文書-単語双クラスタリングを用いた 特許データの概念検索性能向上手法について

青野 雅樹† 土肥 広典‡

豊橋技術科学大学情報工学系 〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1

E-mail: †aono@ics.tut.ac.jp, ‡doi@kde.ics.tut.ac.jp

あらまし 特許データの検索では，先願特許（先例）を漏れなく検索できる再現率を優先する概念検索の有効性が指摘されている．しかし，概念検索では一般に検索質問が長文であることが期待されており，少数の単語（索引語）からでは，検索における再現率を向上するのは難しいとされている．本報告では，ベクトル空間モデルを用いるが，従来型の次元削減手法による概念検索と異なり，前処理として「文書クラスタリング」と「単語クラスタリング」の両方を同時に行う「双クラスタリング」を複数回，異なるクラスタ粒度で実行し，これより「クラスタ粒度階層構造」を作る．この後，クラスタ粒度階層構造を用いて検索質問拡張を行うことで，再現率を向上させる仕組みを検討した．本手法と従来手法(LSI, VSM)での特許データを用いた比較実験をあわせて報告する．

キーワード 概念検索，双クラスタリング， 検索質問拡張

On Improving Conceptual Search for Patent Data Using Co-clustering

Masaki AONO† and Hironori DOI‡

Information of Computer Sciences Department, Toyohashi University of Technology

1-1 Hibarigaoka, Tempaku-cho, Toyohashi-shi, Aichi-ken, 441-8580 Japan

E-mail: †aono@ics.tut.ac.jp, ‡doi@kde.ics.tut.ac.jp

Abstract Co-clustering is a method for producing both document and term clusters simultaneously, assuming abundance of local “co-occurrence” of a set of documents and the associated set of terms. Patent data can be regarded as one such example. We present an algorithm to generate a hierarchy of co-clusters (document and term clusters) for enhancing conceptual search with query expansion using the hierarchy. We also describe a comparative study of our proposed method with an LSI (Latent Semantic Indexing), one of dimensionality reduction technologies, as a typical example of previous methods for conceptual search.

Keyword Conceptual search，Co-clustering, Query expansion

1. はじめに

特許データや論文などの文献データに関する情報検索では，検索質問のキーワード列と完全に一致する全文検索と異なり，比較的長文からなる自然文を与えて内容的な類似性から検索を行う概念検索技術の有効性が指摘されている．概念検索では，入力となる検索質問を，通常，単語列に分解し，その単語列と類似するデータを検索対象とする．

本報告では，日本語特許データを検索対象とし，ベクトル空間モデルに基づく概念検索に焦点をあて，前処理として「双クラスタリング」(co-clustering)を複数回，異なる「クラスタ粒度」で適用し，得られたデータから「クラスタ粒度階層構造」を作成し，このデータ構造に基づく概念検索の検索性能を向上する手法に関して述べる．クラスタ粒度に関しては，双クラスタリングで出力されるクラスタ数を 2 のべき乗（たとえば 16, 32, 64, 128, …）で変更しながら実行した．その上で，特異値分解等による次元削減技術に基づく従来型の概念検索技術と，本手法との比較実験を行った．

1.1. 概念検索の関連技術

特許データに対する検索を中心とした手法は，情報学研究所(NII)にて，NTCIR ワークショップの「特許タスク」[16]として，多くの企業・大学等の研究者が参加して行われている．この背景には，特許検索技術が通常の Web 検索技術と異なり，適合率よりも再現率が重視されるという側面，1 文書あたりのデータ量が Web 検索における Web ページよりもかなり大きいという側面，さらに特許においては，同一の対象物であっても，様々な戦略に基づき，意図的に異なる単語や表現で記述される場合が多いという歴史的な側面がある．

概念検索のツールとして，ベクトル空間モデルの分野でもっとも標準的に使用されるツールは，特異値分解や固有値分解などの線形代数の理論に基づく，LSI (Latent Semantic Indexing) [11] と PCA (Principal Component Analysis) [1,14]である．これらは，いずれも単語の次元を削減する手法として有用であり，次元削減後，もともと言葉そのものは異なるが，意味的・概念的に近い言葉が，低次元空間での「距離」として近くなるという性質に着目した技術である．ただし，多義語(polysemy)に関しては，必ずしも低次元空間で分離できない場合があること，またどの程度の次元に削減すれば，本来の同義語（たとえば「もみじ」と「かえで」など）が近接する理想的な次元削減となるかなど，次元数(k)はアドホックに決められることが多い．

Hoffmann[13]は LSI を改良し，文書と単語の（潜在的な）共起度合いに着目し，“aspect model”（観点モデル）と呼ばれる統計学的なモデルを提案した．このモデルは PLSI (Probabilistic LSI)モデルと呼ばれる．PLSI は，LSI の拡張のひとつであるが，文書と単語の共起に着目した点は興味深い．ただし，潜在的な変数などを仮定するため，潜在変数のパラメータの設定方法，計算時間を要すること，および実装の難しさなどの問題がある．

1.2. クラスタリングの検索への利用関連技術

クラスタリングを用いて情報検索を支援する研究は，90 年代後半から徐々に報告されている．Cutting ら[7]は，検索結果に k-平均アルゴリズムで排他的クラスタリングを行い，検索結果のグルーピングに利用した．江口ら[12]は，Cutting らの使用したクラスタリング手法を用いて，検索結果をグループ化したあと，適合性フィードバックを改良し，検索性能の向上が得られたことを報告している．クラスタリングで検索質問拡張を行う同様のアプローチは，新田ら[5]によって，1000 件程度のデータに対して実験した報告がなされている．

最近では，佐々木ら[3]がランダム射影法による次元削減での検索精度向上を目的とした球面 k 平均アルゴリズムを利用するアプローチや，Chang ら[6]による，前処理として文書データから特徴抽出をしたあと，特徴をクラスタリングして，質問拡張を自動化するアプローチなどが報告されている．なお，検索への応用は言及していないが，Dhillon と Modha [8]は，球面 k 平均アルゴリズムをもとに，各クラスタの「概念ベクトル」を与え「概念」を分割する手法と LSI による次元削減手法を比較している．非排他的クラスタリングを Web 文書検索に利用したものとしては，成田ら[4]の手法が報告されている．また，小西らは，特許の特徴を抽出して検索性能を向上するアプローチ[2]や CDF-ICF（Category Document Frequency － Inverse Category Frequency）を利用した検索手法[15]を報告している．

2. 双クラスタリング

従来の検索性能の向上を目的としたクラスタリングを用いる手法では，クラスタリングはもっぱら文書クラスタリングに限られていた．我々は，特定用途の文書コレクション，とりわけ特許文書に関しては，文書と単語の共起性が高いと考え，Dhillon ら[9,10]の提案した双クラスタリング(co-clustering)を用いて，文書クラスタと単語クラスタの両面から，概念検索の検索性能向上に使う手法を提案する．本節では，双クラスタリングの概説を述べ，次節でこれを拡張した新しいデータ構造に基づく概念検索手法を論じる．

2.1. 双クラスタリングの定義

文書集合 $X=\{x_1,x_2,...,x_m\}$ と単語（索引語）集合 $Y=\{y_1,y_2,...,y_n\}$ が与えられたとき，X と Y をランダムな変数と考える．それぞれのデータ集合の共起を考えると，結合確率密度関数 $p(X,Y)$ を，共起する単語と文書の(正規化された)行列で与えることができる．$\hat{X}=\{h_1,h_2,...,h_s\}$ を s 個のクラスタからなる文書クラスタとし，$\hat{Y}=\{g_1,g_2,...,g_t\}$ を t 個のクラスタからなる単語クラスタとする．$\pi_x \equiv p(x)$ を文書 x が現れる確率分布関数，$\pi_y \equiv p(y)$ を単語 y が現れる確率分布関数とする．

クラスタに関しては，$\pi_h \equiv p(h)=\sum_{x\in h} p(x)$ を文書クラスタ h の確率分布関数，$\pi_g \equiv p(g)=\sum_{y\in g} p(y)$ を単語クラスタ g の確率分布関数とする．文書 x が与えられたとき，単語集合 Y の出現する確率密度関数を $z(x)=p(Y\mid x)$ と表すことにする．同様に，文書クラスタ h が与えられたとき，単語集合 Y の出現する確率密度関数を $\mu_h=p(Y\mid h)$ と表すことにする．

双クラスタリングでは，文書と単語の共起性に着目し，相互情報量を考え，$I(X,Y)-I(\hat{X},\hat{Y})$を最小とする文書クラスタ$\hat{X}$と単語クラスタ$\hat{Y}$を求める．ただし，

$$I(X,Y)-I(\hat{X},\hat{Y})=KL(p(X,Y)\,\|\,q(X,Y))$$

である．上式の右辺の$KL(\cdot\|\cdot)$はカルバック・ライブラー距離で，$q(x,y)=p(h,g)p(x\mid h)p(y\mid g)$である．なお，$p(h,g)=\sum_{x\in h}\sum_{y\in g}p(x,y)$である．

2.2. 双クラスタリングの意義

双クラスタリングでは，クラスタリングする前と，行った後の相互情報量の差を最小化することで，文書クラスタと単語クラスタを同時に生成する．これを文書×単語行列で考えると，相関の高い行と列の局所的な部分行列成分が，双クラスタリングによって低次元の相関の高い部分空間の集合で近似できることを意味する．これは，従来の特異値分解による行列の次元削減手法と対照的である．特異値分解でk次元に次元削減する場合，近似される行列を展開すると，その影響は大域的であり，行列の要素には，通常，正負の値が混在する密行列となる．これに対して，双クラスタリングでは，近似の影響は局所的であり，行列の要素の疎率を保ち，かつ，負の値にならないという特徴を持つ．元来，TF-IDF に代表される重み付けされた文書×単語行列の要素に負の値はないので，双クラスタリングの上記の性質は良好な特徴であると考えられる．

3 提案手法

双クラスタリングの良好な性質は，文献データや特許データなどの概念検索にも有効であると期待される．たとえば特許データの場合，すべての特許には，特許出願分野情報（国際特許分野[International Patent Classification (IPC)]）が付与されており，IPC が同一の特許では，その分野特有の専門用語が頻出する傾向が高いことからも双クラスタリングの有効性が推察される．たとえば，IPC のサブクラス“G11B”では，「ディスクカートリッジ」，「記録担体」，「消磁」といった固有の用語が頻出する．

3.1. クラスタリングの問題点とその対応策

クラスタリングは，対象とするデータによっては有効であることが期待されるが，手法にかかわらず，クラスタリングの結果をそのまま利用すると，必ずしも検索性能を向上できない場合がある．代表的なのは，以下の場合である．

① 生成するクラスタの数を不適切(多すぎ，もしくは少なすぎ)に選択した場合
② クラスタ初期化における乱数の値により，生成されるクラスタの品質が著しく変化する場合
③ クラスタ情報に潜在的に含まれるノイズを含んだまま，検索の支援に使用した場合

これらの代表的なクラスタリング誤使用を避けるため，我々は以下のような対応策を講じた．

①の問題を緩和するために，クラスタ数を 2 のべき乗（たとえば$16,32,64,128,\cdots$）で変更させ，粒度の異なるクラスタリングを実施することで，与えられた特許文書集合全体から判断して，ノイズとなる文書や単語以外は，どこかの粒度のクラスタで吸収させるようにした．また，検索質問拡張時に使用するデータ構造として，粒度の異なるクラスタ間で，適当な閾値以上の類似度を有するクラスタ同士にリンクをつけ，「クラスタ粒度階層構造」を作成した．

②の問題を緩和するために，①で述べたそれぞれの粒度で乱数の初期値を変更し，双クラスタリングを複数回実行し，同一のクラスタに含まれる確率の高い文書同士，単語同士を最終的に同一クラスタに帰着するような平滑化アルゴリズム(図 1 のアルゴリズム参照)を考案した．

③の問題を緩和するために，②で述べた平滑化アルゴリズムで，どこにも含まれない文書は，ノイズ（アウトライヤー）として棄却した．検索質問拡張においては，検索質問が与えられたとき，①で述べた粒度の粗い順に，文書クラスタを代表する単語列とその出現頻度に基づく重みの列（これを「クラスタ平均ベクトル」と呼ぶことにする）と類似度が高く，かつ，その類似度がある閾値を超えている場合のみ，検索質問拡張を行った．上記の閾値未満である場合は，文書クラスタベクトルは使用せず，その代わり，粒度のもっとも細かい双クラスタリングで得られた単語クラスタの中で検索質問ともっとも類似する単語列をもとに検索質問拡張を行った．

②と③の処理により，もともとは排他的なクラスタリングであっても，ノイズに分類される文書はどの文書クラスタにも属さなくなる．一方，単語としては，複数の文書クラスタベクトルや単語クラスタに出現することを許しているため，非排他的なクラスタリングを行ったことになる．

3.2. 提案手法

前節の観察と，クラスタリングにおける問題点の対応策に基づき，双クラスタリングに基づく，検索質問拡張法を考案した．以下に提案手法を述べる．

「アルゴリズム A」

[step1] クラスタ粒度 M，（乱数の）初期値を変更して R 回，双クラスタリングを実行する．得られた文書クラスタを $\mathbb{D}^{r} = \{\mathbf{D}_1^r, \mathbf{D}_2^r, ..., \mathbf{D}_M^r\}(r = 1, ..., R)$，単語クラスタを $\mathbb{W}^{r} = \{\mathbf{W}_1^r, \mathbf{W}_2^r, ..., \mathbf{W}_M^r\}(r = 1, ..., R)$ と表現する．ここで，文書クラスタ $\mathbf{D}_j^r$ は，単語クラスタ $\mathbf{W}_j^r$ と対応するものとする．

[step2] 文書クラスタベクトル $\mathbb{H} = \{\mathbf{H}_1, \mathbf{H}_2, ..., \mathbf{H}_M\}$ を $\mathbb{D}^1$ で初期化する．同様に単語クラスタベクトル $\mathbb{G} = \{\mathbf{G}_1, \mathbf{G}_2, ..., \mathbf{G}_M\}$ を $\mathbb{W}^1$ で初期化する．

[step3] 図 1 のクラスタ平滑化アルゴリズムを実行する．ただし，$\mathbf{K(j)}$ は，r 回目の双クラスタリングで得られた文書クラスタ $\mathbf{D}_j^i$ の要素数を表し，$\mathbf{L(j)}$ は，r 回目の双クラスタリングで得られた単語クラスタ $\mathbf{W}_j^i$ の要素数を表す．

Algorithm ClusterSmoothing($\mathbb{H}, \mathbb{G}, R, M$)

1: **for** $r = 2$ **to** R
2: **for** $j = 1$ **to** M
3: $\mathbf{D}_j^r = \{\mathbf{d}_{j,1}^r, \mathbf{d}_{j,2}^r, ..., \mathbf{d}_{j,K(j)}^r\}$;
4: $\mathbf{W}_j^r = \{\mathbf{w}_{j,1}^r, \mathbf{w}_{j,2}^r, ..., \mathbf{w}_{j,L(j)}^r\}$;
5: **for** $i = 1$ **to** $|\mathbb{H}|$
6: $\mathbf{H} = \mathbf{H}_i$; $\mathbf{D} = \mathbf{D}_j^r$; $\mathbf{W} = \mathbf{W}_j^r$;
7: **if** (Similarity($\mathbf{H}, \mathbf{D}$) > δ) **then**
8: $\mathbf{H}_i$ = Average($\mathbf{H}, \mathbf{D}, \mathbf{W}$);
9: **else** /* 追加 */
10: $\mathbb{H} = \mathbb{H} \cup \mathbf{D}$;
11: $\mathbb{G} = \mathbb{G} \cup \mathbf{W}$;
12: **endif**
13: **end for**
14: **end for**
15: **end for**

図 1．クラスタ平滑化アルゴリズム

提案手法では，まず異なる粒度 $M_1, M_2, ..., M_k$ で，それぞれ R 回双クラスタリングを実行する．（「アルゴリズム A の[step1]」．こうして得られた文書クラスタと単語クラスタから，「アルゴリズム A」の[step2]を適用して，拡大文書クラスタ $\mathbb{H}$ と単語クラスタ $\mathbb{G}$ を得る．この後，[step3]でクラスタの平滑化アルゴリズムを適用する．図 1 の 7 行目の Similarity 関数は，クラスタベクトル $\mathbf{H}$ と文書ベクトル $\mathbf{D}$ との類似度を計算する．類似度がある閾値(δ)より大きければ，8 行目にある Average 関数で，クラスタベクトルを文書ベクトル $\mathbf{D}$ と単語ベクトル $\mathbf{W}$ をもとに，その方向を以下の式で更新する．

$$\mathbf{H}_i = \sum_i (\mathbf{h}_i + \mathrm{merge}(\mathbf{d}_i, \mathbf{w}_i))$$

$$\mathrm{merge}(\mathbf{d}_i, \mathbf{w}_i) = \lambda \mathbf{d}_i + \mu \mathbf{w}_i$$

$$d_i = \sum_{j=1}^{D(i)} \alpha_j w_j$$

$$\mathbf{w}_i = \sum_{j=1}^{W(i)} \beta_j w_j$$

$$\mathbf{H}_i = \frac{\mathbf{H}_i}{\|\mathbf{H}_i\|}$$

ただし，$D(i), W(i)$ は，それぞれ文書クラスタと，それと双対な単語クラスタから追加する単語数の上限値を表し，$\alpha_j, \beta_j, \lambda, \mu$ は非負の実数パラメータである．

乱数の初期値によらないクラスタの平滑化がなされた後，「アルゴリズム B」により，異なる粒度で得られたクラスタ間に階層関係を構築する．なお，「アルゴリズム B」は，文書クラスタをもとに行う．

「アルゴリズム B」

[step1] 粒度の粗い順に，M_i と M_{i+1} の間で $\mathbb{H}_i$ と $\mathbb{H}_{i+1}$ の間で相互クラスタ類似度を，クラスタ平均ベクトルの類似度より求める．

[step2] もし，[step1]で，$\mathbb{H}_i$ の中の文書クラスタ $\mathbf{D}_i$ と $\mathbb{H}_{i+1}$ の中の文書クラスタ $\mathbf{D}_{i+1}$ の類似度がある閾値より大きければ，$\mathbf{D}_i$ と $\mathbf{D}_{i+1}$ の間にリンクをつける．

[step3] [step2]を最も粒度の細かいクラスタ $\mathbb{H}_k$ まで繰り返す．

図 2 は，（アルゴリズム B）により得られたクラスタ階層の例である．粒度の粗いクラスタで検出されるクラスタは，そのクラスタを構成するデータ集合の要素数が大きなクラスタであることが多く，たとえば，粒度=16 では，「画像，画像データ」という単語に代表されるクラスタ，「細胞，酵素」という単語に代表されるクラスタ，「燃料，エンジン」という単語に代表される

クラスタ,「変速，ブレーキ」という単語に代表されるクラスタなどが検出される．

粒度が細かくなると，粗い粒度では検出できなかったクラスタが検出されるようになる．たとえば，粒度=16 では検出できなかった「遊技，パチンコ」という単語に代表されるクラスタが粒度=32 で新たに検出されたり，粒度=64 になると,「免振，免振装置」という単語に代表されるクラスタが新たに検出されたり，粒度=128 では,「ボール，ゴルフ」という単語に代表されるクラスタが新たに検出されたりする．

一方，粗い粒度で検出されていたクラスタの一部は，2 つ以上のサブクラスタに分割されて検出される場合がある． たとえば，粒度=64 の「油圧，ピストン」という単語に代表されるクラスタは，粒度=128 の「油圧，ブレーキ」という単語に代表されるクラスタと,「ピストン，シリンダ」という単語に代表されるサブクラスタに分割される．

逆に，粗い粒度の２つ以上のクラスタが細かい粒度のひとつのクラスタにリンクを持つ場合も生じる．たとえば，粒度=32 の「画像，画像データ」という単語に代表されるクラスタと，同じ粒度の「画像，撮像」という単語に代表されるクラスタは，粒度=64 の「画像，原稿」という単語に代表されるクラスタにともにリンクを有する．

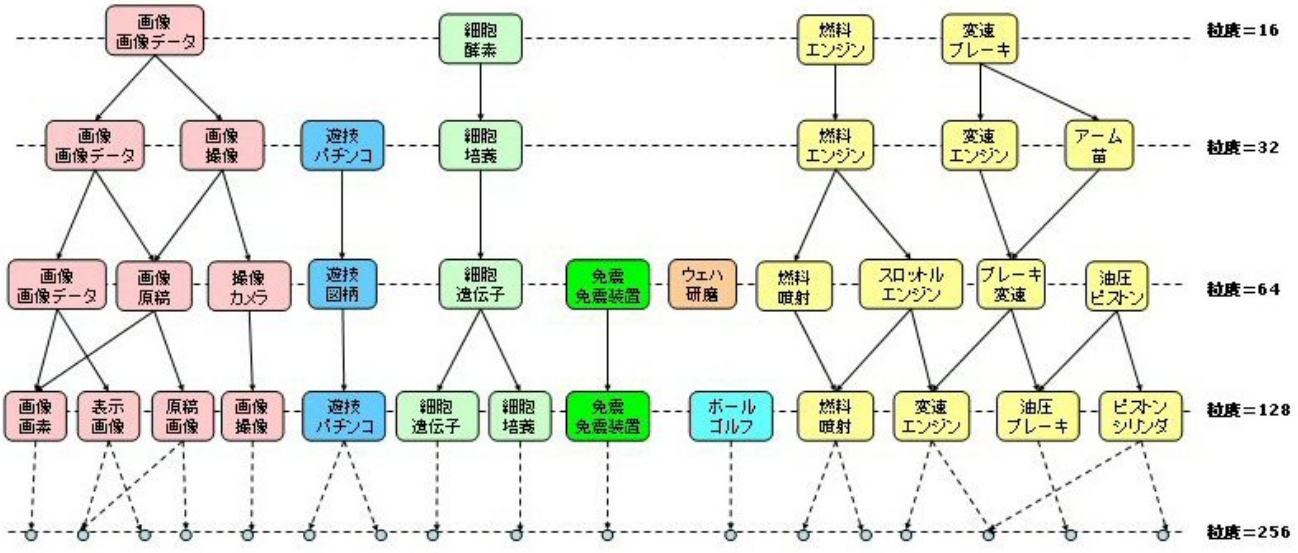

図２．異粒度の双クラスタリングを複数回行い，平滑化後，生成されたクラスタ粒度階層構造の一部．

一般に，クラスタ間の類似度が高い粒度の異なるクラスタ同士にリンクをつけると，図２のような階層構造が得られる．

「クラスタ粒度階層構造」が得られたら，検索質問に対して，この階層構造を粒度の粗い方から順に検索質問ベクトルとクラスタ平均ベクトルとの類似度を計算する．もし類似度が閾値以上であれば，そのクラスタの平均ベクトルを構成する単語と重みを加えて検索質問拡張を行う．粗い粒度の順に行うのは，一般に粒度の粗いクラスタの要素数が多いため，再現率を重視する応用では，より類似したデータを検索しやすくするためである．

閾値以上のクラスタが見つからない場合は，もっとも粒度の細かい単語クラスタの中で，検索質問ベクトルともっとも類似する単語列で検索質問拡張を行う．

4. 実験結果

実験には，NTCIR-3 に含まれる 1998 年の特許文書約 33 万件より，ランダムに選んだ 2 万文書を用いた．比較として，素朴なベクトル空間モデル(VSM)，LSI による次元削減法(k=128,256,512)，および本提案手法の 5 つの方法での再現率･適合率を描画した(図 5-9)．検索質問としては，以下の 5 種類の課題を用いた．

(1) 「遊技機，パチンコ」に関する検索質問
(2) 「田植機，移植機」に関する検索質問
(3) 「免振，耐震」に関する検索質問
(4) 「遺伝子，DNA 組換え」に関する検索質問
(5) 「生ゴミ，ゴミ処理機」に関する検索質問

(1) の検索質問に関しては，特許の IPC のサブクラス “A63F”に対応する特許文書のうち，ゲーム機や図柄表示装置などを除く 120 文書を正解集合とし

た．結果は，図 5 に示すように本提案手法がもっともよい検索性能を示した．LSI に関しては，3 種類ともほぼ同様の性能となった．

(2) の検索質問に関しては，特許の IPC のサブクラス“A01C”に対応する特許文書のうち，田植や種苗などの耕作機に関する 42 文書を正解集合とした．結果は図 6 に示すように，本提案手法がもっともよい性能を示した．LSI では，k=256 のものがもっともよい結果を示した．

(3) の検索質問に関しては，特許の IPC のサブクラス“E04H”, “E04G”, “F16F”, “B32B, “H01L”, “E02D”, “E04F” , “E01D, “B63B” , “E04C”など多岐にわたるサブクラスの中に耐震，免振に関する様々な装置や建築物，部品等に関する特許文書があり，合計 78 文書を正解集合とした．結果は図 7 に示すとおりである．この質問では，複数の IPC に正解が分散するケースであり，本提案手法がもっともよい結果を示した．

(4) の検索質問に関しては，特許の IPC のサブクラス“C12N”, “C07K”,および“A61K” の中の遺伝子や DNA 組換え等に関する 58 文書を正解集合とした．結果は，図 8 に示すようであり，本提案手法がおおむね良好だが，一部 LSI(k=128)がもっともよい性能を示す場所が観測された．LSI でも，k=256 と k=512 の性能は，VSM より性能で劣っていた．図 3 は，この検索質問に関係する粒度=64 の文書および単語クラスタの内容を示したもので，図 2 のクラスタ階層構造における「細胞，遺伝」（粒度=64）に対応するものを表す．

(5) の検索質問に関しては，特許の IPC のサブクラス“B09B”, “B02C” , “C05F”,および“F23G” の中の生ゴミ処理機に関する 41 文書を正解集合とした．結果は図 9 に示すように提案手法がおおむね良好な性能を示したが，再現率が大きい部分では，一部の LSI の手法が優る結果となった．

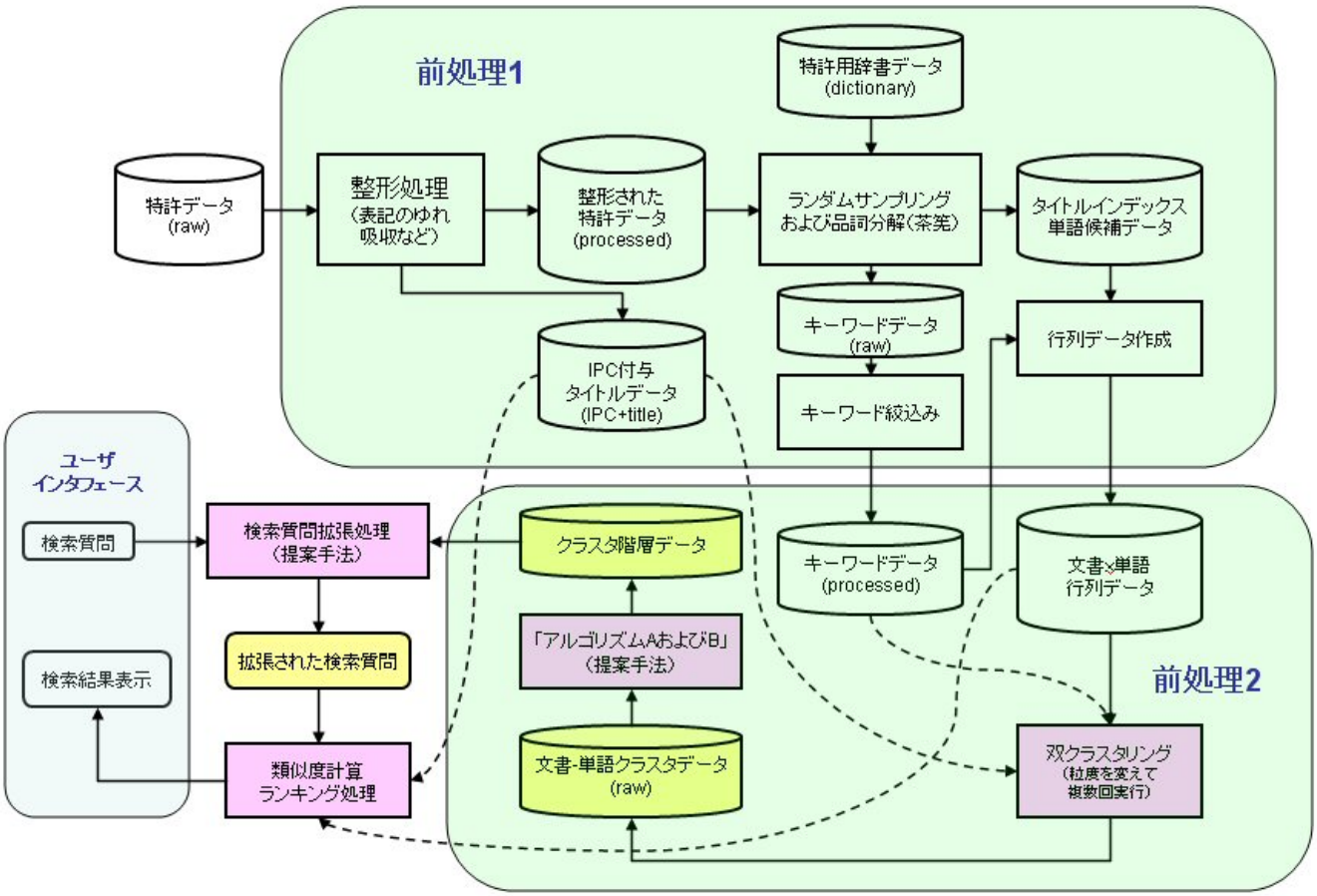

図 3．提案手法を含む特許概念検索プロトタイプシステムの概略

提案手法を含む，特許データの概念検索プロトタイプシステムを図３のような構成で構築した．特許文書は NTCIR から入手したもので，前処理１として整形処理，品詞分解（茶筅）などを行い，最終的に単語 30,370 個を抽出した．双クラスタリングの入力は，文書×単語の疎行列データ，キーワードデータ，および文書タイトルデータである．双クラスタリングを複数回異なる粒度で実行し，クラスタ平滑化と階層化を「アルゴリズム A」と「アルゴリズム B」により実行し，図２に例示したようなクラスタ粒度階層構造を得る．これらの処理は図３の前処理２に含まれる。

一方，ユーザから入力される検索質問は，図３の「検索質問拡張処理」において，クラスタ階層構造をもとに検索質問拡張を行い，類似度の高い順にランク上位の文書に関して再現率と適合率を計算した．

クラスタ粒度は，16, 32, 64, 128, 及び 256 の 5 レベルを用い，それぞれ 5 通りの異なる乱数の初期値で双クラスタリングを行い，クラスタ粒度階層構造を作成した．なお，双クラスタリングの 1 回あたりの実行時間は，平均 20 分程度であった．使用した PC は Pentium4 (2.8GHz)である．

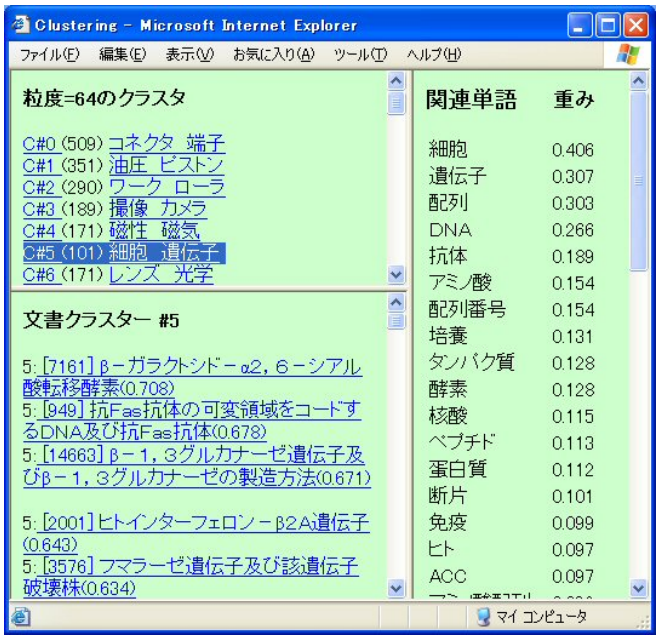

図４．「遺伝子・組換え技術」に関する文書クラスタと単語クラスタ表示例

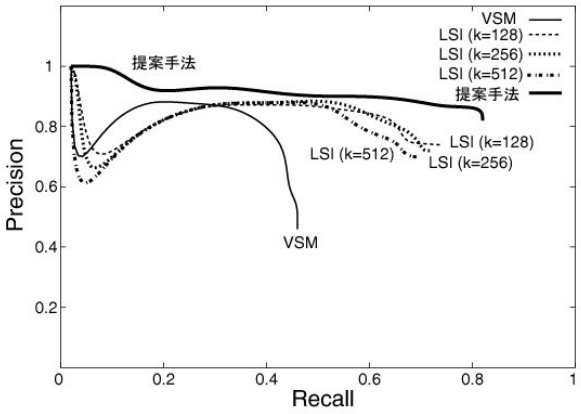

図５．「遊技機，パチンコ」に関する特許検索の再現率・適合率曲線

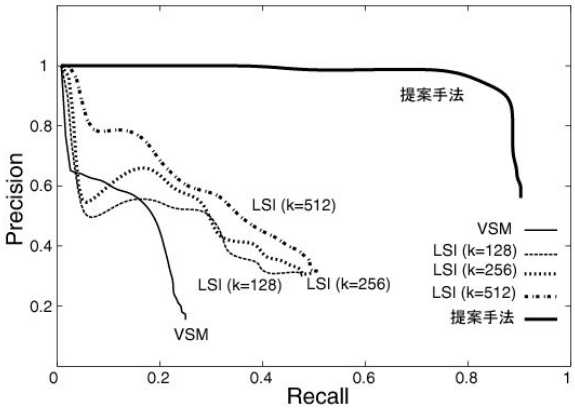

図６．「田植，移植機」に関する特許検索の再現率・適合率曲線

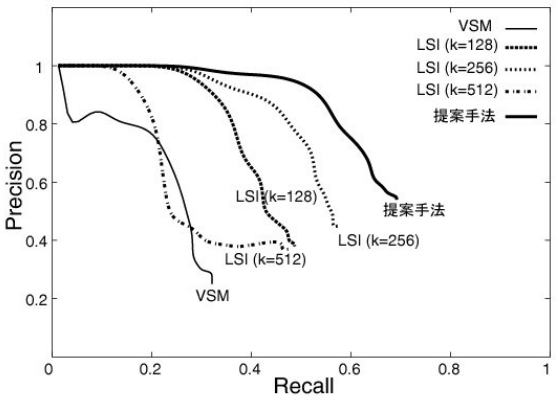

図７．「免振，耐震」に関する特許検索の再現率・適合率曲線

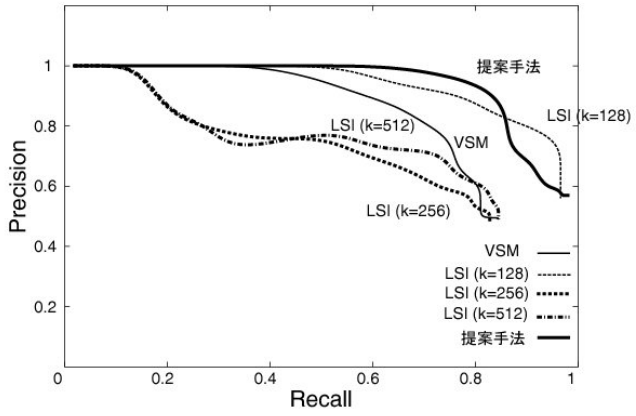

図８．「遺伝子，DNA 組換え」に関する特許検索の再現率・適合率曲線

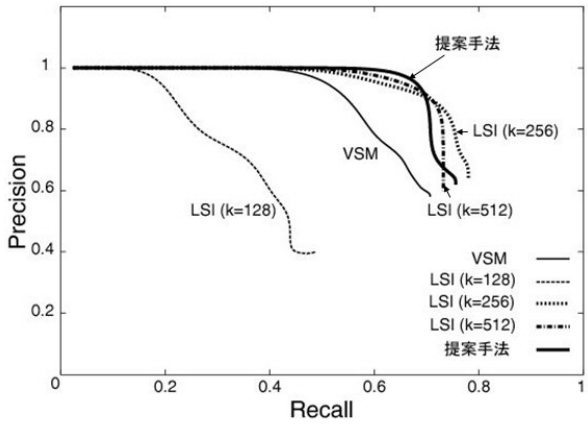

図９．「生ゴミ，ゴミ処理機」に関する特許検索の再現率・適合率曲線

5. おわりに

双クラスタリングを特許データの概念検索に利用する手法を述べた．特に，粒度の異なる「クラスタ粒度階層構造」を利用して，検索質問拡張をするアプローチで，既知の手法である LSI をほとんどの場合凌ぐ結果を得た．LSI がよい性能を出す例もあったが，k（削減する次元数）の値を事前に予測するのは困難であり，本手法のように前処理として，クラスタ階層構造を作っておくことで，k の値を気にしなくてもすむ点は評価できると思われる．

クラスタ粒度に関しては，今回の実験では，２のべき乗で変化させたが，そのような粒度で階層構造を作成するのが最良であるかは，今後の課題である．

今後は，検索質問群で同様の結果が得られるかどうかの実験，また，手法のスケーラビリティ向上のための工夫や，稀少文書の検索の実験などを行う予定である．

文　献

[1] 青野雅樹，小林メイ，“ベクトル空間モデルに基づく次元削減による大規模文書データの検索と可視化，”情報処理学会，*マルチメディアと分散処理研究会*，2002-DPS-108, pp. 79-84, June 2002.

[2] 小西一也，北内啓，高木徹，“発明の特徴に着目した検索語抽出による先願特許検索”，DEWS2004, 2004.

[3] 佐々木稔，北　研二，“ランダム・プロジェクションによるベクトル空間情報検索モデルの次元削減，”*自然言語処理*，８巻，１号，pp. 5-19, 2001.

[4] 成田宏和，太田学，片山薫，石川博，“Web 文書検索のための非排他的クラスタリング手法の提案”，2P-01, DEWS2003, 2003.

[5] 新田清，蓬莱尚幸，園部正幸，“文書クラスタリングを利用した検索質問展開手法の開発と評価，”pp. 9-16, *データベースシステム, 118-2 情報学基礎 54-22*, pp.9-16, May 1999.

[6] Youjin Chang et al, “Conceptual Retrieval based on Feature Clustering of Documents,” *Proc MF/IR 2002*, available at http://dcs.vein.hu/CIR/, 2002.

[7] Douglass R. Cutting, et al. “Scatter/Gather: A Clustering-based Approach to Browsing Large Document Collections,” *Proc .ACM SIGIR’92*, pp.318-329, 1993.

[8] Inderjit S. Dhillon and Dharmendra S. Modha, “Concept Decomposition for Large Sparse Text Data Using Clustering,” *Machine Learning*, Vol.42, pp. 143-175, 2001.

[9] Inderjit S. Dhillon, S. Mallela, D. S. Modha, “Information-Theoretic Co-clustering”, Proc. SIGKDD ’03, pp. 89-98, 2003.

[10] Inderjit S. Dhillon and Yuqiang Guan, “Information Theoretic Clustering of Space Co-Occurrence Data,” *Proc IEEE ICDM’03*, Melbourne, Florida, USA, pp.517-520, November 2003.

[11] Scott Deerwester et al., “Indexing by latent semantic analysis,” *Journal of the American Society for Information Sciences*, vol.41, pp.391-407, 1990.

[12] Koji Eguchi, et al, “Adaptive Query Expansion Based on Clustering Search Results,” *Transactions of Information Processing Society of Japan*, Vol.40, No.5, pp.2439-2449, May 1999.

[13] Thomas Hoffmann, “Unsupervised Learning by Probabilistic Latent Semantic Analysis,” *Machine Learning*, Vol.42, pp.177-196, 2001.

[14] Mei Kobayashi, Masaki Aono, H. Takeuchi, H. Samukawa, “Matrix computations for information retrieval and major and minor outlier cluster detection,” *Journal of Computation and Applied Mathematics*, Vol.143, No.1-1, pp. 119-129, 2002.

[15] Kazuya Konishi, Akira Kitauchi, and Toru Takaki, “Invalidity Patent Search System of NTT DATA”, Working Notes of NTCIR-4, Tokyo, 2-4 June, 2004.

[16] NTCIR (NII-NACSIS Test Collection for IR Systems), http://research.nii.ac.jp/ntcir/