データベースシステム 118－13
情 報 学 基 礎 54－13
（1999．5．17）

コーパス対応の関連シソーラスナビゲーション

梶 博行†, 森本康嗣†, 相薗敏子†, 山崎紀之††, 飯田恵子††, 内田安彦††

†日立製作所中央研究所 ††日立製作所ソフトウエア事業部

あらまし: 電子化されたテキスト情報の増加とともに情報アクセス技術の重要性が高まっている．本稿では，大規模テキストコーパスの探索を支援する，インタラクティブなテキストマイニングシステムを提案する．提案システムは，コーパスから関連シソーラスを自動生成し，コーパスに対応したシソーラスをナビゲーションできるようにする．関連タームのクラスタリング，シソーラスオーバビューの生成，オーバビューから詳細へのズームインという特徴機能によって，漠然とした情報要求しかもたないユーザや専門外のドメインの情報を求めているユーザでも，適切な情報を効率よく獲得することができる．プロトタイプの開発と新聞記事コーパスを用いた実験を通じて，提案システムの有効性を実証した．

Navigation through a Corpus-dependent Association Thesaurus

Hiroyuki Kaji†, Yasutsugu Morimoto†, Toshiko Aizono†, Noriyuki Yamasaki††, Keiko Iida††, Yasuhiko Uchida††

{kaji, morimoto, aizono}@crl.hitachi.co.jp, {yamasa_n, iida_ke, uchida_y}@soft.hitachi.co.jp

†Central Research Laboratory, Hitachi, Ltd.
1-280 Higashi-koigakubo, Kokubunji-shi
Tokyo 185-8601, Japan

††Software Division, Hitachi, Ltd.
549-6 Shinano-cho, Totsuka-ku, Yokohama-shi,
Kanagawa 244-0801, Japan

ABSTRACT: With the growing amount of textual information available in electronic form, information access technologies have become extremely important. This paper proposes an approach to interactive text mining that facilitates exploration through a large corpus of texts. The proposed system automatically generates an association thesaurus from a corpus, and enables users to navigate through this corpus-dependent thesaurus. Its novel functions, including the clustering of related terms, the generation of an overview of the thesaurus, and the zooming-in from the overview to the details of a specific part, allow users to get information efficiently even when their information needs are vague or they seek information in unfamiliar domains. The effectiveness of the system has been demonstrated through prototyping and an experiment with a newspaper article corpus.

1. はじめに

電子化されたテキスト情報が増加するとともに，テキストマイニング技術へのニーズが高まっている．全自動のテキストマイニングは，情報抽出[1)]や，究極的には自然言語理解の技術が必要である．しかし，これらの技術はいまだ揺籃期にある．少なくとも実用の観点からは，インタラクティブなテキストマイニング，すなわちテキストマイニング支援ツールへのアプローチが，より重要である．

インタラクティブなテキストマイニングは，情報検索，テキスト分類などの既存技術により，ある程度可能である．しかし，それらの技術の限界は明らかであり，テキストコーパスの効果的な探索を可能にする新しいアイデアが必要である．現在の情報検索システムは，情報要求を陽に表現することをユーザに要求するが，欲しい情報が何であるかユーザ自身，正確にはわかっていないことも多いという問題がある．有力な代替アプローチとして，コーパスにどんな情報が含まれているかをユーザに提示する，ブラウジング

型のシステムが考えられる.

本稿では, テキストコーパスから関連シソーラスを生成し, コーパスに対応したシソーラスをナビゲーションすることのできるシステムを提案する. 関連シソーラスは, これまで, 検索精度を向上させるため, テキスト検索システムの内部で用いられてきた[2)3)]. 本研究では, コーパスの情報空間を可視化するツールとして関連シソーラスを利用する.

2. システムのコンセプト

2.1 システム概観

提案システムは, **図1**に示すように, シソーラスジェネレータ, シソーラスナビゲータ, テキスト検索エンジンから構成される. シソーラスジェネレータは, コーパスに含まれる共起情報に基づいてシソーラスを生成するバッチ処理プログラムである. 生成されるシソーラスは, タームの連想関係のネットワークであり, 関連シソーラスと呼ばれる. 下位概念語や同義語などを示す概念階層ではない. コーパスの内容を可視化するツールとしては, 概念階層シソーラスより関連シソーラスのほうが有用である. なぜなら, 関連タームの集合がコーパス中に含まれるトピックあるいはサブドメインに対応するからである.

シソーラスナビゲータは, 関連シソーラスのナビゲーションを助けるインタラクティブなソフトウエアツールである. シソーラスのナビゲーションを通じて, ユーザの情報要求が明確になり, さらには新しい興味を呼び起こすことが期待される. 興味が十分に特定化された段階で, スクリーン上に表示されている関連タームがテキスト検索エンジンにわたされ, 文書検索が実行される.

2.2 シソーラスナビゲータに対する要求

図1に示したシステムコンポーネントのうち, シソーラスナビゲータに特に焦点をあてる. 提案システムの有効性は, シソーラスナビゲータが提供する機能とユーザインタフェースによって決まるからである. 従来のシソーラスブラウザ, 例えばイリノイ大学ディジタルライブラリで開発されたシソーラスブラウザ[4)]は, ユーザが指定したタームの関連タームを検索, 表示することにより, タームの間を遷移することを可能にする. しかし, あまりにも素朴な機能であり, テキストマイニングの要求に応えることはできない. テキストマイニングツールへの要求と従来のシソーラスブラウザの問題点は, 以下のようにまとめられる.

(1) ユーザは, コーパスにどんなトピックあるいはサブドメインが含まれているかを知りたい. しかし, タームの間の関連をたどるだけで, トピックやサブドメインを理解することは困難である. トピックやサブドメインはタームの集合によって暗示されるので, シソーラス内部の中間構造である関連タームのクラスタを提示することが効果的であろう.

(2) ユーザは, コーパスに対応するドメイン全体の情報構造をつかみたい. しかし, 多数のタームからなる非階層型のシソーラスの全体構造を示すことは難しく, そのような機能をもつシソーラスブラウザは, これまで開発されていない. 従来のシソーラスブラウザは, ユーザが指定したタームのすぐ近くの構造を見せるだけである.

(3) 漠然とした情報要求をもつユーザが, 自分の情報要求を適切に表現するタームを知りたいということも多い. あるいは, 専門外でよく知らないドメインの用語について知りたいという場合もある. このような場合, タームを探しているユーザにタームを入力させるというのは, 実際的でない.

2.3 提案するシソーラスナビゲータと情報探索

前節で述べた要求を満たすため, 次の機能をもつシソーラスナビゲータを提案する.

- 関連タームのクラスタリング
- シソーラスオーバビューの生成
- 関心のあるトピック, サブドメインのズームイン

これらの機能を利用したシソーラスナビゲーションの概念的なイメージを**図2**に示す. 典型的な情報探索セッションは次のようになる.

最初に, システムがコーパス対応シソーラスのオーバビューを表示し, コーパスの情報空間にユーザが容易に入れるようにする. オーバビューは, コーパスの一種の要約である. ドメインの一般的なタームのクラスタから構成され, コーパスがどんなトピックやサブドメイン

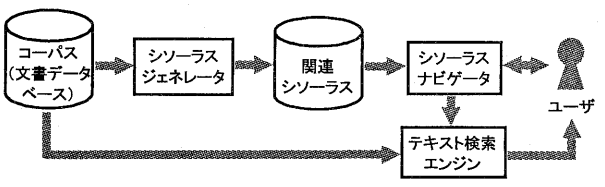


図1 提案システムの概観

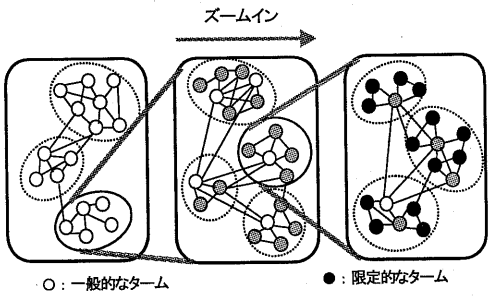


図2 シソーラスナビゲーションのイメージ

を含んでいるかについての理解を助ける.

シソーラスオーバビューをみたユーザは，興味のある一つあるいはいくつかのタームクラスタを選択し，ズームインしてみることができる．すなわち，ユーザが選択したクラスタに関連のあるタームを検索し，もともとクラスタを構成していたタームと併せて，クラスタリングすることにより，サブドメインの詳細なビューが与えられる．十分に特定化されたトピックを表すタームクラスタに到達するまで，ユーザはズームイン操作を繰り返すことができる．なお，非階層型シソーラスに対するズームイン機能を実現することは，一つの技術的な挑戦である．

上の説明では，ナビゲーションはシソーラスオーバビューの表示とともに開始されるとした．しかし，いつもそれがよいというわけではない．ユーザの興味がすでに明確である場合，その興味に関連のあるタームを入力することによってセッションを開始できることが望ましい．これにより，ユーザはより早く目標に到達することができるであろう．

ユーザが介入して，クラスタを修正することも必要である．クラスタリング結果の質が十分に高ければ，ユーザの操作は，シソーラスナビゲータが提示するクラスタの中から選択することだけでよい．しかし，クラスタの中には，はっきりした主題が読み取れないものや，あまり関係のない不適切なタームが混在するものもあるであろう．従って，ズームインする前に，ユーザがクラスタを修正できることが必要である．クラスタの修正は，ユーザの負担のように思われるが，ズームインしたときに意味のあるクラスタが得られる可能性が高くなり，全体としてシソーラスナビゲーションの効率を向上させることができる．

3. コーパスからのシソーラス自動生成

3.1 概 要

コーパスからのシソーラス生成は，図3のように行われる．まず，コーパスからタームを抽出して，出現頻度とともにファイルに格納する．また，共起するタームの組を抽出して，共起頻度とともにファイルに格納する．そして，タームの組ごとに統計的な相関を計算する．予め定めた閾値をこえる相関値をもつタームの組を，相関値とともにシソーラスファイルに格納する．生成したシソーラスは，ユーザに表示するものであり，テキスト検索システムの内部で使用されるものに比べ，高精度のターム抽出が要求される．

3.2 ターム抽出

シソーラスは，ドメインの重要な概念を表すタームから構成されるのが理想である．重要な概念を表すタームの多くは，コーパス中に高頻度で出現する名詞である．しかし，高頻度の名詞が重要な概念を表すとは限らない．そこで，出現頻度が予め定めた値をこえる名詞を選択するが，同時に，ストップワードリストを用いて共通的な語を取り除く．

ドメイン固有の重要概念は，しばしば複合名詞で表現される．従って，出現頻度が予め定めた値をこえる複合名詞も抽出する．複合名詞は，品詞列のパターンを用いたパターンマッチングによって抽出する．品詞列パターンは，当然，言語によって異なる．日本語の複合名詞を抽出するため，次のパターンを用いる．

COMP_NOUN := {{PREFIX} NOUN + SUFFIX} {PREFIX} NOUN + .

しかし，パターンにマッチする単語列がいつも複合名詞であるとは限らず，単なる名詞句であるかもしれない．そこで，複合名詞の先頭要素と末尾要素のストップワードリストを用意することによって，ある種の名詞句を取り除く．例えば，‘上記’を先頭要素のストップワードとすることにより，‘上記システム’というような名詞句を除外する．同様に，‘全体’を末尾要素

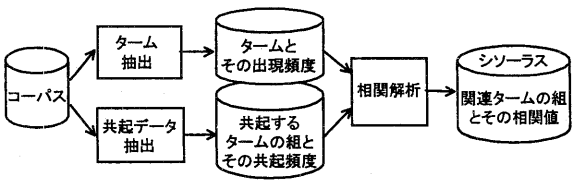


図3 コーパスからのシソーラス自動生成

のストップワードとすることにより，‘システム全体’というような名詞句を除外する．

3.3 共起データ抽出

シソーラスの目的から，共通のトピックに関連するタームの組を抽出することが必要である．そこで，共起の定義として，文書共起や構文共起ではなく，中程度のサイズのウィンドウによるウィンドウ共起を採用する．すなわち，一定数のタームを収容するウィンドウをテキストに沿って移動させながら，各位置においてウィンドウに含まれる任意のタームの組を抽出する．(プロトタイプのウィンドウサイズは，機能語を除いて25タームとした．)

ここで，複合名詞を構成する名詞の組は，当然，ウィンドウ内にはいるが，それらは共起データとみなさない．そのような名詞の組を共起として扱うと，細かすぎる概念あるいは概念の断片を表す名詞がシソーラスに多数含まれることになる．ユーザは，結合可能な名詞の組をさがし出して，頭の中で結合することによって，意味のある概念を認識しなければならない．

3.4 タームの相関解析

タームの相関値として，次式で定義される相互情報量を用いる．

$$A(t_i,t_j)=\log_2\frac{g(t_i,t_j)\Big/\sum_{i,j}g(t_i,t_j)}{\left\{f(t_i)\Big/\sum_i f(t_i)\right\}\cdot\left\{f(t_j)\Big/\sum_j f(t_j)\right\}}.$$

ここに，$f(t_i)$はターム t_i の出現頻度，$g(t_i,t_j)$はターム t_i と t_j の共起頻度である．相互情報量は，低頻度タームには過大な値になるという問題がある[5]．そこで，二つのタームの間に関連があるかどうかを対数尤度比によって検定し，検定に合格したタームの組のみについて相互情報量を相関値として用いる．また，相互情報量の値が予め定めた閾値以下のタームの組も除外する．各タームについて，そのタームとの相関値が大きい順に，予め定めた個数の関連タームを選択して，シソーラスファイルに格納する．

4. オーバビュー表示／ズームイン機能をもつシソーラスナビゲータ

4.1 概 要

シソーラスナビゲータは，図4に示すように主要ターム抽出，ターム集合拡大，タームクラスタリングの各モジュールから構成される．主要ターム抽出モジュールとタームクラスタリングモジュールをこの順に実行することにより，シソーラスオーバビューが生成される．ターム集合拡大モジュールとタームクラスタリングモジュールをこの順に実行することにより，ズームイン機能が実現される．

4.2 オーバビューのための主要ターム抽出

シソーラスオーバビューは，コーパスのドメインの一般的な(ジェネリックな)タームから構成されるべきである．しかし，一般的なタームの明確な判定基準はないので，代わりに，コーパス中の主要タームを集める．主要タームとは，コーパス中のできるだけ多くの文書を特徴づけるタームである．具体的には，以下のようにして主要タームを選定する．ここで，M はオーバビューに含めるタームの総数である．(プロトタイプでは，Mを300とした．)

i) 各文書に対する特徴ターム集合の決定

文書 d_i におけるターム t_j の重み $w_{i,j}$ を tf-idf (term frequency - inverse document frequency) 法により計算する．すなわち，

$w_{i,j}=tf_{i,j}\cdot\log_2(N/n_j)$.

ここに，$tf_{i,j}$は文書 d_i 中のターム t_j の出現頻度，n_jはターム t_j が出現する文書数，N は文書の総数である．

次に，各文書 d_iごとに，重み $w_{i,j}$ の大きい順に $m(i)$ 個のタームを特徴タームとして選択する．文書 d_i の特徴ターム数 $m(i)$は，文書 d_i 中の異なりターム数の20%，ただし5以上，50以下とする．

ii) コーパスの主要タームの選定

文書の特徴ターム集合に含まれ

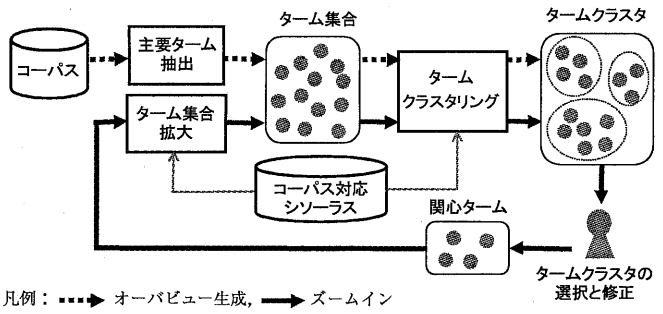


図4 シソーラスナビゲータの構成

る頻度の大きい順に，M個のタームを選定する．

4.3 ズームインのためのターム集合拡大

ズームインでは，ユーザが選択したサブドメインの各部分が詳細化されるべきである．そこで，以下に示すように，ユーザの関心タームの各々に関して均等にターム集合を拡大する．

- 入力：ユーザの関心タームの集合
  $T_o = \{t_1, t_2, \cdots, t_m\}$.
- 出力：M個のタームを含む拡大ターム集合 T_e.
  （プロトタイプではMを300とした．）

i) T_eの初期値を T_oとする．

ii) **While** $|T_e| < M$ **for** $i = 1, 2, \cdots$ **do;**
  While $|T_e| < M$ **for** $j = 1, 2, \cdots, m$ **do;**
    t_jとの相関値が i 番目に高いタームをT_eに追加する;
  end;
end;

上記方法で追加されるタームは，関心タームと相関の高いタームであるから，サブドメインの主要タームに近い性格をもつ．従って，**図2**に示したようなズームインを実現することができる．

4.4 タームクラスタリング

主要ターム集合あるいは拡大ターム集合を，ターム間の相関に基づいてクラスタリングする．さまざまな凝集的クラスタリングアルゴリズム[6]によるタームクラスタリングの予備実験をしたところ，グループ平均法が最も良好であった．ただし，クラスタ数が予め定めた個数になることを停止条件として，クラスタをマージする操作を繰り返すと，本来は別々であるべき複数のクラスタがつながって，一つの大きなクラスタが得られるという傾向がみられた．そこで，クラスタのサイズに制限を設ける．

タームクラスタリングのアルゴリズムは次のとおりである．

- 入力：ターム集合$\{t_1, t_2, \cdots, t_M\}$.
- 出力：タームクラスタ.

i) 各タームをそれぞれ一つのタームクラスタとすることにより，クラスタの初期集合をつくる．すなわち，
  $C_1=\{t_1\}, C_2=\{t_2\}, \cdots, C_M=\{t_M\}$.

ii) 任意のクラスタの組 C_i, C_j に対して $|C_i| + |C_j| > \alpha M$であるなら，停止．

αは，入力ターム集合のサイズに対するクラスタのサイズの比の上限値である．（プロトタイプでは，αを0.1とした．）

iii) $|C_i| + |C_j| \leqq \alpha M$であるようなクラスタの組 C_i, C_jのうちで，類似度が最大の組をマージする．

ここで，タームクラスタ C_i, C_jの類似度 $S(C_i, C_j)$をグループ平均法で計算する．すなわち，ターム t, t' の相関値が $A(t,t')$であるとき，

$$S(C_i,C_j) = \underset{t\in C_i, t'\in C_j}{ave} A(t,t').$$

5. プロトタイプと実験

5.1 プロトタイプ

プロトタイプをクライアント／サーバシステムとして開発した．シソーラスナビゲータは，WWW ブラウザ上で利用することができる．スクリーンは，**図5(a)**に示すように，タームクラスタフレーム（左下），タームリストフレーム（右下），ワークフレーム（上）に分割されている．

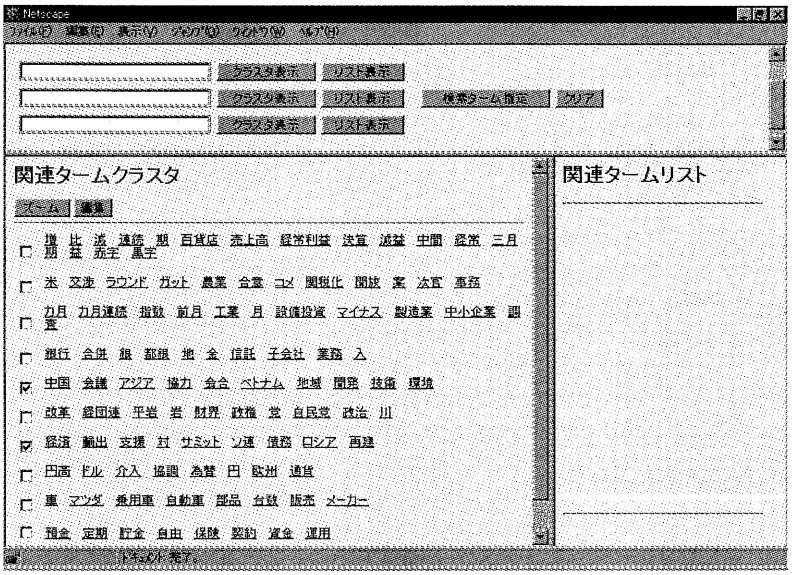
タームクラスタフレームには，シソーラスのオーバビューまたはズームインされたビューが表示され，ユーザは一つまたはいくつかのタームクラスタを選択してズームインすることができる．タームクラスタフレームに付属する別のフレームで，クラスタの修正，すなわちタームの追加，削除を行うことができる．

タームリストフレームは，ユーザが指定したタームの関連タームおよび上位ターム，下位タームをリスト形式で表示する．タームリストフレームは，従来のシソーラスブラウザと同様の機能を提供し，タームクラスタフレームを補完する．ワークフレームは，ユーザが明確な興味をもっているときにタームを入力するエリアであると同時に，テキスト検索エンジンとのインタフェースエリアでもある．

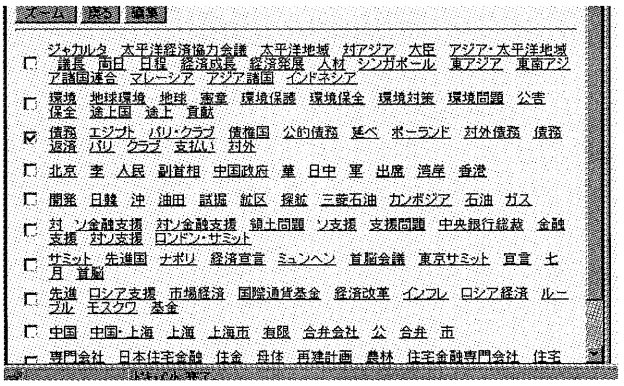
5.2 新聞記事コーパスを用いた実験

毎日新聞経済面の5年分，41,022記事からなる33.5 M バイトのコーパスを用いて実験を行った．シソーラスジェネレータは，28,948ターム（最小出現頻度を10とした），1,570,059組のターム関連からなるシソーラスを生成した．シソーラス生成に要した時間は，HP9000 C200 ワークステーションで4時間であった．

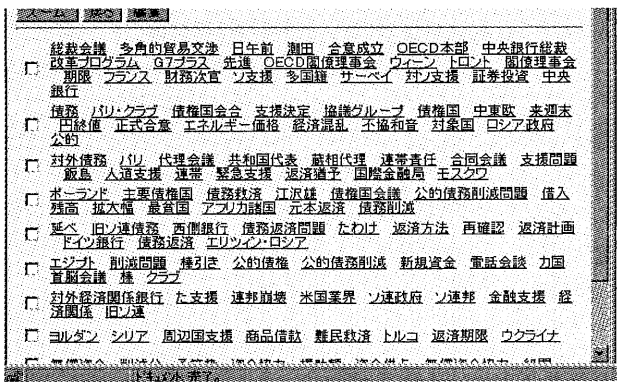
ナビゲーションのセッションにおけるズームイン操作の応答時間は，約10秒であった．タームクラスタリングがユーザに与える情報の有効性を考えると，この応答時間は受け入れることができる．なお，クラスタリングの対象とするタームの数（**4.2** および **4.3** 節に



(a) オーバビュー



(b) ズームイン



(c) さらにズームイン

図5 シソーラスナビゲーションの例

おける M は一定であり，応答時間はシソーラスやコーパスのサイズにほとんど依存しない．

次に，ナビゲーションセッションの例を図5に示す．

このセッションは，図5(a)のシソーラスオーバビューによって始まった．ユーザは，図5(a)の第5，第7の二つのクラスタを選択した．すなわち，ユーザは「開発途上国・地域の開発援助」に興味をもった．

ユーザが選択した二つのクラスタにズームインされ，図5(b)に示すように，より具体的なトピックを示すクラスタが表示された．ユーザは，それぞれのクラスタがどんなトピックを示しているかを容易に理解することができた．図5(b)のクラスタの上から順に，「アジア太平洋地域の経済援助」，「地球環境問題」，「国際債務問題」，「中国関連」，「エネルギー資源開発」等々である．「国際債務問題」に特に興味をもったユーザは，第三のクラスタを選択した．

第三のクラスタにズームインされた図5(c)では，すべてのクラスタがそれぞれ一つのトピックに対応しているわけではないが，全体として「国際債務問題」に関連する多数のタームが表示された．ユーザは，興味のあるタームをスクリーンから選択するだけで，文書を検索することができた．このように，文書検索のフロントエンドとしてのシソーラスナビゲータの有効性を確認した．

ユーザが興味をもつドメインのビューの他の例を図6に示す．これは，技術の進歩とそれが環境に及ぼす影響，環境保護のための技術などに漠然と興味を

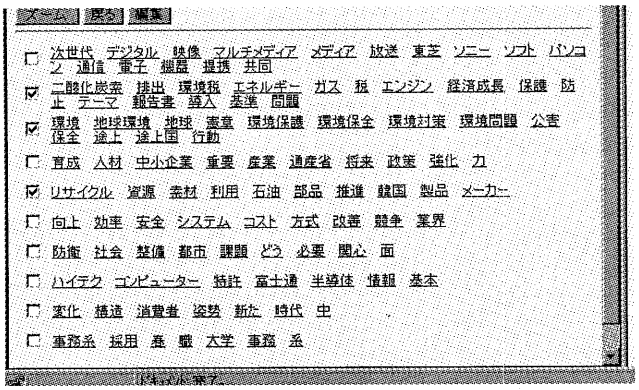


図6 ユーザが興味をもつドメインのビューの例

もったユーザが，二つのターム‘技術’，‘環境’を入力した結果である．このユーザは，興味に適合する第2，第3，第5の三つのクラスタを選択して，ナビゲーションを進めることにより，環境問題のキーワードの理解を深めることができた．

6. 提案方法の有効性

提案方法の有効性は，以下のようにまとめられる．これらは，プロトタイプの利用実験を通じて実証された．

(1) テキストデータへのアクセシビリティ向上

従来のシソーラスブラウザやテキスト検索システムとの大きな違いは，ユーザがタームを全く入力しなくてもよく，スクリーンに表示されるタームのなかから選択するだけよいということである．このため，漠然とした情報要求しかもっていない人やドメインの知識が乏しい人でも，テキストコーパスにアクセスすることが容易になる．

(2) シソーラスナビゲーションの効率向上

従来のシソーラスブラウザでは，ユーザの認知単位が個々のタームである．これに対し，本提案のシソーラスナビゲータでは，認知の単位がタームクラスタすなわちトピックである．タームクラスタを一瞥するだけで，ユーザはトピックを認識することができる．ズームインによって十分に特定化されたトピックに到達するまでは，個々のタームを詳しくみる必要はない．また，個々のタームを詳しくみる段階では，関連タームがまわりに表示されているので，タームの概念を容易に認識することができる．

(3) 知識獲得ツールとしての可能性

シソーラスナビゲータは単なる用語ツールではなく，知的作業の支援ツールとして，その創造性，生産性を向上する可能性をもっている．クラスタリング機能によって，コーパス全体あるいは任意の部分の情報を構造化して提示する．従って，研究者やアナリストは，大規模なコーパスから有用な情報を容易に引き出し，重要な知識を獲得することができる．

7. 関連研究との比較

本稿で提案したシソーラス生成／ナビゲーションと同様に，テキストコーパスの探索を可能にする方法の先行研究がいくつかある．一つは，コーパス全体から部分へと段階的に文書クラスタリングを行う方法で，Xerox PARC の scatter/gather に代表される[7][8]．もう一つは，Kohonen の自己組織化マップアルゴリズムにより，二次元平面（グリッド）上で文書をクラスタリングする方法である[9)-11)]．これらの方法は，基本的にコーパスを文書の集合として扱っている．これに対して，シソーラス生成／ナビゲーションは，コーパスに出現するタームに注目し，コーパスをタームの集合として扱う．この相違の帰結として，シソーラス生成／ナビゲーションは次の(1)(2)の点で優っている．

(1) ユーザの介入と動的な構造化の容易さ

クラスタリングや自己組織化の結果がユーザにとって満足のいくものとは限らないので，ユーザが結果を修正してから詳細レベルに進むということが，実用上，非常に重要である．シソーラス生成／ナビゲーションでは，ターム集合がシステムの処理対象であるので，ユーザによるタームクラスタの修正に応じた動的な構造化が容易に実現できる．これに対し，文書クラスタリングや自己組織化マップでは，文書クラスタのサマリとしてターム集合が表示されるが，内部処理の対象は文書クラスタである．ユーザによるサマリの修正を，文書クラスタに反映させることは容易でない．また，ユーザによる文書クラスタの修正は，ユーザの負担が極端に大きく，現実的でない．

(2) 抽出される情報の詳細さ

テキストマイニングのツールとしては，抽出可能な情報の詳細さのレベルが深いことが重要である．シ

ソーラス生成／ナビゲーションでは，ズーミングによって，かなり限定的な概念を表すタームまで探索していくことができる．文書クラスタリングや自己組織化マップもズーミング機能をもつが，それはより小さな文書クラスタを生成することである．あくまでも文書のクラスタであるので，そのサマリとしてユーザに表示されるのは，一般的なタームにとどまる．

次に，計算機負荷の面で，従来研究と比較する．

(3) バッチ処理の負荷

シソーラス生成／ナビゲーションは，コーパスが小規模のときは有利といえないが，コーパスが大規模になると有利である．その理由は，タームの組ごとに相関を計算するものの，ターム集合全体を構造化することはしないからである．文書クラスタリングや自己組織化マップでは，全ての文書を対象に構造化処理を行うので，文書数の増加とともに計算量が急激に大きくなる．なお，共起データ抽出を含むテキストの処理は，文書クラスタリングや自己組織化マップの場合のタームベクトル計算処理より重く，この点では若干不利である．

(4) オンライン処理の応答時間

文書クラスタリングと比べて，シソーラス生成／ナビゲーションが優位である．なぜなら，ズームイン時にクラスタリングするタームの数が一定であるので，一定の応答時間が確保できる．文書クラスタリングでは，セッションのはじめのうち多数の文書をクラスタリングしなければならない．自己組織化マップでは，構造化済みのマップをブラウジングするだけなので応答は早いが，上記(1)の動的な構造化を犠牲にしたものであり，比較の対象にならない．

8. おわりに

コーパスからのシソーラス自動生成／ナビゲーション機能をもつテキストマイニング支援システムを開発した．シソーラスジェネレータは，複合語を含むタームの抽出と，共起データに基づくタームの相関解析により，関連シソーラスを生成する．シソーラスナビゲータは，関連タームのクラスタリング，シソーラスオーバビューの生成，オーバビューから部分の詳細へのズームインを特徴機能としてもつ．これにより，情報要求が漠然としていたり，専門外のドメインである場合にも，効果的な情報探索を可能にした．

今後の課題として，シソーラスオーバビューのスケーラビリティがある．大きなコーパスに対して生成されるオーバビューのカバー率を評価することが必要である．極めて大きなコーパスに対しては，オーバビューの複数ページへの分割，階層構造の導入を検討することも必要になるであろう．

謝辞：本研究は，一部，通産省／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．また，毎日新聞社から研究目的での使用許諾を受けて，CD-ROM 毎日新聞データ(91,92,93,94,95 の各年度版)を実験に使用した．これらの組織，企業に感謝致します．

参考文献

1) R. Grishman, and B. Sundheim. Message understanding conference - 6: A brief history. Proc. of COLING '96, pp. 466-471.

2) Y. Jing, and W. B. Croft. An association thesaurus for information retrieval. Proc. of RIAO '94, pp. 146-160.

3) H. Schutze, and J. O. Pedersen. A cooccurrence-based thesaurus and two applications to information retrieval. Proc. of RIAO '94, pp. 266-274.

4) E. H. Johnson, and P. A. Cochrane. A hypertextual interface for a searcher's thesaurus. Proc. of Digital Library '95.

5) T. Dunning. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, vol. 19, no. 1, pp.61-74 (1993).

6) A. El-Hamdouchi, and P. Willett. Comparison of hierarchical agglomerative clustering methods for document retrieval. The Computer Journal, vol. 32, no. 3, pp. 220-227 (1989).

7) D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. Proc. of ACM SIGIR '92, pp. 318-329.

8) M. A. Hearst, and J. O. Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. Proc. of ACM SIGIR '96, pp. 76-84.

9) X. Lin, D. Soergel, and G. Marchionini. A self-organizing semantic map for information retrieval. Proc. of ACM SIGIR '91, pp. 262-269.

10) K. Lagus, T. Honkela, S. Kaski, and T. Kohonen. Self-organizing maps of document collections: a new approach to interactive exploration. Proc. of 2nd Int'l Conf. on Knowledge Discovery and Data Mining, pp. 238-243 (1996).

11) T. Kohonen. Self-organization of very large document collections: State of the art. Proc. of 8th Int'l Conf. on Artificial Neural Networks, vol. 1, pp. 65-74 (1998).