メタデータおよびユーザ行動に基づくマルチビュー融合による データセット間類似度の学習手法

程昊陽 [1*]　　早矢仕晃章 [1]
Haoyang Cheng[1]　Teruaki Hayashi[1]

[1] 東京大学 大学院工学系研究科
[1] School of Engineering, The University of Tokyo

Abstract: A dataset's metadata, consisting of structured descriptors such as titles, tags, and descriptions, summarizes the dataset's topic, provenance, and structure without requiring access to its content. We propose a multi-view framework for learning dataset–dataset similarity solely from metadata. The available information is partitioned into two metadata views—Tag and Text—and an auxiliary non-metadata view, User Behavior. For the Tag and Text views, we construct Dataset–Tag/Word bipartite graphs, perform type-constrained random walks, treat the walks as sentences, and train Skip-gram with Negative Sampling (SGNS) to capture contextual co-occurrence, from which we derive per-view dataset–dataset similarities. For the User Behavior view, we model co-usage sequences using Item2Vec to obtain an auxiliary similarity. Finally, the three views are integrated through Adaptive Fusion to produce a unified similarity matrix over datasets. Experiments on the MetaKaggle dataset demonstrate that the proposed method outperforms standard baselines in a series of metrics including nDCG@20 and MAP@20.

1　はじめに

異なる分野のデータを組み合わせた価値創出がイノベーションの源泉として注目されてきている．このような中，Web 上のプラットフォームにおいてデータ提供者がデータまたはデータに関する情報を公開し，利用者がデータを検索・購入するデータ市場が発展してきた．近年の調査では，このようなデータ市場やデータ取引プラットフォームが世界的に登場してきており，金融，都市ガバナンス，医療など幅広い分野において重要インフラとして位置づけられはじめている [1].

活用可能なデータの種類や規模が増大する一方で，異種のデータの探索から発見，活用までのプロセスの随所で様々な問題が表出してきている．例えば，個々のデータセットはしばしば構造が複雑でサイズも大きく，加えてプライバシー規制やコンプライアンス要件を伴う場合も少なくない．そのため，データを実際にダウンロードして内容を確認し，前処理を行ったうえで類似性を判断するには多大なコストがかかる [2]．既存のデータセット検索システムやデータレイク内の探索手法は，タイトルや簡単なキーワード，表層的な特徴量に依存しており，異種性やノイズを含む環境では自身の関心に合致したデータセットを見つけることは困難である．結果として，データ再利用やモデル転移などの応用の妨げとなっている [3].

本研究ではこれらの課題に対し，マルチビュー融合によってデータセット間の類似度を推定する手法を提案する．まず，タグ情報，テキスト情報，ユーザ行動といった異なるデータに関する情報源をビュー（View）として扱う．そして，Tag/Text ビューにおいて二部グラフ上の型制約付きランダムウォークによって得られる系列を文と見なし，Skip-gram with Negative Sampling（SGNS）を用いて表現ベクトルを学習する．さらに，近似最近傍探索（Approximate Nearest Neighbor, ANN）や正規化手法を用いて各ビューの類似度行列を構築する．これら複数の類似度行列は，マルチビュー融合によって統合し，最終的に単一の類似度行列を得る．

提案手法は，従来の実データに基づく手法や単一ビューによる手法と比較して，大規模な実データへのアクセス制限や異なるデータ形式による制約を回避しつつ，タグ，テキスト情報，ユーザ行動という異なる情報源の相補的な特性を統合することができる．その結果，単一の情報源に基づくアプローチでは捉えにくかった，高次の意味的関係や潜在的な利用パターンを抽出でき，より有効なデータセット類似度の推定が可能となる．

*連絡先：東京大学大学院工学系研究科
〒 113-8656 東京都文京区本郷 7-3-1
E-mail: teikoyo@g.ecc.u-tokyo.ac.jp

2 先行研究

2.1 実データによる表形式データ類似度推定

データの内容，すなわち実データに基づく表形式データの類似性推定やデータ探索・発見研究は，セルの値や統計的特徴を利用して表データ同士の類似度を計算し，検索やマッチング，統合タスクに適用されてきた．例えば，Lv らは Ferret を提案し，特徴量が豊富なオブジェクトを対象として，大規模データから内容が類似するオブジェクトを探索する枠組みを提供した．しかし，構造化した表や意味情報の扱いには制約があるという課題があった [4]．また，Yan らは Web クエリに対して実データベースのテーブル検索手法を提案し，表形式データの内容に基づくランク付けによりマッチング精度を改善した．だが，この手法は主に単一クエリによる検索に焦点を当てており，表全体の類似関係を体系的に表現する手法ではない点に課題がある [5]．作本らは，多面的なデータセット分類基準を設定し，類似データセット発見タスクにおいて，様々な類似度指標の有効性を評価した．しかし，この研究の主な関心は指標選択と評価であり，個々のデータのどのような情報が類似度に寄与しているのかということや計算コストの問題には十分に踏み込んでいない [6]．

2.2 メタデータによるデータ類似度推定

メタデータに基づくデータの類似度推定では，タイトル，説明文，タグ，変数情報などの補助的属性を利用してデータセット間の関連性を評価することで，異種のデータ検索や統合を支援する方法が提案されている．例えば，Ravishankar らは，少数のメタデータ属性のみを用いた教師なしクラスタリング手法を提案し，実データへアクセスせずにクラスタリング品質を向上させようとした．しかし，主な対象は一般文書やレコードであり，データセットレベルでの類似性推定は限定的であった [7]．Bernhauer らは，文脈情報を組み込んだオープンデータの類似検索フレームワークを提案し，関連データセットの発見率が高まったと報告している．しかし，メタデータの疎密やノイズに対して脆弱であるという課題が残っている [8]．また，Sakumoto らは，メタデータによるクラスタリングとメタデータ項目選択手法を提案し，複数分野において異なるメタデータ類似度指標とその組み合わせを体系的に比較した．これにより類似データセット発見やデータ連携に有用な知見を提供したが，依然としてメタデータ内部における指標選択や重み付けが中心であり，異種メタデータ間の補完関係を統合的に活用する枠組みの提案には至っていない [9]．

2.3 マルチビュー融合とグラフ類似度の融合

マルチビュー融合とグラフ類似度の統合とは，相補的な複数の情報源であるビューが存在する場合に，それらを統合して一貫した類似度表現を構築し，単一ビューよりも安定したクラスタリングや分類性能を得る方法である．Kumar らは共正則化マルチビュースペクトラルクラスタリングを提案し，各ビューのスペクトル埋め込み間に一貫性のある正則化項を入れることでクラスタリング精度を向上させた．しかし，この手法はビュー間の品質差やノイズビューの影響を自動的に抑制する仕組みが十分ではない [10]．Wang らの Similarity Network Fusion は，各ビューをサンプル類似度グラフとして構築し，反復的なグラフ拡散により統合する手法であるが，ビューごとの類似度行列を事前に固定して与える必要があり，拡散パラメータやビュー重みの設定に敏感であるという課題がある [11]．さらに，Gönen らは複数ビューのカーネルに局所的な重みを付与することでサンプルの局所構造を捉え，クラスタリング性能を改善した．しかし，この手法は事前に設計または計算されたカーネルに依存するため，ビューの種類や構造が大きく異なる場合には直接適用しにくい [12]．

これらの手法はマルチビューとグラフの融合の有効性を示しているものの，ビュー間類似度をあらかじめ適切に構成し，ビュー同士も比較的同質であると仮定することが多い．そのため，出自や構造が大きく異なる複数のメタデータからなるビューや，その品質のばらつきや不均衡を細かく一貫して扱うモデルを統合したフレームワークは十分に整っていない．

3 提案手法

3.1 メタデータのマルチビュー表現

メタデータとは，データセットに関する情報を形式的に記述したものであり，データ名，概要，タグや利用履歴などの情報を含んでいる．メタデータは分析対象そのものではなく，主にデータセットの説明書として機能し，検索に用いられることが多い．実データの扱いと比較し，メタデータを使うことにはいくつかの利点がある．例えば，メタデータは軽量かつ少量の記述のみで大まかなデータセットの内容や関連性を判断するための情報が含まれている．そのため，アクセス性や計算のコストが低く，プライバシーやコンプライアンス上のリスクが小さい．また，異なる種類のデータセットでも，共通の記述項目によって横断的にデータ同士を比較することができる．さらに索引構築・クラスタリング・推薦などの基盤をメタデータ層で整備することができる．そのため，本研究では実データで

はなく，メタデータを対象としてビューを構築するアプローチを採用する．

本研究のビュー（View）とは，データセットに付帯するメタデータ群から抽出された特定の特徴空間を指す．言い換えれば,「ある観点から観測されたデータセットに関する情報」をである．また，実験ではデータセットに Meta Kaggle データセット（4.1 節にて後述）を用いるため，以降のビューの作成については Meta Kaggle データセットの構造をもとに説明する．まず，本研究ではメタデータの特徴や性質を踏まえ，メタデータをタグビュー，テキストビュー，行動ビューの 3 種類に分けて用いる．

タグビューはメタデータの `Tags` の項目を利用する．Tags には，データセットの主たる分野，データ形式，代表的な利用シーンが記述されている．まず，タグ文字列を小文字化し，空白の削除・分割などで正規化する．そして，全データセットでの出現頻度から代表的なタグ集合を抽出する．同一または近しい概念を表すタグ（computer science, tabular, image など）を共有するデータセットペアは，研究領域やデータタイプ，データの用途が近いと見なすことができる．タグビューは「大まかに同種のデータセットか」ということを把握するための情報となるが，概念の粒度が統一されていなく，粗いものもあり，データ提供者の主観的に左右されやすい．

テキストビューは `Title` や `Description` などの自然言語による自由記述の項目で構成される．これらはタグより詳細な情報を有し，内容・構造・利用方法などが記述される．本手法ではこれらを 1 文書として結合し，前処理と分かち書きを行った後，テキストビューの特徴空間を構築する．テキストビューは「内容が似ているか」,「同じタスクに利用できるか」といった詳細な意味や内容の類似性が計算でき，タグビューの不十分さを補完する一方で, 長文・ノイズ・冗長記述も多いため，他のビューとの併用が重要となる．

ユーザ行動ビューは，作成者・組織 ID，閲覧数，ダウンロード数，投票数，利用頻度，利用履歴の時系列などで構成される．同一ユーザによって作成・管理されているデータセットや，利用パターンが類似するデータセットは,「利用方法が似ている」と見なすことができる．そのため，このビューは「誰がどのように当該データセットを利用しているのか」というユーザの類似性を捉え，データセットの内容や他のビューでは得られない機能的・利用文脈的な近さを反映することができる．

3.2 タグ・テキスト二部グラフの構築

3.2.1 タグビュー（Tag View）の二部グラフ構築

タグビューでは，まずデータセット集合（D）とタグ集合（T）から二部グラフ D–T とを生成する．はじめに，タグ文字列に対して小文字化，空白の除去，分割などの前処理を行う．続いて，全データセットにおける出現頻度（w_{tag}）を計算し，10 件以上のデータセットに出現するタグのみを残す．そして，各データセットとそこに含まれるタグの組を三つ組 $(d, t, 1)$ として取り出す．ここで 3 番目の要素 「1」 は,「そのタグが当該データセットのメタデータに現れた場合，対応する辺の重みを 1 とする」ことを表す．これらの三つ組を集約して，$|D| \times |T|$ の疎行列 D–T を得る．

3.2.2 テキストビュー（Text View）の二部グラフ構築

テキストビューでは，はじめに，テキスト形式で提供されているメタデータの項目である `Title`，`Subtitle`，`Description` を連結し，各データセットのテキストのリストを作成する．続いて，小文字化や正規表現を用いた前処理を行った後，トークン列の生成を行う．そして，トークンの文書頻度を算出したうえで，少なくとも 200 データセットに出現し，全データセットの 50% 以下にしか出現しない語のみを残す．最終的に，データ（D）とトークン（W），その出現頻度（w_{word}）の 3 つ組を得，$|D| \times |W|$ の疎行列 D–W を構築する．これにより，テキスト情報を自然言語由来の意味情報に符号化し，タグビューと並列に扱えるグラフの構造であるテキストビューに変換する．

3.3 ランダムウォークによる文生成

本節では，タグビュー（D–T）およびテキストビュー（D–W）に対して，データセット間の高次の共起構造を抽出するためのシーケンス生成手法を説明する．この目的は，二部グラフに埋め込まれたタグと語を介した間接的な関係性をシーケンスとして取り出し，SGNS による表現学習のコーパスとして活用するためである．二部グラフはデータセットとメタ情報間の多対多関係を表現したものであるが，データセット同士の距離は明示的に表現されていない．そこでランダムウォークを用いてデータセット同士の関係を系列情報として抽出し，埋め込みを行うことで距離を計算可能にする．

本手法のランダムウォークでは，D–T および D–W はいずれも「データセット→メタ情報→データセット→...（$d_0 \rightarrow t_1 \rightarrow d_1 \rightarrow t_2 \rightarrow d_2 \rightarrow \ldots$）」の交互遷移のみを許容するため，型制約付きランダムウォー

ク (Type-Constrained Random Walk) を採用する．これにより，直接の共通タグや共通語を持たない場合でも，中間ノードを介して意味的に近いデータセット同士が系列内で近接するようになる．これは DeepWalk や node2vec が一般のグラフで実現している高次近傍の共起抽出を二部グラフに適用したものであり，メタデータ由来のデータセットの意味構造を統計的に取り出すための処理である．

タグビューでは，TF–IDF および PPMI による重み付けの後，行方向に正規化を行うことで遷移確率行列 $P_{\mathrm{D}\to\mathrm{T}}$ および $P_{\mathrm{T}\to\mathrm{D}}$ を構成する．これにより，頻出タグの過度な影響を抑えつつ，タグの情報量を反映した確率的遷移が得られる．ランダムウォークはこれらの遷移行列を交互に適用し，データセットからタグへ，タグからデータセットへと遷移することで文を生成する．最終的にはデータセットノードのみを抽出し，共通するタグを有する確率の高いデータセット同士が近くに現れやすいコーパスを形成する．

テキストビューのランダムウォークもタグと同様であるが，中間ノードとして語を利用する点が異なる．語彙はタグと比較してはるかに多様であるため，BM25 重みに基づく遷移確率を設計することで，文書固有の特徴語を適切に重み付けした類似性を抽出する．語彙の共有はタグの共有よりも粒度の細かいデータセット同士の意味的関連性を有しているため，テキストビュー由来の文は，データセット間の潜在的トピックや内容構造をより詳細に表現できる．

こうして生成されたシーケンスは，タグビューとテキストビューで異なる性質を持つ．タグビューは主にデータセットの領域横断的，カテゴリー的な類似性を，テキストビューは文脈的，データの内容に踏み込んだ類似性をそれぞれ捉えることができる．両者は互いに補完的であり，両方を SGNS に入力することで，タグでは粗すぎる情報，テキストでは冗長になりがちな情報の両方をバランスよく抽出できることが期待できる．

3.4 SGNS によるビュー別データセット埋め込み

本節では，前節で生成したシーケンスを用いて，タグビューおよびテキストビューに対して SGNS を適用し，各データセットの埋め込み表現を学習する方法を述べる．SGNS を用いる理由は，シーケンスに潜む高次の共起構造を低次元の連続表現に圧縮できる点にある．SGNS は中心ノードとその周囲の文脈ノードの共起を最大化しつつ，負例サンプリングによりノイズとなる関係の出現を抑制するため，二部グラフ由来の複雑な構造パターンを柔軟に学習できる．

まず，ランダムウォークによって得られたデータセット列 $[d_0, \ldots, d_{L-1}]$ に対し，スライディングウィンドウを用いて，各中心ノードの前後 w ステップ以内に現れるノードを文脈ノードとすることで正例ペア $(\mathrm{center}, \mathrm{context})$ として取り出す．一方，ウィンドウの中に現れないノードを，シーケンス中の出現頻度に基づいて構成した分布からサンプリングし，負例ノードととして取り出す．高頻度ノードに過度に偏らないよう，頻度の 3/4 乗に比例する平滑化分布を用いる．これにより，本来共起しにくいノードペアを明示的に負例として与えることができる．

続いて，各データセットノートに対して「入力ベクトル」と「出力ベクトル」の 2 種類の埋め込みを保持し，バッチ単位で中心・文脈・負例ノードの ID をまとめて取得する．中心ベクトルと文脈ベクトルの内積が大きくなるように，また中心ベクトルと負例ベクトルの内積が小さくなるように損失を計算し，埋め込み行列を同時に更新する．実装においてはベクトル化した `logsigmoid` により一括計算し，タグビューから得たシーケンスとテキストビューから得たシーケンスに対してそれぞれ独立に SGNS を学習することで，タグ共起に基づく埋め込み Z_{tag} と，説明文共起に基づく埋め込み Z_{text} を得る．どちらも同じデータセット集合を行方向に共有するが，表現している意味的側面はビューごとに異なる．

実験で用いる Meta Kaggle データでは，ノード数（データセット数）が数十万，ランダムウォークで生成したシーケンスが数百万規模となるため，学習には膨大な計算量を要する．本研究では，PyTorch DistributedDataParallel を用いたデータ並列学習および混合精度訓練を併用し，ウォーク列を分割して各プロセスに割り当てることで，大規模コーパスに対しても効率的な学習を実現した．学習後は入力埋め込み行列 E_{in} をデータセットの最終表現として採用し，正規化したうえで近似最近傍探索に入力することで，ビューごとのデータセット類似グラフを構築する．

3.5 行動ビュー（Behavior View）の構築

行動ビューは，データセットそのものの内容やデータセットの内容について記述されたメタデータではなく，利用者の行動パターンや作成者の属性を通じて得られる関係性を捉えるビューである．具体的には，(1) データセットの作成者，組織 ID に基づく共属関係，および (2) 閲覧数，ダウンロード数，投票数，利用回数，経過日数などの利用ログに基づく行動類似度を扱う．これらは，実データ内容や一般的なメタデータとは異なる「誰が，どのような目的で利用しているか」というデータセットの機能的，文脈的な側面を反映しており，タグやテキストでは捉えられない重要な情報源となる．

行動ビューの構築には，まず共属関係グラフと利用行動グラフを別々に構成する．共属関係グラフでは，同一ユーザ（または組織）が作成したデータセット同士を強く連結するように重み付けする．一方，利用行動グラフは，閲覧，ダウンロード，投票，利用といった行動系列に基づく共通パターンを反映し，データセット間の利用パターンの類似性を表す．

続いて，行ごとの特徴強度（ノルム・分散など）を指標とし，各行に対して「どのシグナルをどれだけ信頼するか」を示す重みを算出する．すなわちデータセットごとに，共属情報と利用行動情報の寄与度を判断し，重み付き加重和を取ることで統合した行動ビューを作成する．その後，冗長なエッジを抑制するための行内 top-K の取得と行正規化を行い，最終的に行動ビューのデータセット類似グラフを得る．

3.6 3 つのビューの融合

最後に，タグビュー，テキストビュー，行動ビューの 3 つの類似グラフを統合するマルチビュー融合を行う．提案手法の特徴は，各データセットごとに 3 つのビューの信頼度を推定し，ビュー間で重みを動的に調整する適応的融合（Adaptive Fusion）を採用している点にある．このとき，まず Fused3-RA 手法で 3 つのビューのみから構築した類似度行列と，Fused3-RRF 手法にてタグ・説明文・作成者に基づいて候補近傍を再スコアリングした類似度行列を得る．最後に，両者の長所を合わせるために，Fused3-Blend 手法（$S_{\text{blend}} = (1-\eta)S_{\text{RA}} + \eta S_{\text{RR}}$ によって計算）を用い，ビュー特性の差異やスケールの不一致を吸収しつつ統合類似度行列を得る．

この融合行列は，各データセットに対して「タグが示す主題情報」,「テキストが表す意味情報」,「行動が示す利用文脈・機能特性の情報」の各ビューを総合的に結びつけたものであり，単一ビューでは捉えきれないデータセットに関するより広い観点に基づいたデータセット同士の類似性評価を可能にする．特に，タグが欠落したデータセットや，説明文が短いデータセットについては，行動ビューが代替情報として機能する．また，一方で行動情報の偏りはタグとテキストビューによって補正される．このような相補的なビューの融合により，ノイズの影響を抑えつつ，強いビューが弱いビューを支える構造を実現できる．

4 実験設定

4.1 データセットと設定

本実験では，メタデータのみを用いた類似度推定を検証するという目的のため，Meta Kaggle データセットを採用した[1]．これは Kaggle が公開しているメタデータの集合であり，データセットの識別子，タグ，タイトルとサブタイトル，概要説明，作成日時に加え，閲覧数，ダウンロード数，投票数，Kernel 利用回数といった利用統計情報を含む．全体で約 52 万件のデータセットのメタデータと，597 種類の異なるタグが存在する．

ランダムウォークから得た各データセット列については，長さ 40 のシーケンスがタグビューで約 200 万件，テキストビューで約 400 万件となった．また，実験では，評価指標との整合性とマルチビュー構造の保持とのバランスが最も良かったことから，代表値として $\eta = 0.3$ を採用した．

4.2 実験手順

元データには「データセット間の真の類似度」（Gold Label）を表す正解データが存在しないため，代替指標として以下の (1) タグ類似度，(2) テキスト類似度，(3) 行動類似度を用いることでデータの類似度を計算する提案手法の評価を行う．

1. **タグ類似度**: 前処理・正規化済みのタグ集合を用いる．2 つのデータセットが共有するタグに対し，IDF に基づく重みを付与して加算し，その値をタグ類似度とする．これにより，頻度が低いが識別力の高い概念が重み付けされ，データセットの主題的な近さを捉える．

2. **テキスト類似度**：Title, Subtitle，Description を連結し，BM25 ベクトルに変換する．2 つの BM25 ベクトルのコサイン類似度を計算し，必要に応じて閾値で二値化して MAP などの指標に用いる．データセットの意味的な類似度を捉える．

3. **クリエイター類似度**：CreatorUserId を用い，同一ユーザであれば 1，それ以外は 0 とする単純な二値の関連度とする（ただし，欠損は –1）．作成者が同一であればデータセットの対象や設計が類似しやすいという点を踏まえた関連指標である．

比較実験では，単一ビュー（タグ PPMI ＋コサイン，テキスト BM25 ＋コサイン類似度，行動特徴のコサイン類似度），単純なランキング融合手法（RRF, CombSUM），そして本研究のマルチビュー融合手法（Fused3-RA, Fused3-RRF, Fused3-Blend）を評価対象とする．最終的な総合スコアは，タグ類似度，テキスト類似度，クリエイター類似度にそれぞれ 0.6 / 0.3 / 0.1 の重みを付与した統合指標とし，検索性能を比較する．

さらにアブレーション解析により，各ビューの寄与度を分析する．具体的には，タグ＋テキストのみで行動

[1] 用いたデータセットの最終更新日は 2025 年 10 月 25 日である．

表 1: Result of Comparison with Baseline

Method	nDCG@20	MAP@20	MRR@20	P@20	R@20
Tag-SGNS	0.0327	0.0883	0.0982	0.0328	0.0001
Text-SGNS	0.0329	0.0879	0.0978	0.0326	0.0000
Tag-BM25-Cos	0.1403	0.1854	0.1948	0.1364	0.0169
Text-PPMI-Cos	0.7721	0.7700	0.7681	**0.7725**	0.0232
Behavior-Eng-Cosine	0.0516	0.1000	0.1186	0.0499	0.0042
Fusion-RRF	0.3799	0.4573	0.6453	0.3694	0.0375
Fusion-CombSUM	0.2490	0.3829	0.4108	0.2359	0.0439
Fused3-RA	0.8746	0.8717	0.8607	0.3217	0.8376
Fused3-Blend-eta0.30(Ours)	**0.9119**	**0.9128**	**0.8960**	0.5898	**0.8429**

ビューを使わないなど，あるビューを除外することで，提案手法であるマルチビュー融合の効果を確認する．

4.3 評価指標の設計

実験では，次の 5 つの評価指標を用いて各手法の性能を比較する．

1. **nDCG@20**：各サンプルについて，上位 20 件の近傍において，タグ共有・テキスト BM25 コサイン類似度が高い・クリエイターが同一などの関連候補がどれだけ上位に並んでいるかを，位置に応じた重みで評価し，正規化した指標である．値が大きいほど，関連データセットが優先的に提示されていることを意味する．
2. **MAP@20**：各サンプルの上位 20 件リスト内で，関連項目が出現した時点の精度を計算し，その平均を全サンプルで平均したもの．値が高いほど，関連データセットがリストの上位に密に出現していることを示す．
3. **MRR@20**：各サンプルについて，上位 20 件内で最初に現れる関連近傍の順位の逆数を取り，それを平均した指標である．値が大きいほど，ユーザが最初に目にする候補が関連データセットである可能性が高いことを表す．
4. **Precision@20**：上位 20 件中に含まれる関連項目の割合．
5. **Recall@20**：上位 20 件に含まれる関連項目数が，全関連項目集合のうちどの程度を占めるかを表す．値が大きいほど，関連データセットを取りこぼさずに網羅できていることを示す．

5 実験結果と考察

5.1 ベースライ方法との比較実験

表 1 は，全手法の統合指標に基づく順位を示しており，本研究の手法 Fused3-Blend-eta0.30 が他の全てを大きく上回っていることを表している．この順位から，いくつかの階層的な傾向を読み取ることができる．第一に，融合手法はほとんどの指標において，総じて単一ビュー手法より優れた性能を有している．第二に，適応的融合（Fused3-RA および改良版 Fused3-Blend）は，単純なランキング融合（RRF, CombSUM）より高い性能を発揮している．第三に，行動ベースの単一ビューとテキストベースの単一ビューがいずれも上位に位置しており，データセット推薦においてはタグだけでなくテキスト情報も同程度に重要であることが分かる．

5.2 アブレーション実験

アブレーション実験では，融合手法からビューを個別に除去し，対応する単一ビューのベースラインと比較することで，異なるビューを融合することの有効性を検証する．

5.2.1 タグビューの検証

表 2 より，タグビューのみを用いる Tag-SGNS の nDCG@20 は 0.0300 と低く，タグ共有だけではデータセット類似性を十分に表現できないことが分かる．これは，SGNS が共起統計に依存する一方で，異なるタグ数が 597 個と少なく，分布も偏っており，多くのタグペアで共起回数が極めて低いためである．その結果，意味的な関連とノイズの切り分けが難しく，得られた埋め込みベクトルの類似度表現能力が弱くなってしまっていると考えられる．

表 2: Result of View-Ablation Experiments

Method	nDCG@20	MAP@20	MRR@20	P@20	R@20
Tag-SGNS	0.0300	0.0804	0.0894	0.0299	0.0000
Text-SGNS	0.0003	0.0011	0.0011	0.0003	0.0000
Text-BM25-Cos	0.1425	0.1675	0.1514	0.1400	0.0409
Behavior-Eng-Cosine	0.0081	0.0143	0.0146	0.0076	0.0017
(Behavior-Similarity-only)	0.8654	0.8780	0.8775	0.3420	0.7953
Fusion-RRF (Tag + Text)	0.0277	0.0435	0.0491	0.0205	0.0161
Fused3-RA (Tag + Text)	0.8330	0.8457	0.8370	0.3132	0.7890
Fused3-Blend (Tag + Text)	**0.8944**	**0.8503**	**0.8395**	**0.3199**	**0.8000**
Fusion-RRF (Tag + Behavior)	0.1514	0.2392	0.2724	0.1368	0.0356
Fused3-RA (Tag + Behavior)	0.4173	0.5855	**0.5998**	0.2290	0.3035
Fused3-Blend (Tag + Behavior)	**0.7197**	**0.6323**	0.4331	**0.3298**	**0.3709**
Fusion-RRF (Text + Behavior)	**0.3591**	**0.4375**	**0.6114**	**0.3482**	**0.0023**
Fused3-RA (Text + Behavior)	0.1097	0.2358	0.2667	0.0951	0.0002
Fused3-Blend (Text + Behavior)	0.0712	0.1874	0.1916	0.1278	0.0004

一方，マルチビュー融合手法はベースラインを大きく上回る．タグビューとテキストビューを融合した Fusion-Blend は nDCG@20 を 0.8330 まで向上させ，Tag-SGNS 比で約 30 倍となり，他の指標でも大幅な改善が見られる．タグビューと行動ビューを組み合わせた場合も同様に性能が向上している．これは，タグビューが明示的なデータセットの主題を，テキストビューが潜在的な意味を，行動ビューがユーザー行動パターンをそれぞれ捉え，これらを融合することでデータセット間の類似性をより包括的かつ精度高く表現できることを示している．

5.2.2 テキストビューの検証

テキストビューでも顕著な差が見られる（表 2）．テキストビューのみを用いた Text-SGNS の nDCG@20 はほぼ 0 であり，データセット類似性をほとんど捉えられていない．これは，説明テキストが長いうえにノイズも多く，D–W の二部グラフ上の共起構造が疎であったため，SGNS で有効な統計シグナルが得にくかったことが原因と考えられる．テキストの内容を BM25 ベクトル化しコサイン類似度で評価しても，改善には限界があり，単一ビューでは 0.14 程度にとどまった．

一方，テキストビューとタグ・行動ビューを組み合わせた融合手法は大きな性能向上を示した．タグビュー単体のシグナルは弱くても，他ビューと融合することで，タグを持たないデータセットに対してテキストビューと行動ビューが代替的な関連性シグナルを提供し，タグの不足を補えていることが分かる．

さらに，テキストビューとタグビューの融合は大きな性能の向上を示したのに対し，テキストビューと行動ビューのみの融合は相対的に向上の度合いが小さく，データセット類似性の表現にはタグビューの寄与が不可欠であることが確認された．この結果は，メタデータを異なるビューに分離して扱う本稿の設計が妥当であることを裏づけている．

5.2.3 行動ビューの検証

はじめに行動ビューのみに基づいてデータセット類似度を計算すると，性能は相対的に低くなることが分かった．続いて，クリエイター類似度のみを用いてデータセット間の類似度を直接評価し直すと，高スコアが得られ，行動ビュー自体は「ユーザ活動の観点からの類似性」を的確に捉えていることが確認できた．ただし，行動ビューのみではデータセットが有する内容や主題といった側面は表現できないため，タグ・テキストビューと組み合わせることで情報を補完することが重要になる．

表 2 からも，行動ビューとタグ・テキストビューの融合が大きな性能向上に寄与していることが分かる．強いシグナルを持つビューが弱いビューを補完し，ビュー間の寄与を動的に調整することで全体性能が底上げされるという，マルチビュー融合の本質的な意味であり，単純な平均ではなく，ビュー間の寄与を調整することが各種指標を同時に押し上げているものと考えられる．

6 おわりに

本研究は，データセットの内容，すなわち実データに直接アクセスすることなくデータセット間の類似度を計算するために，タグ，説明テキスト，ユーザ行動

という 3 種類の異なるメタデータビューによるマルチビュー表現と適応的融合の新しいフレームワークを提案した．このようなメタデータベースのマルチビュー類似度学習は，実際のデータ市場やオープンデータプラットフォームにおけるデータ発見と再利用支援において，いくつかのメリットがある．第一に，データ内容にアクセスせずに類似データセット候補を提示できるため，プライバシーやコンプライアンス上の制約が厳しい環境でも，利用者が候補を絞り込むための検索エンジンとして機能し得る．第二に，異なる性質のメタデータを統合することで，単一のキーワード検索や単一ビューでは見落とされがちな関連データセットや代替データセットを推薦でき，様々な下流タスクを支援できる．第三に，提案手法で得られた類似度グラフは，クラスタリングやトピック分析，データカタログの自動整理などにも利用可能であり，データエコシステムの基盤となることが期待できる．

一方で，本研究にはいくつかの限界も存在する．まず，実験は Kaggle という単一のデータプラットフォームのメタデータに基づいており，ドメインや言語，メタデータスキーマの異なるデータカタログに対しても同様の性能が得られるかどうか検証が必要である．また，評価にはタグ・テキスト・作成者情報から構成した代替指標を用いており，真の利用者満足度やタスク達成度を直接測定しているわけではない．そのため，真の類似度を得るためには，表形式データ向け言語モデル（TaLMs）などを援用した類似度比較も必要となる．さらに，ユーザ行動ビューは多数のデータセットを提供する作成者や人気のデータセットにバイアスを受けやすく，公平性・多様性の観点からの検討も重要である．今後は，より大規模かつ多様な外部データポータルを対象とした実験を行うとともに，オンライン評価やユーザスタディによる実運用下での有効性検証，行動ログのバイアスを緩和する融合戦略や，新たなメタデータ種別を取り込んだ拡張モデルの検討も重要な課題である．

謝辞

本研究は JSPS 科研費（JP25K00153）の助成を受けました．

参考文献

[1] Azcoitia, S. A., Laoutaris, N.: A Survey of Data Marketplaces and Their Business Models, *SIGMOD Record*, Vol. 51, No. 3, pp. 18–29 (2022)

[2] Chen, Z.: Challenges and Progress in Dataset Search, *Proc. 8th Symposium on Future Directions in Information Access*, (2020)

[3] Nargesian, F., et al.: Dataset Discovery in Data Lakes (D3L), *arXiv preprint*, (2020)

[4] Lv, Q., Josephson, W., Wang, Z., Charikar, M., Li, K.: Ferret: A Toolkit for Content-Based Similarity Search of Feature-Rich Data, *Proc. EuroSys 2006*, pp. 317–330 (2006)

[5] Yan, Z., Tang, D., Duan, N., Bao, J., Lv, Y., Zhou, M., Li, Z.: Content-Based Table Retrieval for Web Queries, *Neurocomputing*, Vol. 349, pp. 183–189 (2019)

[6] 作本 猛, 早矢仕 晃章, 坂地 泰紀, 野中 尋史: 類似データセット発見課題における詳細なデータセット分類に基づいた有効性の評価, 言語処理学会 第 *29* 回年次大会発表論文集 (2023)

[7] Ravishankar, T. N., Shriram, R.: Metadata Based Clustering Model for Data Mining, *Journal of Theoretical and Applied Information Technology*, Vol. 67, No. 1, pp. 59–67 (2014)

[8] Bernhauer, D., Nečaský, M., Škoda, P., Klímek, J., Skopal, T.: Open Dataset Discovery Using Context-Enhanced Similarity Search, *Knowledge and Information Systems*, Vol. 64, No. 12, pp. 3265–3291 (2022)

[9] Sakumoto, T., Hayashi, T., Sakaji, H., Nonaka, H.: Metadata-based Clustering and Selection of Metadata Items for Similar Dataset Discovery and Data Combination Tasks, *IEEE Access*, Vol. 12, pp. 40213–40224 (2024)

[10] Kumar, A., Rai, P., Daumé, H. III: Co-regularized Multi-view Spectral Clustering, *Advances in Neural Information Processing Systems 24 (NeurIPS 2011)*, pp. 1413–1421 (2011)

[11] Wang, B., San Lucas, A. G., Shah, N., Man-Child, E., Kantarcioglu, M., et al.: Similarity Network Fusion for Aggregating Data Types on a Genomic Scale, *Nature Methods*, Vol. 11, No. 3, pp. 333–337 (2014)

[12] Gönen, M., Margolin, A. A.: Localized Data Fusion for Kernel k-Means Clustering with Application to Cancer Biology, *Advances in Neural Information Processing Systems 27 (NeurIPS 2014)*, pp. 1305–1313 (2014)