図 7(B)，(C)）．

分散型は，エッジデバイスが個々に収集した訓練データを基に，ローカルでモデルを作成

し，必要に応じて他のエッジのローカルモデルを交換して，自身のローカルモデルを更新するアーキテクチャである．この分散型は，図 6 のレベル 5～6 に該当する．

ハイブリッド型は，クラウドで学習を行う集中型（図 7(A)）と分散型（図 7(B)）を結合させたものである．分散型においてエッジデバイスのみであったローカルモデルの交換対象に，クラウドを加えたものである（図 7(C)）．

表 3 に AI の実行形態の得失の一覧を示す．

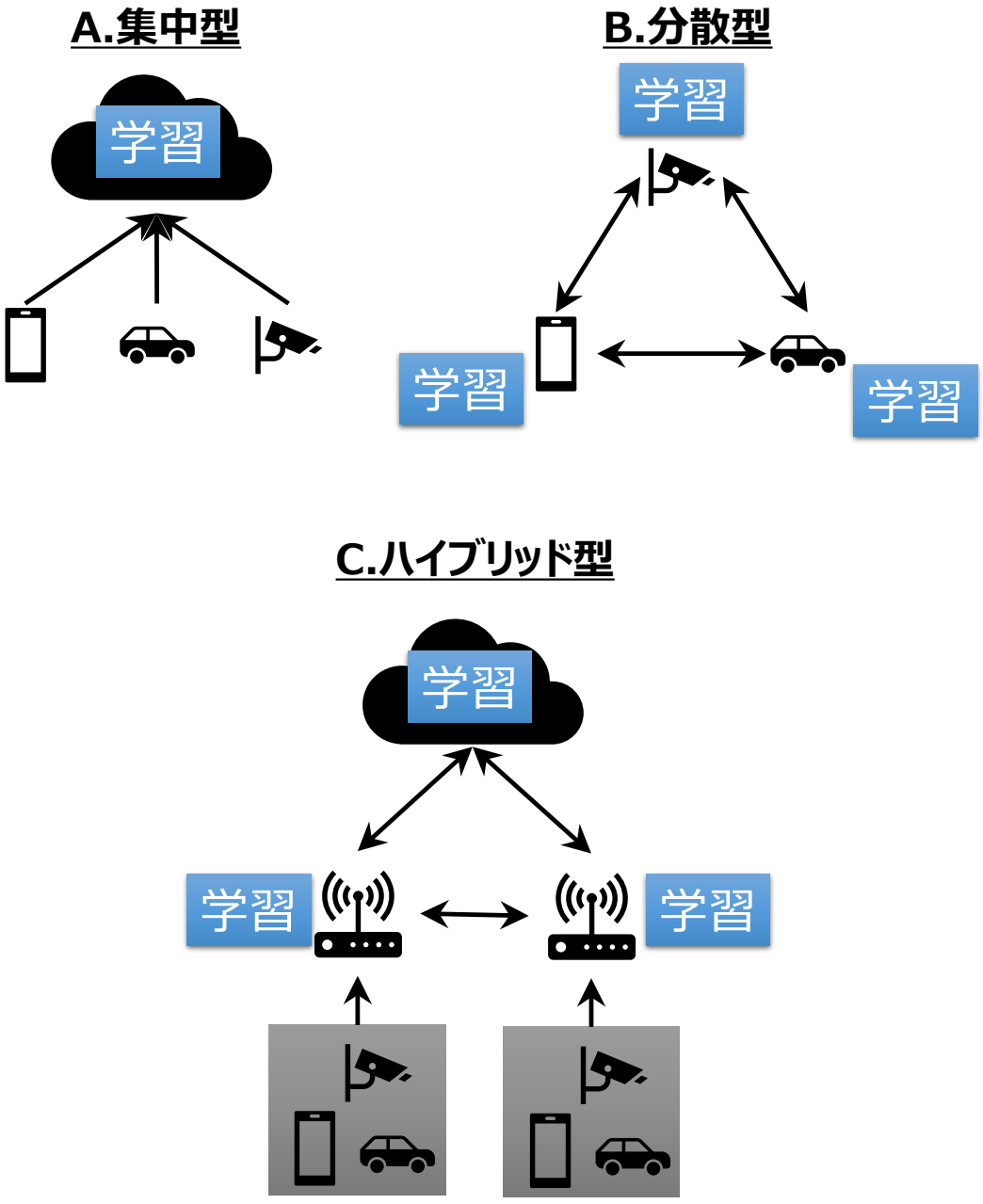


図 7 分散学習のアーキテクチャ[1]

表 3 AI の実行形態の得失

	クラウド AI	エッジ AI	
		推論のみ	学習＋推論
概要	• 学習と推論をクラウドサーバ上で実行	• 学習をクラウド側で実行し，クラウドで作成したモデルをエッジに送信して，エッジ上で推論を実行する形態	• 学習も推論もエッジ上で実行する形態
主なユースケース	• データマイニング	• 工場内のデバイスの制御や故障予測などの製造系の AI • 株価のトレンド予測などの金融系の AI • がけ崩れの監視などの防災系の AI	• 工場の検品作業 • 自動運転 • ドライバーの状態監視 • 人物行動分析 • 顔認証
利点	• 複雑で大規模な AI アルゴリズムを実行可能	• 複雑な AI アルゴリズムを用いた推論結果を低遅延で利用可能	• パーソナルデータが扱える • 低遅延 • 通信帯域の消費少 • モデルの改ざん・誤配送のリスクなし
欠点	• データアップロードに伴う通信帯域の枯渇 • アップロードデータのセキュリティの問題 • エネルギー消費が増大 • データ遅延が発生	• データアップロードに伴う通信帯域の枯渇 • アップロードデータのセキュリティの問題 • モデルの改ざん・誤配送のリスク	• 小規模な AI アルゴリズムが対象

エッジAIの実現を支援するサービスとして，Microsoft, AWS（Amazon Web Services），GCP（Google Cloud Platform）は，クラウドとエッジが連携する Azure IoT Edge[78]，AWS IoT Greengrass[79]，GCP Cloud IoT Edge[80]等のサービスを提供している．これらのサービスを使用するには，以下の 3 つの観点を考慮する必要がある[1].

(1) エッジの限定された計算資源の活用
(2) エッジで判断することによるリアルタイム性の確保
(3) 異なるセキュリティ要件を持つクラウド・エッジの活用

まず，観点(1)を考慮することで，計算資源に制約のあるエッジ上で AI アプリケーションを実行する際の AI の性能低下を防止することができる．これを考慮しない場合，多くの AI アプリケーションは，エッジの計算能力を上回る計算能力を要求するため，AI の性能が低下する可能性がある．

次に，観点(2)を考慮することで，クラウドコンピューティングでは達成できなかったリアルタイム性を確保できる．制御系システムの多くは，入力を得てから出力（アクチュエート）するまでの時間に制限があり，これをリアルタイム性と呼ぶ．クラウド上で AI アプリケーションを実行する形態では，AI の利用のための大量の画像データなどがクラウドへ送信される．そのため，通信帯域を過剰に消費し，それに伴って通信速度が不安定になったり，通信時間が処理時間の大半を占めたりする影響で，通信遅延以下のリアルタイム性を求めるユースケースには対応できない．一方，エッジ AI は，通信遅延の影響をほとんど受けないため，AI の実行に時間を割り当てることができる．

最後に，問題(3)を考慮することで，データの機密性に応じた適切な処理を実現することができる．顧客データや産業上重要な制御データを AI に利用する場合，パブリッククラウド等の完全に信頼できない環境下に機密データを保管できないため，クラウド上で AI を実行することができない．機密データを保護するために，クラウド上で暗号化や復号処理を行うと，処理が増加し，リアルタイム性を守ることができなくなるおそれがある．一方，エッジは，データの生成元に近く，信頼できる顧客側が管理することも多いため，機密データを危険にさらすことなく AI 処理を実行できる．

エッジ AI における訓練データ検索システムでは，クラウドとエッジを次のように使い分ける．クラウドは，広域に点在するエッジ群が保有する訓練データに関するキーワードを集約し，訓練データの候補となる画像を持つエッジを大まかに絞り込むために利用する．この処理により，エッジ単独で広域に点在する転移学習の対象となる訓練データを発見するより，エッジの処理負荷を軽減することができ，観点(1)で述べた計算資源の制約に対応することができる．このとき，クラウド上の訓練データのキーワードと検索処理過程を悪意のあるクラウド管理者から保護するために，1.2 節で述べたようなガイドラインや法律に従って，暗号化などのセキュリティ対策を実施する．一方，エッジは，クラウドで絞り込まれた訓練

画像の候補を保有するエッジから画像をダウンロードし，訓練データとして有効な画像かどうかを判定するために利用する．この処理により，データの越境移転の有無を把握し，1.2節で述べたようなガイドラインや法律に従って，データの所有者から利用許諾を取ることができ，かつ，ダウンロードした画像を正規の利用者以外が利用できないよう物理的に隔離することができる．

1.4 エッジ AI での訓練データ不足に起因するモデルの品質低下

この節では，エッジ AI の利用で問題となる訓練データ不足について説明し，この訓練データ不足によってモデルの品質が低下する恐れがあることを述べる．

分散型や，ハイブリッド型のエッジ AI を利用する場合，集中型に比べて各エッジで使用する訓練データの収集範囲が狭いため訓練データが不足し，ローカルモデルが十分な推論精度を持たないことがある．

エッジ AI でモデルを作成する具体的な研究例として，Few-shot learning の考え方をベースとした FedHealth[81]がある．FedHealth は，ウェアラブルヘルスケアデバイスを対象として，プライバシー保護の関係上データがクラウド上で共有できない場合に，ユーザで共通に使用するモデルと，エッジ上の個別のユーザ環境で取得したデータから作成したモデルを基に，新たなモデルを構成するフレームワークである．エッジ上のデータからモデルを作成するため，エッジで学習に必要な訓練データ量を収集する必要がある．しかしながら，たとえば，身体的特徴や日常の活動パターンが似た人物 A，B の宅内カメラ画像で心拍数や行動認識による健康管理を想定する場合，A の環境下では学習に必要な画像が十分あるが，B の環境下では十分存在しない状況が考えられる．この状況下では，B のモデルを作成するための訓練データが足りず，十分な推論精度が出せない可能性がある．

図 8 は，IoT の構成機器別のデータ収集範囲を示したものである．分散型では，デバイスを束ねるゲートウェイやデバイス単位でモデルを作成する．そのため，ゲートウェイであれば，そのゲートウェイに接続するデバイスのデータが収集対象となる．一方，デバイスで学習する場合は，主にそのデバイスで測定したデータしか利用できない．クラウドが，そのクラウドに接続するすべてのゲートウェイとデバイスからデータを取得できると考えると，ゲートウェイやデバイスが収集するデータ数が圧倒的に少ないことが分かる．

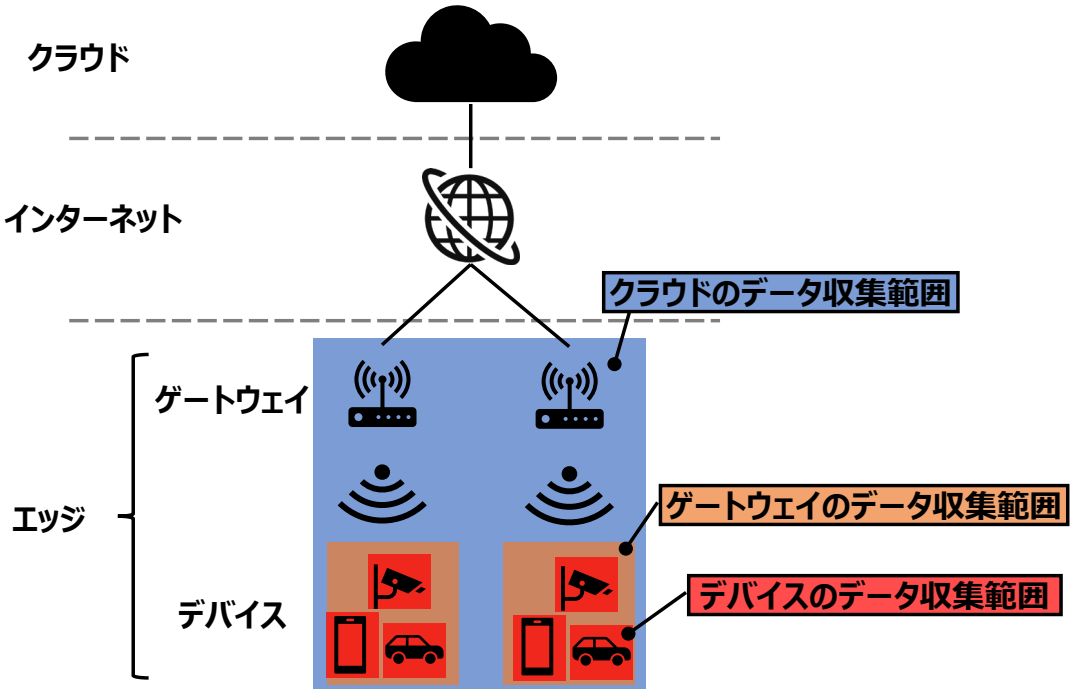


図 8 IoT の構成機器別のデータ収集範囲

収集可能なデータが少ないと，訓練データの候補となるデータも比例して少なくなり，品質の良い訓練データセットの確保が困難となる．品質の良い訓練データセットとは，発見したい事象（以降，識別クラス，あるいは単に，クラスと表記）のサンプルデータを満遍なく含む状態である．たとえば，工場の検品処理を考えると，良品と不良品を識別するモデルを作成したい場合，良品が写っている画像と不良品が写っている画像を 1:1 に近い割合で含む訓練データセットが，品質が良いということになる．

一方，事象のデータ数に偏りがある不均衡な訓練データセットを使用すると，モデルの精度が低下する．一般的にモデルの評価には，ホールドアウト検証やk分割交差検証がよく使用される[82][83][84]．ホールドアウト検証は，初期のデータセットを訓練用と推論用の 2 つに分割して評価する．訓練用のデータセットと推論用のデータセットの要素は，初期のデータセットから無作為に抽出する．推論用のデータセットは，訓練用のデータセットよりも小さくすることが多い．k分割交差検証は，まず，初期のデータセットをk個のグループに分割する．次に，k個のグループのうちの 1 個のグループを推論用のデータセットに，残りの$k-1$個のグループを訓練データセットとして，モデルの出力を得る．同様に，各グループ一回ずつテストデータセットとなるように，合計k回モデルの出力を得る．このようにして得られたk個のモデルの出力値を平均して，モデルの最終的な評価値とする．図 9 は，工場の検品処理における，訓練データの品質の違いによるモデルの評価への影響を示したものである．

良品と不良品の画像を 1:1 の割合で含む，品質の良い訓練データセットを使用して，ホー

ルドアウト検証を行う場合，訓練データセットとテストデータセットの中に，満遍なく良品と不良品の画像が含まれることが分かる．同様に，k分割交差検証を行う場合，2 回目と 3 回目のテストデータセットに良品または不良品の画像のみを含むパターンがあるが，訓練データセットには少なくとも 2 つ以上の良品または不良品の画像が含まれることが分かる．言い換えると，モデルの学習において，どのようなパターンでも少ない方の画像の割合が 1/3 を下回ることがない．これは，モデルが良品と不良品の識別を行う上で，どちらかのサンプル画像が全くなく，学習できない状態を回避できることを意味する．

一方，良品と不良品の画像を 7:1 の割合で含む，品質の悪い訓練データセットを使用してホールドアウト検証を行う場合，図 9 の例では訓練データセットにしか不良品の画像が含まれないことが分かる．つまり，この訓練データセットで学習したモデルは，常に良品と出力すれば，識別精度が 100%の良いモデルと判断されてしまう．このように，品質の悪い訓練データセットを使用すると，未知の入力値に対する識別性能（汎化性能）を正しく判定できない問題がある．同様に，k分割交差検証を行う場合，1〜3 回目の推論においてホールドアウト検証の同じ問題が生じる．4 回目の推論においては，訓練データセットに良品データしか含まれておらず，不良品を学習できない．言い換えると，このモデルは出力として不良品が定義されない状態となり，テストデータセットに含まれる不良品画像が未知の入力となってしまう．通常，モデルで学習しなかった未知の入力は，学習済みの既知のクラスに分類されることが多い．この例では，不良品であるにもかかわらず良品と判断される可能性が高い．

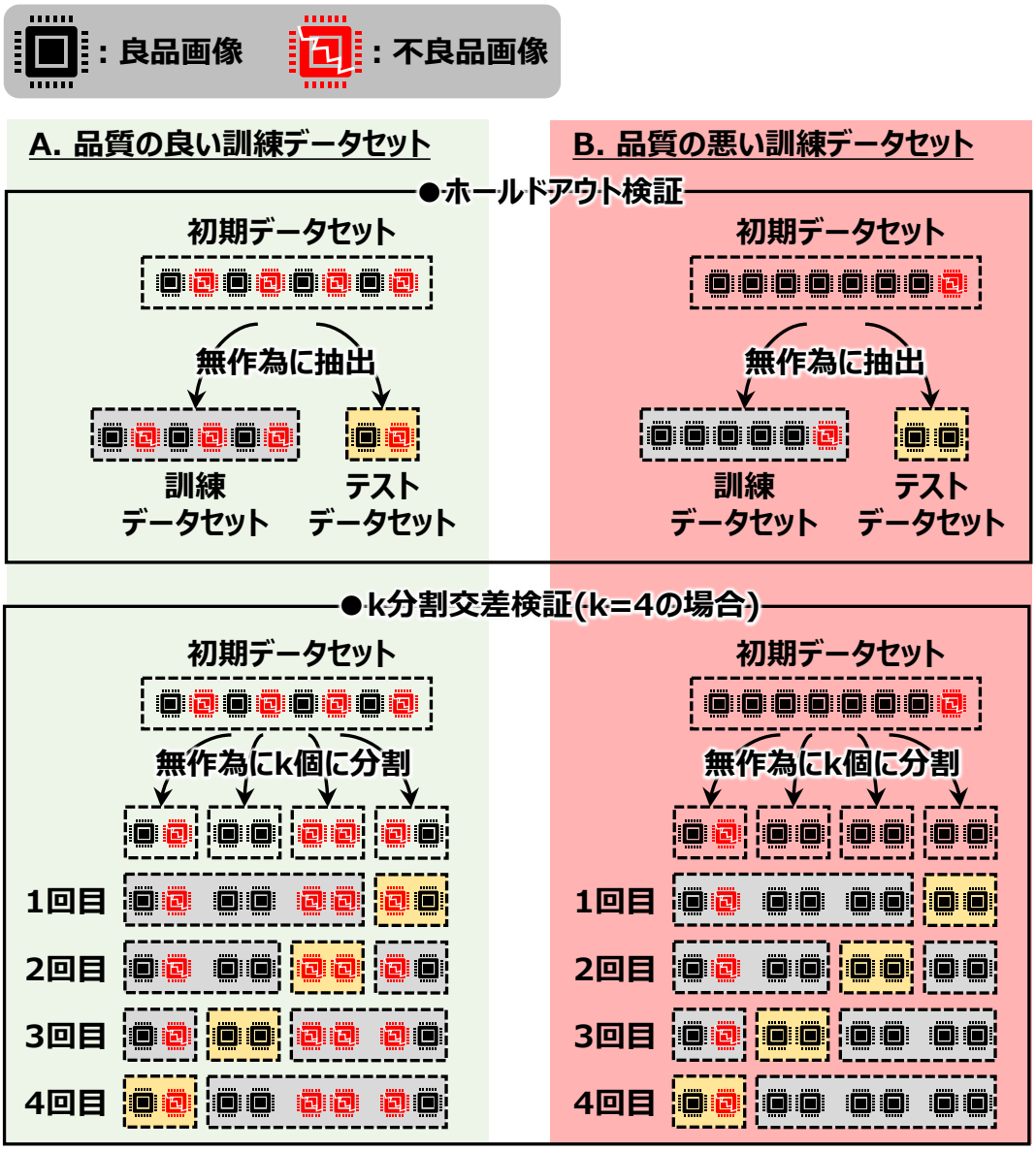


図 9 訓練データの品質の違いによるモデルの評価への影響

識別精度が高い品質の良いモデルを作るためには，クラスを満遍なく含む訓練データセットを用意することのほか，クラスバランス変化と共変量シフトに注意する必要がある[85].

クラスバランス変化は，訓練データセットに含まれる各クラスの訓練データの数と，テストデータセットに含まれる各クラスの訓練データの数が異なる状態を指す．図 10 は，クラスバランス変化が発生した際のモデルの推論精度への影響を示したものである．図 10 のクラスバランス変化がない状態では，訓練データセットとテストデータセットに同じ割合の良品画像と不良品画像が含まれている．このクラスバランス変化がない状態でモデルを作成する場合は，モデルが良品か不良品かを推論精度する割合にバイアスがかからない．一方，クラスバランス変化がある状態では，訓練データセットには，良品：不良品＝3：1 の割合で良品画像と不良品画像が含まれるが，テストデータセットの良品と不良品の割合は良品：不良品＝1：3 となっている．この状態で，訓練データセットを用いて訓練を行うと，良品を多く推定するようなバイアスがかかってしまう．このように作られたモデルを使用して先ほどのテストデータセットに対して推定を行うと，不良品が多いにもかかわらず，モデルは「不良品は稀にしか現れない」と認識して，不良品を良品と判断してしまうことがある．すなわち，クラスバランス変化がある状態では，モデルの推論精度が低下しやすい．クラスバランス変化を防ぐには，識別したい各クラスのサンプルデータ数をできるだけ等しくすることが重要である．

補足として，なぜ推定バイアスが発生するかを説明する．たとえば，ニューラルネットワークを用いる学習アルゴリズムでは，入力層からの情報を基に中間層のノードが重みを学習する．このとき，入力に良品画像が多い場合，良品が出力されやすいような重みが学習される．人間の脳のように，良品画像が多いと良品を識別するネットワークが強化されていくイメージである．

サポートベクターマシン（SVM）[86]のような古典的な機械学習においても，クラスの識別境界面が，サンプルデータ数の多いクラスの方の影響を強く受ける．つまり，サンプルデータ数の多いクラスの識別精度のみが向上するか，過学習[87]が生じる．いずれの場合も，品質の良いモデルとは言えない．

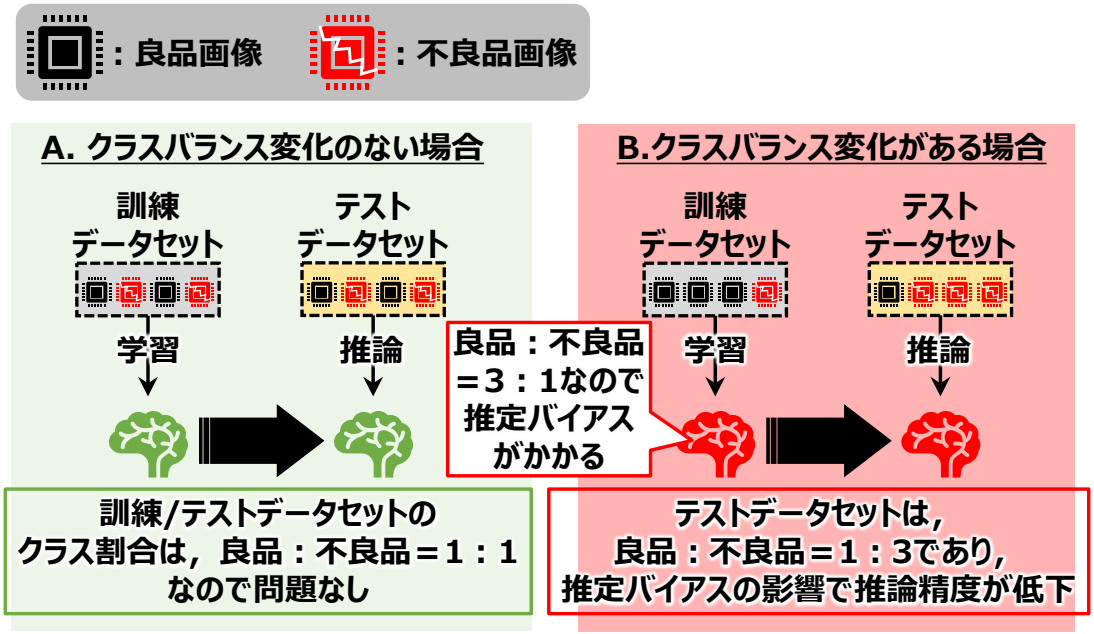


図 10 クラスバランス変化によるモデルの推論精度への影響

共変量シフトは，訓練データセットに含まれるクラスの分布とテストデータセットに含まれるクラスの分布が異なる状態を指す．図 11 は，共変量シフトが発生した際のモデルの推論精度への影響を示したものである．図 11 の例では，優製品，良製品，可製品，不良品を識別するモデルを作成する想定で，訓練データセットには，優製品画像と不良品画像のみが含まれ，テストデータセットには良製品画像と可製品画像が含まれている．この訓練データセットで訓練したモデルは，良製品と可製品をうまく識別できない状態となる．さらに，テストデータセットを使用してこのモデルで推論すると，良製品画像と可製品画像を優製品あるいは不良品と判断するか，ランダムに近い形で良製品か可製品かを判断してしまう．この例は極端であるが，訓練データセットの各クラスのサンプルデータ数の分布と，テストデータセットの各クラスのサンプルデータ数の分布に大きな違いがあると，同様の状態が発生し得る．したがって，共変量シフトが発生する環境では，モデルの品質が低下する．

一般的に，推論時の環境を正確に予測することが難しいため，共変量シフトを防ぐには，同一クラスのサンプルデータにバリエーションを持たせるか，クラスのバリエーション（種類）をできるだけ多くしておくことが重要である．つまり，識別したいクラスと「類似する」ものを収集し，訓練データに加えることが有効な対策となる．

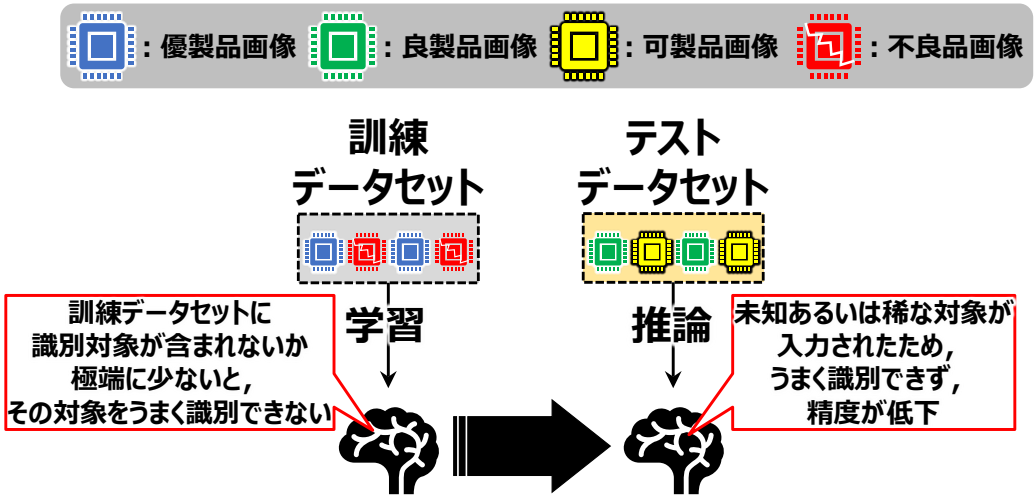


図 11 共変量シフトによるモデルの推論精度への影響

以上の議論から，品質の良いモデルを作成するためには，以下の 2 つの方策が有効である．

- **クラスバランス変化に対応するために，訓練データセットに含まれる各識別クラスのサンプルデータ数をできるだけ等しくする**
- **共変量シフトに対応するために，訓練データセットに識別クラスと類似するデータをできるだけ含ませる**

これらを満たすデータセットが，品質の良い訓練データセットとなる．よって，データの収集範囲が狭まることで，データのバリエーションに偏りが生じ，類似するデータが集めにくい状態となることは，モデルの品質を低下させる原因となる．

さらに，事象の生起確率に偏りがある場合，品質の良い訓練データセットの作成に長い時間を要することがある．たとえば，工場の検品処理を考えると，製品が良品である確率が 99.9%で，不良品である確率が 0.1%の場合，良品と不良品の画像を同数確保しようとすると，不良品の画像を集めるのに良品の 1000 倍の時間を要することになる．

データ収集範囲が広いほど，このような稀なケースのサンプルデータを集めやすくなるが，データ収集範囲が狭いことによる訓練データセットの品質低下あるいは確保時間の増大は，特に AI を利用するビジネスでは，機会損失などの深刻な問題を生む．

1.5 研究目的

本研究の目的は，訓練データ検索システムを実現し，エッジ AI における訓練データ不足を解決することである．

1.4節で,エッジAIでの訓練データ不足に起因するモデルの品質低下について説明した．この問題を解決するために，訓練データ検索システムを用いて，転移学習と呼ばれる技術を適用する．1.4節に記載のとおり，品質の良いモデルを作成するためには，訓練データセットが以下の条件を満たすことを要求する．

- **各識別クラスのサンプルデータ数をできるだけ等しくする**
- **識別クラスと類似するサンプルデータをできるだけ含ませる**

訓練データ不足が発生するエッジAIにおいて，これらの2つの条件を満たすためには，他の環境（以降，転移元と表記）で入手した訓練データを自身の環境（以降，転移先と表記）の学習に流用することが有効である．これは，転移学習[88][89]と呼ばれる技術の一部であり，近年，活発に研究されている．転移学習には，データ転移とモデル転移がある．データ転移は，訓練データを流用するもので，モデル転移は，訓練データではなく学習済みのモデルを流用するものである．データ転移とモデル転移のいずれの場合も，転移先と転移元の訓練データの類似度が高いことが，転移を成功させる一つの条件である．本研究では，特に断りがない限り，転移学習のうちデータ転移を対象とする．

訓練データ検索システムを用いて，転移学習に有効な転移元の訓練データを高精度に特定することができれば，識別クラスのサンプルデータを偏りなく，かつ，識別クラスの類似サンプルデータを含む良い訓練データセットを構成することができる．仮に，識別クラスのサンプルデータ数に偏りがある不均衡データセットを使用して分類器を作成した場合，次のような現象が生じ，クラス識別性能が低下する．たとえば，訓練データのクラス分布が「良品」:「不良品」=99 : 1の場合，常に「良品」と推論する分類器が作られると，分類器の「良品」に対する正解率は99%となる．一方，「不良品」に対する正解率は0%となる．推論時の入力データのクラス分布が，「良品」:「不良品」=5 : 5であった場合，この分類器は「良品」を過検出してしまう．このような状況が発生すると，「良品」と「不良品」で異なる処理を提供するAIサービスの場合，「不良品」に対しても「良品」の処理が割り当てられ，AIサービスの品質が低下する．

一方，転移先画像の単一サンプルデータを訓練データ検索システムに入力し，転移学習に有効な複数の転移元画像を得ることができれば，品質の良い訓練データセットを構成することができる．前述の例においては，転移先画像のサンプルデータに「不良品」が一つでもある場合，転移元画像のサンプルデータ集合から類似度が高いものを機械的に抽出し，必要に応じクレンジングすることで，高品質な「不良品」の集合を構築することができる（図12）．このように，訓練データ検索システム内での精度の高い類似度の算出は，不均衡データセットの生成を抑制し，AIサービスの品質低下を防ぐために重要な役割を果たす．

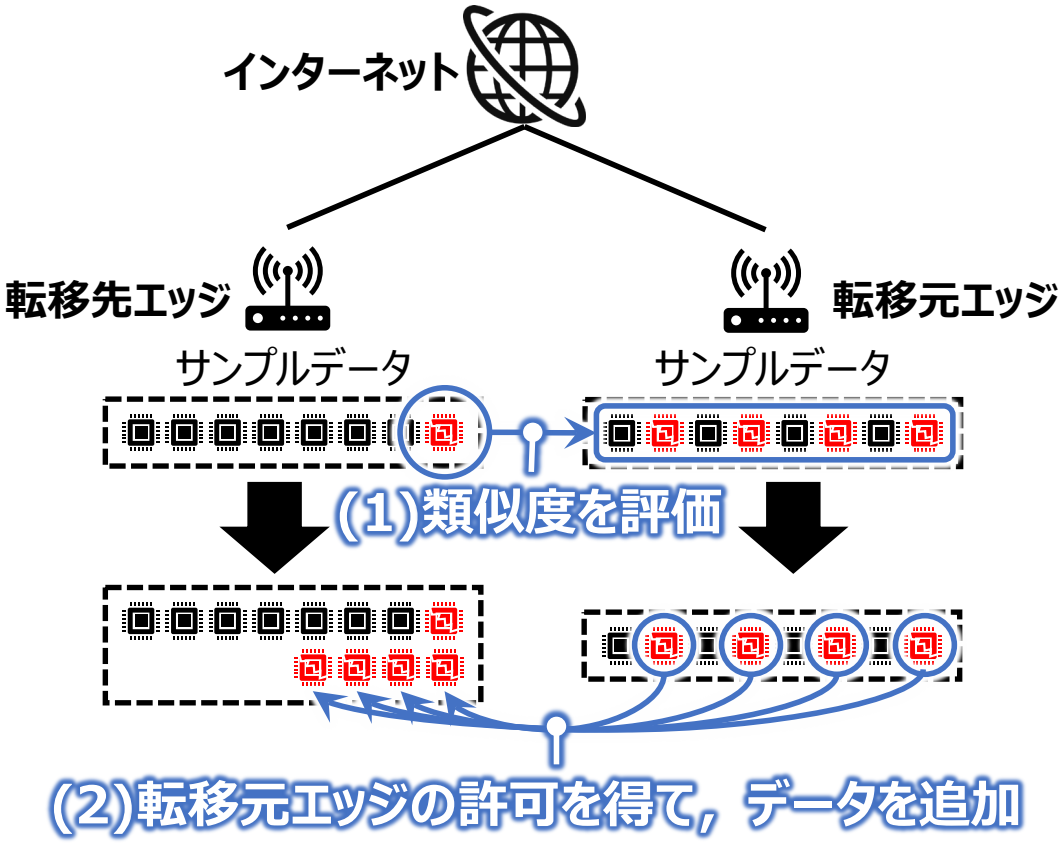


図 12 類似度の評価とデータ転移の関係

1.6 本研究の貢献

1.1 節で述べたように，エッジ AI における訓練画像を検索するシステムのための以下の 2 つの新しい技術を提案する．

新技術(1) クラウドにおいて，キーワード数に対するスケーラビリティとセキュリティを維持しながら，暗号化されたキーワードどうしを検索する技術

新技術(2) エッジにおいて，訓練データとして有効な画像かどうかを判定する技術

新技術(1)の従来の技術は，キーワードのハッシュ値（開示ビット）で検索空間を狭めるが，開示ビット長を手動で設定しなければならなかった．提案手法では，キーワードの最小エントロピーとk-匿名性を用いて，検索速度とセキュリティがバランスする開示ビット長を自動的に求める．

新技術(2)の従来の技術は，高負荷な AI アルゴリズムや単純な類似度指標で画像の類似度を算出しているため，エッジでの動作が難しく，比較画像間の照明変動や幾何学変換の

影響を受け，転移学習に好ましい類似度の算出が困難であった．提案手法では，幾何学変換に強い画像のヒストグラムを画像の特徴として使用し，照明変動に対応するために，ヒストグラムを分割し，分割区間ごとに DTW（Dynamic Time Warping）距離を求め，それらを結合して類似度を得る類似度指標を考案した．

新技術(1)に関する貢献の具体的な説明は，4.2 節で述べる．同様に，新技術(2)に関する貢献の具体的な説明は，5.2 節で述べる．

1.7 論文の構成

2 章では，スケーラビリティとセキュリティを維持しながら暗号化されたキーワードどうしを検索する技術と，訓練データとして有効な画像かどうかを判定する技術の関連研究を述べる．

3 章では，本研究の全体像，本研究の全体に係る前提条件，および，2 つの提案手法の性能要件を述べる．

4 章では，スケーラビリティとセキュリティを維持しながら暗号化されたキーワードどうしを検索する技術として，検索可能暗号における性能とセキュリティを考慮した高速化パラメータの決定手法を述べる．この章では，検索性能とセキュリティがトレードオフとなる検索可能暗号に対して，検索性能とセキュリティがバランスする高速化パラメータを，検索に使用するキーワードの最小エントロピーとk-匿名性を用いて求める方法を提案する．

5 章では，訓練データとして有効な画像かどうかを判定する技術として，データ転移に好ましい訓練データが検索可能な特徴点マッチングの類似度指標の詳細を述べる．この章では，データの類似度を，画像のヒストグラムの形状に着目し計算する方法を提案する．具体的には，訓練画像の類似度計算において，画像のヒストグラムの形状が平行移動したり，伸縮したり，相似形である場合でも類似度が高くなるように，類似度評価区間を極値で分割し，区間ごとに Dynamic Time Warping（DTW）距離を求め，各距離を結合することで類似度を計算する方法を提案する．

6 章では，結論と本研究の成果について述べる．

第2章 関連研究

本章では，訓練データ検索システムで必要とされる，スケーラビリティとセキュリティを維持しながら暗号化されたキーワードどうしを検索する技術と，訓練データとして有効な画像かどうかを判定する技術についての関連研究を述べる．

それぞれの関連研究から，本研究で対象とする以下の 2 つの課題を導出する．

- 検索速度とセキュリティを考慮した検索可能暗号の高速化パラメータの決定手法
- 転移学習に好ましい訓練データが検索可能な特徴点マッチングの類似度指標

2.1 暗号化されたキーワードどうしを検索する技術

2.1.1 データの保護範囲と保護手法

暗号化されたキーワードどうしの検索について述べる前に，本研究で対象とするデータの保護範囲を説明する．図 13 は，データの保護範囲を示したものである．本研究では，(1)エッジとクラウド間の通信（図 13(1)），(2)クラウド上のデータ保管領域（図 13(2)），(3)クラウド上のデータ挿入・検索処理の挙動（図 13(3)），の 3 つに保護範囲を分ける．

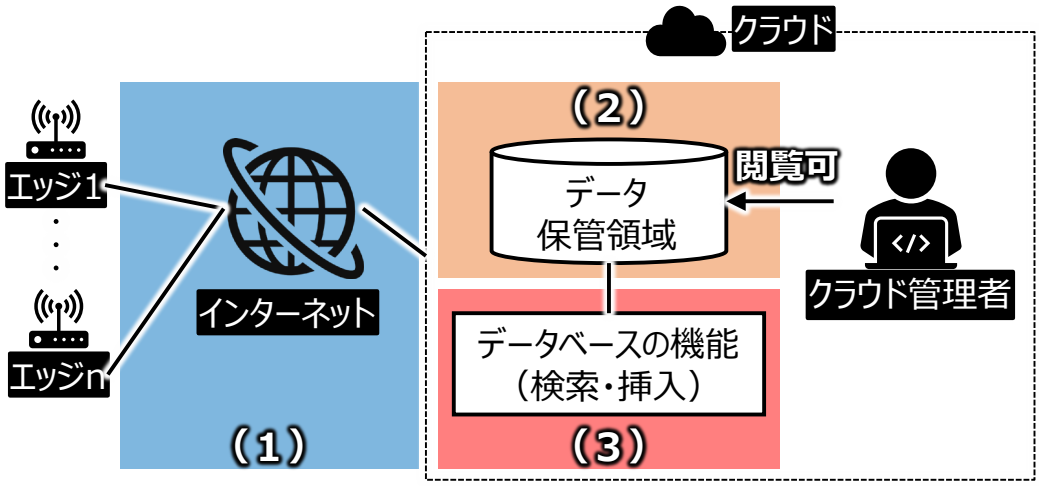


図 13 データ保護範囲

保護範囲(1)は，VPN（Virtual Private Network），TLS（Transport Layer Security），SSL（Secure Socket Layer），HTTPS（Hypertext Transfer Protocol Secure）等の既存のセキュア通信プロトコルで保護する．保護対象(2)は，保護対象のデータに対して，AES（Advanced Encryption Standard）などの既存の暗号スキームによるデータ暗号化で保護する．このとき，データの保管領域を暗号化しても，データの挿入や検索処理の途中で，暗

号化データがバッファやメモリ領域上で一時的に復号されるリスクがあることに注意する．悪意のあるクラウド管理者は，このバッファ領域やメモリの内容を見ることで，平文状態のデータを閲覧することができる．保護対象(3)は，データを暗号化したままデータ挿入や検索処理を実行することで，処理過程そのものを秘匿して保護する．このようにすることで，データ挿入や検索処理中の一時的な復号を防ぐ．保護範囲(1)(2)に関しては，既存方式を適用する前提で本論文の検討対象外とする．したがって，本論文では保護範囲(3)について議論する．

保護範囲(3)を保護するための，データを暗号化したままデータ挿入や検索処理を実行する技術として，秘密計算がある[90][91]．秘密計算は，算術演算と回路計算のいずれかを用いて実現できる．算術演算を用いるものは，秘密分散とマルチパーティ計算（Multi Party Computation，MPC）を用いる方法，準同型暗号や検索可能暗号を用いる方法などがある[92][93]．回路計算を用いるものは，Garbled Circuit（GC）のような論理ゲートを用いる方法がある[91]．なお，算術演算は，a+b や a・b など演算系を指し，回路計算は，AND ゲートや OR ゲートなどを指す．算術演算と回路計算は等価であり，互いに変換可能である．たとえば，a・b は a AND b と同じで，NOT a は 1－a，a XOR b は a+b－2・a・b と表される．

表 4 は，秘密計算の種類と得失をまとめたものである．秘密分散とマルチパーティ計算を用いる方法では，秘匿対象の元データを，秘密分散を用いて複数のシェアに分割する（図 14 上段①）．シェアは，元のデータを断片化した無意味なものであり，秘密分散時に予め指定した個数のシェアを集めない限り，元のデータを復元できないようになっている．そして，秘密計算を要求するクライアントは，各シェアを物理的に隔離したサーバに送る．シェアを受け取ったサーバは，MPC を用いてサーバごとにピアツーピア通信をしながら同じ計算を実行し，計算結果を結果シェアとしてクライアント返却する（図 14 上段②）．クライアントは，結果シェアをサーバから一定数受信すると，計算結果が復号可能な状態となる（図 14 上段③）．この方式の利点は，処理速度が速い，一定数のシェアを入手しない限り元データが復元されないため，攻撃者の共謀に強い，鍵が不要であることが挙げられる．一方，欠点は，物理的に隔離したサーバが 3 台以上必要で，コスト上昇の懸念がある，サーバ台数に秘匿性の強度が依存する，サーバどうしでピアツーピア通信するため，システムの構築難易度が高い，MPC により，サーバ 1 台で計算するより処理のオーバヘッドが大きいことが挙げられる．

表 4 秘密計算の種類と得失

	秘密計算		
	秘密分散+ マルチパーティ計算	**準同型暗号, 検索可能暗号**	**Garbled Circuit**
利点	• 処理速度が速い • 一定数のシェアを入手しない限り元データが復元されないため，攻撃者の共謀に強い • 鍵が不要	• 完全一致のような機能が単純なものであれば，処理速度が速い • 暗号モジュールのみで構築可能であり，導入難易度が低い	• 秘密分散と比較し，ピアツーピア通信の回数が少ない • 2 者間計算が可能(秘密分散は 3 者が必要)
欠点	• 物理的に隔離したサーバが 3 台以上必要で，コスト上昇の懸念がある • サーバ台数に秘匿性の強度が依存する • サーバどうしでピアツーピア通信するため，システムの構築難易度が高い • MPC により，サーバ 1 台で計算するより処理のオーバヘッドが大きい	• 秘密鍵の管理が必要 • 高安全な方式は低速であり，<u>高速化には適切なパラメータ設定が必要</u>	• Garbled Circuit の生成に時間がかかる • ラベルのサイズが大きく，データ量が増える • 演算が複雑で，計算負荷が高い • 元データと Garbled Circuit で用いるラベルの対応付けを秘密に管理する必要がある

図 14 秘密計算の動作の概要

次に，準同型暗号や検索可能暗号を用いる方法を説明する．この方法は，まず，秘密鍵を用いて，秘匿対象の元データを準同型暗号などの暗号アルゴリズムで暗号化する（図 14 中段①）．このとき，暗号データは，暗号化したまま，元データに対する加算あるいは乗算ができる状態となる．秘密計算を要求するクライアントは，サーバに暗号データを送る．サーバは，受け取った暗号データと，サーバが既に受信した他の暗号データとの演算を行う（図 14 中段②）．検索可能暗号の場合，各暗号データが，演算権限のある秘密鍵で暗号化されているならば，演算結果から元のデータどうしが一致しているかどうかのみを判断できる．サーバは，演算結果をクライアントに返却し，クライアントは復号鍵を用いて演算結果を復号し，演算結果を得る．この方式の利点は，完全一致のような機能が単純なものであれば，処理速度が速い，暗号モジュールのみで構築可能であり，導入難易度が低いことが挙げられる．一方，欠点は，秘密鍵の管理が必要であることや，高安全な方式は低速であり，高速化には適切なパラメータ設定が必要なことである．

最後に，Garbled Circuit を用いる方式は，通常の論理ゲートと異なり，入力が 0，1 ではなく，0，1 に対応した乱数 K_0，K_1 を用いて計算を行う回路である[94]．K_0，K_1 はラベルと呼ばれる．たとえば，通常の論理回路は，AND ゲートに 0，1 を入力すると，0 が出力されるが，Garbled Circuit では，K_0，K_1 が Garbled AND ゲートに入力され，結果として K_0 が出力される．すなわち，攻撃者は，特定の Garbled Circuit におけるラベルと 0，1 の対応関係を知らないと結果が正しいものであるかを判断できない．Garbled Circuit を用いる方式では，まず，エッジで Garbled Circuit を作成する（図 14 下段①）．次に，作成した GC をクラウドへ送信する．このとき，GC のラベルと 0，1 の対応関係は，エッジ上で秘密に保持する．エッジが，クラウドに秘密計算を実行させる際は，0,1 の入力をラベルに変換し，クラウドへ送信する．クラウドは，受信したラベルを GC に入力し，結果をエッジに返却する（図 14 下段②）．最後に，エッジは，処理結果をラベルと 0，1 の対応関係を使って復号する（図 14 下段③）．この方式の利点は，秘密分散と比較し，通信量が少ない，2 者間計算が可能（秘密分散は 3 者が必要）なことである．一方，欠点は，Garbled Circuit の生成に時間がかかる，ラベルのサイズが大きく，データ量が増える，演算が複雑で，計算負荷が高い，元データと Garbled Circuit で用いるラベルの対応付けを秘密に管理する必要がある．

データの検索処理を秘匿したい場合に，秘密分散とマルチパーティ計算を用いる方法と Garbled Circuit を用いる方式では，不具合が生じる．たとえば，暗号化キーワードどうしが一致しているかを判定するためには，逐一，エッジ側で復号して結果を得なければならない．そのため，クラウドとエッジ間の通信が多発し，検索速度が低下する．さらに，秘密分散とマルチパーティ計算を用いる方法は，物理的にサーバを分ける必要があるが，一つのクラウドでは，仮想的にしかサーバを分離できないため，秘匿性を担保できない．

以上のことから，データを暗号化したままデータ挿入や検索処理を実行する技術には，準同型暗号や検索可能暗号を用いる方法が向いている．本研究では，エッジとサーバとの通信

回数を減らすために，検索可能暗号を採用する．検索可能暗号は，秘匿性などのセキュリティと検索速度がトレードオフの関係にあり，高安全な方式を用いる場合には，トレードオフを考慮した高速化パラメータの設定が必要となる[103]．そのため，性能とセキュリティをバランスする高速化パラメータの決定手法を検討する．検索可能暗号を利用すると，検索機能のバリエーションが減るが，処理速度を重視し，限定的な検索機能となることを許容するものとする．

2.1.2 検索の方式

この項では，暗号化されたキーワードどうしの検索において，どのような検索方式が存在し，どの検索方式を用いるのが適しているかを述べる．

主な検索方式は，以下の 3 つである．単純化するために，範囲検索や AND，OR，NOT，ワイルドカード検索は対象外とする．

- 完全一致検索
- 部分一致検索
- あいまい検索

完全一致検索は，キーワードが完全に一致するものを探し出すものである．バイナリ比較で済むため，データベースのインデックスが作成でき，非常に高速である．さらに，検索キーワードと関係のないキーワードが検索されないため，ノイズが乗りにくい．一方で，キーワードが厳密に一致することが求められるため，表記ゆれやタイポ，意味的に類似しているキーワードの検索ができない．

部分一致検索は，検索対象のキーワードの中に検索キーワードの一部が含まれているキーワードを探し出すものである．たとえば，検索対象のキーワードが「日本酒」の場合，検索キーワードに「日本」や「酒」と指定すると，ヒットする．完全一致と異なり，一般的には 1 回のキーワードどうしの比較において，複数回のバイナリ一致が実行されるため，検索速度が低下する．また，キーワードの意味を解釈しないため，意図しないキーワードがヒットすることがある．たとえば，検索対象のキーワードが「日本酒」の場合，検索キーワードに「本」と指定するとヒットし，ユーザが求める書籍ではなく，全く関係のないアルコール飲料が結果に含まれてしまう．一方，完全一致検索より柔軟な結果を得ることができるため，検索の取りこぼしは少なくなる．

あいまい検索は，検索対象のキーワードと検索キーワードが意味的に似ているものを探し出すものである．実現方式はいくつかあるが，古典的には検索対象のキーワードや検索キーワード，あるいはその両方に対して形態素解析を実行して形態素と呼ばれる最小単位の単語に分解し，形態素に対して類義語展開を適用し，キーワードを増やしヒットしやすくするものである．ここで重要となるのは，類義語展開の中で使用されるシソーラスである．シ

ソーラスは，形態素に関する類義語や異表記などの関係をまとめたもの[95]で，この情報を使用することで意味的に似ているキーワードを探し出すことができる．シソーラスの情報量が類義語展開の精度に依存するが，本論文では，十分な情報量を持つシソーラスを使用することを想定し，あいまい検索の精度については言及しない．最近では，あいまい検索に AI を用いるものもあるが，暗号化されたキーワードに対しては適用が困難であるため除外する．

本論文では，訓練データの候補をできるだけ正確に特定するという観点から，暗号化されたキーワードどうしの検索にあいまい検索を用いることとした．ただし，エッジは，訓練データの候補となる画像のキーワードを暗号化してからクラウドにアップロードするため，形態素解析や類義語展開をエッジ側で実行する必要があることに注意する．

あいまい検索は，古典的には全文検索と呼ばれる情報検索技術を用いて実現する．全文検索は，複数の文書ファイルから特定の文字列（キーワード）を検索する技術である[96]．1990 年代にオンライン書籍データベースで使用され，その後，Web 検索エンジンや文書作成アプリケーション等に応用された．近年では，チャットツールである Slack[97]やクラウドファイルストレージである Google ドライブ[98]のような情報共有ツール内の電子データの利活用を目的に，全文検索が使われることがある．

全文検索を実現するアルゴリズムには，文字列検索，シグネチャファイル，転置索引などがある[96]．表 5 は，全文検索を実現するアルゴリズムの主な用途，利点や欠点をまとめたものである．

表 5 全文検索を実現するアルゴリズム

	文字列検索	シグネチャファイル	転置索引
概要	• 検索対象文書中に検索キーワードが含まれるかどうか照らし合わせる方法	• 各文書に対して，文書中のキーワードに対するビットパターンからなるシグネチャの和を取ったシグネチャファイルを作成 • 検索時は，シグネチャファイルと検索キーワードのシグネチャとの照合を行い，一致した場合は，その文書に対して文字列検索を実行し結果を確認	• 文書中のキーワードに対して，それを含む文書を対応付ける形で作った索引を作成
主な用途	• UNIX の grep コマンド • Windows explorer などのユーティリティの検索 • 索引を持たないデータベースのテキスト検索	－	• インターネット検索サービス • 企業向け社内検索サービス • デスクトップ検索
利点	• 事前の索引作成が不要 • 数 MB から十数 MB 程度の小規模なデータに対しては，実用的な速度で動作	• 文字列検索に比べ高速 • 転置索引に比べて更新が容易	• 検索速度は，転置索引の中の検索キーワードに対応する文書の識別子（文書 ID）の数のみに依存するため高速 • フレーズ検索，ランキング検索にも応用可能
欠点	• 大規模データに対しては，処理速度の点から実用的でない	• 検索実行時に，シグネチャによって絞り込まれた文書に対する文字列検索が実行されるため，文書サイズによっては速度が実用的でない	• 事前に索引作成が必要

文字列検索は，検索対象文書中に検索キーワードが含まれるかどうかを照らし合わせる方法である．予め索引作成を行う必要がないため，手軽に利用できる．数 MB から十数 MB 程度の小規模なデータに対しては，実用的な速度で動作するが，大規模データに対しては，処理速度の点から実用的でない．UNIX の grep コマンド，Windows explorer などのユーティリティの検索，索引を持たないデータベースのテキスト検索などに使用される．

シグネチャファイルでは，各文書に対して，文書中のキーワードに対するビットパターンからなるシグネチャの論理和を取ったシグネチャファイルを作成する．検索実行時には，このシグネチャファイルと検索キーワードのシグネチャとの照合を行い，一致した場合は，その文書に対して文字列検索を実行し，結果を確認する．文字列検索に比べ高速で，転置索引に比べて更新が容易である．検索実行時に，シグネチャによって絞り込まれた文書に対する文字列検索が実行されるため，文書サイズによっては速度が実用的でない．

転置索引は，文書中のキーワードに対して，それを含む文書を対応付ける形で作った索引を使用して検索する．検索速度は，転置索引の中の検索キーワードに対応する文書の識別子（文書 ID）の数のみに依存するため高速である．事前に索引作成が必要であるが，多くのインターネットの検索サービスや商業パッケージで採用されている．本研究では，商業利用例が多い転置索引を全文検索として使用する．

全文検索システムをオンプレミスで運用すると RASIS[b)]の確保に多大なコストが生じるため，運用負荷低減とサービス導入の容易性からクラウドサービスを利用することが多い．クラウドサービスとして実装する場合の，転置索引を用いた全文検索の処理手順を以下に示す．図 15 は，処理手順を図示したものである．図中の番号は，手順番号と対応する．なお，説明の都合上，クラウドをサーバ，エッジをクライアントと表記する．

b) Reliability（信頼性），Availability（可用性），Serviceability（保守性），Integrity（完全性），Security（機密性）

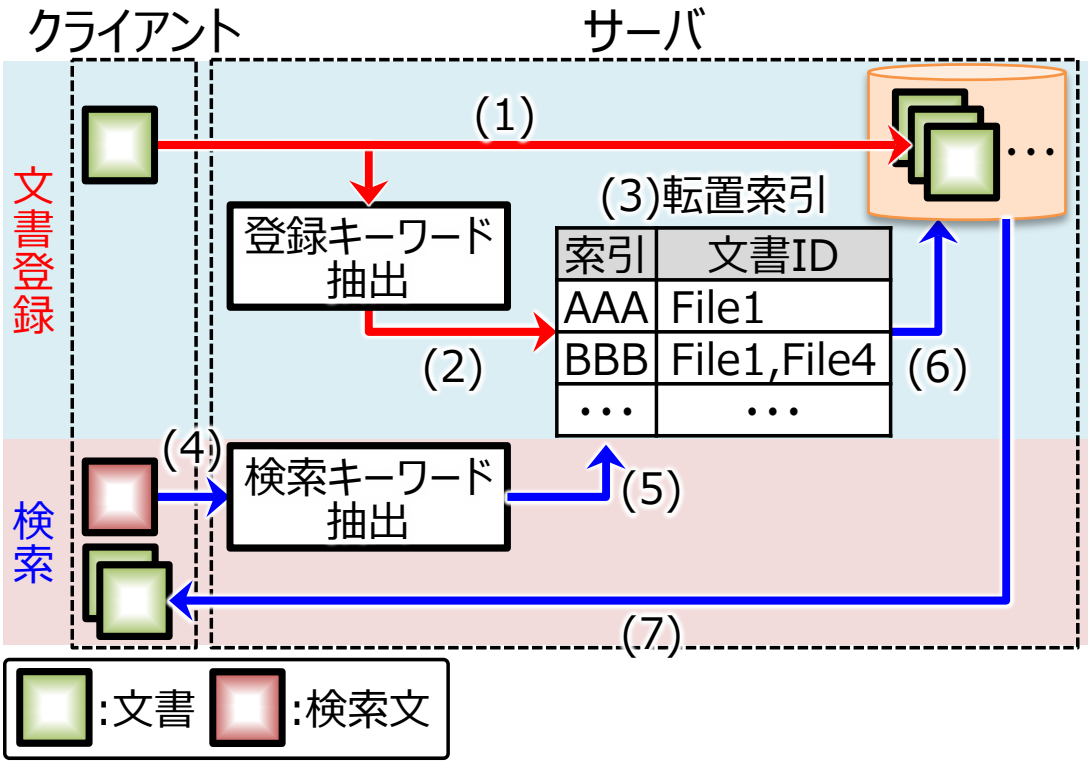


図 15 全文検索の処理手順

文書登録

(1) クライアントは，文書をサーバに送信する．

(2) サーバは，文書を形態素解析し，登録キーワードを抽出する．登録キーワードと近い意味合いの検索キーワードでも文書がヒットするよう，キーワードの類義語や同義語を求める類義語展開を行う場合もある．

(3) サーバは，登録キーワードと文書の ID を関連付けた転置索引と呼ばれる表を作成する．

検索

(4) クライアントは，検索文をサーバに送信する．

(5) サーバは，検索文に対し形態素解析と必要に応じ類義語展開を行い，検索キーワードを抽出する．

(6) 転置索引からこの検索キーワードと一致するレコードを求め，文書 ID を特定する．

(7) 文書 ID から文書を特定し，文書をクライアントに送信する．

以上のように，全文検索では転置索引を作成するため，平文のキーワードがサーバに保存される．その結果，悪意のあるクラウド管理者やクラウド内に侵入したマルウェアによってキーワードが取得され，機密情報が流出するおそれがある．したがって，悪意のあるクラウド管理者から文書を保護するために，キーワードを暗号化した上で検索することが求めら

れる．

2.1.3 検索可能暗号の方式

この項では，2.1.1 項で述べた検索可能暗号の詳細と，検索可能暗号の方式について述べる．

図 16 は，検索可能暗号の概念を表したものである．

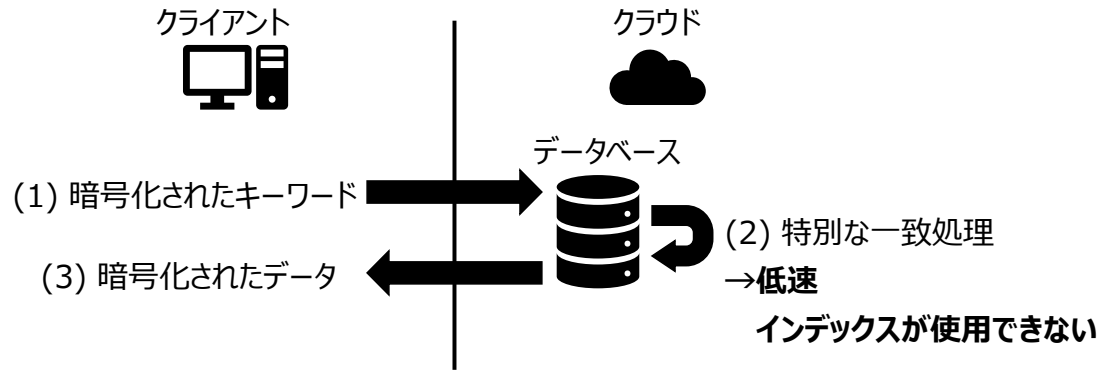


図 16 検索可能暗号の概念

クライアントは，クラウド上のデータベースに暗号化されたキーワードと暗号化されたデータを保存する．検索時は，クライアントから暗号化された検索キーワードをデータベースに送信し，クラウドは，データベース内の暗号化されたキーワードと暗号化された検索キーワードを入力として，特別な一致処理を実行する．この一致処理中に，入力した 2 つのデータが復号されることはなく，2 つのデータが一致しているかどうかしか判定しない．一致した場合，クラウドは暗号化されたキーワードに紐づく暗号化されたデータをクライアントに返却する．

特別な一致処理は，暗号化データに対する特殊な演算を実行するため，平文データの一致処理に比べて低速である．平文データのようにデータベースのインデックス機能が使用できないため，データベース内のすべての暗号化されたキーワードと一致処理を実行しなければならず，大規模データに対して検索可能暗号を使用する場合は，データベースを分散化し，並列的に検索を実行するなどの対策が採られる[99]．

このように，大規模データへの適用の問題はあるが，データベース内の検索対象となるデータはすべて暗号化されるため，パーソナルデータなどの機微な情報を扱うユースケースに適している．

検索可能暗号の詳細な動作を理解するために，全文検索に検索可能暗号を導入した際の処理手順を例に説明する．本研究では，特に断りがない限り，検索可能暗号は，後述する確率的暗号ベースの方式を想定する．確率的暗号ベースの方式は，クラウドに対して高いセキ

ュリティを確保できるが，平文の検索時に比べ検索性能が低下する．

図 17 は，検索可能暗号の処理手順を図示したものである．図中の番号は，処理手順番号と対応する．なお，説明の都合上，クライアントをサーバ，エッジをクライアントと表記する．

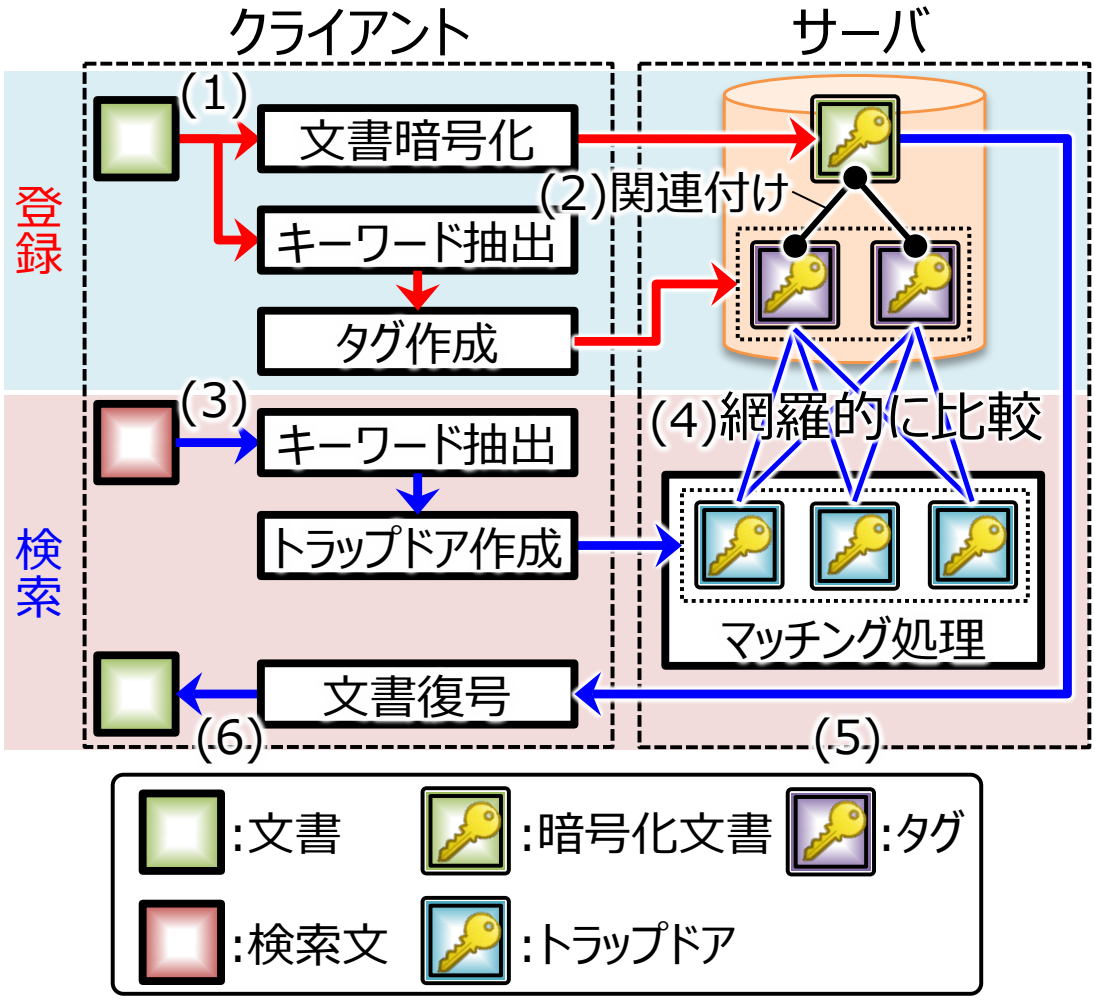


図 17 検索可能暗号の処理手順

登録

(1) クライアントは，文書から登録キーワードを抽出し，登録キーワードを暗号化鍵で暗号化して検索用のタグ（以降，タグと表記）を作成する．並行して，文書を暗号化鍵で暗号化する．暗号化した文書とタグをサーバに送信する．なお，キーワードの暗号化鍵と文書の暗号化鍵は異なっていてもよい．

(2) サーバは，暗号化済みの文書とタグを関連付けてデータベース（以降，DB と表記）に保管する．

検索

(3) クライアントは，検索文から検索キーワードを抽出し，検索キーワードを暗号化鍵で暗号化して検索語（以降，トラップドアと表記）を作成する．このトラップドアをサーバに送信する．

(4) サーバは，トラップドアと DB 内のタグを特殊な演算を用いて網羅的に比較する．この演算は，トラップドアとタグを入力すると，1（：＝一致）または 1 以外（：＝不一致）を出力するため，トラップドアとタグに対応する暗号化前のキーワードが等しいかどうかを判定できる．

(5) サーバは，トラップドアと一致したタグに関連付く暗号化済みの文書をクライアントに返却する．

(6) クライアントは，復号鍵で文書を復号する．

以上のように，暗号化されたキーワードの全文検索では，基本的にサーバ側で転置索引を作成することができないため，タグとトラップドアを一つ一つ比較していきキーワードの一致を確認する必要がある．そのため，DB 内のタグの数が多いと，検索速度が低下する．したがって，全文検索に検索可能暗号を適用する場合，実用的な検索性能を達成するには高速化が必須となる．

検索可能暗号は，方式によって検索性能が異なる．検索可能暗号は，確定的暗号ベースの方式[100]，確率的暗号ベースの方式[101][102]，ハイブリッド方式[103]がある．表 6 は，検索可能暗号の方式と，その利点と欠点をまとめたものである．各方式の説明を以下に示す．

• 確定的暗号ベースの方式：

図 18 上段に示すように，同一の入力からは同一の出力が得られる暗号方式を用いて，タグとトラップドアを作成する．同一のキーワードから常に同一のタグとトラップドアが生成されるため，検索時のマッチング処理はタグとトラップドアのバイナリ一致判定を行うだけでよい．実質的にキーワードを置換するだけであるため，転置索引を作成可能であり，高速な検索が可能なデータベースのインデックスを使用することもできる．よって，サーバの DB テーブルのレコード数nに対して$O(\log n)$で検索できる．一方，登録キーワードとタグ，および，検索キーワードとトラップドアが 1:1 で対応するため，DB テーブル中のタグの頻度分布と平文のキーワードの頻度分布が等しくなり，頻度分析攻撃により平文のキーワードが特定されやすくなってしまう[104]．

• 確率的暗号ベースの方式：

図 18 下段に示すように，同一の入力でも異なる出力が得られる暗号方式を用いて，タグとトラップドアを作成する．検索時のマッチング処理では，同一のキーワードでも異なるタグが生成されるため，タグとトラップドアの間で特殊な演算を行い，キーワードの一

致を判断する必要がある．したがって，トラップドアとデータベース内のすべてのタグを比較する必要があるため，サーバの DB テーブルのレコード数nに対して$O(n)$の検索時間を必要とする．nが大きい場合，実用的な時間で検索を行うことが困難となるおそれがある．一方で，キーワードとタグの出現頻度分布に関連はなく，DB テーブル中のタグの頻度分布からは，キーワードの頻度分布が漏れないため，タグに対して頻度分析攻撃を実行しても，平文のキーワードの情報は一切漏れない．このことから，確定的暗号ベースの方式より安全性が高いとされ，悪意のあるクラウド管理者に対して高いセキュリティを有する．

・ハイブリッド方式：

確率的暗号ベースの方式を基本に，キーワードのハッシュ値の一部（以降，開示ビットと表記）をサーバに開示し，サーバが開示ビットに基づいて検索空間を絞り込むことで高速化を図る．この方式の詳細は後述するが，頻度分析攻撃に対応しつつ，検索性能を向上させることができる．検索時間は，サーバの DB テーブルのレコード数nに対して$O(n/2^c)$である（cは，開示ビットのビット長．以降，開示ビット長と表記する）．ただし，検索性能と頻度分析攻撃への耐性は，cの値によって変化し，トレードオフの関係になっている．したがって，cの適切な設定が不可欠である．

表 6 主な検索可能暗号の方式と利点・欠点

	確定的暗号ベース	確率的暗号ベース	ハイブリッド
代表例	Bellare[100]ら	Song ら[101]，Boneh ら[102]	平野ら[103]
利点	処理が高速	頻度分析攻撃に強い	処理が高速， 頻度分析攻撃に強い
欠点	頻度分析攻撃に弱い	処理が低速	パラメータ調整が必要

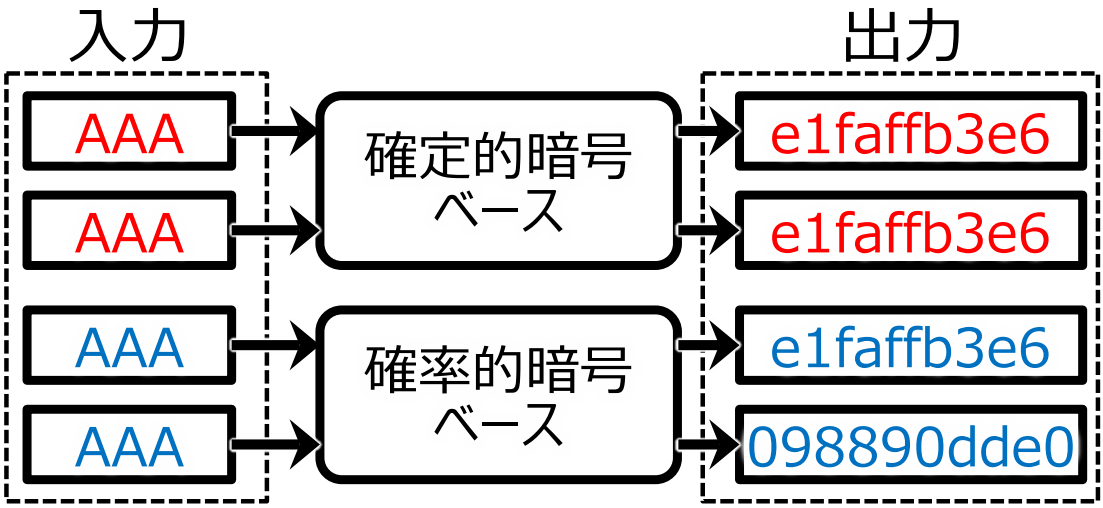


図 18 確率的暗号ベースと確定的暗号ベースの違い

文献[105][106]では，検索可能暗号上のキーワード推定攻撃（KGA：Keyword Guessing Attack）は，キーワード空間が狭い（キーワードの種類の数が少ない）か，キーワードが持つエントロピーが低い（特定のキーワードが出現しやすい）場合に成功しやすいと述べている．したがって，タグや開示ビットに対する頻度分析攻撃によってキーワードの頻度分布が正しく推定されると，KGA が成功しやすくなるため，キーワードの頻度分布を隠蔽することが重要である．

本研究では，クラウドに対し高いセキュリティを確保でき，かつ，高速な検索の実現を見込めるハイブリッド方式に着目する．

2.1.4 ハイブリッド方式の検索可能暗号

この項では，ハイブリッド方式の検索可能暗号の詳細について述べる．

平野らの論文[103]では，開示ビットと呼ぶ，キーワードから確定的に決まる数ビットの値をサーバへ開示し，この開示ビットを使用して DB テーブルの検索空間を絞りこむことで高速化を図る方式を提案している．さらに，提案した方式に対して安全性証明をつけている．

開示ビットは，キーワードのハッシュ値を予め定められた長さでトリミングして求める（**図 19**）．ハッシュ関数には，鍵付きハッシュ関数を用い，クライアントごとに異なる秘密鍵を入力する．鍵付きハッシュ関数を用いることで，秘密鍵を持たない悪意のあるクラウド管理者は，開示ビットを生成できなくなる．したがって，攻撃者は，辞書などを基にキーワードを網羅的に入力して，開示ビットに対応する平文キーワードを探索するような攻撃が不可能となる．開示ビットを使用した検索空間の絞り込み手順の詳細は，2.1.6 項を参照のこと．

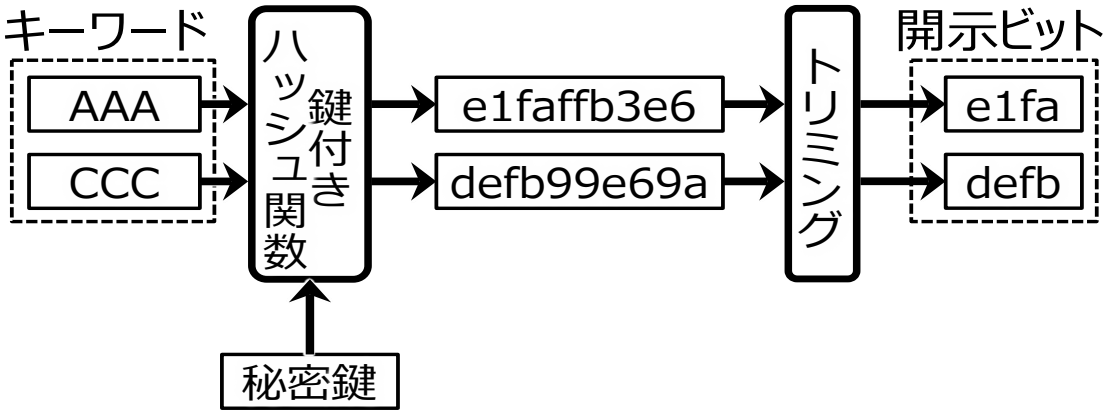


図 19 開示ビットの計算方法

この開示ビットを用いた絞り込み動作により，マッチング処理時の検索空間を減らすことができる（図 20）．2.1.3 項で述べた通り，検索時間は，おおよそ$O(n/2^c)$であるため，開示ビット長を表すパラメータcを大きく設定することで検索時間をより多く削減できる．しかし，cが大きくなるにつれ，絞り込み後のレコード集合の大きさが小さくなっていき，開示ビットと平文のキーワードの出現頻度分布が近づいていく．最終的に，cがある一定以上になると，開示ビットによる絞り込み後のレコード集合の大きさが 1 になり，キーワードの頻度分布と一致する．すなわち，確定的暗号ベースの方式と同等なセキュリティレベルとなってしまう．

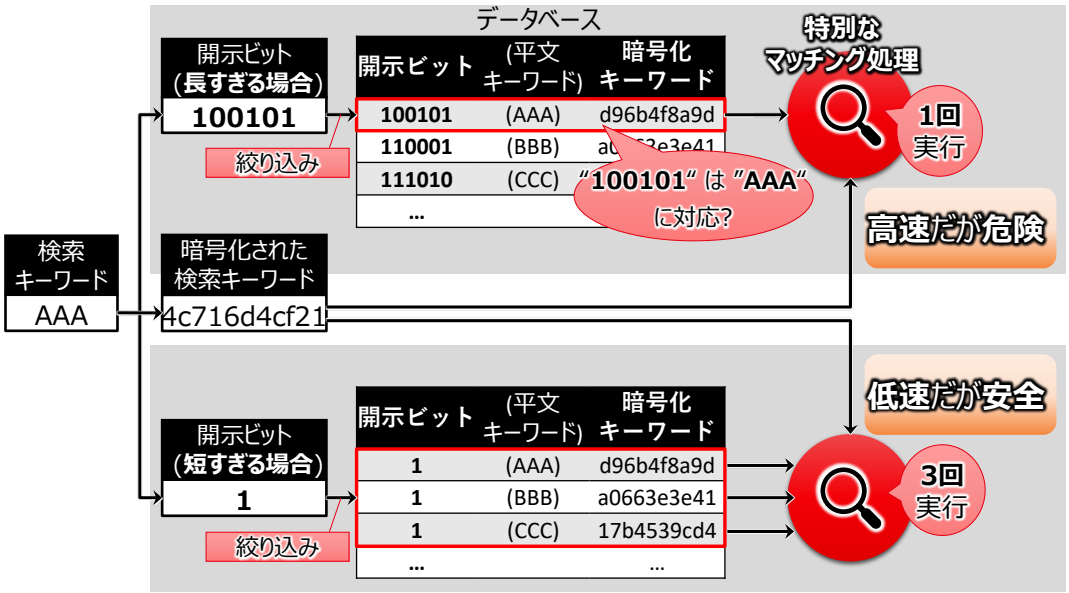


図 20 ハイブリッド方式における頻度分析攻撃のリスク

この開示ビット長は，キーワードの頻度分布に応じて決定されるべきであるが，平野らは開示ビット長の決定方法に関する詳細な議論をしていない．

2.1.5 検索可能暗号を用いた全文検索の既存方式

この項では，検索可能暗号を用いて全文検索を実現している既存研究を紹介する．

検索可能暗号を用いて全文検索を実現する方法として，以下の 2 つの既存研究がある．

(A) タグにあいまい性を持たせる方式

(B) トラップドアにあいまい性を持たせる方式

以降で，これら 2 つの既存研究の詳細を述べる．

A. タグにあいまい性を持たせる方式

Li らの論文[107]では，検索キーワードのタイプミスや表現の揺らぎを許容するために，ファジーキーワード検索技術を適用した方式を提案している．

登録時に，登録キーワードw_iに対して，編集距離[108]を基準としたファジーキーワード集合$\{w_i'\}$を網羅的に求め，$\{\mathrm{w}_i, \{w_i'\}\}$のタグを生成し，サーバに保管しておく．具体的には，図 21 のように，2.1.3 項で述べた処理手順(1)において，文書から抽出したキーワードに対してファジーキーワード集合を求める処理を加える．編集距離は，2 つの文字列がどの程度異なっているかを示す距離の一種であり，1 文字の挿入・削除・置換によって，一方の文字列をもう一方の文字列に変形するのに必要な手順の最小回数として定義される．たとえば，キーワードが CASTLE の場合，1 文字目に対する置換操作は{AASTLE, BASTLE, DASTLE, ⋯, YASTLE, ZASTLE}のようになる．しかし，このように単純に列挙すると$\{\mathrm{w}_i, \{w_i'\}\}$のサイズが著しく増加する．そこで，ワイルドカードを用いて{CASTLE, *CASTLE, *ASTLE, C*ASTLE, C*STLE, ⋯, CASTL*E, CASTL*, CASTLE*}のように表現し，$\{\mathrm{w}_i, \{w_i'\}\}$のサイズを抑える．検索キーワード$w$に対してはファジーキーワード集合を作成せず，通常通りトラップドアを作成し，タグとのマッチング処理を行う．

この方式は，ベースとなる検索可能暗号にワイルドカードを含むタグとトラップドアとの部分一致検索機能を要求する．タグ数は，キーワードとワイルドカードの組み合わせの場合の数に比例して増加するため，文書数が多くなると高速化が必須となる．

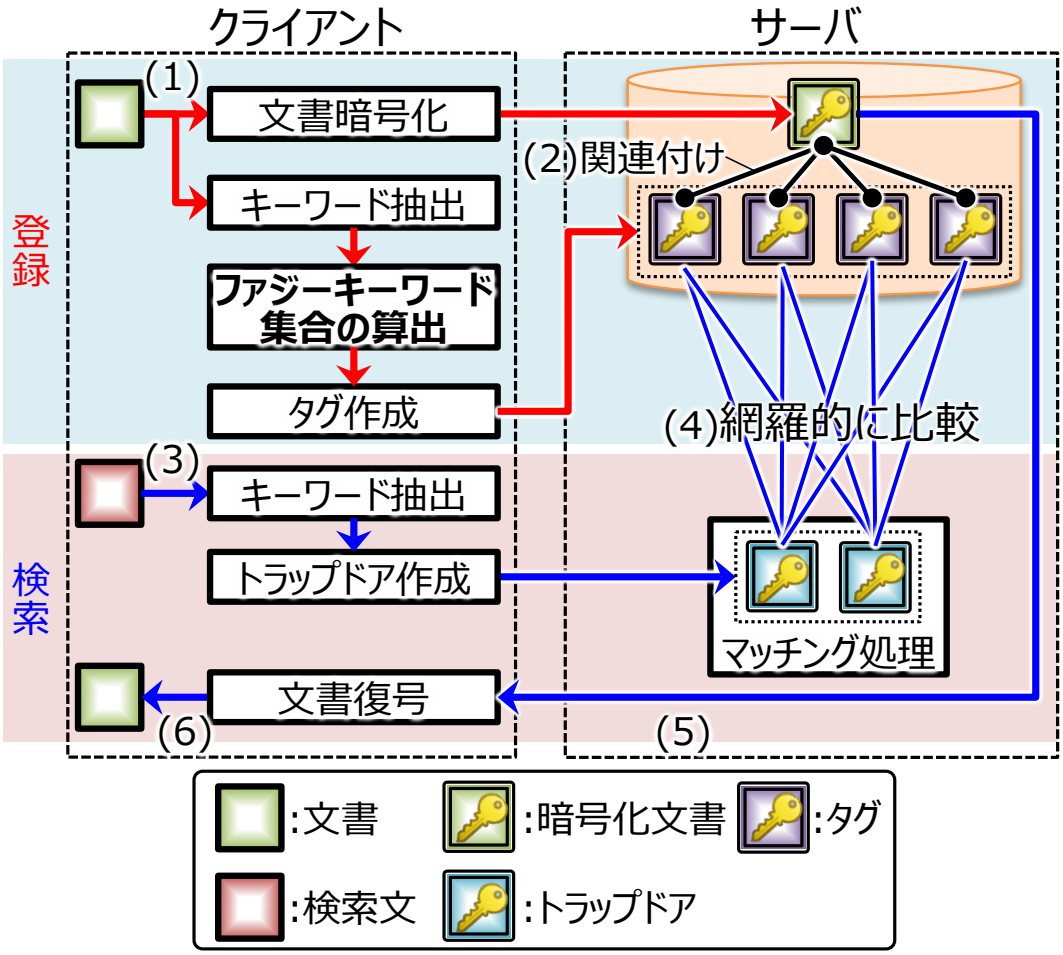


図 21 タグにあいまい性を持たせる方式

B.　トラップドアにあいまい性を持たせる方式

尾形らの論文[109]では，登録時は手を加えず，検索時に検索キーワードwと意味的に近いキーワード集合$S(w)$を求め，$\{w, S(w)\}$に対しトラップドアを生成し，各トラップドアでの検索結果を OR 演算することであいまい検索を実現する方法を提案している．

この方式は，Li らの方式と異なりベースとなる検索可能暗号には，完全一致検索の機能があればよい．一方，文書数に対するスケーラビリティの問題は依然として残る．

表 7 にタグにあいまい性を持たせる方式と，トラップドアにあいまい性を持たせる方式の特徴の一覧を示す．

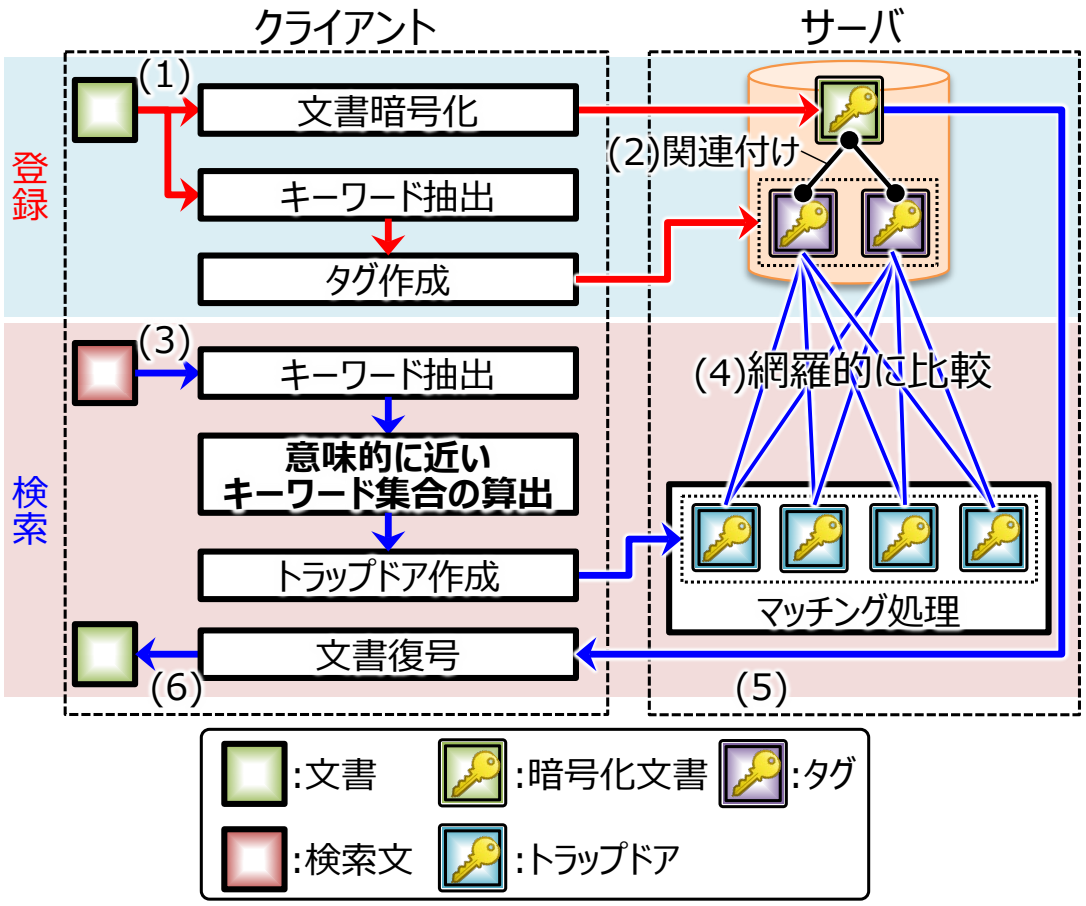


図 22 タグにあいまい性を持たせる方式

表 7 検索可能暗号を用いた全文検索の実現方式の特徴比較

	(A) タグにあいまい性を持たせる方式	(B) トラップドアにあいまい性を持たせる方式
代表例	Li らの方式[107]	尾形らの方式[109]
データ登録速度	オリジナルの検索可能暗号より低速 登録キーワードが増えることで，タグの数がオリジナルの検索可能暗号より多くなるため.	オリジナルの検索可能暗号と同等 データ登録処理はオリジナルの検索可能暗号と変わらないため.
検索の柔軟性	方式(B)に比べ低い あいまい性を動的に変更しようとすると，DB 内のタグをすべて更新しなければならないため.	方式(A)に比べ高い データ登録後でもタグの更新なしにあいまい性の変更が可能で，検索結果をクライアントごとに変えられるため.

タグにあいまい性を持たせる方式は，データ挿入時に登録キーワードが増えるため，全文検索を実装していないオリジナルの検索可能暗号に比べて DB に保管するタグの数が増加する．そのため，データの登録時間が長くなる．一方，検索キーワードは増えないため，トラップドアの数は，オリジナルの検索可能暗号と同等である．あいまい性を制御するために，ファジーキーワード集合の算出アルゴリズムを運用途中で変更すると，DB 内のタグをすべて更新しなければならない．タグを更新しない場合，ファジーキーワード集合の算出アルゴリズムの変更前後で検索ヒット率が変化してしまう．よって，あいまい性を動的に変化させることが実用上困難であり，検索の柔軟性は低い．

トラップドアにあいまい性を持たせる方式は，データ登録時にキーワードを増やさないため，タグの数はオリジナルの検索可能暗号と同等であり，データの登録時間はオリジナルと変わらない．一方，検索キーワードを増やすため，トラップドアの数が増える．検索時のタグとトラップドアの比較回数は，タグにあいまい性を持たせる方式とほぼ変わらない．しかし，検索時にトラップドアを多数作成するため，トラップドアの作成コストが高い場合，タグにあいまい性を持たせる方式より検索時間が増加することがある．検索キーワードに対してキーワードを増やすため，データ登録後でもタグの更新なしにあいまい性の変更が可能で，検索結果をクライアントごとに変えることもできる（検索の柔軟性が高い）．

2.1.6 ハイブリッド方式の検索可能暗号を用いた全文検索の構成方法

この項では，既存方式であるハイブリッド方式の検索可能暗号と全文検索を組み合わせて，暗号化されたキーワードどうしの検索を構成可能であることを述べる．そして，検索速度とセキュリティを考慮した検索可能暗号の高速化パラメータの決定手法が必要になることを説明する．

ハイブリッド方式の検索可能暗号を用いた全文検索の実現方法を，処理手順に沿って説明する．図 23 は，本研究で対象とするデータ転移のユースケースを例に，処理手順を図示したものである．図中の番号は，処理手順番号と対応する．

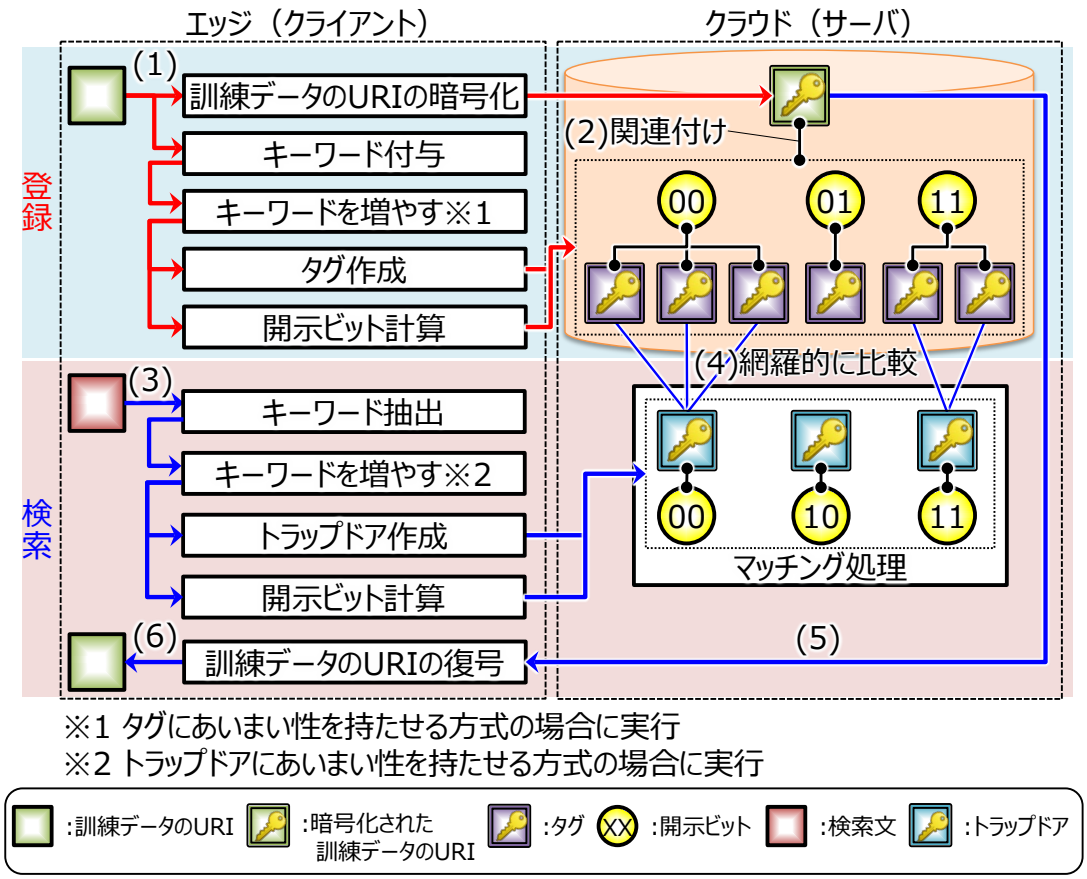


図 23 ハイブリッド方式を用い全文検索の処理手順

登録

(1) エッジ（クライアント）は，訓練データに付与するキーワードを決める．全文検索としてタグにあいまい性を持たせる方式を使用する場合は，キーワードの類義語などを追加する．次に，キーワードからタグと開示ビットを作成する．図 23 の例では，開示ビット値が「00」に対応するタグが 3 つ，「01」に対応するタグが 1 つ，「11」に対応するタグが 2 つ作成される．並行して，訓練データの URI を暗号化鍵で暗号化する．最後に，暗号化された訓練データの URI，タグ，開示ビットをクラウド（サーバ）に送信する．

(2) クラウドは，同じ開示ビット値とタグをグループ化したうえで，暗号化された訓練データの URI と関連付けて DB に保管することで，転置索引を作成する．表 8 に転置索引の例を示す．表中の文書 ID には，別途管理されている暗号化された訓練データの URI のテーブルの主キーが入り，外部キー制約を持つ（表 9）．

表 8 開示ビットを導入した転置索引例

開示ビット	タグ	文書 ID （外部キー制約の子カラム）
00	d96b4f8a9d	File1
00	a0663e3e41	File2
01	17b4539cd4	File2
…	…	…

表 9 文書 ID と暗号化された訓練データの URI を管理するテーブル例

文書 ID （外部キー制約の親カラム，主キー）	暗号化された訓練データの URI
File1	53616c7465645f5f0870…
File2	f2578f7adb62028db543…
File3	ae770d654675fa9ca061…
…	…

検索

(3) エッジは，データ転移の候補となる訓練データを得るために，検索文（自然文）を作成し，その検索文からキーワードを抽出する．全文検索としてトラップドアにあいまい性を持たせる方式を使用する場合は，キーワードの類義語などを追加する．次に，キーワードからトラップドアと開示ビットを作成する．図 23 の例では，開示ビット値が「00」「10」「11」に対応するトラップドアが 1 つずつ作成される．最後に，これらのトラップドアと開示ビットの組をクラウドに送信する．

(4) クラウドは，まず，DB 内からエッジから送信された開示ビットと一致するタグを特する．図 23 の例では，トラップドアに紐づく開示ビット値「00」と一致する DB 内の 3 つのタグを特定する．同様に，「11」と一致するタグ 2 つを特定する．なお，開示ビット値「10」をもつタグは，DB 内に存在しない．そして，開示ビットによって絞られたタグとトラップドアを網羅的に比較する．図 23 の例では，開示ビット値「00」に紐づくトラップドアと，先ほど特定した 3 つのタグをマッチング処理する．同様に「11」に紐づくトラップドアと，先ほど特定した 2 つのタグをマッチング処理する．

(5) クラウドは，トラップドアと一致したタグに関連付く暗号化された訓練データの URI をエッジに返却する．図 23 の例では，開示ビット値「00」あるいは「11」のいずれかのマッチング処理で一致判定が出た場合，タグに紐づく暗号化された訓練データの URI を

返却する．

(6) エッジは，復号鍵で訓練データの URI を復号する．

以上が，ハイブリッド方式の検索可能暗号を用いた全文検索の処理フローである．この構成でも，2.1.4 項で述べた開示ビット長をどのように決定するかの課題が残る．したがって，検索速度とセキュリティを考慮した開示ビット長の決定手法を確立することを本研究の課題とする．

2.2 訓練データとして有効な画像かどうかを判定する技術

2.2.1 類似画像検索の方式

この項では，訓練データとして有効な画像を見つけるための類似画像検索について説明する．

3.1 節において，「見た目が似ている」画像が転移学習に有効であると述べた．3.3.3 項において，照明変動や，画像に対する被写体の割合の変化などが生じても，依然として転移学習に有効な画像であると説明した．したがって，これらの画像が検索できる可能性のある，類似検索技術の既存研究を紹介する．

類似画像検索については，表 10 に示す通り，深層学習を使用して類似性を判定する方法[110][111][112][113]や，特徴点マッチングを使用して類似性を判定する方法[114][115][116][117][118]などがある．

表 10 類似画像検索の実現手段

	類似検索の実現手段	
	深層学習	特徴点マッチング
利点	• 訓練データから特徴を自動的に獲得し，未知のデータとの類似性を判断可能	• 画像の特徴点・特徴量の抽出に**事前学習が不要**
欠点	• 訓練データが大量に必要 • 学習に時間がかかる（高負荷）c) • 画像の拡大縮小や回転に弱い • 特徴量を自動で抽出するため，結果が人の主観と異なる場合がある	• マッチング時に，画像の特徴点・特徴量どうしの類似度を別途計算する必要がある • 類似の度合いを制御しにくい →**類似度指標の選定が重要**

c) 学習済みモデルが存在する場合は，訓練データが不要なこともある．

深層学習を利用する方法は，訓練データから特徴を自動的に獲得し，未知のデータとの類似性を判断できるが，訓練データを大量に要求したり，画像の拡大縮小や回転に弱かったり，特徴量を自動で抽出するため，結果が人の主観と異なる場合がある[119]（図 24 上段）．一方，特徴点マッチングは，画像の特徴点・特徴量の抽出において学習の必要がなく，事前に多くの訓練データを要求しないが，マッチングを行う際に画像の特徴点・特徴量どうしの類似度を別途計算する必要がある．類似度指標を適切に設定しないと，類似の度合いを制御しにくくなる恐れがあり[119]，類似度指標の選定が重要となる（図 24 下段）．

1.2 節で述べたように，訓練データはビジネス上重要な資産であり，その利用に許諾が必要なため，訓練データを大量に要求する深層学習の利用は難しい．加えて，訓練データとして，拡大縮小画像や回転画像を利用すると，出力されるモデルの推論におけるロバスト性の向上が期待できるため[120]，このような訓練データを取り逃す恐れのある深層学習は，転移学習にとって好ましくない．さらに，深層学習の訓練には多くの計算リソースを消費するため[121]，エッジのようなハードウェアリソースが小さい環境での実行が難しい．1.3 節で述べた Few-shot learning などの技術を使用して，訓練データ量の削減や学習の処理負荷を軽減することはできるが，事前学習済のモデルの存在が前提となる．

これらの理由から，本研究では，事前学習が不要な特徴点マッチングをベースに，転移学習に好ましい訓練データが検索できるような類似度指標を検討する．なお，エッジ上にデータが集まり始めた段階で深層学習による方式に切り替えても良い．特徴点マッチングを採用したのは，十分な数のデータが手元にない状態で類似性を判断できるからである．

特徴点マッチングは，特徴点や特徴量の抽出と，マッチング処理（類似度の計算）で構成される．深層学習の自動的に特徴を抽出する部分を，特徴点や特徴量の抽出に使用することができるが，前述のように特徴獲得に多くの訓練データが必要であるため，本研究では，深層学習による特徴抽出を行わない．

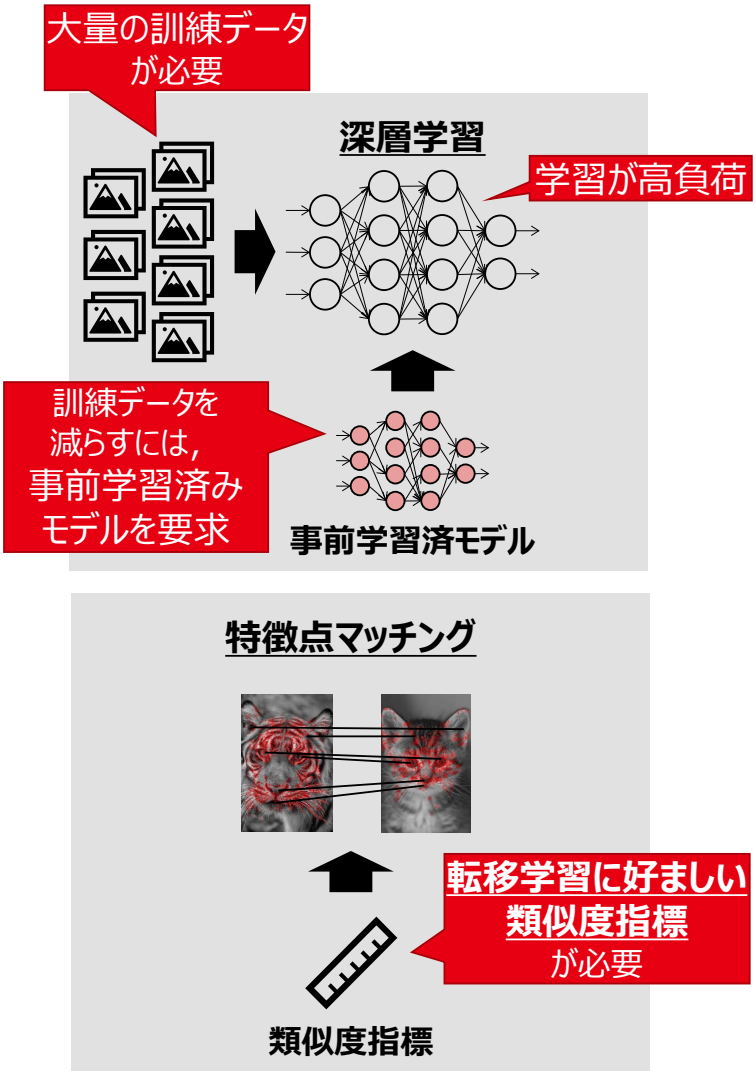


図 24 類似画像検索の実現課題

2.2.2 特徴点マッチング

この項では，特徴点マッチングの概要について説明する．

2.2.1 項で述べたように，特徴点マッチングは，特徴抽出と類似度の計算で構成される（図 25）．まず，比較する 2 つの画像から，それぞれの特徴点や特徴ベクトル，または，画素のヒストグラムなどの特徴を抽出する．次に，類似度指標を使用して，それらの特徴どうしを比較する．比較結果は，数値として出力され，通常，元画像どうしが似ているほど高く，または，低くなる．この関係を用いて，2 つの画像が類似しているかを判断する．

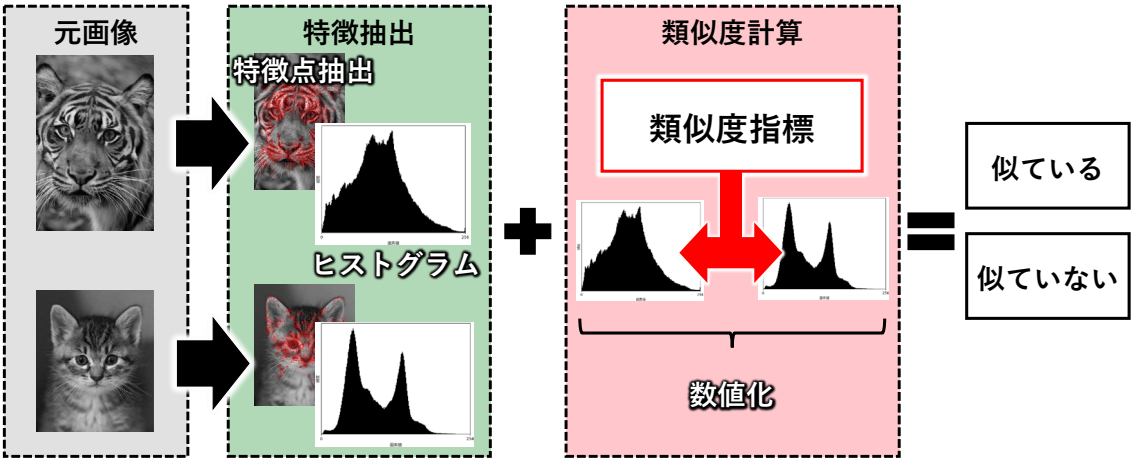


図 25 特徴点マッチング

特徴抽出手法は，多くの手法が存在するが，本研究では AI の訓練でよく用いられる画像を対象に，代表的な手法を説明する．

類似度計算手法は，ヒストグラムの形状変化が生じても類似度を高く算出できる効果が見込める，時系列データの類似度を測る手法に着目して説明する．

データ転移の対象となる訓練画像は，例えば，環境光などの変化に起因する照明変動によって，同一被写体でもヒストグラム等の特徴量が変化する場合がある[122]．昼の画像では，明るい画素値の出現頻度が増え，夜の画像では，暗い画素値の出現頻度が増えるのは，太陽光による照明変動が影響している．しかし，被写体が持つ本来のヒストグラム形状は，照明変動が生じても，ヒストグラム形状が明るい方か暗い方にシフトするだけで，形状の特徴は保存されているはずである．同様に，被写体が画面に占める割合が変化する場合，背景除去などを適切に行うことで，被写体が持つ本来のヒストグラム形状は，相似形として取得できる d)．被写体のコントラストが低い画像は，ヒストグラムの分布が狭くなり，コントラストが高い画像は，ヒストグラムの分布が広くなるため，ヒストグラム形状が伸縮することがある[123]．このように，ヒストグラム形状がシフトしたり，伸縮したり，相似形である時は，データ転移が有効な場合があるため，これらの形状変化に対しては類似度を高く算出するような手法が，本研究のユースケースには適している．

2.2.3 特徴抽出手法

この項では，特徴点マッチングに使用する特徴抽出手法について説明する．

画像検索分野で代表的な特徴抽出として，以下のようなものがある．

(A) 画素値のヒストグラム

(B) Bag-of-Features（BoF）

d) 相似な平面図形の面積比は，相似比の二乗となるため．

(C) Oriented FAST and Rotated BRIEF（ORB）
(D) Histograms of Oriented Gradients（HOG）
(E) Scale Invariant Feature Transform（SIFT）
(F) Speeded-Up Robust Features（SURF）
(G) Accelerated KAZE（AKAZE）

以降で，各特徴抽出について詳細を説明する．

A. 画素値のヒストグラム

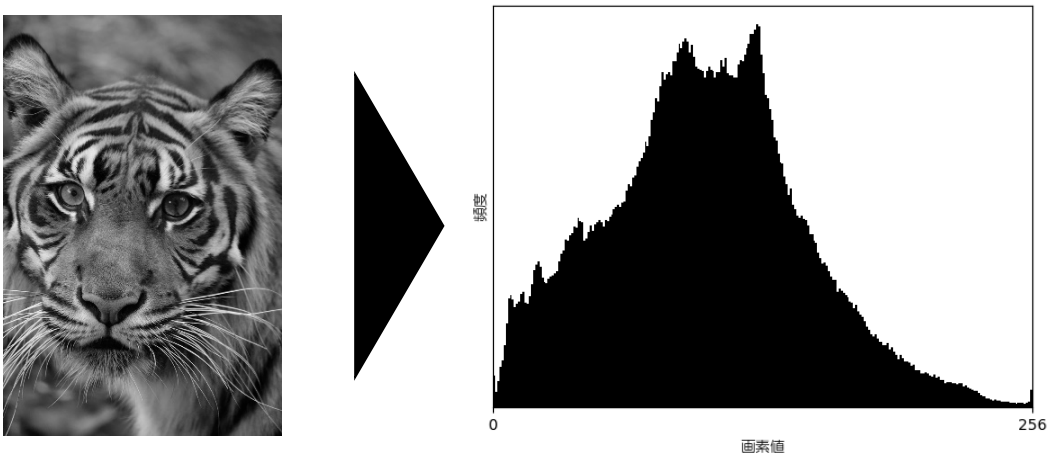
画像の画素値のヒストグラムは，x軸に画素値，y軸に頻度（画素値の個数）を取った統計グラフの一種である．画素値の最大値や最小値，平均値などを画像の特徴量とみなすことができる．

画像をヒストグラム化する場合，カラー画像をグレースケール化してから単一のヒストグラムを生成する方法（図 26 (A)）と，カラー画像の RGB チャネルごとに 3 つのヒストグラムを生成する方法（図 26 (B)）がある．機械学習では，特徴量の抽出が簡単なグレースケール化が良く用いられる．

グレースケール化した後のヒストグラムは，明るさやコントラストを変化させると，その形状が変化する．図 27 (A)がオリジナル画像とそのヒストグラムで，図 27 (B)が明るさを+20%した場合のヒストグラムである．ヒストグラムの形状が全体的に右（明るい方）にシフトしていることが分かる．図 27 (C)はコントラストを+20%した場合のヒストグラムである．コントラストが高いと画素値の分布が広がることが分かる．逆にコントラストを下げると，画像が同系色に近づくため，画素値の分布が狭くなる．図 27 (D)は，明るさとコントラストを各+20%したものである．ヒストグラムの形状が全体的に右（明るい方）にシフトし，画素値の分布が広がっていることが分かる．

このように，特徴量として簡単に利用できる一方で，同一被写体であっても環境要因によって，ヒストグラムの形状が変化し，特徴が変化する可能性がある．よって，ヒストグラムを特徴量として使用する場合は，このようなヒストグラムの形状変化を考慮する必要がある．ヒストグラム化すると，その画像の特徴の位置情報が失われる．たとえば，図 26 のトラの画像にあるような縞模様の特徴は，ヒストグラム化すると，縞を構成する画素値をピークとする形状が現れる．しかし，ヒストグラムからは，その縞模様が画像中のどこに存在するのか，縞模様なのか単に縞を構成する画素値が多く使われているだけなのか，など，特徴の位置に関する情報が失われてしまう．ヒストグラムを使用する際は，特徴の位置情報の欠損も考慮する必要がある．

A.グレースケールヒストグラム



B.カラーヒストグラム

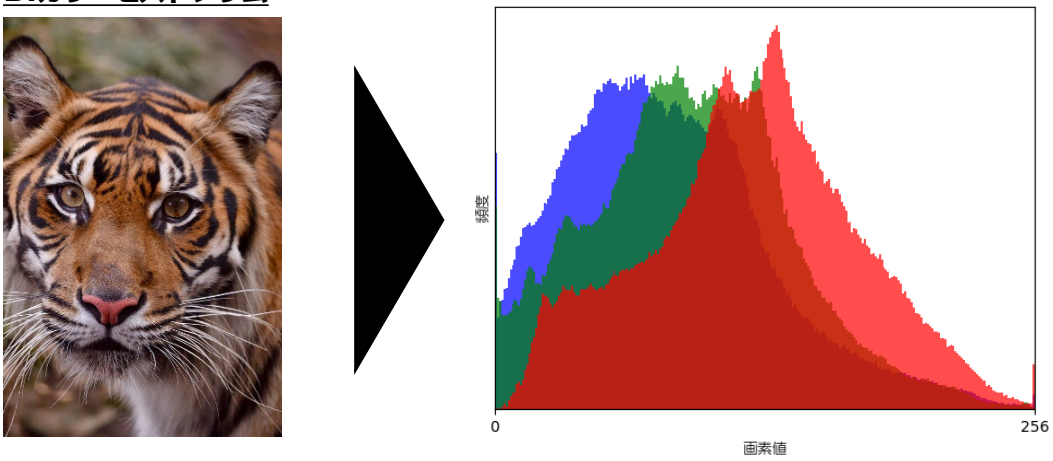
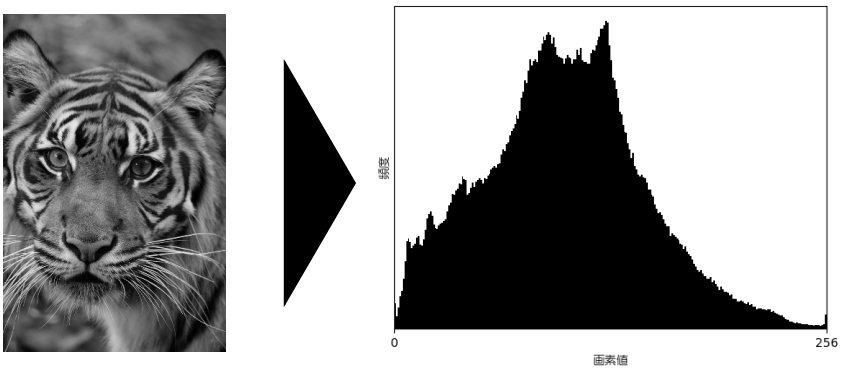
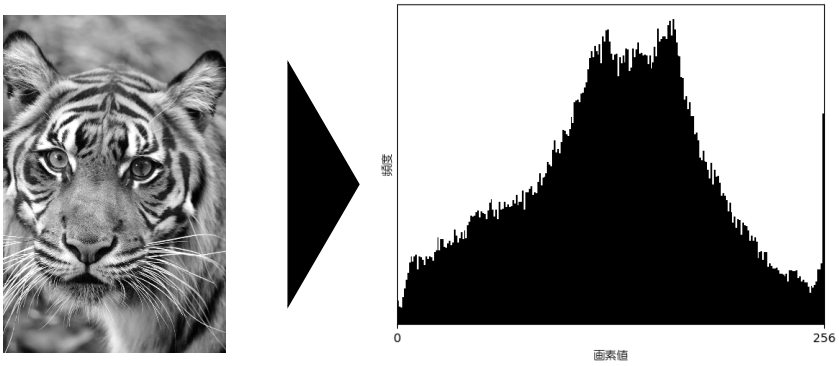


図 26 ヒストグラムの例

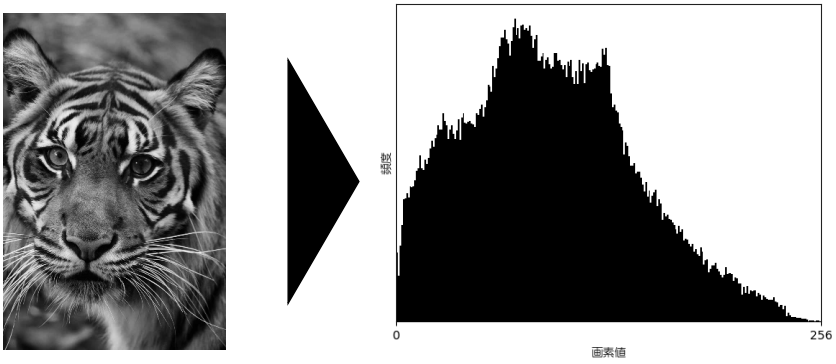
A.オリジナル



B.明るさ+20%



C.コントラスト+20%



D.明るさ+20%，コントラスト+20%

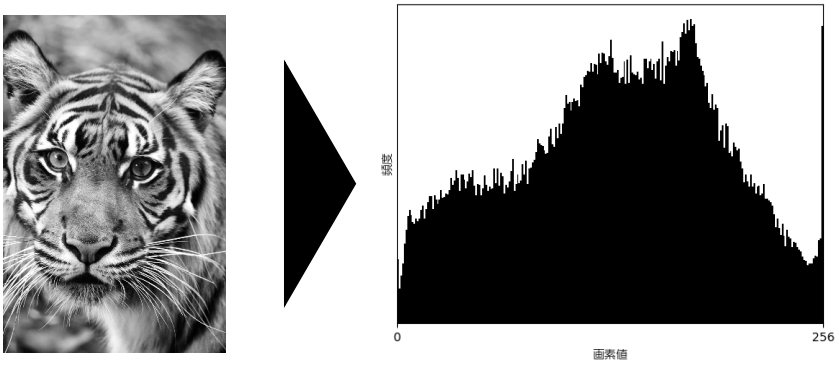


図 27 明るさやコントラストによるヒストグラムの形状変化

B. Bag-of-Features（BoF）[124]

Bag-of-Features（BoF）は，入力画像から局所的な特徴（局所特徴）を抽出し，それをヒストグラム化したものである．最終的に画像がヒストグラムで表現されるため，最大値や最小値，平均値などをその画像の特徴量とみなすことができる．

図 28 は，BoF の処理手順を示したものである．以下の処理手順の番号は，図中の番号と対応する．

(1) 入力画像

入力画像は特徴を検出しやすくするために，グレースケール化することが多い．検出される局所特徴の数をある程度そろえるために，画像サイズや解像度を他の画像と合わせることもある．

(2) 局所特徴ベクトル抽出

入力画像から局所特徴量（局所特徴ベクトル）を抽出する．抽出に用いるアルゴリズムは，後述する ORB，HOG，SIFT，SURF，AKAZE などを用いることが多い．

(3) クラスタリング

局所特徴ベクトルをクラスタリングし，局所特徴ベクトルをクラスごとに分類する．局所特徴ベクトルが高次元であることがあり，そのままだとヒストグラム化が困難である．よって，局所特徴ベクトルを一定数のクラスに分類することで，特徴次元を削減する．

(4) ヒストグラム化

クラス分類した局所特徴量を基に，x軸にクラス ID，y軸に頻度（そのクラスに分類された局所特徴ベクトルの数）を取った統計グラフを作成する．

このように，BoF は，どのような特徴抽出アルゴリズムを用いてもヒストグラムを作成できるので，2 つのヒストグラムの違いを見ることで類似性を計測することができる．

一方で，クラスタリング時の最適なクラス数（クラスタ数）を自動的に求めることは難しい．通常は，クラスタ数を分析者が事前に決定するが，エルボー法[125]やシルエット分析[126]などを用いて，設定したクラスタ数がおおよそ正しいかを確かめることができる．しかしながら，これらの方法は，実際にクラスタリングを試す必要があり，データ規模が大きい場合は非常に時間がかかる．さらに，分析が必ずうまくいくとは限らず，分析結果を最終的に人が判断するため，最適値かどうかは判断が難しい．

クラスタリングは，事前学習が必要なことが多い．その場合，ある程度の数の訓練データが要求されるので，訓練データ数を十分確保できない場合は，クラスタリングを実行できないか，十分な精度が出ない．事前学習が不要な場合でも，ある程度の数の訓練データがないとクラスタを構成できないことがある．さらに，訓練データを追加した際にクラスタ間の境

界が変化し，クラス分類の結果が変わる可能性があり，出力されるヒストグラムの形状がクラスタリングの実行タイミングによって異なってしまう．よって，BoF を安定して動作させるには，事前に十分な数の訓練データを用意しなければならない．

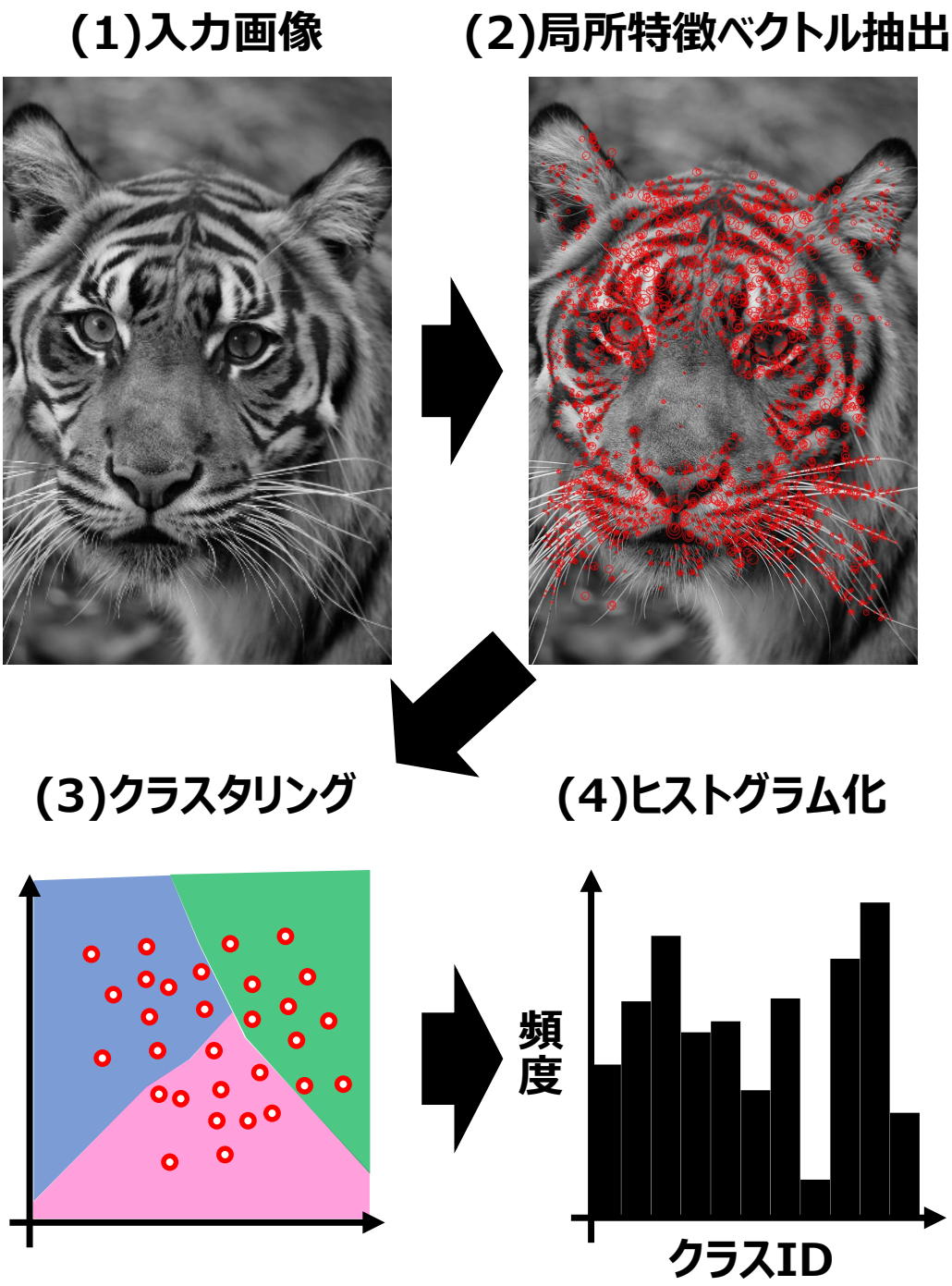


図 28 BoF の処理手順

C. Oriented FAST and Rotated BRIEF（ORB）[114]

ORB は，基本的に FAST[127]による特徴点検出（コーナー検出）と BRIEF[128]による特徴量記述子を組合わせたものである．決定木を利用したコーナー判定によるキーポイント抽出を行う FAST に回転不変性 e)を導入し，さらに，特徴ベクトルの計算に使用する BRIEF にも回転不変性を持たせた rBRIEF と呼ばれる改良を施している．コーナー判定に，スケール不変性 f)を持たせるために，画像ピラミッドと呼ばれる段階的に画像を縮小（ダウンサンプリング）したものを生成する．特徴ベクトルはバイナリ化[129][130][131][132]されており，高次元の実数ベクトルで表現する場合に比べ，省メモリ化されている．

SIFT や SURF より高速に動作し，SURF より精度が良い[114]．元の画像から画像ピラミッドを生成するので，メモリ消費量は若干高い．後述する SIFT や SURF と異なり特許権で保護されていない．

D. Histograms of Oriented Gradients（HOG）[115]

HOG は，局所領域（セル）の画素値の勾配方向をカウントし，ヒストグラム化したものを特徴量として使用する．主に物体検出に使用される．HOG は，画像内の局所的な物体の外観と形状が，勾配またはエッジ方向の分布によって記述できるという考えに基づいている．

勾配を特徴量としているため，照明の変化と，局所特徴の平行移動に不変性を持つが，回転やスケール変化に弱い．画像に含まれる背景が複雑であると，勾配が多数現れてヒストグラムにノイズを含んでしまう可能性がある．勾配計算の過程で正規化を行うため，若干計算負荷が高い．ORB と異なり元の画像のみを用いるが，セルを少しずつずらしながら網羅的に勾配を計算するため，適切に並列化しないと実行速度が低下する．基本となる考え方は，特許で保護されている．

E. Scale Invariant Feature Transform（SIFT）[116][133]

SIFT は，スケールスペースと呼ばれる，元の 2 次元画像に対して段階的に平滑化（ぼかし）した画像群（スケール画像）を加えた 3 次元空間を利用して特徴点の検出を行う．スケールスペースにおいて，スケール方向に差分を取った画像（DoG : Difference-of-Gaussian 画像）内の極値を取る点を特徴点の候補とする．つまり，平滑化に対し不変な点（ぼかしに強い点）を候補としている．候補のうち，DoG 出力値が閾値以下のもの（極値を取るものの，変化量が少ない点）などを除外し，最終的な特徴点とする．次に，特徴点の周辺の輝度勾配ヒストグラムを作成する．この処理は HOG とほぼ同じである．

---

e) 画像内の局所特徴に対して任意の回転変換（対象の中心を通る任意の軸に対する任意角度の回転）が加えられたとしても，特徴抽出が可能であること

f) 画像内の局所特徴が拡大，または，縮小した場合も，特徴抽出が可能であること

これらの処理により，SIFT は，照明，回転，スケールに対して不変性を持つ．スケールスペースの計算負荷が高いため，実行速度は遅い．特徴点を高次元の実数ベクトルで表現するため，メモリ消費量は多い．特許権で保護されていることに注意する．

F. Speeded-Up Robust Features（SURF）[134]

SURF は，SIFT を高速化したものである．高速化のために，DoG 画像の代わりに Box filter を利用する．Box filter の畳み込み（ぼかしを強くすることに相当）は，積分画像を使用することで簡単に求められる．積分画像は，画素値を積分したものなので，DoG 画像のように逐次的にぼかしの強度を変化させて算出する必要がない．そのため，スケール方向の画像の生成を並列化できるため非常に高速である．その他にも，SIFT における各処理で高速化を施すことで SURF は，SIFT に比べ 3 倍の高速化を実現している[134].

SURF も SIFT と同様に，特許権で保護されている．

G. Accelerated KAZE（AKAZE）[117][135]

AKAZE は，SIFT や SURF の欠点を改善した KAZE[136]と同等の高い認識精度を持ちつつ，計算に必要な時間を大幅に短縮したものである．AKAZE は，スケールスペースの算出に非線形で非等方的なフィルタによる平滑化を用いる．SIFT や SURF で使われている Gaussian filter によるスケールスペースでは，平滑化が等方的であるため，物体のエッジもぼやかしてしまい，局所的な特徴をうまくとれないことがある．AKAZE の非線形フィルタは，この問題を解決している．さらに，特徴点記述子に，Modified-Local Difference Binary (M-LDB)という独自の記述子を使用しており，ピラミッド構造の計算を高速化するための独自の工夫を組み入れることで，ロバスト性の向上と高速化を図っている．

AKAZE は，照明不変，スケール不変，回転不変である．なお，AKAZE は特許権で保護されていない．

以上が，画像検索分野で代表的な特徴抽出手法である．なお，(C)～(G)の特徴抽出法の多くは，高次元の特徴ベクトルを出力するため，マッチング処理（類似度の算出）にそのまま入力するには，高次元データを扱えるユークリッド距離（L^2ノルム）やハミング距離を使用する g).

表 4 に，その一覧と特徴をまとめたものを示す．3.3.3 項で示したように，エッジ上で処理を実行するため，実行速度（CPU 使用量），メモリ消費量を考慮しつつ，特徴抽出の精度の高いものを選択する必要がある．なお，特徴抽出の精度は，カバーする不変性が多いほど高いと考えられる．さらに，ビジネス化を考える上では，特許保護の有無も重要な要素となる．特許料の支払いは出来るだけ避けたほうが良い．したがって，本項で挙げた特徴抽出手

g) 実数の高次元特徴ベクトルの場合は，ユークリッド距離を，特徴ベクトルがバイナリ化されている場合は，ハミング距離を使用することが多い．

法のうち，CPU 使用量とメモリ消費量が少なく，多くの不変性を持ち，特許保護がされていないものは，(A)ヒストグラムのみとなる．本研究では，特徴抽出に画像のヒストグラムを用いることにする．

ただし，(C)～(G)の特徴抽出法を使用して高次元の特徴ベクトルを得る場合でも，BoF や主成分分析（PCA：Principal Component Analysis）によって次元を削減し，ヒストグラムとして扱うことができる．よって，リソース消費量の問題や特許保護はあるが，本質的には，本研究で提案する時系列データの類似度測定法をベースとした手法を適用可能である．

表 11 画像検索分野で代表的な特徴抽出手法の一覧

	画像の統計グラフ化		特徴点の抽出と特徴量記述子（特徴ベクトル）の作成				
	(A)ヒストグラム	(B)BoF	(C)ORB	(D)HOG	(E)SIFT	(F)SURF	(G)AKAZE
実行速度（CPU消費量）	**高速**	低速※	**高速**	普通	低速	普通	**高速**
	• 画素値のカウントのみ（消費小）	※クラスタリングが高負荷な傾向あり（消費大）	• (E), (F)より高速 • (F)より高精度	• 正規化処理負荷が若干高い（消費中） • セルをスライドしなら勾配を計算（消費中）	• スケールスペースの計算負荷が高い（消費大）	• (E)の3倍高速	• 並列化による高速化が可能 • (E), (F)より高速
メモリ消費	**小**	大※	中	**小**	大	大	中
	• 元画像のみ使用（消費小）	※特徴ベクトルの表現形式による	• 特長ベクトルをバイナリ化（消費小） • 元画像から画像ピラミッドを生成（消費大）	• 元画像のみ使用（消費小）	• 特徴点を高次元実数ベクトルで表現（消費大） • 元画像からスケールスペースを生成（消費大）	同左	• 特長ベクトルをバイナリ化（消費小） • 元画像からスケールスペースを生成（消費大）
事前学習	不要	**必要**	不要	不要	不要	不要	不要
不変性	• **回転** • **平行移動** • **照明やスケール変化はヒストグラム形状から検知可能**	• 選択する特徴抽出アルゴリズムに依存	• 回転 • スケール	• 照明 • 平行移動	• 照明 • 回転 • スケール	• 照明 • 回転 • スケール	• 照明 • 回転 • スケール
特許による保護	なし	なし	なし	**あり**	**あり**	**あり**	なし
備考		• 特徴抽出に(C)～(G)などを使用	• 高次元特徴ベクトルのままマッチング処理（類似度の計算）を行う場合は，ユークリッド距離やハミング距離等を使用 • 特徴ベクトルの次元削減には，BoF や PCA を使用				

2.2.4 類似度計算手法

この項では，この項では，特徴点マッチングに使用する類似度計算手法について説明する．

類似度計算手法は，数多く存在するが，ヒストグラムの形状変化が生じても類似度を高く算出できる効果が見込める，時系列データの類似度を測る手法に絞って説明する．

Aghabozorgi らのサーベイ[137]によれば，時系列データの類似度計算手法は *Shape-based*，*Compression-based*，*Feature-based*，*Model-based* の 4 つの方式がある．

Shape-based は，2 組の時系列データの形状を，時間軸に対する非線形な伸縮によって可能な限り一致させようとするもので，ユークリッド距離，ハミング距離，DTW[148]，LCSS[138]，MVM[139]などがある．これらは，短い時系列データに適用することが多い．この方式は，時間軸の影響を無視するため，ヒストグラムの形状のシフトに対応できるが，シフト量が大きすぎると画像としての類似性が低下する恐れがある．ヒストグラムの形状のうち高さを厳密に判定するために，幅広く類似画像を検索するには極値の高さの変化に対し許容を持たせる必要がある．

Compression-based は，時系列データを何らかの形で圧縮（変換）し，圧縮後の情報から類似性を求めるもので，CDM[140]，自己相関，カルバック・ライブラー情報量[141]，ピアソン相関係数などがある．これらは，時系列の長さによらず使用できる．この方式は，圧縮方法に精度が左右されるため，扱うデータによって適切な圧縮方法をユーザが選択する必要がある．

Feature-based は，時系列データを低次元の特徴量に変換し，特徴量から類似性を求めるもので，統計学（統計的仮説検定など），係数（Jaccard 係数[142]，Dice 係数[143][144]，Simpson 係数[145]など），コサイン類似度などがある．これらは，長い時系列に適用することが多い．この方式も *Compression-based* と同様に，適切な次元削減方法を選択する必要がある．

Model-based は，時系列データの予測モデルがどの程度類似しているかを求めるもので，HMM[146]，ARIMA[147]などのモデルを利用する．これらは，長い時系列に適用することが多い．この方式も *Compression-based*，*Feature-based* と同様に，モデルの選定が精度に大きく影響する．

本研究では，ユーザによる調整が不要な *Shape-based* 方式を用いる．*Shape-based* 方式でよく使用されるのは，ユークリッド距離，ハミング距離，DTW である[137]．以降で，ユークリッド距離，ハミング距離，DTW について詳細を述べる．

A. ユークリッド距離

ユークリッド距離は，L^2ノルムとも呼ばれ，2 つの点（ベクトル）をつなぐ線分の長さを表す．ベクトル$\vec{x}$=$(x_1, \dots, x_n)$，$\vec{y} = (y_1, \dots, y_n)$のユークリッド距離$\mathrm{d}(\vec{x}, \vec{y})$は，数式(1)のように定義される．

$$\mathrm{d}(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

ユークリッド距離は，データの差の二乗和の平方根であり，ベクトルの成分間の重みを考慮しない．本研究では，画像のヒストグラムの比較におけるヒストグラムの形状のシフトに対しては類似度への影響を少なくし，ヒストグラムの形状が相似形である場合は，類似度を高く算出する性質を持つことが望ましい．しかしながら，ユークリッド距離は，座標系の各軸の値のずれの程度が異なる意味を持つ状況に対応することができない．

B. ハミング距離

ハミング距離は，等しい文字数を持つ 2 つの文字列の中で，対応する位置にある異なった文字の個数である．換言すると，ある文字列を別の文字列に変形する際に必要な置換回数を計測したものとなる．

ハミング距離の例は以下のようになる．

- 1011101 と 1001001 の間のハミング距離は 2
- 2173896 と 2233796 の間のハミング距離は 3
- 「toned」と「roses」の間のハミング距離は 3

たとえば，等しい長さを持つ 2 つのバイナリ列を$\boldsymbol{a}$=$(a_1,\dots,a_n)$，$\boldsymbol{b} = (b_1,\dots,b_n)$とすると，ハミング距離$\boldsymbol{d}(\boldsymbol{a},\boldsymbol{b})$は数式(2)のように定義できる．

$$\boldsymbol{d}(\boldsymbol{a},\boldsymbol{b}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(a_i + b_i)\mathbf{mod}\ 2 \tag{2}$$

ただし，$a_i, b_i \in \{0,1\}, i = 1,\dots,n$ である．$(0+0)\mathbf{mod}\ 2 = (1+1)\mathbf{mod}\ 2 = 0$，$(0+1)\mathbf{mod}\ 2 = (1+0)\mathbf{mod}\ 2 = 1$となることから，$a_i \neq b_i$の場合のみ 1 が出力されるので，ハミング距離の定義に合った計算ができることが分かる．

一般的なハミング距離の定義は，長さの等しい 2 つの任意の文字列を$\boldsymbol{a}$=$(a_1,\dots,a_n)$，$\boldsymbol{b} = (b_1,\dots,b_n)$とすると，数式(3)のようになる．

$$\boldsymbol{d}(\boldsymbol{a},\boldsymbol{b}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{d_i}, \qquad \boldsymbol{d_i} = \begin{cases}\mathbf{0}, & a_i = b_i \\ \mathbf{1}, & a_i \neq b_i\end{cases} \tag{3}$$

ハミング距離は，高次元の特徴ベクトルをバイナリ化[129][130][131][132]したものに対

する類似度の算出に利用できる．しかし，本研究で想定するような画像のヒストグラムの形状のシフトなどに対しては，ハミング距離が大きく算出されてしまい類似性を正しく判断できない．

C. Dynamic Time Warping (DTW)

DTW は，２つの時系列データの各点の距離を総当たりで比較し，２つの時系列データの距離が最短となるパスを見つける手法である．ここでの距離指標に特に制限はないが，誤差の絶対値などを使用することが多い．基本的な概念は 1959 年に提案され[148]，その後，音声認識，筆跡照合，手話認識，データマイニング，時系列データクラスタリングなど幅広い分野に応用されている[149]．

２つの時系列データから類似波形を探索するため，データ分布の形状が異なる場合は類似度を低く算出し，波形が時間軸に対し平行移動したようなデータ分布の形状がシフトするパターンや形状が伸縮するパターンの類似度は高く算出する．このことから，本研究で求める性質に近い性質を持つ．

表 12 に，ユークリッド距離，ハミング距離，DTW の得失の一覧を示す．本研究では，ヒストグラムの形状のシフトや伸縮に強い，DTW をベースとして用いる．

表 12 *Shape-based* 方式の得失

	ユークリッド距離	ハミング距離	DTW
概要	• データの差の二乗和の平方根	• 等しい文字数を持つ 2 つの文字列の中で，対応する位置にある異なった文字の個数	• 2 つの時系列データの距離が最短となるパス長
利点	• 高次元ベクトル間の比較が可能	• バイナリ化したベクトル間の比較を高速に算出可能	• 時間的なスケールが異なっていても比較ができる
欠点	• ベクトルの成分間の重みを考慮しない	• ベクトルの成分が，異なっているか否かのみで距離を算出するため，異なっている度合いを考慮できない	• 値の変動に鋭敏であり，ピークの高さが異なると距離が長く算出される恐れがある
本研究への適正	• 画像のヒストグラムの形状変化の種類によって，類似度への影響を変えることができないため 不向き	• 画像のヒストグラムの形状のシフトなどに対しては，距離が大きく算出されてしまい類似性を正しく判断できないので 不向き	• 画像のヒストグラムのピークの数や高さが異なる場合などのヒストグラムの形状が異なる場合はペナルティが大きく，形状がシフトするパターンや形状が伸縮するパターンには，ペナルティが低いので，向いている

2.2.5 DTW の改良手法

この項では，DTW の改良手法について述べる．そして，転移学習に好ましい訓練データが検索可能な特徴点マッチングの類似度指標が必要になることを説明する．

2.2.4 項において，DTW が幅広い分野に応用されていることを述べた．それに伴い，様々な改良手法が提案されている[150][151]．しかしながら，ほとんどの改良手法は，DTW における 2 つの時系列データの各点の距離を総当たり的に探索する処理を高速化したものである．

DTW を画像の検索に応用すると，いくつか問題が生じる．DTW は，検索対象となる長い時系列波形から，短い部分時系列波形を発見するものである．このことから，時間軸（横軸）方向のずれに対しては，寛容である．一方，画像のヒストグラムに対して DTW を適用しようとすると，たとえば，横軸の範囲を[0,255]に絞ることになる．すると，横軸の値が 0 に近いところでピークを持つヒストグラムと，255 に近いところでピークを持つヒストグラムの類似度が高いと判定される場合がある．詳しくは 5.3.2 項で説明するが，このようなヒストグラムを持つ画像は，見た目が著しく異なってしまう．

このように，横軸方向のずれの度合いを制御するような DTW の改良方式は，著者が調べた限りでは存在しなかった．前述のとおり，DTW の目的が「検索対象となる長い時系列波形から，短い部分時系列波形を発見する」こと起因しており，時系列データに対して DTW を適用する限りでは，問題とならなかったためであると考えられる．したがって，DTW において横軸方向のずれの度合いの制御を必要とする問題は，横軸の範囲を絞ったことによる特有の問題である．

これらの事実から，訓練データとして有効な画像かどうかの判定するための，特徴点マッチングの類似度指標を確立しなければならない。よって，本研究の 2 つ目の課題を，「転移学習に好ましい訓練データが検索可能な特徴点マッチングの類似度指標」の開発とする．

2.3 既存研究の課題と本研究の狙い

この節では，2.1 節と 2.2 節で述べた既存研究の課題と，本研究の狙いを述べる．

暗号化されたキーワードどうしを検索する技術については，ハイブリッド方式の検索可能暗号を使用して構成可能であることを述べた．ただし，検索速度とセキュリティをバランスする開示ビット長を自動的に決定する手法が必要となる．

訓練データとして有効な画像かどうかを判定する技術については，類似画像検索のアルゴリズムに特徴点マッチングを使用するのが良いと述べた．さらに，画像のヒストグラムを特徴として用い，その特徴どうしの比較に必要な類似度指標は，DTW をベースとする方法が有力であった．ただし，ヒストグラムに対して DTW をそのまま適用すると，過剰なフィッティングにより転移学習に必要な類似性を正しく判断できない恐れがあるため，転移学習に好ましい訓練データが検索可能な特徴点マッチングの類似度指標を新たに作成する必

要がある．

以上から，本研究では，以下の 2 つの新しい手法を提案し，訓練データ検索システムを実現していく．

- 検索速度とセキュリティを考慮した検索可能暗号の高速化パラメータの決定手法
- 転移学習に好ましい訓練データが検索可能な特徴点マッチングの類似度指標

これらの提案手法の詳細は，それぞれ 4 章，5 章で説明する．

第3章 本研究の全体像

本章では，訓練データ検索システムにおける，2.3 節で述べた以下の 2 つの新しい提案手法の位置づけを述べる．

- 検索速度とセキュリティを考慮した検索可能暗号の高速化パラメータの決定手法
- 転移学習に好ましい訓練データが検索可能な特徴点マッチングの類似度指標

さらに，本研究の全体に係る前提条件や 2 つの提案手法の性能要件を述べる．

3.1 訓練データ検索システムにおける提案手法の位置づけ

訓練データ検索システムにおける，本研究で提案する 2 つの提案手法の位置づけを図 29 に示す．

1 つ目の提案手法である，検索速度とセキュリティを考慮した検索可能暗号の高速化パラメータの決定手法は，クラウド上で実行される暗号化されたキーワードどうしの検索に必要な，検索可能暗号のパラメータの設定に使用される．これにより，キーワード数に対するスケーラビリティとセキュリティを維持しながら，暗号化されたキーワードどうしを検索することが可能となる．すなわち，訓練データの候補となる画像を保有するエッジを，高速かつ安全に特定できるようになる．

2 つ目の提案手法である，転移学習に好ましい訓練データが検索可能な特徴点マッチングの類似度指標は，エッジ上で訓練データとして有効な画像かどうかを判定するための，特徴点マッチングの類似度指標に使用される．これにより，高精度で転移学習に有効な訓練データが特定できるようになり，品質の良い訓練データセットを構築可能となる．したがって，エッジで使用するモデルの推定精度を向上することができる．

上記 2 つの提案手法は，1.1 節で述べたように独立した 2 つの技術であり，各々の提案手法が実現されることで，結果的に訓練データ検索システムが構築可能となる．

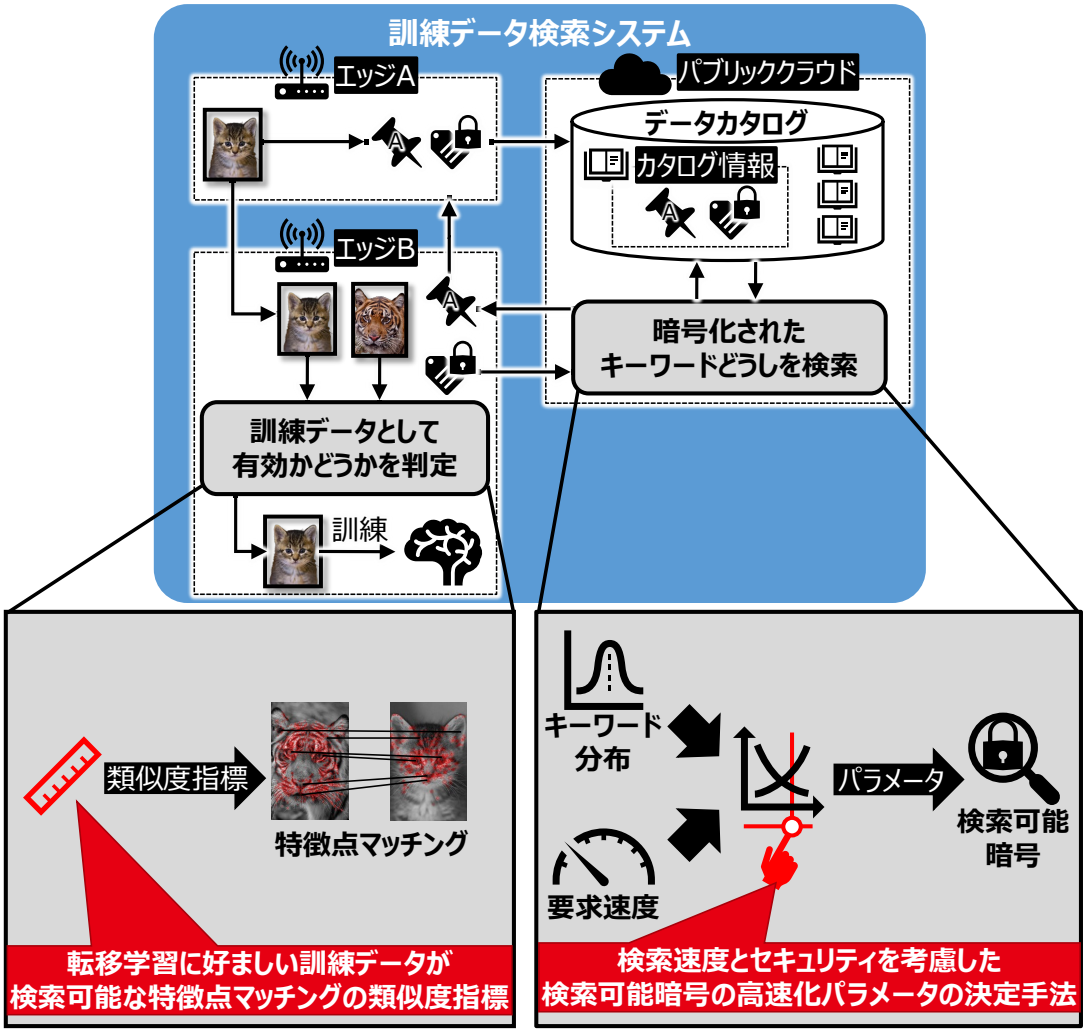


図 29 訓練データ検索システムにおける提案手法の位置づ

3.2 訓練データ検索システムを使用した転移学習の導入の効果

「見た目」が似ている画像を転移学習することで，モデルの推定精度が向上する事例を紹介し，訓練データ検索システムの利用が効果的であることを示す．

転移学習が成功する一つの条件は，1.5 節で述べたように，画像であれば見た目，時系列データであれば波形が似ているなどの類似度が高いデータを使用することである．訓練データ検索システムでは，画像どうしの照明変動や幾何学変換の影響を無視するため，オリジナルの画像の特徴が似ているものが検索される．簡単に言えば，「見た目」が似ている画像が選ばれやすくなる．

転移学習の利用で，AI の推論精度が向上した事例を紹介する．半導体製造プロセスの欠陥分類において，ある種の欠陥の生起確率が低いために，欠陥判定用の画像が不足し，推論精度が低下する問題がある．このような状況下での既存手法の平均正解率が 77.2%である

のに対し，転移学習の利用で 87.3%に向上したとの報告がある[152]．この報告において，熟練作業者が転移学習に有効な欠陥画像を選定し，転移先の環境下で生起確率の低い欠陥の画像数を増やすことが述べられている．転移学習に有効な欠陥画像は，熟練作業者が「見た目」を基に欠陥の種別を判断していると考えられる．すなわち，この報告の中の転移学習は，転移先の生起確率の低い欠陥画像と似ている欠陥画像を検索することと同義である．

あるクラスの画像の生起確率が低いようなユースケースは，たとえば，「定点監視カメラの画像から希少な動物を見つける」タスクや，「零下 30 度以下で、屋外装置の破損を発見する」タスクのようなものが考えられる．したがって，このようなユースケースでは，転移学習が有効であり，「見た目が似ている」画像を検索するニーズが潜在的にかなり存在すると予想できる．

前述のユースケースでは，訓練データ検索システムを使用した転移学習が有効であると予想でき，本研究成果の活用が期待できる．

3.3 本研究の全体に係る前提条件

訓練データ検索システムでは，まず，訓練データに付与されているキーワードを使用して，データ転移の候補となる画像を言語的に発見する．次に，転移元の候補画像と転移先のサンプルデータを使用して，非言語的（画像的）に類似度を算出することで，転移先エッジに対して転移学習が成功しやすいかどうかの判断材料を提示する．これら 2 つの処理は，理想的には転移先エッジからの要求に基づいて一気通貫で実行される．

訓練データ検索システムの評価は，2 つの検索処理をまとめた総合的な検索性能を見るべきではあるが，これら 2 つの処理は独立性が高く，学習フェーズで使用されることを踏まえると，高いリアルタイム性を要求しない．さらに，クラウド側で実行される暗号化されたキーワードどうしの検索は，キーワード数に対するスケーラビリティが重要視されるのに対し，エッジ側で実行される訓練データとして有効な画像かどうかを判定する処理は，画像の類似度の算出精度が重要視される．したがって本研究では，クラウドとエッジそれぞれで新しく開発する技術を個別に評価し，それらの評価結果を統合することで，訓練データ検索システムの有効性を示す．

以降の項では，2 つの評価において共通する前提条件と，各評価の性能要件について述べる．各評価の個別の前提条件については，それぞれ 4.1 節，5.1 節で述べる．性能要件は，4.1 節，5.1 節に再掲する．

3.3.1 訓練データ

本研究で扱う訓練データは，画像であり，その画像にはキーワードが付与されていることを前提とする．このキーワードは，AI で用いる教師データ（正解ラベル）のほか，訓練データを取得した日時，場所等の周辺環境の情報や，画像サイズや解像度などのメタデータ，

人物であれば性別や人種，年齢などのパーソナルデータも含む．ただし，本研究では，これらの機微なデータを研究用途で使用することが困難であるため，著作権フリーな書籍データから抽出したキーワードを類義語で拡張したもの代替として使用し，データ規模のみを合わせる．このように変更しても，キーワード数に対するスケーラビリティの評価に何ら影響はない．

暗号化されたキーワードどうしの検索の評価は，検索性能のボトルネックとなるデータ規模のみに着目し，検索の精度については対象外とする．この理由は，暗号化されたキーワードどうしの検索の精度は，既存の類似検索技術の一つである類義語展開の精度に依存するためである．本研究では，類義語展開は十分な精度で行われるとみなす．

検索対象となる訓練画像は，ビジネスシーンを想定したデータを想定するのが理想的である．たとえば，個別最適化が必要なユースケースにおける，工場の生産ラインでの検品作業で使用される良品・不良品の画像や，住宅内の人物の行動認識用に記録された動画，自動運転向けのドライブレコーダの動画が挙げられる．しかしながら，これらのデータはパーソナルデータを含むか，ビジネス上の機微なデータであるため研究用途では使用できない．そのため，本研究では，研究用途で公開されている動物の画像を代替として用いる．特に，画像どうしの類似性を一目で判断できるように，体表に縞模様などの特徴的な柄を持つ猫科の動物を中心にデータを集め，評価に用いる．

3.3.2 信頼モデル

本研究では，図 30 に示すような信頼モデルを想定する．この想定は，暗号化されたキーワードどうしの検索を実現する技術の一つである，検索可能暗号を利用する場合の想定と同じであり，すでに検索可能暗号を利用した製品化事例[153][154]や実用化予定[155]がいくつかあるため，妥当なものである．

本研究の信頼モデルでは，エッジは信頼できるものとする．エッジの構成要素であるゲートウェイは，データ生成元であるデバイスに近く，デバイスの管理者が管理を兼ねていることも多いため，ゲートウェイやデバイスに悪意のある攻撃者が侵入することは想定しない．インターネットは，盗聴等の恐れがあるため信頼できないため，インターネットへ流れるデータはすべて暗号化する必要がある．クラウドは，信頼できる範囲とできない範囲が共存するものとする．具体的には，データベースの機能であるデータの検索や挿入は，正常に実施されるものとして信頼する．一方，クラウド管理者には，悪意のある者がおり，データベースに保管されているデータを閲覧できるものとする．ただし，データの改ざんや破棄はできないものとする．

クラウドの多くは，SaaS（Software as a Service）としてデータストアの機能を外部に公開している．悪意のあるクラウド管理者によってデータベース内部のデータが改ざん・破棄されたり，データベースへのクエリの結果が異常であったりすると，SaaS の品質が低下し，利用者が減少することになる．そうすると，結果的には，悪意のあるクラウド管

理者の攻撃対象であるデータが減少してしまう．したがって，攻撃を高い確率で成功させたい悪意のあるクラウド管理者は，より多くの SaaS 利用者を集めるために，できる限り正常な SaaS を装うと想定できる．このような，「正常なサーバのように振る舞うが，盗み見のみを行う」攻撃者を passive な攻撃者と呼ぶ[91].

暗号化に必要な秘密鍵は，エッジによって秘密に管理され，漏洩することがないものとする．さらに，秘密鍵は十分な強度を持ち，攻撃者によって現実的な時間内で秘密鍵を復元したり，暗号化データを秘密鍵なしに復号したりすることができないものとする．

エッジやクラウドに関する上記以外の想定については，1.2 節で示した IoT セキュリティに関するガイドラインやプライバシー保護に対する法令を遵守するものとし，十分な対策がなされているものとする．

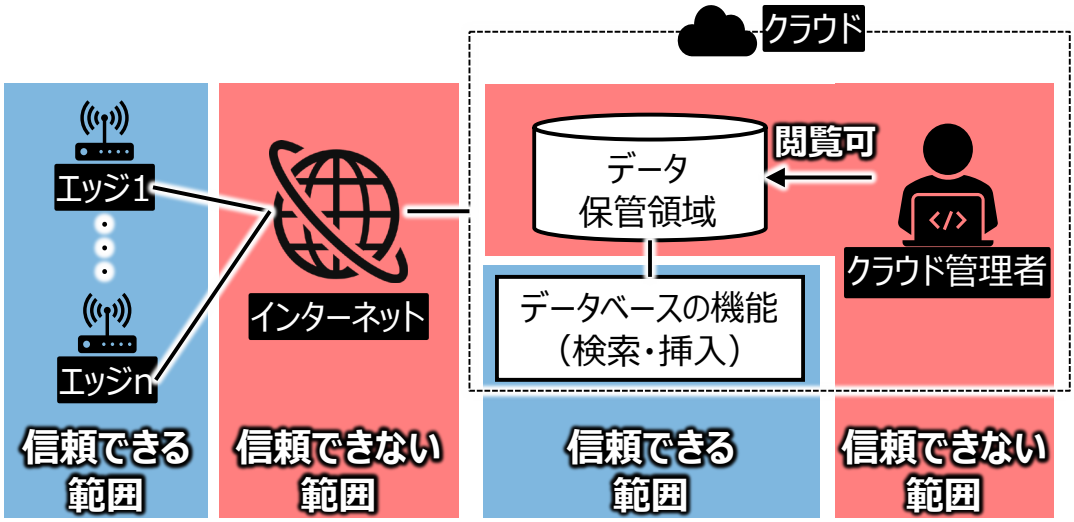


図 30 信頼モデル

3.3.3 性能要件

暗号化されたキーワードどうしの検索の対象となるデータの規模は，文献[156]によれば，1,000 個のデータセットが最低限の推論品質を確保するために必要な数で，10,000 個が優れたデータセット，10 万～100 万個のデータセットが極めて優れたデータセットとなる．したがって，1 つの画像に複数のキーワードが付与されていると想定すると，少なくとも数十万キーワード規模を想定しておく必要がある．さらに，データ転移では，エッジが訓練データの URI などをアップロードする処理は，非同期かつリアルタイム性は要求しない．一方，訓練データの候補を絞り込むための検索は，クラウドに保存されている膨大なデータカタログに対して，ある程度対話的に検索を実行する必要があると予想され，利用者にとってストレスとならない数秒オーダのリアルタイム性が必要となる．

これらのことから，暗号化されたキーワードどうしの検索で要求される性能要件は，以下

の 2 つである．

(1) 数十万キーワード規模での検索時間が，数秒オーダであること
(2) キーワード数に対してスケーラブルであること

一方，訓練データとして有効な画像かどうかを判定する処理は，画像数に対するスケーラビリティよりも，画像の類似度の算出精度が重要視される．さらに，計算資源に限りがあるエッジ上で実行されることと，転移学習は AI の学習フェーズで実行されるので，高いリアルタイム性は要求されないことを考慮する．

これらのことから，訓練データとして有効な画像かどうかを判定する処理で要求される性能要件は，以下の 2 つである．

(1) 類似する特徴を持つ画像に対し，類似度を高く算出できること
（似ている順に画像を順位付けできること）
(2) 限られた計算資源で，実行可能であること

ただし，上記(2)の限られた計算資源での実行に関しては，後述する特徴点マッチングという事前学習が不要で比較的軽量な手法をベースとして採用する関係上，性能評価によって改めて検索速度などを計測することはしない．提案する類似度指標についても，機械学習などの処理負荷が高いものは選択せず，事前学習が不要な軽量なものをベースとする．したがって，本研究では，(1)のみを評価する．ただし，本論文での類似する特徴とは，画像の特徴量を「波形」と見たときに，波形どうしが平行移動したり，伸縮したりしているものと定義する．波形にこのような変化が生じる原因は，環境光などの変化に起因する照明変動や，画像に対する被写体の割合の変化などであるため，ここで定義する類似する特徴を持つ画像は，データ転移に有効と判断できる．

第4章 検索速度とセキュリティを考慮した検索可能暗号の高速化パラメータの決定手法

2.1.3 項において，高安全な検索可能暗号は，平文データに対する検索より低速となることを述べた．高速化の手法として，キーワードのハッシュ値の一部（開示ビット）をクラウドに開示し，クラウドが開示ビットに基づいて検索空間を絞り込むハイブリッド方式があった．この方式は，開示ビットの長さ（開示ビット長）が長いほど大きい高速化効果が得られるが，一方で平文のキーワードの頻度分布と開示ビットの頻度分布が近づいていきキーワードが特定される危険性がある．この章では，平文のキーワードの頻度分布を隠しつつ，十分な高速化効果が得られる開示ビット長を，最小エントロピーとk-匿名性を用いて求める方法を提案する．さらに，データベースの分散化により，スケーラビリティを強化する．既存の検索可能暗号を用いた全文検索システムに提案方式を適用して評価した結果，数十万キーワード規模の場合で検索時間を最大 97.2%削減でき，提案手法によってデータベースが最低でも 31-匿名性を持つような設計に対して，実際は2,598-匿名性を持ち，高速化と安全性を両立することを示す．

4.1 前提条件

4.1.1 評価に使用するキーワードセット

この章の評価で使用するキーワードセットは，著作権フリーな書籍データから抽出したキーワードを類義語で拡張したもの代替として使用し，データ規模のみを合わせる．本来は，キーワードに，AI で用いる教師データ（正解ラベル）のほか，取得日時，場所等の周辺環境の情報や，画像サイズや解像度などのメタデータ，人物であれば性別や人種，年齢などのパーソナルデータを使用すべきだが，これらの機微なデータを研究用途で使用することが困難である．しかしながら，クラウドでの暗号化されたキーワードどうしの検索では，検索性能のボトルネックとなるデータ規模に対する検索速度の評価が重要である．よって，著作権フリーな書籍データから抽出したキーワードを類義語で拡張したもの代替として使用しても，キーワード数に対するスケーラビリティの評価には何ら影響はない．

4.1.2 ハイブリッド方式に対する攻撃者の想定

3.3.2 項で述べた信頼モデルから，ハイブリッド方式に対して攻撃者が可能な行動と，不可能な行動は，以下の 3 つである（図 31）．

① 攻撃者は，正規のクライアントの秘密鍵を入手できないため，開示ビット，タグ，トラップドアを計算できない．したがって，既存の正規のクライアントになりすますことが

できない．

② **攻撃者は，データベース内の開示ビットにのみアクセスできる悪意のあるクラウド管理者である．**

開示ビット以外のタグやトラップドア，訓練データの URI はすべて暗号化されており，攻撃者が閲覧しても一切情報を得ることができない．よって，攻撃者は，開示ビットからキーワードの推定を試みる．

③ **攻撃者は，マッチング処理の結果を監視することができない．**

攻撃者は，マッチング処理の結果を継続的に監視することで，キーワードを推定することができる．しかし，検索頻度が高い場合でも，成功までに長い時間を要する．よって，攻撃が成功する確率は無視できるほど小さい．マッチング処理の結果を継続的に監視するキーワード推定攻撃は，開示ビットを用いない検索可能暗号においても可能で，検索可能暗号そのものの脆弱性である．本研究では，この攻撃を想定しない．

これらの想定から，攻撃者は，データベース内に保管されている開示ビットから，平文のキーワードを推定するものとする．

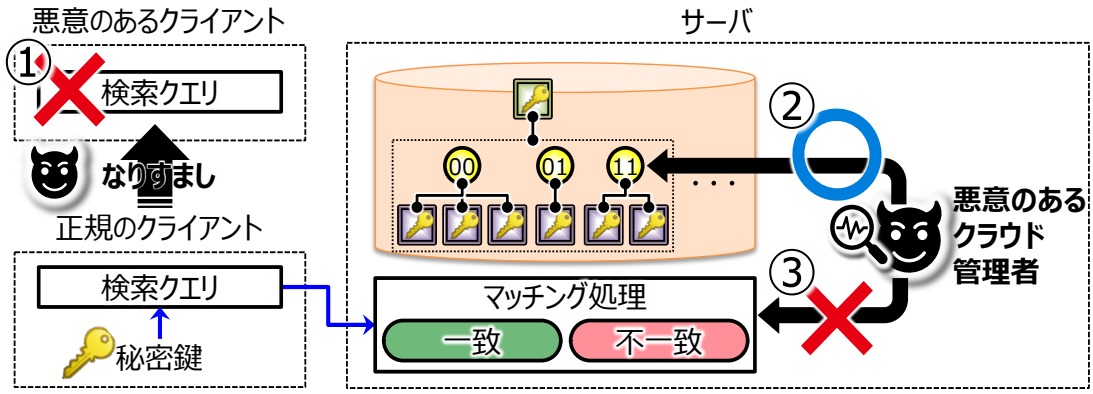


図 31 ハイブリッド方式に対して攻撃者が可能な行動

4.1.3 性能要件

3.3.3 項で示した暗号化されたキーワードどうしの検索で要求される性能要件は，以下の 2 つである．

(1) 数十万キーワード規模での検索時間が，数秒オーダであること

(2) キーワード数に対してスケーラブルであること

なお，検索の精度については対象外とする．この理由は，暗号化されたキーワードどうしの検索の精度は，2.1.2 項で述べたように，既存の類似検索技術の一つである類義語展開の

精度に依存するためである．本研究では，類義語展開に必要なシソーラスが十分な情報量を持ち，高い精度が保たれているとみなす．

4.2 貢献

ハイブリッド方式において，キーワードの頻度分布を隠しつつ，数十万キーワード規模での検索時間が数秒オーダとなる高速化効果が得られる開示ビット長を，最小エントロピーとk-匿名性[157]を用いて求める方法を提案する．具体的には，k-匿名性のkができるだけ大きくなるように，次の手順で開示ビット長を決定する．まず，kの最小値である2-匿名性を満たす十分条件から，最小エントロピーの下限を求める．次に，所望の高速化効果が得られる最小の開示ビット長を特定し，その時の最小エントロピーを算出する．最後に，この最小エントロピーが前述の下限以上であるならば，この時の開示ビット長を採用する．

検索可能暗号を用いた全文検索システムとして，Li らの方式[107]と尾形らの方式[109]と同等なシステムを実装し，それらに提案方式を適用して数十万規模のキーワード数で性能評価して，検索時間を最大 97.2%削減できることを示す．さらに，提案方式によって開示ビットの値が同じキーワードが最低でも 31-匿名性を持つような設計に対し，実際には2,598-匿名性を持ち，キーワードの頻度分布の推定が困難であることも示す．なお，評価においては，スケーラビリティを確保するために，DB を分散化する．

本提案方式は，キーワード分布に依らず，開示ビット長を決定できる．ただし，出現頻度が高いキーワードの開示ビットが，検索キーワードの開示ビットに含まれる場合，検索空間の削減率が低くなるおそれがある．

4.3 提案方式

開示ビットへの頻度分析攻撃の対策を行いつつ，ユースケースで要求される高速化効果が得られる開示ビット長を，最小エントロピーとk-匿名性を用いて求める方法を提案する．

この章では，以下の手順で提案方式の詳細を説明する．表 13 は，この節で使用するシンボルの定義である．

(1) 頻度分析攻撃への対策アプローチ
(2) k-匿名性が成り立つための十分条件
(3) 最小エントロピーの算出
(4) 開示ビット長の決定方法
(5) 分散データベースの適用

表 13 シンボル定義

N	DB 内のタグの総数
c	開示ビット長
k	k-匿名性のパラメータk
$w \in W$	キーワード集合W（または，キーワードの確率変数)とその元w
$P(W)$	Wの出現確率（出現頻度）関数
B_c	開示ビット長がcである開示ビットの確率変数
$P(W\|B_c = b)$	$b \in B_c$が与えられたときのキーワードの確率変数Wの条件付き出現確率関数
$H_\infty(W\|B_c)$	$b \in B_c$が与えられたときのキーワードの確率変数Wの条件付き最小エントロピー

4.3.1 頻度分析攻撃への対策アプローチ

提案方式は，開示ビット長をできるだけ短く設定することで，開示ビットの頻度分布を隠すアプローチを採る．

開示ビットの頻度分布を隠蔽するためには，以下の 2 つの方法が考えられる．

- **開示ビット長を短く設定する**
- **開示ビットの算出方法を変更する**

開示ビット長を短くする方法は，開示ビットの空間を小さくすることで，開示ビットが重なる確率を高くし，複数のキーワードに対して同一の開示ビットが割り当てられる状態を作り出す（図 32)．この動作により，開示ビットの頻度分布は，平文のキーワードの頻度分布と異なるようになり，頻度分析攻撃への耐性が高くなる．

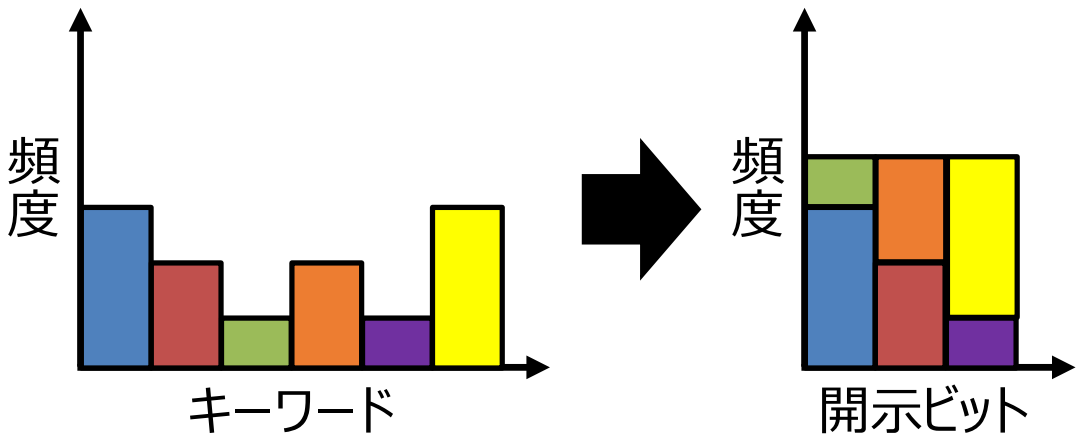


図 32 開示ビット長を短くする方法での開示ビット分布例

一方，開示ビットの算出方法を変更する方法は，同一のキーワードから複数の開示ビットを生成するようにして，開示ビットの頻度分布と平文のキーワードの頻度分布を対応させないようにする．たとえば，開示ビットの頻度分布が一様分布に近づくように，平文のキーワードの出現頻度の高いキーワードには，より多くの開示ビットが割り当てられるように調整し，逆に出現頻度の低いキーワードは，まとめて同じ開示ビットが出力されるよう調整する（図 33）．開示ビットの値を制御する最も単純な方法は，元のキーワードにランダムな文字列を付与し，このランダムな文字列を付与したキーワードから開示ビットを試算して，うまく開示ビットを分割または併合できたかを確かめ，うまくいかなければ再度付与する文字列を更新して，所望の結果が得られるまで繰り返すものである．このランダムな文字列をソルトと呼ぶことがある．

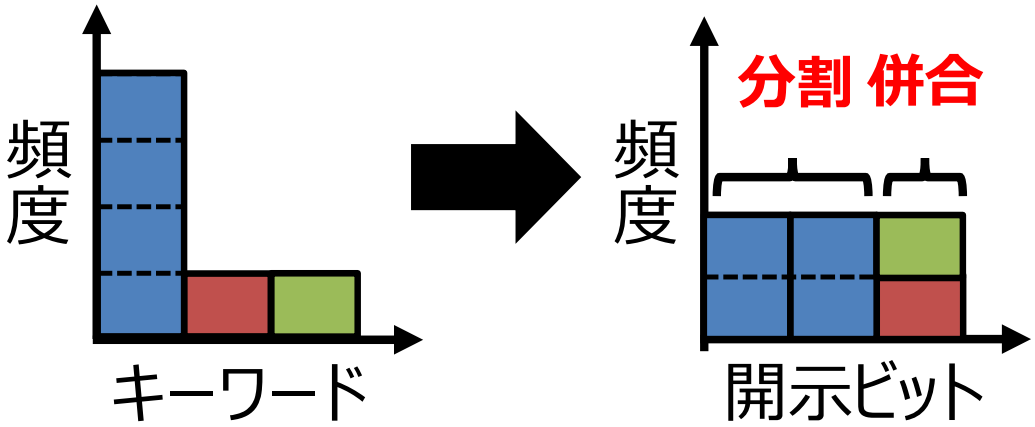


図 33 開示ビットの算出方法を変更する方法での開示ビット分布例

開示ビットの算出方法を変更する方法には，2 つの問題点がある．1 つ目は，通常，事前に完全なキーワードの頻度分布が得られないため，開示ビットを分割または併合するか否かを，クラウドのデータベースに送信済みの開示ビットの頻度分布を使用して判断しなければならないことである．2 つ目は，検索キーワードの開示ビットを計算する際に，過去にそのキーワードの開示ビットの計算に使用されたすべてのソルトを知る必要があることである．

1 つ目の問題に対しては，発行済みの開示ビットの頻度分布を記録しておくデータベースを用意する．次に，2 つ目の問題に対しては，キーワードと，そのキーワードに対して使用されたソルトのリストを記録するデータベースを用意する．そして，これらのデータベースを秘密に保持する．秘密に保持する理由は，攻撃者がこれらのデータベースの情報を手に入れると，この情報から平文のキーワードの頻度分布が推定される恐れがあるためである．本研究の前提とする信頼モデルでは，パブリッククラウドのデータ保管領域は，信頼できない

範囲に設定されており，先ほどの秘密に保管すべき 2 つのデータベースを配置することができない．よって，新たにサーバを追加し，そこへ 2 つのデータベースを配置して，信頼できる範囲に設定するという変更が必要になる（図 34）．

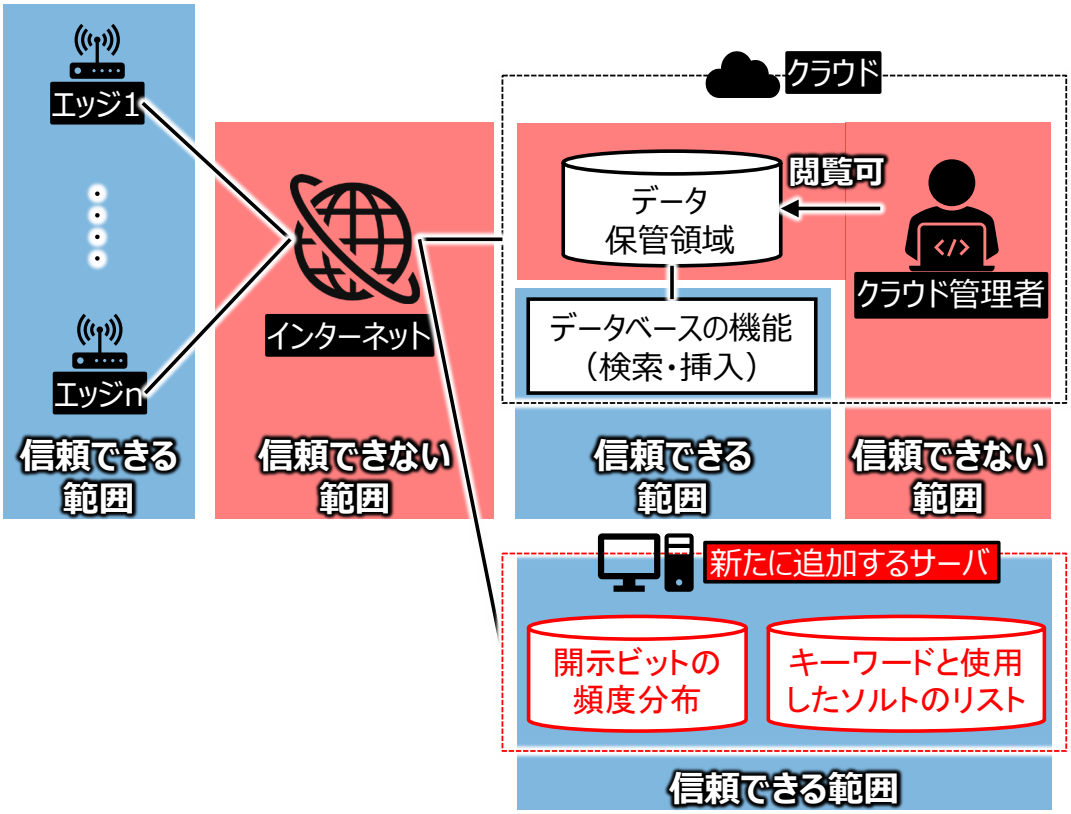


図 34 開示ビットの算出方法を変更する方法を使用する際に必要な信頼モデル

このような変更は，利用者にとって負担が大きいものとなる．したがって本研究では，信頼モデルの変更が不要な，変更開示ビット長を短く設定する方法を採用する．

開示ビットの頻度分布に対する攻撃法として，伊藤らは，キーワードの出現数または生起確率（以降，これらを頻度と表記）とタグの頻度をそれぞれ降順に並べ，頻度の高い順にキーワードとタグを対応付けることで，最尤推定によるキーワードの特定が可能と述べている[104]．確率的暗号ベースの方式では，サーバ側で検索結果を継続的に監視し，タグの頻度の統計を取る必要があり，攻撃成功までには長い時間を要する．一方，確定的暗号ベースの方式では，サーバが保管する DB 中のタグの種類とその出現数から頻度情報が得られるため，攻撃が成功しやすい．

開示ビットの導入によって，確定的暗号ベースの方式の場合と同様に，開示ビットの頻度分布に対して最尤推定が行われキーワードが特定される可能性がある．前述の最尤推定に対しては，伊藤らのように複数の DB にタグを確率的に振り分けてタグの頻度を攪拌する

か，タグ（開示ビット）の頻度分布を隠す方法が有効である．

ハイブリッド方式の検索時間が$O(n/2^c)$であることから，開示ビット長cを増やすと，検索空間の削減量が増えるので検索性能は向上する．一方で，検索空間が削減されるということは，ある開示ビット値を持つキーワードの数が減るということであり，開示ビットの頻度分布とキーワードの頻度分布が近づき，最尤推定によるキーワード特定の攻撃を受け易くなる．

したがって，検索性能と最尤推定による攻撃に対する安全性のトレードオフをとるために，1 つの開示ビットに対応付けられるキーワードがk個以上存在するように開示ビット長cの最大値を検索対象のデータから求め，この最大値を上限として，必要な検索性能を得るための開示ビット長を設定する．

このように，「データセット中のどのようなカラムを参照してもk以上のレコードが該当する」状態をk-匿名性を持つという[157]．たとえば，表 14 は4-匿名性を持つハイブリッド方式の転置索引の例である．表中の括弧書きのキーワードは，実際の転置索引に記載されていないが，説明の都合上，表記する．この例では，開示ビットが「10」に対応するキーワードが「AAA」，「BBB」，「CCC」，「DDD」の 4 種類あることを確認できる．もし，転置索引中のすべての開示ビットに対しても同様に，最低でも 4 種類のキーワードを含むとすると，この転置索引は4-匿名性を持つといえる．

表 14 4-匿名性を持つ転置索引の例

DB内の開示ビット	タグ	(キーワード)	
10	d96b4f8a9d	AAA	1
10	a0663e3e41	AAA	
10	17b4539cd4	AAA	
10	77963b7a93	BBB	2
10	6a9cd718aa	BBB	
10	3007b906b9	CCC	3
10	b9468d5ee1	DDD	4
01	2c47f2a939	EEE	
01	cf7c9541c2	EEE	
・・・	・・・	・・・	

本研究では，どのような開示ビットで検索したとしても，絞り込み後の DB テーブルレコードが最低でも2-匿名性を持てば，攻撃者は 2 つの候補までキーワードを絞り込めるが，最終的には，くじ引きのように確率 50%でしか正しいキーワードを特定できない状態となるため，$k = 2$をkの最小値とした．ただし，ユースケースによっては，$k = 2$よりも大きな値を設定する必要があると考えられる．このような場合においても提案方式が簡単に使用できるよう，提案方式のユーザが調整可能な値は，できるだけ少なく，単純にする．

図 35 は，2-匿名性を持つハイブリッド方式の転置索引において，悪意のあるクラウド管理者がキーワードを推定しようとしたときの例である．表中の括弧書きのキーワードは，実際の転置索引に記載されていないが，説明の都合上，表記する．悪意のあるクラウド管理者は，DB 内の開示ビットとタグからキーワードを推測しようとするが，2-匿名性を持つ場合，開示ビットが「10」に対応するキーワードが 2 つ以上存在するため，正しくキーワードの出現頻度を算出できない．したがって，悪意のあるクラウド管理者は，高々50%の確率でしか正しいキーワードを特定することができない．50%という確率が安全であるかはユースケースによるが，本研究では，攻撃者が 100%の確信をもってキーワードを推定できない状態を安全であると定義する．

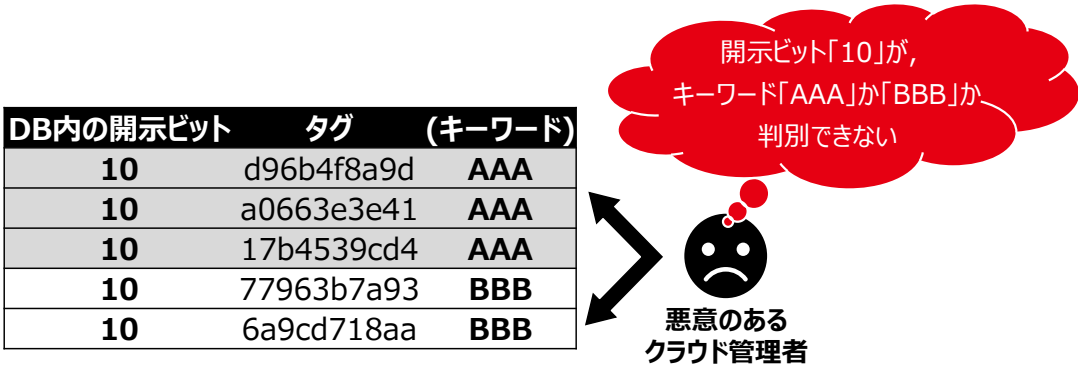


図 35 2-匿名性の安全性

なお，ある 1 つの開示ビットにk個以上のキーワードが対応するときに，その開示ビットのキーワードを推定できる最大の確率が$1/k$以下となるとは限らないことに注意する．たとえば，表 15 のように1/3の確率で出現するキーワード「AAA」と，2/3の確率で出現するキーワード「BBB」が開示ビット「10」に対応づくとき，開示ビット「10」のキーワードを推定できる確率は「AAA」が1/3，「BBB」が 2/3 である．よって，最大値である「BBB」の推定確率は 2/3 で，1/2より大きくなる．

表 15 推定できる確率が$1/k$にならない場合の例

DB内の開示ビット	(キーワード)	
10	AAA	$P(\mathrm{AAA}\|10)=\frac{1}{3}$
10	AAA	
10	BBB	$P(\mathrm{BBB}\|10)=\frac{2}{3}$
10	BBB	
10	BBB	
10	BBB	

しかし，ある開示ビットに対応するキーワードの出現頻度の最大値が$1/k$以下であれば，必ずk-匿名性以上を持つ．これを証明する．

いま，表 16 のようなk-匿名性を持つ転置索引があるとする．

表 16 転置索引の例

DB内の開示ビット	(キーワード)	
b	w_1	n_1
・・・	・・・	
b	w_1	
b	w_2	n_2
・・・	・・・	
b	w_2	
・・・	・・・	
b	w_k	n_k
・・・	・・・	
b	w_k	

開示ビットが$\boldsymbol{b}$に対応する$\boldsymbol{k}$種類のキーワード$\boldsymbol{w_i}(\boldsymbol{i}=\boldsymbol{1},\ldots,\boldsymbol{k})$があり，それぞれ$\boldsymbol{n_i}(\boldsymbol{i}=\boldsymbol{1},\ldots,\boldsymbol{k})$個あるとする．ただし，$\boldsymbol{n_i}\in\mathbb{N}$である．$\boldsymbol{w_i}$の出現頻度は，数式(4)で表すことができる．

$$\frac{\boldsymbol{n_i}}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{k}}\boldsymbol{n_j}} \tag{4}$$

ここで，$\boldsymbol{n_i}$がすべて等しい（$\boldsymbol{n_1}=\boldsymbol{n_2}=\cdots=\boldsymbol{n_k}$）場合，開示ビットが$\boldsymbol{b}$に対応する$\boldsymbol{w}$の出現頻度を$P(\boldsymbol{w})$と表すと，$P(\boldsymbol{w})$の最大値は数式(5)となる．よって，$P(\boldsymbol{w})$の最大値が$1/k$以下を満たすことが分かる．

$$\max_{i=1,\ldots,k} P(\boldsymbol{w_i})=\frac{1}{k} \tag{5}$$

次に，$\boldsymbol{n_m}$が最も大きい（$\boldsymbol{n_m}>\boldsymbol{n_i}(\boldsymbol{i}\neq\boldsymbol{m})$）場合，$P(\boldsymbol{w})$の最大値は数式(6)となる．

$$\max_{i=1,\ldots,k} P(\boldsymbol{w_i})=P(\boldsymbol{w_m})=\frac{\boldsymbol{n_m}}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{k}}\boldsymbol{n_j}} \tag{6}$$

出現回数の最小値である$\boldsymbol{n_i}$=1 ($\boldsymbol{i}\neq\boldsymbol{m}$)，$\boldsymbol{n_m}>\boldsymbol{1}$と置くと，数式(6)は数式(7)のように変形できる．

$$\frac{\boldsymbol{n_m}}{\boldsymbol{n_m}+\sum_{\boldsymbol{j=1,j\neq m}}^{\boldsymbol{k}}\boldsymbol{n_j}}=\frac{\boldsymbol{n_m}}{\boldsymbol{n_m}+(\boldsymbol{k}-\boldsymbol{1})} \tag{7}$$

さらに，数式(5)が$1/k$以下であるためには，数式(8)の条件を満たさなければならない．

$$\begin{gathered}\frac{\boldsymbol{n_m}}{\boldsymbol{n_m}+(\boldsymbol{k}-\boldsymbol{1})}\leq\frac{\boldsymbol{1}}{\boldsymbol{k}}\\ \boldsymbol{1}+\frac{\boldsymbol{k}}{\boldsymbol{n_m}}-\frac{\boldsymbol{1}}{\boldsymbol{n_m}}\geq\boldsymbol{k}\\ \boldsymbol{1}-\frac{\boldsymbol{1}}{\boldsymbol{n_m}}\geq\boldsymbol{k}-\frac{\boldsymbol{k}}{\boldsymbol{n_m}}\\ \boldsymbol{1}-\frac{\boldsymbol{1}}{\boldsymbol{n_m}}\geq\boldsymbol{k}\left(\boldsymbol{1}-\frac{\boldsymbol{1}}{\boldsymbol{n_m}}\right)\\ \boldsymbol{1}\geq\boldsymbol{k}\end{gathered} \tag{8}$$

数式(8)より$\boldsymbol{k}=\boldsymbol{1}$のみ成立するが，本研究で定義した$\boldsymbol{k}\geq\boldsymbol{2}$と矛盾する．

$\boldsymbol{k \geq 2}$を満たすためには，キーワードを増やす必要がある．$\boldsymbol{n_{k+1}}$個の$\boldsymbol{w_{k+1}}$を増やす場合，$P(\boldsymbol{w})$の最大値は数式(9)となる．

$$\max_{i=1,\ldots,k+1} P(\boldsymbol{w_i}) = \frac{\boldsymbol{n_m}}{\sum_{j=1}^{k+1} \boldsymbol{n_j}} = \frac{\boldsymbol{n_m}}{\boldsymbol{n_m + n_{k+1} + (k-1)}} \leq \frac{\boldsymbol{1}}{\boldsymbol{k}} \tag{9}$$

数式(9)を変形すると数式(10)のようになる．

$$\begin{aligned} \boldsymbol{1 + \frac{n_{k+1}}{n_m} + \frac{k}{n_m} - \frac{1}{n_m} \geq k} \\ \boldsymbol{1 - \frac{1}{n_m} + \frac{n_{k+1}}{n_m} \geq k\left(1 - \frac{1}{n_m}\right)} \\ \boldsymbol{1 + \frac{n_{k+1}}{n_m - 1} \geq k} \end{aligned} \tag{10}$$

$\boldsymbol{k}$の最小値である$\boldsymbol{2}$を想定すると，数式(11)を満たせば$\boldsymbol{2}$-匿名性をもつ．

$$\begin{aligned} \boldsymbol{\frac{n_{k+1}}{n_m - 1} = 1} \\ \boldsymbol{n_{k+1} = n_m - 1} \end{aligned} \tag{11}$$

$\boldsymbol{n_m > 1}$より，$\boldsymbol{n_m}$の最小値は 2 であるので，$\boldsymbol{n_{k+1} \geq 1}$となることに注意する．よって，$\boldsymbol{w_{k+1}}$を$\boldsymbol{n_m - 1}$個以上増やすことで，2-匿名性の条件を満足する．ここで，キーワードの種類を 1 つ増やすということは，転置索引が$(k+1)$-匿名性を持つことと等しい．したがって，$\boldsymbol{n_m}$が最も大きい場合，開示ビットが$\boldsymbol{b}$に対応するキーワードの出現頻度の最大値が$1/k$以下を満たすためには，転置索引が$(k+1)$-匿名性以上を持つことが条件となる．

以上から，ある開示ビットに対応するキーワードの出現頻度の最大値が$1/k$以下であれば，必ずk-匿名性以上を持つことが示された．実際にこの条件が成立するか，具体的な例を用いて確認する．

表 17 は，転置索引が2-匿名性以上を持つとき，キーワードの出現頻度の最大値が1/2以下になる例を表したものである．表 17(A)は，キーワードの出現頻度が等しいとき，キーワードの出現頻度の最大値が1/2以下になる場合の一例である．この場合は，いずれのキーワードの出現確率も1/2以下となる．一方，表 17 (B)は，キーワードの出現頻度に偏りがあるとき，キーワードの出辺頻度の最大値が1/2以下になる場合の一例である．この場合は，新たなキーワード「CCC」を，最も出現回数が多いキーワード「AAA」の個数より 1 つ少な

い 3 個追加する．これにより，キーワード「AAA」の数に対し，「AAA」以外の数が増え，1/2以下を満たすようになる．以上より，証明が正しいことを確認できた．

提案方式は，ある開示ビットに対応するキーワードの出現頻度の最大値が$1/k$以下であるか否かを調べるだけで，k-匿名性以上を持つことを容易に確認できる．この事実を利用して，適切な開示ビット長を決定する．

表 17 転置索引が2-匿名性以上で，キーワードの出現頻度の最大値が1/2以下になる例

A. キーワードの出現頻度が等しい場合

DB内の開示ビット	(キーワード)
10	AAA
10	BBB

$P(\mathrm{AAA}|10) = \frac{1}{2} \leq \frac{1}{2}$

B. キーワードの出現頻度に偏りがある場合

DB内の開示ビット	(キーワード)
10	AAA
10	AAA
10	AAA
10	AAA
10	BBB
10	BBB
10	CCC
10	CCC
10	CCC

$P(\mathrm{AAA}|10) = \frac{4}{9} \leq \frac{1}{2}$

AAAの数-1個追加

4.3.2 k-匿名性が成り立つための十分条件

k-匿名性を満たすための十分条件を，暗号理論においてよく利用される尺度である最小エントロピーを用いて表現する[158]．確率変数Xの最小エントロピーは，数式(12)のように定義され，Xの取り得る値の中で最も生起しやすい値の曖昧さを表すものである．

$$H_\infty(X) := -\log_2 \max_{x \in X} P(X) \tag{12}$$

最小エントロピーが 0 に近いほど曖昧さがないため，本研究の場合ではキーワードが特定されやすくなることを意味する．この定義に従い，k-匿名性が成立するための十分条件を求める．Wをキーワードの確率変数，B_cを開示ビット長がcの開示ビットの確率変数，$P(W|B_c = b)$をある開示ビット$b \in B_c$が与えられた時のキーワードWの条件付き生起確率とすると，k-匿名性が成立するための十分条件は，条件付き最小エントロピー$H_\infty(W|B_c)$が数式(13)の不等式を満たすことである．

$$H_\infty(W|B_c) := -\mathbf{log_2} \max_{w \in W, b \in B_c} P(W|B_c = b) \geq \mathbf{log_2}\, k \tag{13}$$

ただし，$c = 0$の場合は開示ビットで絞り込まれないため，数式(14)のように読み替える．$P(W)$は，キーワードの生起確率である．

$$H_\infty(W) := -\mathbf{log_2} \max_{w \in W} P(W) \tag{14}$$

式(13)の意味について説明する．$H_\infty(W|B_c)$が$\mathbf{log_2}\, k$以上とは，数式(13)の真数部分が数式(15)を満たす場合である．

$$\max_{w \in W, b \in B_c} P(W|B_c = b) \leq \frac{1}{k} \tag{15}$$

数式(15)は，どのような開示ビット値bに対しても，キーワードが少なくともk個存在しないと成立しない．しかしながら，4.3.1 項で示したように，$P(W|B_c = b)$が$1/k$以上となる場合でもk-匿名性を持つことがある．図 36 は，2-匿名性を満たすが，2-匿名性が成立するための十分条件である数式(13)を満たさない例である．図 36 (A)を見ると，DB 内の開示ビットが「10010」であるキーワードが「AAA」,「BBB」の 2 つあるにもかかわらず，$H_\infty(W|B_c)$の値は約 0.74 となり$\mathbf{log_2}\, 2\, (= \mathbf{1})$以上にならない．一方，図 36 (B)のように，DB 内の開示ビットが「10」であるキーワードが「AAA」,「BBB」,「CCC」,「DDD」の 4 つある場合は，$H_\infty(W|B_c)$の値が約 1.22 となり，$\mathbf{log_2}\, 2\, (= \mathbf{1})$以上で数式(13)を満たす．

A. 2-匿名性が成立するための十分条件を**満たさない**状態

DB内の開示ビット	タグ	(キーワード)
10010	d96b4f8a9d	**AAA**
10010	a0663e3e41	**AAA**
10010	17b4539cd4	**AAA**
10010	77963b7a93	**BBB**
10010	6a9cd718aa	**BBB**
10111	3007b906b9	**CCC**
10011	b9468d5ee1	**DDD**

$H_\infty(\mathrm{AAA}|10010) = -\log_2 \frac{3}{5}$
$\fallingdotseq \mathbf{0.74} \not\geq 1$

$H_\infty(\mathrm{CCC}|10111) = 0 \not\geq \mathbf{1}$

$H_\infty(\mathrm{DDD}|10011) = 0 \not\geq \mathbf{1}$

B. 2-匿名性が成立するための十分条件を**満たす**状態

DB内の開示ビット	タグ	(キーワード)
10	d96b4f8a9d	**AAA**
10	a0663e3e41	**AAA**
10	17b4539cd4	**AAA**
10	77963b7a93	**BBB**
10	6a9cd718aa	**BBB**
10	3007b906b9	**CCC**
10	b9468d5ee1	**DDD**

$H_\infty(\mathrm{AAA}|10\ ) = -\log_2 \frac{3}{7}$
$\fallingdotseq \mathbf{1.22} \geq \mathbf{1}$

図 36 2-匿名性を満たすが，2-匿名性が成立するための十分条件を満たさない例

このように，k-匿名性が成立するための十分条件は，k-匿名性を厳密に判断しているわけではないが，十分条件を用いるとk-匿名性以上を持つかどうかの判定を簡潔に判定できるメリットがある．たとえ図 36 に示した状況が発生し，クラウド管理者が開示ビット長を短くしたとしても，セキュリティレベルが下がるわけではないため問題ない．したがって，$P(W|B_c = b)$が$1/k$以上となる場合でもk-匿名性を持つことがある状況は，k-匿名性の判定を簡潔することを重視し，考慮しなくて良い．以上のことから，検索対象のデータが数式(15)を満たすことを確認できれば，セキュリティを確保することができる．

以降の節では，数式(15)に着目して$P(W|B_c)$の求め方について述べる．

4.3.3 最小エントロピーの算出

4.3.2 項で述べた条件付き最小エントロピー$H_\infty(W|B_c)$を求めるには$P(W|B_c)$が必要になるが，一般的には，キーワードは徐々に追加されていくため，初期状態で完全なキーワードの頻度分布を事前に知るのは困難である．そこで，少量のサンプルキーワードから$P(W|B_c)$の分布形状を推定する．以降，この推定した分布を「サンプル分布」，全文検索システムで

実際に使用するキーワードから求めた$P(W|B_c)$を「実際の分布」と呼ぶ．サンプル分布は，キーワードを特定することなく，キーワードの頻度分布と，キーワードに割り当てられる開示ビットの頻度分布を推測し，その推測に基づいて分布を求める．このことから，サンプル分布と実際の分布のそれぞれの開示ビットの頻度分布にずれが生じ，最小エントロピーの値に誤差が出るおそれがあること注意する．

本研究では，実際のキーワードが決まっていない初期状態でも，実際の分布の近似値であるサンプル分布を使用して最小エントロピーを求めることができるメリットを重視する．サンプルキーワード集合の頻度分布が，母分布となる実際のキーワード集合の頻度分布をよく表していることを前提とする．以降で，サンプル分布の具体的な計算手順を述べる．

サンプル分布の計算手順

(1) 実際のキーワードの頻度分布をよく表す代表的な文書 h)のサンプルからキーワードを抽出し，それらキーワードを数値化する．

(2) 手順(1)のキーワードの数値を用いて，カーネル密度推定を行い，数値化したキーワードの確率密度関数を推定する．

(3) 手順(2)で推定した連続型確率変数の確率密度関数を，離散型確率変数の確率質量関数に変換する．この離散型確率変数は，実際のキーワード集合と同程度な大きさを持つ，キーワード文字列は特定しない仮想的なサンプルキーワード集合の元に対応付けられる．すなわち，サンプルキーワード集合の元である各キーワードに，確率変数の異なる数値が対応付けられる．

(4) 手順(3)で求めた数値で定義されたサンプルキーワード集合のすべての元に対して，元（数値）を文字列とみなして開示ビットを計算し，サンプルキーワードの開示ビットの頻度分布を求める．この開示ビットの頻度分布とサンプルキーワードの確率質量関数を用いて，サンプル分布を求める．

手順(1)では，手順(2)の事前準備のために，キーワードを数値化する．具体的には，全文検索システムへの入力が想定される，実際のキーワードの頻度分布をよく表している代表的な文書サンプルをいくつか選定し，その文書サンプルから全文検索システムと同じ方法でキーワードを抽出する．こうして抽出したキーワード集合は，キーワードの重複を許容した多重集合で定義される．この多重集合の大きさをmとすると，キーワードの出現頻度の降順に$1,2,\dots,n$（$n \leq m$）と数値を割り当てることによって，要素数がmの$\{1,1,\dots,1,2,2,\dots,n\}$のような多重集合となる．この数値に置換した多重集合を$\{w_i\}_{i=1}^{m}$表記する．

手順(2)では，実際のキーワードの頻度分布の形状が未知であるため，ノンパラメトリック手法であるカーネル密度推定を用いる．$\{w_i\}_{i=1}^{m}$を用いて，カーネル密度推定量$\hat{p}_h(w)$を数式(16)のように求める．

---

h) 実際には訓練データを説明した自然文であるが，説明の都合上，文書と表現する．

$$\hat{p}_h(w) = \frac{1}{mh}\sum_{i=1}^{m} K\left(\frac{w - w_i}{h}\right) \tag{16}$$

hはバンド幅であり，$\{w_i\}_{i=1}^{m}$が実際のキーワードの分布をよく表すと仮定したため，$h = 0.001$と設定した．カーネル関数Kは，平均が 0 で分散が 1 のガウス関数（$K(w) = \exp(-w^2/2)/\sqrt{2\pi}$）を用いた．

手順(3)では，連続型確率分布である$\hat{p}_h(w)$を離散型確率分布に変換する．変換の際，実際の分布と分布形状が似るように，離散型確率変数の数を，実際のキーワードの頻度分布のキーワード数に近づける方がよい．本方式では，$\hat{p}_h(w)$のwの定義域の中からM個（$n < M$）の値を等間隔で選ぶことで，キーワードを拡張する．このように拡張した値を$\widehat{w}_1, \dots, \widehat{w}_M$と表記する．ただし，$\widehat{w}_1 = 1$，$\widehat{w}_M = n$とする．たとえば，$n = 10, M = 100$とすると，$\widehat{w}_1 = 1, \widehat{w}_2 = \widehat{w}_1 + 1/11, \widehat{w}_3 = \widehat{w}_2 + 1/11, \dots, \widehat{w}_{100} = \widehat{w}_{99} + 1/11 = 10$となる．このように拡張した$\widehat{w}_i$の頻度$\hat{P}(\widehat{W} = \widehat{w}_i)$を数式(17)のように定義する．

$$\hat{P}(\widehat{W} = \widehat{w}_i) := \boldsymbol{C}\int_{\widehat{w}_{i-\frac{1}{2}}}^{\widehat{w}_{i+\frac{1}{2}}} \hat{p}_h(w)dw \tag{17}$$

$\boldsymbol{C}$は$\hat{P}(\widehat{W})$の総和を 1 にするための規格化定数であり，数式(18)となる．

$$\boldsymbol{C} = \frac{1}{\int_{\widehat{w}_{1-\frac{1}{2}}}^{\widehat{w}_{M+\frac{1}{2}}} \hat{p}_h(w)dw} \tag{18}$$

図 37 は，$\hat{P}(\widehat{W} = \widehat{w}_i)$の算出を図示したものである．$\hat{P}(\widehat{W} = \widehat{w}_i)$は，$\widehat{w}_i$を中心に$\widehat{w}_{i-1}$，$\widehat{w}_{i+1}$との中間地点の範囲で積分する．なお，$\widehat{w}_{1/2}$，$\widehat{w}_{M+1/2}$は，カーネル密度推定の外挿による予測値を利用する（図 37(A)の部分）．

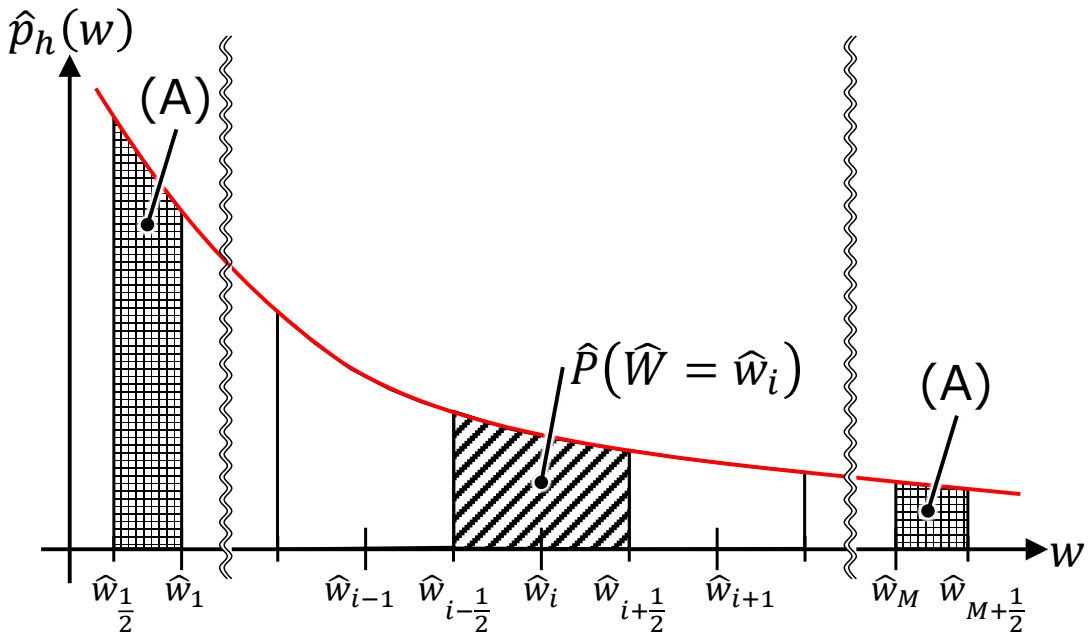


図 37 連続型確率分布から離散型確率分布への変換

手順(4)では，手順(3)で求めた$\hat{P}(\hat{W})$から，$P(W|B_c)$の近似値$\hat{P}(\hat{W}|B_c)$を計算する．$\hat{w}_i$から開示ビットを計算する関数を$f(\hat{W})$とするとき，開示ビットbが与えられた時の$\hat{w}_i$の条件付き確率質量関数$\hat{P}(\hat{W}=\hat{w}_i|B_c=b)$は，数式(19)となる．

$$\hat{P}(\hat{W}=\hat{w}_i|B_c=b)=\frac{\hat{P}(\hat{w}_i)}{\sum_{\{\hat{w}_j|f(\hat{w}_j)=f(\hat{w}_i)\}}\hat{P}(\hat{w}_j)} \tag{19}$$

数式(19)を基に数式(15)が具体的に計算できるため，最小エントロピー$H_\infty(\hat{W}|B_c)$も同様に求めることができる．この$H_\infty(\hat{W}|B_c)$を開示ビット長の決定に用いる．

4.3.4 開示ビット長の決定方法

開示ビット長は，以下の手順で決定する．

開示ビット長の決定手順

(1) 開示ビット長を変化させながら，4.3.3 項で述べた方法で条件付き最小エントロピー$H_\infty(\hat{W}|B_c)$を計算する．

(2) 手順(1)で算出した最小エントロピーが数式(13)を満たす最大の開示ビット長を求める．

(3) 手順(2)で求めた最大の開示ビット長以下で，ユースケースが要求する性能を満たす検索空間削減率（$=1-1/2^c$, cは開示ビット長）を達成する最小の開示ビット長を決定する．

上記手順の詳細を，本評価環境下に適用した場合を例に説明する．

まず，手順(1)に従い最小エントロピーを求める．表 18 は，本評価環境で想定するキーワード規模の一覧である．Li らの方式で約 65.5 万キーワード規模，尾形らの方式で約 25.5 万キーワード規模となる．この例では，Li らの方式での最小エントロピーの近似値をサンプル分布から求めると仮定する．まず，4.3.3 項で示したようにサンプルキーワード集合の頻度分布が，母分布となる実際のキーワード集合の頻度分布をよく表していることが前提である．4.3.3(1)では，1,000 個の文書からこのような条件を満たす文書を数個選ぶようにしているが，この例では理想的な文書が選択され，サンプルキーワード集合が得られたとみなして，実際のキーワードの頻度分布を母分布として標本を選ぶ．このように変更しても，前提条件を満たすように文書を厳密に選択するという条件下では，最終的に得られるサンプルキーワードはほぼ同じとなる．よって，この例では，次のようにサンプルキーワードを選ぶ．表 18 で示したキーワードの頻度分布を母分布として，この母分布に従いキーワードの重複を許して標本を 1,000 個抽出する．続いて，4.3.3 項で示した方法で，$m = 1{,}000$，$M = 42{,}000$とおき，数値に置換したキーワードを，Li らの方式で抽出したキーワード数に近い 42,000 個に拡張したときの頻度分布$\hat{P}(\widehat{W})$を計算する．さらに，開示ビット長cを 1～12 まで変化させて$\hat{P}(\widehat{W}|B_c)$を求め，最小エントロピー$H_\infty(\widehat{W}|B_c)$を計算すると，図 38 緑線のようになる．

次に，手順(2)に従い最大の開示ビット長を求めると，$k = 2$の場合，最小エントロピーが$\mathbf{log}_2 2 = 1$（図 38 緑破線）以上で最も長い開示ビット長は 8 であることが分かる．

最後に，手順(3)に従い開示ビット長を決定する．仮に本ユースケースで 10 倍の高速化を得ようとする場合，検索空間を 90%ほど削減でき，かつ，匿名性が上がるようにできるだけkの値が大きくなるようにcを設定すればよい．

図 38 青線が検索空間削減率で，図 38 青破線が 90%削減ラインである．この場合，$c = 4, k = 31$で達成可能である（$\because H_\infty(\widehat{W}|B_4) \fallingdotseq 4.97 \geq \mathbf{log}_2 31 \fallingdotseq 4.95$）．よって，設定すべき開示ビット長は 4 と決定する．

尾形らの方式に提案方式を適用した場合も同様に計算すると，設定すべき開示ビット長は 4 であった．

表 18 登録キーワードの種類とキーワードの出現回数

Li らの方式			尾形らの方式		
	キーワード	出現回数		キーワード	出現回数
1	自分	1863	1	自分	706
2	愛情	1466	2	親	689
3	娘	1311	3	人	678
…	…	…	…	…	…
42,392	宿縁	1	32,721	宿縁	1
	合計	685,323		合計	255,362

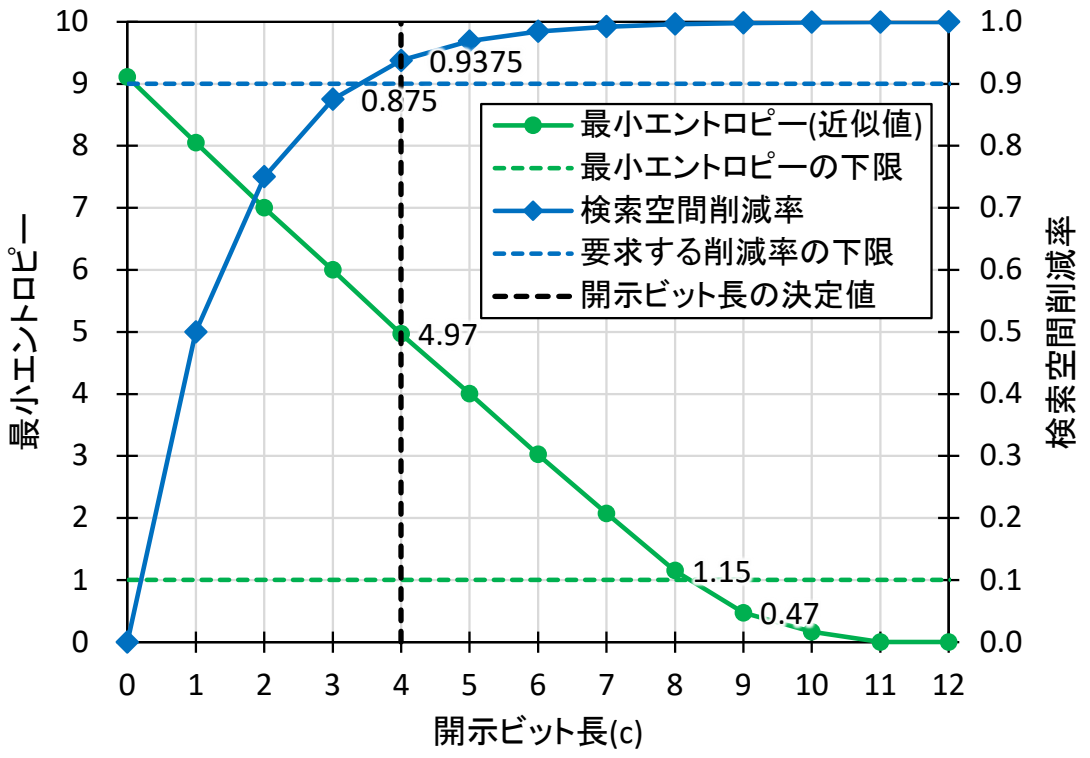


図 38 最小エントロピーと検索空間削減率

4.3.5 分散データベースの適用

提案方式は，データベースを分散化しても適用可能である．各データベースに対して均等にデータが振り分けられている場合，各データベース内のキーワードの頻度分布はほぼ等しくなる．したがって，キーワードのエントロピーも単一のデータベースをした場合と等しく，分散されたすべてのデータベースで同じ開示ビット長を設定することができる．

たとえば，キーワードw_1の出現総数が 9，キーワードw_2が 6, キーワードw_3が 3 で，3 つのデータベースに分散させるとき，各データベースのキーワードw_1，w_2，w_3の出現回数はそれぞれ 3，2，1 となる．よって，w_1，w_2，w_3の出現割合は 3 : 2 : 1 となり，この値はデータベース分散させない場合と一致する．

4.4 評価

1,000 個の文書に対して Li ら，尾形らの方式と提案方式を実装したものを用いて性能評価し，キーワードの頻度分布を隠しつつ，検索時間を短縮可能であることを示す．さらに，データベースの分散化により，数十万キーワード規模の検索へのスケーラビリティが確保されていることも示す．

具体的には，以下の 2 つの項目を計測し，高速化効果と安全性を確認する．

(1) 文書 1,000 冊に対する平均検索時間と検索時間の削減率
(2) 最小エントロピーの計算において，実際の分布とサンプル分布を使用したときの誤差と，
実際の DB テーブルレコードのk-匿名性

4.4.1 評価条件

表 19 に評価条件の一覧を示す．提案方式の適用対象である Li らの方式で用いるファジーキーワード集合$\{w_i'\}$と，尾形らの方式で用いる検索キーワードと意味的に近いキーワード集合$S(w)$を，キーワードの類義語の集合に変更する．

このように変更しても，タグあるいはトラップドアにあいまい性を持たせ，全文検索を実現するという本質的な動作は変わらない．

提案方式を適用しない Li らの方式と尾形らの方式は，単一の DB を使用し，開示ビットを使用しない．

評価に使用する文書は，訓練データを説明する自然文が理想だが，そのような自然文を準備できなかったため，青空文庫[159]の書籍を代替とした．

評価項目は検索時間で，10 回の試行における標本平均と，95%信頼区間を算出する．

表 19 評価条件一覧

提案方式の適用対象	Li らの方式，尾形らの方式
文書	青空文庫の宮本百合子の書籍 1,000 冊
形態素解析エンジン	Janome[160] （固有名詞，一般名詞のみ抽出，MeCab IPA 辞書[161]を使用）
類義語展開エンジン	word2vec[162]（日本語学習済みモデル[163]を使用）
類義語展開数	1 語につき 3 つの類義語に展開
検索文	愛（ヒット率：210 冊/1,000 冊）
開示ビット長	4
分散 DB	PostgreSQL 9.6 + Citus Extension
分散数	8
試行回数	10（標本平均，95%信頼区間を算出）

4.4.2 評価環境

図 39 にプログラムスタックを示す．サーバは，Citus Docker（v7.0.3）[164][165]をベースイメージとして用い，平野らの方式を C 言語で実装した．クライアントは，Python ベースであるが，C 言語で実装した平野らの方式（図中の暗号コア）を呼び出している．評価環境のサーバ側プログラムは，Oracle VM 上に Ubuntu をゲスト OS としてインストールし，さらに Docker エンジンを実行して提案手法が動作する全文検索の Docker コンテナを動作させている．Oracle VM ホスト環境を表 20 に，Oracle VM ゲスト環境を表 21 に示す．

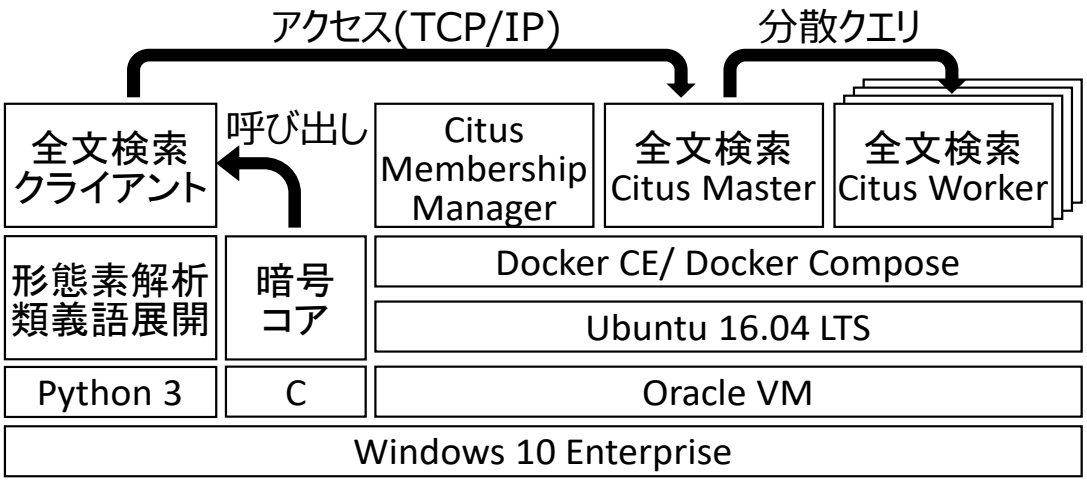


図 39 プログラムスタック

表 20 Oracle VM ホスト環境

CPU	Intel Core i5-8500 CPU@3GHz (6 コア 6 スレッド)
メモリ	64GB
ディスク	Samsung SSD 860 EVO 1TB × 2 (Read 550 MB/s , Write 520 MB/s)

表 21 Oracle VM ゲスト環境

CPU	Intel Core i5-8500 CPU@3GHz (4 コア 4 スレッド)
メモリ	16GB
ディスク	100GB（シンプロビジョニング）

4.4.3 評価結果（高速化効果）

Li らの方式および尾形らの方式における，提案方式の適用有無による平均検索時間と信頼区間の評価結果を表 22 に示す．表 22 の信頼区間は，平均検索時間を起点とした範囲である．

提案方式を適用しない Li らの方式は，平均検索時間が 155.80 秒であり，本研究で想定する社内情報検索のユースケースの応答性能を満たしているとは言えない．

一方，提案方式を適用した場合は，4.39 秒に短縮でき，提案方式を適用しない場合と比較し 97.2%削減できることを確認した．この値であれば，ユーザが対話的に社内情報検索を使用しても問題ない範囲である．同様に，提案方式を適用しない尾形らの方式は，平均検索時間が 227.75 秒であったが，提案方式を適用した場合は，19.89 秒に短縮され，提案方式を適用しない場合と比較し 91.3%削減できることを確認した．

表 22 平均検索時間（カッコ内が信頼区間）

	提案方式適用前	提案方式適用後
Li らの方式	155.80 秒 (±2.95 秒)	4.39 秒 (±0.55 秒)
尾形らの方式	227.75 秒 (±1.33 秒)	19.89 秒 (±0.33 秒)

4.4.4 評価結果（安全性）

Li らの方式における，最小エントロピーの誤差と実際の DB テーブルレコードのk-匿名性の実測値を図 40，図 41 に示す．以降，実際のキーワードの頻度分布を使用して得られた最小エントロピーを「最小エントロピー（実際）」，サンプルキーワードの頻度分布を使用して得られた最小エントロピーを「最小エントロピー（近似値）」と表記する．

最小エントロピー（近似値）は，最小エントロピー（実際）よりも高く，特に開示ビット長が 9 以下で誤差が大きい．最小エントロピー（実際）と最小エントロピー（近似値）の平均平方二乗誤差 i)は，約 0.44 であり，数式(13)を満たす最大の開示ビット長を 9 と判断する恐れがある（図 38）．

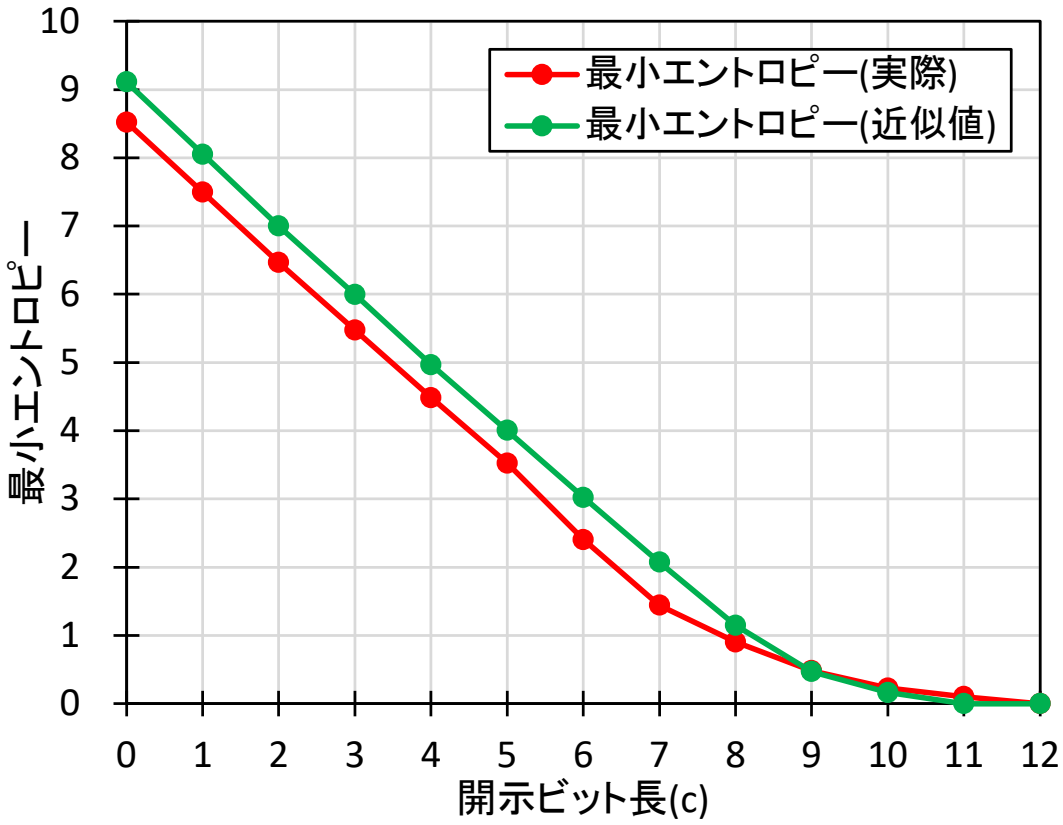


図 40 最小エントロピーの実際の値と近似値との誤差

i) Root Mean Square Error（RMSE）．x_iを真値，$\hat{x}_i$を実測値とする場合，RMSE は次のように定義される．

$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}$ ．

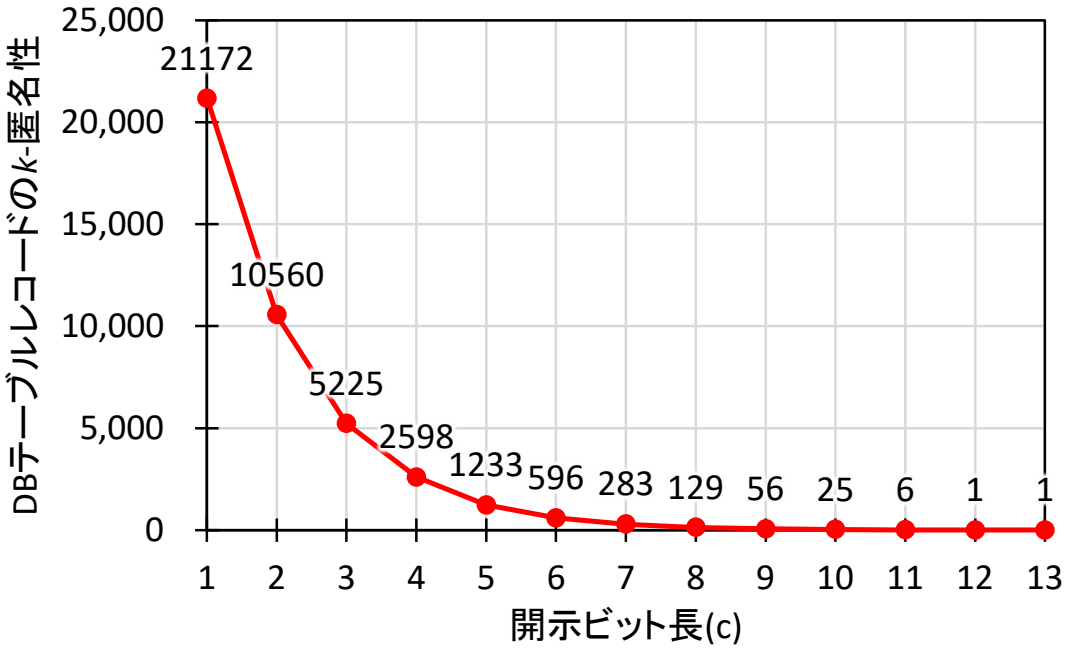


図 41 DB テーブルレコードのk-匿名性の実測値

図 42 は，実際のキーワードの頻度分布（$P(W)$）とサンプルによる近似値（$\hat{P}(\hat{W})$）を図示したものである．カーネル密度推定のバンド幅が小さいため，近似値が滑らかでないが，大まかな分布形状は類似していることが分かる．特に，最小エントロピーの算出に影響するのが，頻度が最も大きくなるキーワードの数値が 1 付近の頻度である．近似値の最大頻度が，実際の値より小さいことが，誤差の一因となっている．

しかしながら，提案方式は，kをできるだけ大きくするアプローチから，開示ビット長cを 4 に設定しており，31-匿名性以上を持つように設計している．そのため，先述の誤差が生じたとしても，本評価環境下では，2-匿名性以上を持つことから，この誤差は許容範囲内である．

一方，実際の DB テーブルレコードのk-匿名性の実測値は，$c = 4$で$k = 2{,}598$であり，31-匿名性以上を持つことを確認できる．

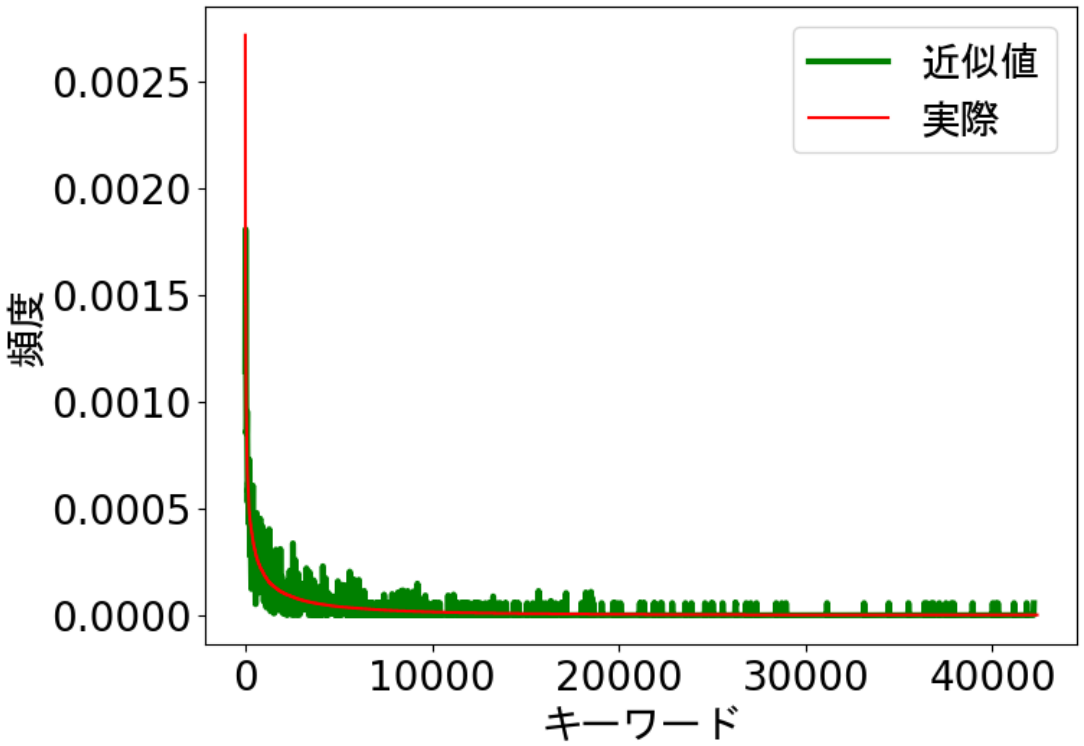


図 42 実際のキーワードの頻度分布（$P(W)$）とサンプルによる近似値（$\widehat{P}(\widehat{W})$）

4.5 考察

評価の結果から，Li ら，尾形らの方式に対して提案方式を適用すると，数十万キーワード規模において，検索時間が 4.39 秒～19.89 秒に短縮可能（91.3%～97.2%削減）であることを確認できた．よって，提案方式の使用で，3.3.3 項で述べた数十万キーワード規模での検索時間が数秒オーダであることと，キーワード数に対してスケーラブルであることを満たすことができる．

提案方式は，開示ビットによる絞り込み後の DB テーブルのレコードが最低でも 2-匿名性を持つように設計しているため，キーワードの頻度分布が漏洩する可能性も少ない．キーワードの頻度分布を推定しているため，最小エントロピーの計算結果に誤差が生じるが，本評価環境下では許容できる範囲内である．

さらに，分散 DB を使用することにより 1 つあたりの DB に格納されるタグ数が減るので，スケーラビリティが強化されると同時に，確率的暗号ベースの方式での頻度分析攻撃の成立に必要なタグの頻度の観測を困難にすることができる（DB どうしが結託しない場合）．

なお，平野らの方式と伊藤らの方式は，組み合わせることができる．この場合も開示ビットとキーワードが 1:1 に対応するケースを完全に排除するために，本研究で示した開示ビット長の設定が必要である．

今回得られた性能は，書籍データから抽出したキーワードに対してのものである．訓練画像のキーワードに対しても，同じような性能を出せるのかを考察する．経験則ではあるが，「ある領域での語の出現頻度は、その出現頻度の順位と反比例する」というジップの法則[166][167]がある．たとえば，「ある訓練データに関係するキーワード」がジップの法則に従うならば，キーワードの頻度分布は，ジップ分布の形状となると予想できる．なお，ジップ分布は，指数分布の一種であり，連続化するとパレート分布に変換可能である．今回使用したような書籍データのキーワードの出現頻度は，ジップ分布に従うことが知られているので，提案方式の適用対象であるキーワードの頻度分布が，ジップの法則に従う，または，それにジップ分布に近い分布をとるならば，本評価と同じような検索性能を出すことができる可能性がある．

4.6 まとめ

平野らの開示ビットに基づく検索可能暗号の高速化を使用する際に必要な開示ビット長を，最小エントロピーとk-匿名性を用いて決定する方法を提案した．数十万キーワード規模の入力に対して Li ら，尾形らの方式と提案方式を実装したものを用いて性能評価した結果，それぞれ 97.2%と 91.3%の検索時間の短縮が可能であることを示した．さらに，提案方式によって DB テーブルが最低でも 31-匿名性を持つように開示ビット長を設計したが，実際には2,598-匿名性を持ち，キーワードの頻度分布の推定が困難であることを確認した．

本研究では，$k = 2$を安全と定義したが，ユースケースによっては，kを大きく設定した方が適切な場合もある．そのような場合でも，提案方式は，数式(13)の不等式のkの値を変更するだけで，最適な開示ビットを計算することができる．このように，提案方式は，確保すべきセキュリティレベルをk-匿名性を用いてわかりやすく簡潔に表すことができる．

第5章 転移学習に好ましい訓練データが検索可能な特徴点マッチングの類似度指標

特徴点マッチングでは，転移学習に好ましい訓練データが検索できるような類似度指標を考える必要があると述べた．この章で提案する類似度評価手法は，データの類似度を，画像のヒストグラムの形状に着目し計算する．ヒストグラムの類似度計算において，ヒストグラムの形状が平行移動（シフト）したり，伸縮したり，相似形である場合でも類似度が高くなるように，類似度評価区間を極値で分割し，区間ごとに Dynamic Time Warping 距離を求め，各距離を結合することで類似度を得る．画像認識でのユースケースを想定した評価環境下では，ヒストグラムの形状にシフトや伸縮，相似形が存在するヒストグラムの比較においても，類似画像を抽出可能であり，さらに，ユーザによる調整が必要なパラメータが不要であることを示す．

5.1 前提条件

5.1.1 評価に使用する訓練画像

この章の評価で使用する訓練画像は，研究用途で公開されている動物の画像を用いる．本来は，工場の検品時の製品画像や，行動認識用に記録された画像，自動運転向けのドライブレコーダの画像など，ビジネスシーンを想定した画像を想定するのが理想的である．しかしながら，これらのデータはパーソナルデータを含むか，ビジネス上の機微なデータであるため研究用途では使用できない．代替として動物の画像を用いるが，見た目が似ている画像を検索するという目的に沿うよう，画像どうしの類似性を一目で判断できるような，体表に縞模様などの特徴的な柄を持つ猫科の動物を中心にデータを集めた．

5.1.2 類似性の定義と確認方法

3.3.3 項で述べたように，転移学習に有効な画像を検索できるかを目的として，画像の特徴量を「波形」と見たときに，波形どうしが平行移動したり，伸縮したりしているものほど類似性が高いと定義する．波形にこのような変化が生じる原因は，環境光などの変化に起因する照明変動や，画像に対する被写体の割合の変化などであるため，ここで定義する類似する特徴を持つ画像は，データ転移に有効と判断できる．

類似性の確認方法は，まず，転移元相当の画像と，転移先相当の複数の画像を一つずつ比較し，類似度を算出する．そして，算出した類似度を基にクラスタリングし，最後に，類似度が高いクラスタ内の画像のヒストグラムが平行移動や伸縮していることなどを目視で確認する．

5.1.3 性能要件

3.3.3 項で示した，訓練データとして有効な画像かどうかを判定する処理で要求される性能要件は，以下の 2 つである．

(1) 類似する特徴を持つ画像に対し，類似度を高く算出できること
（似ている順に画像を順位付けできること）
(2) 限られた計算資源で，実行可能であること

ただし，上記(2)の限られた計算資源での実行に関しては，処理が軽量なアルゴリズムである特徴点マッチングをベースとして採用する関係上，性能評価によって改めて検索速度などを計測することはしない．したがって，この章では，(1)のみを評価する．

5.2 貢献

この章では，転移元画像と転移先画像の類似度評価を転移元画像のヒストグラムと転移先画像のヒストグラムの類似度に基づいて行う場合において，ヒストグラムの形状がシフトしたり，伸縮したり，相似形であっても類似度を高く算出できる効果が見込める Dynamic Time Warping（DTW）距離[148]を使用した類似度評価手法を提案する．

基本的な考え方は，時系列データの類似度指標である DTW を，画像の類似度の比較に対して適用することである．ヒストグラムの形状変化のうち形状のシフトについては，2 つのデータを少しずつずらしながら類似度を算出するスライディングウィンドウ法を用いても良いが，類似度算出に大きく影響するスライド幅をユーザが調整する必要がある．

具体的には，転移先の画像の画素値を基にしたヒストグラムを作成し，そのヒストグラムをカーネル密度推定によって滑らかな曲線状に近似した後，極値で区間分割する．この分割区間ごとに転移元の画像のヒストグラムを先ほどと同様の方法で曲線近似したものとの DTW 距離を算出後，それぞれの分割区間での DTW 距離を線形結合することで類似度評価を行う．区間分割を行うことで，転移元画像のヒストグラムと転移先画像のヒストグラムの全体に対して DTW を適用したときに生じる，過度なフィッティングを防止する．

有効性の評価では，画像認識での転移学習のユースケースを想定し，類似の特徴量を持つ画像に対して，提案手法が高い類似度を持つものとして抽出できるかを検証する．さらに，この評価環境下において，スライディングウィンドウ法のスライド幅のような，ユーザによる調整が必須のパラメータが不要であることを確認する．

5.3 提案方式

転移元画像と転移先画像それぞれのヒストグラムの形状のシフトや伸縮，相似形に対応

するために，まず，転移先画像のヒストグラムの極値で評価区間を分割し，分割区間ごとに転移元画像のヒストグラムとの DTW 距離を算出後，それぞれの結果を結合し，最終的な類似度を得る方法を提案する．

提案手法では，次の手順で類似度評価を行う．

(1) ヒストグラムの導出
(2) 区間分割
(3) 分割区間の類似度評価と結合

5.3.1 ヒストグラムの導出

DTW 距離を類似度指標とし，転移元画像と転移先画像の類似度を求めるために，転移元画像と転移先画像それぞれの特徴量ベクトルを求め，特徴量ベクトルから 2 次元データ分布であるヒストグラムを計算する（図 43）．このヒストグラムどうしを DTW で比較することで，類似度を高く算出すべきヒストグラムの形状変化に対応する．本研究では，プライバシーの関係上，人物ではなく動植物の画像を評価に用いる．評価に使用するデータは，柄（模様）に特徴があるため，特徴量をグレースケール化した画素とし，画素値の出現頻度をヒストグラムとする．正確には，グレースケール化した画像の画像値からカーネル密度推定によって画素値xを確率変数とする確率質量関数$p(x)$を求め，$(x, p(x))$をヒストグラムとする．本研究では，カーネル密度推定によって関数化されたものをヒストグラムと呼ぶ．確率質量関数を使用することで，画像の母集団のデータを推定するため，画像の標本に含まれるノイズの影響を抑えることができる．

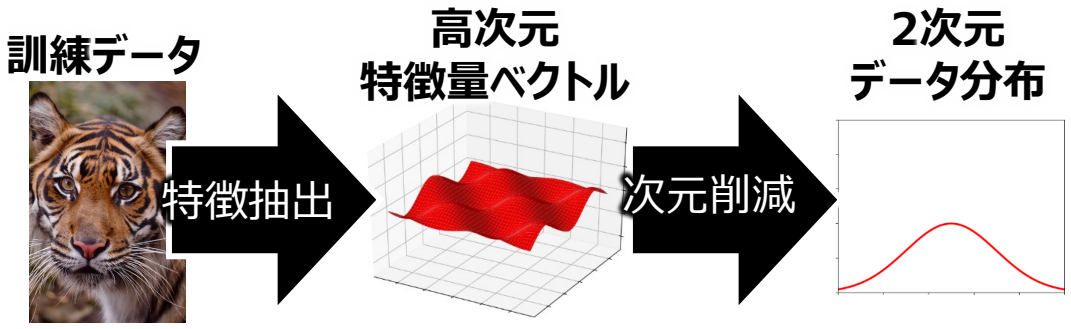


図 43 特徴抽出とヒストグラムの導出の流れ

ただし，$(x, p(x))$は転移元画像と転移先画像の特徴を十分反映しており，波形状に近似できればどのようなものでも構わない．たとえば，$(x, p(x))$に Bag-of-Features（BoF）[124]を用いる方法がある．BoF は，訓練データから抽出した特徴ベクトル群をクラスタリングし，そのクラス ID と出現頻度を生成する手法である．よって，xにクラス ID，$p(x)$にその

クラスの出現頻度を設定すればよい.

特徴抽出は，画像であればORB[114]，HOG[115]，SIFT[116]，AKAZE[117]，音声であればMFCC[118]などを用いる．BoFは特徴が出現する位置情報が損失するが，特徴次元が下がり推論時の計算量を抑えられるため，計算資源の限られたエッジ環境でも利用可能と考えられる．

このように，BoFにおける特徴抽出方法を変更することで使用することで，動植物などの画像分類の他，スマートスピーカで使用される音声認識などのユースケースにも応用可能である．

5.3.2 区間分割

5.3.1項で求めた転移元と転移先画像の$p(x)$について，ヒストグラムの形状のシフトや伸縮，相似形に対応するために，転移先の$p(x)$の極小値で分割する．極小値から次の極小値の間には必ず1つの極大値が含まれる．極大値は，$p(x)$の形状の特徴の一つと捉えることができ，この特徴的な形状が分割区間内，あるいはその近傍に含まれていれば，部分的な類似性があると判断する．

5.3.1節で述べた通り，画像をグレースケールに変換するため，画素の表現空間が8ビットである場合，画素値は0から255の間で変化する．よって本研究では，xの区間を[0,255]とする．ただし，本項では区間を$[a,b] \coloneqq \{x \in \mathbb{Z} | a \leq x \leq b\}$と定義する．転移先の$p(x)$（以降，$p_t(x)$と表記）の極小値の集合が，$\{p_t(x_1), p_t(x_2), \dots, p_t(x_n)\}$（ただし，$x_i < x_{i+1}$）であるとき，分割区間を数式(20)のように定義する．ただし，$b_0 = 0$，$b_i = x_i\,(i = 1, \dots, n)$，$b_{n+1} = 255$である．

$$[b_0, b_1], [b_1, b_2], \dots, [b_n, b_{n+1}] \tag{20}$$

ここで，分割区間に転移元の$p(x)$（以降，$p_s(x)$と表記）の極大値が入りやすくなるよう，各区間にマージンm_iを導入することを考える（数式(21)）．ただし，m_iは必須パラメータではないため，必要に応じユーザが設定する．

$$[b_0, b_1 + m_0], [b_1 - m_1, b_2 + m_1], \dots, [b_n - m_n, b_{n+1}] \tag{21}$$

本研究では，m_iを数式(22)のように定め，パラメータ$overlap$を使用して比率で表現する．

$$\left\{ m_i = overlap \cdot (b_{i+1} - b_i) \middle| \begin{matrix} i = 0, \dots, n \\ 0 \leq overlap \leq 1 \end{matrix} \right\} \tag{22}$$

そして，m_iの導入によってxの区間が[0,255]を超えないよう，分割区間Iv_iを数式(23)のよう

に修正する（図 44）.

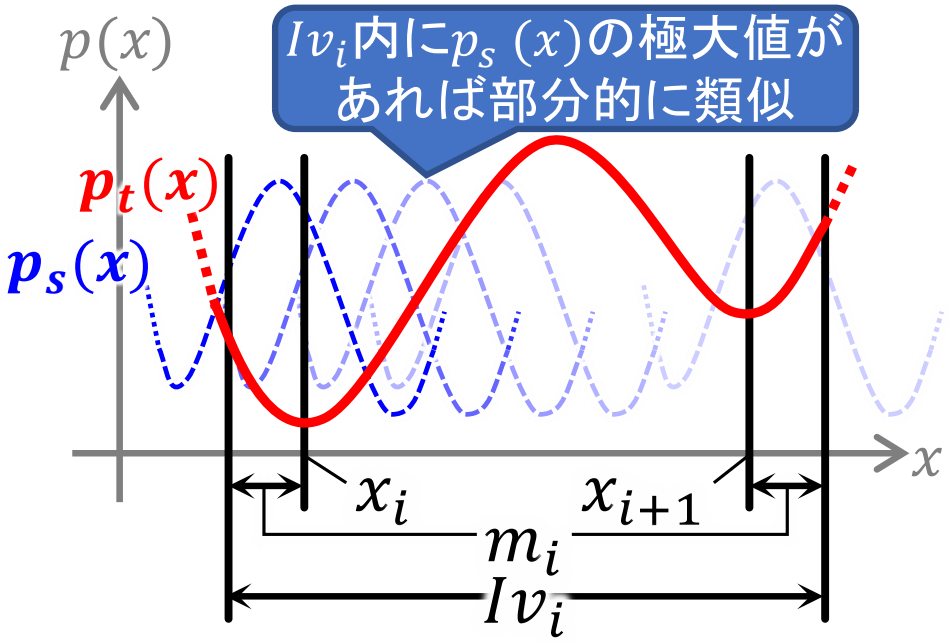


図 44 区間分割の考え方

$$Iv_i = \begin{cases} [b_i, \min(b_{i+1} + m_i, 255)], & i = 0 \\ [\max(0, b_i - m_i), \min(b_{i+1} + m_i, 255)], & i = 1,..n-1 \\ [\max(0, b_i - m_i), b_{i+1}], & i = n \end{cases} \quad (23)$$

本研究では，m_iを分割区間長に比例するように設定したが，原理的には固定値でも良い．m_iを分割区間長に比例させる利点は，$overlap$パラメータにより分割区間長が最大でも分割区間の 3 倍となるため，過剰な形状一致を抑制できることである．単純には，スライディングウィンドウ法のように$p_s(x)$の元波形全体をずらしながら DTW 距離を算出する方法もあるが，波形をずらす際に DTW 距離算出の対象とならない部分が発生し，類似度算出精度が低下する恐れがある．加えて，ユーザによるスライド幅の調整が必要になる．

図 45 は，提案手法の区間分割と DTW の適用イメージである．赤実線が$p_t(x)$の波形で，青破線が$p_s(x)$の波形である．区間分割を行わない通常の DTW では，元波形の位相のずれが大きい場合でも形状を一致させている（図 45 上段右）．しかし，位相のずれが大きすぎる場合，転移元と転移先の画像の類似性が著しく低下する恐れがある．図 46 は，不適切な形状一致の例である．画像 1 と画像 2 は，見た目が異なり，ヒストグラムの極大値のシフト量も大きいが，DTW では，このような場合でも波形の大部分が一致してしまう．区間分割を行うと，$p_t(x)$の波形の極小値で分割され，その区間内で$p_s(x)$の波形との形状一致判定が行われる（図 45 中段右）．$overlap$ =0 の場合のIv_1，Iv_2に着目すると，$p_s(x)$の波形の極

大値が部分的にしか含まれないため，半分程度の一致に留まる．一方，$overlap = 0.4$では，Iv_1，Iv_2内に$p_s(x)$の波形の極大値を含むようになるため，大部分が一致するようになる（図 45 下段右）．このように，m_iの値を変化させることで，波形の形状一致の度合いを制御することができるが，m_iの導入によって必要以上に形状一致が許容される恐れがあり，m_iの要否も含め調整の余地がある．

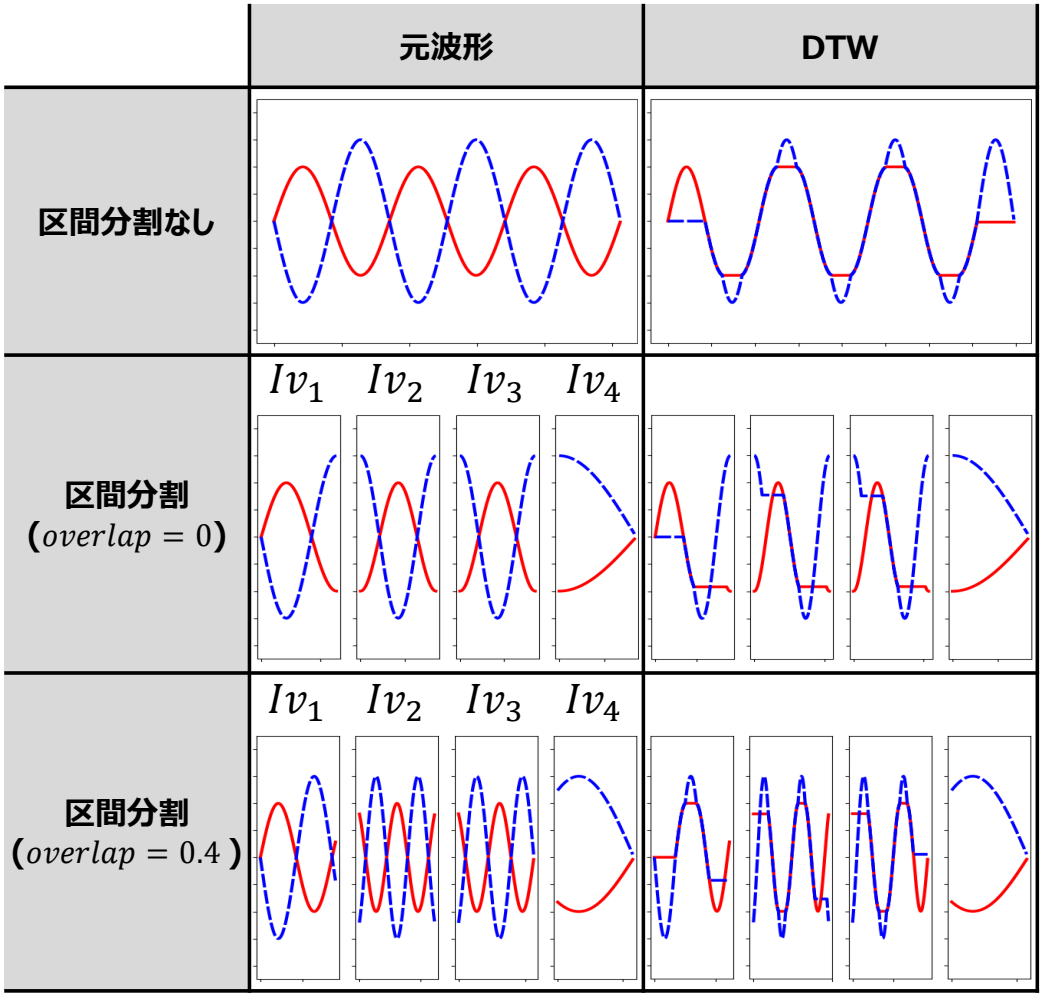


図 45 提案手法の区間分割と DTW 適用イメージ

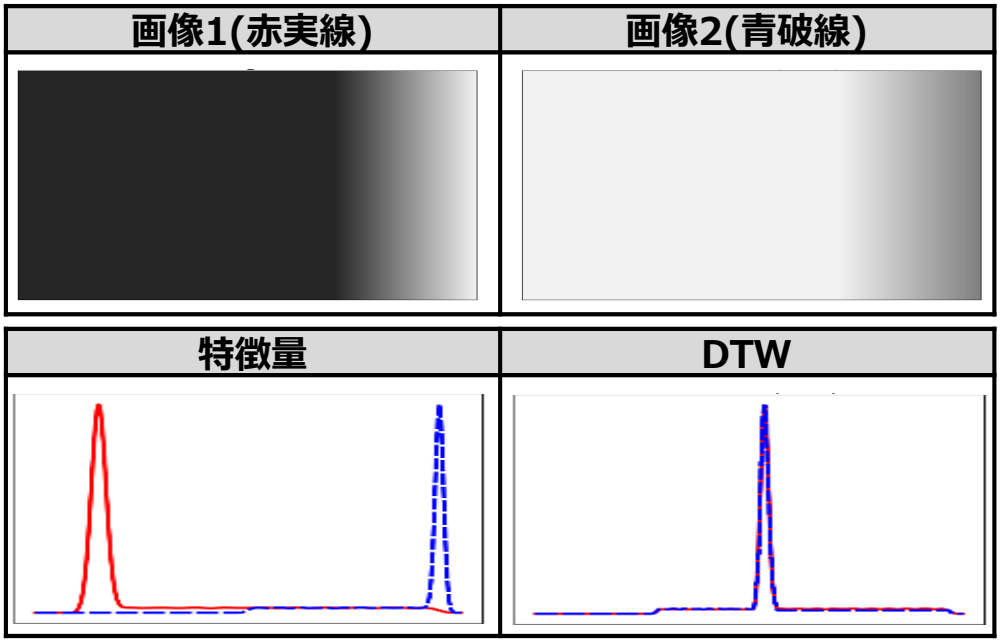


図 46 不適切な形状一致例

5.3.3 分割区間の類似度評価と結合

特徴として重視する極大値を含む区間の DTW 距離を類似度に強く反映させるために，各分割区間の DTW 距離に重み付けをして結合する．

類似度評価の前に$p_t(x)$，$p_s(x)$の値を[0,1]に正規化する．この処理は，画像どうしの大きさの違いや，照明変動に伴うヒストグラムの形状変化を軽減するためのものである．図 47 は，2 つの画像を提案手法の手順で比較する際の，正規化の効果を示したものである．まず，2 つの画像を入力として，それぞれヒストグラムを作成する．次に，これらのヒストグラムを波形化するために，カーネル密度推定を行う．このとき，図中の赤の波形の最大ピークが低く，青の波形の最大ピークが高いことが分かる．この状態で比較してしまうと，照明変動などの影響を強く受け，類似度を必要以上に低く算出してしまう．提案手法において，正規化後の波形の比較に使用する DTW は，x軸方向のシフトや伸縮は許容するが（図 48 (1)），$p(x)$軸方向の伸縮（極値の高さ）は許容しないためである．そこで，これらの波形を正規化して，最大ピークの高さを合わせる．これによりに，前述の照明変動などの影響を最小限に抑える．

一方，転移学習では，転移元画像を変換してから転移することがある[88]．たとえば，$p_t(x)$，$p_s(x)$の全体形状について，スケールのみが異なる場合，$p_s(x)$を定数倍する変換を介すことで転移が可能となる．そこで，$p_t(x)$，$p_s(x)$をそれぞれ最大値と最小値を基準にスケール変換し，$p(x)$軸方向の伸縮（図 48 (2)）に定数倍の許容を持たせる．正規化後の$p(x_j)$の値$p(x_j)_{norm}$を，数式(24)に示す．ここで，$p(x)_{min}$は$p(x)$の最小値，$p(x)_{max}$は$p(x)$の最大値である．

$$p(x_j)_{norm} = \frac{p(x_j) - p(x)_{min}}{p(x)_{max} - p(x)_{min}} \tag{24}$$

正規化済の$p_t(x)$を$p_t(x)_{norm}$，転移元画像集合$\mathbb{S}$の元$s_k \in \mathbb{S}$の正規化済の$p_{s_k}(x)$を$p_{s_k}(x)_{norm}$とすると，重みパラメータ$w_i \in [0,1]$を用いて，類似度$\boldsymbol{s}(t, s_k)$を数式(25)のように定義する．ただし，$\boldsymbol{DTW}(t,s)_{Iv_i}$は分割区間$Iv_i$における時系列データ$t$，$s$の DTW 距離である．

$$\boldsymbol{s}(t, s_k) = \sum_{i=0}^{n} w_i \cdot \boldsymbol{DTW}(p_t(x)_{norm}, p_{s_k}(x)_{norm})_{Iv_i} \tag{25}$$

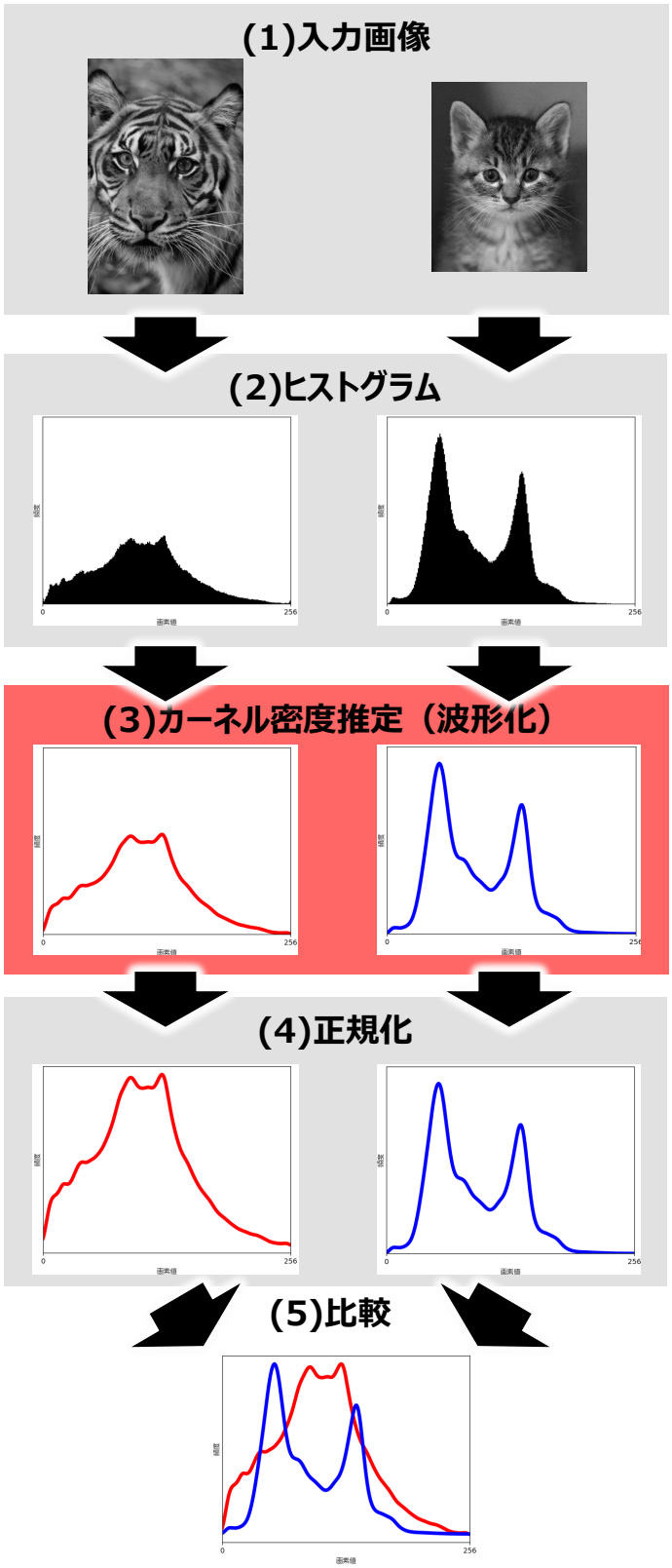


図 47 正規化の効果

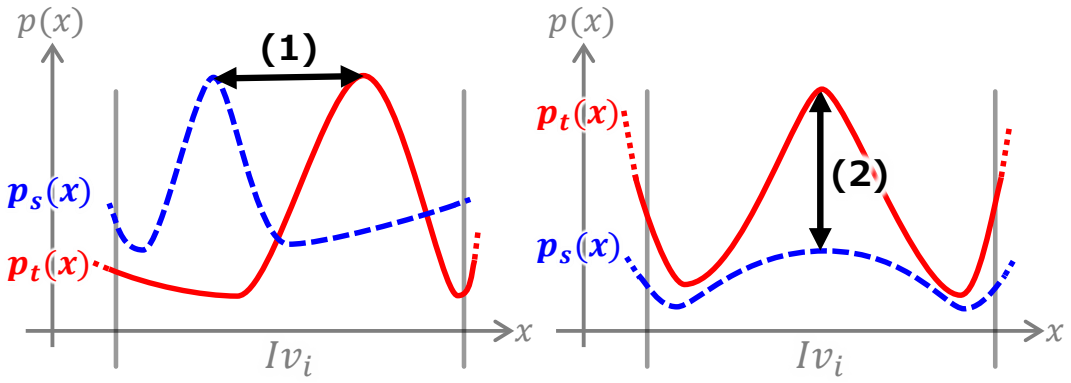


図 48 極大値の伸縮の許容範囲のイメージ

5.4 評価

5.4.1 評価方法

提案手法の有効性を評価するために，AI がよく用いられる画像認識分野での転移学習のユースケースを想定して，類似する特徴量を持つ画像に対し，提案手法が高い類似度を持つものとして順位付けできるかを確認する．

提案手法で用いている区間分割には調整の余地があり，一旦マージンm_i（$overlap$パラメータ）を導入するが，m_iの変化に対して類似度が大きく変化せず，$overlap$の調整は実用上不要であることを評価により示す．

A. 前提条件

評価用の画像は，5.3.1 項で述べたようにプライバシーの関係上，人物ではなく動植物の画像を用いる．これらの画像は，光源やコントラスト，画像に対する被写体の占有率が異なることに加え，被写体が似ているが同一の動植物でないケースを想定する．たとえば，個人の健康管理での転移学習では，異なる住居に同一人物が居る状況が少なく，環境ノイズが存在する条件下で身体的特徴が似た人物を探すのが現実的である．さらに，立つ，座る，転ぶなどの人物の動作認識の学習では，完全に同一被写体でなくても，転移学習が有効である可能性が高い．

本評価では，転移先画像中の動物と似ている動物を「分類学的に似ている生物」と定義する．転移元画像セットには，転移先の動物画像と似ている動物の他，分類系統上の距離が離れた動物や植物画像をいくつか入れる．この転移元画像セットから転移先画像の動物と似ている動物を含む画像の類似度を高く算出できるかを評価する．

本研究で想定する画像認識分野での転移学習では，類似画像の検索においてリアルタイム性を求めない夜間バッチ等での実行を想定する．そのため，検索時間は評価対象外とする．加えて，転移学習に必要な画像を収集する範囲をある程度広くする必要があるため，転移元

の画像を保存・検索する環境は，ある程度の計算能力のある PC 相当のエッジデバイスとする．

評価対象の画像から得られる特徴量は，画像をグレースケール化後，カーネル密度推定によって求めた画素値の確率質量関数を[0,1]に正規化したもの（$p(x)_{norm}$）とする．$p(x)_{norm}$は，画像の特徴を良く抽出できているものとする．即ち，$p_t(x)_{norm}$と$p_s(x)_{norm}$が類似の形状を持つ場合，元の画像どうしの類似度も高くなる．

画像は背景などの識別対象以外の領域（ノイズ）を含むが，本評価ではノイズ除去は実施しない．加えて，画像解像度や画像サイズの変更も行わない．

B. 評価項目

転移元の画像集合を，転移先の画像（*target* 画像）との類似度が大きい順に順位付けできているかを確認する．本評価では，動植物の画像を用いて，以下を評価する．

(1) *target* 画像と似た画像ほど順位が上位であるか
(2) *overlap*パラメータによって，類似度が著しく変動しないか

提案手法と比較する方式は，筆者らのピアソン相関係数を用いる方式[168]（以降，「ピアソン相関係数」と表記），区間分割を行わず DTW を適用する方式（以降，「DTW」と表記）の 2 方式とする．

評価(1)における類似度評価手法ごとの評価関数と順位付け規則を，表 23 に示す．各手法は，$p_t(x)_{norm}$と$p_s(x)_{norm}$を入力として評価関数を用いて類似度を出力し，順位付け規則に従い順位を決定する．

提案手法の$overlap$パラメータによる順位変動とその時の類似度の変化を確認する．理論上は，$overlap$が 0 に近いほどピアソン相関係数での類似度評価結果に近づき，$overlap$が 1 に近いほど DTW での類似度評価結果に近づくはずである．なお，評価(1)における順位は，$p_t(x)_{norm}$と$p_s(x)_{norm}$の形状が近い画像ほど高くなると定義したため，背景などの環境ノイズによって，*target* 画像の動植物の分類系統と近いものが上位に来ない場合があることに注意する．

評価(2)では，$overlap$=0 と$overlap > 0$の場合で，類似度が著しく変動しないことを確認し，マージンm_iの導入が実用上不要であることを示す．提案手法，ピアソン相関係数，DTW がどの程度の距離（類似度の差）をもって順位付けしているかを可視化し，順位付けのロバスト性を確認する．たとえば，似ている画像どうし（猫の画像どうしやチーターの画像どうしなど)は，多少順位が入れ替わっても問題ないため，似ている画像どうしは，近い距離に配置されるのが望ましい．一方，明らかに似ていない画像どうしは，遠い距離に配置するのがよい．即ち，上記の特性を持っていれば，類似画像の類似度がノイズによって多少変化したとしても，順位変動を抑えられる（ロバスト性がある）といえる．

表 23 に示した通り，評価関数の変域および順位付け規則が類似度評価手法ごとに異なるため，転移元画像間の類似度の距離を共通の基準で確認できるよう評価関数を次のように変更する．まず，ピアソン相関係数の評価関数に対してのみ出力値に-1を乗算し，順位付け規則を昇順にする．そして，全ての手法の評価関数の出力値を，数式(24)と同様の方法で[0,1]に正規化する（以降，この値を「相対距離」と表記する）．
この相対距離を基にクラスタリングし，$overlap$パラメータを変化させても，クラスタリング結果が変わらないことを確認する．さらに，転移用の訓練データセットの構築において最も重要なのは，転移先画像との類似度が高い転移元画像が検索されることである．つまり，相対距離が最も近いクラスタが変動しなければ，有効な訓練データセットの構築に必要なデータが，検索によって抽出されやすくなると考えられる．

表 23 評価関数と順位付け規則

類似度評価手法	評価関数	順位付け規則
提案手法	$s(t, s_k)$ （変域:= $[0, \infty]$）	昇順
ピアソン相関係数	ピアソン相関係数 （変域:= $[-1,1]$）	降順
DTW	DTW 距離 （変域:= $[0, \infty]$）	昇順

5.4.2 評価条件

表 24 に評価条件の一覧を示す．評価に用いる類似度評価手法は，提案方式，ピアソン相関係数，DTW の 3 つである．$overlap$パラメータは，0，0.05，0.1，0.2，0.4 と変化させる．w_iは 1 に設定し，本評価では重みを付けないことにする．評価用画像は，ImageNet[169]からダウンロードし 16 枚の画像（図 49）で，1 枚を転移先画像，残りの 15 枚を転移元画像に設定した．

表 24 評価条件一覧

類似度評価手法	提案手法，ピアソン相関係数，DTW
$overlap$	0，0.05，0.1，0.2，0.4
w_i	1（重み付けなし）
評価用画像	動植物画像 16 枚（図 49） （*target* 画像 1 枚，転移元画像集合 15 枚） （ImageNet の API から入手[169]）

図 49 評価用画像一覧

5.4.3 評価環境

PC のハードウェア環境を表 25 に示す．エッジとしては高スペックではあるが，本評価では類似度の性能を見るだけなので，処理速度の速いサーバを用意した．

ソフトウェア環境を表 26 に示す．Windows 10 上に Python 3.6 をインストールし，評価用のプログラムを動作させた．評価プログラムで使用した主なライブラリは，OpenCV，scipy，dtw である．

表 25 PC ハードウェア環境

CPU	Intel Core i5-8500 CPU @ 3GHz （6 コア 6 スレッド）
メモリ	64GB
ディスク	Samsung SSD 860 EVO 1TB × 2 （シーケンシャル Read 550 MB/s） （シーケンシャル Write 520 MB/s）

表 26 PC ソフトウェア環境

OS	Windows 10 Enterprise バージョン 1909 (64bit)
ランタイム	Python 3.6
主な使用ライブラリ	OpenCV[170]（グレースケール化） scipy.stats[171]（カーネル密度推定） scipy.signal[172]（極値算出） dtw[173]（DTW 距離算出）

5.4.4 評価結果

A. 手法ごとの類似順位の変化

target 画像の特徴量（赤実線）と転移元画像の特徴量（青破線），$overlap = 0$時の区間分割位置（縦実線）をプロットしたグラフを図 50 に示す．手法ごとの類似順位の変化をプロットしたものを図 51 に示す．

動植物の表面の柄によって画素値の確率質量関数が決まると期待するため，*target* 画像（虎）は，*tiger1*，*tiger2* や *cheetah1*，*cheetah2* などの順位が高いことが望ましい．
ピアソン相関係数は増減関係を重視するため，*target* 画像の特徴量と形状が重なるような *cheetah1*（1 位），*tiger2*（2 位），*tiger1*（4 位）は上位だが，*cheetah2*（7 位）のように極大値を引き延ばした形状の順位が下がり，x軸方向の伸縮への許容性が低い．一方，*giraffe1*（5 位）や *cat1*（6 位）のように，ヒストグラムの増減関係が似ているものの順位が高く出ており，$p(x)$軸方向の伸縮への許容性が高すぎる．

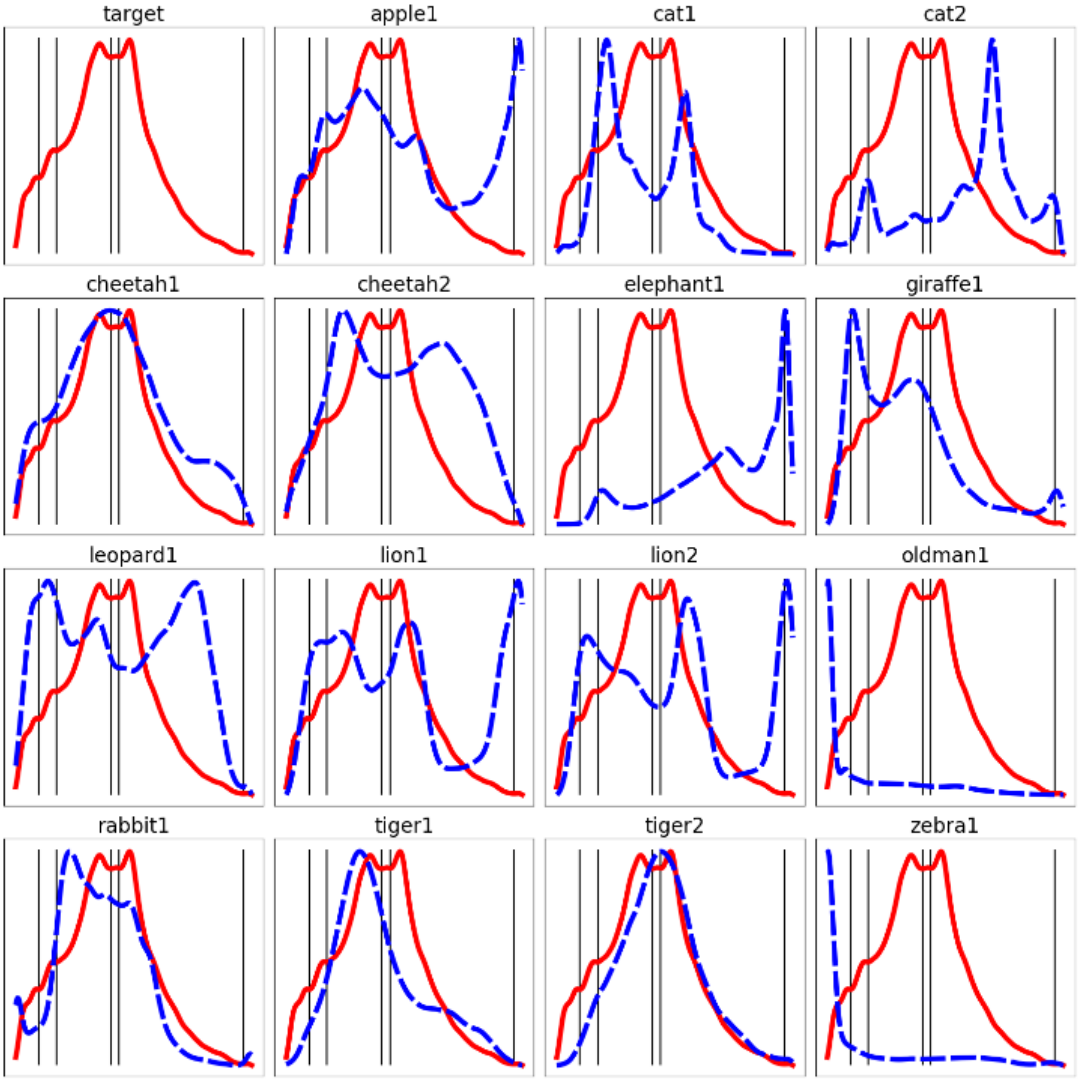


図 50 特徴量と分割区間

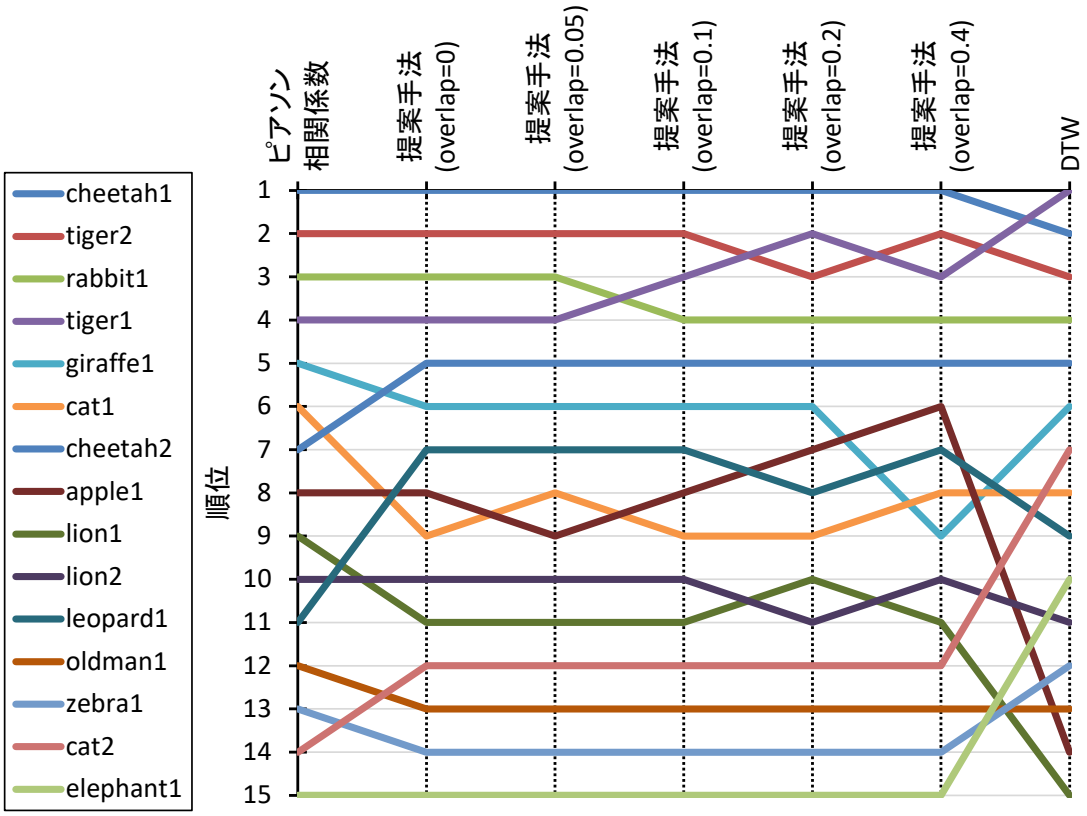


図 51 手法ごとの類似順位の変化

提案手法（*overlap* = 0）は，*cheetah2*（7 位→5 位），*leopard1*（11 位→7 位）の順位が上がり，*giraffe1*（5 位→6 位）や *cat1*（6 位→9 位）の順位が下がっており，期待した結果に近づく．*overlap* を 1 に近づけると，極値のシフトに寛容となり，似た特徴量を持つ *cheetah1*, *tiger2*, *rabbit1*, *tiger1* の順位が変動するようになる．同様に，*giraffe1*, *leopard1*, *apple1*，*cat1* の間と，*lion1*，*lion2* の間の順位にも変動が生じる．

DTW は，*cat2* や *elephant1* のような極値が大きくシフトするパターンの順位を高くしている．一方で，*apple1* や *lion1* のような極値の数が異なるパターンの順位を低くする．提案手法の*overlap*が 1 に近いパターンよりもかなり極値のシフトに寛容となっていることが分かる．

B.　overlapパラメータによる類似度の変化

target 画像との類似度の相対距離の変化を手法ごとにプロットしたものを図 52 に示す．図 53 は，*target* 画像との類似度の相対距離を用いて，*k*-*means*（*k*=6）でクラスタリングした例である．図中の星印は，各クラスタの中心点である．なお，相対距離が 0 に近いほど *target* 画像と似ていることを示している．

重要なのは相対距離が最も近いクラスタが変動しないことである．本評価条件での類似

画像検索結果で期待する範囲を図 54 に示す．すなわち，分類系統において近いものどうしが近くになるように，トラ，ライオン，チーター，ネコの画像が，相対距離の最も近いクラスタに配置されればよい．

ピアソン相関係数，DTW，提案手法とも，相対距離の最も近いクラスタにトラ，チーター，ネコが含まれ，図 54 で示した期待する範囲内である．よって，時系列データの類似度計算手法が有効に機能しており，本評価条件においては類似画像検索の最も重要な要件を満たしているといえる．

しかし，ピアソン相関係数は，{*cheetah2* , *giraffe1* , *cat1*}や{*zebra1* , *cat2*}をクラスタリングしており，分類系統においてやや離れているものが同一クラスタに配置されている．

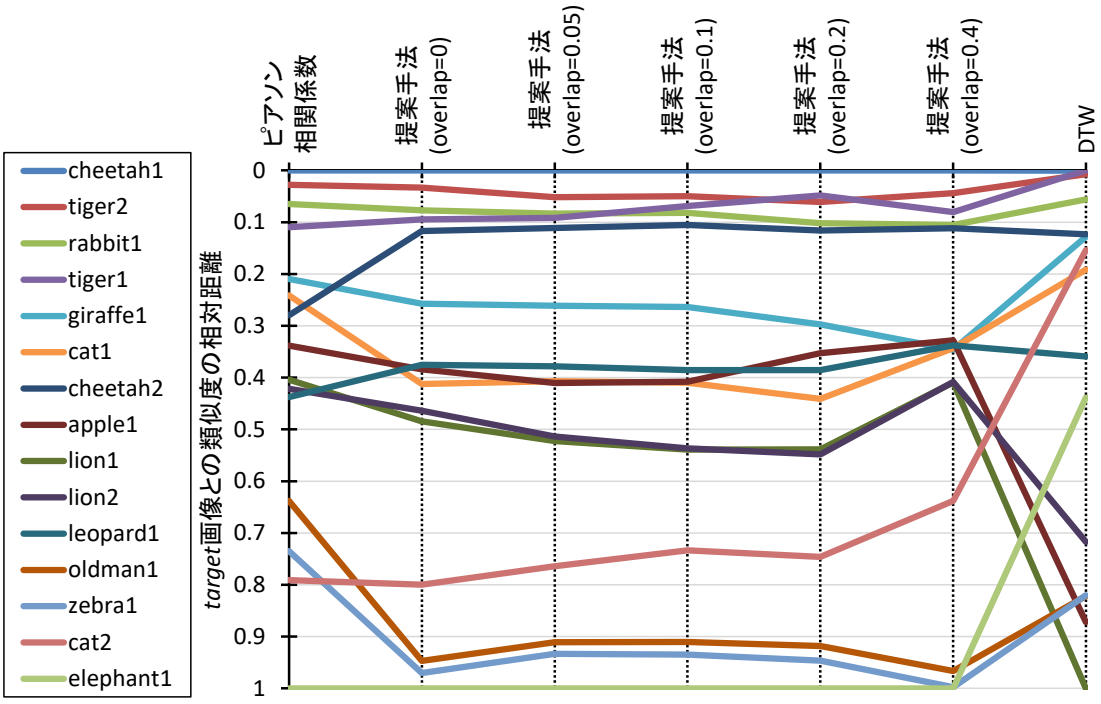


図 52 手法ごとの類似度の相対距離の変化

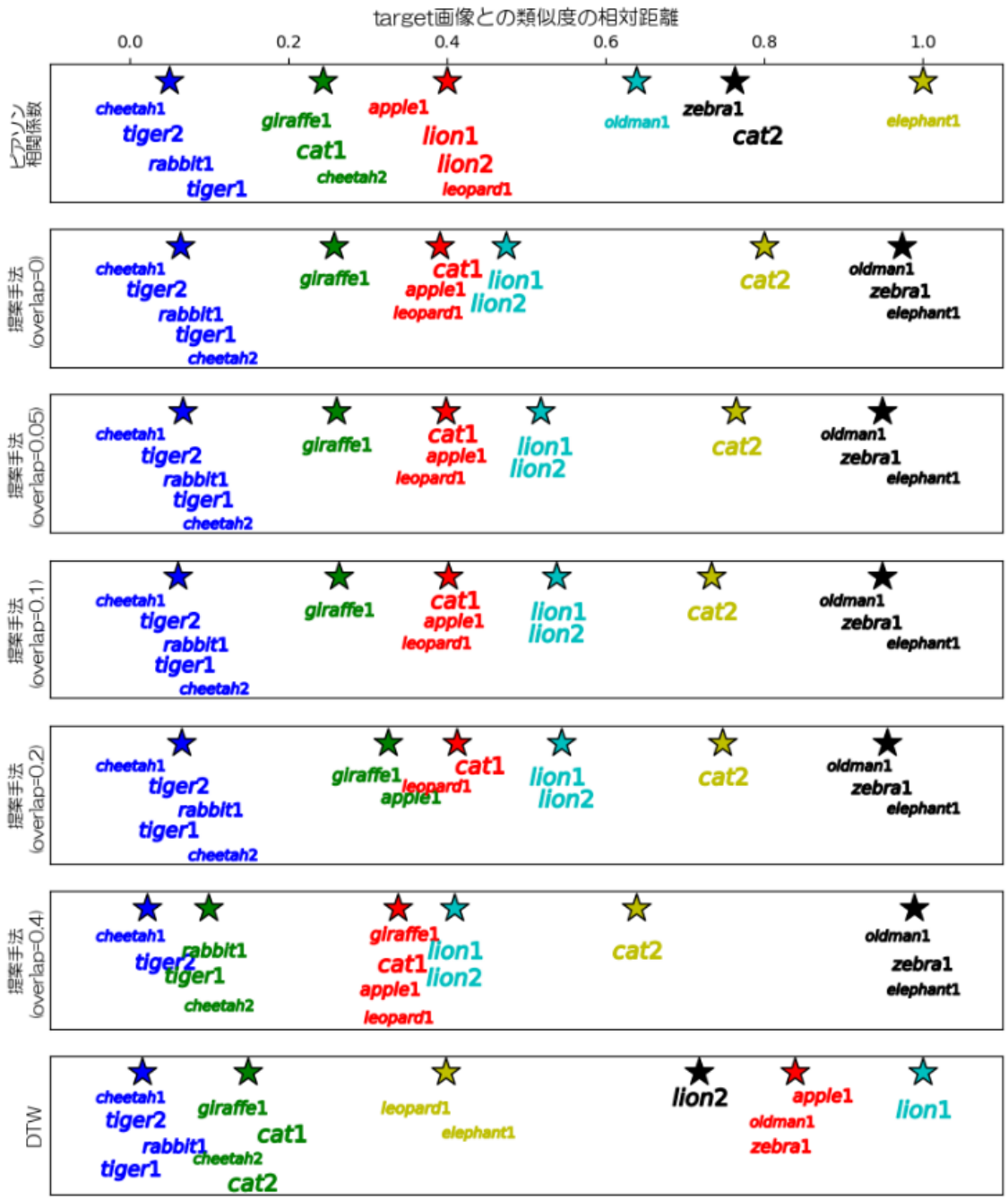


図 53 *k-means*（*k*=6）クラスタリング例

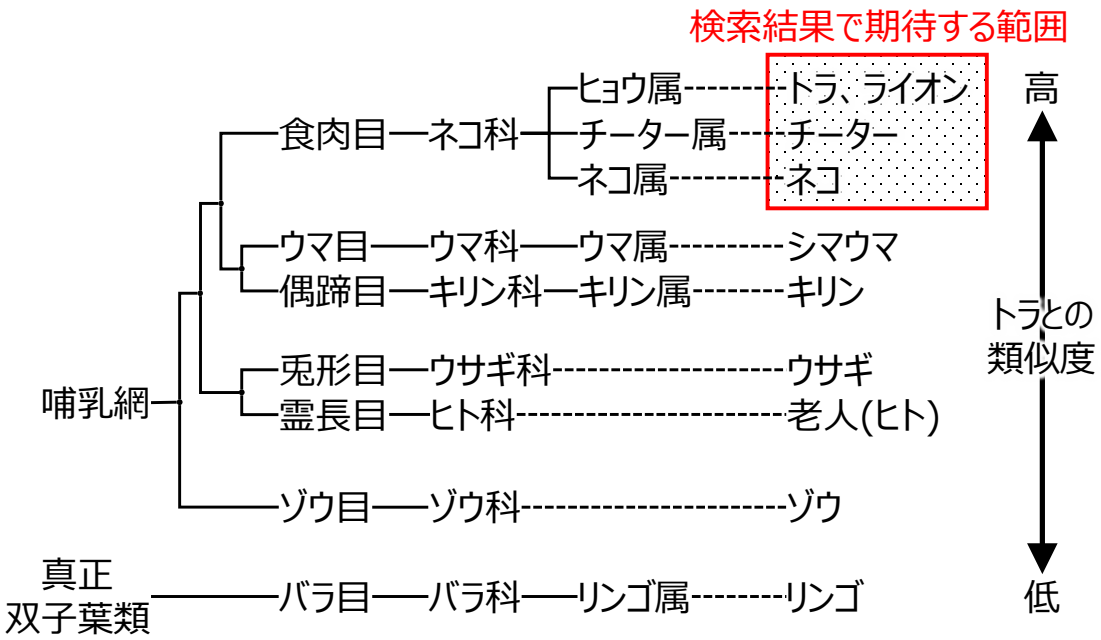


図 54 類似画像検索の結果で期待する範囲

同様に，DTW でも，*lion1*，*lion2* が異なるクラスタに分けられ，{*leopard1*, *elephant1*} をクラスタリングするなど，分類系統の近いものどうしを近くに配置していない．一方，提案手法は*overlap* = 0～0.1 の場合，{*cheetah1*, *cheetah2*, *tiger1*, *tiger2*, *rabbit1*}や{*lion1*, *lion2*}のようにネコ科のものどうしをクラスタリングしており，{*giraffe1*}や{*oldman1*, *zebra1*, *elephant1*}のようなトラと分類系統上離れているものは，異なるクラスタに配置しているため，想定する動作に近い．

rabbit1 や *apple1* がネコ科のクラスタに入っているが，これは画像に識別対象以外の環境ノイズ（背景など）が混入して特徴量のヒストグラムの形状が *target* 画像のものと意図せず似たことが影響していると考えられる．一方で，*overlap*を大きくしても，おおむねクラスタリング結果は変わらないが，クラスタ間距離の近いものどうしが近づく傾向にある．ただし，順位の変動はクラスタ内に留まっている．

本評価では，類似度を基に順位付けしたが，*k-means* のようなクラスタリングアルゴリズムを利用し，*target* 画像との類似度が近い順にクラスタを特定した後，各クラスタ内の画像を提示すれば順位を安定させることが可能である．

5.5 考察

本評価においては，転移先と転移元画像の特徴量の極値のシフトや極値の形状の相似形に対して，提案手法がピアソン相関係数と比較し，適切なロバスト性をもって類似度の評価が可能であることを確認した．さらに，DTW で問題となる過剰な形状一致による類似度の誤判定を，提案手法の極値による区間分割によって解決することができた．

提案手法は，$overlap$パラメータの導入により，類似順位と類似度が若干変化したが，変動の範囲は限定的であった．実用上は，クラスタリングアルゴリズムの使用によって順位変動を吸収できる見込みで，本評価条件下では，$overlap$パラメータは不要である．

制約事項として，類似性評価対象データの特徴量は，本評価で用いた$p(x)_{norm}$のように，x_iに対し$p(x_i)_{norm}$がただ一つ定まるものを使用する必要がある．そして，提案手法は特徴量のヒストグラムの極値を重視するため，訓練データから特徴量を抽出する段階で，外れ値やノイズを含むとその影響を受けやすい．本評価では，画素値の確率質量関数を特徴量としたため，画像全体の色彩を特徴として強く反映する．したがって，背景画像の割合が大きいと意図しない特徴が抽出されることがある．以上のことから，一般的な機械学習における訓練データの前処理と同じく，特徴抽出時に外れ値やノイズの除去を行い，意図する特徴のみが含まれるように留意しなければならない．

本評価では，画像の特徴量としてヒストグラムを用いたが，画像から得られる値を変形していき，最終的に 2 次元の波形状になれば提案方式を適用可能である．さらに，提案方式は，画像のほか，2 次元の波形状の特徴を持つ時系列データでも適用可能である．ただし，特徴の値の数が少なく波形とならない場合は提案方式を適用できないことに注意する．

5.6 まとめ

本研究では，転移学習のためのデータ類似度評価手法を提案した．提案手法は，特徴量のヒストグラムの比較における極値のシフトや極値の形状の相似形に対応するために，データ類似度評価区間を極値で分割し，区間ごとに DTW 距離を算出し，各距離を結合して最終的な類似度を得る．画像認識での転移学習のユースケースを想定した評価の結果，極値のシフトや極値の形状の類似形が存在する特徴量のヒストグラムの比較においても，類似画像を抽出可能であった．同評価環境下では，提案手法は，ユーザによる調整が必要なパラメータが不要であることを確認した．

第6章 結論

エッジ AI における訓練データ検索システムを実現するためには，クラウドにおいてスケーラビリティとセキュリティを維持しながら暗号化されたキーワードどうしを検索する技術と，エッジにおいて訓練データとして有効な画像かどうかを判定する技術の開発が必要であった．

クラウドとエッジは，担保すべきセキュリティ対策とプライバシー保護のレベルが異なり，処理も分離可能であることから，これらの技術開発を独立した 2 つの問題ととらえ，各々の問題を解決することで訓練データ検索システムが構築可能であることを示した．

本研究では，スケーラビリティとセキュリティを維持しながら暗号化されたキーワードどうしを検索する技術と，訓練データとして有効な画像かどうかを判定する技術の 2 つに対して，それぞれ以下の 2 つの新しい手法を提案し，その有効性を確認した（図 55）．

提案手法(1) 検索速度とセキュリティを考慮した検索可能暗号の高速化パラメータの決定手法

提案手法(2) 転移学習に好ましい訓練データが検索可能な特徴点マッチングの類似度指標

提案手法(1)は，検索性能とセキュリティがトレードオフとなる検索可能暗号において，検索に使用するキーワードの最小エントロピーとk-匿名性を用いて，検索性能とセキュリティがバランスする高速化パラメータを自動的に求める．評価の結果，数十万キーワード規模の場合で検索時間を最大 97.2%削減でき，提案手法によってデータベースが最低でも 31-匿名性を持つような設計に対して，実際は2,598-匿名性を持ち，高速化と安全性を両立することを示した．

提案手法(2)は，時系列データの類似 2 つの訓練画像データの類似度を，画像のヒストグラムの形状に着目して算出する．具体的には，2 つのヒストグラムの類似度計算において，ヒストグラムの形状が平行移動（シフト）したり，伸縮したり，相似形である場合でも類似度が高くなるように，ヒストグラムを波形とみなして類似度評価区間を極値で分割し，区間ごとに Dynamic Time Warping（DTW）距離を求め，各距離を結合することで類似度を得る．評価の結果，画像認識でのユースケースを想定した評価環境下では，2 つのヒストグラムの形状にシフトや伸縮，相似形が存在する場合の比較においても，類似画像を抽出可能であり，さらに，ユーザによる調整が必要なパラメータが不要であることを示した．

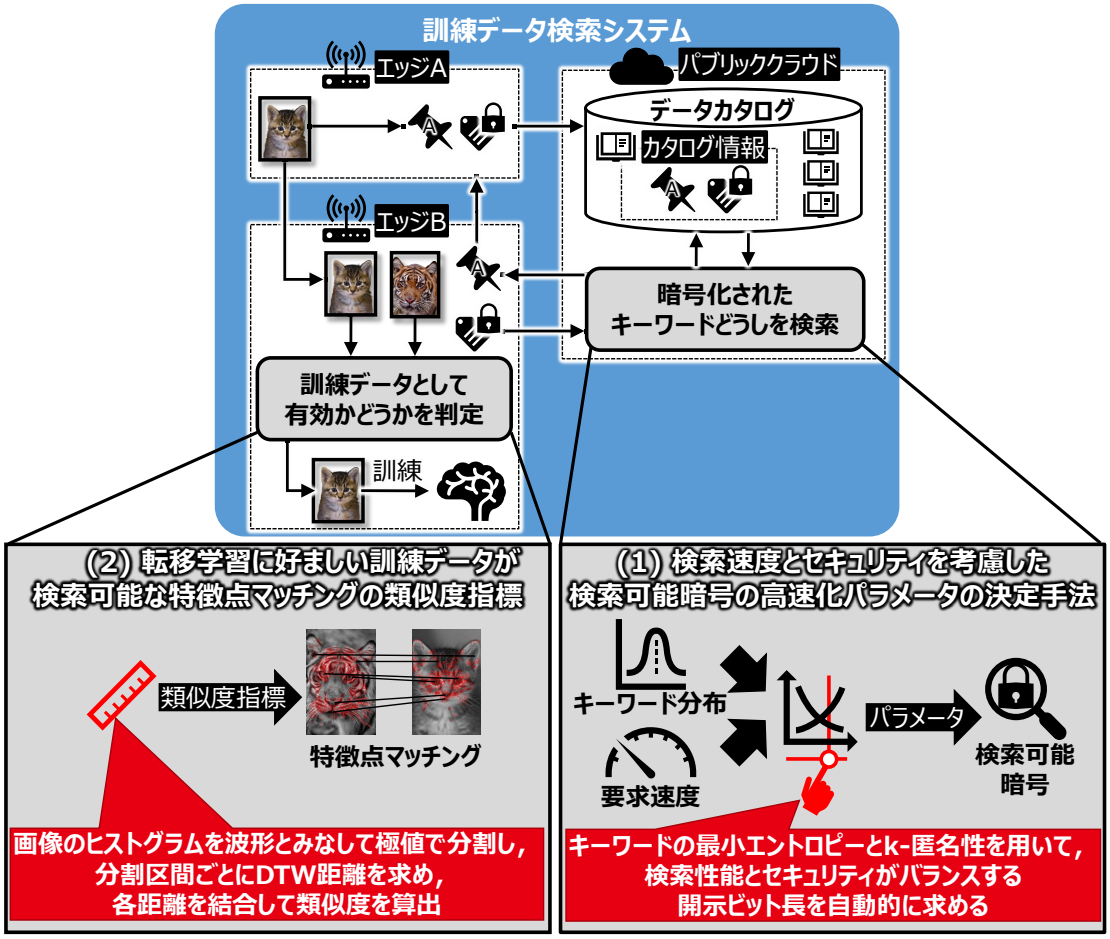


図 55 2 つの新しい提案手法

提案手法(1)により，十分なスケーラビリティを持ち，高いセキュリティを維持しつつ，検索速度が向上する秘匿検索技術を確立した．提案手法(2)により，環境ノイズに対して不変性を持つ訓練画像データを検索できるようになり，エッジが収集した，ドメインが似ている画像データを AI の学習に相互利用できる技術を確立した．これらの結果から，エッジ AI における訓練データ検索システムの実現課題である 2 つの独立した問題が解決され，システムが構築可能であることが示された．

今後は，エッジ AI における訓練データ検索システムの運用方法を検討する．具体的には，システムを継続的に改善する方法や，プライバシー保護の観点でのデータリネージの実現方法などである．データリネージは，データの所出や流れを追跡できるようにするための履歴情報である．このデータリネージは，パーソナルデータの利用の終了に伴う確実なデータ消去に不可欠である．一方，システムの継続的な改善方法については，たとえば，訓練データのキーワードの分布が運用中に変化する状況で，どのように開示ビットを適応的に変更していくかや，訓練データの検索結果に関するユーザのフィードバックを基に，ユーザの好みに合わせて類似度指標をカスタマイズする方法などに取り組む．

付録

付録 1.用語定義

本論文で使用する，主な用語の定義を表 27 に示す.

表 27 用語定義

用語	説明
エッジ	• デバイスと，そのデバイスを束ねるゲートウェイの総称 • ゲートウェイが多段となっている場合，企業あるいはユーザの完全管理下にある信頼できる範囲までとする
デバイス	• センサやアクチュエータを持つ末端の機器 • 計算能力は，高々PC 程度 • 企業あるいはユーザの完全管理下にあり，信頼できるものとする
ゲートウェイ	• デバイスを束ねるルータやサーバの総称 • ゲートウェイをさらに束ねるゲートウェイも想定する • 計算能力は，高々PC 程度 • 企業あるいはユーザの完全管理下にあり，信頼できるものとする
クラウド	• Microsoft Azure，AWS（Amazon Web Services），GCP（Google Cloud Platform）などのパブリッククラウド • 第三者機関であり，完全に信頼はしないが，悪意がある可能性があるのは，クラウドの管理者のみとし，クラウド内のシステムは信頼できるものとする
クライアント	• クライアントサーバモデルにおけるクライアント • サーバが提供する機能や情報（サービス）を利用する • エッジ AI では，エッジがクライアントに相当
サーバ	• クライアントサーバモデルにおけるサーバ • 機能や情報（サービス）を提供する • エッジ AI では，クラウドがサーバに相当 • クラウド上のサーバは，論理的に分離された計算領域上に配置される場合がある
AI	• 人工知能のことであり，機械学習と深層学習を含む概念
機械学習	• AI のうち，SVM（Support Vector Machine）やk近傍法（k-Nearest Neighbor algorithm）などの古典的なアルゴリズムを使用して，コンピュータが学習を行うもの

深層学習	• 機械学習のうち，多層の層からなるニューラルネットワークを使用して，コンピュータが学習を行うもの
学習	• 訓練データを入力して学習し，モデルを作成する • 学習の方法は，AI のアルゴリズムによって異なる
推論	• テストデータを入力して，モデルを使用して推論する • 推論の方法は，AI のアルゴリズムによって異なる
訓練データ	• 学習に使用するデータ • 画像，時系列データなどがある • 訓練データには，キーワードが付与されている • キーワードは，教師データとなることがある
テストデータ	• 推論に使用するデータ • 訓練データに画像を用いた場合は，テストデータも画像を用いるなど，通常，訓練データと同じ種類のデータを使用する • テストデータには，通常，教師データとなるキーワードは付与されない
モデル	• 学習によって得られるパラメータ群や関数，座標群などをまとめたもの • とくに，テストデータを分類するためのモデルは，分類器と呼ばれる • モデルは，AI のアルゴリズムによって異なる
ローカルモデル	• エッジ上で学習して得られたモデル
クラス， 識別クラス	• 分類器が出力する，分類の集合全体，または，分類の集合の一つ
サンプルデータ	• IoT デバイスが収集したデータ • データには，画像やセンサデータなどがある • 訓練データの候補となる
パーソナルデータ	• 個人の属性情報，移動・行動・購買履歴，ウェアラブル機器から収集された個人情報 • 特定の個人を識別できないように加工された人流情報，商品情報等も含む • 個人情報に加え，個人情報との境界が曖昧なものを含む，個人と関係性が見出される広範囲の情報を指す

個人情報	• 生存する個人に関する情報 • 氏名，生年月日，その他の記述等により特定の個人を識別することができるもの • 他の情報と容易に照合することができ，それにより特定の個人を識別することができるものを含む（個人識別符号も対象）
個人データ	• 個人情報データベース等を構成する個人情報 • 個人情報データベース等とは、一定の規則に従って整理され，容易に検索可能な状態になっているもの ➢ メーラーのアドレス帳，ID と個人情報を関連付けて保管されているデータベースなど

参考文献

[1] Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K. and Zhang, J.：Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing, Proc. IEEE, Vol. 107, No. 8, pp. 1738-1762 DOI: 10.1109/JPROC.2019.2918951 (2019).

[2] Karim , A.：Mobile Computing Opportunities, Challenges and Technology Drivers (online), IEEE DAC 2014 Keynote, available from 〈http://www2.dac.com/events/videoarchive.aspx?confid=170&filter=keynote&id=170-103--0&#video〉 (accessed 2021-09-22).

[3] Karim, A.：Trends, Opportunities and Challenges Driving Architecture and Design of Next Generation Mobile Computing and IoT Devices, MIT MTL Seminar 2015, available from 〈https://www.mtl.mit.edu/seminars/trends-opportunities-and-challenges-driving-architecture-and-design-next-generation-mobile〉 (accessed 2021-09-22)

[4] 株式会社アイスマイリー：AI 画像認識システムで工場や倉庫の検品業務を効率化，株式会社アイスマイリー(オンライン)，入手先 〈https://aismiley.co.jp/ai_news/manufacturing-industry-inspection-image-recognition/〉(参照 2021-12-15).

[5] AUDI AG : Audi optimizes quality inspections in the press shop with artificial intelligence，AUDI AG(online)，available from 〈https://www.audi-mediacenter.com/en/press-releases/audi-optimizes-quality-inspections-in-the-press-shop-with-artificial-intelligence-10847〉 (accessed 2021-12-16).

[6] SKYDISC : AI 外観検査機，SKYDISC(オンライン)， 入手先 〈https://skydisc.jp/visual_inspection/〉(参照 2021-12-16).

[7] 三菱電機：「雪見だいふく」工場で進むスマート化、“完全自律運転” を目指すその第一歩，三菱電機(オンライン)，入手先 〈https://www.mitsubishielectric.co.jp/fa/compass/casestudies/custom/ccase43/index.html〉(参照 2021-12-16).

[8] NEC：エッジコンピューティングのソリューション事例 3. 画像解析技術を活用した人物行動分析サービス，NEC(オンライン)，入手先〈https://jpn.nec.com/techrep/journal/g17/n01/170106.html〉(参照 2021-12-15).

[9] OKI : AI エッジコンピューティング 行動解析 AI プラットフォーム構築サービス，OKI(オンライン)，入手先〈https://www.oki.com/jp/AIedge/partner/intro/granvalley.html l〉(参照 2021-12-15).

[10] ITmedia：自動運転車をファーストクラスに、NVIDIA が最新プラットフォームを発表，ITmedia(オンライン)， 入手先 〈https://monoist.itmedia.co.jp/mn/articles/2111/11/news052.html〉(参照 2021-12-16).

[11] 自動運転 LAB：自動運転に必須の「エッジ AI」とは？，自動運転 LAB(オンライン)，入手先 〈https://jidounten-lab.com/u_edgeai-autonomous〉(参照 2021-12-16).
[12] amnimo：注目が集まるエッジ AI とは？クラウド AI との違いや活用事例， amnimo (オンライン)，入手先 〈https://amnimo.com/column/007/〉(参照 2021-12-16).
[13] 株式会社 日立ソリューションズ・クリエイト：エッジ AI とは？ 人工知能が産業にもたらす可能性, 株式会社 日立ソリューションズ・クリエイト(オンライン)， 入手先 〈https://www.hitachi-solutions-create.co.jp/column/technology/edge-ai.html〉(参照 2021-12-16).
[14] パソナテック：エッジ AI とは？クラウド AI との違いや利用シーンについても解説，パソナテック(オンライン)，入手先 〈https://www.pasonatech.co.jp/workstyle/column/detail.html?p=2376〉(参照 2021-12-16).
[15] ITmedia：なぜクラウドではダメなのか？ いま「エッジ AI」が注目されるワケ， ITmedia (オンライン)，入手先 〈https://www.itmedia.co.jp/news/articles/2003/19/news050.html〉(参照 2021-12-16).
[16] サイバートラスト株式会社：急成長するエッジ AI 環境に必要なセキュリティ機能をパッケージ化し「EM+PLS」を提供開始， サイバートラスト株式会社(オンライン)，入手先 〈https://www.cybertrust.co.jp/pressrelease/2021/1104-secure-edge-empls.html〉(参照 2021-12-16).
[17] 株式会社ヘッドウォータース：エッジ AI とは， 株式会社ヘッドウォータース(オンライン)，入手先 〈https://www.headwaters.co.jp/service/ai/edge.html〉(参照 2021-12-16).
[18] Geekroid：エッジ AI とは何か？クラウド AI との違いから特徴まで紹介， Geekroid (オンライン)，入手先 〈https://mynavi-agent.jp/it/geekroid/2020/11/post-305.html〉(参照 2021-12-16).
[19] 日経 BP：AI 活用を阻むデータ不足、それを解決する「秘密分散学習」の仕組み， 日経 BP (オンライン)，入手先 〈https://active.nikkeibp.co.jp/atcl/act/19/00255/011400002/〉(参照 2021-12-16).
[20] ITmedia：教師データが足りないと「異常予測」は難しい、ならば「異常検知」から始めよう，ITmedia(オンライン)， 入手先 〈https://monoist.itmedia.co.jp/mn/articles/1912/04/news020.html〉(参照 2021-12-16).
[21] 富士電機：予知保全システムの導入とその課題解決について， 富士電機(オンライン)，入手先 〈https://www.fujielectric.co.jp/products/column/fa/fa_05.html〉(参照 2021-12-16).
[22] TechCrunch ：“ディープラーニングで解決できない課題”に独自 AI で挑むハカルスが 1 億円を調達，TechCrunch(オンライン)， 入手先 〈https://jp.techcrunch.com/2018/10/05/hacarus-raises-100m-yen/?guccounter=1&guce_referrer=aHR0cHM6Ly93

d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAJYllXF5gEhBNhAMSKcAcJvmpMRgYkAHMqR6PoCOdGWGfWUiiiV9gItgmwcA4ZmA9Dbc1swNCfQukpjU-RFjp6L_OLcPE2maUk7ow_i5JXBltZOiJeqTRH4VdCbzDrmPPKOoiwIZpLaef9GR0OmR-fXPETaZX6yrmWapsfLZlnsx〉(参照 2021-12-16).

[23] IPA：1.5 ビッグデータ時代の知識処理，IPA(オンライン)， 入手先 〈https://www.ipa.go.jp/files/000082706.pdf〉(参照 2021-12-16).

[24] IIJ：世界で広がる「プライバシー保護規制」私たちが取るべき対策は？ ，IIJ(オンライン)，入手先 〈https://www.iij.ad.jp/interview/06.html〉(参照 2021-12-16).

[25] 寺田眞治：個人情報保護関連の海外の法制度の概要，JIPDEC(オンライン)，入手先 〈https://www.jipdec.or.jp/archives/publications/J0005156.pdf〉(参照 2021-12-16).

[26] JIPDEC：OECD8 原則 ，JIPDEC (オンライン)，入手先 〈https://www.jipdec.or.jp/library/word/csm0kn0000000egq.html〉(参照 2021-12-16).

[27] 高橋克巳：公的統計の高度な二次的利用のための秘密計算技術の応用の研究，総務省統計局(オンライン)， 入手先 〈https://www.stat.go.jp/training/2kenkyu/research/pdf/resear34.pdf〉(参照 2021-12-16).

[28] 花岡悟一郎，他：プライバシー保護データ解析技術の社会実装，JST(オンライン)， 入手先〈https://www.jst.go.jp/kisoken/crest/research/activity/1111094/ai_sympo3/pdf/crestai_10.pdf〉(参照 2021-12-16).

[29] NTT：秘密計算，NTT(オンライン)， 入手先〈https://www.rd.ntt/sil/project/sc/secure_computation.html〉(参照 2021-12-16).

[30] JEITA スマートホーム部会：JEITA スマートホームデータカタログ，JEITA スマートホーム部会(オンライン)，入手先 〈https://www.jeita.or.jp/japanese/pickup/category/2020/smarthome.html〉(参照 2021-12-16).

[31] ITU：ITU-T Y.4000/Y.2060 (06/2012)，ITU (オンライン),入手先 〈https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-Y.2060-201206-I!!PDF-E&type=items〉(参照 2021-11-19).

[32] AWS：IoT とは？，AWS(オンライン),入手先 〈https://aws.amazon.com/jp/iot/what-is-the-internet-of-things/〉(参照 2021-11-19).

[33] Oracle：What is IoT?，Oracle(オンライン),入手先 〈https://www.oracle.com/internet-of-things/what-is-iot/〉(参照 2021-11-19).

[34] Trend Micro：Internet of Things (IoT)，Trend Micro (オンライン),入手先 〈https://www.trendmicro.com/vinfo/us/security/definition/internet-of-things〉(参照 2021-11-21).

[35] 内閣府：Society 5.0，内閣府(オンライン)，入手先 〈https://www8.cao.go.jp/cstp/society5_0/〉(参照 2021-11-22).

[36] 経済産業省：Connected Industries，経済産業省(オンライン)，入手先 〈https://www.

meti.go.jp/policy/mono_info_service/connected_industries/index.html〉(参照 2021-11-22).

[37] 株式会社 DTS : コネクテッド インダストリーズ (Connected Industries) とは, 株式会社 DTS(オンライン), 入手先 〈https://dts-bigdata.jp/media/useful/a36〉(参照 2021-11-22).

[38] 一般社団法人データ社会推進協議会 : 団体概要, 一般社団法人データ社会推進協議会(オンライン),入手先 〈https://data-society-alliance.org/about/association/〉(参照 2021-11-22).

[39] 経済産業省 :「Connected Industries」関連政策の進捗等について, 経済産業省(オンライン), 入手先 〈https://www.meti.go.jp/policy/mono_info_service/connected_industries/pdf/201806_progress.pdf〉(参照 2021-11-22).

[40] 経済産業省 商務情報政策局 サイバーセキュリティ課 : サイバー・フィジカル・セキュリティ対策フレームワーク Version 1.0, 経済産業省(オンライン), 入手先 〈https://www.meti.go.jp/policy/netsecurity/wg1/CPSF_ver1.0.pdf〉(参照 2021-11-22).

[41] ISO : ISO/IEC 27001 INFORMATION SECURITY MANAGEMENT, ISO(online), available from 〈https://www.iso.org/isoiec-27001-information-security.html〉 (accessed 2021-11-25).

[42] IEC : Understanding IEC 62443, IEC(online), available from 〈https://www.iec.ch/blog/understanding-iec-62443〉 (accessed 2021-11-25).

[43] NIST : NISTIR 7628 Revision 1, NIST(online), available from 〈https://nvlpubs.nist.gov/nistpubs/ir/2014/NIST.IR.7628r1.pdf〉 (accessed 2021-11-25).

[44] (株)日立製作所 : 制御システムセキュリティの標準化動向～IEC 62443 の最新状況と認証制度の紹介～, (株)日立製作所 (オンライン), 入手先〈https://www.jpcert.or.jp/present/2020/ICSR2020_04_HITACHI.pdf〉(参照 2021-11-26).

[45] 八山幸司 : 米国における電力インフラと IT をめぐる動向, JETRO(オンライン), 入手先 〈https://www.jetro.go.jp/ext_images/_Reports/02/2016/c370a43cce9fa649/rpNY_ITinfra201606.pdf〉(参照 2021-12-20).

[46] 大崎人士, 坂根広史, 半田剣一 : スマートメーターシステムに係るセキュリティ基準と評価の現状, 安全工学, Vol.54, No.6, pp. 442-451(2015).

[47] IPA : 米国電力関係の基準の概要, IPA (オンライン), 入手先 〈https://www.ipa.go.jp/files/000077731.pdf〉(参照 2021-12-20).

[48] NIST : About NIST, NIST(online), available from 〈https://www.nist.gov/about-nist〉 (accessed 2021-11-25).

[49] ENISA : About ENISA - The European Union Agency for Cybersecurity, ENISA (online), available from 〈https://www.enisa.europa.eu/about-enisa〉 (accessed 2021-11-25).

[50] NIST : NIST Special Publication 800-53 Revision 5, NIST(online), available from 〈https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r5.pdf〉 (accessed 2021-11-25).
[51] NIST : NIST Special Publication 800-53 Revision 5, IPA（オンライン), 入手先 〈https://www.ipa.go.jp/files/000092657.pdf〉(参照 2021-11-25).
[52] ENISA : Guidelines for Securing the Internet of Things, ENISA (online), available from 〈https://www.enisa.europa.eu/publications/guidelines-for-securing-the-internet-of-things〉 (accessed 2021-11-25).
[53] GSMA : About Us, GSMA(online), available from 〈https://www.gsma.com/aboutus/〉 (accessed 2021-11-25).
[54] GSMA : GSMA IoT Security Guidelines and Assessment, GSMA(online), available from 〈https://www.gsma.com/iot/iot-security/iot-security-guidelines/〉 (accessed 2021-11-25).
[55] GSMA : IoT セキュリティ・ガイドライン概要説明書バージョン 2.0, GSMA(オンライン), 入手先 〈https://www.gsma.com/iot/wp-content/uploads/2017/11/CLP.11-v2.0_JPN.pdf?utm_source=Website&utm_medium=Main_Security_Download&utm_campaign=IoTSG_Overview_Japanese_CLP.11-v2.0_JPN.pdf&utm_term=IoT+Security&utm_content=Downloads〉(参照 2021-12-20).
[56] IIC : About Us, IIC(online), available from 〈https://www.iiconsortium.org/about-us.htm〉 (accessed 2021-11-26).
[57] IIC : Industrial Internet Security Framework Technical Document, IIC(online), available from 〈https://www.iiconsortium.org/pdf/IIC_PUB_G4_V1.00_PB-3.pdf〉 (accessed 2021-11-26).
[58] 中尾昌善 :「つながる世界の開発指針」の 産業展開に向けて, IPA(オンライン), 入手先 〈https://www.ipa.go.jp/files/000057704.pdf〉(参照 2021-12-20).
[59] e-Gov ポータル : 個人情報の保護に関する法律, e-Gov ポータル(オンライン), 入手先 〈https://elaws.e-gov.go.jp/document?lawid=415AC0000000057〉(参照 2021-12-20).
[60] 佐藤一郎 : IoT 時代のパーソナルデータの保護と利活用～個人情報保護法改正とその影響～, IPA(オンライン), 入手先 〈https://www.ipa.go.jp/files/000046424.pdf〉(参照 2021-12-20).
[61] 個人情報保護委員会 : GDPR（General Data Protection Regulation : 一般データ保護規則), 個人情報保護委員会(オンライン), 入手先〈https://www.ppc.go.jp/enforcement/infoprovision/laws/GDPR/〉(参照 2021-10-23).
[62] DLA Piper : DLA Piper GDPR fines and data breach survey: January 2021, DLA Piper (online), available from〈https://blogs.dlapiper.com/privacymatters/dla-

piper-gdpr-fines-and-data-breach-survey-january-2021/〉(accessed 2021-10-23).
[63] 個人情報保護委員会：California Consumer Privacy Act of 2018 米国「カリフォルニア州消費者プライバシー法　2018 年」, 個人情報保護委員会(オンライン), 入手先〈https://www.ppc.go.jp/enforcement/infoprovision/laws/CCPA/〉(参照 2021-12-20).
[64] JETRO：米国連邦データプライバシー法案の概要（2021 年 6 月）, JETRO(オンライン）入手先〈https://www.jetro.go.jp/world/reports/2021/01/7f744522a1ddc8eb.html〉(参照 2021-12-20).
[65] ソフトバンク：中国サイバーセキュリティ法とは, ソフトバンク(オンライン), 入手先〈https://www.softbank.jp/biz/services/platform/alibabacloud/solution/china/cybersecurity/〉(参照 2021-12-20).
[66] 総務省：令和 2 年版　情報通信白書,総務省(オンライン),入手先 〈https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r02/pdf/02honpen.pdf〉(参照 2021-09-21).
[67] 総務省：令和元年版　情報通信白書,総務省(オンライン),入手先 〈https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r01/pdf/01honpen.pdf〉(参照 2021-12-22).
[68] ビジネス+IT：AI の「画像・音声・動画認識」のカラクリ、機械学習のプロセスを解説, ビジネス+IT (オンライン), 入手先〈https://www.sbbit.jp/article/cont1/66309〉(参照 2021-12-22).
[69] 矢野正基, 大賀隆裕, 大西正輝：深層学習を用いた画像識別タスクの精度向上テクニック, 電子情報通信学会論文誌 D, Vol. J102-D, No. 2, pp.34-52 (2019).
[70] Cheng, T-Y.：Building a One-shot Learning Network with PyTorch, Towards Data Science(online), available from〈https://towardsdatascience.com/building-a-one-shot-learning-network-with-pytorch-d1c3a5fafa4a〉(accessed 2021-12-22).
[71] Song, H., Torres, M. T., Ozcan, E., and Triguero, I. ：L2AE-D: Learning to Aggregate Embeddings for Few-shot Learning with Meta-level Dropout, arXiv preprint arXiv:1904.04339v1 (2019).
[72] 梅田弘之：水増しと転移学習 (Vol.7), AISIA(オンライン), 入手先 〈https://products.sint.co.jp/aisia/blog/vol1-7〉(参照 2021-12-22).
[73] 佐藤敦：少量データ向け深層学習技術, NEC 技報, Vol.72, No.1, pp.113-117 (2019).
[74] Wang, Y., Yao, Q. Kwok, J., and Ni, L. M.：Generalizing from a Few Examples: A Survey on Few-Shot Learning, arXiv preprint arXiv:1904.05046v3 (2020).
[75] 人工知能学会：AI マップ β 2.0 （2020 年 6 月版）, 人工知能学会(オンライン), 入手先〈https://www.ai-gakkai.or.jp/resource/aimap/〉(参照 2021-12-22).
[76] 土方 細秩子：AI・IoT が“米国の原子力発電”レベルの電力を食い尽くす, ビジネス+IT(オンライン), 入手先〈https://www.sbbit.jp/article/cont1/36775〉(参照 2021-10-22).

[77] 織田 浩一：クラウド利用でのデータ遅延を救うエッジ AI～Edge Computing World レポート～，wisdom(オンライン)，入手先〈https://wisdom.nec.com/ja/series/orita/2020112501/index.html〉(参照 2021-10-22).

[78] Microsoft Azure : Azure IoT Edge とは，Microsoft (オンライン),入手先〈https://docs.microsoft.com/ja-jp/azure/iot-edge/about-iot-edge〉(参照 2018-12-14).

[79] Amazon Web Services : AWS IoT Greengrass とは，Amazon Web Services (オンライン),入手先〈https://docs.aws.amazon.com/ja_jp/greengrass/latest/developerguide/what-is-gg.html〉(参照 2018-12-14).

[80] Google ： Edge TPU，Google (オンライン)，入手先〈https://cloud.google.com/edge-tpu〉(参照 2021-09-22).

[81] Chen, Y., Qin, X., Wang, J., Yu, C. and Gao, W.：FedHealth: A Federated Transfer Learning Framework for Wearable Healthcare, IEEE Intelligent Systems, doi: 10.1109/MIS.2020.2988604 (2020).

[82] Kohavi, R.：A study of cross-validation and bootstrap for accuracy estimation and model selection, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Vol.2, No.12, pp.1137-1143 (1995).

[83] Chang, J., Luo, Y., and Su, K.：GPSM: a Generalized Probabilistic Semantic Model for ambiguity resolution, In Proceedings of the 30th Annual Meeting on Association For Computational Linguistics, pp.177-184 (1992).

[84] Devijver, P. A., and Kittler, J.：Pattern Recognition: A Statistical Approach (1982).

[85] 杉山 将, 山田 誠, ドゥ・プレシ マーティヌス・クリストフェル, リウ ソン：非定常環境下での学習：共変量シフト適応，クラスバランス変化適応，変化検知，日本統計学会誌，Vol.44，No.1，pp.113-136(2014).

[86] Vapnik, V. and Lerner, A.：Pattern recognition using generalized portrait method, Automation and Remote Control, Vol.24, pp. 774-780 (1963).

[87] Tetko, I. V., Livingstone, D. J., Luik, A. I. :. Neural network studies. 1. Comparison of Overfitting and Overtraining, . J. Chem. Inf. Comput. Sci., Vol.35, No.5, pp.826–833 (1995).

[88] 神嶌 敏弘：転移学習，人工知能学会誌，Vol.25，No.4，pp.572-580 (2010).

[89] 中山英樹：深層畳み込みニューラルネットワークによる画像特徴抽出と転移学習，電子情報通信学会技術研究報告，Vol.115，No.146，pp.55-59(2015).

[90] NEC セキュリティ研究所：NEC の秘密計算技術のご紹介，NEC (オンライン)，入手先〈https://jpn.nec.com/rd/technologies/201805/pdf/mpc_introduction.pdf〉(参照 2021-09-24).

[91] 菊池 亮, 五十嵐 大：秘密計算の発展－データを隠しつつ計算する仕組みとその発展－，

電子情報通信学会 基礎・境界ソサエティ 解説論文, Vol.12, No.1, pp.12-20 (2018).
[92] 岡村利彦, 寺西勇:FinTech のセキュリティ強化に貢献するマルチパーティ計算技術, NEC 技報, Vol.69, No.2, pp.50-54(2017).
[93] Acompany :【超入門】秘密計算って何？図で概要をわかりやすく解説！, Acompany (オンライン), 入手先〈https://acompany.tech/blog/secure_computing/〉(参照 2021-12-23).
[94] 菊池 亮 : 安全なデータ活用を実現する秘密計算技術 : 4. Garbled circuit を用いた秘密計算と混合的構成, 情報処理, Vol.59, No.10, pp.893-897(2018).
[95] 山本 : 分かった気になる自然言語処理概観 その 1, YDC(オンライン), 入手先〈https://www.ydc.co.jp/column/0002/20190913.html〉(参照 2020-12-24).
[96] 松井くにお, 難波功, 井形伸之 : 全文検索エンジン (情報検索の新潮流), 情報の科学と技術, Vol.50, No.1, pp. 9-13 (2000).
[97] Slack Technologies : Slack, Slack (オンライン), 入手先〈https://slack.com/intl/ja-jp/〉(参照 2020-09-02).
[98] Google : Google ドライブ, Google (オンライン), 入手先〈https://www.google.com/intl/ja_ALL/drive/〉(参照 2020-09-02).
[99] O'Brien, J. and Marakas, G.M. : Management Information Systems, pp. 185-189, Irwin Professional Pub (2008).
[100] Bellare, M., Boldyreva, A. and O'Neill, A.: Deterministic and Efficiently Searchable Encryption, Proc. Advances in Cryptology - CRYPTO 2007, LNCS 4622, pp.535-552 (2007).
[101] Song, D.X., Wagner, D. and Perrig, A.: Practical Techniques for Searches on Encrypted Data, Proc. IEEE Symposium on Security and Privacy, IEEE Computer Society, pp.44-55 (2000).
[102] Boneh, D., Di Crescenzo, G., Ostrovsky, R. and Persiano, G.: Public Key Encryption with Keyword Search, Proc. Advances in Cryptology - EUROCRYPT 2004, LNCS 3027, pp.506-522 (2004).
[103] Hirano, T., Kawai. Y. and Koseki, Y. : DBMS-Friendly Searchable Symmetric Encryption: Constructing Index Generation Suitable for Database Management Systems, Proc. IEEE Conference on Dependable and Secure Computing (DSC 2021), IEEE, pp. 1-8 (2021).
[104] 伊藤隆, 平野貴人, 森拓海, 服部充洋 : 秘匿検索の頻度分析対策としての複数 DB 活用について, 情報処理学会論文誌, Vol.60, No.1,pp.240-249 (2019).
[105] N. Pramanick and S. T. Ali, "A comparative survey of searchable encryption schemes," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-5, 2017.

[106] Y.Zhou, N.Li, Y.Tian, D.An, and L.Wang, "Public Key Encryption with Keyword Search in Cloud: A Survey," Entropy 2020, Vol.22, No.4, 2020.

[107] Li, J., Wang, Q., Wang, C., Cao, N., Ren, Kui., and Lou, W.：Fuzzy Keyword Search over Encrypted Data in Cloud Computing, Proc. IEEE Conference on Computer Communications, IEEE, pp. 441–445 (2010).

[108] Levenshtein, V.：Binary codes capable of correcting spurious insertions and deletions of ones, Problems of Information Transmission, Vol.1, No.1, pp.8–17(1965).

[109] 尾形わかは，金岡晃，松尾真一郎：実用的な多機能検索可能暗号方式～身も蓋もない方式を考えてみた～，暗号と情報セキュリティシンポジウム（SCIS2015）予稿集，3F1-3 (2015).

[110] Krizhevsky, A., Sutskever, I., and Hinton, G. E.：ImageNet classification with deep convolutional neural networks, Communications of the ACM, Vol.60, No.6, pp 84–90 (2017).

[111] Simonyan, K., Zisserman, A.：Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv preprint arXiv:1409.1556 (2014).

[112] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A.：Going Deeper with Convolutions, arXiv preprint arXiv: 1409.4842 (2014).

[113] He, K., Zhang, X., Ren, S., and Sun, J.：Deep Residual Learning for Image Recognition, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, doi: 10.1109/CVPR.2016.90 (2016).

[114] Rublee, E., Rabaud, V., Konolige K., and Bradski, G.：ORB: An efficient alternative to SIFT or SURF, 2011 International Conference on Computer Vision, pp. 2564-2571, doi: 10.1109/ICCV.2011.6126544 (2011).

[115] Dalal, N. and Triggs, B.：Histograms of Oriented Gradients for Human Detection, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 1, pp. 886-893, doi: 10.1109/CVPR.2005.177 (2005).

[116] Lowe, D.G.：Distinctive Image Features from Scale-Invariant Keypoints, Int. J.Comput. Vision, Vol.60, No.2, pp.91–110 (2004).

[117] Alcantarilla, P., Nuevo, J., and Bartoli, A.：Fast explicit diffusion for accelerated features in nonlinear scale spaces, Trans. Pattern Anal. Machine Intell, Vol.34, No.7, pp.1281–1298, (2011).

[118] Bridle, J. S. and Brown, M. D.：An Experimental Automatic Word-Recognition System, JSRU Report No. 1003, Joint Speech Research Unit, Ruislip, E

ngland (1974).

[119] i Magazine : 類似画像検索の 3 つの手法と精度向上のテクニック, アイマガジン株式会社(オンライン), 入手先〈https://www.imagazine.co.jp/%E9%A1%9E%E4%BC%BC%E7%94%BB%E5%83%8F%E6%A4%9C%E7%B4%A2%E3%81%AE3%E3%81%A4%E3%81%AE%E6%89%8B%E6%B3%95%E3%81%A8%E7%B2%BE%E5%BA%A6%E5%90%91%E4%B8%8A%E3%81%AE%E3%83%86%E3%82%AF%E3%83%8B%E3%83%83%E3%82%AF/〉(参照 2021-09-25).

[120] Shorten, C., Khoshgoftaar, T.M. : A survey on Image Data Augmentation for Deep Learning, Journal of Big Data Vol.6, No.60 (2019).

[121] Thompson, N.C., Greenewald, K., Lee, K., and Manso, G.F. : The Computational Limits of Deep Learning, arXiv preprint arXiv:2007.05558 (2020).

[122] 喜多 泰代 : 二次元濃度ヒストグラムを用いた画像間変化抽出, 電子情報通信学会論文誌, Vol.J90-D, No.8, pp.1957-1965 (2007).

[123] 大美 英一, 斉藤 文彦 : 最頻度エッジ方向に基づく画像の傾き検出と視覚的傾斜角との関係, 電気学会論文誌C（電子・情報・システム部門誌）, Vol.124, No.1, pp.112-118, https://doi.org/10.1541/ieejeiss.124.112 (2004).

[124] Csurka, G., Dance, C.R., Fan, L., Willamowski, J. and Bray, C.: Visual Categorization with Bags of Keypoints, ECCV International Workshop on Statistical Learning in Computer Vision, pp. 1-22 (2004).

[125] Thorndike, R.L. : Who belongs in the family?, Psychometrika, Vol.18,No.4, pp.267-276 (1953).

[126] Peter, J. Rousseeuw : Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, Vol. 20, pp. 53-65 (1987).

[127] Rosten, E. and Drummond, T. : Machine learning for high speed corner detection, European Conference on Computer Vision, Vol.3951 (2006).

[128] Calonder, M., Lepetit, V., Strecha, C., and Fua, P. : Brief: Binary robust independent elementary features, European Conference on Computer Vision, Vol.6341 (2010).

[129] Min, K., Yang, L., Wright, J., Wu, L., Hua, X.-S., and Ma, Y. : Compact projection: Simple and efficient near neighbor search with practical memory requirements, 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3477–3484 (2010).

[130] Wang, J., Kumar, S., and Chang, S.-F. : Sequential projection learning for hashing with compact codes, Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 1127-1134 (2010).

[131] Achlioptas, D. : Database-friendly random projections, Proceedings of the t wentieth ACM SIGMODSIGACT-SIGART symposium on Principles of database systems, pp. 274–281 (2001).

[132] Li, P., Hastie, T. J. and Church, K. W. : Very sparse random projections, Proceedings of the international conference on Knowledge Discovery and Data mining, pp. 287–296 (2006).

[133] 藤吉弘亘,安倍 満 : 局所勾配特徴抽出技術:-SIFT 以降のアプローチ-, 精密工学会誌, Vol. 77, No. 12, pp. 1109-1116 (2011).

[134] Bay H., Tuytelaars T., and Van Gool L. : Speeded-Up Robust Features (S URF), Computer Vision and Image Understanding, Vol 110, No. 3, pp. 346-35 9 (2008).

[135] Alcantarilla, P. F., Bartoli, A., and Davison, A. J. : KAZE features, Europ ean Conference on Computer Vision (ECCV), Vol.7577, pp. 214–227 (2012).

[136] Alcantarilla, P. F., Bartoli, A., and Davison, A. J. : KAZE Features, Euro pean Conference on Computer Vision (ECCV), Vol.7577 (2012).

[137] Aghabozorgi, S., Shirkhorshidi, A.S. and Wah, T.J.: Time-series clustering. A decade review, Information systems, Vol. 53, pp.16-38 DOI:10.1016/j.is.2015. 04.007 (2015).

[138] Vlachos, M., Kollios, G. and Gunopulos, D. : Discovering Similar Multidim ensional Trajectories, Proceedings of 18th International Conference on Data E ngineering, pp.673-684(2002).

[139] Latecki, L.J., Megalooikonomou, V., Wang, Q., Lakaemper, R., Ratanamah atana, C. and Keogh, E. : Elastic Partial Matching of Time series, Knowledge Discovery in Databases : PKDD2005, pp.577-584(2005).

[140] Keogh, E., Lonardi, S., Ratanamahatana, C., Wei, L., Lee, S.H. and Hand ley, J. : Compression-based data mining of sequential data, Data Mining and Knowledge Discovery, Vol. 14, pp.99-129(2007).

[141] Kullback, S. and Leibler, R.A. : On Information and Sufficiency, Annals of Mathematical Statistics, Vol. 22, No. 1, pp.79-86 DOI:10.1214/aoms/117772969 4 (1951).

[142] Jaccard, P. : Distribution de la Flore Alpine dans le Bassin des Dranses et dans quelques régions voisines, Bulletin de la Société Vaudoise des Science s Naturelles, Vol. 37, No. 140, pp.241-272(1901).

[143] Sørensen, T. : A method of establishing groups of equal amplitude in plan t sociology based on similarity of species and its application to analyses of th e vegetation on Danish commons, Kongelige Danske Videnskabernes Selskab,

Vol. 5, No. 4, pp.1-34(1948).

[144] Dice, Lee R.：Measures of the Amount of Ecologic Association Between Species, Ecology, Vol. 26, No. 3, pp.297-302 DOI:10.2307/1932409 (1945).

[145] Szymkiewicz, D.：Une contribution statistique a la geographie floristique, Acta Societatis Botanicorum Poloniae, Vol.34, No. 3, pp.249-265(1934).

[146] Smyth, P.：Clustering Sequences with Hidden Markov Models, Proceedings of the 9th International Conference on Neural Information Processing Systems (NIPS'96), pp. 648-654(1996).

[147] Kalpakis, K., Gada, D. and Puttagunta, V.：Distance Measures for Effective Clustering of ARIMA Time-Series, Proceedings 2001 IEEE International Conference on Data Mining ,pp. 273-280(2001).

[148] Bellman, R. and Kalaba, R.E.：On adaptive control processes, IRE Transactions on Automatic Control, Vol. 4, No. 2, pp. 1-9 (1959).

[149] Senin, P.：Dynamic Time Warping Algorithm Review, Information and Computer. Science Department University of Hawaii at Manoa Honolulu, USA, Vol. 855, pp. 1-23 (2008).

[150] Lou, Y., Ao, H., and Dong, Y. : Improvement of Dynamic Time Warping (DTW) Algorithm, 2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), pp. 384-387 (2015).

[151] Silva, D.F., Giusti, R., Keogh, E. et al. : Speeding up similarity search under dynamic time warping by pruning unpromising alignments, Data Mining and Knowledge Discovery,Vol.32,No.2,pp.988-1016(2018).

[152] 井本和範, 中居友弘 : 深層学習を用いて自動化した半導体製造プロセスの欠陥分類システム, 東芝レビュー, Vol.74, No.5, pp.13-16(2019).

[153] NTT テクノクロス : クラウドデータ暗号化 TrustBind/Secure Gateway, NTT テクノクロス(オンライン), 入手先〈https://www.ntt-tx.co.jp/products/trustbind/sgw/〉(参照 2021-10-24).

[154] 日立製作所 : 秘匿情報管理サービス 匿名バンク, 日立製作所(オンライン), 入手先〈https://www.hitachi.co.jp/Prod/comp/app/tokumei/index.html〉(参照 2021-10-24).

[155] 株式会社富士通研究所 : 暗号化した機密情報の類推を防止する技術を開発, 株式会社富士通研究所(オンライン), 入手先〈https://pr.fujitsu.com/jp/news/2019/10/16.html〉(参照 2021-10-24).

[156] V7：An Introductory Guide to Quality Training Data for Machine Learning, V7 (online), available from 〈https://www.v7labs.com/blog/quality-training-data-for-machine-learning-guide#how-much-data〉 (accessed 2021-10-14).

[157] Samarati, P. and Sweeney, L.：Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression, SRI International (1998).

[158] 四方順司, 渡邉洋平：情報理論的暗号技術について，情報処理，Vol.55, No.3, pp. 260-267（2014）.

[159] 青空文庫，入手先〈https://www.aozora.gr.jp/〉(参照 2020-01-20).

[160] 打田 智子：Janome v0.3 documentation(ja)，入手先〈https://mocobeta.github.io/janome/〉(参照 2020-01-20).

[161] MeCab: Yet Another Part-of-Speech and Morphological Analyzer，入手先〈http://taku910.github.io/mecab/〉(参照 2020-01-20).

[162] Gensim：models.word2vec – Word2vec embeddings, Radim Řehůřek，入手先〈https://radimrehurek.com/gensim/models/word2vec.html〉(参照 2020-01-20).

[163] Kyubyong / wordvectors：Pre-trained models , github，入手先〈https://github.com/Kyubyong/wordvectors〉(参照 2020-01-20).

[164] Docker：citusdata/citus, dockerhub，入手先〈https://hub.docker.com/r/citusdata/citus〉(参照 2020-01-06).

[165] Citus Data：Citus，入手先〈https://github.com/citusdata/docker/tree/v7.0.3〉(参照 2020-01-06).

[166] 金明哲：統計的テキスト解析(5)～統計法則と指標～，同志社大学(オンライン)，入手先〈https://mjin.doshisha.ac.jp/R/60/60.html〉(参照 2021-12-24).

[167] 横川壽彦：ジップの法則(Zipf's Law)，日本ファジィ学会誌，Vol.14，No. 6, p. 604(2002).

[168] 森郁海，伊藤岳広，中村嘉隆，稲村浩：AI/IoT ソリューションへの転移学習適用における学習モデル検索システムの提案と評価，第 27 回マルチメディア通信と分散処理ワークショップ論文集，pp. 141-148 (2019).

[169] ImageNet：API documentation , ImageNet (online), available from〈http://image-net.org/download-API〉(accessed 2020-03-24).

[170] OpenCV：Changing Colorspaces , OpenCV (online), available from〈https://docs.opencv.org/4.1.1/df/d9d/tutorial_py_colorspaces.html〉(accessed 2020-03-24).

[171] SciPy：Kernel density estimation , scipy.stats (online), available from〈https://docs.scipy.org/doc/scipy/reference/tutorial/stats.html〉(accessed 2020-03-24).

[172] SciPy：Peak finding , scipy. signal (online), available from〈https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.argrelmin.html〉(accessed 2020-03-24).

[173] Pierre Rouanet：dtw 1.4.0 , pypi (online), available from〈https://pypi.org/project/dtw/〉(accessed 2020-03-24).

謝辞

本研究に取り組むにあたり，最後まで親身なご指導と励ましを賜りました指導教員である公立はこだて未来大学システム情報科学研究科究科長稲村浩教授に心より感謝いたします．また，研究遂行にあたり有益なご討論ご助言を戴いた京都橘大学工学部情報工学科中村嘉隆准教授，および，公立はこだて未来大学の各先生方に感謝いたします．特に議論と論文執筆にあたり貴重な知見と綿密なレビューを頂きました高橋修名誉教授，藤野雄一教授，白石陽教授，新美礼彦教授に感謝いたします．

研究の遂行，実験，論文発表にあたり，三菱電機株式会社の松田哲史さん，平野貴人さんをはじめ，協力いただいた方々に，この場を借りて感謝の意を述べさせていただきます．

最後に，社会人大学院生としての会社業務と研究の両立を支えてくれた，家族に深く感謝いたします．

本研究に関する論文発表

論文誌(査読有)

[1] 森郁海, 中村嘉隆, 稲村浩: 極値分割と Dynamic Time Warping によるデータ類似度評価手法, 情報処理学会論文誌, Vol.62, No.2, pp.497-507, 2021 年 2 月.

国際会議(査読有)

[1] Ikumi Mori, Takato Hirano, Yoshitaka Nakamura, and Hiroshi Inamura: Determination of parameters balancing between security and search performance on searchable encryption, Proceedings of the 13th International Conference on Mobile Computing and Ubiquitous Networking (ICMU2021), November 2021, Tokyo, Japan.

国内口演(査読なし)

[1] 森郁海, 平野貴人, 中村嘉隆, 稲村浩: 確率的暗号ベースの検索可能暗号を用いた全文検索の高速化検討, 情報処理学会研究報告, Vol.2021-DPS-186, No.14, pp.1-7, 2021 年 3 月.

[2] 森郁海, 伊藤岳広, 中村嘉隆, 稲村浩: AI/IoT ソリューションへの転移学習適用における学習モデル検索システムの提案と評価, 第 27 回マルチメディア通信と分散処理ワークショップ (DPSWS2019) 論文集, pp.141-148, 2019 年 11 月.

関連した論文、特許

国内口演(査読なし)

[1] 森郁海，齊藤志保，折本拓真，伊藤岳広 : 分散エッジ環境における機械学習実現最適化の検討〜エッジ上で動作するアルゴリズム・オントロジーの決定と転移学習適用による最適化検討〜，情報処理学会研究報告，Vol. 2019-DPS-177，No. 2，pp. 1-8 (2019).

特許

[1] 発明者：板垣弦矢，森郁海，国際出願番号：PCT/JP2021/001949，発明の名称：情報処理装置，情報処理方法及び情報処理プログラム

[2] 発明者：森郁海，板垣弦矢，国際出願番号：PCT/JP2020/046743，発明の名称：類似度算出装置，類似度算出方法，及び，類似度算出プログラム

[3] 発明者：森郁海，国際出願番号：PCT/JP2019/040614，発明の名称：検索装置，検索方法，検索プログラム及び学習モデル検索システム

[4] 発明者：折本拓真，森郁海，国際出願番号：PCT/JP2019/019905，発明の名称：オントロジー生成システム，オントロジー生成方法およびオントロジー生成プログラム

共著論文

[1] 樫原渉，撫中達司，森郁海，板垣弦矢：オントロジーを用いたユーザ嗜好の多面的特徴抽出方法の提案， マルチメディア，分散協調とモバイルシンポジウム 2021 論文集， Vol. 2021， No. 1， pp. 205-213 (2021).

[2] 板垣弦矢，森郁海，撫中達司：ソリューションと機器推薦のためのオントロジースキーマ及び推薦方法の検討，人工知能学会研究会資料， Vol. 53， No. 1， pp.1-8 (2021).

[3] 板垣弦矢，撫中達司，齊藤志保，森郁海，折本拓真，伊藤岳広：ノード間類似度を用いたオントロジーの切り出し手法についての提案， 情報処理学会研究報告，Vol. 2019-CDS-26， No. 18， pp. 1-8 (2019).

[4] 板垣弦矢，撫中達司，齊藤志保，森郁海，折本拓真，伊藤岳広：フォグコンピューティング向け水平統合型 IoT プラットフォームの提案とその評価， 情報処理学会研究報告， Vol. 2019-CDS-24， No. 28， pp. 1-8 (2019).