修士学位論文

題目

Binary Relevance とヒューリスティクスを用いたゲームタグ推薦

指導教員

楠本 真二 教授

報告者

玉置 文人

令和 8 年 2 月 2 日

大阪大学 大学院情報科学研究科

コンピュータサイエンス専攻

令和 7 年度 修士学位論文

Binary Relevance とヒューリスティクスを用いたゲームタグ推薦

玉置 文人

内容梗概

Steam や Google Play などのゲームストアでは，開発者はゲームにジャンルや世界観を表すゲームタグを付与する．ゲームタグは検索やレコメンドに使われる情報であり，日々リリースされる豊富なゲームの中から好みのゲームを探すためには欠かせない情報である．しかしながらゲームタグの豊富さや一貫性の確保の難しさから，開発者が適切にゲームタグを付与するのは容易ではない．また機械学習を用いたゲームタグ推薦手法も研究されているが，ゲームタグの持つ不均衡が課題となり全てのゲームタグへの適用は困難である．本研究では開発者によるゲームタグ付与の支援を目的とし，機械学習を用いた全てのゲームタグに適用可能なゲームタグ推薦手法の提案に取り組む．提案手法では Binary Relevance を用いてゲームタグ推薦を単純な二値分類へ分解し，またタグの共起関係に基づくヒューリスティクスを導入し推薦精度を向上させる．本研究では提案手法の推薦性能を確認するため Steam を題材としたゲームタグ推薦を実験した．実験の結果，全体的には提案手法に有意な推薦精度の改善は確認できなかったが，一部のタグやモダリティでは有意な推薦精度の改善が確認できた．今後は個々のゲームタグに合わせたモデル設計や，より効果的なヒューリスティクスの設計が課題となる．

主な用語

ゲームタグ，タグ付与支援，機械学習，マルチラベル分類，Binary Relevance，ヒューリスティクス

目次

図目次

表目次

1　はじめに

Steam[*1]や Google Play[*2]などのゲームストアでは，ゲームの属性を表す要素としてゲームタグが用意されている．ゲーム開発者はゲームリリース時に，そのゲームに適切なゲームタグを選んで付与する．ゲームタグが表す属性は幅広く，Action タグのようなジャンルを表すタグや，Fantasy タグのような世界観を表すタグ，Pixel Graphics タグのような視覚属性を表すタグなど様々な側面のタグが様々な粒度で用意されている．ゲームタグの用途には，付与されたタグをもとにゲームを調べる検索機能や，プレイヤーがよくプレイするゲームと似たタグが付与されたゲームを紹介するレコメンド機能がある．ゲームストアでは日々豊富なゲームが公開されており，その中から好みのゲームを探すにはゲームタグを用いた検索・レコメンド機能が必須である．従って検索機能やレコメンド機能を有意義にするためにも，開発者による適切なゲームタグの付与が求められる．

ゲーム開発者による適切なゲームタグの付与は容易ではない．タグ付与を困難とする要因の 1 つにゲームタグの豊富さが挙げられる．ゲームストアに用意されているゲームタグは種類が多く，例えば Google Play では 160 種類以上，Steam では 400 種類以上用意されている．ゲームタグの豊富さはゲーム内容の表現をより豊かにするメリットがある反面，ゲームタグの全貌把握を困難にもする [1]．また一貫したタグ付けが開発者に求められる点もタグ付けを難しくする要因の 1 つである．ゲームタグが検索やレコメンドで使われる点を考慮するとタグ付けの基準は一貫しているべきであり，開発者が主観的にタグを付けるとプレイヤーの誤解や不適切な検索・レコメンド結果を招いてしまう [2][3]．しかしゲームジャンルや世界観といったゲームの属性のほとんどは明確な定義を持っておらず，開発者の主観を排除しきれない．

ゲームのジャンルやタグの一貫した分類・付与を支援する既存手法として，機械学習を用いた分類・推薦手法が提案されている．Jiang と Zheng[4] はゲームのジャンル分類を目的として，テキストと画像を用いたマルチモーダルなジャンル分類手法を提案している．Jiang と Zheng は研究の中で，ゲームのジャンル分類においてはテキストと画像を単体で用いるよりも 2 つを組み合わせて用いた方が分類精度が向上したと報告している．また Rubei と Di Sipio[3] はジャンルを表すタグとそれに関連するタグの推薦を目的としたゲームタグ推薦手法を提案している．Rubei と Di Sipio は提案手法の中で，主要なジャンルの推薦結果を用いて共起関係を持つ関連タグを推薦するという手法を採り入れて関連タグの推薦を実現している．しかしこれらの既存手法は全てのゲームタグに適用するのは難しいと考えられる．その要因は，ゲームタグの持つ不均衡にある．ゲームタグには，正例より負例の方が圧倒的に多いという個々のタグの不均衡と，タグの間でも出現頻度の差が大きいというタグ間の不均衡という 2 つの不

[*1] https://store.steampowered.com/
[*2] https://play.google.com/

均衡が存在する．既存手法では主要なジャンルやそれに関連するタグの推薦が目的であるため，個々のタグの不均衡は小さく，またタグ間の不均衡も顕著ではない．全てのゲームタグの推薦を実現するには正例が負例に対し極めて少ないタグや，他のタグに比べて出現頻度に大きな差があるタグも扱う必要がある．

そこで本研究ではゲーム開発者のタグ付与支援を目的とし，全てのゲームタグへ適用可能なタグ推薦手法を提案する．提案手法ではマルチラベル分類手法の一つである Binary Relevance を採り入れる．Binary Relevance の導入により個々のタグに特化したモデル学習を行い，個々のゲームタグの推薦精度を向上させる．またゲームタグ推薦の二値分類問題への分解により，タグ間の不均衡による影響をなくし，個々のタグの不均衡への対策を容易にする．加えて提案手法では共起関係に基づくヒューリスティクスも導入する．Binary Relevance で考慮できないタグ間の共起関係を，ヒューリスティクスによって補いさらに推薦精度の向上を狙う．

提案手法のゲームタグ推薦性能を評価するため，Steam のデータセットにある 29,022 件のゲームを題材として実験を行った．実験では題材となる Platformer タグと，不均衡の度合いや関連するモダリティ，Platformer タグとの共起関係が異なる 5 種類のタグの計 6 種類のタグを用いて，Binary Relevance と共起に基づくヒューリスティクスの性能を評価した．Binary Relevance と単一モデルの推薦精度を比較したところ，一部のタグやモダリティでは Binary Relevance が単一モデルに比べ有意な推薦精度の向上を実現していたが，全体的には推薦精度に有意な差は認められなかった．またヒューリスティクス導入前後の推薦精度を比較したところ，適合率と再現率の一方は改善できたものの，もう一方の指標の悪化を抑えられず結果として有意な推薦精度改善には至らなかった．提案手法の問題点を探るため実験結果を分析したところ，モデルやヒューリスティクスの設計が不適切だった点が原因だったと考えられる．そのため異なるモダリティの採用やモデルの学習方法の改善，より効果的なヒューリスティクスの設計が今後の課題である．

2　準備

2.1　ゲームタグ

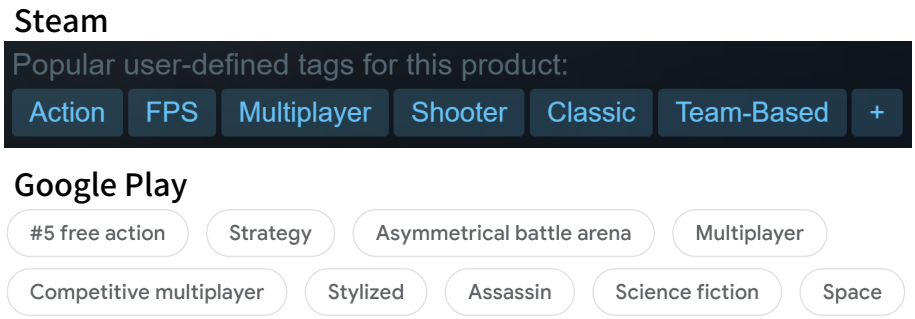


図 1: ゲームタグの例

Steam や Google Play といったゲームストアでは，各ゲームに開発者がジャンルや世界観などを表すゲームタグを付ける．図 1 に Steam と Google Play におけるゲームタグの一例を示す．図 1 にあるように 1 つのゲームにはジャンルや世界観などを表すゲームタグが複数個付与される．

表 1: ゲームタグが表す属性の例

属性	ゲームタグ例
主たるジャンル	Action, Strategy, RPG, Puzzle
詳細なジャンル	FPS, Platformer, Tactical, Battle Royale
限定的なジャンル	Asymmetrical Battle Arena, Metroidvania, City Builder, Souls-Like
ゲームシステム	Team-Based, Side Scroller, PvP, Minigames
世界観・雰囲気	Classic, Space, Retro, Fantasy
見た目・視覚情報	Stylized, Pixel Graphics, VR, Anime
主要な登場要素	Dragon, Train, Ninja, Gun
プレイヤー情報	Competitive Multiplayer, Single Player, Co-op, Massively Multiplayer

ゲームタグが表す属性は多岐にわたる．表 1 にゲームタグが表す主な属性とゲームタグの例を示す．表 1 にあるようにゲームタグが表す情報はゲームのジャンルだけではない．どんな世界観を舞台にしたゲームなのか，どのように世界やキャラクターが描画・表現されるのか，そのゲームで鍵となる要素は何か，プレイヤー同士の関係は敵か味方か，などの情報もゲームタグで表される．また一つの属性の中でも様々な粒度のゲームタグが用意されており，特にゲームジャンルを表すタグは豊富に用意され

ている．例えば Steam ではアクションゲームを表す Action タグが用意されている．しかし Steam ではそれに加えて，アクションゲームの中の 1 ジャンルであるプラットフォームゲームを表す Platformer タグが別途用意されている．さらにはプラットフォームゲームの中の 1 ジャンルであるメトロイドヴァニアゲームを表す Metroidvania タグも用意されている．このようにゲームストアでは，ゲームタグだけでゲームの概要が表現できるほど様々な粒度のゲームタグが幅広く用意されている．

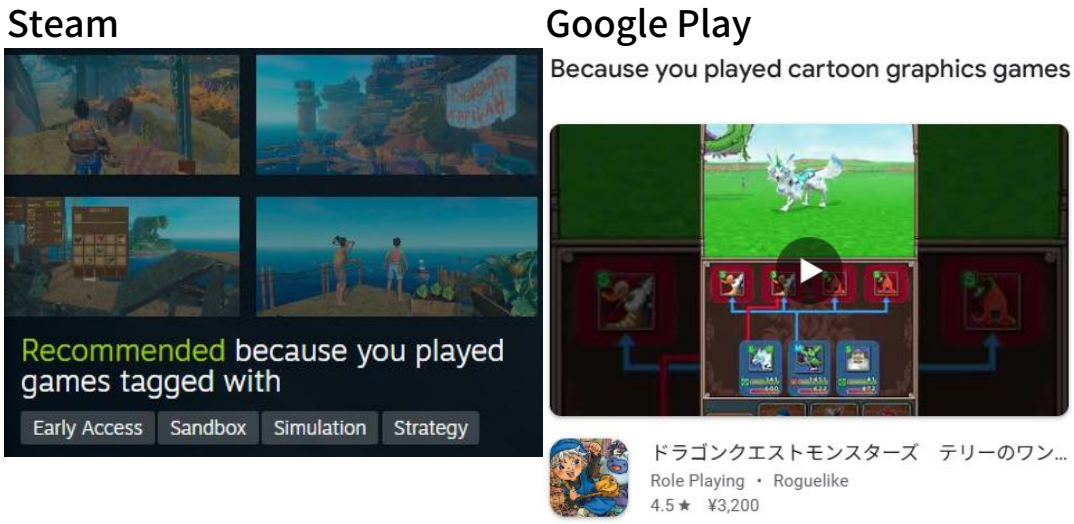


図 2: ゲームタグを利用したレコメンド

ゲームタグの主な用途としては検索とレコメンドの 2 つが挙げられる．多くのゲームストアではゲームタグを利用した検索・絞り込み機能が用意されている．そのためプレイヤーは自身が興味のあるゲームをタグを利用して検索できる．例えばアクションゲームに興味があるなら Action タグが付いたゲームを探したり，ホラーゲームを避けたいなら Horror タグが付いていないゲームに絞り込むといった活用ができる．またゲームストアでは，ゲームタグをプレイヤーの興味関心の分析に利用し，そのプレイヤーに推奨のゲームを提案している．図 2 は Steam と Google Play における，ゲームタグを利用したレコメンドの例である．図 2 のように，ゲームストアではプレイヤーが遊んだゲームに付いているタグを参照し，同様のタグが付いたゲームを推薦している．

ゲームタグは検索やレコメンドで使われるので，適切にタグを付けないとプレイヤーの誤解や不自然な検索・レコメンド結果を招きかねない [2][3]．そのためゲーム開発者には適切なタグ付けが求められる．

2.2 ゲームのタグ付けの難しさ

2.2.1 開発者によるタグ付けの課題

2.1 節で述べたようにゲーム開発者はリリース時に適切にゲームタグを付ける必要があるが，適切なタグ付けは容易ではない．ここではタグ付けを難しくしている要因を大きく 2 つに大別して考える．タグが豊富にある点と，一貫したタグ付けが求められる点である．

ゲームタグが表す属性はジャンルや世界観，プレイスタイルなど広範にわたるが，その分用意されているタグも多い．例えば Google Play では 160 種類以上[*3]，Steam では 400 種類以上[*4]のタグが現時点で用意されており，今後もゲームの多様化に伴いその数は増加すると考えられる．この豊富さはゲームタグによるゲーム内容の表現をより詳細にできるメリットがある一方で，タグの全貌の把握を困難にするデメリットもある．特に 400 種類以上ものタグが用意されている Steam では，その豊富なタグを把握しきれないという声もあがっている [1].

主観を排除しきれず一貫したタグ付けができない点もタグ付けの難しさの 1 つである．ゲームタグはそのゲームストア内全体で使われる情報である．従って有意義な検索やレコメンドを実現するためには，なるべく一貫した基準でタグを付ける必要がある．しかしジャンルや世界観といったゲームの属性のほとんどは明確な定義を持っておらず，開発者の主観を排除しきれない [4][5].

2.2.2 既存手法とその課題

一貫性のあるタグ付与を支援する手法の 1 つに，機械学習を用いた自動推薦が挙げられる．機械学習を用いてジャンルを予測する研究は文学や音楽，映画など幅広い分野で行われており [6][7][8]，ゲーム分野においても機械学習を用いてジャンル予測やタグ推薦を行う手法がいくつか提案されている [3][4]. Jiang と Zheng[4] はゲームの主要なジャンル 15 種類の分類に取り組んでいる．その提案手法には，ゲームの説明文とカバー画像の 2 つのモダリティを用いたマルチモーダルな手法が採り入れられている．また Rubei と Di Sipio[3] は主要なジャンルを表す 7 種類のタグと，それに関連する 100 種類のタグを対象としたタグ推薦システム AURYGA を提案している．その中ではタグの共起関係を考慮するために，関連タグの推薦に主要なジャンルの推薦結果を採り入れるというアイデアを用いている．

しかし既存手法の課題として，ゲームタグが持つ不均衡への対応が不十分である点が挙げられる．マルチラベル分類で扱うデータには大きく，各ラベルの不均衡，ラベル間の不均衡，そしてラベルセットの不均衡の 3 種類の不均衡が存在する [9]．ゲームタグ推薦においては，このうち各ラベルの不均衡とラベル間の不均衡の 2 つが課題となる．各ラベルの不均衡とは，それぞれのラベルの正例/負例の数に関

[*3] https://www.apptweak.com/en/aso-blog/complete-list-of-available-google-play-store-tags
[*4] https://partner.steamgames.com/doc/store/tags

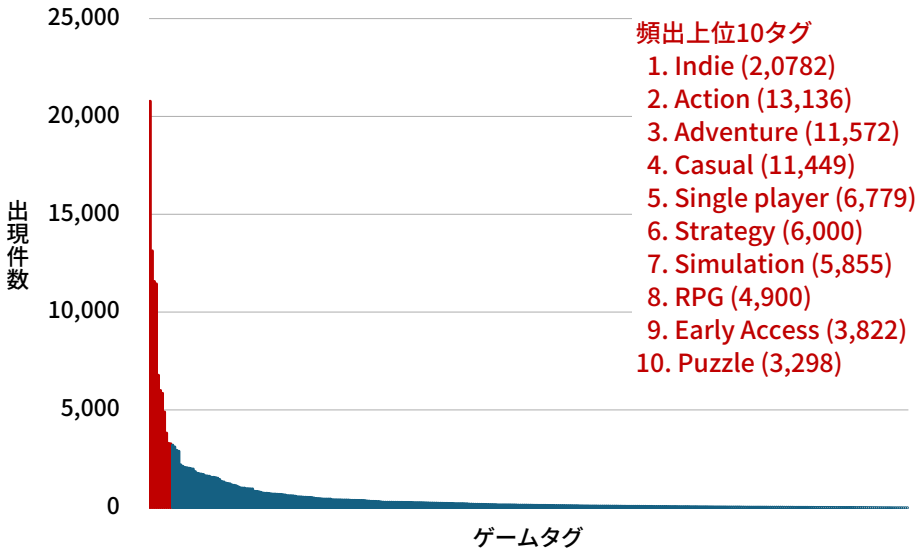


図 3: Steam のゲームタグの出現件数と頻出上位 10 タグ

する不均衡である．一般的にマルチラベル分類のデータは，どのラベルも正例よりも負例の方が圧倒的に多い [10][11]．もう 1 つのラベル間の不均衡とは，各ラベルの出現頻度の差による不均衡である．マルチラベル分類のデータの中には，特定のラベルだけが頻出し，その他のラベルはほとんど出現しないといった偏りを持つデータもある．

ゲームタグには各ラベルの不均衡とラベル間の不均衡の 2 つの不均衡が両方存在する．図 3 は Steam Store Games[*5]というデータセットに含まれる 29,022 タイトルのゲームと 371 種類のゲームタグについて，横軸を各タグ，縦軸をそのタグの出現件数として，左側から出現頻度が多い順に並べて示した図である．図に示す通り，Indie タグや Action タグといった一部の頻出タグは半分近くのゲームに出現する．その一方でほとんどのタグは出現件数が少なく，371 種類のタグの中で出現件数が 10 番目に多い Puzzle タグでも，29,022 件中 3,298 件と全データの 1/10 程度しか出現しない．このような不均衡を持つデータでは希少なタグを学習・推薦できなかったり，逆に頻出するタグを必要以上に推薦してしまう恐れがあるため，2 つの不均衡に対して何らかの策を講じる必要がある[*6]．

しかし既存手法は対象としたジャンルやタグが限られており，不均衡対策が十分でない．Jiang と Zheng[4] の研究は主要なジャンルの分類に焦点を当てたシングルラベル分類の研究である．そのため

[*5] https://www.kaggle.com/datasets/nikdavis/steam-store-games

[*6] Tarekegn ら [9] が述べているもう 1 つの不均衡であるラベルセットの不均衡とは，同時に出現しやすい/しにくいラベルの組の間に生まれる不均衡である．Label Powerset[12][13] のようなラベルを付与された組単位で扱う手法ではラベルセットの不均衡が課題となりやすい．しかしラベルセットの不均衡が存在するということは，言い換えれば共起関係を持つラベルが存在するということであり，マルチラベル分類時の判断材料として活用できる情報になる．実際 Rubei と Di Sipio[3] はラベルセットの不均衡を利用して主要なジャンルタグと共起する関連タグの推薦を行っている．上記の理由から，本研究ではラベルセットの不均衡は課題として取り上げていない．

個々のタグの不均衡とタグ間の不均衡のどちらも顕著ではなく，提案手法にも不均衡対策は取り込まれていなかった．Rubei と Di Sipio[3] の研究はマルチラベル分類の研究であり，個々のタグの不均衡への対策として学習データを絞り込むアンダーサンプリングを採り入れていた．しかし Rubei と Di Sipio の研究も比較的出現頻度の高い関連タグまでを対象とした研究である．そのため，個々のタグの不均衡が過度に強くはなく，タグ間の不均衡もある程度抑えられている．全てのタグへ適用可能な推薦手法を実現するには，個々のタグおよびタグ間の不均衡がより顕著になるタグも扱う必要がある．そのため，既存手法をそのまま全てのタグに拡張するのは理論上可能であっても効果的ではないと考えられる．

3 提案手法

3.1 概要

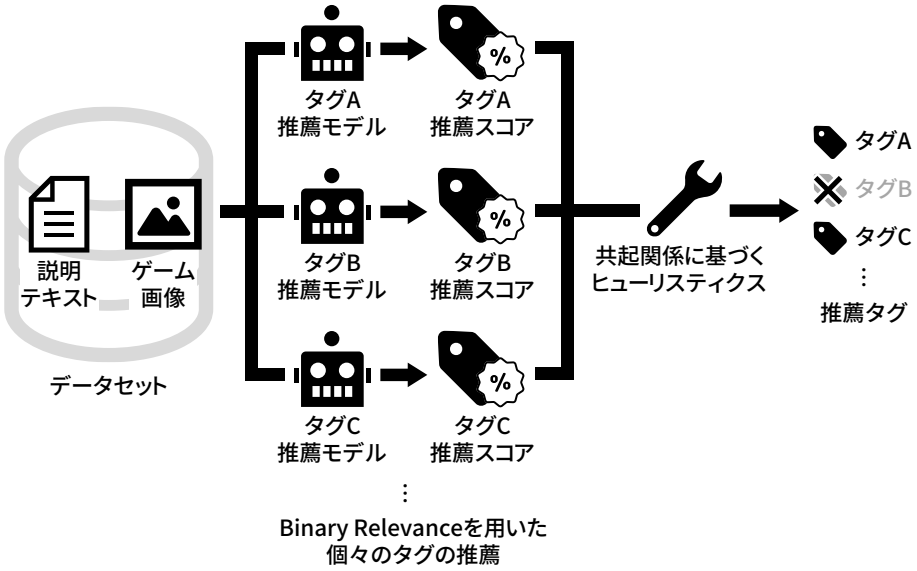


図 4: 提案手法の全体像

本研究の目的は，ゲーム開発者によるゲームリリース時点でのタグ付与の支援である．そのために本研究では，全てのタグに拡張可能なタグ推薦手法を提案する．図 4 に，提案手法の全体像を示す．提案手法ではゲームタグ推薦のために，2 つの手法を採り入れる．個々のタグの推薦に特化する Binary Relevance と，共起関係に基づくヒューリスティクスである．以降では，2 つの手法についてそれぞれ説明する．

3.2 個々のタグの推薦に特化する Binary Relevance

3.2.1 Binary Relevance の概要と狙い

提案手法では Binary Relevance を用いて個々のタグの推薦を行う．Binary Relevance はマルチラベル分類で用いられる手法の 1 つである [12][13]．マルチラベル分類では通常，単一のモデルで全てのタグを分類する．一方 Binary Relevance では，ある 1 つのタグについて二値分類するモデルをラベルの数だけ用意してマルチラベル分類を行う．

Binary Relevance を導入する理由は，個々のタグの推薦精度の向上を狙うためである．Binary Relevance にはマクロ平均で求める指標の最適化に適している [14][15]．マクロ平均は各クラス別に求めた指標の平均であり，各クラスの頻度差に依らず全てのクラスが等しく反映される．対になる概念に

マイクロ平均があるが，マイクロ平均は全てのクラスのデータを一括して扱う平均であり，出現頻度の高いクラスが強く反映される．本研究はタグ間の不均衡が存在する中で全てのタグの推薦精度向上を目標としており，頻出タグに左右されやすいマイクロ平均ではなく，個々のタグの出現頻度に依存しないマクロ平均に基づく最適化が必要である．そのためマクロ平均で求める指標の最適化に適した Binary Relevance は，他のタグに左右されない学習が必要な本研究に適する手法の 1 つだと言える．

さらに Binary Relevance で個々のゲームタグごとにモデルを作成すれば，各ゲームタグに強く結びつく要素を捉えやすくなると考えられる．ゲームタグの中には，Action タグに対する “jump” や Shooter タグに対する銃のような，密接に関わりそうな単語やアイテムを持つタグがある．このようなタグ特有の要素に強く反応する学習ができれば，個々のタグの推薦性能を高められると考えられる [16].

また独立した二値分類問題への分解により不均衡対策がしやすくなるのも，Binary Relevance を導入した理由の 1 つである．2.2.2 節で述べたようにゲームタグ推薦では考慮すべき 2 種類の不均衡があるが，個々のタグを完全に切り離して扱うためタグ間の不均衡について考慮する必要がなくなる．もう 1 つの不均衡である個々のタグの不均衡は依然として残るため対策が必要であるが，他のタグの状態（頻度差，共起関係など）について考慮する必要がないため，従来の不均衡な二値分類で使われる不均衡対策を導入すればよい [14].

3.2.2 Binary Relevance への複数モダリティの組み込み

提案手法では既存手法 [4] を参考に，テキストと画像の 2 つのモダリティを使用する．図 5 に提案手法でのマルチモーダル化の流れを示す．提案手法ではあるタグ X を推薦する際，テキストから推薦するモデル（ 図 5 ①）と画像から推薦するモデル（ 図 5 ②）の 2 つのモデルを用意する．2 つのモデルはそれぞれ独立に学習を行い，学習が完了したらモデルのパラメータは固定する．そして 2 つのモデルが出力する特徴ベクトルを結合して全結合層に流す（ 図 5 ③）．上記の流れで作成されたモデルが，提案手法でタグ X の推薦に使われるマルチモーダルなモデルである．なおこの流れから分かる通り，提案手法では 1 つのタグにつき 2 つのモデルが必要になる．

複数のモダリティを使う理由は，特定のモダリティだけでは捉えられないタグがあるためである．ゲームを表す主な情報としては Jiang と Zheng[4] が注目しているようにテキストと画像の 2 つが挙げられる．しかしこの 2 つのモダリティには表現できる情報に向き不向きがあり，片方だけでは判断できない場合がある．例えば，テキストはゲームシステムを表現しやすい反面，見た目に関する情報は表現しにくい．他方，画像は見た目を表現しやすいが，ゲームシステムは表現しにくい．

複数のモダリティが必要となる例として，図 6 に Steam で公開されている『Level Devil[*7]』というゲームの説明文とヘッダー画像を示す．このゲームには昔風な世界観や見た目であることを表す Retro

[*7] https://store.steampowered.com/app/3242750/Level_Devil/

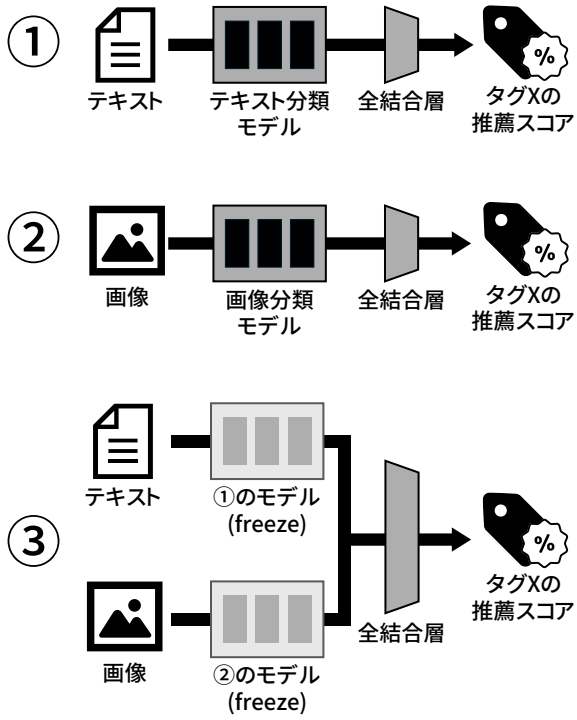


図 5: テキスト分類モデルと画像分類モデルを用いたマルチモーダル化の流れ

というタグが付いている．ヘッダー画像を見るとキャラクターや床のトゲがドット絵で描かれているため，このゲームがドット絵を用いたレトロ風な見た目のゲームであると推測できる．しかし説明文には見た目に関する記述がなく Retro タグが適するかは判断できない．またこのゲームには，ブラックジョークのようなネガティブなネタが含まれることを表す Dark Humor というタグも付いている．説明文を見ると “troll” や “unfair” といった単語が含まれているため，このゲームはプレイヤーを騙す理不尽な要素を楽しむゲームであると推測できる．しかしそういった騙し要素や理不尽な要素はヘッダー画像からは判断できない．このように 1 つのモダリティだけでは捉えられないタグがあるため，提案手法ではゲームを表す 2 つの主なモダリティであるテキストと画像を使う．

3.2.3 Binary Relevance 導入による学習コスト増加

Binary Relevance の問題点の 1 つに，他のマルチラベル分類手法と比べて計算コストが大きくなりやすい点が挙げられる [17]．Binary Relevance は扱う問題が二値分類であるため，個々のモデルは通常のマルチラベル分類と比べ単純である．しかし必要なモデル数がラベル数に比例して増えるため，結果的に計算コストは高くなりやすい．2.2.1 節で述べたようにゲームタグは種類が多いため，提案手法の適用を検討する際は計算資源や時間に注意する必要がある．

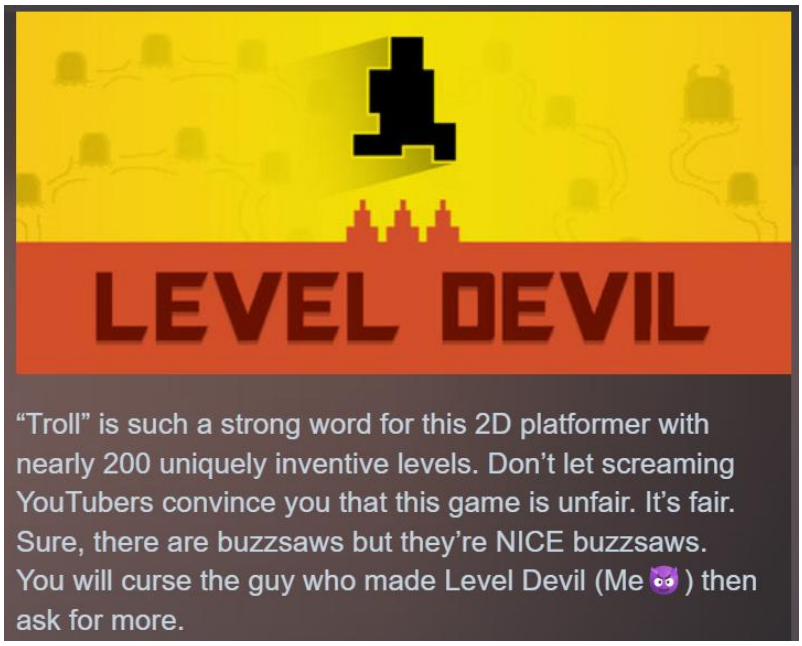


図 6: 1 つのモダリティだけではタグを予測しきれないゲームの例

3.3 共起関係に基づくヒューリスティクス

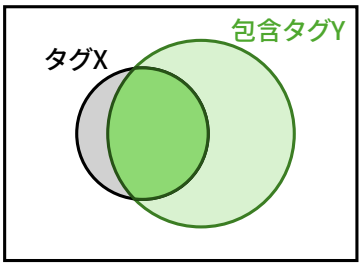
提案手法ではゲームタグの共起関係を考慮するため，共起関係に基づくヒューリスティクスを導入する．マルチラベル分類ではラベルが共起関係を持ちやすい．ラベルが持つ共起関係は分類精度の向上に有用な情報であり，実際マルチラベル分類の研究の中にはこの共起関係を活かして分類精度を高める手法を提案している研究もある [18][19][20]．本研究で扱うゲームタグにおいてもそれは例外ではなく，ゲームタグが持つ共起関係を上手く活用すれば分類精度を向上させられると考えられる．

しかし 3.2.1 節で紹介した Binary Relevance は個々のタグを完全に切り離すため，共起関係を考慮できない [17]．そこで提案手法では共起関係に基づくヒューリスティクスを導入し，個々のタグの推薦スコアに重み付けを行い共起関係を反映する．具体的には，提案手法では包含・被包含・非共起の 3 種類の共起関係を扱う．以降ではそれぞれの共起関係の内容と重み付け方法について説明する．

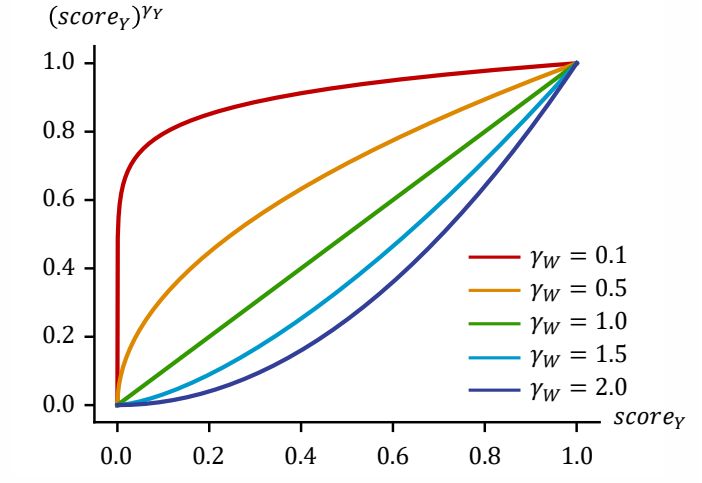
3.3.1 共起関係：包含

あるタグ X に対して，タグ X と共起しやすく，かつタグ X よりも高い頻度で出現するタグ Y がある場合，タグ Y をタグ X の包含タグと呼ぶ．図 7a にタグ X と包含タグ Y の関係を示す．

包含は不適切な推薦を減らし，適合率を上げるための共起関係である．あるゲームに包含タグ Y が該当しないならば，タグ X も該当しづらいと考えられる．そのため誤って推薦してしまっているタグ X



(a) タグ X と包含タグ Y の関係



(b) 包含タグ Y の推薦スコアとタグ X にかかる重み

図 7: 共起関係：包含

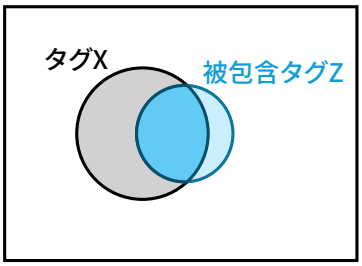
の推薦スコアを下げるのに使える．一方で，包含タグ Y が該当する場合については特に考慮しないこととした．

一例として，Shooter タグと FPS タグが挙げられる．Shooter タグはシューティングゲーム全般を表すタグで，FPS タグはその名の通りシューティングゲームの中の 1 ジャンルである FPS ゲームを表すタグである．そのため Shooter タグは FPS タグと共起しやすく，かつ FPS タグより高い頻度で出現する．従って Shooter タグは FPS タグの包含タグと言える．この場合，Shooter タグが付かないゲームには FPS タグは付きづらいと考えられる．一方で Shooter タグが付くゲームがあったとしても，FPS 以外のシューティングゲームの可能性があるため，FPS タグが付くとは言い切れない．

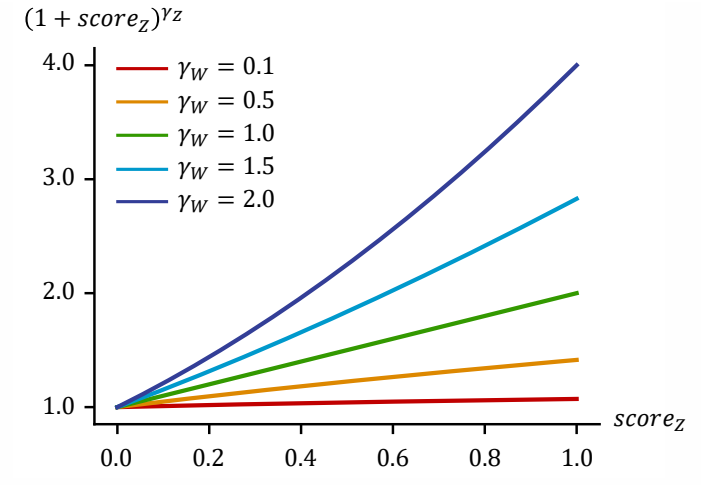
タグ X と包含タグ Y の推薦スコアを $score_X, score_Y (\in [0,1])$ ，包含タグ Y に対応するハイパーパラメータを $\gamma_Y (> 0)$ としたとき，提案手法では以下の式で重み付けを行い調整後のタグ X の推薦スコア $score_{X'}$ を求める．

$$score_{X'} = (score_Y)^{\gamma_Y} score_X$$

包含タグ Y の推薦スコアとタグ X にかかる重みを 図 7b のグラフに示す．包含タグ Y の推薦スコアが高いときは重みが 1 に近く，タグ X の推薦スコアは変化しづらい．一方で包含タグ Y の推薦スコアが低くなると重みが 0 に近づき，タグ X の推薦スコアが大きく減少するようになる．またハイパーパラメータ γ_Y の値が大きいほどタグ X の推薦スコアは減少しやすくなる．



(a) タグ X と被包含タグ Z の関係



(b) 被包含タグ Z の推薦スコアとタグ X にかかる重み

図 8: 共起関係：被包含

3.3.2 共起関係：被包含

あるタグ X に対して，タグ X と共起しやすく，かつタグ X と同程度もしくは低い頻度で出現するタグ Z がある場合，タグ Z をタグ X の被包含タグと呼ぶ．図 8a にタグ X と被包含タグ Z の関係を示す．

被包含は推薦見逃しを減らし，再現率を上げるための共起関係である．あるゲームにタグ X の被包含タグ Z が該当するならば，タグ X も該当しやすいと考えられる．そのため推薦できていない見過ごされたタグ X の推薦スコアを上げるのに使える．一方で，被包含タグ Z が該当しない場合については特に考慮しないこととした．

一例として，包含タグの説明にて例示した Shooter タグと FPS タグを考える．Shooter タグは FPS タグの包含タグであるが，裏を返せば FPS タグは Shooter タグの被包含タグだと言える．この場合，FPS タグが付くゲームには Shooter タグが付きやすいと考えられる．一方で FPS タグが付かないゲームがあったとしても，FPS 以外のシューティングゲームの可能性があるため，Shooter タグが付かないとは言い切れない．

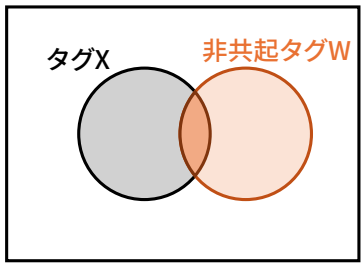
タグ X と被包含タグ Z の推薦スコアを $score_X, score_Z (\in [0,1])$，被包含タグ Z に対応するハイパーパラメータを $\gamma_Z (> 0)$ としたとき，提案手法では以下の式で重み付けを行い調整後のタグ X の推薦スコア $score_{X'}$ を求める．

$$score_{X'} = (1 + score_Z)^{\gamma_Z} score_X$$

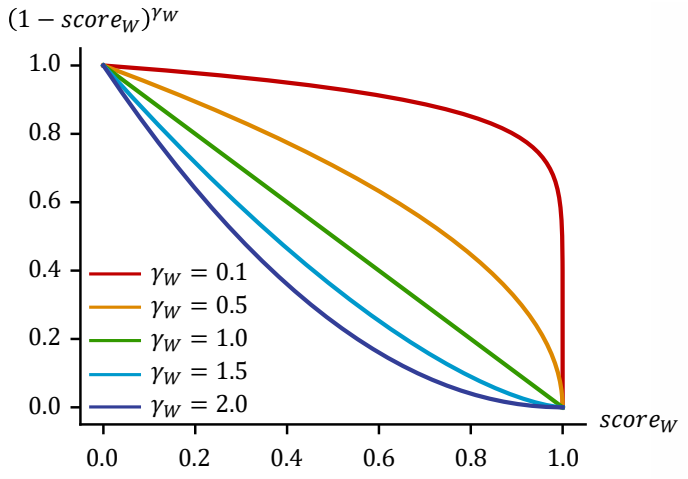
被包含タグ Z の推薦スコアとタグ X にかかる重みを 図 8b のグラフに示す．被包含タグ Z の推薦スコアが低いときは重みが 1 に近く，タグ X の推薦スコアは変化しづらい．一方で被包含タグ Z の推薦ス

コアが高くなると重みが大きくなり，タグ X の推薦スコアが大きく増加するようになる．またハイパーパラメータ γ_Z の値が大きいほどタグ X の推薦スコアが増加しやすくなる．なお被包含タグによる調整のみ，調整後のスコア $score_{X'}$ が 1 を超える場合がある．

3.3.3 非共起



(a) タグ X と非共起タグ W の関係



(b) 非共起タグ W の推薦スコアとタグ X にかかる重み

図 9: 共起関係：非共起

あるタグ X に対して，タグ X と共起しにくいタグ W がある場合，タグ W をタグ X の非共起タグと呼ぶ．図 9a にタグ X と非共起タグ W の関係を示す．

非共起は包含と同様に不適切な推薦を減らし，適合率を上げるための共起関係である．あるゲームにタグ X の非共起タグ W が該当するならば，タグ X は該当しづらいと考えられる．そのため，誤って推薦してしまっているタグ X の推薦スコアを下げるのに使える．一方で，非共起タグ W が該当しない場合については特に考慮しないこととした．

一例として，Horror タグと Family Friendly タグが挙げられる．Horror タグはホラー要素が含まれるゲームに付くタグなのに対し，Family Friendly タグは小さな子どもでも安心して遊べるような家族向けのゲームに付くタグである．そのため Horror タグと Family Friendly タグは共起しにくいので，Family Friendly タグは Horror タグの非共起タグと言える．この場合，Family Friendly タグが付く家族向けゲームには Horror タグは付きづらいと考えられる．逆に Family Friendly タグが付かないゲームがあったとしても，ホラー以外の要素で家族向けでない可能性があるため，Horror タグが付くとは言い切れない．

タグ X と非共起タグ W の推薦スコアを $score_X, score_W (\in [0,1])$ ，非共起タグ W に対応するハイ

パーパラメータを $\gamma_W(>0)$ としたとき，提案手法では以下の式で重み付けを行い調整後のタグ X の推薦スコア $score_{X'}$ を求める．

$$score_{X'} = (1 - score_W)^{\gamma_W} score_X$$

非共起タグ W の推薦スコアとタグ X にかかる重みを 図 9b のグラフに示す．非共起タグ W の推薦スコアが低いときは重みが 1 に近く，タグ X の推薦スコアは変化しづらい．一方で非共起タグ W の推薦スコアが高くなると重みが 0 に近づき，タグ X の推薦スコアが大きく減少するようになる．またハイパーパラメータ γ_W の値が大きいほどタグ X の推薦スコアが減少しやすくなる．

3.4 Classifier Chains と COCOA の採用に対する検討

Binary Relevance のように各ラベル毎に学習しつつ Binary Relevance で考慮できなかった共起関係も学習する手法として，Classifier Chains[21] と COCOA (cross-coupling aggregation)[11] がある．Classifier Chains は Binary Relevance と同様にラベルの数だけ二値分類モデルを構築する手法である．一方で Binary Relevance とは違い Classifier Chains ではあるラベルに対するモデル学習時に，既に学習が完了したモデルの出力も加える．これにより，既に学習が完了したモデルに対応するラベルとの共起関係も反映される．しかし本研究では Classifier Chains は採用しないこととした．Classifier Chains はその性質上，学習できる共起がモデル学習の順番に依存しており，かつ一方向の共起しか学習できない．順序依存を解消するために様々な順序でモデル学習を行う Ensembles of Classifier Chains[21] という手法も存在するが，Ensembles of Classifier Chains は反映させる共起関係に応じた分だけモデルを構築する必要がある．従って本研究で扱うようなラベルが多く共起関係も複雑な分類タスクだと必要なモデルの数が非常に多くなる恐れがある [21]．またラベルの数が多く共起関係も複雑なマルチラベル分類タスクでは Ensemble of Classifier Chains の分類精度が十分向上せず，結果として Binary Relevance と同程度かそれ以下の分類精度になると言われている [15]．上記の点を踏まえ，本研究では豊富なゲームタグにおける片方向/双方向両方の共起関係を反映させるために，Classifier Chains と Ensemble of Classifier Chains は採用しないこととした．

もう 1 つの COCOA も Binary Relevance 同様に各ラベルに応じたモデルを作成する手法である．一方 COCOA では共起関係と各ラベルの不均衡も考慮する．COCOA では Binary Relevance で生じる各ラベルの不均衡を軽減するため，あるラベル X を推薦するモデルを作る際に別のラベル Y も利用して「ラベル X が付く」「ラベル X は付かないがラベル Y は付く」「ラベル X もラベル Y も付かない」の三値分類を行うモデルを作る．この三値分類への分解により，各ラベルの不均衡の原因である負例が 2 つのクラスに分解されるので，比較的均衡な 3 クラスの分類問題に変換できる．またラベル Y の分類がラベル X の分類に反映されるため，ラベル X とラベル Y の共起関係も学習できる．しかし COCOA も Classifier Chains ほどではないが必要なモデル数が増えやすい．COCOA では 2 つのラベルの組み

合わせごとにモデルを作る必要があるので，合計でラベル数の二乗に近い数のモデルが必要になる．また本研究で扱うようなほとんどのタグで個々の不均衡が強いデータだと，COCOA を採用して三値分類に分割しても個々のタグの不均衡は十分には解消されないと考えられる．そのため COCOA についても本研究では採用しないこととした．

4 実験

4.1 Research Questions

提案手法がどの程度ゲームタグ推薦の精度を向上できるか検証するために実験を行う．本実験では以下の 2 つの Research Questions (RQ) を定める．

- RQ1: Binary Relevance の採用により推薦精度は向上するか？
- RQ2: 共起関係に基づくヒューリスティクスの導入により推薦精度は向上するか？

RQ1 では提案手法に採り入れた主な手法の 1 つである Binary Relevance の性能を評価する．単一のモデルで全てのタグを分類する通常のマルチモデル分類をベースラインとして，Binary Relevance を用いる提案手法の精度と比較する．なお RQ1 には共起関係に基づくヒューリスティクスを用いた推薦調整は含めない．RQ2 では提案手法に採り入れたもう 1 つの手法である共起関係に基づくヒューリスティクスの性能を評価する．ヒューリスティクス導入前後の推薦精度を比較し，共起関係に基づくヒューリスティクスの有効性を確認する．

4.2 題材：Steam

4.2.1 実験の題材となるゲームストアとデータセット

本研究ではゲームストアの 1 つである Steam を題材として実験を行う．図 10 に Steam のストアページ例として『Counter Strike[*8]』のストアページを示す．図 10 に示すように Steam のストアページにはゲームを表すモダリティとして，ゲームをテキストで表す短文説明や，ゲームを画像で表すヘッダー画像などが掲載されている．これらのテキスト・画像は開発者がゲームリリース前に設定する情報である．本研究では 図 10 に示す短文説明とヘッダー画像を用いたゲームタグ推薦を実験する．

データセットには Steam Store Games[*5] を使用する．Steam Store Games には全部で 29,022 タイトルのゲームデータがあるが，その中から短文説明が英語でないデータや欠損値が含まれるデータは除外した．最終的に残ったデータは 23,642 件であった．この 23,642 件のデータは学習データ，検証データ，テストデータの 3 つに層化分割して使用する．本実験では学習，検証，テストデータの比率はおよそ 8:1:1 にした（学習データが 18,331 件，検証データ 2,684 件，テストデータが 2,627 件）．

[*8] https://store.steampowered.com/app/10/CounterStrike/

図 10: Steam のストアページ例

4.2.2 実験で使用するタグの選定

本実験では学習・検証データを参考に実験に使用するタグを限定する．3.2.3 節で述べた通り，提案手法を用いて全てのタグの推薦を行うには，全てのタグの分だけモデルが必要になる．そこで本実験では簡略化のため，実験で使用するタグを提案手法の評価に必要なタグに限定してモデルを準備する．実験ではまず推薦対象となるタグ X を 1 つに定める．次にタグ X に対して 3 種類の共起関係（包含，被包含，非共起）を持ついくつかのタグを選定する．そして選定したタグ X と共起関係タグそれぞれに対応する推薦モデルを準備する．最後に共起関係タグを推薦する各モデルの推薦結果を用いてタグ X の推薦内容を調整し，調整後のタグ X の推薦精度を評価する．

使用するタグを選定する際，ゲームタグの共起関係はアソシエーション分析で用いられる Confidence と Lift を利用して評価した．データセットの全てのゲームの集合を G，そのうちタグ X, Y が付いてい

るゲームの集合を $G_{\mathrm{X}}, G_{\mathrm{Y}}$ としたとき，タグ X, Y の Confidence と Lift は以下の式で表される．

$$\mathrm{Confidence}(\mathrm{X}, \mathrm{Y}) = \frac{|G_{\mathrm{X}} \cap G_{\mathrm{Y}}|}{|G_{\mathrm{X}}|} \tag{1}$$

$$\mathrm{Lift}(\mathrm{X}, \mathrm{Y}) = \frac{\frac{|G_{\mathrm{X}} \cap G_{\mathrm{Y}}|}{|G_{\mathrm{Y}}|}}{\frac{|G_{\mathrm{X}}|}{|G|}} \tag{2}$$

Confidence は，あるタグが別のタグにどれだけ包含されているかを表す片方向の指標である．一方 Lift は，あるタグと別のタグが互いにどれだけ共起するかを表す双方向の指標である．そこで共起関係のうち片方向の関係である包含は Confidence で，双方向の関係である非共起は Lift で測る．被包含はタグの出現頻度の差によって片方向か双方向かが変わるので，タグの出現頻度に差があるなら Confidence を，タグの出現頻度に差がない（同程度の出現頻度である）なら Lift を用いて評価した．

表 2: 実験で使うゲームタグ

ゲームタグ	共起関係	データ件数	共起関係の指標値
Platformer	基準	1,508	-
Action	包含	9,815	Confidence(Platformer, Action) = 0.761
Pixel Graphics	被包含	1,500	Lift(Platformer, Pixel Graphics) = 4.069
Side Scroller	被包含	458	Confidence(Side Scroller, Platformer) = 0.609
Metroidvania	被包含	186	Confidence(Metroidvania, Platformer) = 0.726
VR	非共起	2,077	Lift(Platformer, Pixel Graphics) = 0.087

表 2 は本実験で使うゲームタグを，各タグが持つ基準タグとの共起関係，学習・検証データに含まれるデータ件数，共起関係の指標値とともに示した表である．以降では本実験で使うタグについて順に説明する．

基準タグ：Platformer　プラットフォームゲーム（アクションゲームの中でも，移動とジャンプを駆使して足場を乗り移ったり障害物を避けて進むゲーム）に付けられるタグである．Platformer タグには包含や被包含，非共起タグがいくつか存在するため，提案手法の有効性評価に適すると考えた．本実験ではこの Platformer タグの推薦精度を向上させられるかを評価する．

包含タグ：Action　アクションゲームに付けられるタグである．付与されている数が多くデータセットの半分近くのデータで付与されており，本実験で扱うタグの中で唯一個々のタグの不均衡がないタグである．Action タグは意味上で Platformer タグの親に当たるタグだと考えられるため，包含タグとし

て選んだ．実際 Confidence の値は 0.761 であり，全てのタグの中で 2 番目に高かった[*9]．

被包含タグ 1：Pixel Graphics　ドット風の見た目のゲームに付けられるタグである．Platformer タグと同程度の出現頻度のタグの中で Lift が高いタグの 1 つがこの Pixel Graphics であった．またゲームのグラフィックに関するタグであり，本実験で用いるタグの中では画像モダリティが効きやすいタグである．そのため，画像モダリティの性能を評価する狙いも含め採用した．

被包含タグ 2：Side Scroller　プレイヤーの移動などに応じて画面がスクロールしていくゲームに付けられるタグである．データ件数は 458 件と少ないが Confidence は高いため，Platformer タグに包含されやすいタグと言える．

被包含タグ 3：Metroidvania　メトロイドヴァニアゲーム（アイテムや武器などを探索・獲得しながら進めていくアクションゲーム）に付与されるタグである．データ件数は Side Scroller タグよりさらに少なく全ての学習・検証データの中の 1/100 程度のゲームにしか付与されていないが，Platformer タグとの Confidence は全てのタグの中で 2 番目に高い[*10]．意味が限定的であるため特定の要素との結びつきが強いと考えられるので，密接に関わる要素をモダリティから適切に抽出・学習できれば正しく推薦できると考えられる．

非共起タグ：VR　VR を活用したゲームに付けられるタグである．VR タグは Platformer タグと同程度の出現頻度のタグの中で Lift が低いタグの 1 つであり，Platformer タグとは共起しにくいと言える．実際激しい移動やジャンプを伴うプラットフォームゲームとゲーム酔いを起こしやすい VR の相性の悪さを踏まえると，Lift が低いのは妥当だと考えられる．

4.3　Focal Loss による個々のタグの不均衡への対策

本実験では不均衡データセット用の損失関数を用いて個々のタグの不均衡を対策する．2.2.2 節や 3.2.1 節で述べたように個々のタグは正例より負例が圧倒的に多い．そのため何らかの不均衡対策を取らないと負例ばかりを学習してしまい，モデルは何も推薦しない方向に学習を進めてしまう．不均衡対

[*9] Platformer タグを最も包含しているタグは Indie タグで，Confidence(Platformer, Indie) = 0.927 であった．しかし Indie タグはほとんどのゲームに付いているタグ（16,044 件）であり，共起に関係なく出現頻度の高さから Confidence が高くなったと考えられる．またタグの持つ意味を踏まえると Platformer タグと特に共起関係があるとは考え難い．従って想定している共起関係は持っていないと判断し，本実験では扱わなかった．

[*10] Platformer タグに最も包含されているタグは Lara Croft で，Confidence(Lara Croft, Platformer) = 0.727 であった．しかし Lara Croft タグはデータセットが作成された当時の Steam でのみ用いられていたタグである．現在の Steam や他のゲームストアでは登場しない特殊なタグであるため，本実験では扱わなかった．

策にはリサンプリングやコストセンシティブ学習，アンサンブル学習など様々な手法 [22][23][24] があるが，本実験ではコストセンシティブ学習として Focal Loss[25] の導入による不均衡対策する．

Focal Loss は損失関数の 1 つで，特に不均衡なデータを扱う際に使われる損失関数である [26]．あるデータにラベル X が付くか否かを予測する二値分類において，モデルが予測した確率値を $p(\in [0,1]$，1 に近いほどラベル X が付くと予想)，Focal Loss のハイパーパラメータを γ としたとき，Focal Loss は次の式で求められる．

$$\text{Focal Loss} = -(1-p_t)^{\gamma}\log(p_t)$$

$$p_t = \begin{cases} p & \text{（真にラベル X が付いている場合）} \\ 1-p & \text{（真にラベル X が付いていない場合）} \end{cases}$$

式中の $p_t(\in [0,1])$ は予測の正しさを表す度合いと解釈でき，予測が正しいほど 1 に，予想が誤っているほど 0 に近づく．この Focal Loss には損失の係数 $(1-p_t)^{\gamma}$ が予測の難易度に応じて動的に変化するという性質がある．例えば自信を持って正しく予測できる簡単なデータでは，p_t は大きくなり損失の係数 $(1-p_t)^{\gamma}$ は 0 に近づくため，学習に寄与しにくくなる．つまり負例が多い不均衡データでは，簡単に予測できる大部分の負例データは学習でほぼ無視される．一方で自信を持って間違えたりそもそも自信を持って予測できないような難しいデータでは，p_t は小さくなり損失の係数 $(1-p_t)^{\gamma}$ は 1 に近づくため，学習に寄与しやすくなる．つまり負例が多い不均衡データでは，少ない正例データや予測を誤ってしまった負例データなどの予測が難しいデータは学習に強く反映できる．なお本実験では Focal Loss にはハイパーパラメータ γ は，Focal Loss の原論文 [25] で紹介されている $\gamma = 2.0$ に設定して実験した．

4.4 実験設定

本実験ではテキスト分類モデルには BERT の事前学習済みモデル[*11]を，画像分類モデルには ConvNeXt の事前学習済みモデル[*12]を用いた．どちらもエポック数は 100 エポック[*13]，学習率は 1.0×10^{-5} に設定してファインチューニングした．100 エポックの中で，PR-AUC が最も高いエポックのモデルをマルチモーダル化するモデルに使用した．また BERT と ConvNeXt の結合に使う全結合層についても，エポック数は 100 エポック，学習率は 1.0×10^{-5} に設定して学習を行った．結合後のモデルも 100 エポックの中で PR-AUC が最も高いエポックのモデルを評価に使用した．推薦精度の評

[*11] https://huggingface.co/google-bert/bert-base-uncased

[*12] https://huggingface.co/facebook/convnext-tiny-224

[*13] 本実験では 100 エポックでファインチューニングしたが，BERT や ConvNeXt のような事前学習済みモデルのファインチューニングに必要なエポック数はもっと小さく設定しても十分な場合が多い．例えば，BERT の原論文 [27] では 2, 3 エポック，ConvNeXt の原論文 [28] では 30 エポックでファインチューニングが行われている．本実験でも結果的にはどのモデルのファインチューニングにおいても 10 エポック以内で PR-AUC が最大になっており，以降は過学習で PR-AUC が低下していた．

価指標には，個々のタグの適合率，再現率，F1 スコアを用いた．テストデータで推薦の判断に使う推薦スコアの閾値は，検証データでの F1 スコアが最も高くなる閾値を使用した．また各ヒューリスティクスのハイパーパラメータは $0.1 \leq \gamma \leq 2.0$ の範囲で 0.1 刻みで検証し，推薦スコアの閾値と同様に検証データでの F1 スコアが最も高くなるパラメータ値をテストデータで使用した．

4.5 実験結果

4.5.1 RQ1: Binary Relevance の性能

表 3: 各タグの推薦精度（テストデータ・テキストと画像を両方使用）

	適合率		再現率		F1 スコア	
タグ	ML	BR	ML	BR	ML	BR
Platformer	**0.521**	0.510	0.424	**0.477**	0.468	**0.493**
Action	0.715	**0.740**	**0.784**	0.775	0.748	**0.757**
Pixel Graphics	0.367	**0.385**	0.480	**0.530**	0.416	**0.446**
Side Scroller	0.141	**0.191**	**0.567**	0.194	**0.226**	0.193
Metroidvania	0.406	**0.462**	**0.419**	0.387	0.413	**0.421**
VR	**0.907**	0.873	**0.669**	0.666	**0.770**	0.755

最初に Binary Relevance の性能を評価するため，ベースラインである単一モデルによる推薦と Binary Relevance による推薦の精度を比較する．表 3[*14]に通常のマルチラベル分類 (ML) と Binary Relevance (BR) それぞれの各タグの推薦精度を示す．表 3 の F1 スコアを見ると，Platformer タグ，Action タグ，Pixel Graphics タグ，そして Metroidvania タグで Binary Relevance の方が高い精度を記録した．しかし全てのタグで精度が高いわけではなく，Side Scroller タグと VR タグではベースラインを下回る精度であった．また全体的にベースラインとの差は顕著ではなかった．適合率と再現率について見てみると，Side Scroller タグの再現率のみ顕著な差が生じているが，それ以外は F1 スコア同様顕著と言える差は見られなかった．

モダリティ別に提案手法とベースラインの推薦精度の違いを確認するため，テキストのみを用いた推薦と画像のみを用いた推薦の精度も確認する．テキストのみを用いた推薦の精度を 表 4[*14]に，画像のみを用いた推薦の精度を 表 5[*14]に示す．表 4 の F1 スコアを見ると，Pixel Graphics タグと Metroidvania タグの推薦で Binary Relevance がベースラインに対して有意な差を記録した．また 表 5

[*14] 表 3–5 では 3 つの指標（適合率，再現率，F1 スコア）それぞれで，ML と BR のうち値が高い方を太文字で表示している．

表 4: 各タグの推薦精度（テストデータ・テキストのみ使用）

	適合率		再現率		F1 スコア	
タグ	ML	BR	ML	BR	ML	BR
Platformer	0.523	**0.526**	0.471	**0.529**	0.495	**0.528**
Action	0.739	**0.743**	0.728	**0.762**	0.733	**0.753**
Pixel Graphics	0.244	**0.344**	0.302	**0.307**	0.270	**0.325**
Side Scroller	**0.207**	0.148	**0.448**	0.313	**0.283**	0.201
Metroidvania	0.375	**0.520**	0.387	**0.419**	0.381	**0.464**
VR	**0.887**	0.869	0.648	**0.666**	0.748	**0.754**

表 5: 各タグの推薦精度（テストデータ・画像のみ使用）

	適合率		再現率		F1 スコア	
タグ	ML	BR	ML	BR	ML	BR
Platformer	**0.201**	0.179	**0.308**	0.285	**0.243**	0.220
Action	**0.534**	0.509	0.847	**0.890**	**0.655**	0.648
Pixel Graphics	0.435	**0.529**	0.332	**0.361**	0.376	**0.429**
Side Scroller	0.076	0.076	**0.090**	0.075	**0.082**	0.075
Metroidvania[*15]	0.000	**0.012**	0.000	**0.032**	-	**0.017**
VR	0.209	**0.219**	0.536	**0.550**	0.301	**0.314**

の F1 スコアを見てみると，Pixel Graphics タグの推薦で Binary Relevance がベースラインに対して有意な差を記録した．これら 3 組のモダリティとタグの組み合わせについて適合率と再現率を見てみると，いずれも Binary Relevance はベースラインに比べ適合率が大きく向上している．つまりこれら 3 組のモダリティとタグの組み合わせにおいて Binary Relevance は，誤推薦が少ない確実な推薦ができていると言える．

RQ1 への回答

ほとんどのタグとモダリティの組み合わせでは，Binary Relevance と単一モデルの推薦精度に有意な差は見られなかった．しかし一部のタグとモダリティの組み合わせでは Binary Relevance が適合率の高い確実な推薦を行い，単一モデルに対して有意な推薦精度の向上を実現していた．

[*15] ML で画像のみ使用して Metroidvania タグを推薦した結果は，True Negative が 2,561 件，False Positive が 35 件，False Negative が 31 件で，True Positive は 1 件もなかった．

4.5.2 RQ2: 共起関係に基づくヒューリスティクスの性能

表 6: 共起関係に基づくヒューリスティクス導入後の Platformer タグ推薦精度（テストデータ・一部抜粋）

	ハイパーパラメータ							
	γ_{act}	γ_{pg}	γ_{ss}	γ_{mtr}	γ_{vr}	適合率	再現率	F1 スコア
ヒューリスティクスなし	-	-	-	-	-	0.509	0.477	0.493
H1-1: Action のみ	0.1	-	-	-	-	**0.556**	0.436	0.489
H1-2: Pixel Graphics のみ	-	0.1	-	-	-	0.494	**0.506**	**0.500**
H1-3-1: Side Scroller のみ	-	-	1.3	-	-	**0.528**	**0.494**	**0.511**
H1-3-2: Side Scroller のみ	-	-	1.4	-	-	**0.528**	**0.494**	**0.511**
H1-3-3: Side Scroller のみ	-	-	1.5	-	-	**0.525**	**0.488**	**0.506**
H1-4: Metroidvania のみ	-	-	-	0.4	-	**0.563**	0.465	**0.510**
H1-5-1: VR のみ	-	-	-	-	0.1	0.503	**0.500**	**0.502**
H1-5-2: VR のみ	-	-	-	-	0.2	**0.512**	**0.500**	**0.506**
H2: 5 タグ全て	0.2	0.1	0.2	1.4	2.0	**0.568**	0.436	0.493
H3: Action タグ以外の 4 タグ	-	0.1	0.2	0.5	0.4	**0.570**	0.424	0.487

次に共起関係に基づくヒューリスティクスを用いて Platformer タグの推薦スコアを調整した結果を確認する．表 6[*16]にヒューリスティクス導入後の推薦結果を一部抜粋して示す．他のヒューリスティクスの組み合わせで実験した結果は付録 B に記載している．表 6 の H1-1〜H1-5 は各タグのヒューリスティクスを単独で使用した場合の結果である．Side Scroller タグのヒューリスティクスと VR タグのヒューリスティクスは検証データの F1 スコアが最大になるハイパーパラメータが複数あったため，テストデータの結果も検証データの F1 スコアが最大になる各ハイパーパラメータで載せている．H1-1〜H1-5 の F1 スコアを見ると，いくつかのタグでヒューリスティクス導入前からの精度向上が見られるが，その差は RQ1 の時と同様有意な差ではなかった．また H1-1 にある Action タグについては，ヒューリスティクスを導入した結果推薦精度が悪化した．

H1-1〜H1-5 の適合率と再現率について確認すると，一部のヒューリスティクスで想定と異なる推薦調整が行われているのが分かる．H1-1 の Action タグを用いたヒューリスティクスは包含関係を用いたヒューリスティクスであり，適合率を改善するための共起関係タグである．そして H1-1 を見ると再現

[*16] 表 6 では 3 つの指標（適合率，再現率，F1 スコア）それぞれで，ヒューリスティクス導入前よりヒューリスティクス導入後の方が値が大きい場合に太文字で表示している．

率と F1 スコアは悪化しているが，適合率は想定通り改善できている．H1-2〜H1-4 の Pixel Graphics タグ，Side Scroller タグ，Metroidvania タグは被包含関係のタグであり，再現率を改善するための共起関係タグである．実際 Pixel Graphics タグはその狙い通り，適合率は低下したものの再現率の改善には成功している．一方で Side Scroller タグは，再現率は改善できているが，それ以上に適合率が改善される結果となった．また Metroidvania タグでは再現率が低下し，代わりに適合率が大きく改善されている．H1-5 の VR タグは非包含関係のヒューリスティクスであり，包含タグ同様適合率の改善を目的としたヒューリスティクスである．しかし実験結果では適合率の変化は僅かであり，それ以上に再現率が改善されている．このように Side Scroller タグ，Metroidvania タグ，VR タグではそのタグが持つ共起関係の意図に反した推薦調整が起きている．

複数のヒューリスティクスを同時に用いて推薦調整をした場合についても確認する．表 6 の H2 は 5 つのタグのヒューリスティクスを全て用いた場合，H3 は H1-1 でうまくいかなかった Action タグのヒューリスティクス以外の 4 つのタグのヒューリスティクスを用いた場合の結果である．どちらも適合率は単一タグによるヒューリスティクスを使用した場合以上に改善できているが，その分再現率が低下してしまっている．その結果 F1 スコアはヒューリスティクス導入前と同程度あるいはそれ以下となった．

RQ2 への回答

単一のヒューリスティクスについては，ヒューリスティクス導入前後で有意な推薦精度の改善は見られなかった．より具体的には，提案手法で設計したヒューリスティクスはいずれも適合率と再現率の一方は改善できたが，その分もう片方の指標が悪化したため，F1 スコアの有意な改善には至らなかった．複数のタグのヒューリスティクスを同時に適用した場合は適合率の改善と再現率の悪化がより顕著となった．その結果，複数タグのヒューリスティクス導入後の推薦精度はヒューリスティクスを導入しなかった場合と同程度かそれ以下となった．

5　議論

5.1　考察

本節では個々のタグの推薦精度が向上しなかった要因と，ヒューリスティクスによる推薦調整がうまくいかなかった要因をそれぞれ考える．

5.1.1　個々のタグの推薦精度が向上しなかった要因

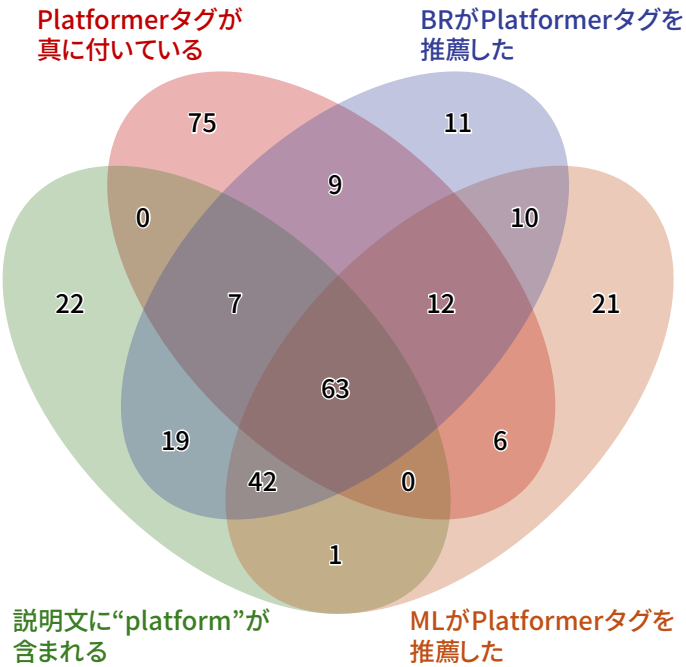


図 11: Platformer タグの説明文と推薦の関係

まず個々のタグの推薦精度が向上しなかった要因としては，Binary Relevance の導入により特定の要素へ過剰に反応してしまった点が挙げられる．3.2.1 節で述べた通り，Binary Relevance の狙いの 1 つとして特定の要素に反応させて推薦精度を高めるという狙いがあった．実際，表す属性・側面が限定的な Platformer タグ，Pixel Graphics タグ，Side Scroller タグ，Metroidvania タグの 4 つのタグにおいてはこの狙いは達成できたと考えている．例えば Platformer タグが付くゲームの説明文には “platform” という単語（“platformer” や “platforming” なども含む）が出現しやすいが，Binary Relevance はベースラインである通常のマルチラベル分類に比べてこの “platform” という単語に反応しやすくなっていた．図 11 は，Platformer タグをテキストのみを用いて推薦した結果を，真に Platformer タグが付いているか，説明文に “platform” を含むか，提案手法の Binary Relevance で Platformer タグを推薦した

か，ベースラインのマルチラベル分類で Platformer タグを推薦したかの 4 つの観点で集計して表した図である．図 11 を見ると，説明文に “platform” が含まれる 154 件のゲームのうち，ベースラインのマルチラベル分類が推薦したのは 106 件だが，Binary Relevance が推薦したのは 131 件と増えている．その結果ベースラインのマルチラベル分類では推薦できなかったゲームを Binary Relevance では 7 件推薦できた反面，同時に 18 件の誤推薦も発生してしまった．他の希少タグについても Binary Relevance は単一モデルに比べて，Side Scroller タグでは “platform” ，Metroidvania タグでは “explore” といった単語に反応しやすい傾向にあった．また Pixel Graphics タグについても，筆者が目視で確認したところドット絵で描かれた要素が映っているヘッダー画像に反応しやすい傾向にあった．このように Binary Relevance の導入により特定の要素に対する反応が過度になり，柔軟な推薦をできなかったのが推薦精度が向上しなかった要因の 1 つだと考えられる．

ただし，特定要素に反応させる手法自体が不適切だとは考えていない．特定要素へ反応するような学習は過度に学習しすぎると前述の通り柔軟に推薦できなくなる問題を生むが，適度に行えば誤推薦の少ない推薦手法の構築に役立つ．実際 3.2.1 節で提案手法がうまくいった例として挙げている通り，テキストを用いた Metroidvania タグ推薦や画像を用いた Pixel Graphics タグ推薦では，ベースラインと比べ Binary Relevance では再現率を維持しつつ適合率を大きく改善できている．適合率が高いタグ推薦は，開発者が付与していないタグの中から付与すべきタグを推薦する手法として活用できる．従って Binary Relevance を用いた特定要素に反応させる手法は，実現したいタグ推薦の内容によっては単一モデルによる推薦よりも適する手法となる可能性がある．

Binary Relevance 以外で個々のタグの推薦精度が向上しなかった要因としては，使用したモダリティやマルチモーダル化の方法が不適切であった点が挙げられる．表 3–5 を見ると，Action タグ以外のタグの推薦精度は，通常のマルチラベル分類か Binary Relevance かに関係なく高いとは言えない精度であった．つまり Binary Relevance 採用以前の問題で，根本的にタグ推薦の精度が良くなかったと言える．3.2.2 節で述べた通り，本研究では既存手法 [4] を参考に説明文とヘッダー画像を Late Fusion で結合する方法を利用した．しかし既存手法はあくまで主要なジャンル分類を目的とした手法であり，主要なジャンル以外を表すタグの推薦にそのまま流用するのは不適切であったと考えられる．全てのゲームタグの推薦を実現するには，説明文やヘッダー画像以外のモダリティや，Late Fusion 以外の結合方法の検討が必要である．

5.1.2 ヒューリスティクスによる推薦調整がうまくいかなかった要因

提案手法で設計した共起関係に基づくヒューリスティクスがうまくいかなかった要因は，どのゲームでも推薦スコアが変化してしまうヒューリスティクスを設計してしまった点にあると考えられる．共起関係を反映するヒューリスティクスには満たすべき特性が 2 つある．共起関係に該当するデータインス

タンスでは推薦スコアを変化させること，そして共起関係に該当しないデータインスタンスでは推薦スコアを変化させないことの 2 つである．前者の共起関係に該当する際の推薦スコア変化は，そもそものヒューリスティクスを導入する目的であり，必要であるのは自明である．しかし共起関係に該当する際に推薦スコアを変化させるだけでは適切な推薦はできない．ゲームタグを推薦する/しないの判断は他のゲームの推薦スコアとの比較により決まる相対的な判断である．そのため共起関係に該当しない場合にも推薦スコアが変化してしまうと，ヒューリスティクスによる推薦調整が意味をなさなくなる．なので共起関係に該当する際に推薦スコアを変えるだけでなく，共起関係に該当しない際にはなるべく推薦スコアを変えないようにする必要もある．

提案手法で設計したヒューリスティクスは共起関係に該当しないデータインスタンスでも推薦スコアが多少変化するヒューリスティクスであった．そのため 2 つの満たすべき特性のうち，共起関係に該当しない際は推薦スコアを変化させるべきでないという特性を満たせていない．共起関係に該当しないデータインスタンスでの推薦スコア調整が悪影響につながった例が，4.5.2 節で述べた共起関係の意図に反した推薦調整である．意図した適合率・再現率の改善と反する推薦調整が行われたのは，ヒューリスティクスの設計が不適切で全体的に推薦スコアが変化してしまい，閾値の再計算が必要になったためである．本来共起関係を利用できるデータインスタンスだけ Platformer タグの推薦スコアを変化させるはずが全てのデータインスタンスで Platformer タグの推薦スコアが動いているため，元の閾値が意味を持たなくなる．すると閾値の再計算が必要となり推薦内容が一新されてしまうため，ヒューリスティクス導入前とは大きく異なる推薦結果になる．その結果ヒューリスティクス導入前後で比較しても推薦内容の変化が共起関係の意図に適するとは限らなくなる．

このように共起関係に該当しないデータインスタンスでも推薦スコアが変化してしまうヒューリスティクスは，ヒューリスティクスの狙いに反する推薦結果をもたらす恐れがある．閾値を変えずに利用するためにも，共起関係を利用する場合にのみ推薦スコアが変化するヒューリスティクスの設計が，推薦精度向上に必要な課題の 1 つだと考えられる．

5.2 今後の課題

本節では今後の課題について，個々のタグの推薦精度に関する課題と，ヒューリスティクスに関する課題の 2 つに分けて考える．

5.2.1 個々のタグの推薦精度に関する課題

個々のタグの推薦精度を高めるための課題の 1 つが，Late Fusion 以外の結合方法の検討である．特に Early Fusion をはじめとした，テキストと画像の特徴を同時に学習できるマルチモーダル手法の検討が挙げられる．本研究の実験では単純な Late Fusion を用いたが，Late Fusion はテキストと画像が

独立して学習されるため，テキストと画像の間の相互作用が学習に反映されない [29]．Early Fusion に対応したモデルの採用などによりテキストと画像の相互作用を適切に考慮できれば，さらなる精度改善が見込める．

次に，使用するモダリティの拡充が挙げられる．4.5.1 節の結果より Pixel Graphics のような一部のタグはテキストだけでは推薦できず，ヘッダー画像などテキスト以外のモダリティが必要となるのが分かった．また Side Scroller のようなテキストでもヘッダー画像でも捉えるのが困難なタグの存在も判明した．そのため全てのタグに推薦範囲を拡張していくには，本研究で用いたテキストとヘッダー画像以外のモダリティも検討する必要がある．例えば同じ画像でもヘッダー画像ではなくゲームのスクリーンショット画像であればゲーム内容を直接反映しているため，ヘッダー画像では表現しきれない詳細なゲームシステムなどが捉えられる可能性がある．またゲームストアではなくゲーム配信においての研究ではあるが，テキスト・画像以外のモダリティとしてゲームのプレイ動画を用いて主要なジャンルを分類する研究もある [30][31]．個々のタグに合わせたモダリティの採用による推薦精度の改善が今後の課題である．

提案手法では様々な属性や粒度のタグを推薦するゲームタグ推薦を独立した扱いやすい二値分類に分解するため，個々のタグを推薦するモデルに対してアイデアや手法を容易に導入できる．本研究では十分に活かせなかった Binary Relevance のカスタマイズ性の高さを活かせば，個々のタグに合わせた推薦モデルを構築して推薦精度を改善できると考えられる．

5.2.2　ヒューリスティクスに関する課題

ヒューリスティクスに関する課題としてまずは，共起関係に該当しない場合に推薦スコアを動かさないヒューリスティクスの設計が挙げられる．5.1.1 節で述べたようにどのデータインスタンスでも推薦スコアが変化するヒューリスティクスでは閾値の再計算が必要になり，その結果推薦内容の解釈も変化する．包含・被包含・非共起の関係が該当する状況でのみ推薦スコアが変動するヒューリスティクスを設計し，想定通りの推薦調整ができるようにすべきである．

次に共起関係の拡充が今後の課題として挙げられる．提案手法では 2 つのタグの間に成立する 3 つの共起関係（包含，被包含，非共起）を考えた．2 つのタグの間に成立する共起関係は基本的にこの 3 つだが，2 つのタグの間ではなく 3 つ以上のタグの間まで拡張すれば他にも共起関係が考えられる．例えば本研究ではメトロイドヴァニアゲームに付く Metroidvania タグを扱ったが，このメトロイドヴァニアゲームはアクションゲームとアドベンチャーゲームの 2 つの側面を持つゲームである．従って，アクションゲームを表す Action タグと，アドベンチャーゲームを表す Adventure タグが両方付与される場合は，この 2 つの側面を持つ Metroidvania タグの付与も検討すべきと考えられる．反対に Action タグと Adventure タグのどちらか片方でも付与されないのであれば，Metroidvania タグも付与されない

と考えられる．このような 3 つ以上のタグに基づく共起関係は提案手法で扱った包含，被包含，非共起のヒューリスティクスだけでは表現しきれないため，新たな共起関係としてヒューリスティクスを設計する必要がある．

タグ間の共起関係以外のヒューリスティクスの設計も今後の課題である．その中の 1 つとして，推薦内容全体に関するヒューリスティクスの導入が挙げられる．ゲームのタグはそのゲームの属性を表すため，適切な範囲内で様々な側面から多角的なタグ付与をすべきだと考えられる．従って主要なジャンルを表すタグのみを付与するのではなく，限定的なジャンルや世界観，視覚属性などを表す希少なタグも付与すべきである．実際に Steam のレコメンド機能では多くのゲームに付与されるような一般的なタグよりもあまり付与されない希少なタグを重視するアルゴリズムが採用されている[*4]ため，希少なタグの適切な付与はセールスにも影響を与えると考えられる．そこで推薦内容全体を踏まえて希少なタグや属性の偏りを調整できるヒューリスティクスを導入すれば，より自然で有意義なタグ推薦ができると考えられる．例えば「限定的なジャンルを表すタグが推薦されていない場合は，限定的なジャンルを表すタグの推薦スコアを上げる」「主要なジャンルを表すタグが 4 つ以上推薦された場合は，その中から推薦スコアが特に高い 3 つのタグに絞る」といった推薦結果全体を踏まえたヒューリスティクスの導入により，そのゲームに付与されるタグのバランスを調整できる可能性がある．

5.2.1 節にて提案手法の強みの 1 つが Binary Relevance の導入による各モデルのカスタマイズ性の高さであると述べたが，ヒューリスティクスについても同様にカスタマイズ性は高い．ヒューリスティクスはモデル学習部分とは独立しており，かつ設計時の制約は少ないため自由なルールを導入できる．重み付けの計算式も損失関数のように微分可能な数式に基づく必要はなく，離散的であったり条件分岐を含んだり，特定のタグにしか成立しないような計算式でもよい．各ゲームストアのゲームタグの付け方に適したヒューリスティクスを設計していくのが今後の課題である．

5.3 妥当性の脅威

本実験では使用するタグを，題材となる Platformer タグと，Platformer タグと共起関係を持つ 5 種類のタグに限定している．そのため真に全てのタグに適用して推薦精度が改善されるかどうかは別途実験し確認する必要がある．また双方向の共起関係についても検証が必要である．例えば本実験では Side Scroller タグの推薦結果を用いた Platformer タグの推薦調整は確認した．しかしその逆の，Platformer タグの推薦結果を用いた Side Scroller タグの推薦調整は確認できていない．

本実験では学習・検証データを固定して実験しているため，汎化性能を十分に評価できていない．特に Binary Relevance を導入した狙いの 1 つである個々のゲームタグ特有の要素の把握は学習データに強く依存するため，学習データを変更すると各モデルが捉える要素が大きく変わる可能性がある．汎化性能を評価するためにも，10 分割交差などで学習・検証データを固定せずに評価する必要がある．

本実験では使用・評価したモデルは BERT と ConvNeXt の 2 つだけある．そのため実験結果は BERT や ConvNeXt の事前学習内容に影響を受けている可能性がある．モデルによる偏りのない評価のためには他のモデルでも同様に実験する必要がある．

Steam 以外のゲームストアのタグでの実験もできていない．2.1 節で述べたゲームストアやゲームタグの特徴自体は Steam 以外のゲームストアやゲームタグでも概ね共通しているため，他のゲームストアに対しても提案手法は適用可能である．しかしゲームストアによって販売されているゲームの内容や用意されているタグの種類，タグの傾向，説明文や画像のフォーマットなどが大きく変わるため，本実験と同様の結果になるとは限らない．

5.4 関連手法：Steam の Tag Wizard

本実験で題材にした Steam では，ゲーム開発者のタグ付けを支援するシステムとして Tag Wizard[*17] が提供されている．Steam が公開している情報から類推すると，Tag Wizard は提案手法と同様にゲームの説明文やタグの共起関係を利用してゲームタグを推薦していると考えられる．しかし内部は非公開であり提案手法との比較が不可能であるため，本研究では扱っていない．

[*17] https://store.steampowered.com/news/group/4145017/view/2246679902968741149

6　おわりに

本研究ではゲーム開発者によるゲームタグの付与支援を目的とし，Binary Relevance とヒューリスティクスを用いたゲームタグ推薦手法を提案した．提案手法のアイデアは，Binary Relevance による個々のタグの推薦精度向上，および共起関係に基づくヒューリスティックな推薦調整による推薦精度向上である．

実験では提案手法の有効性を評価するため，Steam のゲームタグを題材にゲームタグ推薦を行い推薦精度を確認した．まず Binary Relevance と単一のモデルの推薦精度を比較したところ，Binary Relevance による有意な推薦精度の改善は見られなかった．しかし一部のタグとモダリティの組み合わせでは Binary Relevance が単一のモデルに対して有意な推薦精度の改善を示していた．また Binary Relevance を導入した狙いの 1 つである個々のタグ特有の要素の把握は達成できているのが確認できた．そのためモデルの学習方法や使用するモダリティ，実現したい推薦内容によっては Binary Relevance が単一のモデルに対し有意な推薦精度の改善を達成できる可能性がある．次にヒューリスティクス導入前後の推薦精度を比較したところ，こちらも有意な推薦精度の改善は見られなかった．特に提案手法で用いたヒューリスティクスの設計が不適切であったため，適合率と再現率の一方は大きく改善できたが，同時にもう一方が大きく悪化してしまった結果が多く見られた．

今後の課題として，個々のタグの推薦精度を向上させるためのモデル構築方法の検討が挙げられる．本実験の結果より全てのゲームタグ推薦に向けては使用するモダリティやその結合方法を工夫する必要があると分かった．そのため Early Fusion のような複数のモダリティの相互関係を捉えられる結合方法や，説明文やヘッダー画像以外のモダリティを用いた推薦の導入が必要である．また提案手法のヒューリスティクスは共起関係に該当しない場合にも推薦スコアを変動させてしまい，閾値の再計算や共起関係の狙いに反する推薦結果を招いてしまった．各共起関係の狙いを適切に実現できるよう，ヒューリスティクスを再設計する必要がある．提案手法で扱わなかった新たなヒューリスティクスの導入も今後の課題である．ヒューリスティクスによる調整は，制約が少なく自由にアイデアを採り入れやすい．そのため例えば 3 つ以上のタグに関する共起関係や推薦内容全体を踏まえたヒューリスティクスの導入も可能であり，より自然で有意義なタグ推薦が柔軟に実現できると考えられる．モデルの自由度とヒューリスティクスの自由度という提案手法が持つ 2 つの自由度の高さを活かした改善案の検討が今後の課題である．

謝辞

本研究の遂行にあたり，多くの方々にご指導とご支援を賜りました．

楠本真二教授には研究のための環境をご提供いただきました．また中間報告や研究発表の練習では研究内容に対する的確なご指導，ご助言をいただきました．いただいたご指導，ご助言は研究活動を進めるにあたり非常に参考になりました．他にも楠本教授には差し入れとしてみかんをいただくなど研究以外でのご支援もいただきました．心より感謝申し上げます．

柗本真佑准教授には学部 4 年時から 3 年間もの長い期間，研究活動の全ての場面においてご指導いただきました．その丁寧かつ熱心なご指導のおかげでここまで研究活動を進められました．また研究に直接関わる知識や技能以外にも，柗本准教授には人として学ぶべき礼節やソーシャルスキルもご指導いただきました．感謝に堪えません．

楠本研究室事務補佐員の橋本美砂子氏には出張や論文投稿に関わる事務作業でご助力いただきました．事務作業に煩わされずに研究活動を進められたのは，ひとえに橋本氏のご助力のおかげです．さらに橋本氏には研究以外でも，世間話にお付き合いいただいたり，差し入れとしてお弁当を作っていただいたりと多大なサポートをいただきました．深く感謝いたします．

研究室の学生の方々には，研究内外ともに様々な面で支えていただきました．卒業された先輩方は，私にとって学べることが多い模範となる存在で，研究活動ではいつも先輩方を見習わさせていただきました．また TA などの仕事では積極的に前に立ち私を引っ張ってくれる，とても頼りになる存在でした．同期の皆様には研究活動について互いに議論しあうことが多く，その議論を通して様々な知見が得られました．他にも普段から何気なく歓談したり，食事に誘ってもらうなど，私の研究室生活に彩を与えてくれました．後輩の皆様も優秀で，時には私が後輩の皆様から学ぶこともありました．また後輩の皆様の模範になれるよう努力することで，私自身が成長できました．誠にありがとうございました．

最後にここまで学生を続け研究活動を遂行できたのは，家族の支えがあったからです．心よりお礼申し上げます．

参考文献

[1] Reddit, : Dissecting the Steam tag maze: https://www.reddit.com/r/Steam/comments/1ay38r1/dissecting_the_steam_tag_maze/ (Accessed at 2026-01-11).

[2] How To Market A Game, : Steam 101: What Steam tags should I watch out for?: https://howtomarketagame.com/2020/11/18/steam-101-what-steam-tags-should-i-watch-out-for/ (Accessed at 2026-01-11).

[3] Rubei, R. and Di Sipio, C.: AURYGA: A Recommender System for Game Tagging, in *In Proceedings of Italian Information Retrieval Workshop*, pp. 1–7 (2021).

[4] Jiang, Y. and Zheng, L.: Deep learning for video game genre classification, *Multimedia Tools and Applications*, Vol. 82, No. 14, pp. 21085–21099 (2023).

[5] Heintz, S. and Law, E. L.-C.: The Game Genre Map: A Revised Game Classification, in *In Proceedings of Annual Symposium on Computer-Human Interaction in Play*, pp. 175–184 (2015).

[6] Worsham, J. and Kalita, J.: Genre Identification and the Compositional Effect of Genre in Literature, in *In Proceedings of International Conference on Computational Linguistics*, pp. 1963–1973 (2018).

[7] Silla, C. N., Koerich, A. L. and Kaestner, C. A. A.: A Machine Learning Approach to Automatic Music Genre Classification, *Journal of the Brazilian Computer Society*, Vol. 14, No. 3, pp. 7–18 (2008).

[8] Simões, G. S., Wehrmann, J., Barros, R. C. and Ruiz, D. D.: Movie genre classification with Convolutional Neural Networks, in *In Proceedings of International Joint Conference on Neural Networks*, pp. 259–266 (2016).

[9] Tarekegn, A. N., Giacobini, M. and Michalak, K.: A review of methods for imbalanced multi-label classification, *Pattern Recognition*, Vol. 118, No. 107965, pp. 1–12 (2021).

[10] Tahir, M. A., Kittler, J. and Yan, F.: Inverse random under sampling for class imbalance problem and its application to multi-label classification, *Pattern Recognition*, Vol. 45, No. 10, pp. 3738–3750 (2012).

[11] Zhang, M.-L., Li, Y.-K., Yang, H. and Liu, X.-Y.: Towards Class-Imbalance Aware Multi-Label Learning, *IEEE Transactions on Cybernetics*, Vol. 52, No. 6, pp. 4459–4471 (2022).

[12] Boutell, M. R., Luo, J., Shen, X. and Brown, C. M.: Learning multi-label scene classification, *Pattern Recognition*, Vol. 37, No. 9, pp. 1757–1771 (2004).

[13] Tsoumakas, G. and Katakis, I.: Multi-label classification: An overview, *International Journal of Data Warehousing and Mining*, Vol. 3, No. 3, pp. 1–13 (2007).

[14] Zhang, M.-L., Li, Y.-K., Liu, X.-Y. and Geng, X.: Binary relevance for multi-label learning: an overview, *Frontiers of Computer Science*, Vol. 12, No. 2, pp. 191–202 (2018).

[15] Luaces, O., Díez, J., Barranquero, J., del Coz, J. J. and Bahamonde, A.: Binary relevance efficacy for multilabel classification, *Progress in Artificial Intelligence*, Vol. 1, No. 4, pp. 303–313 (2012).

[16] Wang, J. and Gan, K. H.: A Supervised Genre-based Recommendation Model for Game Review, *Pertanika Journal of Science & Technology*, Vol. 33, No. 1 (2025).

[17] Zhang, M.-L. and Zhou, Z.-H.: A Review on Multi-Label Learning Algorithms, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 8, pp. 1819–1837 (2014).

[18] Zhang, M.-L. and Zhang, K.: Multi-label learning by exploiting label dependency, in *In Proceedings of International Conference on Knowledge Discovery and Data Mining*, pp. 999–1008 (2010).

[19] Wu, B., Jia, F., Liu, W., Ghanem, B. and Lyu, S.: Multi-label Learning with Missing Labels Using Mixed Dependency Graphs, *International Journal of Computer Vision*, Vol. 126, No. 8, pp. 875–896 (2018).

[20] Zhu, Y., Kwok, J. T. and Zhou, Z.-H.: Multi-Label Learning with Global and Local Label Correlation, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 30, No. 6, pp. 1081–1094 (2018).

[21] Read, J., Pfahringer, B., Holmes, G. and Frank, E.: Classifier chains for multi-label classification, *Machine Learning*, Vol. 85, pp. 333–359 (2011).

[22] Krawczyk, B.: Learning from imbalanced data: open challenges and future directions, *Progress in Artificial Intelligence*, Vol. 5, No. 4, pp. 221–232 (2016).

[23] Thabtah, F., Hammoud, S., Kamalov, F. and Gonsalves, A.: Data imbalance in classification: Experimental evaluation, *Information Sciences*, Vol. 513, pp. 429–441 (2020).

[24] Johnson, J. M. and Khoshgoftaar, T. M.: Survey on deep learning with class imbalance, *Journal of big data*, Vol. 6, No. 27, pp. 1–54 (2019).

[25] Lin, T.-Y., Goyal, P., Girshick, R., He, K. and Dollar, P.: Focal Loss for Dense Object Detection, in *In Proceedings of International Conference on Computer Vision*, pp. 2980–2988 (2017).

[26] Wang, Q., Ma, Y., Zhao, K. and Tian, Y.: A Comprehensive Survey of Loss Functions in

Machine Learning, *Journal on Annals of Data Science*, Vol. 9, No. 2, pp. 187–212 (2022).

[27] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in *In Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186 (2019).

[28] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T. and Xie, S.: A ConvNet for the 2020s, in *In Proceedings of Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986 (2022).

[29] Snoek, C. G. M., Worring, M. and Smeulders, A. W. M.: Early versus late fusion in semantic video analysis, in *In Proceedings of International Conference on Multimedia*, pp. 399–402 (2005).

[30] De Souza, R. A., De Almeida, R. P., Moldovan, A.-N., Patrocínio, do Z. K. G. and Guimarães, S. J. F.: Gameplay Genre Video Classification by Using Mid-Level Video Representation, in *In Proceedings of Conference on Graphics, Patterns and Images*, pp. 188–194 (2016).

[31] Göring, S., Steger, R., Rao Ramachandra Rao, R. and Raake, A.: Automated Genre Classification for Gaming Videos, in *In Proceedings of International Workshop on Multimedia Signal Processing*, pp. 1–6 (2020).

付録

A　RQ1 の検証データの結果

Table A1: 各タグの推薦精度（検証データ・テキストと画像を両方使用）

	適合率		再現率		F1 スコア	
タグ	ML	BR	ML	BR	ML	BR
Platformer	0.566	0.535	0.500	0.535	0.531	0.535
Action	0.718	0.726	0.804	0.804	0.758	0.763
Pixel Graphics	0.328	0.337	0.533	0.587	0.406	0.428
Side Scroller	0.140	0.215	0.569	0.215	0.225	0.215
Metroidvania	0.486	0.643	0.447	0.237	0.466	0.346
VR	0.897	0.920	0.724	0.728	0.802	0.813

Table A2: 各タグの推薦精度（検証データ・テキストのみ使用）

	適合率		再現率		F1 スコア	
タグ	ML	BR	ML	BR	ML	BR
Platformer	0.560	0.511	0.517	0.552	0.538	0.531
Action	0.747	0.738	0.762	0.788	0.754	0.762
Pixel Graphics	0.264	0.337	0.413	0.371	0.322	0.353
Side Scroller	0.217	0.229	0.508	0.539	0.304	0.321
Metroidvania	0.436	0.524	0.447	0.290	0.442	0.373
VR	0.911	0.886	0.718	0.748	0.803	0.811

Table A3: 各タグの推薦精度（検証データ・画像のみ使用）

	適合率		再現率		F1 スコア	
タグ	ML	BR	ML	BR	ML	BR
Platformer	0.226	0.195	0.349	0.320	0.274	0.242
Action	0.533	0.509	0.836	0.903	0.651	0.651
Pixel Graphics	0.406	0.410	0.335	0.353	0.367	0.379
Side Scroller	0.113	0.111	0.108	0.092	0.110	0.101
Metroidvania	0.091	0.072	0.079	0.158	0.085	0.099
VR	0.230	0.239	0.518	0.558	0.319	0.335

B　全共起使用パターンの推薦精度

ここでは各ヒューリスティクス導入パターンで検証データの F1 スコアが最大になった際の各ハイパーパラメータの値，検証データでの精度，テストデータでの精度を紹介する．なお，4.4 節で述べたように本実験では各ヒューリスティクスのハイパーパラメータを 0.1〜2.0 の範囲で 0.1 刻みで実験したが，いくつかのヒューリスティクス導入パターンでは検証データの F1 スコアが複数のハイパーパラメータで同じ最大値になっていた．その場合については検証データの F1 スコアが最大になるハイパーパラメータの値全ての結果を載せている．

Table B1: 共起関係に基づくヒューリスティクス導入後の Platformer タグ推薦精度 (1/4)

					検証データ			テストデータ		
γ_{act}	γ_{pg}	γ_{ss}	γ_{mtr}	γ_{vr}	適合率	再現率	F1 スコア	適合率	再現率	F1 スコア
-	-	-	-	-	0.535	0.535	0.535	0.509	0.477	0.493
-	-	-	-	0.1	0.525	0.547	0.536	0.503	0.500	0.502
-	-	-	-	0.2	0.525	0.547	0.536	0.512	0.500	0.506
-	-	-	0.4	-	0.571	0.512	0.540	0.563	0.465	0.510
-	-	-	0.5	0.3	0.583	0.512	0.545	0.565	0.430	0.488
-	-	-	0.5	0.4	0.583	0.512	0.545	0.574	0.430	0.492
-	-	1.3	-	-	0.545	0.529	0.537	0.528	0.494	0.511
-	-	1.4	-	-	0.545	0.529	0.537	0.528	0.494	0.511
-	-	1.5	-	-	0.545	0.529	0.537	0.525	0.488	0.506
-	-	0.6	-	0.3	0.531	0.541	0.536	0.518	0.494	0.506
-	-	0.7	-	0.3	0.531	0.541	0.536	0.518	0.494	0.506
-	-	0.1	0.4	-	0.579	0.512	0.543	0.562	0.448	0.498
-	-	0.2	0.4	-	0.579	0.512	0.543	0.558	0.448	0.497
-	-	0.1	0.5	0.2	0.587	0.512	0.547	0.568	0.436	0.493
-	-	0.1	0.5	0.3	0.587	0.512	0.547	0.565	0.430	0.488
-	-	0.2	0.5	0.3	0.587	0.512	0.547	0.565	0.430	0.488
-	-	0.3	0.5	0.3	0.587	0.512	0.547	0.565	0.430	0.488

Table B2: 共起関係に基づくヒューリスティクス導入後の Platformer タグ推薦精度 (2/4)

					検証データ			テストデータ		
γ_{act}	γ_{pg}	γ_{ss}	γ_{mtr}	γ_{vr}	適合率	再現率	F1 スコア	適合率	再現率	F1 スコア
-	-	-	-	-	0.535	0.535	0.535	0.509	0.477	0.493
-	0.1	-	-	-	0.519	0.552	0.535	0.494	0.506	0.500
-	0.1	-	-	0.1	0.519	0.547	0.533	0.494	0.506	0.500
-	0.1	-	-	0.2	0.519	0.547	0.533	0.500	0.506	0.503
-	0.1	-	0.9	-	0.571	0.512	0.540	0.569	0.454	0.505
-	0.1	-	0.5	0.3	0.584	0.506	0.542	0.566	0.424	0.485
-	0.1	-	0.5	0.4	0.584	0.506	0.542	0.575	0.424	0.488
-	0.1	0.1	-	-	0.516	0.552	0.534	0.494	0.506	0.500
-	0.3	0.7	-	0.1	0.557	0.512	0.533	0.544	0.436	0.484
-	0.1	0.6	1.5	-	0.544	0.535	0.540	0.543	0.483	0.511
-	0.1	0.6	1.6	-	0.544	0.535	0.540	0.536	0.483	0.508
-	0.1	0.2	0.5	0.4	0.583	0.512	0.545	0.570	0.424	0.487

Table B3: 共起関係に基づくヒューリスティクス導入後の Platformer タグ推薦精度 (3/4)

					検証データ			テストデータ		
γ_{act}	γ_{pg}	γ_{ss}	γ_{mtr}	γ_{vr}	適合率	再現率	F1 スコア	適合率	再現率	F1 スコア
-	-	-	-	-	0.535	0.535	0.535	0.509	0.477	0.493
0.1	-	-	-	-	0.566	0.500	0.531	0.556	0.436	0.489
0.1	-	-	-	0.3	0.573	0.500	0.534	0.555	0.413	0.473
0.1	-	-	0.7	-	0.553	0.517	0.535	0.549	0.459	0.500
0.2	-	-	1.5	2.0	0.567	0.517	0.541	0.571	0.442	0.498
0.2	-	1.0	-	-	0.553	0.517	0.535	0.519	0.471	0.494
0.2	-	0.9	-	0.4	0.557	0.512	0.533	0.535	0.448	0.487
0.1	-	0.6	1.3	-	0.528	0.547	0.537	0.515	0.488	0.502
0.2	-	0.1	1.3	2.0	0.567	0.517	0.541	0.568	0.436	0.493
0.2	-	0.1	1.4	2.0	0.567	0.517	0.541	0.567	0.442	0.497
0.2	-	0.1	1.5	2.0	0.567	0.517	0.541	0.571	0.442	0.498
0.2	-	0.2	1.3	2.0	0.567	0.517	0.541	0.571	0.442	0.498
0.2	-	0.2	1.4	2.0	0.567	0.517	0.541	0.567	0.442	0.497
0.2	-	0.2	1.5	2.0	0.567	0.517	0.541	0.567	0.442	0.497
0.2	-	0.2	1.6	2.0	0.567	0.517	0.541	0.571	0.442	0.498
0.2	-	0.3	1.4	2.0	0.567	0.517	0.541	0.567	0.442	0.497
0.2	-	0.3	1.5	2.0	0.567	0.517	0.541	0.567	0.442	0.497

Table B4: 共起関係に基づくヒューリスティクス導入後の Platformer タグ推薦精度 (4/4)

					検証データ			テストデータ		
γ_{act}	γ_{pg}	γ_{ss}	γ_{mtr}	γ_{vr}	適合率	再現率	F1 スコア	適合率	再現率	F1 スコア
-	-	-	-	-	0.535	0.535	0.535	0.509	0.477	0.493
0.1	0.1	-	-	-	0.562	0.500	0.529	0.548	0.430	0.482
0.1	0.1	-	-	0.2	0.570	0.500	0.533	0.550	0.413	0.472
0.3	0.2	-	0.9	-	0.635	0.465	0.537	0.633	0.401	0.491
0.1	0.1	-	1.5	1.9	0.563	0.517	0.539	0.563	0.442	0.495
0.1	0.1	-	1.5	2.0	0.563	0.517	0.539	0.563	0.442	0.495
0.1	0.1	0.9	-	-	0.542	0.523	0.533	0.513	0.477	0.494
0.1	0.1	1.0	-	-	0.542	0.523	0.533	0.519	0.477	0.497
0.1	0.1	0.6	-	0.2	0.582	0.494	0.535	0.550	0.413	0.472
0.3	0.2	0.1	0.8	-	0.635	0.465	0.537	0.620	0.390	0.479
0.3	0.2	0.1	0.9	-	0.635	0.465	0.537	0.626	0.390	0.480
0.2	0.1	0.2	1.4	2.0	0.563	0.517	0.539	0.568	0.436	0.493