情報の機密度推定による データ利便性とセキュリティのトレードオフ解消

江田 智尊[1,a)]　吉橋 仁[1]　横田 拓也[1]　清水 俊也[1]　樋口 裕二[1]　辻 健太郎[1]　兒島 尚[1]

概要：生成 AI の普及により，組織に蓄積されたデータをビジネスに利活用したいという要求が高まっている．組織には様々な機密度を持つ情報（機密情報，社外秘情報，公開情報）が混在するが，情報の機密度が適切に設定されていない場合，情報漏洩や情報取得機会の損失が発生することになる．例えばセキュリティ確保を重視してファイルの機密度が高めに設定された場合，情報漏洩リスクは下がるが活用できる情報が制限され，データを利用する生成 AI アプリケーションの利便性があがらない．このような「データ利便性とセキュリティ」のトレードオフを解消するアプローチを本稿では提案する．本アプローチは，同一ファイル内においても様々な機密度の情報が混在することに着目し，ファイルを細かい情報単位（チャンク）に分割し，チャンク毎の機密度を機械学習を用いて推定する．これにより，同一ファイル内でも機密度に応じて情報を再分類でき，セキュリティを確保しつつ情報の利便性を高めることが可能になる．

Resolving the Data Usability-Security Trade-off through Information Confidentiality Estimation

Satoru Koda[1,a)]　Hitoshi Yoshihashi[1]　Takuya Yokota[1]　Toshiya Shimizu[1]　Yuji Higuchi[1]
Kentaro Tsuji[1]　Hisashi Kojima[1]

Abstract: With the proliferation of generative AI, there is a growing demand to utilize the data accumulated by organizations for business purposes. Organizations possess a mix of information with various levels of confidentiality (e.g., secret, proprietary, and public), and if the confidentiality levels are not appropriately configured, it can lead to information leaks or loss of opportunities to access information. For instance, elevating a file's confidentiality level to prioritize security effectively mitigates the risk of information leakage. Conversely, this restricts the accessibility of information, which may consequently diminish the user experience of generative AI applications using organizational data. This paper proposes an approach to resolve such a trade-off between data usability and security. This approach involves dividing files into small information units (chunks) and estimating the confidentiality level of each chunk using machin learning. This enables the reclassification of information within a single file based on its confidentiality level, thereby enhancing information usability while simultaneously ensuring security.

1. はじめに

本稿では，組織内のデータリポジトリに蓄積されたドキュメント群を，視覚言語モデル（VLM）や検索拡張生成（RAG）技術を用いて，組織内部やお客様向けに利活用する状況を想定する．例えば，生成 AI を用いた資料生成アプリや，組織特化型 QA チャットアプリ等を介して，データを活用したサービスを提供する．

このような状況においては，アプリの出力を情報の機密度に応じて制御することが求められる状況がある．例えば，ある社員が資料生成アプリを用いてお客様への説明資料を作成する際，アプリは公開可能な情報のみを出力することが安全上望ましい．このような制御は実装上は，RAG の情報源となるファイル（もしくはそれを分割したチャンク）のメタデータとしてファイルの機密度を記録し，VLM が参照できる情報を機密度に応じてフィルタリングするこ

[1] 富士通株式会社 (Fujitsu Limited)
[a)] koda.satoru@fujitsu.com

とで実現できる．

この制御を正しく機能させるためには，情報機密度を正しく設定する必要がある．一般に機密度ラベルはファイル単位で付与されることが多いが，それはアプリ利便性の面では必ずしも最適ではない．例として，ある社員が役員向けにプロジェクト紹介資料を作成した場面を想定する；本資料のうち技術紹介パートは，公開可能情報を基に図表化や要点整理をして，役員向け説明に練り上げたものであった．しかし本資料は役員向け説明用途のため，ファイル単位では機密情報として取り扱われた．この例では，練成した技術紹介パートをお客様向けに使いたくとも，該当部分は公開可能情報にも関わらずファイル自体の機密度は上位レベルに設定されているため，アプリが情報をフィルターアウトする可能性がある．結果，アプリ利用者の情報取得機会が失われ，ひいてはビジネス機会損失も起こり得る．

このようなケースで無くとも，セキュリティ重視ポリシーのためにファイルの機密度は高く設定される傾向があり，本来はアクセス可能な情報にアプリがアクセスできないケースが多発する．このような場合セキュリティは保たれるが，データ（を用いたアプリ）の利便性があがらない．逆に，利便性を重視する場合はセキュリティが脅かされる．このように，データのセキュリティと利便性はトレードオフの関係にある．

本稿ではこのようなトレードオフ問題を解消するため，データを細かい情報単位（チャンク）に分割して機密度を推定するアプローチを提案する．これにより，安全性を保ちながら利活用できる情報を増やす．また本アプローチを高精度に実現するため，以下の 3 要素から成る手法を提案する：1. チャンク埋め込みモデルの選定フレームワーク，2. 情報固有度に基づく教師チャンクの選定，3. グラフニューラルネットワークを用いた機密度推定モデル．

本研究の貢献は以下の通りである：

- 本研究は，テキストベースではないリッチドキュメントに対して機密度推定を行う初の研究である．
- 提案手法はチャンク単位の機密度推定タスクにおいて，疑似データで F1-score 93.18%, 実データで F1-score 94.30% を達成した．

2. 準備

第 1 節で述べたような，組織が保有するデータリポジトリ内のドキュメントファイル群を組織内部向けやお客様向けに利活用する状況を想定し，前提と課題を整理する．

前提　リポジトリにファイル群 $D = \{\text{file}^i = (x^i, y^i)\}_{i=1}^N$ が存在する．ここで，file^i は PowerPoint, Word, テキスト等のドキュメントファイルを，$x^i \in \mathcal{X}$ はファイルのコンテンツそのものを，$y^i \in \mathcal{Y}$ はファイルの機密度を表す．機密ラベル集合 $\mathcal{Y}$ は，例えば SECRET（最上位機密），CONFIDENTIAL (CONF., 関係者外秘), RESTRICTED (REST., 社外秘), PUBLIC（公開可）で構成される．機密度は各ファイル所有者が事前に設定したラベルとする．

課題　第 1 節で述べたように，本来であれば利用可能な情報にアクセスできない事象が発生し，結果的にデータを利用するアプリケーションの利便性があがらないといった問題が生じる．これには以下のような要因が考えられる：

(1) 一般に組織はセキュリティをより重視するため，ファイル機密度が高めにラベル付けされる傾向がある．
(2) ある時点での機密情報が経過により下位ラベル相当となるが，ファイルの機密度ラベルが更新されない．
(3) 機密度が「ファイル単位」に設定されるため，ファイル内で利用可能/利用不可能な情報を分類できない．

3. 提案アプローチ

上記課題を解決するアプローチを提案・定式化する．

3.1 アプローチ概要

図 1 上部にアプローチ概要を図示する．本稿では，各ファイルを細かい情報単位（チャンク）に分割し，チャンク毎の機密度を機械学習により推定するアプローチを提案する．この分割は，PowerPoint ファイルであればページ単位，Word ファイルであればパラグラフ単位，他にもレイアウト単位（図表，テキストボックス等）といったように，ある程度の情報量を持つ情報の塊とする．これにより，機密度が不当に高く設定されたファイルや，様々な機密度の情報が混在するファイルからでも利用可能な情報のみを抽出することができるようになり，セキュリティレベルを維持しつつ利用可能な情報を増やすことができると考える．また反対に，機密度が下位のファイルの中から上位レベルの機密情報を検出することも可能になる．

3.2 アプローチ定式化

3.2.1 問題設定

各ファイルのコンテンツは $x^i = [x^i_1, ..., x^i_{M_i}] \in \mathcal{X}^{M_i}$ のように，順序を持つ M_i 個のチャンクに分割されるとする．各チャンクが由来するファイル（x^i_j を含む file^i）を親ファイルと呼称する．チャンク毎のラベル推定とは，各 x^i_j の機密度を推定するタスクを指す．ここで記号として，ラベル推定器（Labeler）$L : \mathcal{X} \to \mathcal{Y}$ を導入する．この関数は入力チャンク x に，背景情報 B（メタデータ，コンテキスト等）を基に，予測ラベル $\hat{y} = L(x; B)$ を返す．

3.2.2 ラベル推定ロジック

本ラベル推定に採用可能な推定ロジックについて記述し定式化する．推定ロジックは絶対的推定と相対的推定の 2 つに分類できると考える．

絶対的推定　情報取扱規則等に記載された基準（CRITERIA）を基に機密度を推定するロジックを指す．例えば，「顧客情報（顧客名，取引情報等）を含む情報は

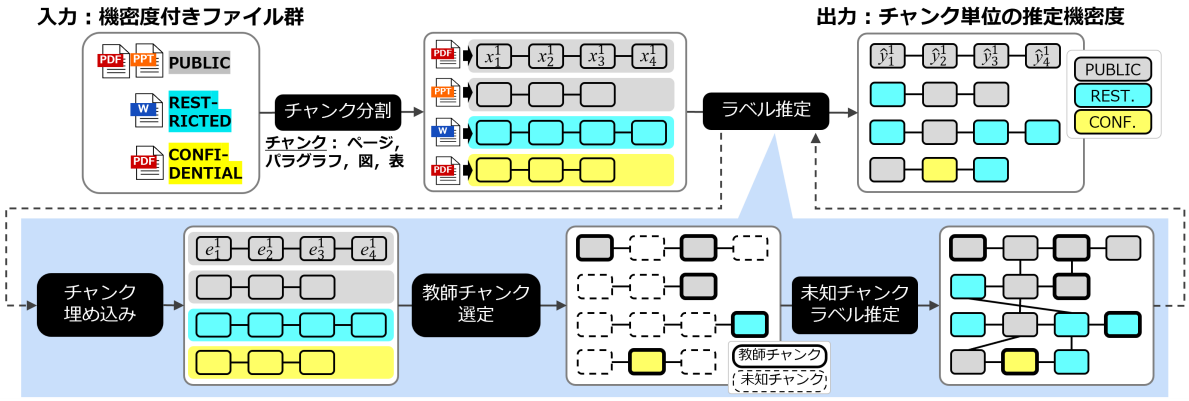


図 1: 提案アプローチのフローチャート：機密度ラベル付きのファイル群を入力に，チャンク分割，チャンク埋め込み，教師チャンク選定，機械学習（グラフニューラルネットワーク）を経て，チャンク単位で推定機密度を出力する．

Fig. 1 Flowchart of Proposed Approach

CONF. とする」のような基準があれば，チャンクが顧客情報を含む場合に CONF. と推定する．このような推定は $L_{\mathrm{DET}}(x; B = \mathrm{CRITERIA})$ と定式化できる．

相対的推定 もし，機密度が有る程度信頼できるチャンクの集合（教師チャンクと呼ぶ） $D_T = \{(x_j^i, y_j^i)\}_{ij} \subset D$ が存在するのであれば，それらとの相対比較や帰納的推定によるラベル推定が可能になる．例えば「あるラベル未知のチャンクに対し，そのチャンクに最も類似する教師チャンクのラベルを付与する」といった推定を実現できる．このような推定ロジックを相対的推定と呼び，教師チャンク D_T を背景情報として $L_{\mathrm{REL}}(x; B = D_T)$ と定式化する．教師有り機械学習はこれの一つの実現手段にあたる．

上記 2 つの推定ロジックは互いに長所・短所を補完するものであるが，本稿では相対的推定ロジックの検討を展開する．絶対的推定はユースケースに応じて基準が変わるため、相対的推定の方が広く適用可能と考えるためである．

4. 提案手法

前節で述べた提案アプローチの実現手法を具体化して提案する．まず前提として，相対的推定ロジックに基づきチャンク単位のラベル推定を行うために，我々は以下の 3 つの技術要素が必要と考えた：1. 埋め込みモデル，2. 教師チャンク選定手法，3. ラベル推定モデル．図 1 下部のフローチャートに，これらの技術要素を配置する．本節は各小節で技術要素ごとに，1. 技術要件の整理，2. その要件を満たす提案手法の記述，を行う形で構成する．簡単のため，以降では PowerPoint 形式のファイルを主に想定する．

4.1 埋め込みモデル

埋め込みモデルは，チャンク $x \in \mathcal{X}$ を埋め込み空間 $\mathcal{H}$ に埋め込む関数 $E : \mathcal{X} \to \mathcal{H}$ であり，チャンクを数式上取り扱い可能な実数空間に射影する．機械学習やチャンクの相対比較を行うにあたり必要となる．

4.1.1 埋め込みモデルの要件

まず，標題について以下のように整理する：

(1) マルチモーダル対応：テキストや図表を含むマルチモーダルドキュメントに対応できる．

(2) 文字認識（OCR）能力：チャンクに文字として何が書かれているかを理解できる．

(3) 内容理解能力：チャンク内の文字と文脈から何が書かれているかの “内容” を理解できる．

(4) 流用関係にあるチャンクペアに高い類似度を出力できる．

この中で要件 4 がラベル推定タスクにおいて独特な要件であると考える．これは相対的推定にあたり，参照情報となる教師チャンクとの類似度がラベル判断材料の一つになるためである．背景として，組織内のデータは様々な形で流用される．PowerPoint ファイルであれば，あるページ（チャンク）を，図を編集・縮小・拡大したり，文章を追加・削除したり，テキストと図のレイアウトを変更したり，といった処理を行い流用することが多い．この際，流用元・先の関係にあるチャンクペアは，改変の度合いが大きくない限り，同じ機密度を持つ可能性が高い．従って，このような流用関係にあるチャンクペアの埋め込みが高い類似度を持つことが，ラベル推定精度向上に寄与すると考える．

4.1.2 提案手法：埋め込みモデル選定フレームワーク

そこで我々は，「要件 1-3 を満たすと想定されるモデル候補に対して要件 4 の評価を行い，その結果が最良のモデルを選択する」という戦略で，埋め込みモデルを選定することを提案する．

要件 1-3 を満たすモデルの例に，Document Visual Question Answering (DocVQA) タスクを解くモデルがある．DocVQA は，テキストや図表を含むドキュメントと質問文から成るペアを入力とし，回答文を生成するタスクである．従って高度な DocVQA モデルは上記の要件 1-3 を満たすと考える．

要件 4 については以下の方法で評価を行う．始めに評価データとして，アンカーとなるチャンク x_{anc} と，それと流用関係にあるチャンク x_{pos}（正例チャンク）から成るペア (x_{anc}, x_{pos}) を複数準備する．更にここから，チャンクトリプレット $t = (x_{anc}, x_{pos}, x_{neg}) \in \mathcal{X}^3$ を複数構築し，その集合 T を得る．ここで x_{neg} は，アンカーチャンク x_{anc} と流用関係にないチャンク（負例チャンク）を表す．

続いて，各モデルの要件 4 を評価データを用いて評価する．提案手法は評価基準にトリプレット損失を用いる．まず，チャンクの埋め込み同士の類似度を算出する関数 $Sim : \mathcal{H} \times \mathcal{H} \to \mathbb{R}_{[-1,1]}$ を定義する．埋め込みモデル $E \in \mathcal{F}$ は，トリプレット損失 $l : \mathcal{X}^3 \to \mathbb{R}_{[-2,2]}$ に基づく評価関数 $S : \mathcal{F} \to \mathbb{R}_{[-2,2]}$ を用いて，以下のように評価される：

$$S(E;T) = \frac{1}{|T|} \sum_{t \in T} l(t), \tag{1}$$

$$l(t) := Sim\left(E(x_{anc}), E(x_{neg})\right) - Sim\left(E(x_{anc}), E(x_{pos})\right).$$

この損失は，1. 流用関係にあるチャンクペアに高い類似度を，2. 流用関係にないチャンクペアに低い類似度を，出力するほど小さくなる．即ち，$S(E;T)$ の値が小さいほど，関数 E は本タスクに最適であると判断される．

4.2 教師チャンクの選定

教師チャンクは，相対的アプローチで基準を成すチャンクである．その選定タスクは，ファイル群 D から部分集合 D_T を抽出することとして定式化される．

4.2.1 教師チャンクが満たすべき要件

無論，教師チャンクは正しい機密度ラベルを持っていることが保証されたチャンクであることが望ましい．これは，人手の介入や仮定次第では可能であるが，現実には実現困難なケースが多い．

4.2.2 提案手法：情報固有度に基づく教師チャンク選定

そこで我々は要件を緩和し，「ある情報がラベル y のファイルに “固有に” 出現するならば，その情報はラベル y であるべきである」という原理で教師チャンクを選定する手法を提案する．例えば，ある図 A がラベル CONF. のファイルには出現するがそれ以外の機密度のファイルには出現しない場合，図 A はラベル CONF. を持つべきである，という考え方である．より公式には，親ファイルのラベルが y であるチャンク x が，y 以外のラベルを持つファイルのいずれのチャンクにも類似しなければ，チャンク x をラベル y を持つ教師チャンクとして選定する．この方法は情報の中身に依存しないため，絶対的な評価ほど十分ではないが，少なくともこのようなチャンクは親ファイルと同一の機密度であると判定されざるを得ないチャンクと言える．

上記の原理に基づき教師チャンクを選定する手法を具体的に以下に記述する：

(1) 各チャンクの機密度ラベルを，親ファイルの機密度で初期化 (つまり $y_j^i \leftarrow y^i$).

(2) 各チャンクを埋め込みモデルに入力し，その埋め込み $e \in \mathcal{H}$ を取得.

(3) 埋め込み同士の距離を算出し，距離行列 $(\in \mathbb{R}_+^{M \times M})$ を取得（M は総チャンク数）.

(4) 距離が $\delta\,(>0)$ 以下のチャンクを集約 → クラスタ C_k $(k = 1, ..., K)$ を構成.

(5) 各クラスタ C_k に対し，1. 最近傍クラスタとの距離 d_k が $\eta\,(>0)$ 以上（つまり他のいずれのクラスタにも類似しない），かつ，2. クラスタ内のチャンクで機密度ラベルの不一致が無い（つまりそのラベルに固有である），かを判定．本条件を満たすクラスタの和集合を取り，教師チャンクとする.

最終的に（かつ簡易的に），教師チャンクは $D_T = \bigcup_{k:\text{Step(5) is True}} C_k$ で表現できる.

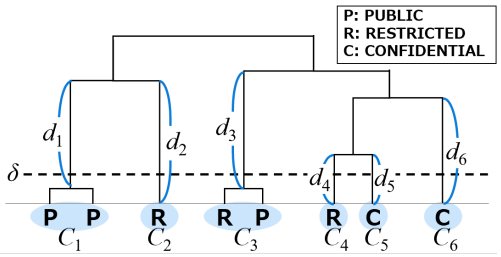


図 2: 教師チャンク選定プロセスの図例（デンドログラム）
Fig. 2 Example of Teacher Chunks Selection (Dendrogram)

本選定プロセスを，階層型クラスタリングで用いられるデンドログラムを用いて**図 2** に例示する．図では，8 つのチャンク（PUBLIC × 3, REST. × 3, CONF. × 2）がある．クラスタ間距離が δ 以下のチャンクを集約すると，クラスタ $C_1, ...C_6$ が構成される．その後，ステップ (5) の条件によりクラスタ C_1, C_2, C_6 が抽出され（ここでは d_4, d_5 のみ η 以下と仮定），これらのクラスタに属する 4 つのチャンクが教師チャンクとして選ばれる．

4.3 ラベル推定モデル

ラベル推定モデル $L(x;B)$ は，背景情報 B を基に，チャンク x の機密度を推定するモデルである．（厳密には，入力に x ではなく e をとる.）

4.3.1 ラベル推定モデルが満たすべき要件

まず，標題について以下のように整理する：

(1) ファイルやチャンク同士の関係（例：バージョン，時系列，チャンク構造）をモデル化できる.

(2) 隣接するチャンクを文脈として利用できる.

(3) 隣接するチャンクのラベルが相関を持つ.

要件 1 について，チャンクは独立ではなく他のチャンクと様々な関係性を持つため，それらをラベル推定時に利用できることが望ましいと考える．その中でも特に，チャンクの構造上の隣接関係（x_j^i と $\{x_{j-1}^i, x_{j+1}^i\}$）は重要である．

チャンクの内容は一般に隣接チャンク（例：前後のページ）の内容と関連するため，ラベル推定時には隣接チャンクの情報を文脈として利用すべきである（要件 2）．また隣接チャンクは共通のラベル持つことが多いため，そのような傾向を制御できるできることが望ましい（要件 3）．

4.3.2 提案手法：グラフニューラルネットワーク（GNN）

上記の要件を鑑み，提案手法は GNN ベースのモデルを採用する．GNN はグラフ構造を持つデータの分析によく用いられる．グラフ $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ は，ノードの集合 $\mathcal{V}$ と，ノード間の関係性を表すエッジの集合 $\mathcal{E}$ から成る．GNN はノード特徴量を入力として受け取る．各層では，各ノードに接続するノード（近傍ノード）から情報を集約し、ノード特徴量を更新する操作を行う．簡易的に，第 k 層の処理（集約，更新）は以下のように表現される：

$$\mathbf{h}_v^{(k)} = \sigma_{\text{update}}^{(k)}\left(\mathbf{h}_v^{(k-1)}, \sigma_{\text{aggregate}}^{(k)}\left(\mathbf{h}_u^{(k-1)} \mid u \in \mathcal{N}(v)\right)\right). \tag{2}$$

ここで，$\mathbf{h}_v^{(k)}$ はノード v の第 k 層特徴量を，$\mathcal{N}(v)$ はノード v の近傍ノード集合を，$\sigma_{\text{aggregate}}^{(k)}, \sigma_{\text{update}}^{(k)}$ は第 k 層集約・更新処理を表す．GNN は教師を持つノードを基に最適化され，未ラベルノードに推定結果を出力する．

GNN は，グラフ特性（例：無向/有向グラフ，ハイパーグラフ）には依存するものの，チャンクの様々な関係性をグラフで表現できるため，要件 1 を満たす．また，構造的隣接関係にあるチャンクにエッジを結ぶことで，順伝播時に隣接チャンクに情報が伝達するため，要件 2 を満たす．加えて，GNN と Label Propagation (LP) [3] を併用することで，ノードのラベル情報を直接グラフ上で伝播して隣接ノードのラベル推定を行うことも可能であり，要件 3 を満たす．

提案手法においては，チャンクをノード，チャンク同士の関係性をエッジに見立てグラフ $\mathcal{G}$ を構成する．具体的には，チャンク同士が「構造的隣接関係にある」または「類似度が一定以上である」場合に，チャンク間にエッジを張る．後者の条件により，異なるファイル間に存在する類似チャンク同士にエッジが結ばれることになるため，LP を効率化できると考える．また用途に応じてエッジを有向化する．教師チャンクに相当するノードは教師ラベルを持つ．また，提案手法で用いる LP アルゴリズム [3] はノードに初期ラベルを設定できるため，教師チャンク以外にも親ファイルと同一のラベルを付与する（これらは訓練課程で更新される）．ノードの第 0 層特徴量 $\mathbf{h}_j^{i(0)}$ には，チャンク埋め込みモデルの出力 $e_j^i = E(x_j^i)$ を用いる．そして式 (2) を用いて特徴量を伝播し，半教師有り学習を行う．最終的に GNN と LP により，GNN 層数に依存する k 次隣接ノードの情報（特徴量，ラベル）と初期ラベルの情報を総合して，未ラベルチャンクの機密度ラベルが再推定される．

表 1: 埋め込みモデル評価結果
Table 1 Result of Embedding Model Evaluation

モデル	モデルサイズ	埋め込みサイズ	評価値
ColPali-v1.3 [5]	3B	パッチ数 × 128	-0.367
ColQwen2-v1.0 [4]	2B	パッチ数 × 128	**-0.428**
LLaVE-2B [6]	2B	1 × 1536	-0.297

GNN を用いたラベル推定器は $L_{\text{GNN}}(x; B = \{D_T, \mathcal{G}\})$ と表現でき，SVM のような古典的機械学習推定器 $L_{\text{SVM}}(x; B = D_T)$ と比較し，グラフトポロジーが背景情報に加わったものと見ることができる．

5. 評価実験

本節では，始めに第 5.1 節にて，埋め込みモデル選定のための評価を行う．その後第 5.2 節にて，選定された埋め込みモデルを用いてラベル再推定評価を行う．

5.1 埋め込みモデル選定

データセット　筆者らは本評価に際し実データセットを作成した．対象としたファイル形式は PowerPoint である．まず始めに，スライドの典型的な改変流用パターンを 31 通り定義した．これには，文字を減らす，全体を縮小する，文字と図を入れ替える，等の改変を含む．3 ソースから成る 106 個のチャンク（ページ）にそれぞれ適用可能な改変を適用し，計 635 枚の流用スライドを作成した．この一部を用いて，3175 個のチャンクトリプレットを作成し評価データとして用いた．

評価対象モデルの選定　始めに，DocVQA タスクを含むベンチマークである MMEB Leaderboard [14] 及び Vidore Leaderboard [13] の上位にランクするモデルを抽出した（2025 年 4 月時点）．更にそのモデルとその派生・亜種モデルの中から，「ライセンスが問題ない」かつ「計算機上で動作するサイズ」という条件を満たすモデルを選定した．結果的に，ColPali-v1.3 [5], ColQwen2-v1.0 [4], LLaVE-2B [6] の 3 モデルが選定された．これらのモデルはいずれも，チャンクを画像（“textual images”）に変換した後モデルに入力することで，埋め込み行列を取得できる．埋め込み行列同士の類似度計算には，Late Interaction [5] を用いる．

結果　これらのモデルを，第 4.1.2 節で提案したモデル選定フレームワークで評価した．結果を**表 1** に記す．表中の評価値列はモデル選定基準 (式 1) の値を表す．結果，モデル候補の中では ColQwen2-v1.0 が最適と判定された．この結果の妥当性は第 5.2.3.1 節で更に分析する．

5.2 ラベル再推定評価

5.2.1 実験設計

5.2.1.1 データセット

疑似データ　異なる機密度間で共通する情報が存在しな

表 2: 実験データセットの統計
Table 2 Statistics of Datasets

データセット	ファイル数	チャンク数	正解ラベル分布
	(CONF., REST., PUBLIC)		
疑似データ	(2, 2, 2)	(20, 20, 20)	(8-20, 8-20, 20-44)
実データ	(3, 8, 5)	(182, 167, 78)	(172, 148, 107)

いように意図して作成した疑似データセットである．機密度ラベルは 3 種類存在する（CONF., REST., PUBLIC）．機密度毎に 2 ファイル，計 6 ファイル存在する．各ファイルは 10 ページから成り，計 60 チャンクのデータである．

実データ　実際の PowerPoint ファイルを，個人識別可能情報（PII）のマスク処理を例外に，そのまま利用したデータセットである．本ファイル群は，ある研究プロジェクトの立ち上げから成果公開に至る過程で生成された実スライドである．機密度ラベルは 3 種類存在する（CONF., REST., PUBLIC）．公開情報はブログ記事の原稿や学会発表でのプレゼンテーション資料等を含む．

両データともスライド 1 ページを 1 チャンクとみなす．データセットの統計を**表 2** に記す．

5.2.1.2 正解ラベル作成手順

疑似データ　各ファイルのチャンクは親ファイルと同一のラベルを持ち，かつそれらは真であると仮定する．そして以下の処理で合成データセットを作成する．PUBLIC ファイルのあるチャンク x^i_j $(1 \leq j \leq 3)$ で，PUBLIC ファイル以外のファイル k の，4 ページ目以降のとあるチャンク x^k_l を上書きする．つまり，x^k_l は親ファイルのラベルは $y^k (\neq \text{PUBLIC})$ であるが，正解ラベルは PUBLIC となる．この操作により，機密度が混在するファイルを含む合成データセットを作成する．更に，上書き位置・枚数を様々に変えることで，ファイル内のチャンク機密度混在度合いが様々なレベルで変化する，計 12 種の合成データセットを作成する．

実データ　人手でラベル付け作業を行う．本データ内の情報は時系列的に CONF. → REST. → PUBLIC と変遷したデータである．従って，最終的に PUBLIC となった情報が機密度上位（{CONF. REST}）ファイルに含まれる場合，その情報を含むチャンクの正解ラベルを PUBLIC とみなす（最終的に REST. となった情報も同様）．

つまり両データとも推定タスクは，機密度上位のファイルに含まれる機密度下位情報を抽出することに相当する．

5.2.1.3 評価手順

データ前処理　各ファイルをチャンク分割し，前述の実験にて選定された埋め込みモデル ColQwen2-v1.0 を適用し埋め込み行列 $e \in \mathbb{R}^{768\times128}$ を取得する（768 は最大パッチサイズに相当）．更にそれをパッチ方向に平均化し埋め込みベクトル $\bar{e} \in \mathbb{R}^{128}$ を取得する．

ラベル推定モデル訓練　疑似データは，各ファイル先頭 3 ページ（計 18 チャンク）を教師チャンクとみなしてモデルを訓練し，末尾 7 ページ（計 42 チャンク）のラベル推定を行う．実データは，第 4.2 節に記載の提案方法で教師チャンクを選定してモデルを訓練し，それ以外のチャンクにラベル推定を行う．

評価基準　各ラベルのサンプル数で重み付き平均化した Precision, Recall, F1-score 基準を用いる．いずれの基準も高いほど良い分類であることを表す．重み付き平均のため，F1-score は必ずしも Precision と Recall の間の値を取るとは限らないことに注意されたい．

5.2.1.4 ラベル推定モデル

以下に，評価するラベル推定モデルを記述する．

保守的推定器　各チャンクに対し，そのチャンクの親ファイルの機密度ラベルを機械的に返すモデル $L_{\text{CONS.}}(x^i_j; B = \{y^i\}) = y^i$.

（古典的）機械学習推定器　教師チャンクで訓練された SVM モデル $L_{\text{SVM}}(x; B = D_T)$．加えて，派生モデルとして，Leave-One-Out Cross-Validation（LOOCV）形式で機密度を推定するモデル $L_{\text{SVM-CV}}(x; B = D \setminus \{(x, y)\})$ も比較する．本モデルでは，全てのチャンクを教師チャンクとみなす．教師ラベルは親ファイルと同一とする．そして LOOCV 形式で未ラベルチャンクを 1 つずつラベル推定する．従って，モデル L_{SVM} と比較して教師データは増えるが，ラベル誤りを含むノイズチャンクが訓練データに混入する可能性が高くなることに注意されたい．

またグラフ推定器との公平な比較のため，SVM ベースの 2 モデルに以下の処理を適用する：1. 隣接チャンクの文脈を考慮するため，事前に隣接チャンク間で埋め込みを平滑化する（つまり $\bar{e}^i_j \leftarrow (\bar{e}^i_{j-1} + \bar{e}^i_j + \bar{e}^i_{j+1})/3$）．2. PUBLIC ラベルへの推定を比較的重視する（PUBLIC クラスの Recall を高める）ために，PUBLIC に相当するクラス重みを大きくしてモデルを訓練する．

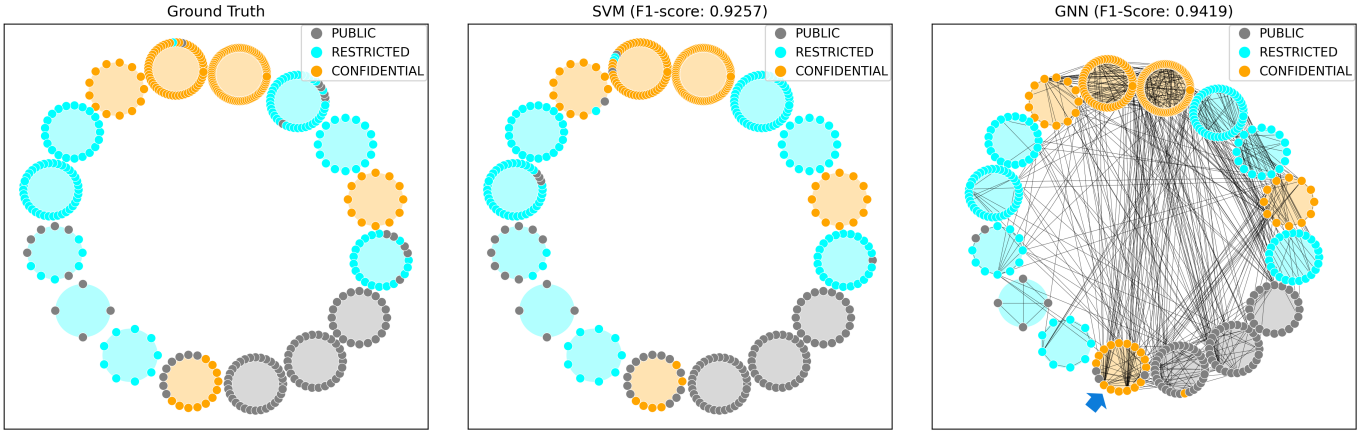
グラフ推定器　提案手法である GNN ベースの推定モデル $L_{\text{GNN}}(x; B = \{D_T, \mathcal{G}\})$．本モデルは重み初期化にランダム性があるため，実データにおいては訓練を 5 回試行し結果の平均を取る．疑似データは複数の合成データセットから成るため，それらの結果を平均化する．

5.2.2 実験結果

表 3 に評価結果をまとめる．両データセットにおいて，提案手法が既存手法を定量的に上回る結果となった．以下，各データセットにおける結果を考察する．

疑似データの結果は，ファイル内の機密情報の混在度合いがどのようなレベルであれ，GNN が総合的に良い結果を出すことを示す．なぜならば疑似データは，ファイル内の機密度混在度合いが様々なレベルで変化する，複数の合成データセットで構成されるからである（第 5.2.1.2 節参照）．

続いて実データの結果分析のために，**図 3** に正解ラベルと，各モデルのラベル推定結果を図示する．**図 3a** が示す



(a) 正解ラベル　(b) SVM-CV による推定結果　(c) GNN による推定結果

図 3: 正解ラベルと各モデルによるラベル推定結果：各ノードはチャンクに対応し，それらの色は正解・予測ラベルの機密度（橙色: CONF.，水色: REST.，灰色: PUBLIC）を表す．背景の複数ノードから成る円は 1 ファイルに対応し，その色は同様にファイル自体の機密度を表す．またチャンクは円状に順番に配置されている．図 (c) のエッジは実際に構築されたグラフを表現する．図 (c) の青矢印は，推定失敗要因の分析対象ファイルを指す（第 5.2.3.2 節参照）．

Fig. 3 Ground truth and prediction labels by each model.

表 3: 実験結果
Table 3 Evaluation Results

データセット	手法	評価基準		
		Precision	Recall	F1-score
疑似データ	CONS.	88.11	81.11	80.96
	SVM	89.62	88.89	88.73
	SVM-CV	89.58	86.53	86.63
	GNN	**94.17**	**93.19**	**93.18**
実データ	CONS.	93.64	92.97	92.62
	SVM	71.31	50.58	49.84
	SVM-CV	92.72	92.51	92.57
	GNN	**94.77**$\pm$0.14	**94.47**$\pm$0.21	**94.30**$\pm$0.23

表 4: アブレーション分析結果 (F1-score, %)
Table 4 Ablation Study (F1-score, %)

		モデル選定	
		なし (ColPali)	**あり** (ColQwen2)
教師チャンク選定	ランダム	91.00$\pm$2.06	92.36$\pm$1.48
	提案	91.48$\pm$1.09	**94.30**$\pm$0.23

ように，実験に用いた実データは，ファイル内の機密度混在度はそれほど高くなく，チャンクは親ファイルと同一のラベルを持つパターンが多かった．故に保守的推定器でも数値的には良い結果を得られる．SVM-CV と GNN の推定結果を比較すると，GNN の方が隣接チャンクが同一のラベルを共有する割合が高くなっていることがわかる（**図 3c**）．つまり，隣接チャンク間で情報の文脈が保たれるケースにおいては，GNN は優位に働くと考えられる．しかしそうではない場合，例えば CONF. ラベルのファイルの 1 ページにだけ無関係な文脈で PUBLIC ラベルを持つチャンクが現れる場合，そのラベルを正しく推定することは比較的苦手である．これは GNN に頻発する over smoothing 現象（ノード特徴量が近傍ノードで均一化され、個々のノードの識別性が失われる現象）に起因すると考えられる．

5.2.3 実験結果分析

5.2.3.1 アブレーション分析

上記実験では，全推定モデルで共通の埋め込みモデル（ColQwen2-v1.0）と教師チャンク選定方法を用いた．そこでこれら 2 つの提案要素の貢献度合いを分析するために，GNN 推定器に対し，これらの要素を一部除いたモデルを訓練し評価を行った．具体的には，埋め込みモデルは評価結果第二位となった ColPali-v1.3（表 1 参照）に，教師チャンク選定方法はランダムサンプルに，それぞれ置き換えて評価実験を行った．その結果を**表 4** に記す．結果，両要素とも精度改善に寄与していることがわかった．ここでは特に埋め込みモデルに着目し，グラフの “ラベルホモフィリー”（エッジを結ぶチャンク同士が同一の機密度ラベルを持つ割合）を分析する．一般にラベルホモフィリーが高いほど GNN の推定は容易になる [3]．提案 GNN では，グラフ構築時に「類似度が一定以上である」条件でエッジを結ぶため，ラベルホモフィリーは埋め込みモデルに依存する．実際に埋め込みモデル毎にその値を算出すると，ColQwen2-v1.0 が 0.82, ColPali-v1.3 が 0.69 であった．つまり，提案した埋め込みモデルフレームワークは，GNN 推定に有利なラベルホモフィリーが高くなる埋め込みモデルを選択する能力があることが示唆される．

5.2.3.2 失敗例分析・対策

続いて GNN 推定の，over smooting 以外に起因する失敗例を分析する．ここでは，図 3c の 6 時半方向にある矢印

が指す CONF. ラベルを持つファイルに着目する．本ファイルは CONF. ラベルであるが，一部チャンク（円の 8-12 時部分）は PUBLIC が真であり，これらの再推定に失敗した．この原因の一つは，グラフ構築時に，これらのチャンクに関連する PUBLIC ラベルを持つチャンクにエッジが結ばれなかったことと考える．そのため，本来であれば類似する PUBLIC チャンクからラベル情報が LP によって伝達されるが，それが機能しなかった．データを分析したところ，エッジが結ばれなかった要因として，(1) チャンクの内容自体は関連するが，表現が大きく異なっていたため，類似度が低く判定された，(2) チャンクの情報が複数チャンクの合成であるため，単体ペアで見ると類似度が低く判定された，ことが発生していたと分析する．対策としては，「失敗の要因となる類似度が低く判定される流用関係」のペアで埋め込みモデルをファインチューニングし，ラベルホモフィリーを高めることが挙げられる．

最後に，情報漏洩リスクになり得る機密度下位への誤ラベルを分析する．実験では 1 つのチャンクが，正解が CONF. に関わらず PUBLIC と推定された（矢印が指すファイルの 3 時半方向）．しかし本チャンクの中身は実際には背表紙（会社ロゴのみが書かれたスライド）であった．今回はアノテーションの際，セキュリティ安全側に倒し，表紙類は全て親ファイルと同一のラベルを正解ラベルとして付与した．しかし実際には背表紙は様々な機密度のファイルに存在するため，モデルにとってラベル判定が困難なチャンクであったと考える．対策としては，表紙類をテンプレートとして用意し，事前に類似度マッチングによりそれらを推定対象チャンクから除外することが挙げられる．

6. 関連研究

ドキュメントの機密度推定　ドキュメントのコンテンツからその機密度を推定する研究は多からず存在する．どのような情報を機密対象とするかは文献に依って異なり，外交関連文書 [1], [2], [8], PII [7], [8]，医療記録 [11] などが扱われている．しかし既存研究は総じてテキスト形式の文書を扱っており，企業データに多く見られるリッチドキュメントを対象とした研究・ベンチマークデータセットは無い．また一部研究では，ドキュメント単位より細かい単位（パラグラフ）での分類の必要性を主張している [1], [2]．しかしこれらの研究において，GNN のようなパラグラフの関係性をモデル化して機密度を推定する手法は評価されていない．故に本研究はファイル単位より細かい粒度（チャンク単位）の機密度推定において，チャンクの関係性をモデル化することの優位性を示した初の研究と言える．

RAG セキュリティ　組織データの安全利用の観点は，近年は RAG セキュリティの文脈で語られることが多い．それらの多くは RAG システムそのものをセキュアに設計することで，機密情報の流出をジェイルブレイク等の攻撃から守ることに主眼を置いている [10]．RAG に与えるデータの安全性の側面では，PII 検出 [9] やデータベース暗号化 [9], [12] 等の対策を主張する既存研究が存在する．一方で本稿のように，情報の機密度を再推定することにより安全性を確保することを意図した研究は存在しない．

7. まとめ

本稿では，組織データを生成 AI 等を用いて利活用する際に直面し得る「データのセキュリティ–利便性トレードオフ」の問題を議論した．提案アプローチは，データをチャンクに分割して機密度を推定する戦略を採る．また，チャンク単位の機密度推定を高精度に実現するため，1. チャンク埋め込みモデルの選定フレームワーク，2. 情報固有度に基づく教師チャンク選定，3. グラフアプローチを用いた機密度分類モデル，から成る手法を提案した．以上により，機密度が高い情報を安全かつ利便的に利用することが可能になる．今後の課題として，チャンク内のごく一部に存在するような PII 等の重要情報に対して推定モデルの感度を高めることで，相対的推定と絶対的推定を兼用するようなモデルを構築することが挙げられる．

参考文献

[1] K. Alzhrani, E. M. Rudd, T. E. Boult, and C. E. Chow.: Automated Big Text Security Classification. *ISI*, 2016.

[2] E. Bass, M. Albanese, and M. Zampieri.: Disc: A Dataset for Information Security Classification. *SECRYPT*, 2024.

[3] Y. Cheng, C. Shan, Y. Shen, X. Li, S. Luo, and D. Li.: Resurrecting Label Propagation for Graphs with Heterophily and Label Noise. *KDD*, 2024.

[4] vidore/colqwen2-v1.0. `https://huggingface.co/vidore/colqwen2-v1.0`

[5] M. Faysse, et al.: ColPali: Efficient Document Retrieval with Vision Language Models. *ICLR*, 2025.

[6] Z. Lan, L. Niu, F. Meng, J. Zhou, and J. Su.: LLaVE: Large Language and Vision Embedding Models with Hardness-Weighted Contrastive Learning. *arXiv*, 2025.

[7] J. McKechnie and G. McDonald.: SARA: A Collection of Sensitivity-Aware Relevance Assessments. *arXiv*, 2024.

[8] H. Narvala, G. McDonald, and I. Ounis.: RelDiff: Enriching Knowledge Graph Relation Representations for Sensitivity Classification. *EMNLP*, 2021.

[9] M. Panda and S. Mukherjee.: Advancing Privacy and Security in Generative AI-Driven Rag Architectures: A Next-Generation Framework. *IJAIA*, 2025.

[10] A. RoyChowdhury, M. Luo, P. Sahu, S. Banerjee, and M. Tiwari.: ConfusedPilot: Confused Deputy Risks in RAG-based LLMs. *arXiv*, 2024.

[11] M. F. Sayed and D. W. Oard.: Jointly Modeling Relevance and Sensitivity for Search Among Sensitive Content. *SIGIR*, 2019.

[12] P. Zhou, Y. Feng, and Z. Yang.: Privacy-Aware RAG: Secure and Isolated Knowledge Retrieval. *arXiv*, 2025.

[13] Vidore Leaderborad. `https://huggingface.co/spaces/vidore/vidore-leaderboard`

[14] MMEB Leaderboard. `https://huggingface.co/spaces/TIGER-Lab/MMEB-Leaderboard`