グラフ構造を用いた探索プロトタイプの開発

Development of a Graph-based Search Prototype

植　木　達　彦，牧　原　幸　伸

要　約　現代の情報化社会において，蓄積される情報資源の量は増加している一方で，データは組織や個人のもとに分散して保管されており，データのサイロ化が問題となっている．この課題を解決するため断片化した情報を，グラフ構造を用いてネットワーク化することで，網羅的な探索を行うことを目指す．本取り組みではこのアプローチの第一歩として，データをハイパーグラフへと変換し，重み付けランダムウォークを用いて網羅的な探索を行うプロトタイプを構築した．実験として，知的財産分野における特許調査業務をユースケースに選定し，特許のデータをグラフに変換し，探索の精度を評価した．その結果，提案するプロトタイプは参考値としたキーワード検索を上回る精度を達成した．今後は，探索パラメータ調整の自動化や異種データの結合の研究を進める予定である．

Abstract　In today's information society, while the volume of accumulated information resources is burgeoning, data remains dispersed across organizations and individuals, leading to the issue of data siloing. To address this challenge, we aim to facilitate exhaustive information discovery by networking fragmented data through graph-based structures. As an initial step in this approach, we developed a prototype that transforms data into hypergraphs and performs comprehensive searches using weighted random walks. In our experiment, we selected patent search within the field of intellectual property as a primary use case, mapping patent data onto graph structures to evaluate search precision. The results demonstrate that the proposed prototype achieves accuracy superior to that of the keyword search. Future research will focus on the automation of search parameter tuning and the integration of heterogeneous data sources.

1. はじめに

情報化が進展する社会環境において，蓄積される情報資源の量は急速に増加の一途を辿っている．しかしながら，これらの膨大なデータの多くは組織や個人のもとに分断された状態で保管されており，「データのサイロ化」という構造的な問題となっている．結果として，ユーザが検索によって発見し活用できるデータは，特定のプラットフォームやインターネット上に一般公開された極めて限定的な資源に留まっているのが実情である．このように分散し断片化したデータベース群を横断的かつ網羅的に探索し，求める情報を高精度に抽出することは，キーワードマッチングに代表されるような既存の検索手法では困難である．

上記のような課題を根本的に解決するためには，検索の文字列などのユーザのバイアスに過度に依存しない客観的な網羅性を担保しつつ，日々増大するデータを低コストで処理できるスケーラブルな探索技術が求められる．本取り組みでは，この探索を実現するアプローチとして，データを「ノード（頂点）」と「エッジ（枝）」からなるネットワーク構造として表現し，要素間の結びつきの強さを元に解析を行うグラフ構造に着目した．断片化した情報を，グラフ構造

を用いてネットワーク化することで，網羅的な探索が可能となるだけでなく，グラフのトポロジー（接続形態）から，従来の手法では見出すことのできなかった潜在的な関係性や新たな洞察を抽出することが可能となる．

本取り組みでは，上記のアプローチを実現するための第一歩として，単一のデータソースからグラフを構築し探索を行うプロトタイプを開発した．具体的には，企業に広く普及しているリレーショナルデータベース（RDB）形式のデータをグラフ構造に変換し，ノードおよびエッジに独自の重み付けを施したランダムウォーク（Random Walk：RW）ベースのグラフ探索を行うプロトタイプの構築，およびその実証実験の成果について報告する．実証実験においては，本来であれば多様なユースケースに対して適用可能であるが，本取り組みでは，プロトタイプの有効性を定量的に確認するため，知的財産分野における「特許調査業務」をユースケースとして取り上げる．特許のデータをグラフに変換，探索を実施し精度を評価する．

本稿の構成は以下の通りである．2 章では関連研究と RDB のグラフ変換における既存手法の課題，および本研究の前提技術となる重み付けフレームワークについて整理する．3 章では前提技術を含むプロトタイプシステムの構築について詳述する．4 章では実際の特許データを用いた実証実験の設計と結果を示す．5 章では得られた結果に基づく考察と技術的課題を述べ，6 章で本稿のまとめと今後の展望を総括する．

2. 解決のアプローチと前提となる技術

情報資源に対し，グラフを用いて構造化し，そこから適切な関係性を引き出すアプローチには多数のバリエーションが存在する．本取り組みでは，ノードやエッジの属性情報に依存せず，グラフの接続構造そのものをベースに重要度を測る PageRank 中心性を採用する．

また，ベースとなる情報資源は一般的な企業で広く利用されているリレーショナルデータベース（RDB）形式を前提とした．

本章では，前提となる PageRank 中心性について説明し，既存手法の課題，および本取り組みの前提技術となる藤井ら[1]の「重み付けフレームワーク」のモデルについて説明する．

なお本取り組みは，大日本印刷株式会社と慶應義塾大学金子研究室との共同研究に基づくものである．

2.1 PageRank 中心性とランダムウォークによるグラフ解析

PageRank（PR）は，ネットワーク上においてランダムウォーク（確率的な遷移）によって到達しやすい，ハブとして重要なノードやエッジを評価する中心性指標である．この指標はもともとウェブページの相対的な重要度を評価するために考案されたアルゴリズムであるが，現在ではソーシャル・ネットワーキング・サービス（SNS）の解析など，多様なアプリケーションにおいて利用されている．

PR の特徴は，単なるノード同士の接続数（次数）の多寡を評価するだけでなく，ネットワーク全体のグローバルなトポロジーを反映して各要素の重要度を算出できる点にある．これにより，直接的な繋がりを持たないノード間であっても，ネットワークの構造を介した間接的な影響力や関連性を定量的に測ることが可能となる．

さらに，PR を特定の探索に特化させた指標として，Personalized PageRank（PPR）が存在する．通常の PR が「ネットワーク全体における各ノードのグローバルな重要度」を測るの

に対し，PPR は特定の起点ノード（またはノード群）からランダムウォークを開始したと想定し，その「起点ノードに対する相対的な関連性の強さ」を評価する指標である．本研究で対象とする特許探索タスクにおいては，検索クエリとなる特定の特許（起点ノード）に焦点を当て，そこから関連の深い特許をランキング形式で見つけ出す必要があるため，通常の PR ではなく PPR を用いて探索処理を実行する．なお，本稿では特記しない限り，PPR を用いた評価も含めて広義の PR として表記する．

2.2　RDB のグラフ変換手法

RDB のデータをグラフ解析のアルゴリズムに適用するためには，テーブル内に格納されたレコード群とカラムの関係性を，ノードとエッジからなるネットワーク構造に適切に変換する必要がある．代表的な既存の変換手法として，クリーク展開とハイパーグラフ変換の二つが存在するが，それぞれが課題を抱えている．

2.2.1　クリーク展開

クリーク展開は，RDB の各レコードを一つのノードに変換し，同一のカラム値を持つレコード間を全て通常のエッジ（二つのノードのみを結ぶ一対一のエッジ）で相互に結合してグラフを形成するアプローチである．この変換手法は直感的であるが，例えば人事情報のようなデータベースには「性別」や「国籍」，あるいは特許データにおいては「最上位の国際特許分類（IPC）」のように，極端に多くのレコードが該当する巨大なカラムが存在するという特性がある．

この手法でグラフを構築した場合，巨大なカラムに属するレコード群の間に膨大な数のエッジが生成される．その結果，ランダムウォーカーがそれらの巨大なクラスタによる引力に捕らわれやすくなり，PR の確率分布が「サイズの大きいカラム」や「次数の高い有名ノード」に極端に引きずられる現象が発生する．

これにより，ノードの PR 値と次数の順位相関（以降，NPR 相関度）およびエッジの PR 値とエッジサイズの順位相関（以降，EPR 相関度）が過剰に高くなる．結果として，PR が巨大な属性による単純な次数効果以外の微細な特徴をほとんど反映できなくなり，探索のための評価指標としての価値が著しく一面的なものになってしまう．

2.2.2　ハイパーグラフ変換

ハイパーグラフとは，通常のエッジが二つのノード間のみを結ぶのに対し，図 1 のように一つのエッジが二つ以上の複数ノードを同時に包含できる拡張されたグラフ構造である．ハイパーグラフ変換では，同一のカラム値に属するレコード群を一つのハイパーエッジとして定義し，該当する全レコードをその内部のノードとして接続する．

ハイパーグラフ上のランダムウォーク処理では，現在地のノードを含む全エッジ群の中から等確率で一つのエッジを選択したのち，その選択されたエッジに含まれる全ノードの中からさらに等確率で次の遷移先ノードを選択するという二段階のステップを踏む．図 2 の例であれば，ノード a から開始したランダムウォーカーがエッジ b を選択し，さらにエッジ b の中からノード e を選択し遷移している．この手法では，最初のエッジ選択ステップにおいてエッジサイズ（内包するノードの絶対数）の大小による遷移確率への影響が排除されるため，2.2.1 項のク

リーク展開のように巨大なエッジにランダムウォーカーが過剰に吸い込まれる問題を回避できる．

しかし，ハイパーグラフ変換では逆に，NPR 相関度と EPR 相関度が極端に低くなりすぎるという弊害が生じる．ネットワークの解析において「次数が高い（多くの共通属性を共有している，あるいは特許であれば多数の文献から引用されている基本技術である）」という事実は，情報伝播の要となる重要な指標の一つである．しかし，HT 手法では PR アルゴリズムがこの次数の重要性を構造的にほとんど加味できなくなり，やはり探索のための評価指標としての有用性が損なわれてしまう．

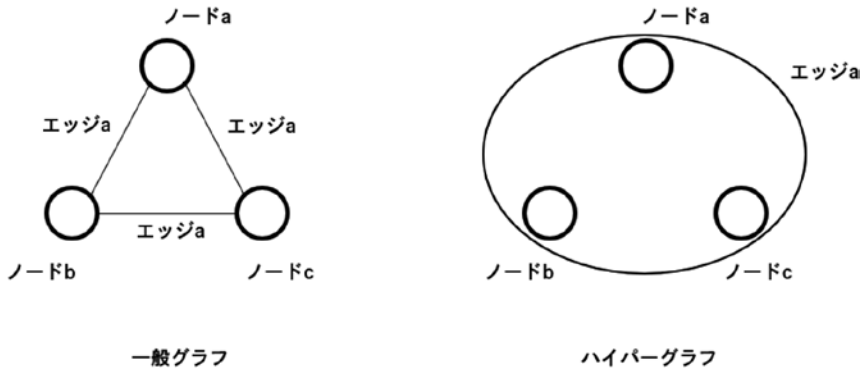

図 1　一般グラフとハイパーグラフ

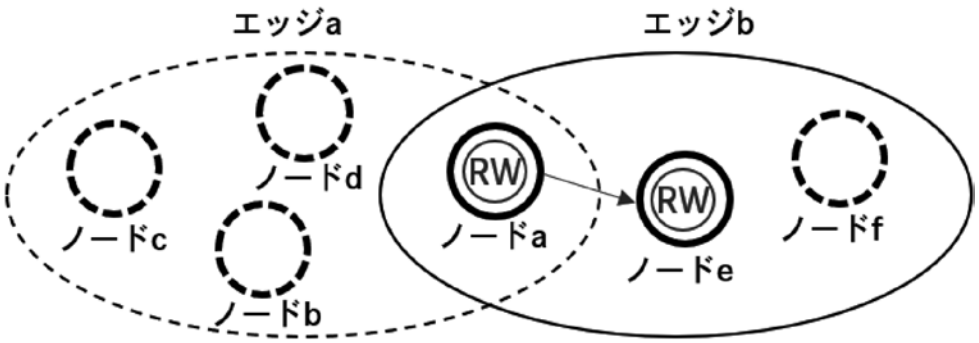

図 2　ハイパーグラフ上のランダムウォーク処理

2. 3　解決すべき技術的課題

既存のグラフ変換手法では，PR 相関度の値が両極端（高すぎる，または低すぎる）となり，ユーザの探索要件に合致した適切な依存度を持つ PR を提供できないことが最大の課題である．特許調査のような複雑な探索タスクにおいては，特許分類などの巨大なクラスタを形成する属性の影響を適度に抑えつつ，特有の専門用語などのニッチな属性が持つ関連性を適切に拾い上げるバランスが求められる．また，基本特許と呼ばれるような次数（被引用数など）の大きな重要なノードの影響力も完全に無視することはできない．

したがって，NPR 相関度と EPR 相関度を自由に調節し，探索の遷移確率をデータセットや探索目的に応じて最適化することが必要となる．

2.4　重み付けフレームワーク

前節で述べた技術的課題を解決するため，本取り組みでは前提技術として，データをハイパーグラフ構造に変換した上でエッジサイズやノードの次数に応じた「重み付け」を動的に行うフレームワークを採用する．

本項では，このフレームワークを説明する

2.4.1　ハイパーグラフ上のランダムウォーク遷移確率

重み付きハイパーグラフ上のランダムウォーカーは，以下の二段階の独立した確率的ステップを実行してネットワークを巡回する．

1）エッジの選択

ランダムウォーカーが現在滞在しているノードに接続されている全エッジの中から，エッジの重みに応じた確率で遷移先のエッジを選択する．エッジ e_i が選択される確率 P_{e_i} は，以下の式(1)で定義される．

$$P_{e_i} = \frac{w_{e_i}}{\sum w_e} \tag{1}$$

ここで，w_{e_i} は該当する遷移先エッジに付与された個別の重みであり，分母の $\sum w_e$ は現在地ノードを含む全エッジの重みの総和である．重みの大きいエッジへの遷移確率が相対的に高くなり，重みの小さいエッジへの遷移確率は低く制御される．

2）ノードの選択

次に，第一段階で選択されたエッジの内部に含まれる全ノードの中から，各ノードの重みに応じた確率で最終的な遷移先のノードを選択する．ノード v_i が選択される確率 P_{v_i} は，以下の式(2)で定義される．

$$P_{v_i} = \frac{w_{v_i}}{\sum w_v} \tag{2}$$

ここで，w_{v_i} は該当ノードに付与された個別の重みであり，分母の $\sum w_v$ は現在地のエッジが包含する全ノードの重みの総和である．

2.4.2　エッジサイズに応じた重み付けモデル

RDB において極端に多くのレコードを保持する巨大なカラム（例：多くの特許を含む技術分類など）への過剰な遷移を抑制し，EPR 相関度を低下させるため，エッジサイズの大小に応じた重み付けを導入する．

エッジ e_k に対して，式(3)のように重み w_{e_k} を付与する．

$$w_{e_k} = \left(S_{e_k} - 1\right)^{1-\alpha} \tag{3}$$

ここで，S_{e_k} はエッジ e_k のエッジサイズ（エッジが包含するノード数）であり，α は $0 \leq \alpha \leq 1$ の範囲を取るエッジ重み付けパラメータである．

図 3 に示すように，$\alpha = 0$ に設定した場合，エッジサイズに比例してランダムウォーカーが吸い込まれるクリーク展開と同様の振る舞いを示す．一方，α の値を徐々に大きくするにつれ

て，包含ノード数が多いサイズの大きいエッジの重みが減衰する．これにより，巨大なエッジによるランダムウォーカーの流量の独占が抑えられ，EPR 相関度を単調に減少させることが可能となる．$\alpha = 1$ の場合は，全てのエッジの重みは 1 となりハイパーグラフ変換と同様の振る舞いとなる．

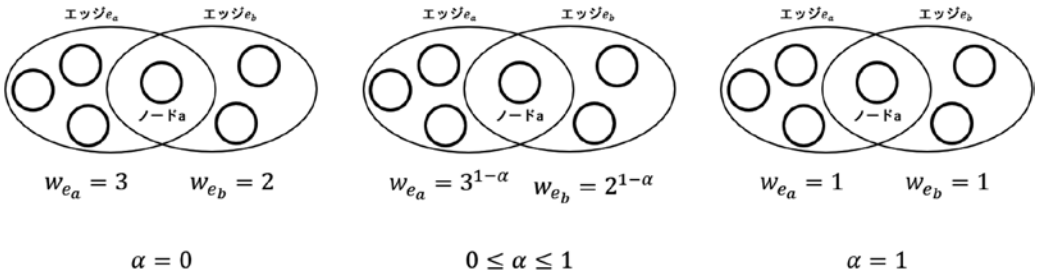

図 3　エッジサイズに応じた重み付け

2. 4. 3　ノードの次数に応じた重み付けモデル

巨大なエッジに属することで結果的に高次数となってしまったノード（例：様々な分類に重複して登録されている汎用的な特許など）への過剰な到達を抑制し，NPR 相関度を低下させるため，ノードの次数の大小に応じた重み付けを導入する．

ノード v_k に対して，式(4)のように重み w_{v_k} を付与する．

$$w_{v_k} = d_{v_k}{}^{-\beta} \tag{4}$$

ここで，d_{v_k} はノード v_k の次数であり，β は $0 \leq \beta \leq 1$ の範囲を取るノード重み付けパラメータである．

図 4 に示すように，$\beta = 0$ に設定した場合，全てのノードの重みは 1 となり，選択されたエッジ内において等確率でノードが選択される．β の値を大きくするにつれて，高次数ノードの重みが相対的に小さくなり，高次数ノードへのランダムウォーカー流量が抑制される．これにより，ノード間の次数による遷移確率の差が平滑化され，NPR 相関度を単調に減少させることが可能となる．

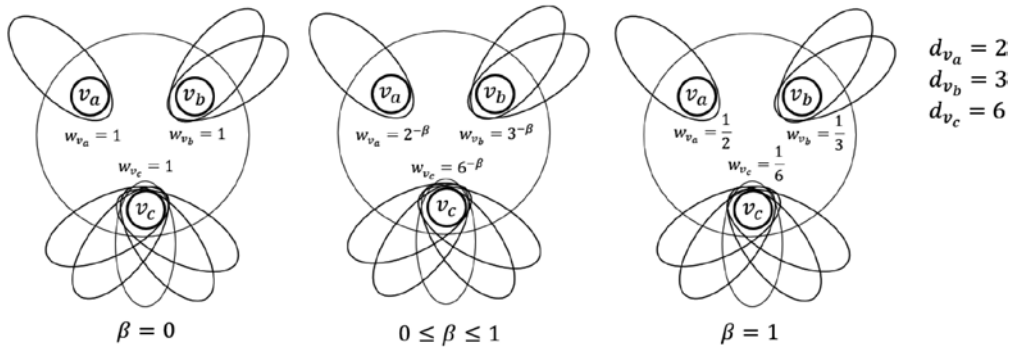

図 4　ノードの次数に応じた重み付け

これら二つのパラメータ（α,β）の強弱をユーザの探索要件に合わせてチューニングすることで，両極端な既存手法（クリーク展開とハイパーグラフ変換）の間を結び，最適な相関度を持つ PR を出力することが可能となる．

3. プロトタイプの構築

2 章の重み付けフレームワークを前提として採用し，RDB に格納されたデータをハイパーグラフに変換し，探索を実行する「特許探索プロトタイプ」を構築した．本章では，その構成について説明する．

3.1 プロトタイプの全体アーキテクチャ

開発したプロトタイプは以下の四つのフェーズで構成される（図 5）．

1）グラフ構築

指定したデータソースからデータを読み込み，ノードとするデータの基本単位と，ハイパーエッジとして結びつける共通属性を定義してグラフ構造を構築する．本システムでは，自然言語で記述されたテキストを形態素解析したものと，分類番号などを同一のグラフに統合する．

2）探索指示

ユーザが探索の起点となる単一ないし複数のノード（基準となる手持ちの特許など）をシステムに入力する．従来の検索エンジンのように特定のキーワード群を入力する必要はなく，ノードそのものを探索クエリとして扱う．ここでランダムウォークの試行回数や終了確率，前提技術で定義したエッジ・ノード重み付けパラメータ（α,β）を指定する．

3）グラフ探索処理

指定された起点ノードから，式(1)および式(2)の遷移確率に基づくランダムウォーカーがネットワーク全体を巡回し，各ノードの PPR 値を計算する．

4）結果提示

計算された PPR 値に基づき，起点ノードから関連性が高いと推定されるノード群をランキング形式でリスト化し，提示する．

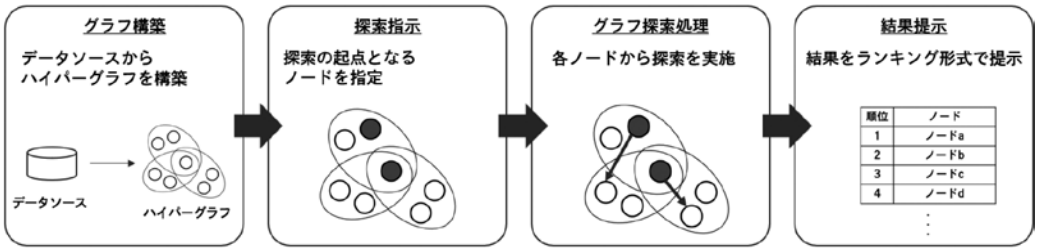

図 5　プロトタイプの全体アーキテクチャ

4. 実験

構築したプロトタイプの探索精度と有効性を定量的に評価するため，知的財産分野における特許先行技術調査業務を取り上げて実証実験を行った．

4.1　ユースケース選定の背景と目的

本プロトタイプで実装したグラフベースの探索技術は，テキストや数値ベースであれば特定のデータソースやフォーマットに特化したものではなく，様々な用途に対して汎用的に利用可能である．しかし，出力されるランキング結果は「ノード間の関係性の強さ（PPR 値）」という指標であるため，一般的なデータセットにおいては「何が正しい探索結果であるか」を客観的かつ定量的に評価（正解データを定義）することが非常に難しいという課題があった．

そこで，定量的な評価が可能なユースケースとして特許調査業務を選定した．通常，ユースケースとして商品レコメンドなどの個人によって嗜好が異なる（正解のない）ものが考えられるが，定量的な評価を行うことが難しい．特許調査においては，特定の特許からの関係性の有無を，特許庁の審査官の判断を正解として扱うことで定量的な評価を行った．

例えば企業が新たな技術について特許出願を行う際，事前に特許庁の審査官と同様の観点から先行技術調査（サーベイ）を実施することが不可欠である．出願審査において，審査官から既存特許との抵触を指摘されて拒絶理由通知書が発行され，最終的にそのまま拒絶査定となった場合，出願までに要した多大な時間と資金が浪費される結果となる．

特許探索においては，技術分野が少しでも異なれば同一の概念や技術的課題に対しても全く異なる語彙が用いられることが多く，申請者が自身の専門外の技術領域にまで精度の高い調査を及ぼすことは単純なキーワード検索では極めて困難である．本実験では，ある特許に対し審査官が引用した「引用文献」を正解ペア，つまりクエリ特許に対して関連性があると審査官に判断された特許と定義する．当プロトタイプが起点特許から探索を開始し，提示したランキングリストの上位に，この引用文献がどれだけの確率で含まれているかを測定することで，システムの探索精度を評価する．

4.2　データセットの諸条件

特許データベースより，以下の条件にて特許データを抽出し実験用データセットを構築した．なお，データセットの諸条件については，異なる手法であるが類似特許検索の研究である中山[2]らの研究条件を参考にしている．

特許データベースより，以下の条件にて特許データを取得した．

1）国際特許分類（IPC）：D（繊維・紙）
2）対象期間：全期間
3）データクレンジング：「要約」項の存在しない特許レコードを除外し，最終的な全ノード件数を 164,075 件とした．

取得した特許データの項目は以下の通りである．

1）発明の名称
2）要約
3）請求項
4）IPC
5）F ターム
6）出願人識別番号

7）代理人識別番号
8）引用文献情報

IPC（国際特許分類）とは，世界各国が共通に使用できる特許分類である．

F タームとは，特許審査のための先行技術調査を迅速に行うために開発された検索インデックスのことであり，特許庁が開発している．

出願人識別番号は出願人本人が付与される 9 桁の番号であり，代理人識別番号は弁理士などの代理人が付与される 9 桁の番号である．

中山[2]らが用いているデータ項目に対し，本取り組みでは請求項や出願人識別番号，代理人識別番号などを追加で用いている．

4. 3　ハイパーグラフの構築

グラフ構築においては，ノードとエッジを定義する必要がある．今回の実験では，ノードは特許 1 レコードとした．エッジに関しては図 6 のように二通りのエッジ生成処理を行い，同一のハイパーグラフへと統合した．

1）自然言語で記載されたテキスト項目に関する処理

特許レコードから「発明の名称」「要約」および「請求項」の自然言語テキストを抽出し，形態素解析を用いて解析，「名詞」のみを抽出する．抽出された各名詞に対し，その名詞が出現した特許レコード同士をハイパーエッジで接続する．図 6 の例であれば，特許 A の「要約」から複数の名詞を抽出し，同じ名詞を有する他の特許とハイパーエッジで接続している．

2）分類番号に関する処理

「IPC（国際特許分類）」「F ターム」「引用文献情報」「出願人識別番号」「代理人識別番号」といった項目は，そのまま属性値として扱う．名詞と同様に，各属性値がどの特許レコードに出現したかを検索し，共有するレコード同士をハイパーエッジで構築する．また，「引用関係」については引用及び被引用特許を含んだハイパーエッジとして構築する．図 6 の例であれば，特許 A と同じ IPC に属する他の特許とハイパーエッジで接続している．

これにより，テキスト内の共通単語から生じるエッジと，分類番号から生じるエッジが同一のハイパーエッジとしてグラフ内に存在することとなる．

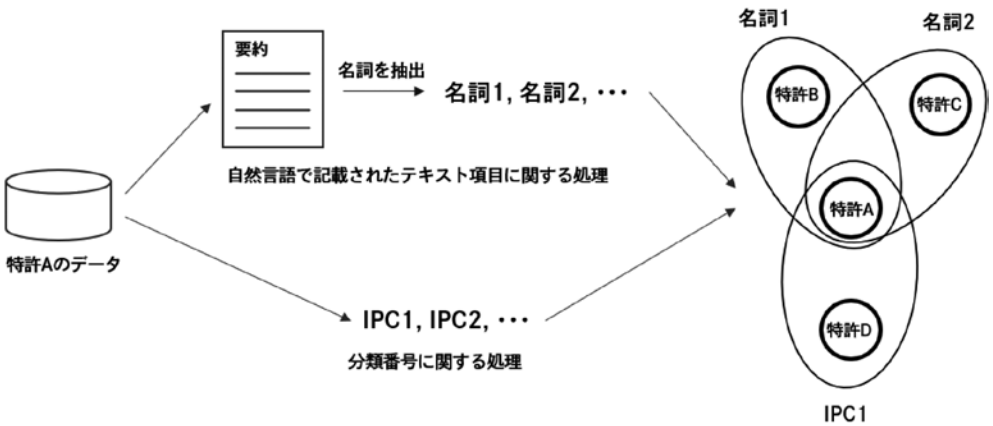

図 6　特許データからハイパーグラフの構築

また事前の検証において，上記で構築手順したグラフは重み付けが無視されるほどにノードに対する次数が過度に多くなってしまい（遷移確率がほぼ同一となり，ランダムに遷移先を選ぶことと同じとなってしまい）結果が収束せず精度が出ない，という課題が発生した．例えば一つのノードを含むエッジが 1,000,000 個存在する場合は，各エッジへの遷移確率が $\alpha = 1$ において 0.0001％となり，α を 1 以外に調整し重み付けを行ったとしても，ランダムにエッジを選び遷移することとほぼ同じとなり，収束しない結果となる．よって今回のグラフについては，エッジの大きさに上限値を設け，上限値を超過した大きさのエッジを削除した．

4. 4　実験条件と評価結果

精度評価の指標として，中山[2]らの先行研究を参考に Recall@k を採用した．Recall@k は，クエリとなる特許に対してシステムが予測出力した上位 k 個のリストの中に，実際の正解特許（審査官が提示した引用文献）が含まれている割合を算出し，最終的に全クエリにおける平均を取った指標である．特許調査においては，適合率（Precision）よりも見落としを防ぐ網羅率（Recall）が重要と考え，本取り組みではこの指標を採用している．

テスト手法として，データセット全体（164,075 件）のうち，時系列的に最も新しい 10％にあたる特許群を探索の起点（クエリ）とし，残りの過去 90％のデータ群を探索対象の母集団とした．クエリとなる 10％のデータ群については「引用関係」のハイパーエッジを削除した状態で実験を行った．探索実行時には，ランダムウォークの試行回数，終了確率，重み付けパラメータ（α, β）について，精度が最大化されるようチューニングを行った．

表 1 に提案手法の精度を記載する．（引用関係有り）と記載した手法は，4. 2 節で述べたデータ項目全てを用いてハイパーグラフを構築した場合の結果である．（引用関係無し）と記載した手法は，4. 2 節で述べたデータ項目から「引用関係」を除外してハイパーグラフを構築した際の結果である．

完全に同一のデータセットではなく用いているデータ項目も異なるため，あくまで参考値ではあるが，中山ら[2]の実験結果よりキーワードによる全文検索の精度と中山ら[2]の提案手法である機械学習ベースの類似特許検索手法による精度も併記する．

表 1　各手法の精度

No.	探索手法	k＝50	k＝100
1	本取り組みの提案手法（引用関係有り）	37.5%	48.3%
2	本取り組みの提案手法（引用関係無し）	36.8%	47.8%
3	キーワードによる全文検索	25.2%	31.0%
4	中山ら[2]の提案手法	37.1%	48.1%

注記：精度算出に用いたデータセット等が完全に同一でないため No. 3 と No. 4 は参考値

ハイパーグラフを用いた提案手法は，k＝50,100 のどちらにおいてもキーワードによる全文検索の精度を大幅に上回った．

提案手法のうち引用関係有りの結果は，中山ら[2]の提案手法と同程度の水準となった．

引用関係無しの結果においても，引用関係有りの結果や中山ら[2]の結果から大きくは劣らな

い結果となった．これは，「引用関係」のような正解のデータが無い状態であっても，ある一定の精度で正解を推測することが可能であることを示している．また，正解の存在しないような課題に対して適用した場合においても，ある程度もっともらしい結果を提示することができると考えられる．

5. 考察

得られた結果に基づき，一定の精度が得られた要因を考察する．

5.1 グラフベース探索の精度の要因について

表1に示された通り，キーワード全文検索の精度（Recall@100 で 31.0%）は他の手法に比べて低い．これは，特許文書特有の「言い換え」や「専門用語のブレ」など，技術領域ごとの専門用語を単純な文字列一致では吸収できないためと推測される．

一方，提案手法が高い精度を記録した最大の要因は「ハイパーグラフによる多面的なトポロジー」にあると考えられる．特許間の関連性は，単に明細書のテキストが似ているかという一側面的な要素だけでなく，「特定のニッチなFターム群を共有している」「同じ代理人が担当しており，記述が似ている」「同じ出願人が過去に出した一連の技術ポートフォリオに含まれる」といった，複数のメタデータの交差によって特徴づけられる．

提案手法は，名詞の共有というテキスト的側面に加え，IPC やFターム，出願人といった多様な文脈を，「ハイパーエッジ」として単一のグラフ空間に投影している．その上でランダムウォークを実行することで，熟練した審査官が経験的に行うような「複数の要素を経由して目的の特許に至る」といった連想的な推論プロセス（マルチホップな探索）を，網羅的に再現できたのではないかと考えられる．

5.2 パラメータ（α, β）が探索の精度に与えた影響

特許データベースにおいては，「G06F（計算処理）」のような極めて汎用的で範囲の広いIPC や，明細書中に頻出する「装置」「方法」といった一般名詞が無数に存在する．クリーク展開のようにこれらを単純なエッジで結合してしまうと，これらの巨大な属性が過度にランダムウォーカーを吸い込み，結果として次数に基づく一面的な結果になってしまう．

本プロトタイプで導入したエッジ重み付け（α）は，式(3)によってこのような「大きすぎる分類コード」や「汎用すぎる単語」が持つエッジの引力を数学的に減衰させる機能を果たしたと推測される．これにより，ノイズとなる汎用的な繋がりが排除され，マイナーではあるがクエリと確かな共通項（特定のニッチなFタームや専門用語の組み合わせなど）を持つ，真に関係性のある先行特許へとランダムウォーカーを効率的に到達させることに寄与したと考えられる．

5.3 プロトタイプの残課題と将来展望

本プロトタイプについて，技術的な残課題が存在する．

第一に，重み付けパラメータ（α, β）のチューニングである．例えば，EC サイトの商品レコメンドで用いる場合であれば，個人毎に探索の趣味嗜好が異なるため，結果を参照しつつ自身の好みのパラメータに調整すればよい．このような場合であればユーザにパラメータ調整を

委ねることが可能であるが，本実験のような絶対的な正解が存在し精度を求められるケースにおいては，精度が高まるようパラメータのチューニングを行う必要がある．クエリの性質，目的や対象データセットに応じて最適な（α,β）の値は動的に変動するはずであり，なんらかのチューニングのアルゴリズムが必要である．例えば，事前のデータ分布やネットワークの次数分布から，これらのパラメータを自動的に推定・調整を行うアルゴリズムが考えられる．

第二に，異種データベースの接続である．今回は特許データベースという複数の項目はあるが単一のデータベース内で探索を完結させているが，実際の調査業務においては，特許文献だけでなく，学術論文データベースなど，異なる外部の情報資源と横断的に探索を行うことが求められる．異なるデータベースから生成されたグラフ同士を「共通」もしくは「近しい」ノードやエッジを介して自動的に接続する方式（グラフ間の自動接続）を確立する必要がある．

6. お わ り に

本稿では，企業や社会全体に散在しサイロ化が進む情報資源を統合し，網羅的かつスケーラブルな情報探索を実現するためのアプローチの第一歩として，データソースからグラフを構築，重み付けされたランダムウォークの探索フレームワークを提案し，プロトタイプを構築した．知的財産分野における特許調査業務を対象とした実験を通じて，本手法が一定の探索精度を達成できることが分かった．

今後は，本実証実験の考察で明らかになったパラメータチューニングの自動化や，学術論文等を含めた異種データベースとのグラフ接続方式の確立に向けて研究を進展させる予定である．

謝辞

本取り組みは，大日本印刷株式会社と慶應義塾大学金子研究室との共同研究に基づくものである．

参考文献

［1］ 藤井洸太朗，山下剛志，金子晋丈，リレーショナルデータベースのグラフ利用のための重み付けフレームワーク，信学技報，電子情報通信学会，vol. 123，no. 342，2024年1月，https://ken.ieice.org/ken/paper/20240119wcar/

［2］ 中山樹，菊地良将，中西宏和，佐々木勇和，荒瀬由紀，鬼塚真，グラフ深層学習を用いた類似特許検索の精度向上，第16回データ工学と情報マネジメントに関するフォーラム，2024年1月，https://pub-files.atlas.jp/fs/public/deim2024/ver_4/abstract/ja/T2-B-4-02.pdf

※ 上記参考文献に示したURLのリンク先は，2026年4月1日時点での存在を確認．

執筆者紹介 植 木 達 彦（Tatsuhiko Ueki）

2014年大日本印刷(株)入社．インフラエンジニアとしてサーバやネットワークの設計，構築，保守，運用までを担当した後，2023年より先進技術のリサーチ業務，特に企業が保有するデータの利活用についてのリサーチに従事．

牧　原　幸　伸（Yukinobu Makihara）
2011年大日本印刷(株)入社．プリント基板および電装システムの設計開発業務に従事．リハビリテーション用歩行アシストロボットの研究開発を担当した後，ヘルスケア領域のデータ分析技術の研究開発を経て，2024年より先進技術のリサーチ業務に従事．博士（工学）．