テキストコーパスからのトピック階層の抽出

梶 博行[†] 森本康嗣[†] 相薗敏子[†]

タームの共起性に基づく関連シソーラスは，コーパスから自動生成することができるという利点を持つ．しかし，情報検索システム内部の処理に向いたシソーラスであり，情報空間を人間が理解するという目的には必ずしも適していない．関連シソーラスを人間向きのシソーラスに高めるため，関連シソーラスからトピック階層とトピックの代表タームリストを抽出する方法を開発した．最初に，比較的頻度の高いタームをクラスタリングすることにより，包括的なトピックの集合を抽出する．次に，トピックの代表タームと関連が強いタームをクラスタリングすることにより，各トピックをより限定的なトピックに分割する．また，タームのトピック代表性の指標として新たに提案したタームグループ内累計相互情報量を用いて，トピックの代表タームを抽出する．これらの処理により，トピックを表すタームリストが階層的に配置されたシソーラスを生成する．日本語の新聞記事コーパスを用いた評価実験では，最上位のタームリストの 89%，下位のタームリストの 76%が有効なトピックを示唆し，代表タームの 3 分の 2 近くがトピックのコアタームであった．この結果，本方法が大規模コーパスの情報空間を可視化する効果的な手段であるとの結論を得た．また，プロトタイプの試用を通じて，シソーラス作成ツールおよび文書データベースのブラウジングツールとしての実際的な効果を確認した．

Extracting a Topic Hierarchy from a Text Corpus

Hiroyuki Kaji,[†] Yasutsugu Morimoto[†] and Toshiko Aizono[†]

Co-occurrence-based association thesauri have the advantage that they can be generated from text corpora automatically. However, they are rather machine-oriented thesauri. To augment association thesauri to human-oriented ones, a method for extracting a hierarchy of topics and a list of representative terms for each topic was developed. First, a set of generic topics is extracted by clustering terms of relatively high frequency. Then, each topic is divided into more specific topics by clustering terms related to its representative terms. The representative terms are extracted by using a new measure of topic-representativeness defined as the sum of mutual information within a term group. The resultant thesaurus is a hierarchy of term lists each of which represents a topic. An experiment using a Japanese newspaper article corpus demonstrated that the method is an effective means of visualizing the information space of a large corpus; 89% of the top-level term-lists and 76% of the lower level ones were informative, and almost two out of three representative terms were core terms of the topics. Furthermore, the practical effectiveness of the method not only as a thesaurus construction tool but also as a document-database browsing tool was confirmed through trial use of the prototype.

1. はじめに

パーソナルコンピュータの普及とともに文書の電子化が急速に進んでいる．ネットワークを通じて送られてくるメールやニュースも増加するいっぽうである．個人や企業が保有する大量のテキスト情報を有効に利用できるようにすることが重要な課題である．

テキストデータベースへの効果的なアクセスを可能にするために，データベースの情報空間をコンパクトに表現するシソーラスが必要である．しかし，シソーラスが整備されている分野は，医学や工学など特定の専門分野に限られる．また，手作りのシソーラスでは情報内容の変化に追随することが困難である．このため，テキストデータベースすなわちコーパスからシソーラスを自動生成する技術が望まれている．

シソーラスにはいくつかのタイプがあるが，本論文では，タームの共起性に基づく関連シソーラス（association thesaurus）に着目した．関連シソーラスは，コーパスの統計的な処理により自動生成することが可能である．しかし，質問拡張など，情報検索システム内部の処理で利用することを主目的とするシ

† 日立製作所中央研究所
Central Research Laboratory, Hitachi, Ltd.

ソーラスである[1)~3)]．人間が情報空間を理解するのを助けるという目的には適さない．本論文では，構造化とタームの精選により，関連シソーラスを人間向きのシソーラスに高める方法を提案する．

以下，2 章で基本的なアイデアを述べ，3 章で提案方法を詳細に述べる．4 章で代替案と比較したあと，5 章で新聞記事コーパスを用いて評価し，6 章で実際的な利用と効果について述べる．さらに，7 章で今後の課題を述べ，8 章で関連研究と比較する．

2. 基本アイデア

2.1 関連シソーラスとその問題点

関連シソーラスは，2 つのタームとそれらの間の関連度を表す数値からなる三つ組データの集合である．本論文では，関連度として相互情報量を用いる．2 つのターム t，t' の間の相互情報量 $MI(t,t')$ は次式で定義される[4)]．

$$MI(t,t') = \log_e \frac{g(t,t') \Big/ \sum_{t,t'} g(t,t')}{\left\{ f(t) \Big/ \sum_{t} f(t) \right\} \cdot \left\{ f(t') \Big/ \sum_{t'} f(t') \right\}}.$$

ここに，$f(t)$ は t の出現頻度，$g(t,t')$ は t と t' の共起頻度である．

コーパスに対応する情報空間を人間が理解するという目的からみると，関連シソーラスには次の問題点がある．

(1) 構造

3 つ組データの単なる集合であるため，全体的な構造が理解しにくい．タームの数が多いので，タームの関連ネットワークとして分かりやすく図示することも困難である．

(2) タームの質

人間が知りたいのは索引づけや検索のキーとなるタームであるが，自動生成された関連シソーラスは，通常，一般的な語を多数含んでいる．一般語を除去するため，ストップワードリストが用いられている．しかし，一般語は非常に数が多いので，網羅的なストップワードリストを作成することは難しい．ストップワードとすべきかどうかが分野に依存する語もある．このため，ストップワードリストの効果は限られる．

2.2 関連シソーラスの構造化とタームの精選

1 つのトピックに関係のあるタームはテキスト中で共起することが多いので，それらの間の相互情報量は大きな値になる．したがって，相互情報量をタームの近さと考えてクラスタリングすれば，トピックに対応するタームのクラスタを抽出することができる．このクラスタは，特定のトピックに関係の深いタームの集合であり，類似の意味を表すタームの集合ではない．狭義には類似要素の集合を意味する “クラスタ” という用語を用いると，誤解をまねくおそれがある．本論文では，クラスタリング処理を記述する際にはクラスタという用語を用いるが，結果として得られるタームのクラスタを “タームグループ” と呼ぶ．

タームグループを抽出することにより，関連シソーラスに中間的な構造を与えることができる．しかし，大規模なコーパスには多数のトピックが含まれるので，タームグループの数も多くなる．また，トピックの粒度，すなわち 1 つのトピックとしてまとめるべきタームグループの大きさに最適な値があるわけではない．そこで，包括的なトピックから限定的なトピックまでさまざまなレベルのトピックを抽出し，トピックの階層として関連シソーラスを構造化する．

構造化の次には，タームグループからトピックが認知できるかどうかが問題になる．クラスタリング処理で混入したノイズを除去する必要がある．また，瞬時にトピックが認知できるよう，ターム数を一瞥できる程度に絞るべきである．そこで，タームグループから “代表ターム” を選択し，“代表タームリスト” でトピックを表現する．

本論文では，代表タームリストを選択するまでの一連の処理を “トピック抽出” と呼ぶ．厳密には，トピックを 1 つの語または短いフレーズで表現して，初めてトピック抽出というべきである．しかし，精選された代表タームを出力すれば，人間は容易にトピックを認知することができる．したがって，代表タームリストの選択までででも，トピック抽出と呼んでよいであろう．

2.3 トピック階層の段階的な抽出

トピックの階層を抽出する素朴な方法を**図 1** に示す．関連シソーラスに含まれるタームの集合に階層的クラスタリングアルゴリズムを適用して，入れ子になったタームグループ群を抽出したあと，各タームグループから代表タームを選択する．この方法には次の問題点がある．

第 1 の問題点は，ターム数が増加すると計算量が急激に増大することである．ターム数を P としたとき，一般的な階層的クラスタリングアルゴリズムの計算量は P^2 に比例する[5)]．このため，大規模な関連シソーラスに図 1 の方法を適用することは困難である．

第 2 の問題点は，クラスタ間のオーバラップが制限されることである．階層的クラスタリングアルゴリズ

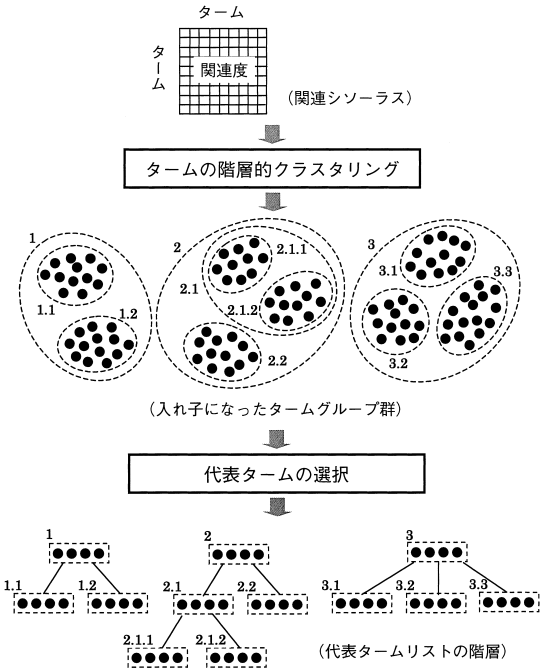


図 1 トピック階層抽出の素朴な方法
Fig. 1 A naive method for extracting a topic hierarchy.

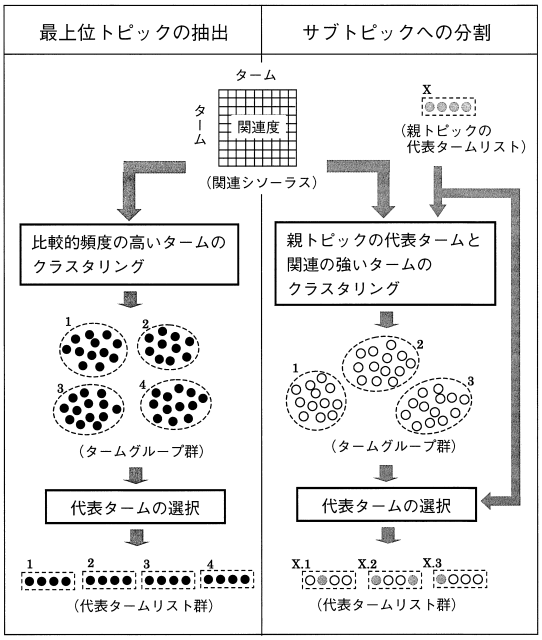


図 2 トピック階層抽出の提案方法
Fig. 2 Proposed method for extracting a topic hierarchy.

ムにおける 2 つのクラスタの関係は,「一方が他方を完全に包含する」,「共通部分がまったくない」のいずれかである．部分的にオーバラップするクラスタが生成されることはない．複数のトピックに関係するタームが存在することを考えると，この制限は不都合である．クラスタ間の部分的なオーバラップを許す非階層的クラスタリングアルゴリズムも考案されているが，計算量の点でトピック階層の抽出には適用困難である．たとえば，文献 6) に記述されているアルゴリズムの計算量は，P 個の要素をクラスタリングし，最大 $(k-1)$ 個の要素のオーバラップを許すとき，P^{k+2} に比例する．

これらの問題点を解消するため，本論文では，上位のトピックから下位のトピックへと段階的に抽出する方法を提案する．**図 2** に示すように，(1) 最上位トピックの抽出と，(2) トピックのサブトピックへの分割の 2 つのステップから構成される．ステップ (2) で得られるサブトピックをステップ (2) で再分割することもできるので，任意の深さのトピック階層を抽出することができる．各ステップの処理は次のとおりである．

(1) 最上位トピックの抽出

最上位のトピックは包括的なトピックであるから，その代表タームは出現頻度が比較的高いタームである．したがって，最上位トピックの抽出のために，関連シソーラスに含まれるすべてのタームをクラスタリングする必要はない．出現頻度が一定値以上のタームをクラスタリングし，得られたタームグループから代表タームを選択する．

(2) トピックのサブトピックへの分割

サブトピックを特徴づけるタームは，一般に親トピックの代表タームと関連が強い．そこで，親トピックの代表タームと関連が強いタームをクラスタリングし，得られたタームグループから代表タームを選択する．ここで，親トピックの代表タームは，複数のサブトピックに関係することが多いので，特別な扱いをする．すなわち，クラスタリングの対象タームからは除外するが，各サブトピックの代表タームの候補には含める．

本方法によれば，前述の問題点は次のように解決される．

(a) クラスタリング処理の計算量

ステップ (1) あるいは (2) におけるクラスタリングの対象ターム数を p，生成されるクラスタ数を c，また全体として抽出するトピック階層の深さを d とすると，ステップ (1) とステップ (2) の繰返しを合わせた全体の計算量は $(1+c+c^2+\cdots+c^{d-1})p^2$ に比例する．関連シソーラスの総ターム数を P とすると，$1+c+c^2+\cdots+c^{d-1} < p \ll P$ であるから，P が非

常に大きくなると $(1+c+c^2+\cdots+c^{d-1})p^2 \ll P^2$ である．したがって，本方法は，図 1 の方法が適用できない大規模な関連シソーラスにも適用可能である．

(b) クラスタ間のオーバラップ

ステップ (2) では，分割すべきトピックに応じてクラスタリング対象タームを決定する．その際，他のトピックのタームグループとの重複や，他のトピックを分割するためのクラスタリング対象タームとの重複も排除しない．したがって，ステップ (1) とステップ (2) の繰返しを合わせた全体の結果としては，オーバラップしたクラスタ群が得られる．

各ステップで用いるクラスタリングアルゴリズム自体はオーバラップを許さないので，1 つのステップで得られる兄弟クラスタの間ではオーバラップは生じない．しかし，ステップ (2) の説明で述べたように，親トピックの代表タームをすべてのサブトピックの代表ターム候補とする．したがって，親トピックの代表タームに限って，兄弟クラスタ間でもオーバラップが許される．

2.4 代表タームの選択

タームグループ（クラスタ）の代表タームは，クラスタのセントロイドに近いタームを選択するのが最も一般的な考え方であろう．セントロイドは，クラスタのメンバータームとの関連度の総和（あるいは平均）が大きい仮想的な要素である．したがって，メンバータームとの関連度の総和が大きいタームがセントロイドに近いタームである．そこで，ターム t のタームグループ C 内の累計相互情報量 $SMI(t,C)$ を次式で定義する．

$$SMI(t,C) = \sum_{t' \in C, t' \neq t} MI(t,t').$$

そして，タームグループ内累計相互情報量が大きいタームをタームグループの代表タームとして選択する．

この方法によれば，タームグループ内の多くのタームと関連を持つターム，いい換えるとタームグループのトピックを連想させるタームが選択される．比較的少数のメンバータームと非常に強い関連を持つタームも選択されるが，それも望ましいことである．タームグループに含まれる重要なサブトピックを代表するタームと考えられるからである．共通部分の少ない複数のトピックを表すタームグループ，いわば複数のセントロイドを持つクラスタに対してもタームグループ内累計相互情報量は有効と思われる．各セントロイドに近いタームのタームグループ内累計相互情報量が大きくなるからである．

タームグループ内累計相互情報量の定義式において，t が C のメンバータームである必要はない．すなわち，非メンバータームに対してもタームグループ内累計相互情報量を計算することができる．このことが，トピックのサブトピックへの分割における親トピック代表タームの特別な扱いを可能にしている．すなわち，親トピックの代表タームは，クラスタリングの対象タームから除外されるので，どのタームグループのメンバーにもならない．しかし，どのタームグループに対してもタームグループ内累計相互情報量を計算することができるので，代表タームの候補としてメンバータームと同様に扱うことができる．

3. 提案方法の詳細

提案方法は，(I) 関連シソーラスの生成，(II) 最上位トピックの抽出，(III) トピックのサブトピックへの分割の 3 つのステップから構成される．大規模なコーパスを対象にするので (I) と (II) はバッチ処理とせざるをえないが，(III) は実時間処理を想定する．

3.1 関連シソーラスの生成

コーパスから関連シソーラスを生成するステップは，(1) ターム抽出，(2) 共起データ抽出，(3) タームの関連度計算の 3 つのサブステップからなる．

(1) ターム抽出

テキストを形態素解析し，名詞，未知語，複合名詞を抽出する．コーパス中での出現頻度をカウントし，あらかじめ定めた閾値以上の出現頻度を持つ語を以降のサブステップの処理対象とする．

タームとは，本来，専門分野で特定の意味を持つ語である．タームの自動抽出は，近年，さかんに研究されているが，基礎的な研究の段階にとどまっている[7),8)]．そこで，本研究では，上に述べたようにタームの選定基準として品詞と頻度のみを用いることにした．タームの多くは名詞である．専門用語や固有名詞は辞書に登録されていないことが多いので，未知語も抽出する．また，専門用語には複合名詞が多いので，名詞と未知語の並びを複合名詞として抽出する．なお，分野によっては動詞のタームも重要であるが，サ変動詞語幹を名詞として扱うほかは対象外とした．

(2) 共起データ抽出

コーパスからウィンドウ内に共起するタームのペアを抽出し，共起頻度をカウントする．ウィンドウとは一定数の語を収容する枠である．ウィンドウをテキストに沿って移動させながら，ウィンドウ内に同時に出現しているタームのペアを抽出する．

関連シソーラスの生成では，従来，文書共起を採用することが多かった．同一文書に出現することを共起

と考える方法である．単一のトピックについて記述した文書を対象とする場合は，文書共起でよかった．本研究では，複数のトピックを含む文書への適用を考慮して，ウィンドウ共起を採用した．ここで，ウィンドウのサイズが問題である．一般にパラグラフが 1 つのまとまった内容を表すので，パラグラフのサイズが 1 つの目安となる．しかし，ウィンドウサイズを大きくすると，意味的な関連のないタームのペアも増加する．妥協点として，2〜3 文をカバーする程度のウィンドウを用いる．

(3) タームの関連度計算

ターム間の関連度として相互情報量を計算し，あらかじめ定めた閾値で足切りする．すなわち，t と t' の相互情報量 $MI(t,t')$ が閾値に満たないなら，$MI(t,t')$ を 0 にする．また，相互情報量は，出現頻度の小さいタームに対して過大評価される傾向がある[9]．そこで，対数尤度比による検定を行い，t と t' の関連が有意でなければ，$MI(t,t')$ を 0 にする．

3.2 最上位トピックの抽出

タームのクラスタリングには，凝集的なクラスタリング方法の 1 つであるグループ平均法を用いる[5]．また，実行時に次のパラメータを指定する．

- p：クラスタリングの対象ターム数
- q：クラスタサイズの上限
- r：1 つのトピックの代表ターム数

アルゴリズムは次のとおりである．

1) クラスタリング対象タームの選択

コーパス中での出現頻度の大きい順に p 個のタームを選択する．

2) タームクラスタリング

a) 各タームをそれぞれ 1 つのクラスタとする．

b) 関連度が最大のクラスタのペアを選択し，1 つのクラスタにマージする．ただし，マージするとクラスタサイズが q を超えるペアは対象外とする．ここで，クラスタ C と C' の関連度 $R(C,C')$ は次式で計算する．

$$R(C,C') = \underset{t\in C, t'\in C'}{\text{ave}}\, MI(t,t').$$

すなわち，C 中のタームと C' 中のタームの関連度をすべてのタームの組合せについて平均する．

関連度が 0 でなく，マージしてもクラスタサイズが q を超えないペアがある限り，b) を繰り返す．

c) サイズが r 以上のクラスタを選択する．

3) 代表タームの選択

タームグループ（クラスタ）C_i の代表タームとして，タームグループ内累計相互情報量 $SMI(t,C_i)$ の大きい順に r 個のタームを C_i から選択する $(i=1,2,\cdots)$．

3.3 トピックのサブトピックへの分割

タームのクラスタリングには，最上位トピックの抽出ステップと同様にグループ平均法を用いる．また，実行時に次のパラメータを指定する．

- p'：クラスタリングの対象ターム数の上限
- q'：クラスタ数
- r'：1 つのトピックの代表ターム数

アルゴリズムは次のとおりである．

1) クラスタリング対象タームの選択

親トピックの代表タームリストが $\{t_1,t_2,\cdots,t_n\}$ であるとする．各代表ターム t_i に対して，t_i との関連度（相互情報量）の大きい順に最大 p'/n 個のタームを集める．それらの和集合から $\{t_1,t_2,\cdots,t_n\}$ を除いた集合 T をクラスタリングの対象とする．

2) タームクラスタリング

a) T の各タームをそれぞれ 1 つのクラスタとする．

b) 関連度が最大のクラスタのペアを選択し，1 つのクラスタにマージする．ただし，マージするとクラスタサイズが $|T|/q'$ を超えるペアは対象外とする．ここで，クラスタ C と C' の関連度 $R(C,C')$ は次式で計算する．

$$R(C,C') = \underset{t\in C, t'\in C'}{\text{ave}}\, MI(t,t').$$

関連度が 0 でなく，マージしてもクラスタサイズが $|T|/q'$ を超えないペアがある限り，b) を繰り返す．

c) サイズの大きい順に q' 個のクラスタを選択する．

3) 代表タームの選択

タームグループ（クラスタ）C_i の代表タームとして，タームグループ内累計相互情報量 $SMI(t,C_i)$ の大きい順に r' 個のタームを $C_i \cup \{t_1,t_2,\cdots,t_n\}$ から選択する $(i=1,2,\cdots,q')$．

4. 予 備 実 験

4.1 代替案との比較

提案方法に対してさまざまな代替案が考えられるので，それらと比較する実験を行った．実験に用いたコーパスは，毎日新聞の経済面記事 5 年分（35.5 M バイト，約 41,000 記事）である．関連シソーラス生成のパラメータは以下のように設定した．

– タームの出現頻度の閾値：10
– ウィンドウサイズ：25 語（内容語のみの語数）
– 相互情報量の閾値：2.0

その結果得られた，約 29,000 タームの関連シソー

ラスを実験に利用した．

4.1.1 クラスタリングアルゴリズム

凝集的クラスタリングの代表的な方法として，グループ平均法のほかに単一連結法，完全連結法がある[5]．これらの比較実験を行い，タームクラスタリングにはグループ平均法が適していることを確認した．単一連結法では，クラスタが芋づる式につながりやすいというこの方法の欠点が顕著に現れ，関係の薄い複数のトピックが 1 つのクラスタに融合してしまった．完全連結法では，クラスタがトピックを表すほど大きくならないうちに，マージできるクラスタのペアがなくなった．シソーラスデータが疎であることがその原因である．

次に，グループ平均法で，(i) クラスタサイズに上限を設けてクラスタ数を間接的に制御する方法と，(ii) クラスタサイズに上限を設けずクラスタ数を直接制御する方法を比較した．提案方法である (i) は，関連度の閾値がクラスタによって違ってしまうという点で変則的な方法である．関連度の閾値がすべてのクラスタに共通となる (ii) が一般的な方法である．しかし，(ii) では，1 つのクラスタにトピックが融合してしまい，残りはたかだか 2，3 個のタームからなるクラスタであった．シソーラスデータが疎であることが主な原因である．いっぽう，(i) では，トピックの融合が適当な段階で停止した．

4.1.2 最上位トピックの抽出のためのクラスタリング対象ターム

最上位トピックの抽出のためにクラスタリングするタームの選択方法として次の 4 案を比較した．

(i) 頻度順（2,000/4,000 ターム）：提案方法．
(ii) 特徴ターム頻度順（2,000/4,000 ターム）：文書の特徴を表すタームを重視する方法である．具体的には，tf-idf（term frequency-inverse document frequency）[10] 順に上位 20%のタームを各文書の特徴タームと考え，特徴タームとして出現する文書の多い順にタームを選択した．
(iii) ランダム（2,000/4,000 ターム）
(iv) 頻度順（10,000 ターム）：提案方法でターム数を多くした場合である．関連シソーラスの全タームを対象とするクラスタリングが処理時間的に困難であるため，その代用として実行した．

これらの方法の結果を，タームグループの代表タームリストからトピックが認知できるか，またトピックの粒度は揃っているかという観点から評価した．その結果，(i) と (ii) が良好であった．どちらも，トピックを表すタームグループが多く，トピックの粒度も揃っていた．抽出されたトピックの集合も両者は類似していた．(iii) では，トピックを表すタームグループがほとんど得られなかった．ランダムサンプリングによって，疎なシソーラスデータがいっそう疎になるためである．(iv) では，トピックの粒度が不揃いであった．すなわち，関係が薄いトピックが融合したタームグループとともに，サブトピックとして抽出されるべき限定的なトピックを表すタームグループが散見された．計算量だけでなく質の面からも，クラスタリング対象タームを適切に限定することが必要といえる．(i) と (ii) の間で質の優劣はつけ難く，処理の軽い (i) を採用することにした．

4.1.3 サブトピックへの分割のためのクラスタリング対象ターム

トピックのサブトピックへの分割のためにクラスタリングするタームの選択方法として次の 3 案を比較した．

(i) 親トピックの代表タームの関連ターム（親トピックの代表タームを除外）：提案方法．
(ii) 親トピックの代表タームの関連ターム：親トピックの代表タームの扱いのみが提案方法と異なる．
(iii) 親トピックのタームグループ：図 1 の方法，すなわちクラスタ間のオーバラップを許さない方法．

これらの方法の比較は，同一の親トピックを用いて行うべきである．(iii) の入力として，ターム数が多いタームグループが必要であるので，4.1.2 項の (iv) で得られたトピックを親トピックとした．全体的に，(i) や (ii) では数個のタームグループが得られるのに，(iii) では 1〜2 個のタームグループしか得られないというケースが多かった．すなわち，(iii) はサブトピックの抽出能力が低いことが分かった．

サブトピックが抽出できた場合でも，タームグループの質の面で，(iii) は (i) や (ii) より劣っていた．同一の親トピックを各方法で分割した結果の例を**図 3** に示す．(a) のトピックを (i)，(ii)，(iii) の各方法で分割した結果が (b)，(c)，(d) である．(b) や (c) はサブトピックが明確であるが，(d) は焦点がややぼやけている．関連タームを集め直したうえでクラスタリングするのが効果的であることが分かる．

(i) と (ii) については，さらに 4.1.2 項の (i) で得られたトピックを親トピックとして比較実験を行った．(i) と (ii) の相違点は親トピックの代表タームの扱いである．親トピックの代表タームはサブトピックでも重要なタームであるので，これをクラスタリング対象タームから除外するとサブトピックの抽出が困難になる可能性がある．逆に，複数のサブトピックに関係す

機種，パソコン，デジタル，携帯，音声，画面，通話，電話，携帯電話，映像，シャープ，通信，ソフト，アナログ，ＮＥＣ，液晶，回線，端末，画像，メモリー

(a) 親トピックの代表ターム

携帯電話，移動通信，通話，通話料金，日本移動通信，新規加入，移動電話，音声，アナログ，売り切り，ＰＨＰ，ＰＨＳ，携帯，自動車電話，電話，通信，電電，回線，デジタル，新電電

映像，液晶，画像，薄膜トランジスタ，左右，画面，スクリーン，ＤＶＤ，デジタル，ワイドテレビ，パソコン，ソフト，音声，ハイビジョン放送，ＴＦＴ，端末，シャープ，ＮＥＣ，立体，アナログ

(b) 代表タームの関連ターム（代表タームを除外）のクラスタリング

新規加入，新規加入料，移動通信，移動電話，電話，携帯電話，通話，回線，日本移動通信，自動車電話，通信，ＰＨＳ，通話料金，売り切り，携帯，アナログ，通話料，新電電，電電，ＰＨＰ

ビデオカメラ，液晶，画面，静止，液晶モニター，液晶ビデオカメラ，ビューカム，ソフト，液晶ビューカム，シャープ，立体，端末，画像，電話回線，映像，音声，カラー液晶，左右，ワイドテレビ，デジタル

(c) 代表タームの関連タームのクラスタリング

インテル，宅配便，他社，ＩＢＭ，モトローラ，ＩＤＯ，ベンティアム，基地局，マイクロプロセッサー，分社化，ＣＰＵ，日本ＩＢＭ，中央演算処理装置，ふた，インテル社，日本移動通信，従来機，パワー，半導体メーカー，周波数

速さ，映像，立体映像，ヘッドホンステレオ，心臓部，速度，視聴者，世界最高速，電子部品，音声，ＡＶ，ディスプレー，コードレス電話，再生，電子，テープ，液晶，ＩＣ，高速，松下

(d) タームグループのクラスタリング

図 3 トピック分割時のクラスタリング対象ターム選択方法の比較
Fig. 3 Comparison of methods for selecting terms to be clustered for topic decomposition.

るタームを除外してクラスタリングすることにより，サブトピックの分離性が向上する可能性もある．実験の結果は，どちらも顕著には現れず，全体として (i) と (ii) は同等であった．

最後に，提案方法 (i) の特徴である，サブトピック間のタームのオーバラップについて述べる．図 3 (b) にオーバラップの例がみられる．すなわち，“アナログ” と “デジタル” が 2 つのサブトピック「通信」と「AV」に共通の代表タームとなっている．これに対して，図 3 (c) では，“アナログ” は「通信」のみの代表ターム，“デジタル” は「AV」のみの代表タームになっている．サブトピックの認知という意味では (b) と (c) の間に大きな差はない．しかし，サブトピックをさらに分割したり，代表タームリストを検索要求として用いたりすることを考えると，オーバラップに対する制限は少ないほうがよい．

4.1.4 代表タームの選択

タームグループの代表タームを選択する方法のベースラインとして，クラスタリング対象タームの選択方法に準拠した方法が考えられる．最上位トピック抽出のためのクラスタリング対象タームを選択する方法として，(i) 頻度順と，(ii) 特徴ターム頻度順が良好で

大蔵省，政府，見方，声，効果，意見，拡大，見直し，規模，予算，議論，国民，土地，景気対策，負担，要求，対策，説明，当面，個人

(a-1) 頻度順に選択した代表ターム

減税，消費税率，所得税減税，所得税，税率，税制改革，財源，増税，政府税調，税制，連立与党，税制改正，政府税制調査会，消費税，赤字国債，与党，補正予算，税，予算編成，加藤

(a-2) タームグループ内累計相互情報量の順に選択した代表ターム

(a) 頻度順に選択したタームのクラスタリングで得られたタームグループ

統合，ＥＵ，欧州，マルク，通貨，ＧＤＰ，四半期，商務，上海，中国，橋本，Ｇ，黒字，プラ，ベトナム，銘柄，品，制度，銀，自由化

(b-1) 特徴ターム頻度順に選択した代表ターム

長プラ，表面利率，表面，利率，長，プラ，長期プライムレート，貸出金利，プライムレート，貸出，長信銀，ＥＭＳ，ポンド，ＥＲＭ，欧州通貨，ＥＣＵ，中央銀行，貿易収支，通貨，通貨統合

(b-2) タームグループ内累計相互情報量の順に選択した代表ターム

(b) 特徴ターム頻度順に選択したタームのクラスタリングで得られたタームグループ

図 4 代表ターム選択方法の比較
Fig. 4 Comparison of methods for selecting representative terms.

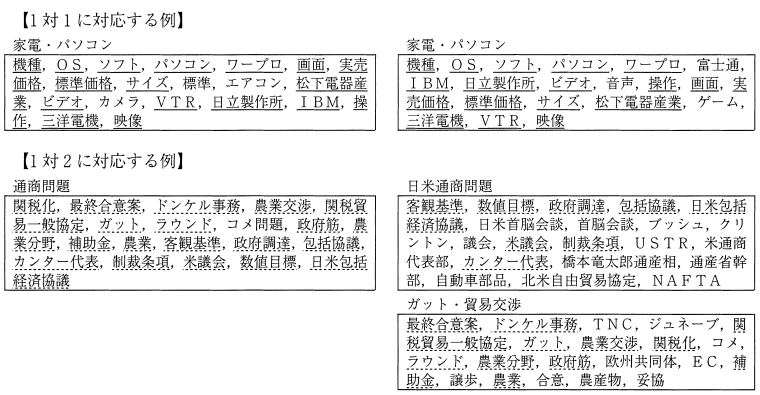
あったので，それらに準拠した方法とタームグループ内累計相互情報量に基づく方法を比較した．

図 4 (a) は，頻度順に選択したタームのクラスタリングによって得られたタームグループに対する代表タームリストの比較例である．代表タームも頻度順に選択した結果が (a-1)，タームグループ内累計相互情報量に基づいて選択した結果が (a-2) である．同一のタームグループに対する代表タームリストとは思えないほど，それらの結果は異なっている．(a-2) はトピック「税・財政」を表すが，(a-1) がトピックを表すとはいい難い．図 4 (b) は，特徴ターム頻度順に選択したタームのクラスタリングによって得られたタームグループに対する代表タームリストの比較例である．代表タームも特徴ターム頻度順に選択した結果が (b-1)，タームグループ内累計相互情報量に基づいて選択した結果が (b-2) である．(b-2) はトピック「金利・通貨」を表すが，(b-1) がトピックを表すとはいい難い．

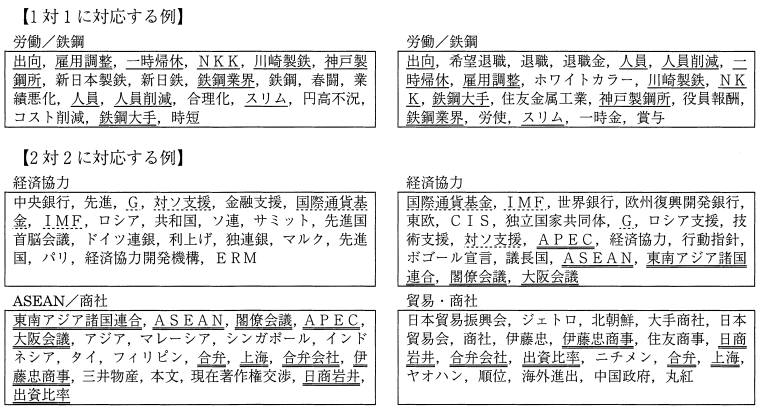
ベースラインとの比較を通じて，タームグループ内累計相互情報量の有効性が明らかになった．頻度や特徴ターム頻度がコーパス全体でのタームの重要度を表すのに対し，タームグループ内累計相互情報量はタームグループにおけるタームの重要度を表す．この違いが代表タームの選択における有効性の差を生じさせたと思われる．

4.2 パラメータ値の有効範囲

提案方法が広い範囲のパラメータ値で動作することと，パラメータ値の変更による結果の変化に連続性があることを確認する実験を行った．



(a) p が同じで q が異なるケース [左：(p=2000, q=150, r=20)，右：(p=2000, q=100, r=20)]



(b) p/q が同じケース [左：(p=2000, q=100, r=20)，右：(p=4000, q=200, r=20)]

（注 1）タームグループが表わすトピックを代表タームリストの上部に記した．
（注 2）左右の代表タームリストに共通のタームに下線を付した．

図 5 異なるパラメータ値で抽出されたトピックの比較
Fig. 5 Comparison of topics extracted with different parameter values.

4.2.1 最上位トピックの抽出

前節の実験と同じ経済面記事対応の関連シソーラスの場合，クラスタリング対象ターム数 p が 500〜4,000，クラスタサイズの上限 q が $\max\{50, p/30\}$〜$p/10$ の範囲で良好な結果が得られた．また，トピックの認知という観点から，代表ターム数 r は 15〜20 が必要十分であることが分かった．

4.1.1 項で述べたように，抽出するトピック数を直接指定することはできないが，ほぼ p/q 個のトピックが抽出される．トピック数による抽出結果の変化をみるため，p が同じで q が異なるケースの間で，抽出されたトピックを比較した．たとえば，$(p = 2{,}000,\ q = 150,\ r = 20)$ のケースと $(p = 2{,}000,\ q = 100,\ r = 20)$ のケースの間では，1 対 1 対応が 8 組，1 対 2 対応が 4 組，2 対 3 対応が 1 組，0 対 1 対応が 1 つであった．**図 5** (a) に 1 対 1 対応の例と 1 対 2 対応の例を示す．これらの例からも分かるように，抽出するトピック数を変えたとき，結果の変化には連続性がある．

次に，パラメータ値は異なるがほぼ同数のトピックを抽出するケースを比較した．たとえば，$(p = 2{,}000,\ q = 100,\ r = 20)$ のケースと $(p = 4{,}000,\ q = 200,\ r = 20)$ のケースの間では，1 対 1 対応が 12 組，1 対 2 対応が 2 組，2 対 1 対応が 1 組，2 対 2 対応が 2 組，0 対 1 対応が 1 つであった．図 5 (b) に 1 対 1 対応の例と 2 対 2 対応の例を示す．これらの例からも分かるように，p/q の値が同じケースでは類似の結果が得られる．

4.2.2 トピックのサブトピックへの分割

サブトピックへの分割ステップでは，最上位トピックの抽出ステップと異なり，クラスタリング対象ターム間の関連度データが密であるので，指定された数のクラスタを生成することが可能である．このため，パラメータ q' で分割数を指定する方式としている．サブトピックへの分割は必要に応じて繰り返せばよいの

で，$q' = 2$〜7 に対する動作を確認した．クラスタリング対象ターム数の上限 p' は 400 とした．$p' = 400$ とすれば，$q' = 2$〜7 のとき，クラスタサイズの上限 (p'/q') が最上位トピックの抽出ステップにおける値 (q) に近くなるからである．また，代表ターム数 r' は親トピックの代表ターム数 r に一致させた．

同一の親トピックに対して q' の値を変えた結果を比較した．たとえば,「パソコン，家電」のトピックは，$q' = 2$ で「AV」と「パソコン」に分割され，$q' = 6$ で「ビデオ」,「テレビ」,「ゲーム機，映像」,「パソコン」,「パソコンソフト」,「家電」に分割された．このように，分割数を変えたとき，結果の変化には連続性がある．親トピックによっては，q' を大きくしても意味のある結果が得られなかった．たとえば,「株・為替」のトピックは，$q' = 2$ で「株式」と「外国為替市場」に分割されたが，それ以上小さなサブトピックを抽出することはできなかった．パラメータ q' の有効範囲は親トピックによって異なるので，サブトピックへの分割ステップはインタラクティブに利用することが必要である．

4.3 処理時間

各ステップの処理時間を UNIX ワークステーション（CPU クロック周波数：400 MHz，主記憶：512 MB）上で実測した．

4.1 節で述べた関連シソーラスの生成に要した時間は 226.7 分であった．内訳は，ターム抽出が 112.3 分，共起データ抽出が 94.7 分，タームの関連度計算が 19.7 分であった．ターム抽出と共起データ抽出の処理時間はコーパスの大きさに比例する．実際のシステムの運用では，テキストの蓄積と同時にこれらの処理を実行し，結果を累積していく．したがって，さらに大きいコーパスに対しても，関連シソーラス生成の処理時間がネックになることはない．

最上位トピックの抽出ステップの処理時間は，主にクラスタリングの対象ターム数 p に依存する．$p = 500$, $1,000$, $2,000$, $4,000$ として実測したところ，それぞれ 0.4，1.2，7.1，52.4 分であった．性能面からの p の上限が約 4,000 であるといえる．

サブトピックへの分割ステップの処理時間は，クラスタリングの対象ターム数 p' に依存する．$p' = 200$, 300，400，500 として実測したところ，それぞれ 1.5, 2.4，3.0，4.0 秒であった．したがって，このステップは実時間で利用することができる．

5. 評価

5.1 トピック認知テスト

コーパスの情報空間を可視化する手段としての提案方法の有効性を評価するため，トピック認知テストを行った．使用したシソーラスは，4 章の予備実験と同じ毎日新聞経済面記事コーパスから生成したシソーラスである．被験者は筆者の周辺から 10 名を選んだ．全員，知識労働者であるが，経済が専門ではない．文系出身者が 3 名，理系出身者が 7 名であった．年齢別では，20 歳台が 2 名，30 歳台が 5 名，40 歳以上が 3 名であった．本テストのシソーラスを見るのは全員が初めてであった．ただし，4 名は他分野のシソーラスによる利用経験（3 カ月から 1 年，ときどき利用）があった．

被験者のタスクは次のとおりである．

(1) トピックの記述

タームグループの代表タームリストから認知されるトピックを単語または短いフレーズで記述する．複数のトピックが認知される場合はそれらを併記する．単独でもトピックを連想させるタームがあるが，提案方法の趣旨に従い，少なくとも 5 個程度の代表タームが関係するトピックに限定する．

(2) 同一トピックを表すタームグループの同定

兄弟タームグループから認知されるトピックの差異をチェックする．広い意味では同一トピックであっても，焦点や観点が異なる場合は別のトピックと考える．なお，このチェックは，サブトピックのタームグループに対してのみ行った．同一トピックを表すタームグループがないことを予備実験で確認した最上位トピックに対しては省略した．

テストデータとして，最上位トピックの抽出ステップの実行結果 8 ケースとサブトピックへの分割ステップの実行結果 3 ケースを用いた．**図 6** に，代表タームリストと 2 名の被験者 A，B が記述したトピックを例示する．テスト結果を集計するため，被験者ごとにタームグループを ‘有効’ と ‘無効’ に分類した．他のタームグループとは異なるトピックを表すタームグループのほか，同一トピックを表すタームグループのうちで最初に提示されたタームグループが ‘有効’ である．たとえば，図 6 (c) の被験者 A の場合，(i) と (ii) のうち (i) だけが ‘有効’ である．なお，すべての兄弟タームグループが同一トピックを表す場合は，親トピックを分割することができなかったと考えられるので，すべての兄弟タームグループを ‘無効’ とした．

最上位トピックの抽出の 8 ケースとサブトピックへ

(i)	東南アジア諸国連合，ＡＳＥＡＮ，閣僚会議，ＡＰＥＣ，大阪会議，アジア，マレーシア，シンガポール，インドネシア，タイ，フィリピン，合弁，上海，合弁会社，伊藤忠商事，三井物産，本文，現在著作権交渉，日商岩井，出資比率	A：東アジア経済 B：アジア・商社・合弁
(ii)	市場金利，定額貯金，郵便貯金，金利自由化，定期預金，金利，短期金融市場，短期金利，金融市場，短期，貸出金利，長期プライムレート，長プラ，表面利率，利率，利回り，引き下げ，通貨供給量，マネーサプライ，代表的	A： 金利動向 B： 金融市場
(iii)	機種，ＯＳ，ソフト，パソコン，ワープロ，富士通，ＩＢＭ，日立製作所，ビデオ，音声，操作，画面，実売価格，標準価格，サイズ，松下電器産業，ゲーム，三洋電機，ＶＴＲ，映像	A： パソコン B： パソコン・家電
(iv)	総務庁，消費支出，世帯，年収，サラリーマン，庁，報告，金融制度調査会，諮問機関，審議，中間報告，カ年計画，規制緩和，建設省，道路，大都市圏，調査結果，略，マンション，地方自治体	A： ？ B： ？

(a) 最上位トピック（p=2000, q=100, r=20）

(i)	投資自由化，太平洋経済協力会議，事務レベル会合，高級事務レベル会合，行動指針，高級事務レベル，日章野靖夫，東南アジア諸国連合，ＡＳＥＡＮ，マニラ，フィリピン，閣僚会議，非公式首脳会議，ボゴール，大阪会議，経済協力，ＡＰＥＣ，ジャカルタ，経済閣僚会議，マハティール	A： APEC B： アジア経済協力
(ii)	企業化調査，ガス開発，石油ガス開発，マクダーモット，三井物産，サハリン沖，企業化，ペトロナス，マレーシア，鉱区，日商岩井，伊藤忠商事，合弁会社，合弁会社設立，上海市，上海，合弁，本文，現在著作権交渉，著作権	A： ？ B： エネルギー資源開発

(b) サブトピック（p'=400, q'=2, r'=20; 親トピックは(a)の(i)）

(i)	経済閣僚会議，チェンマイ，タイ北部，日章野靖夫，ＡＳＥＡＮ自由貿易地域，ＡＦＴＡ，自由貿易地域，東南アジア諸国連合，ＡＳＥＡＮ，大野俊，マニラ，スビック，フィリピン，タイ，ジャポニカ，シンガポール，リーソン，閣僚会議，リーソン容疑者，日本米	A： ASEAN B： 自由貿易
(ii)	投資自由化，太平洋経済協力会議，事務レベル会合，ＳＯＭ，高級事務レベル，特別会合，行動指針，高級事務レベル会合，経済協力，非公式首脳会議，スハルト，インドネシア，ジャカルタ，ボゴール，ボゴール宣言，大阪会議，ＡＰＥＣ，ＡＳＥＡＮ，東南アジア諸国連合，マハティール	A： ASEAN B： アジア経済協力
(iii)	ガス開発，石油ガス開発，サハリン沖開発，サハリン沖，ＳＯＤＥＣＯ，マクダーモット，企業化調査，鉱区，サラワク，国営石油会社，マレーシア，インドネシア，伊藤忠商事，日商岩井，サテライトジャパン，三井物産，ＩＣＤＣ，企業化，ＪＣＳＡＴ，ペトロナス	A： ガス田 B： エネルギー資源開発
(iv)	合弁会社，グンゼ，子供用，有限公司，上海市，上海，公司，出資比率，有限，合弁，重慶，合弁会社設立，広州，中国企業，ヤオハン，現地企業，伊藤忠商事，マレーシア，フィリピン，汽車	A： 海外生産 B： 日本企業の合弁会社

(c) サブトピック（p'=400, q'=4, r'=20; 親トピックは(a)の(i)）

図 6 代表タームリストとトピック記述の例
Fig. 6 Examples of representative-term list with topic descriptions.

の分割の 3 ケースの結果をそれぞれ**表 1** (a)，(b) にまとめた．すなわち，10 名の被験者それぞれに対して有効なタームグループの比率を算出し，その平均値，最低値，最高値を示している．それによると，平均で最上位トピックのタームグループの 89.2%，サブトピックのタームグループの 76.0%が有効であった．また，有効なタームグループの比率が被験者によって大きく変動することも分かる．表 1 (c) には，'有効' に分類した被験者数別のタームグループの分布を示す．最上位トピックでは被験者による判定のばらつきが小さいが，サブトピックで分割数を大きくすると被験者によるばらつきが大きくなる．詳細な内容のサブトピックの理解には，その分野の知識や想像力が要求されるためと思われる．

本テストでは，制限時間を設けなかったが，1 タームグループあたりの所要時間は，タスク (1) のみの最上位トピックの場合，全被験者の平均で 34 秒，最長の被験者で 51 秒であった．タスク (2) が加わったサブトピックの場合，全被験者の平均で 55 秒，最長の被験者で 90 秒であった．システムの利用経験者は所要時間が短いという傾向がみられたが，所要時間の長短や利用経験の有無と有効なタームグループの比率の間に相関は認められなかった．被験者の関心分野や知識の深さの影響が大きかった．本テストの結論として，提案方法は，情報を欲しているユーザに，それぞれの知識レベルに応じた有益な情報を提供するといえる．

5.2 代表タームの質の評価

タームグループ内累計相互情報量の有効性を評価するため，トピックとの関係によってタームを 3 つのカテゴリに分類し，タームグループ内累計相互情報量の値とカテゴリの関連を調べた．タームのカテゴリは次のとおりとし，分類作業は筆者らが行った．

表 1 トピック認知テストの結果
Table 1 Results of the topic recognition test.

(a) 有効なタームグループの比率（最上位トピック）

パラメータ (p, q, r)	タームグループ数	有効なタームグループの比率 (%)		
		平均	最低	最高
(500, 50, 15)	8	91.3	87.5	100
(1000, 50, 15)	20	86.0	75.0	100
(2000, 150, 20)	14	95.7	85.7	100
(2000, 100, 20)	22	87.3	68.2	100
(2000, 67, 20)	31	81.9	64.5	93.5
(4000, 300, 20)	17	93.5	70.6	100
(4000, 200, 20)	23	93.0	69.6	100
(4000, 133, 20)	32	91.3	71.9	100
合計	167	89.2	74.3	98.2

(b) 有効なタームグループの比率（サブトピック）

パラメータ (p', q', r')	タームグループ数	上段：有効なタームグループの比率 (%) 下段：ﾄﾋﾟｯｸが認知できたﾀｰﾑｸﾞﾙｰﾌﾟの比率 (%)		
		平均	最低	最高
(400, 2, 20)	44	83.6	68.2	95.5
		94.1	77.3	100
(400, 4, 20)	88	78.2	53.4	95.5
		93.3	84.1	98.9
(400, 6, 20)	132	72.2	45.0	93.2
		90.5	81.8	97.0
合計	264	76.0	53.0	91.7
		92.0	82.2	98.1

（注）親トピックは(a)の(p=2000, q=100, r=20)の結果を用いた．

(c) ‘有効’に分類した被験者数別のタームグループ数

ケース	‘有効’に分類した被験者数				
	0〜2	3〜5	6〜7	8〜9	10
最上位トピック (8 ケースの合計)	4 (2.4%)	7 (4.2%)	9 (5.4%)	53 (31.7%)	94 (56.3%)
サブトピック (p'=400, q'=2, r'=20)	0 (0.0%)	4 (9.1%)	8 (18.2%)	13 (29.5%)	19 (43.2%)
サブトピック (p'=400, q'=4, r'=20)	6 (6.8%)	9 (10.2%)	11 (12.5%)	35 (39.8%)	27 (30.7%)
サブトピック (p'=400, q'=6, r'=20)	13 (9.8%)	19 (14.4%)	28 (21.2%)	39 (29.5%)	33 (25.0%)

(i) コアターム：トピックの主要なオブジェクトや事象を表すターム．当該トピックに関する専門用語や企業/団体/国名などが典型的なコアタームである．

(ii) 関係ターム：コアタームと結び付くことによって，トピックをより明確，具体的にする効果を持つターム．コアタームの構成要素でコアタームを連想させるターム，コアタームを含む句もこのカテゴリに分類した．

(iii) 無関係ターム：非常に一般的な語およびトピックに関係のないターム．

表 2 は，各タームにタームグループ内累計相互情報量の値とカテゴリを付記したタームグループの例である．表 2 (a) は最上位トピックのタームグループである．表 2 (b) は，サブトピックのタームグループに親トピックの代表タームを追加したリストである．タームの順序はタームグループ内累計相互情報量の降順で，太枠で囲んだ部分が代表タームリストである．

最上位トピック 20 個とサブトピック 20 個を選び，それぞれに対して表 2 と同様な表を作成した．タームグループ内累計相互情報量の降順に第 1 位から第 n 位までのターム集合（$n = 10, 20, 30, 40, 60, 80$）のカテゴリ別ターム数を集計した結果を**表 3** に示す．n が大きくなると，コアタームの比率が減少し，無関係タームの比率が増加している．関係タームの比率は両者の中間で，初めは増加するが途中で減少に転じている．太枠で囲んだ $n = 20$ に対する値が代表タームリストの質を表す．すなわち，最上位トピックの代表タームの 67.0%，サブトピックの代表タームの 59.5%がコアタームであった．この比率は，タームグループ全体でのコアタームの比率に比べて非常に高い．タームグループ内累計相互情報量が代表タームの選択に有効な指標であることが分かる．

6. 実際的な利用の方法と効果

6.1 シソーラス作成ツールとしての利用

コーパス対応のシソーラスを作成するツールとして提案方法を利用する場合，人間が行う作業は次のとおりである．

(1) 最上位トピックの抽出ステップ，サブトピックへの分割ステップの順にいくつかのパラメータ値で実行し，最も適切と思われる結果を選択する．

(2) 代表タームリストのトピックを簡単な名詞（句）で記述する．

(3) 代表タームリストから不適当なタームを削除する．また，タームグループの中から適当なタームを代表タームリストに加える．

この方法により，大規模なシソーラスを短時間で，したがって低コストで作成することができる．たとえば，4，5 章で用いたコーパスから新聞の経済面対応のシソーラスを約 15 時間で作成することができた．うち，シソーラスの骨格を決める (1) と (2) に要した時間は約 4 時間であった．作成したシソーラスの規模は，トピック数が 3 階層で合計 135，ターム数が約 2,600 であった．

6.2 文書データベースブラウジングツールとしての利用

提案方法を文書データベースのブラウジングに利用するため，代表タームリストを検索要求として文書検索エンジンに渡すインタフェースを付加したプロトタイプを開発した[12),13)]．ユーザとシステムの対話は次のように進む．最初に，事前に抽出された最上位トピック群が表示される．ユーザが興味を持ったトピックを選択すると，サブトピック群が表示される．これはいわばズームインである．ズーム倍率（パラメータ

表 2 タームグループの例
Table 2 Examples of term group.

(a) 最上位トピックのタームグループ

ターム	SMI	CAT	ターム	SMI	CAT	ターム	SMI	CAT	ターム	SMI	CAT
東南アジア諸国連合	69.67	**	中国	34.46	**	商社	21.25	**	米国企業	8.67	
ASEAN	65.64	**	住友商事	33.84	**	北京	21.07	**	研究開発	8.30	
マレーシア	63.74	**	三菱商事	32.30	**	自由化	20.47		危険	7.21	
インドネシア	60.34	**	現地	31.85		北朝鮮	20.44	**	技術	6.68	
合弁会社	60.26	**	インド	31.08	**	韓国	20.20	**	実現	6.46	
合弁	59.73	**	現地法人	30.99	*	オーストラリア	19.89	*	ブランド	6.32	
シンガポール	57.14	**	バンコク	30.36	**	協力	19.33		研究	5.86	
APEC	56.51	**	大手商社	29.75	**	市	19.08		局	5.40	
閣僚会議	54.41	*	城山	29.36		出資	18.92	*	リスク	4.57	
タイ	54.29	**	石坂泰	28.85		豪州	18.69	*	扱い	4.55	
大阪会議	49.90	*	会議	28.18		設立	17.64		準備	4.43	
フィリピン	48.88	**	連星	27.35		核	16.25		原則	4.30	
三井物産	45.65	**	経済観測	26.64		日米関係	15.38		伊勢丹	3.73	
アジア	44.91	**	議長	26.09		本	14.29		関係	2.69	
本文	44.32		日本企業	25.84	*	羽田	13.56		要素	2.48	
上海	43.98	**	診断	25.46		日本人	12.28		現実	2.31	
伊藤忠商事	42.96	**	東南アジア	24.80	**	設計	12.24		不可欠	2.29	
現在著作権交渉	41.91		特許	24.69		経済成長	11.32	*	実情	2.29	
日商岩井	40.85	**	現在	24.33		成功	10.25		懸念	2.28	
出資比率	40.18	*	行動	23.89		構想	10.05		カギ	2.20	
ベトナム	36.56	**	事務局	23.60		不満	9.81		意向	2.14	
丸紅	36.27	**	資本金	22.50		決意	9.00		注目	2.07	
アジア諸国	35.54	**	生活設計	21.84		表明	8.82		管理	2.07	
表示	35.09		香港	21.69	**	行方	8.74		場所	2.07	
台湾	34.62	**	進出	21.67		夢	8.69		要請	2.05	

(b) サブトピックのタームグループ

ターム	SMI	CAT	ターム	SMI	CAT	ターム	SMI	CAT	ターム	SMI	CAT
(三井物産)	142.74	**	製油	48.54	*	(シンガポール)	34.23		タス通信	24.55	
企業化調査	111.99		液化天然ガス	47.25	**	石油鉱区	34.07	*	(出資比率)	24.38	*
ガス開発	92.24	*	サハリン石油開発			伊藤忠	33.60	**	尋俊邦	22.75	
(伊藤忠商事)	88.14	**	協力	46.64	**	サテライト	32.82		室伏稔	21.89	
ペトロナス	85.68	**	ロシア最高会議	46.47		サハリン開発	32.73	**	(大阪会議)	20.78	
企業化	84.71	*	製油所	46.30	*	連結対象会社	32.48		連結対象	20.35	
(マレーシア)	84.46	**	日サ	45.80		タイム	32.19		米倉功	19.95	
石油ガス開発	84.26	**	インドネシア石油	45.42	**	(フィリピン)	31.52		石油精製	19.93	*
マクダーモット	78.16	**	シェル	43.66	**	IJPC	31.44	**	カロルコ	15.44	
(日商岩井)	77.97	**	(合弁)	43.56	*	(閣僚会議)	31.25		タシケント	14.64	
サハリン沖	77.38	*	熊谷直彦	43.54	*	住友商事	30.37	**	関連機器	14.32	
鉱区	77.14	*	保険対象	40.67		(合弁会社)	29.48	*	協力協定	11.52	
国営石油会社	74.71	*	西シベリア	40.44	*	最高会議	29.43		(上海)	11.45	
SODECO	74.16	**	ワーナー	39.96		事業許可	29.27		伊藤忠インター		
サテライトジャパン	57.97		海外投資保険	38.88	*	衛星通信	28.68		ナショナル	11.38	*
サハリン沖開発	57.74	**	請求額	35.74		TW	28.15		CI	10.19	
JCSAT	55.90		チュメニ	35.59	*	サラワク州	26.58	*	(現在著作権交渉)	4.94	
(インドネシア)	54.40	**	(東南アジア諸国			鈴木精	26.30		(本文)	4.94	
サラワク	51.12	**	連合)	35.01		(アジア)	25.83				
ICDC	50.28	**	LNG	34.75	**	(APEC)	25.75				
SAJAC	48.70	**	(ASEAN)	34.66		(タイ)	25.15				

(注1) (a), (b)とも，表全体がタームグループで，太枠内が代表タームリストである．また，(b)には親トピックの代表タームが追加されている．( ) 付きのタームが親トピックの代表タームである．
(注2) SMI：タームグループ内累計相互情報量．
(注3) CAT：タームのカテゴリー．**はコアターム，*は関係ターム，無印は無関係ターム．

q'）を変えたり，代表タームリストを修正してズームの方向を微調整したりすることができる．ズームインを繰り返し，興味（トピック）が十分に絞られた段階で文書検索を実行する．

このシステムの狙いは，検索要求を記述するのが難しい，一度で満足できる検索結果を得ることは難しい，といったサーチ型検索の問題点を解消することである．プロトタイプの試用を通じて，次のような効果が確認された．第1に，漠然とした興味からブラウジングしていくうちに，興味が明確になり，また新しい興味がわいてくる．第2に，自分自身では思いつかないが興味を的確に表現するタームを見つけることができる．知識が乏しい専門外の分野の情報を検索するとき，特に有効である．第3に，文書中で実際に使用されているタームを用いて検索するので，確実に検索結果が得られる．

本システムは，さまざまな分野に適用することができるが，どのような内容の文書が含まれているかが不明確な文書データベースに特に有効である．各種の報告書や提案書，メモ，議事録などを含む企業の文書データベースに適用することにより，部門間にまたがる情報の共有・利用を促進する．コールセンターへの問合せやクレーム情報も有力な適用分野である．この場合，本システムは，単なる検索ツールというより情報の分析・可視化ツールとして位置づけられる．

7. 今後の課題

本論文はトピック階層に着目したマクロな構造化を

表 3 タームグループ/代表タームリストの質の評価
Table 3 Qualitative evaluation of term groups and representative-term lists.

(a) 最上位トピックのタームグループ

タームグループ内累計相互情報量の順位	コアターム	関係ターム	無関係ターム	計
1〜10	144 (72.0%)	40 (20.0%)	16 (8.0%)	200
1〜20	268 (67.0%)	86 (21.5%)	46 (11.5%)	400
1〜30	358 (59.7%)	138 (23.0%)	104 (17.3%)	600
1〜40	402 (50.3%)	193 (24.1%)	205 (25.6%)	800
1〜60	439 (36.6%)	272 (22.7%)	489 (40.8%)	1200
1〜80	443 (28.8%)	299 (19.4%)	798 (51.8%)	1540

(注) (p=2000, q=100, r=20)のケースから選んだ 20 個のタームグループに対する評価結果を集計した．太枠内が代表タームリストに対応する．

(b) サブトピックのタームグループ

タームグループ内累計相互情報量の順位	コアターム	関係ターム	無関係ターム	計
1〜10	126 (63.0%)	46 (23.0%)	28 (14.0%)	200
1〜20	238 (59.5%)	89 (22.3%)	73 (18.3%)	400
1〜30	299 (49.8%)	174 (29.0%)	127 (21.2%)	600
1〜40	327 (40.9%)	259 (32.4%)	214 (26.8%)	800
1〜60	371 (30.9%)	394 (32.8%)	435 (36.3%)	1200
1〜80	376 (25.3%)	439 (29.5%)	672 (45.2%)	1487

(注) (p'=400, q'=4, r'=20)のケースから選んだ 20 個のタームグループに対する評価結果を集計した．太枠内が代表タームリストに対応する．

扱った．今後の課題として，個々のトピックの内部，すなわち代表タームリストの内部を構造化することがあげられる．以下のような機能によって，より洗練されたシソーラスを得ることができるであろう．

(1) 同義語の抽出

重要な同義語はすでに代表タームリスト内に得られている．図 6 (a) の (i) 中の “東南アジア諸国連合” と “ASEAN”，同じく図 6 (a) の (ii) 中の “金利” と “利率”，“通貨供給量” と “マネーサプライ” はその例である．しかし，同義語であることが陽に認識されているわけではない．代表タームリスト中の同義語ペアを抽出し，同義語として表示することが望ましい．

(2) 意味カテゴリによる分類

代表タームリストには，さまざまな意味カテゴリのタームが含まれている．意味カテゴリごとにタームを分類して表示することにより，代表タームリストをより理解しやすくする．図 6 (a) の (i) を例にとると，{マレーシア，シンガポール，$\cdots$}，{伊藤忠商事，三井物産，$\cdots$} などのサブグループを抽出することが考えられる．

(3) タームの包含関係の表示

代表タームリストには，包含関係を持つタームのペアが多数含まれている．図 6 (a) の (i) 中の “合弁” と “合弁会社”，同じく図 6 (a) の (ii) 中の “金利” と “市場金利” や “金利自由化” はその例である．これらの関係を表示し，トピック内部の構造を理解しやすくする．

8. 関連研究との比較

8.1 タームクラスタリング

タームのクラスタリングにより関連タームの集合を抽出するという考え方は古くから知られている[11]．しかし，大規模コーパスの可視化に有効であることを実証した研究は報告されていない．本論文では，計算量と質の両面で大規模コーパスに適用し得る方法を新たに開発した．

一方，同義語抽出を目的としたタームクラスタリングの研究が多数報告されている[14)〜19)]．本論文のクラスタリングがタームの一次相関（関連度）を利用するのに対し，同義語抽出のためのクラスタリングはタームの二次相関を利用する．すなわち，共起タームの集合で表現される出現文脈の類似度に基づいてタームをクラスタリングする．7 章で述べたトピック内部の構造化には，これらの研究が参考になる．なお，同義語抽出の研究のなかで，オーバラップしたクラスタを生成するアルゴリズムが提案されているが，10 語あまりをクラスタリングする実験にとどまっている[20]．

8.2 重要タームの抽出

長尾らは，χ^2 検定を利用して文書カテゴリを特徴づけるタームを抽出する方法を提案した[21]．カテゴリに分類された文書集合を与える必要があるので，本論文の代表ターム抽出にこの方法を適用することはできない．tf-idf など，文書の特徴タームを抽出するための指標も，文書を文書集合に置き換えることによって，コーパスからの重要ターム抽出に利用することができる[8]．しかし，その場合も，分類された文書集合が必要である．本論文の方法は，分類された文書集合を必要としない点でこれらの従来技術と異なっている．

Hisamitsu らは，タームが出現する文書集合における単語分布と全文書集合における単語分布との距離をタームの代表性（representativeness）の指標とすることを提案した[22]．タームが出現する文書の集合は自動的に求められるので，前述の問題点は解消されている．しかし，特定のトピックやトピックに関する文書の集合とは無関係に計算される指標であるので，トピックの代表ターム抽出には適していない．本論文の方法は，Hisamitsu らの方法と比較したとき，トピックごとにグループ化された重要タームリストを抽出することが特徴である．

8.3 文書データベースのブラウジング

文書データベースのブラウジングの素直な実現方法として，文書クラスタリングを用いる方法がある．Scatter/gather[23]，WEBSOM[24] など，いくつかのシステムがすでに提案されている．Scatter/gather は階層的クラスタリングアルゴリズムを用い，WEBSOM は自己組織化マップアルゴリズムを用いるという違いはあるが，両システムとも文書クラスタの階層

をトップダウンにたどる機能を提供する．文書クラスタはそれを特徴づけるタームのリストとして表示されるので，本研究のシステムと見かけは非常に似ている．

Scatter/gather や WEBSOM などの従来システムと本研究のシステムとの違いは内部の処理にある．従来システムが文書をクラスタリングするのに対し，本研究のシステムはタームをクラスタリングする．本研究のシステムは，この違いに起因する次の長所を持つ．

(1) 大規模文書データベースに対する計算量

クラスタリング処理の計算量は，クラスタリング対象の要素数が増加すると，急激に増大する．文書数は増加するいっぽうであるが，1 つの分野で使用されるタームの数は飽和する．したがって，文書データベースが一定の規模を超えると，文書クラスタリングよりタームクラスタリングが有利になる．

文書クラスタリングでも計算量を減らす工夫がされている．ここでは，本研究のシステムの計算量と Scatter/gather の工夫されたアルゴリズムの計算量を比較する．前者は $k \cdot p^2$（k：比例係数，p：クラスタリング対象タターム数），後者は $K \cdot c \cdot d$（K：比例係数，c：クラスタ数，d：クラスタリング対象文書数）であるが，比例係数の大きさに注意して比較する必要がある．Scatter/gather のアルゴリズムでは，文書間の類似度計算を事前に行わず（事前に計算すると，総文書数 D の 2 乗に比例する計算量になる），クラスタリング処理の中で文書間あるいは文書とクラスタ間の類似度を計算する．いっぽう，本研究のシステムでは，ターム間の関連度を事前に計算し（計算量は，1 つのタームと閾値以上の頻度で共起するタームの数 m と総ターム数 P の積に比例する），クラスタリング処理の中では計算済みの関連度を参照する．したがって，$k \ll K$ である．d は $D, D/c, D/c^2, \cdots$ と段階的に小さくなるが，D が大きい場合，初期の段階では $k \cdot p^2 < K \cdot c \cdot d$ である．

(2) 雑多な文書の集合への適用性

文書のサイズが一様でない場合，あるいは複数のトピックを記述した文書が含まれる場合，文書クラスタリングの精度は低下する．これに対し，ウィンドウ共起に基づくタームのクラスタリングでは，文書という単位は意味を持たないので，精度の低下は小さい．

(3) ユーザインタラクションの取り入れやすさ

6.2 節で述べたシステムでは，ズーム前にユーザが代表タームリストを修正することができる．クラスタリングの精度を考えると，実用上重要な機能である．ところが，文書クラスタリングによるシステムでは，同様な機能を実現することが困難である．タームリストを修正しても，システム内部の処理対象である文書クラスタを修正したことにはならないからである．

9. むすび

関連シソーラスからトピック階層とトピックの代表タームリストを抽出する方法を開発した．最初に，比較的頻度の高いタームをクラスタリングすることにより，最上位トピックを抽出する．次に，トピックの代表タームと関連が強いタームをクラスタリングすることにより，トピックをサブトピックに分割する．このような段階的な方式により，大規模な関連シソーラスの構造化を可能にした．また，トピックが認知しやすい代表タームリストを抽出するため，タームのトピック代表性を表す指標としてタームグループ内累計相互情報量を考案した．

新聞記事コーパスを用いた評価実験では，最上位の代表タームリストの 89%，下位の代表タームリストの 76%が，新しいトピックを示唆するという意味で有効であった．また，代表タームの 3 分の 2 近くがトピックを強く示唆するコアタームであった．これらの結果から，コーパスの情報空間を可視化する手段として本方法が有効であるとの結論を得た．プロトタイプの試用を通じて実際的な効果も確認した．シソーラス作成コストを大幅に低減することに加え，文書データベースブラウジングツールとして，ユーザの興味を喚起し，検索タームを思いつかせる効果がある．

今後の課題として，代表タームリスト内部の構造化があげられる．応用面では，文書分類への展開も興味深い．本方法は，分類体系と分類知識を同時に学習する方法とみることができる．

謝辞 本研究は，一部，通商産業省（現，経済産業省）/情報処理振興事業協会（IPA）/日本情報処理開発協会（JIPDEC）の「次世代電子図書館システム研究開発事業」の支援を受けた．(株）毎日新聞社からは，CD-ROM 毎日新聞データ（91〜95 の各年度版）の使用許可をいただいた．また，評価実験において（株）日立システムアンドサービスの中原章義氏からプログラミングの支援をいただいた．以上の関係者に感謝いたします．

参考文献

1) Jing, Y. and Croft, W.B.: An association thesaurus for information retrieval, *Proc. RIAO '94, Conference on Intelligent Text and Image Handling*, pp.146–160 (1994).
2) Schuetze, H. and Pedersen, J.O.: A cooccur-

rence-based thesaurus and two applications to information retrieval, *Proc. RIAO '94, Conference on Intelligent Text and Image Handling*, pp.266–274 (1994).
3) Mandala, R., Tokunaga, T. and Tanaka, H.: Combining multiple evidence from different types of thesaurus for query expansion, *Proc. 22nd Annual International ACM SIGIR Conference on Research & Development of Information Retrieval*, pp.191–197 (1999).
4) Church, K.W. and Hanks, P.: Word association norms, mutual information, and lexicography, *Computational Linguistics*, Vol.16, No.1, pp.22–29 (1990).
5) El-Hamdouchi, A. and Willett, P.: Comparison of hierarchical agglomerative clustering methods for document retrieval, *The Computer Journal*, Vol.32, No.3, pp.220–227 (1989).
6) Jardine, N. and Sibson, R.: The construction of hierarchic and non-hierarchic classifications, *The Computer Journal*, Vol.11, No.2, pp.177–184 (1968).
7) Justeson, J.S. and Katz, S.M.: Technical terminology: some linguistic properties and an algorithm for identification in text, *Natural Language Engineering*, Vol.1, No.1, pp.9–27 (1995).
8) Kageura, K. and Umino, B.: Methods of automatic term recognition: a review, *Terminology*, Vol.3, No.2, pp.259–289 (1996).
9) Dunning, T.: Accurate methods for the statistics of surprise and coincidence, *Computational Linguistics*, Vol.19, No.1, pp.61–74 (1993).
10) Salton, G. and McGill, M.J.: *Introduction to Modern Information Retrieval*, p.63, McGraw-Hill (1983).
11) *ibid*, p.80.
12) 梶 博行，森本康嗣，相薗敏子，山崎紀之，飯田恵子，内田安彦：コーパス対応の関連シソーラスナビゲーション，情報処理学会データベースシステム研究会/情報学基礎研究会研究報告，DBS-118-13/FI-54-13 (1999).
13) Kaji, H., Morimoto, Y., Aizono, T. and Yamasaki, N.: Corpus-dependent association thesauri for information retrieval, *Proc. 18th International Conference on Computational Linguistics*, pp.404–410 (2000).
14) Hindle, D.: Noun classification from predicate-argument structures, *Proc. 28th Annual Meeting of the Association for Computational Linguistics*, pp.268–275 (1990).
15) Pereira, F., Tishby, N. and Lee, L.: Distributional clustering of English words, *Proc. 30th Annual Meeting of the Association for Computational Linguistics*, pp.183–190 (1992).
16) Grefenstette, G.: Use of syntactic context to produce term association lists for text retrieval, *Proc. 15th Annual International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp.89–97 (1992).
17) Ushioda, A.: Hierarchical clustering of words and applications to NLP tasks, *Proc. 4th Workshop on Very Large Corpora*, pp.28–41 (1996).
18) Li, H. and Abe, N.: Word clustering and disambiguation based on co-occurrence data, *Proc. 17th International Conference on Computational Linguistics*, pp.749–755 (1998).
19) Lin, D.: Automatic retrieval and clustering of similar words, *Proc. 17th International Conference on Computational Linguistics*, pp.762–768 (1998).
20) Fukumoto, F. and Tsujii, J.: Automatic recognition of verbal polysemy, *Proc. 15th International Conference on Computational Linguistics*, pp.762–768 (1994).
21) 長尾 真，水谷幹男，池田浩之：日本語文献における専門用語の自動抽出，情報処理，Vol.17, No.2, pp.110–117 (1976).
22) Hisamitsu, T., Niwa, Y. and Tsujii, J.: A method of measuring term representativeness — Baseline method using co-occurrence distribution, *Proc. 18th International Conference on Computational Linguistics*, pp.320–326 (2000).
23) Cutting, D.R., Karger, D.R., Pedersen, J.O. and Tukey, J.W.: Scatter/gather: A cluster-based approach to browsing large document collections, *Proc. 15th Annual International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp.318–329 (1992).
24) Lagus, K., Honkela, T., Kaski, S. and Kohonen, T.: Self-organizing maps of document collections: a new approach to interactive exploration, *Proc. 2nd International Conference on Knowledge Discovery and Data Mining*, pp.238–243 (1996).

(平成 13 年 9 月 10 日受付)
(平成 14 年 12 月 3 日採録)

梶　博行（正会員）

1973 年京都大学工学部電気工学第二学科卒業．1975 年同大学院修士課程修了．同年（株）日立製作所入社，システム開発研究所を経て，現在，同社中央研究所勤務．自然言語処理，機械翻訳，情報検索などの研究開発に従事．電子情報通信学会，人工知能学会，言語処理学会，ACM，Association for Computational Linguistics 各会員．

森本　康嗣（正会員）

1986 年名古屋大学工学部電気工学第二学科卒業．1988 年同大学院工学研究科電気工学専攻博士前期課程修了．同年（株）日立製作所入社，システム開発研究所を経て，現在，同社中央研究所勤務．自然言語処理，機械翻訳，情報検索などの研究開発に従事．言語処理学会会員．

相薗　敏子（正会員）

1989 年聖心女子大学文学部教育学科心理学専攻卒業．1992 年東京工業大学大学院総合理工学研究科システム科学専攻修士課程修了．同年（株）日立製作所入社，システム開発研究所を経て，現在，同社中央研究所勤務．自然言語処理，情報検索などの研究開発に従事．人工知能学会会員．