共起に基づく類似性尺度

相澤　彰子

情報を伝達するための文字の並びを「テキスト」と呼ぶ．テキスト中に出現するさまざまな構成要素は，その出現位置によって他の要素と関係づけられている．この要素どうしの関係が織りなす空間はどのようなものになるだろうか？言語処理の分野では，このような空間は「意味」と密接な関係があると考える．そして空間上での距離が近いものは，何らかの意味的な近さをもつという前提のもとに，さまざまな類似度尺度が適用される．本稿では，テキストをめぐるさまざまな「共起」事象について述べ，これらを扱うための統計的手法や適用事例を紹介する．

キーワード：共起行列，シソーラス自動構築，類似度尺度，言語コーパス

1．はじめに

テキストにおける「共起」を一般的に説明するならば，「ある区間の中で2つの要素どうしが同時に観察されること」となる．ここで「区間」とは，ある基準で区切られたテキストの領域，「要素」とは，その領域に含まれるテキストの構成要素である．区間や要素の単位を何に定めるかは，想定する問題によってさまざまである．例えば，区間を文書，要素を文書中の語に対応させると，検索語入力に対して関係が深い文書を出力する情報検索の問題になる．また，区間を対訳テキスト，要素を各言語の単語に対応づけると，適切な訳語ペアを見つける対訳抽出の問題になる．区間を辞書の項目，要素を見出し語および定義文中に出現する語にすると，語の上位下位関係を表すシソーラスの自動構築に結びつく．このように共起情報の扱いは，多くの言語処理において必要となる基本的な技術である．

類似性尺度には多様なバリエーションが存在する．これらはいずれも形式的には共起を表す行列表現の上で定義可能であるが，実際の適用では，(1)テキストの上で観察される共起をどのように共起行列の形に表現するか，(2)計算した類似度に基づきどのような処理を実現するか，の2つが重要なポイントとなる．以下，本稿では，まず2節で類似度計算の基本になる共起行列の構成について述べる．次に3節で，要素間の共起のしやすさを測り，共起の度合が強いペアを抽出するための統計的な尺度をまとめる．さらに4節で，区間内で共起する要素（文脈）に基づき，意味的な類似性や距離（非類似性）を測るための尺度を紹介する．最後に5節で，共起行列の変換や行列成分のクラスタリングに基づく分析法について述べる．

2．共起行列

共起事象を行列の形で表したものを「共起行列」と呼ぶ．共起行列の構成法としては，(a)要素と区間の共起を表現する場合，(b)異なる2種類の要素どうしの共起を表現する場合，(c)同じ種類の要素どうしの共起を表現する場合，の3通りが考えられる（図1）．前出の例でいえば，情報検索が(a)，対訳抽出が(b)，シソーラス自動構築が(c)に対応する．以下，それぞれについて簡単にまとめる．

(a)　区間と要素の共起

区間の集合を $S=\{s_1, \cdots, s_r\}$，テキストの構成要素の集合を $W=\{w_1, \cdots, w_p\}$ として，区間 $s_i(\in S)$ で要素 $w_j(\in W)$ が観察される回数を $f(s_i, w_j)$ と表記する．このとき，値 $f(s_i, w_j)$ を第 (i, j) 成分にもつ $r \times p$ 行列 C_{SW} を W と S の共起頻度行列と呼ぶ．具体的には，S は文書，W は文書中に出現する語，$f(s_i, w_j)$ は語 w_j の文書 s_i での文書内頻度などである．

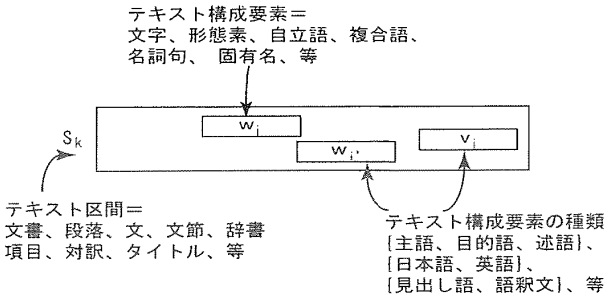

図1　テキストにおける区間と要素

あいざわ　あきこ
国立情報学研究所
〒 101-8430 千代田区一ッ橋 2-1-2

(b) 異なる種類の要素どうしの共起

異なる種類のテキスト構成要素の集合をそれぞれ $W=\{w_1, \cdots, w_p\}$, $V=\{v_1, \cdots, v_q\}$, 区間の集合を $S=\{s_1, \cdots, s_r\}$ とする．要素 $w_i(\in W)$, $v_j(\in V)$ に対して，w_i, v_j がともに出現する区間の数を $f(w_i, v_j)$ とするとき，値 $f(w_i, v_j)$ を第 (i, j) 成分にもつ $p\times q$ 行列 C_{WV} を W と V の共起頻度行列と呼ぶ．(a)で与えられる C_{SW}, C_{SV} の非零成分を1で置き換えた行列を C'_{SW}, C'_{SV} とするとき，定義から $C_{WV}=C'^{T}_{SW}C'_{SV}$ である．具体的には，S は対訳関係にある日本語と英語の文，W は日本語の単語，V は英語の単語などである．

(c) 同じ種類の要素どうしの共起

(b)において，W と V が同じである場合を考え，2つの要素 w_i, $w_j(\in W)$ の共起回数を $f(w_i, w_j)$ とする．このとき，値 $f(w_i, w_j)$ を第 (i, j) 成分にもつ $p\times p$ 正方行列 C_{WW} を W の共起頻度行列と呼ぶ．(b)の場合と同様に $C_{WW}=C'^{T}_{SW}C'_{SW}$ である．具体的には，S は辞書の見出し語と定義文，W は辞書の見出し語，V は定義文中に出現する語などである．

以下では簡単のため，共起行列を C, C の行および列に対応する要素や区間を $X=\{x_1, \cdots, x_m\}$ および $Y=\{y_1, \cdots, y_n\}$ で表記する．なお，上記では頻度を行列成分の値としたが，確率モデルを適用する場合には，同時出現確率 $P(x_i, y_j)$ や条件付き確率 $P(x_i|y_j)$ を値として用いる．ここで，テキスト構成要素の頻度分布は，いわゆるジップの法則にしたがうことが広く知られている．相対頻度と出現確率は必ずしも一致しないので注意が必要である．確率モデルの適用では，統計的言語モデルの分野でディスカウンティングやスムージングと呼ばれる手法[1]によって，確率の補正を行う場合も多い．

3. 共起の度合を測るための尺度

本節では，2つの要素 $x_i\in X$, $y_j\in Y$ に注目して共起の度合いを測るための代表的な尺度を紹介する．このような計算の目的は，例えば共起ペアどうしを比較して，与えられたテキスト集合の中から結びつきが強いペアを抽出することである．このために，x_i と y_j が互いに無関係に出現する場合と比較した偏りが尺度として用いられる．

3.1 2×2分割表の上で定義される類似性尺度

計算の基本となるのは，は2×2分割表（contingency table）あるいはクロス表（cross tabulation）と呼ばれる表である（図2）．2×2分割表の作成では，

	y_j	$\bar{y_j}$	Σ
x_i	f_{11}	f_{12}	$f_{1\cdot}$
$\bar{x_i}$	f_{21}	f_{22}	$f_{2\cdot}$
Σ	$f_{\cdot 1}$	$f_{\cdot 2}$	F

$f_{1\cdot}=f_{11}+f_{12}$
$f_{2\cdot}=f_{21}+f_{22}$
$f_{\cdot 1}=f_{11}+f_{21}$
$f_{\cdot 2}=f_{12}+f_{22}$
$F=f_{11}+f_{12}+f_{21}+f_{22}$

図2 2×2分割表

2つの要素 $x_i\in X$, $y_j\in Y$ に注目し，「x_i が観察された」「x_i 以外が観察された（$\bar{x}_i$）」，「y_j が観察された」「y_j 以外が観察された（$\bar{y}_j$）」の組合せ4通りについて出現頻度を集計する．

分割表は要因間の関係を分析するため広く用いられており，数多くの統計的尺度が存在する[2]．その中で言語処理の分野で多く使われる代表的な尺度を表1にまとめる．f_{11}, f_{12}, f_{21}, f_{22} の4つの数値に対して，統計分析や情報理論を背景にした多様な尺度が用いられていることがわかる．

まず，基本となる2つの尺度として，①頻度と②自己相互情報量（pointwise mutual information, PMI）の2つがある．①では，数多く観察される共起が重要であるとみなす．一方②では，偏って共起するものほど重要であるとみなす．①と②を比較すると，前者では全体としての頻度（$f_{1\cdot}$ や $f_{\cdot 1}$）が高い要素が有利であり，後者では逆に頻度が低い要素が有利であるといわれる．例えば与えられたテキスト中で，「言語」が1,000回，「処理」が1,000回，「言語　処理」が100回観察されたとする．また，「コルモゴロフ」「スミルノフ」「コルモゴロフ　スミルノフ」が各々1回ずつ観察されたとする．①では，「情報処理」の方が高く評価されるが，②では「コルモゴロフ　スミルノフ」が高く評価される．

ところが現実の適用においては，抽出したい共起ペアはその中間にある場合も多い[3]．表1の③～⑯の尺度は，これら中間的な共起ペアの抽出に用いられる．適用分野が多岐に渡るため，これらをすべて体系的に位置付けることはむずかしいが，大きく分類すれば以下のようになる．まず③～⑧は，共起頻度が全体の頻度の中に占める割合に基づくものである．このうち⑦のTスコアや⑧のZスコアは平均値による補正を含む．⑨および⑩では，①と②の掛け合わせ効果が期待される．⑪～⑯は，分割表のすべてのセルについての偏りを考慮した尺度である．例えば「共起回数が目立って少ない組み合わせ」があれば，それも偏りとして

検出されることになる．各適用分野において，異なる尺度の比較や組合せ効果に関する報告もあり，実際の適用において選択肢は，多くても数種類程度である．

3.2 言語処理における適用例

表1の共起尺度を用いる言語処理の例を以下にあげる．

用語抽出 与えられたテキスト集合（コーパス）から，意味的なまとまりをもつ語の並びを用語として抽出する．得られた用語は，様々な言語解析に利用できる．例えば，「イベント」「駆動」を1つの情報処理用語「イベント駆動」として辞書登録すると，検索性能の向上が期待できる．

コロケーション抽出 用語抽出では隣接した2語の共起を考慮するのに対して，コロケーション抽出では，一定幅のウィンドウ内での共起を調べて，共起の度合の高いものを候補とする．コロケーションの定義は研究者によってさまざまであるが，一般には「風邪を引く」など慣習的な表現を指し，翻訳などの応用には欠かせない．

対訳抽出 意味的な対応がある2言語のテキストから対訳関係にある要素ペアを抽出する．例えば，「計算機アーキテクチャ」と「computer architecture」のような対訳から｛計算機，computer｝の対応を求めるなどである．対訳テキストの単位は利用可能なコーパス資源によって決まる．伝統的には，文単位での対応を想定する場合が多いが，近年では，同一分野の文書を多言語で集めたコンパラブルコーパスを用いる場合もある．

関連語抽出 同時に出現するものは互いに関連が深いという前提のもとに，自己相互情報量やJaccard係数などを用いて関連が強い語を抽出する．ウェブ検索エンジンのヒット数を利用して関連性を求める場合もある．例えば，「樋口一葉」，「夏目漱石」，「樋口一葉　夏目漱石」を検索質問としてエンジンに投げ，各々に対して得られるヒット数を使って共起の度合を数値化する．

言語的な解析を目的として処理を行う場合には，共起尺度による計算結果は，コーパスの基礎的な統計データとして利用されることになる．一方で，用語や辞書の自動抽出を目的とする場合には，共起尺度の計算だけで実用的な結果を得ることは一般にはむずかしい．そこで，有効な統語的・意味的手がかりを求めて工夫が凝らされることになる．例えば用語抽出では，語の独立性，他の語と結びつく力，他の分野との相対的な頻度の差などが手がかりとして用いられる．対訳抽出では，対訳辞書など既存の言語資源を用いてあいまい性の解消を行う場合も多い．ウェブ検索エンジンを用いる場合については利用上での制約から単純な尺度が多く用いられるが，スニペット（検索エンジンが上位文書に対して返す簡潔表示）の解析を使って，より精度を高める研究も行われている．

表1　共起の度合を測る代表的な尺度

類似性尺度	*定義式*
①頻度	f_{11}
②自己相互情報量（PMI）	$\log\left(\frac{p_{11}}{p_{1\cdot}p_{\cdot 1}}\right)=\log\left(\frac{f_{11}F}{f_{1\cdot}f_{\cdot 1}}\right)$
③ダイス係数	$\frac{2f_{11}}{f_{1\cdot}+f_{\cdot 1}}$
④Jaccard係数	$\frac{f_{11}}{f_{11}+f_{12}+f_{21}}$
⑤Simpson係数	$\frac{f_{11}}{\min(f_{1\cdot},\ f_{\cdot 1})}$
⑥コサイン	$\frac{f_{11}}{\sqrt{f_{1\cdot}}\sqrt{f_{\cdot 1}}}$
⑦Tスコア	$\frac{f_{11}-f_{1\cdot}f_{\cdot 1}/F}{\sqrt{f_{11}}}$
⑧Zスコア　Sは領域長	$\frac{f_{11}-f_{1\cdot}f_{\cdot 1}/F}{\sqrt{f_{1\cdot}f_{\cdot 1}/F}}$　または　$\frac{f_{11}-Sf_{1\cdot}f_{\cdot 1}/F}{\sqrt{Sf_{1\cdot}\times f_{\cdot 1}/F\times(1-f_{\cdot 1}/F)}}$
⑨修正ダイス係数	$(\log f_{11})\times\frac{2f_{11}}{f_{1\cdot}+f_{\cdot 1}}$
⑩重みつきPMI	$p_{11}\times\log\left(\frac{p_{11}}{p_{1\cdot}p_{\cdot 1}}\right)=\frac{f_{11}}{F}\times\log\left(\frac{f_{11}F}{f_{1\cdot}f_{\cdot 1}}\right)$
⑪オッズ比	$\frac{f_{11}f_{22}}{f_{12}f_{21}}$
⑫補完類似度（CSM）	$\frac{f_{11}f_{22}-f_{12}f_{21}}{\sqrt{f_{1\cdot}f_{\cdot 1}}}$
⑬カイ二乗	$\sum_{i=1,2}\sum_{j=1,2}\frac{\left(f_{ij}-f_{i\cdot}f_{\cdot j}/F\right)^2}{f_{i\cdot}f_{\cdot j}/F}=\frac{F(f_{11}f_{22}-f_{12}f_{21})^2}{f_{1\cdot}f_{2\cdot}f_{\cdot 1}f_{\cdot 2}}$
⑭イェーツの補正をしたカイ二乗	$\frac{F\left(\lvert f_{11}f_{22}-f_{12}f_{21}\rvert-F/2\right)^2}{f_{1\cdot}f_{2\cdot}f_{\cdot 1}f_{\cdot 2}}$
⑮対数尤度比	$2\sum_{i=1,2}\sum_{j=1,2}f_{ij}\log\left(\frac{f_{ij}F}{f_{i\cdot}f_{\cdot j}}\right)$
⑯相互情報量（エントロピー）	$\sum_{i=1,2}\sum_{j=1,2}p_{ij}\log\left(\frac{p_{ij}}{p_{i\cdot}p_{\cdot j}}\right)=\sum_{i=1,2}\sum_{j=1,2}\frac{f_{ij}}{F}\log\left(\frac{f_{ij}F}{f_{i\cdot}f_{\cdot j}}\right)$

4. 文脈で類似度を捉える

本節では，共起行列の各行をベクトル $\overrightarrow{x_1}=(x_{r1},$

$\cdots x_{iq}$) とみて，行ベクトルどうしの類似度を計算したり，新たな入力ベクトル $z=(z_1, \cdots z_q)$ に対して最も近い行を求めたりするための類似性尺度について紹介する．

4.1 文脈ベクトルによる類似性尺度

ここでの基本的な考え方は，「類似した文脈で使われる語は意味的に類似している」というものである．これは，共起行列の列要素を「特徴（素）」に対応させて特徴空間を構成することに相当する．ベクトル表現 $\overrightarrow{x_i}=(x_{i1}, \cdots x_{iq})$ は，共起要素の分布によって x_i を特徴付けるものであることから「文脈ベクトル（context vector）」と呼ばれる．確率モデルを用いる場合には，ベクトルの代わりに，x_i が観察された場合の列要素の分布を表す条件付き確率 $P(Y|x_i)$ を用いる．類似度の計算が2つの分布の近さに基づくことから，文脈ベクトルに基づく類似性尺度は，しばしば「分布間類似度（distributional similarity）」と参照される．

文脈ベクトルに基づく代表的な尺度を表2にまとめる．これらの計算方法は大別して以下の3つにまとめることができる．

まず①は，共起要素の重なりに注目する尺度である．特徴空間上の2点 x_i, x_j に対して，各々の共起要素の集合 $S(x_i), S(x_j)$ の重なりを類似度とする．ただし，$S(x)$ は x と共起する Y の部分集合で，$S(x)=\{y_k|f(x, y_k)>0\}$ とする．代表例として，表2の①ではJaccard係数をあげたが，その他ダイス係数をはじめとする表1の尺度が広く適用可能である．

次に②～④は，特徴ベクトル間の距離に基づく尺度である．特徴空間上の2つのベクトル $\overrightarrow{x_i}=(x_{i1}, \cdots, x_{ir})$, $\overrightarrow{x_j}=(x_{j1}, \cdots, x_{jr})$ の間の距離を非類似度として用いるもので，x_{ik} の値としては，単純な共起頻度だけではなく，分布の特性や応用を考慮した重みがかけられる場合も多い．特に情報検索の分野では，伝統的にtf-idf (term frequency inverse document frequency) と呼ばれる重みづけ法にコサイン尺度を適用することが一般的である．Tf-idfの基本的な定義を以下にあげる．

$$x_{ik}=f(x_i, y_k)\log\frac{F}{df(y_k)}$$

ここで $df(y_k)$ は文書頻度と呼ばれるもので，y_k が共起する X の要素数である．$df(y_k)$ の値が小さいほど，y_k は x_i に特徴的であると考えられる．情報検索分野においては，類似の重みづけが多数提案されており，経験的によいとされるものとしてOkapi BM 25[4]などがある．

表2　文脈ベクトルに基づく代表的な尺度

類似性尺度	*定義式*
①Jaccard係数	$\frac{\lvert S(x_i)\cap S(x_j)\rvert}{\lvert S(x_i)\cup S(x_j)\rvert}$
②相関係数	$\frac{\sum_{k=1}^{r}(x_{ik}-\bar{x_i})(x_{jk}-\bar{x_j})}{\sqrt{\sum_{k=1}^{r}(x_{ik}-\bar{x_i})^2}\sqrt{\sum_{k=1}^{r}(x_{jk}-\bar{x_j})^2}}$, $\bar{x_i}=\frac{1}{r}\sum_{k=1}^{r}x_{ik}, \bar{x_j}=\frac{1}{r}\sum_{k=1}^{r}x_{jk}$
③ユークリッド距離	$\sqrt{\sum_{k=1}^{r}(x_{ik}-x_{jk})^2}$
④コサイン尺度　確率ではBhattacharyya distance	$\frac{\sum_{k=1}^{r}x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^{r}x_{ik}^2}\sqrt{\sum_{k=1}^{r}x_{jk}^2}}$
⑤カルバックライブラー情報量	$KLI(p_i \,\|\| \, p_j)=\sum_{k=1}^{r}P(y_k\|x_i)\log\frac{P(y_k\|x_i)}{P(y_k\|x_j)}$
⑥Jensen-Shannonエントロピー	$\sum_{k=1}^{r}\left(P(y_k\|x_i)-P(y_k\|x_j)\right)\log\frac{P(y_k\|x_i)}{P(y_k\|x_j)}$ $=KLI(p_i \,\|\| \, p_j)+KLI(p_j \,\|\| \, p_i)$
⑦information radius	$KLI\left(p_i \,\|\| \, \frac{p_i+p_j}{2}\right)+KLI\left(p_j \,\|\| \, \frac{p_i+p_j}{2}\right)$
⑧Skew divergence	$KLI(p_i \,\|\| \, \alpha p_j+(1-\alpha)p_i)$

⑤～⑧は，確率モデルに基づく尺度である．確率変数 p_i，p_j が条件付き確率 $P(y_k|x_i), P(y_k|x_j)$ にしたがうものとして，まず，p_i，p_j 間の距離尺度の基本となるのは⑤のカルバックライブラー情報量（クロスエントロピー）である．⑤は p_i，p_j に対して非対称となるため，これを双方向に加え合わせたものが，⑥のJensen-Shannonエントロピーである．⑤や⑥の計算では，分母の確率がゼロになることを避けるため，スムージングによる確率の補正が必要である．これに対して⑦や⑧は，スムージングを考慮せずに用いることができる．

4.2 言語処理における適用例

表2の文脈ベクトルに基づく尺度を用いる言語処理の例を以下にあげる．

情報検索　文書を行に対応させ，文書中に出現する語を要素とする文書ベクトルを作成する．各文書ベクトルと検索語ベクトルとの類似度を求め，検索要求に対して文書が適合するかどうかを判定する．既に述べたように伝統的にtf-idfに代表される重みづけが用いられる．近年では確率的言語モデルに基づき文書と検索語の確率分布を推測し，カルバックライブラー情報

量で分布間の距離を調べる方法も一般的である．

テキスト分類　文書カテゴリ（クラス）を一定の割合で語を生成する情報源であると考えて行に対応させる．与えられた文書がいずれのカテゴリに由来するものかを類似性尺度を用いて判定する．単純な語の出現確率に基づくNaïve Bayesにはじまり，k近傍法（kNN）やサポートベクタマシン（SVM）等，数多くの機械学習が盛んに適用されてきた．少数事例からの学習や多重トピック抽出など，近年も多くの展開がみられる．

あいまい性解消　テキスト中に出現した語の意味を文脈ベクトルで捉え，あらかじめ辞書に登録された複数の語義の中から，その文脈での用法に該当するものを類似度に基づき選択する．また，シソーラスや品詞体系が与えられたとき，未知語の品詞や分類を文脈から推定する問題でも，文脈ベクトルが活用される．

確率モデルは理論的な基盤が明確で汎用的であるが，一方で現実の応用では，事象（特徴素）の排他性の仮定が必ずしも成立しないという問題もある．テキストの構成要素は様々な依存関係をもち，「情報」と「情報検索」の場合のように領域に重なりがあったり，単語と品詞のように同一の語に対して複数の特徴が与えられたりといった場合も考えられる．サポートベクタマシンや条件付き確率場（CRF）などの機械学習手法は，このような前提のもとでも適用可能である．また，格関係にある名詞と動詞など，限定された領域での共起関係を扱う場合には，集合の重なりに基づく単純な尺度も多く用いられる．

5．共起行列の分解に基づくアプローチ

本節では，共起行列全体の構造を分析するアプローチについて述べる．まず一般的に行われるのは，前節で述べた類似度や距離を用いて，階層的クラスタリングやk-means等の汎用的なクラスタリング手法を適用し，意味的なまとまりを分析することである．一方，共起行列の特性を利用したものとして，行と列を同時に考慮して特徴空間の構造を分析したり，要素をグループ化したりする手法がある．以下では，後者に焦点をあて紹介をする．

5.1　特異値分解に基づく方法

特異値分解（Singular Value Decomposition, SVD）による行列変換を基本とする手法として，統計分野におけるコレスポンデンス分析（correspondence analysis）があり[5]，対応して言語処理分野では潜在的意味インデクシング（Latent Semantic Indexing, LSI）や潜在的意味解析（Latent Semantic Analysis, LSA）と呼ばれる手法が用いられる[1]．

簡単に概要を紹介すると，特異値分解では$p\times q$共起行列Cを以下の形に分解する．

$$C=UDV^T,\quad U^TU=VV^T=1$$

ただし，行列Cのランクをrとして，Uは$p\times r$，Vは$r\times q$，Dは$r\times r$型の対角行列である．また，Dの対角成分を$(\sigma_1,\sigma_2,\cdots,\sigma_r)$として$\sigma_1\geq\sigma_2\geq\cdots\geq\sigma_r>0$とする．これらは行列$C$の特異値と呼ばれ，$CC^T$の（正の）固有値の平方根に等しい．

共起行列に対して特異値分解を適用する理由はいくつか考えられる．まず，コレスポンデンス分析で行われるように，同一平面上に$x\in X, y\in Y$を同時に配置することで，両者の関係を視覚的に把握することが考えられる（図3）．$CC^T=(UDV^T)(UDV^T)^T=(UD)(DU)^T$より，$U$の列ベクトルは$CC^T$の固有ベクトルとなり，$U$の第$(i,k)$成分は第$k$番目の特異値に対応する軸への$x_i$の重みを示す．同様に$V$の列ベクトルは$C^TC$の固有ベクトルとなり，$V$の第$(k,j)$成分は第$k$軸への$y_j$の重みを示す．これにより$x_i$と$y_j$を共通の空間に配置して考えることが可能になる．ただし，コレスポンデンス分析とLSAとでは，基準化の方法に違いがある．コレスポンデンス分析ではZスコアを行列成分の値として，得られた固有ベクトルを周辺確率（$f_{i\cdot}/F$や$f_{\cdot j}/F$）を用いて基準化するが，LSAでは頻度をそのまま行列成分の値として，得られた固有ベクトルを特異値（特異値の2乗は第k軸に対する分散）で重みづけしたUDやDV^Tを特徴ベクトルとする．目的が要因分析にあるのか，距離関係を維持した変換にあるのか，という違いが現われているといえる．

特異値分解を適用する理由の2つ目は，「概念」ベクトルの抽出である．元来，テキスト上の「表記」とそれが伝える「概念」とは一対一に対応するものではない．そこで特異値分解を適用して特徴空間の潜在的な構造を分析することで，「表記」と「概念」を対応づけようというものである．これによって，例えば，直接は共起していない要素どうしの距離を測ることが可能になる．情報検索や対訳テキストへの適用例などが知られる．

特異値分解を適用する理由の3つ目は，次元数の削減である．特異値分解の結果得られる行列の添え字の大きな要素を取り除き，次元数をrからkとして得

られる行列を

$$C_k = U_k D_k V_k^T$$

とするとき，C_kはランクkの行列の中で二乗誤差（フロベニウス・ノルム）についてCのもっともよい近似になることが知られている．UDやDV^TのかわりにU_kD_kや$D_kV_k^T$を用いて類似度を計算することで，ノイズや冗長な成分を取り除く効果が期待される．なお，確率的な解釈に基づくものとして尤度最適化基準に基づくpLSA（probabilistic LSA）も提案されている．pLSAにおける行列分解は特異値分解ではなく最大エントロピー法により行われる．

5.2 共クラスタリング法

最後に，行列表現上で関連のある行と列をまとめ，同時にクラスタリングを行う手法を紹介する．共起行列は零成分が大半を占める疎な行列であるが，特異値分解により得られるベクトルは非零成分を多く含む密なものになる．自動生成されるこのようなベクトルは，必ずしも人間にとってわかりやすい「意味」を反映するとは限らない．これに対して，互いに関連が深い行および列をまとめた部分集合$\{X_k \in X, Y_k \in Y\}$を想定して，これを1つの概念やグループに対応させる「共クラスタリング（coclustering）」は，よりわかりやすい形で意味の「成分」を示すものであるといえる．

共クラスタリングは，行や列を入れ替えて図4に示すような密な矩形領域を作る作業と考えると直観的にわかりやすい．図3と比較すると，特異値分解の適用後に新しく構成された空間の軸上での順番にしたがって行や列を並べ替えることで，密な領域が作られることが予想できる．

共クラスタリングの場合も通常のクラスタリングと同様に，排他的なクラスタを生成するか/重なりのあるクラスタを生成するか，クラスタ数をあらかじめ定めるか/何らかの停止基準を設定するか，分枝型（divisive）であるか/凝集型（agglomerative）であるか等を含めて，様々なバリエーションを考えることができる．Li and Abeは情報量基準を用いて，まず行，次に列を1つずつ統合することでクラスタリングを行う方法を提案した[6]．情報量を用いて同様にグループ化を行う方法として，Pereiraらによる分布クラスタリング（distributional clustering）[7]やTishbyらによる情報ボトルネック法（information bottleneck method）[8]がある．個々の要素の分布をクラス分布に置きかえることによる歪み（情報量の観点からの損失）を最小にするよう，階層的にクラスタリングを行う方法である．ただし，これらの方法では行要素のクラスタリングに焦点が与えられており，必ずしも列要素をあわせてグループ化するものではない．

これに対してDhillonらによるスペクトラルグラフ分割（spectral graph bipartitioning）[9]やZhaらによる2部グラフ分割[10]では，クラスタリングを，共起行列から作成した2部グラフを2つに分断する問題として定式化する．グラフを分断するリンク集合（カット）を用いて目的関数を定めると，その近似最適解は特異値分解による第2軸によって与えられる．スペクトラルグラフ分割ではさらに，得られた第2軸の2つの端点をシードとしてk-meansクラスタリングを適用して，要素全体を2分割する．この際に，行と列

webのacドメインから収集した形容詞＋名詞の共起ペア

	飲み物	画像	海	言葉	光	写真	情報	色	星	声	風	雰囲気
暗い	0	16	32	3	12	9	0	110	124	4	1	46
小さい	0	430	2	0	8	68	323	6	32	114	0	0
大きい	0	737	5	4	8	376	13	7	20	205	3	0
暖かい	25	0	108	78	20	0	0	10	0	9	62	90
明るい	0	12	3	3	202	8	4	251	387	91	1	304
冷たい	80	0	62	23	18	0	0	3	7	16	344	6

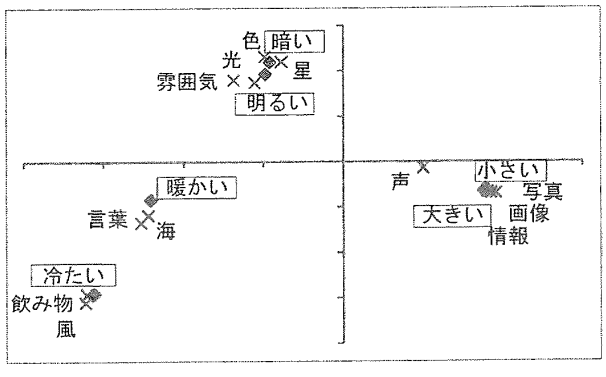

図3 LSAによる平面への配置例

webのacドメインから収集した形容詞＋名詞の共起ペア

	情報	画像	写真	声	星	色	光	雰囲気	海	言葉	飲み物	風
小さい	323	430	68	114	32	6	8	0	2	0	0	0
大きい	13	737	376	205	20	7	8	0	5	4	0	3
暗い	0	16	9	4	124	110	12	46	32	3	0	1
明るい	4	12	8	91	387	251	202	304	3	3	0	1
暖かい	0	0	0	9	0	10	20	90	108	78	25	62
冷たい	0	0	0	16	7	3	18	6	62	23	80	344

クラスタ： $X_k = \{$暗い, 明るい$\}$, $Y_k = \{$星, 色, 光, 雰囲気$\}$

図4 行列の入れ替えと共起クラスタ

は同時にクラスタリングされる．このプロセスを繰り返すことで分枝型のクラスタリングを実現する．

5.3 言語処理における適用例

共起行列の分解や共クラスタリングについて，言語処理における適用例をあげる．

下位範疇化 名詞と動詞の共起関係に注目して，各々を意味的なカテゴリにまとめるなどの下位範疇化や，語を話題別にまとめる語彙分割の問題に適用されている．

文書クラスタリング もともと1990年当初にLSIが適用されたのは情報検索の分野であったが，計算コスト等の問題から，大規模な検索システムにおける適用は一般的ではない．近年ではむしろ，確率的解釈に基づく発展で，話題抽出やイベント追跡，文脈適応等の分野における適用が盛んである．

スペクトラルグラフ分割の例にみられるように，共起行列は，その要素を枝の重みに対応させると，グラフ表現に変換できる（図5）．本稿の2節では共起行列を3つのタイプにわけて定義したが，これらはグラフ的な観点からみると違いがある．まず，2節(a)で行をノード，列をリンクに対応させると，共起行列はグラフの隣接行列となる．また(b)で，行と列を異なる種類のノードに対応させ，要素をリンクの重みとすると2部グラフとなる．(c)で行と列が同じ種類のノードである場合には，一般のグラフとなる．共起行列の場合と同様に，グラフ上でもさまざまな距離（類似度）を定義することが可能である．このようにして構成したグラフ上での関係抽出はリンクマイニングと呼ばれ，近年盛んに研究されているところである．また共起行列の上で，相関が強い要素だけを残してグラフ表現に変換する考え方は，グラフィカルモデルの構築にも通じる．言語処理における適用事例はまだ多いとはいえないが，リンクマイニングの急速な展開を考えると，今後が期待される．

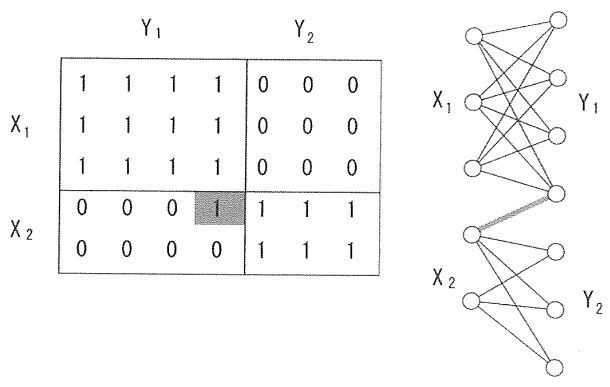

図5 共起行列と2部グラフ

6. まとめ

本稿では共起に基づくさまざまな類似性尺度について概観してきた．これらの共起尺度は，テキストの表記と意味とを対応づけるための分析手段として，さまざまな言語処理の場面で使われている．数理的な解釈に基づくと，これらの問題はコスト関数や尤度の最適化問題として定式化されるが，他の分野にも共通するように，想定するモデルや結果の妥当性の検証にはノウハウや経験が必要である．

参考文献

[1] 北研二「確率的言語モデル」，言語と計算/辻井潤一編；4，東京大学出版会（1999）．

[2] Christopher D. Manning and Hinrich Schütze: "Foundations of Statistical Natural Language Processing," MIT Press (1999).

[3] Akiko Aizawa: "The Feature Quantity: an Information Theoretic Perspective of Tfidf-like Measure," proc. of the 23st SIGIR, 104-111 (2000).

[4] Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull and Marianna Lau: "Okapi at TREC-3," proc. of the 3rd Text REtrieval Conference, 21-30 (1994).

[5] Ludovic Lebart, Andre Salem and Lisette Berry: "Exploring Textual Data," Text, Speech and Language Technology; 4, Kluwer Academic Publishers (1998).

[6] Hang Li and Naoki Abe: "Word Clustering and Disambiguation Based on Co-occurrence Data," proc. of COLING-ACL '98, 749-755 (1998).

[7] Fernando Pereira, Naftali Tishby and Lillian Lee: "Distributional Clustering of English Words," proc. of the 31st ACL, 183-190 (1993).

[8] Noam Slonim and Naftali Tishby: "Document Clustering Using Word Clusters via the Information Bottleneck Method," in proc. of the 23rd ACM SIGIR, 208-215 (2000).

[9] Inderjit S. Dhillon: "Coclustering Documents and Words using Bipartitle Spectral Graph Partitioning," proc of the 7th ACM SIGKDD, 269-274 (2001).

[10] Hongyuan Zha, Xiaofeng He, Chris H. Q. Ding, Ming Gu and Horst D. Simon: "Bipartite Graph Partitioning and Data Clustering," proc. of CIKM '01, 25-32 (2001).