メタデータ間の隠れコミュニティ距離を用いた近傍検索

鳥井 修† 白井 智† 金井 達徳†
†（株）東芝 研究開発センター

あらまし近年ドキュメントにメタデータを付与して管理・運用するケースが急増しており，メタデータを有効に活用したドキュメント検索手法に対する需要が高まっている．本稿では，隠れコミュニティ距離と呼ぶ新しい距離定義を導入し，これを利用した近傍検索手法を提案する．隠れコミュニティ距離は，メタデータ値間の距離をメタデータ値が基準の特徴ベクトル間の距離で定義する手法である．隠れコミュニティ距離の導入により，従来型のドキュメントが基準の特徴ベクトルから定義される距離や事前知識から定義される距離では見えてこなかったメタデータ値間の有用な関係を浮かび上がらせ，メタデータに関するユーザーの知識が十分でなくても，所望のドキュメントが発見可能となる．

キーワード メタデータ，近傍検索，データの可視化，メタデータ管理，多次元 DB

1. はじめに

近年ドキュメントに「データについてのデータ」であるメタデータを付与して管理・運用するケースが急増しており，メタデータを有効に活用したドキュメント検索手法に対する需要が高まっている．

１９９４年に米国オハイオ州 Dublin で開催されたメタデータワークショップで，インターネット上のさまざまなリソースを効率的に検索するための汎用的なメタデータが提案され，その中でメタデータの要素として，タイトル，作成者，出版社，データフォーマット，言語など 15 の基本要素が設定された．その内容は，Dublin Core[2]と呼ばれ，現在インターネットや電子図書館における標準的なメタデータとして広く認められている．

例えばウェブの要約をメタデータとして比較的簡単に記述する RSS（RDF Site Summary または Really Simple Syndication）は，Dublin Core の基本要素を公式モジュールとして採用しており，ニュース，日記などのサイト，そして最近急増している Blog（Weblog）で利用されている．

また，ウェブの世界以外でもメタデータ活用が進んでおり，例えば MPEG-7 は，マルチメディア・コンテンツに対するメタデータの表記方法に関する国際標準規格である．その他，電子政府，放送，地理・観光情報，フィルタリング，ユーザプロファイルなど様々な分野において数々のメタデータ規格の実例が存在し，挙げればきりがない．

我々は前述の通り世界的な広がりを見せているメタデータに注目し，メタデータ付きドキュメントを効率よく検索する方式の検討，システムの試作を行ったので，本稿においてその内容について説明する．

メタデータはドキュメントに関する情報を整理したものであるので，これをうまく用いれば効率のよい検索の実現が期待される．しかしながら実際には，規格策定が進んでいる割には検索にメタデータを有効に活用しているケースは少ないのが実情である．これは「ドキュメントにどのような種類のメタデータ項目と，どのような種類のメタデータ値が付与されているかが分かりにくい」ことに原因があると考えられる．

我々はこのような状況を鑑み，近傍検索の手法をメタデータ付きドキュメントの検索に適用することで，メタデータを分かり易くユーザーに提示し，またメタデータに関するユーザーの知識が十分でなくても検索を可能にする技術の確立を目指している．

近傍検索は，ユーザーが必要としている情報に関連する情報があらかじめ分かっている場合に用いられる手法であり，第一に関連情報にフォーカスし，その近傍情報を求め，第二にフォーカスしている情報とその近傍情報の関係を適宜活用しながら，本当に必要な情報を検索する．これをメタデータ検索に当てはめて考えると，現在フォーカスしているメタデータとその近傍メタデータの密接な関係を適宜活用しながら，本当に必要なメタデータを検索する方法と言える．

近傍検索において近傍メタデータを活用して情報を効率よく発見する具体的な方法は以下の二つが挙げられる．

メタデータの可視化：

現在フォーカスしているメタデータとその近傍メタデータとの関係を可視化する．ユーザーは近傍メタデータの中から自分が検索しているメタデータにより関連が深いと判断するメタデータを選択し，フォーカスを移動する，という操作を繰り返し行う．メタデータ集合を目視しながら，本当に必要なメタデータに少しずつ近付いて行くことで，目的のメタデータに到達することが期待される．

検索式の自動展開：

検索の検索条件に指定されているメタデータを，メタデータとその近傍メタデータの和集合に自動的に展

開する．近傍メタデータが自動的に検索条件に追加されるので，ユーザーの手間や知識を必要とせずに，漏れの少ない検索を実現することが期待される．

近傍検索を行う際には，近傍メタデータをいかにして定義するか，言い換えるとメタデータ間の距離をいかに適切に定義するかによって発見効率が大きく左右される．

本稿では，メタデータ間の適切な近傍関係を定義するために，隠れコミュニティ距離と呼ぶ新しい距離定義を導入し，これを利用した近傍検索手法を提案する．隠れコミュニティ距離は，メタデータ値間の距離をメタデータ値が基準の特徴ベクトル間の距離で定義する手法であり，複数の異なる視点に立ったメタデータの近傍関係をユーザーに提供する．隠れコミュニティ距離の導入により，従来型のドキュメントが基準の特徴ベクトルから定義される距離や事前知識から定義される距離では見えてこなかったメタデータ値間の有用な関係を浮かび上がらせ，メタデータに関するユーザーの知識が十分でなくても，所望のドキュメントが発見可能となる．

本研究では，特許検索をモチーフとし，隠れコミュニティ距離を用いた近傍検索システムの試作を行ったので，本稿では試作内容についても合わせて説明する．

2. 近傍検索

近傍検索は，[1]において「ある情報を元にしてその近傍情報を選択するという操作を繰り返して行くことにより必要な情報への到達を可能にする方法のことである」と紹介されている．近傍検索を行えば，ユーザーが検索したいと思っている情報に関する知識が完全でなくても，知識の断片をたどりながら，目的の情報にたどり着くことが可能である．本節では「隠れコミュニティ距離を用いた近傍検索」のベースとなる近傍検索について，[1]に即しながら紹介を行う．

[1]に実例として，ある製品を見た展示会を思い出したい場合について紹介されている．第一に，会場で誰かとその話を議論したことを思い出す．第二に，展示会に一緒に行った人の情報を参照して，それが誰であったか特定する．第三に，その人に関連する写真リストの中から展示会で撮ったものを特定する．第四に，写真を撮った日時から展示会を特定する．以上四段階の近傍検索により，ある製品を見た展示会を特定している．

上記の例は，人間が実生活の中で行っている近傍検索の例であり，誰もが頭の中でこのような関連する情報を一つ一つ順番にたどる検索を常に実行していると思われる．もし同様の近傍検索を計算機上で実現できれば，効率のよい検索を実行することが可能であると期待される．

人間が行っている近傍検索と完全に同一の検索を計算機上で実現することは容易ではないが，適切な設計を行えばその一部は実現可能である．その際，近傍をどのように定義するか，つまり距離をどうやって定義するかが非常に重要であり，近傍距離の定義のよしあしによって目的とする情報へ到達可能であるかどうか大きく左右される．

[1]では，二つのドキュメントの間の距離を，ドキュメントの内容に注目して定義する方法が紹介されている．これは二つのドキュメントがどの程度共通のキーワードを含んでいるかによって決まる距離であり，二つのドキュメントが共通のキーワードを数多く含んでいる場合には二つのドキュメントの距離は小さな値を，逆の場合には大きな値を取る．

具体例として，ドキュメント D1 がキーワードとして K1, K2, K 3 , K4 を含んでおり，同様に D2 が K1, K2, K3 を，D3 が K4 のみを含んでいる場合を考えると，キーワードとドキュメントの関係は，以下に示す 4×3 のキーワード―ドキュメント行列 K で表現可能である．

$$K=\begin{pmatrix}1&1&0\\1&1&0\\1&1&0\\1&0&1\end{pmatrix}.$$

ここで，キーワード―ドキュメント行列 K の ij 成分は，ドキュメント j がキーワード i を含んでいれば 1，含んでいなければ 0 である．

$$d_1=\begin{pmatrix}1\\1\\1\\1\end{pmatrix},d_2=\begin{pmatrix}1\\1\\1\\0\end{pmatrix},d_3=\begin{pmatrix}0\\0\\0\\1\end{pmatrix}\begin{matrix}\leftarrow K1\text{の重み}\\\leftarrow K2\text{の重み}\\\leftarrow K3\text{の重み}\\\leftarrow K4\text{の重み}\end{matrix}.$$

キーワード―ドキュメント行列 K の各列ベクトルは上記の通りであるが，その第 i 成分は，キーワード Ki のこのドキュメントにおける重みを表しており，各列ベクトルは各ドキュメントの(キーワードを基準とした時の)特徴ベクトルになっていると言える．従って，ドキュメント間の距離は，列ベクトル間の距離で計算可能である．例えば，ドキュメント D1 とドキュメント D3 の距離は $1-\cos(d_1,d_3)=0.5$[1]と計算できる．

$$k_1=\begin{pmatrix}1\\1\\0\end{pmatrix},k_2=\begin{pmatrix}1\\1\\0\end{pmatrix},k_3=\begin{pmatrix}1\\1\\0\end{pmatrix},k_4=\begin{pmatrix}1\\0\\1\end{pmatrix}\begin{matrix}\leftarrow \text{D1の重み}\\\leftarrow \text{D2の重み}.\\\leftarrow \text{D3の重み}\end{matrix}$$

また，キーワード―ドキュメント行列 K の各行ベク

[1] 本稿では二つのベクトルの余弦を cos で表す．つまり，$\cos(x,y)=\dfrac{x\bullet y}{|x||y|}$ と定義する．

トルは上記の通りであるが，その第 i 成分は，ドキュメント Di のこのキーワードにおける重みを表しており，各行ベクトルは各キーワードの(ドキュメントを基準とした時の)特徴ベクトルになっていると言える．従って，キーワード間の距離は行ベクトル間の距離で計算可能である．例えば，キーワード K１とキーワード K２の距離は$1-\cos(k_1,k_2)=0$，キーワード K１とキーワード K４の距離は$1-\cos(k_1,k_4)=0.5$と計算できる．

以上本節では，近傍検索の紹介と近傍検索に用いられるドキュメント間，キーワード間の距離の紹介を行った．なお，本節で紹介した距離のことを本稿の以下において単純距離と呼ぶことにする．

3. 隠れコミュニティ距離

本節では，メタデータ値間の関係を測る基準として隠れコミュニティ距離を導入し，導入の効果について説明を行う．

本稿はメタデータ付きドキュメントを対象としている．2節において一般のドキュメントの近傍検索においてキーワード間，ドキュメント間の距離定義が必要であったことと同様に，メタデータ付きドキュメントの近傍検索においてメタデータ値間の距離定義が必要である．メタデータ値間の距離定義が適正になされれば，それを用いてメタデータの可視化や検索式の自動展開を行い，効率的な検索の実現が可能であると期待される．では，メタデータ値間の距離はどのようにして定義すればよいであろうか．各ドキュメントに作成者とジャンルの二種類のメタデータが付与されている場合を具体例に取って考える．

第一の方法として，メタデータ値のターミノロジーとそれらの間の関係を定義する（オントロジー辞書などの）知識データベースを別途準備するやり方がある．この場合には，適切な知識データベースを準備，メンテナンスするためにコストがかかり，そもそも適切な知識データベースが存在しない場合もある．また，知識データベースを利用しても適切な距離が定義できない場合もある．例えばドキュメントの作成者間の関係を定義する知識には，会社や学校の組織図がある．組織図を用いる場合には，組織改変に伴うメンテナンスコストが無視できないことは言うまでもなく問題である．また，組織的には隣同士だが，業務内容や研究内容がかけ離れていてほとんど交流がないことはよくある話であり，組織図を用いて定義される組織や人物間の距離は，現実的な距離になっていない場合が多く，これもまた問題である．

第二の方法として，単純距離を用いるやり方がある．これは2節でキーワード間の距離を定義した方法と同様に，メタデータとドキュメントの関係をメタデータ―ドキュメント行列で表現し，メタデータ値間の距離を行ベクトル間の距離によって定義する方法である．

メタデータ―ドキュメント行列の各行ベクトルは各メタデータ値の特徴ベクトルになっていると言えるので，メタデータ値間の距離を行ベクトルで測る方法は一見よさそうに思えるが，同時に以下に示す問題も含んでいる．

組織の壁:

二人のドキュメント作成者は，業務内容や研究内容が似通っている場合であっても，ドキュメントを数多く共著していない限りお互いの距離が近いとはみなされない．この問題は，作成者が別の組織に属している場合などによく起こることである．

方言・表記の壁:

例えば「ユビキタス」と「パーベイシブ」などのように，また「インターフェース」と「インタフェイス」などのように，意味的には近い（同じ）言葉であっても，同一のドキュメントに対するメタデータ値として同時に付与されない限りお互いの距離が近いとはみなされない．この問題は，意味が近い（同一の）用語が複数の方言や表記を持っている場合であっても，個々のドキュメントにおいてはそれらのうち特定のもののみを用いることが多いことから起こることである．

言語の壁:

例えばカタカナ表記の「ビル・ゲイツ」とアルファベット表記の「Bill Gates」が近いなどのように，意味的には近い（同じ）言葉であっても，同一のドキュメントに対するメタデータ値として同時に付与されない限りお互いの距離が近いとはみなされない．

本稿では，これらの問題を解消するために，第三の方法として，隠れコミュニティ距離と呼ぶ距離定義を導入する．単純距離を用いる第二の方法では，メタデータ値をドキュメントが基準の特徴ベクトルで表現したが，隠れコミュニティ距離では，メタデータ値をメタデータ値が基準である特徴ベクトルで表現する．

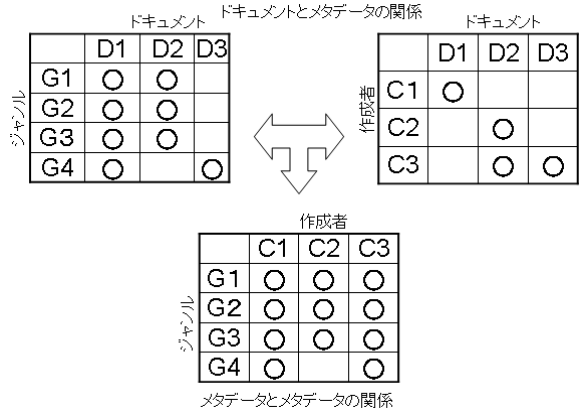

図 1 ドキュメント，メタデータの関係

具体例として，ドキュメント D１のジャンルが G1，G２，G３，G4，作成者が C1，同様に D２のジャンルが

G1, G2, G3, 作成者が C2, C3, さらに D3 のジャンルが G4, 作成者が C3 である場合を考える（図 1）.

この例においてジャンルメタデータとドキュメントの関係を表すジャンルメタデータ―ドキュメント行列 G は以下に示す 4×3 行列で, 作成者メタデータとドキュメントの関係を表す作成者メタデータ―ドキュメント行列 C は以下に示す 3×3 行列でそれぞれ表される.

$$G=\begin{pmatrix}1&1&0\\1&1&0\\1&1&0\\1&0&1\end{pmatrix},C=\begin{pmatrix}1&0&0\\0&1&0\\0&1&1\end{pmatrix}.$$

ここで, G(C)の i,j 成分は, ドキュメント j がジャンル(作成者)メタデータ i を含んでいれば 1, 含んでいなければ 0 である.

これらメタデータ―ドキュメント行列 G, C を用いてジャンルメタデータと作成メタデータの関係を行列表現するためには, G と C の転置行列を掛け合わせる.

$$GC^t=\begin{pmatrix}1&1&1\\1&1&1\\1&1&1\\1&0&1\end{pmatrix}.$$

ここで, この行列の ij 成分はジャンルメタデータ値 i と作成者メタデータ j が同一のドキュメントに付与された回数を表しており, 行列の各行ベクトルは各ジャンルの作成者を基準とした時の特徴ベクトル, 各列ベクトルは各作成者のジャンルを基準とした時の特徴ベクトルになっていると言える. これらより, 作成者 C1, C2, C3 は以下に示す特徴ベクトルで表現可能である.

$$c_1=\begin{pmatrix}1\\1\\1\\1\end{pmatrix},c_2=\begin{pmatrix}1\\1\\1\\0\end{pmatrix},c_3=\begin{pmatrix}1\\1\\1\\1\end{pmatrix}\begin{matrix}\leftarrow\text{「メタデータ値」 G1の重み}\\\leftarrow\text{「メタデータ値」 G2の重み}\\\leftarrow\text{「メタデータ値」 G3の重み}\\\leftarrow\text{「メタデータ値」 G4の重み}\end{matrix}.$$

上記特徴ベクトルを用いて, 例えば C1, C2 の隠れコミュニティ距離は, $1-\cos(c_1,c_2)=0.25$ と計算できる.

本稿では, メタデータ値をメタデータが基準の特徴ベクトルで表現し, メタデータ値間の距離を特徴ベクトル間の距離によって測る距離定義のことを隠れコミュニティ距離と呼ぶことにする. また, この例におけるジャンルのように距離の基準となる項目のことを隠れコミュニティ距離の基準となるメタデータ項目と呼ぶことにする. 一般に, メタデータ項目 M1 の二つのメタデータ値 i, j 間のメタデータ項目 M2 を基準とする隠れコミュニティ距離は以下の方法で求めることができる.

メタデータ M1 とドキュメントの関係を表すメタデータ―ドキュメント行列 X, メタデータ M2 とドキュメントの関係を表すメタデータ―ドキュメント行列 Y から, XY^t を計算し, その第 i 行ベクトルを x, 第 j 行ベクトルを y とする. この時 $1-\cos(x,y)$ が求めるべきメタデータ値 i, j 間の隠れコミュニティ距離である. M1 と M2 が等しい場合もあり得る.

隠れコミュニティ距離という名前は, 前述の事前知識を用いる第一の方法, 単純距離を用いる第二の方法で定義される距離では見えてこなかったメタデータ値間の有用な関係を浮かび上がらせることから名付けたものである. 隠れコミュニティ距離は, 頻繁に同一のドキュメントに登場するメタデータ値間に近い距離を付与する単純な方法とは異なり, あるメタデータ値がどんなメタデータ値と同時に用いられたかに注目して距離を測る方法である. したがって隠れコミュニティ距離は, 同一のドキュメントに一度も同時に登場しないメタデータ値間にも, 相互に関連が深ければ近い距離を付与し, この結果, 組織の壁, 方言・表記の壁, 言語の壁などの問題を解決する.

前記 C1, C2, C3 をドキュメントが基準である特徴ベクトル表現ですると以下の通りになる(これらベクトルは作成者メタデータ―ドキュメント行列 C の行ベクトルに等しい).

$$c_1=\begin{pmatrix}1\\0\\0\end{pmatrix},c_2=\begin{pmatrix}0\\1\\0\end{pmatrix},c_3=\begin{pmatrix}0\\1\\1\end{pmatrix}\begin{matrix}\leftarrow\text{「ドキュメント」 D1の重み}\\\leftarrow\text{「ドキュメント」 D2の重み}\\\leftarrow\text{「ドキュメント」 D3の重み}\end{matrix}.$$

特徴ベクトルから計算した作成者 C1 と作成者 C2 の単純距離は 1 であり, 組織の壁が障害となり, これらの特徴ベクトルから C1 と C2 の密接な関係を読み取ることは困難である. 一方ジャンルメタデータを基準とした時, 作成者 C1 と作成者 C2 の隠れコミュニティ距離は前述の通り 0.25 と計算でき, この値からなら両者が互いに近い関係にあることが読み取れる.

先の例では, 隠れコミュニティ距離の基準となるメタデータ項目としてジャンルを用いたが, 一般にドキュメントに付与されているメタデータ項目が N 種類ある場合には, N 通りの中から任意のメタデータ項目を隠れコミュニティの基準として選択することが可能であり, N 通りの異なる基準を持った距離を定義できる[2]. これは隠れコミュニティ距離が, 複数の異なる基準の距離を提供できることを意味し, ユーザーはそれらの中から検索の趣旨に最も合った基準を自分で選択することが可能である.

また, 定義から分かる通り, 二つのメタデータ値が異なるメタデータ項目に属する場合であっても, これらの間の隠れコミュニティ距離を計算することが可能

[2] 隠れコミュニティ距離の基準として複数のメタデータ項目を選択することも可能であるが, 本稿では簡単のため一項目のみを選択するものとして以下の説明を進める.

である. 例えば, ジャンルメタデータ G1 と作成者メタデータ C1 の間の隠れコミュニティ距離は, G1 と C1 の特徴ベクトルを同一のメタデータ項目(例えば, ジャンルメタデータなど)を基準にして求め, それらの距離を計算することで求めることができる.

以上説明した通り, 隠れコミュニティ距離によると, メタデータに関する事前知識を一切用いることなく, 複数の異なる視点に立ったメタデータの近傍関係を定義可能であり, この結果メタデータ値間の有用な関係を浮かび上がらせることが可能である.

4. 近傍メタデータ

近傍メタデータは, 「あるメタデータ値を中心として, このメタデータから隠れコミュニティ距離が近いメタデータ値の集合である」と定義できる. より厳密には, 以下の四つのパラメータが決まると近傍メタデータ値集合が一意に定まる.

近傍中心:

近傍の中心メタデータ値を何にするかを表す.

近傍半径:

近傍中心から隠れコミュニティ距離が近いメタデータ値のうち何件を近傍メタデータ値として取るかを表す.

近傍対象:

近傍メタデータ値をどのメタデータ項目の中から取るかを表す.

近傍基準:

隠れコミュニティ距離の基準をどのメタデータ項目にするかを表す.

以下これら四つのパラメータが近傍メタデータ値にどんな影響を与えるか説明を行う.

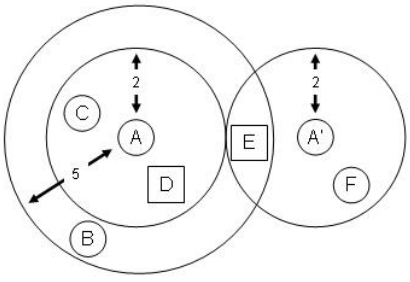

図 2 近傍中心, 近傍半径

図 2は, あるメタデータ項目を基準とした時のメタデータ値の分布を例示したものであり, A, A', B, C, D, E, F はメタデータ値を, メタデータ値を囲む○, □の記号はメタデータ項目をそれぞれ表している.

この図において, 中心 A, 半径 2 の近傍メタデータは C, D である. ここで, 中心は A のままで半径を 4 に変更すると近傍メタデータ値集合は B, C, D, E に変わり, 半径は 2 のままで中心を A'に変更すると近傍メタデータ値集合は E, F に変わる.

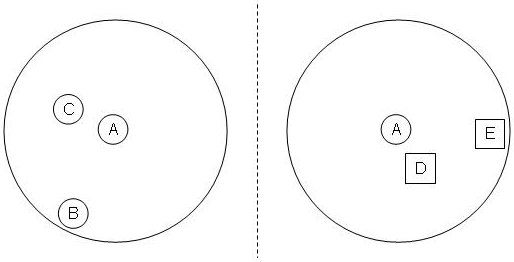

図 3 近傍対象

図 2において, 中心 A, 半径 2 の近傍メタデータは C, D であった. ここで, 中心, 半径ともそのままで, 近傍メタデータ値を取ってくる対象を○メタデータ項目に限定すると, 近傍メタデータ値集合は B, C に変わり, 中心, 半径ともそのままで, 近傍メタデータ値を取ってくる対象を□メタデータ項目に限定すると, 近傍メタデータ値集合は D, E に変わる.

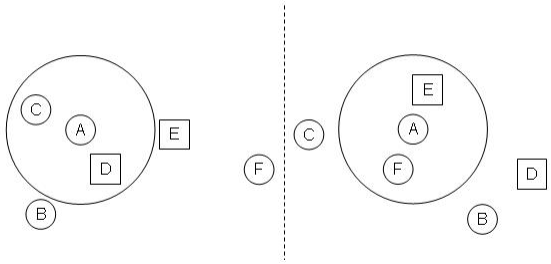

図 4 近傍基準

図 2において, 中心 A, 半径 2 の近傍メタデータ値集合は C, D であった(図 4左). 3節で説明した通り, 隠れコミュニティ距離の基準は複数存在するので, 図 2とは異なるメタデータ項目を基準とすると, 図 2とは異なるメタデータ値の分布が得られる. 例えば, 図 4右に図 2とは異なるメタデータ項目を基準とした分布の一例が示してある. この図において中心 A, 半径 2 の近傍メタデータ値集合は E, F であり, 図 2とは異なる近傍メタデータ値集合が得られる.

以上, 説明を行った通り, 中心, 半径, 対象, 基準を切り替えることで隠れコミュニティ距離をベースとした様々な近傍メタデータ値集合を得ることが可能である. これは各ユーザーが様々な視点に立った検索を実行することが可能であることを意味する. これらの近傍距離を用いた近傍メタデータをデータの可視化や検索式の自動展開に活用すれば, 効率的なメタデータ検索を実現可能であると期待される.

5. 検索システムの試作

前節までに説明を行った隠れコミュニティ距離と近傍メタデータを用い, 特許検索をモチーフとした検索システムの試作を行ったので, 本節ではその説明を行う.

本試作で利用した特許セットの各特許には, メタデータとして, タイトル, キーワード（リスト）, 発明者

（リスト）の情報が付与されているものとする．本検索システムは，隠れコミュニティ距離をベースとする近傍検索システムであり，1節で説明を行ったメタデータの可視化機能や検索式の自動展開機能を保持する．システムの検索画面は，ブラウズパネルとクエリパネルの二つのパネルから構成され，ユーザーはブラウズパネルを用いて(主に)メタデータ値の検索を行い，その結果を用いて近傍中心の指定を実行し，クエリパネルを用いて(主に)近傍半径，近傍対象，近傍基準の指定を実行し，特許ドキュメントの検索を行う．

4節で説明を行った通り，メタデータ値には視点が異なる(近傍中心，近傍半径，近傍対象，近傍基準が異なる)複数の近傍集合が存在し，これらがメタデータ値間の複数種類の関係ネットワークを形成する．ユーザーは，ブラウザパネル上にグラフ描画される関係ネットワーク上を動き回ることで，目的のメタデータやドキュメントに到達可能であると期待される．

関係ネットワーク上を動き回るとは，具体的には，以下に示す三通りの動作を意味する．

メタデータ項目内での近傍中心の移動:

フォーカスをあるメタデータ値から「同一」メタデータ項目の別のメタデータ値へ移動する．

メタデータ項目間の近傍中心の移動:

フォーカスをあるメタデータ値から「別の」メタデータ項目のメタデータ値へ移動する．

近傍基準の変更:

隠れコミュニティ距離の基準となるメタデータ項目を変更する．

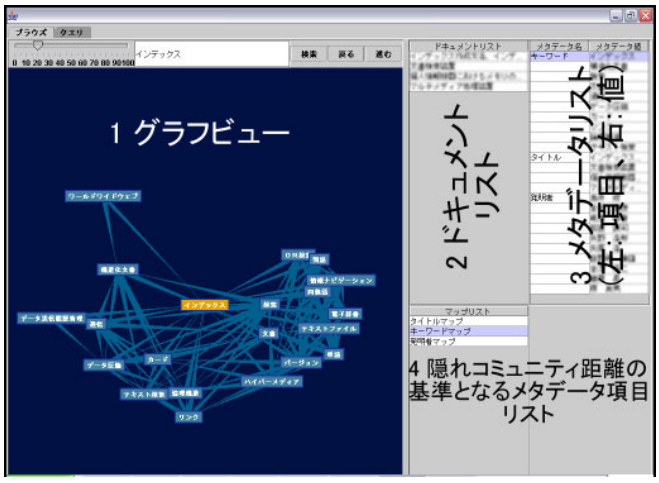

図 5 ブラウズパネル

図 5にブラウズパネルのスナップショットを示した．図 5を適宜参照しながら上記三通りの動作に関して説明を行う．

メタデータ項目内での近傍中心の移動

本試作は，メタデータをグラフ形式で可視化する機能を有する．ブラウズパネルの「グラフビュー」（１）に，現在フォーカスしているメタデータ値から隠れコミュニティ距離が近いメタデータデータ値上位数件を表示する．隠れコミュニティ距離から導出されるこれらメタデータ値の関係ネットワークをグラフ描画し，現在フォーカスしているメタデータ値をグラフの中心に橙色のノードとして配置し，また相互に隠れコミュニティ距離が近いメタデータ値同士を互いに近い位置に配置する．ユーザーは現在フォーカスしているメタデータ値の近傍メタデータ値が何であるか，またそれらのメタデータ値がどのような近傍関係にあるか直感的につかむことが可能である．

このグラフ中の任意のノードをクリックすると，そのノードがグラフの中心に移動し，このノードを中心とするメタデータ値の関係ネットワークを新たにグラフ描画する．このクリック操作が「メタデータ項目内での移動」に相当し，クリックされたメタデータ値にフォーカスが移り，新たにフォーカスしているメタデータ値の近傍関係をグラフより視覚的につかむことが可能である．

なお図 5の例では，キーワード「インデックス」の近傍メタデータ値２０件がグラフ表示してあるが，この件数はユーザーの好みに応じてグラフビュー左上のスライドバーにより調整することが可能である．

メタデータ項目間の近傍中心の移動

ブラウズパネルの「メタデータリスト」（３）に，現在フォーカスしているメタデータ値の関連メタデータ値リストを表示する．

このリスト中の任意のメタデータ値をクリックすると，そのメタデータ値がグラフビュー上でグラフの中心ノードに移動し，このノードを中心とするメタデータ値の関係ネットワークを新たにグラフ描画する．このクリック操作が「メタデータ項目間の移動」に相当し，クリックされたメタデータ値にフォーカスが移り，新たにフォーカスしているメタデータ値の近傍関係を視覚的につかむことが可能である．

なお，現在フォーカスしているメタデータ値の関連メタデータ値リストとは，本試作ではフォーカスしているメタデータ値が付与されているドキュメントに付与されているすべてのメタデータ値を指す[3]．一般に，あるメタデータ値は複数のドキュメントに付与されるので，メタデータリストはこれらドキュメントに付与されているすべてのメタデータ値を列挙する．図 5の例では，キーワードに「インデックス」を含んでいる４つの特許に付与されている２４件のメタデータ値

[3] これ以外にも，「現在フォーカスしているメタデータ値に隠れコミュニティ距離が近い上位数件のメタデータ値のリスト」などが考えられる．

（キーワード１０件，タイトル４件，発明者１０件）を表示している．

近傍基準の変更

ブラウズパネルの「隠れコミュニティ距離の基準となるメタデータ項目リスト」（４）に，特許に付与されているメタデータ項目のリスト(つまりこれが3節でも説明を行った通り，隠れコミュニティ距離の基準である）を表示する．

このリスト中の任意のメタデータ項目をクリックすると，グラフビュー上に，フォーカスは変更せずクリックしたメタデータ項目を近傍基準に変更したメタデータ値の関係ネットワークをグラフ描画する．

このクリック操作が「近傍基準の変更」に相当し，フォーカスしているメタデータ値は変更せず，新しい基準の下でフォーカスしているメタデータ値の近傍メタデータ値が何に変化したか，またそれらのメタデータ値がどのような近傍関係にあるかグラフより視覚的につかむことが可能である．

なお，ブラウズパネルのドキュメントリスト（２）には，現在フォーカスしている（グラフの中心に描画されている）メタデータ値を保持する特許のタイトルが表示される．図 5の例では，キーワードが「インデックス」である特許４件が表示されている．本試作では，タイトルをクリックすることで，特許データベースに接続し，明細書を閲覧することも可能である．

以上説明を行った通り，隠れコミュニティ距離から形成される関係ネットワーク上を（１）メタデータ項目内での近傍中心の移動，（２）メタデータ項目間の近傍中心の移動，（３）近傍基準の変更，という三種類の操作を用いて動き回ることで，目的のメタデータ値を効率よく発見可能であると期待される．

本試作から3節で説明を行った組織の壁，方言・表記の壁，言語の壁などの問題が隠れコミュニティ距離の導入によって解決されることが確認できた．キーワード項目を基準とする隠れコミュニティ距離は，組織を超え業務内容や研究内容が近い人物を明らかにする．例えば，筆者自身にフォーカスした場合の近傍検索を実行してみることで，一緒に業務を行っているメンバーに加えて，名前は初めて聞くが業務内容が近い人物を発見できた．また，キーワード項目を基準とする隠れコミュニティ距離は，方言・表記・言語を超え意味が近いキーワードも明らかにする．例えば，ワールドワイドウェブと WWW など同一のドキュメントではあまり同時には用いられないが，意味が同一のキーワードの近傍関係を発見できた．

4節で説明を行った通り，メタデータ値には視点が異なる(近傍中心，近傍半径，近傍対象，近傍基準が異なる)複数の近傍メタデータ値集合が存在する．本システムは検索の近傍中心に指定されているメタデータ値を，メタデータ値とその近傍メタデータ値の和集合に自動的に展開するが，この際ユーザーが，クエリパネル上に表示される近傍パラメータを調整することで，目的のメタデータやドキュメントを少ない漏れで絞り込むことが可能であると期待される．

近傍パラメータを調整するとは，具体的には，以下に示す四通りの動作を意味する．

近傍中心の追加，削除:

検索式に新たなメタデータ値を近傍中心として追加し，また検索式からメタデータ値を削除する．

近傍半径の変更:

隠れコミュニティ距離が近いメタデータの中から何件選択するかを変更する．

近傍基準の変更:

隠れコミュニティ距離の基準となるメタデータ項目を変更する．

近傍対象の変更:

隠れコミュニティ距離が近いメタデータの中からどのメタデータ項目を選択するか変更する．

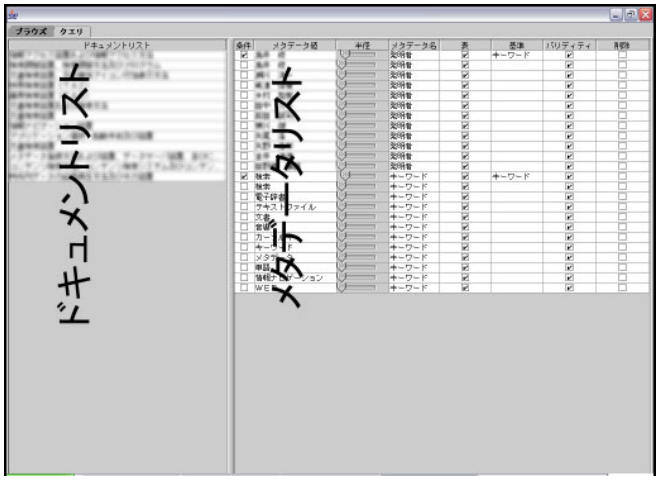

図 6 クエリパネル

図 6にクエリパネルのスナップショットを示した．図 6を適宜参照しながら上記四通りの動作に関して説明を行う．

近傍中心の追加・削除

クエリパネル右のメタデータリストに検索条件のメタデータ値のリストを，左のドキュメントリストに検索結果の特許リストを表示する．

本検索システムは，近傍中心に指定されたメタデータ値を近傍メタデータ値集合に自動展開した上で検索を実行する機能を保持する．

図 6の例は，検索式の近傍中心に発明者「鳥井修」とキーワード「検索」を指定した結果である．発明者「鳥井修」は「鳥井修」に隠れコミュニティ距離が近い上位１０人の発明者に自動展開され，またキーワー

ド「検索」は「検索」に隠れコミュニティ距離が近い上位１０件のキーワードに自動展開され，その結果１１人の発明者によって発明され，かつ１１件のキーワードを持つ特許である１３件が列挙してある．

検索結果は，ユーザーが直接近傍中心に指定したメタデータ値に合致する特許に加えて，指定されたメタデータ値の近傍メタデータに合致する特許も含むため，漏れの少ない検索を実現する．

検索式の近傍中心にメタデータ値を追加するためには，ブラウズパネルのグラフビューで，そのメタデータ値を右クリックする．再度右クリックすると検索条件から削除される．

このクリック操作が「近傍中心の追加・削除」に相当し，単純なマウス操作により検索条件の調整が可能である．

近傍半径の変更

近傍中心に指定されたメタデータ値を近傍メタデータ値集合に自動展開する際，ユーザーは，クエリパネルのメタデータリスト中のスライドバーをスライドすることで，近傍中心に指定したメタデータ値を何件（図 6の例では１０件）まで自動展開するか調整することが可能である．

このスライド操作が「近傍半径の変更」に相当し，目的の特許に付与されているメタデータ値が事前にはっきりと分かっている場合には近傍半径を小さく指定し（０でもよい），逆に目的の特許に付与されていると期待されるメタデータ値の揺らぎが大きい場合には近傍半径を大きく指定するとよい．

近傍基準の変更

近傍中心メタデータ値を近傍メタデータ値集合に自動展開する際，ユーザーは，クエリパネルのメタデータリスト中の「基準」項目をクリックすることによって，近傍基準となるメタデータ項目を調整することが可能である．

このクリック操作が「近傍基準の変更」に相当し，メタデータ値はそのままで，近傍基準を新しいメタデータ項目に切り替える．

近傍対象の変更

近傍中心に指定されたメタデータ値を近傍メタデータ値集合に自動展開する際，ユーザーは，クエリパネルのメタデータリスト中の「表」項目をクリックすることによって，近傍メタデータ値をどのメタデータ項目の中から選択するか調整することが可能である．

このクリック操作が「近傍対象」の変更に相当し，クエリパネルのメタデータリスト上で，メタデータ値の「表」フィールドをクリックすることで，近傍対象を近傍中心と同じメタデータ項目，または近傍基準のメタデータ項目に切り替える．

以上説明を行った通り，隠れコミュニティ距離を用いて構成される検索式を（１）近傍中心の追加・削除，（２）近傍半径の変更，（３）近傍基準の変更，（４）近傍対象の変更，という四種類の操作を用いて調整する．ユーザーは似通ったメタデータ値を一つ一つ自力で検索条件に加えていく手間が省け，何らかの方法で一つだけメタデータ値を指定すれば，それに類似するメタデータ値を複数指定した効果を生む．例えば，ある人物の業務内容や研究内容と関連が深い特許リストを検索する場合に，この人物がどんな業務や研究に携わっているかを確認することなく，隠れコミュニティ距離を用いて特許を発見できた．

ブラウズパネルとクエリパネルはともにドキュメントリストを保持するが，両者の間には違いがある．ブラウズパネルのドキュメントリストには，現在フォーカスしている単一のメタデータ値を保持する特許のタイトルが表示される．一方，クエリパネル左には，複数の近傍メタデータ値によって絞りこまれたドキュメントリストが表示される．

以上説明を行った通り，試作した特許検索システムにおいて，ユーザーはブラウズパネルとクエリパネル二つのパネルへのインタラクティブな操作を組み合わせることにより，隠れコミュニティ距離から形成されるメタデータ値の関係ネットワーク上を動き回り，また隠れコミュニティ距離から構成される検索式を調整することで，目的のメタデータやドキュメントを効率的に発見することが可能であると期待される．

6. まとめ

近年のメタデータを有効に活用したドキュメント検索手法に対する需要の高まりを鑑み，本稿では隠れコミュニティ距離と呼ぶ新しい距離定義を導入し，これを利用した近傍検索手法を提案した．隠れコミュニティ距離の導入により，従来型のドキュメントが基準の特徴ベクトルから定義される距離や事前知識から定義される距離では見えてこなかったメタデータ値間の有用な関係を浮かび上がらせ，メタデータに関するユーザーの知識が十分でなくても，所望のドキュメントの検索が可能となると期待される．現在のところ特許検索をモチーフとした検索システムの試作を行っているが，隠れコミュニティ距離導入の効果に関してはまだ評価が十分に行われていない．今後は，各種データに対して本手法を適用し，定量的な評価を行う予定である．

文　献

[1] 増井俊之，近傍関係を活用した情報検索，情報処理学会研究報告，Vol. 2003-HI-104，pp. 53–58 (2003)

[2] Dublin Core Metadata Initiative, http://dublincore.org/.