社会技術研究論文集 Vol.5, 216-226, Mar. 2008

文書クラスタリングによるトピック抽出および課題発見

TOPIC EXTRACTION AND SOCIAL PROBLEM DETECTION BASED ON DOCUMENT CLUSTERING

橋本　泰一[1]・村上　浩司[2]・乾　孝司[3]・内海　和夫[4]・石川　正道[5]

[1]博士（工学）東京工業大学統合研究院 (E-mail: hashimoto@iri.titech.ac.jp)
[2]博士（工学）東京工業大学統合研究院 (E-mail: murakami@iri.titech.ac.jp)
[3]博士（工学）東京工業大学統合研究院 (E-mail: inui@iri.titech.ac.jp)
[4]修士（工学）東京工業大学統合研究院 (E-mail: utsumi@iri.titech.ac.jp)
[5]博士（工学）東京工業大学統合研究院 (E-mail: ishikawa@iri.titech.ac.jp)

自然言語処理技術を応用し，分析対象となる新聞記事を取得して，記事を自動的に分類し，トピックとなる社会事象を抽出，さらにはトピックを構造化して課題の発見に至る手続きを可能とする手法を開発した．この手法によれば，多数のトピックを含む文書集合に対して階層的クラスタリングを施し，クラスタ間の語彙使用の類似性に基づく構造化を行い，個々のクラスタについてこれを要約するキーワードおよび関係する主体（組織名）を自動抽出することによって内容を効率的かつ経済的に俯瞰できることを示した．本論文では，産業活動に伴う事故・災害に関する社会の課題発見を事例として，これらのテキストマイニング技術を統合した社会変化の定量的分析手法の有効性について検証した．

キーワード：テキストマイニング，文書クラスタリング，文書要約，組織名抽出，課題発見

1. はじめに

社会の複雑化によって社会に不安と不信を引き起す要因が増加している．しかも，一つの問題に多くの主体が関与し，事件が起こるとその波及範囲が想定外の分野にも波及している[1), 2)]．このような社会課題の時系列変化の分析は，新聞記事をもとに行われる．しかし，新聞記事は出来事の発生から関連する具体的な事柄が日々蓄積され，時間の経過とともにこれが多方面に波及して膨大な量となるため，記事を収集し，分類，分析することは大きな労力を要する．また，新聞記事から情報収集するには，情報検索技術を利用して，キーワードにより関連した記事を発見する方法が一般的である．この場合，得られる情報は入力するキーワードによって制約され，課題の多様性を把握することが難しい．キーワードに制約されることなく，多くの新聞記事から関連する情報を多面的に獲得し，これをもとに変化する社会課題の発見に導く俯瞰的な分析手法の確立が望まれている[3)]．

一方，コンピュータやインターネットの普及に伴い，電子化テキストが増加の一歩を辿っており，新聞記事を含め，手軽に大量の文書を入手することが可能になった．そのため，大量の文書から欲しい情報を獲得し，何らかの傾向を発見したいというニーズが高まっている．このニーズを満たすためにテキストマイニングに関する研究・開発が盛んに行われている．テキストマイニングと混同されやすい概念にデータマイニングと情報検索がある．Rajmanは，データマイニングとテキストマイニングの違いについて，「データマイニングは，構造化されたデータからの情報抽出に関する技術である．テキストマイニングは，構造化されていないテキストデータからの情報抽出に関する技術である．」と述べている[4)]．また，Hearstは，テキストマイニングと情報検索の違いについて，「情報検索の目的は，ユーザが必要とする情報を含む文書の発見を助けることである．情報検索で得られる情報は，その文書の著者にとっては既知の情報である．テキストマイニングの目的は，まったく新しい情報を発見することである．」と述べている[5)]．テキストマイニングとは，テキストから新しい情報を発見する技術である．

テキストマイニングでは，計算機による文書中の語彙の出現分布により文書を表現する．2 つの文書が互いに類似した語彙の出現分布を持つ場合には，同一の話題（トピック）を扱っていると考えられ，大量の文書を自動的に類似した話題の文書群に分類したり，文書群の関係を構造化したりすることができる．Uramoto らは，大量の新聞記事に対して，語彙の使用分布の類似した記事を関連づけることにより新しい発見を支援するシステムを提案した[9)]．このシムテムでは，新聞記事を単語ベクトルとして表現し，ベクトルの類似性に基づき記事を関連づけ，グラフ構造として表示する．

本論文で提案するテキストマイニング手法は，大量の記事集合から互いに類似した内容をもつ記事を自動的に処理し，分析者による課題発見の作業を容易とすること

を目的とする．提案手法は，Uramoto らのシステムと同様に記事を単語ベクトルとして表現する．そして，そのベクトルの類似度を基にいくつかの記事集合(クラスタ)に分類し，構造化を行う（階層型文書クラスタリング)．さらに，重要なクラスタを特定するための技術として，クラスタのグループ化，クラスタの指標（密度・中心度）を提案する．

本論文では，産業活動に伴う事故・災害に関する社会の課題発見を事例として，これらのテキストマイニング技術を統合した社会変化の定量的分析手法の有効性について検証した．

2. 提案手法

提案手法は，次に述べる手順によりトピック抽出および課題発見を行う．１）日本経済新聞記事データベースより，日経シソーラスを用いて検索し，分析の対象となる記事文書集合を取得．２）得られた文書集合に階層型クラスタリングを施し，文書を記事群（クラスタ）へ分類，構造化（系統樹）する．３）系統樹の構造に基づくクラスタのグループ化．４）クラスタを要約するキーワードおよび主体（組織名）の抽出．５）個々のクラスタについて中心度および密度を算出し，話題性の強い重要クラスタを判別．５）重要クラスタについて要約キーワードを用いて文書を KWIC（Key Word In Context）検索することにより記事の内容を把握する．６）トピックを関連付ける俯瞰的な課題を分析者の視点を加えて発見する．

2.1. 記事文書集合の取得

分析対象とした文書は，日本経済新聞記事データベースの 1990 年から 2005 年の日本経済新聞本誌とした．新聞記事検索のための用語集「日経シソーラス」[18]の中分類「生産，品質管理」に含まれる語(61 語)の OR 結合と「災害，事件，犯罪」に含まれる語(259 語)の OR 結合の AND 結合により新聞記事の検索を行なった. 検索の結果得られた年単位の記事集合（D）をクラスタリングの対象とした．なお，記事集合からは，会社人事情報および選挙当選結果など事故，事件，災害などと関係に乏しい記事は削除した.

2.2. 文書クラスタリング

複数のトピックをもつ大量な文書集合に対して，文書クラスタリング手法を用いて，いくつかの類似した文書群（クラスタ）に分類する．文書クラスタリングとは，文書に出現する語の分布の類似性に基づいて文書を自動的に類別する手法である．この手続きは，人が記事を分類する方法に類似していることが望ましい．この要件を満たしかつ文書分類で精度が高いことが知られている UPGMA 法（Unweighted Pair-Group Method using arithmetic Average，群平均法）を採用した[10]．UPGMA 法は，階層型クラスタリング法であるため，類別したクラスタの結合もしくは分割関係を系統樹として可視化できる利点もある．分析には，クラスタリングソフトウェア CLUTO[11]を用いた.

クラスタリングにおいて，文書は文書に含まれる語を要素とするベクトルとして表現する. ベクトルの要素は，次の手続きで決定される．１）ある記事の先頭から n 文を抽出する．２）抽出された文に対して形態素解析器 ChaSen [12]を用いて品詞が名詞および名詞の複合からなる語のみを抽出する．３）抽出された語について分析対象文書全体における頻度を算出し，あらかじめ定める閾値 th_f 以上の語をベクトルの要素とする．ベクトルの要素となる語および複合語を語 e_i と呼ぶ．各要素の値は，(1)式に従って決定され，文書は(2)式のようなベクトル $\mathbf{d}$ として表現される.

$$w_d(e_i) = tf_d(e_i) \times |e_i| \times \frac{1}{1+\log(first_d(e_i))} \quad (1)$$

$$\mathbf{d} = (w(e_1), w(e_2), \cdots, w(e_n)) \quad (2)$$

ここで，$tf_d(e_i)$ は文書 d での語 e_i の出現頻度，$|e_i|$ は語 e_i を構成する文字数（文字数が多い語を重視するため)，$first_d(e_i)$ は文書 d で e_i が初めて現れた文の位置（１～n）である．なお，本論文では，n=5 および $th_f = 5$ とした.

文書ベクトル間の類似度は，(3)式で定義されるコサイン類似度関数を用いた．また，一般にクラスタリングアルゴリズムは，あらかじめいくつのクラスタに分類するかを分析者が与えることが必要となる．本論文では，全記事件数／クラスタ数が 20～30 となることを目安とした.

$$sim(d_i, d_j) = \cos(\mathbf{d}_i, \mathbf{d}_j) = \frac{\mathbf{d}_i \cdot \mathbf{d}_j}{|\mathbf{d}_i \| \mathbf{d}_j|} \quad (3)$$

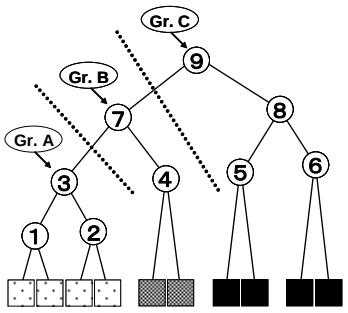
2.3. クラスタのグループ化

文書クラスタリングによって生成される系統樹を見ると，直感的に記事内容の近いクラスタが密に集積しているノードを識別することができる．次に，このような系統樹上において，各ノードの下にある類似した記事を含むクラスタの束なり具合を定量的に表すための指標を導入した．この指標は，あるノードの子孫となる任意の２つのクラスタの組み合わせにおいて，互いにそのノードを共有する場合の数（頻度）によって表すことがきる．この指標を用いることによって，クラスタの類似性によってクラスタを階層的にグループ（群）化することが可能となる．このグループ化操作は，類似したクラスタ（個々の社会事象）の集合として成立するトピックを抽出するために，まとまりのあるクラスタの境界を設定する強力な指標となる.

グループ化の手順は次のとおりである．系統樹上のノ

ードに注目し，このノードの子供を祖先とするクラスタの数の積（頻度）を算出する．例えば，Fig. 1 においてノード7に注目すると，ノード7の子供（ノード3と4）を祖先とするクラスタの数がそれぞれ4と2であるため，ノード7の頻度はその積8となる．この頻度は，系統樹において，注目したノードが記事数の多いクラスタ同士を結合し．より大きなトピックを形成するかどうかを量る指標となる．

次に，頻度の値が多い順にノードを並べる（ランキング）．そして，頻度の比率（当該ノード頻度／上位ノード頻度）を算出し，その比率が最も小さい値をもつ頻度を閾値として以下のノードを無視する．比率が最も小さくなる閾値は，祖先の共有に乏しいクラスタ群を意味し，独立したグループとはみなされないからである．Fig. 1 の例では，10 個のクラスタに対してノードが 9 ある系統樹について，頻度が 4 未満となる 5 番目以下のノードを無視することになる．選定した高頻度ノードに対して，系統樹上で最もクラスタに近い位置にあるノードを Gr.A, 次に頻度の大きいノードを Gr.B などとしてクラスタ群を識別する．この時，既に定義されたグループのクラスタを差し引き，以上の手続きを繰り返すことによって任意の系統樹を構成するクラスタをグループ化することができる．



(2)グループ化のためのスコア

ランク	ノード	頻度	頻度比	グループ
1	⑨	24	—	C
2	⑦	8	0.33	B
3	③	4	0.50	A
3	⑧	4	1.0	(C)
5	①	1	0.25	(A)
5	②	1	1.0	(A)
5	④	1	1.0	(B)
5	⑤	1	1.0	(C)
5	⑥	1	1.0	(C)

Fig. 1　系統樹上でのグラスタのグループ化

2.4.　要約キーワードおよび組織名抽出

クラスタ c の要約として，クラスタ内の記事に含まれる語 e_i ごとに(4)式のスコアを計算し，最大上位 20 語を要約キーワードとした．

$$score_c(e_i) = \sum_{d \in c} w_d(e_i) \qquad (4)$$

要約キーワードを系統樹上に表記することによってクラスタに含まれる記事の内容を容易に把握できる．

組織名抽出は，山田らが提案した教師あり機械学習に基づく固有表現抽出の手法[13]を用いた．山田らの報告によれば，組織名の抽出精度は約 87%である．

2.5.　クラスタの中心度および密度

文書の内容を表すキーワードをもとに，その共有性に基づいて文書を分類する共語分析法が提案された[6) 7) 14)]．共語分析において，文書の相対的な関係を定量的に把握する目的で，中心度（Centrality）および密度（Density）を指標が導入された．我々は，階層型クラスタリングによって生成された系統樹上での祖先 - 子孫の関係に注目して，中心度および密度の新たな定義を提案する．

中心度 $cent(c_i)$ は，系統樹上で，クラスタ c_i の先祖 $anc(c_i)$ と他クラスタ c_j の祖先 $anc(c_j)$ に共通する先祖の内，最も葉に近い位置にある祖先 x の系統樹の根からの深さ $dep(x)$ の平均値であらわす．

$$cent(c_i) = \frac{1}{|C|-1} \sum_{c_j \in C / c_i} \frac{\max\{dep(x) \mid x \in anc(c_i) \cap anc(c_j)\}}{\max\{dep(y) \mid y \in C\}} \qquad (5)$$

C はクラスタの集合，$|C|$はクラスタ数を表す．この時，$cent(c_i)$ は，0 から 1 の値をとり，クラスタ c_i が多くのクラスタと系統樹上で葉に近い位置に共通する祖先がある時 1 に近づく．そのため，中心度が 1 に近づくとき，多くのクラスタと強い関係を有すると考えられる．

密度 $dens(c_i)$ は，クラスタ c_i に含まれる語（要素）の数に対する 2 つ以上の文書に共通に出現する語（要素）の数の割合を表す．

$$dens(c_i) = \frac{\text{クラスタ}c_i\text{内の２つ以上の文書に出現する語の数}}{\text{クラスタ}c_i\text{内の文書に出現する語の数}} \qquad (6)$$

密度は，値が 1 に近い程，クラスタ内で内容が類似した文書が多いことを意味する．

2.6.　課題発見

Swanson は医療分野でのデータベース検索の手法として，文書集合の関係性から仮説生成を可能とする CBD（Complementary But Disjoint）の考え方を提案し[15)]，これに基づいて医療用データベース MEDLINE 専用の検索インタフェース “Arrowsmith”を開発した[16)]．この方法により，ある概念を取り扱う科学文献の集合に対して，内容的関係をもち，かつ，集合内の文献の引用を共有しない異なった文献集合が特定できる場合，両者の文献集合が主張する内容とは互いに異なる第３の観点(研究仮説)を見出すことができる．本研究では，CBD 法の考え方を新聞記事の文書クラスタに対して適用した．

これまで述べてきたように，関連ある検索語によって収集された記事集合のクラスタは，系統樹上でのクラスタ間の類似性に基づいてグループ化(クラスタ群を形成)することができる．このグループ化を密度 - 中心度図上に表すと Fig. 2 のようになる．CBD 法の概念をこれらクラスタ群に適用するには次のように考えればよい．クラスタは, 2.3 節で説明した方法により自動的にクラスタ群に分けられる．この時，同一クラスタ群のトピックは，群中で高い中心度をもつクラスタによって特徴づけられ，分析者は容易に識別することが可能である(後出の Table 1 および Table 2 参照)．これらクラスタ群は互いに重複することなく排他的（Disjoint）に分類されており，系統

樹上で近い位置にある場合には内容的に相補的（Complementary）な関係が成り立つ．異なったグループ間でトピックを互いに比較，評価することによってこれらトピックを関係付ける俯瞰的な課題（仮説）を見出すことが可能となる．

なお，文書集合にはノイズとなる記事が含まれることが避けられないことから，密度 - 中心度図中でどこまでのクラスタ群を分析の対象とするかを判定する必要がある．このため，新たに CBD 因子を定義し，相補的と判断した観点の広がりを定量化する尺度を導入した．

$$CBD\text{因子} = \frac{\Delta cent_{\max}}{disj_{\min}} \tag{7}$$

ここで，分析の対象となるクラスタ群の中心度の平均値の最大差分を $\Delta cent_{\max}$，最小差分を $disj_{\min}$ とした．

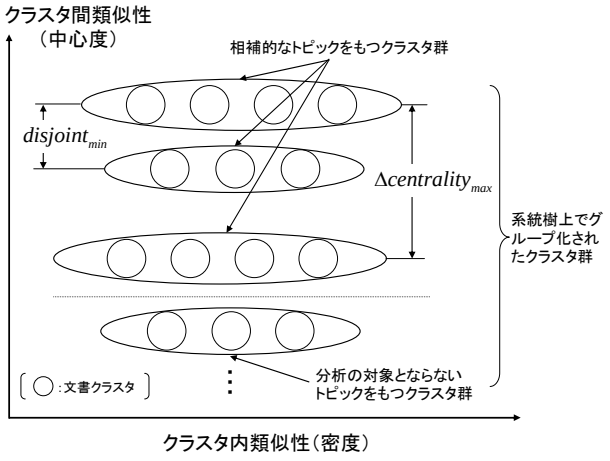


Fig. 2　密度 - 中心度図に基づく CBD 法の概念図

3.　新聞記事からの課題発見

前節で述べた分析手法を用い，新聞記事を対象として実際に「事件・事故・災害」に関する記事の分析を行なった．人事，選挙当選結果など明らかに「事件・事故・災害」に関する記事とは異なる記事を除き，最終的に取得した記事は1990年から2005年の間で4601件であった．Fig. 3 に，この期間の記事数の推移を示す．1995 年に発生した阪神淡路大震災の記事が最大（707 件）を示し，これ以前の 1990 年代前半では，報道記事は 160～188 件で爆発火災事故を中心とした記事が目立った．また，1995 年以降は次第に報道件数が増加しており，突出した変化が 1995 年および 2000 年に見られた．1995 年の阪神淡路大震災は，わが国が過去経験したことのない規模で産業活動に影響を及ぼしたこと，および 2000 年の雪印集団食中毒事件により社会が企業の品質安全管理を徹底する方向へ導いたことを考慮して 1995 年および 2000 年を対象として詳細な分析を行った．

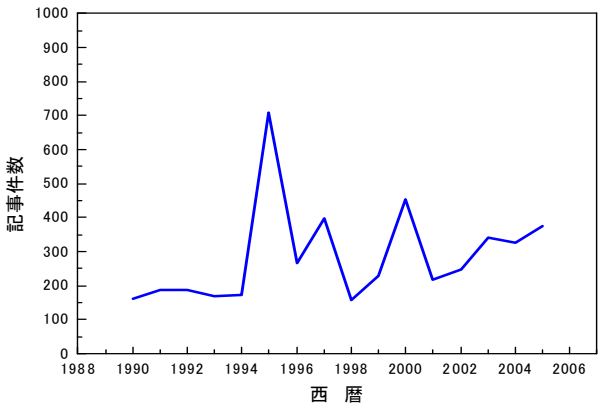


Fig. 3　「事件・事故・災害」に関する記事数の推移

3.1.　1995 年

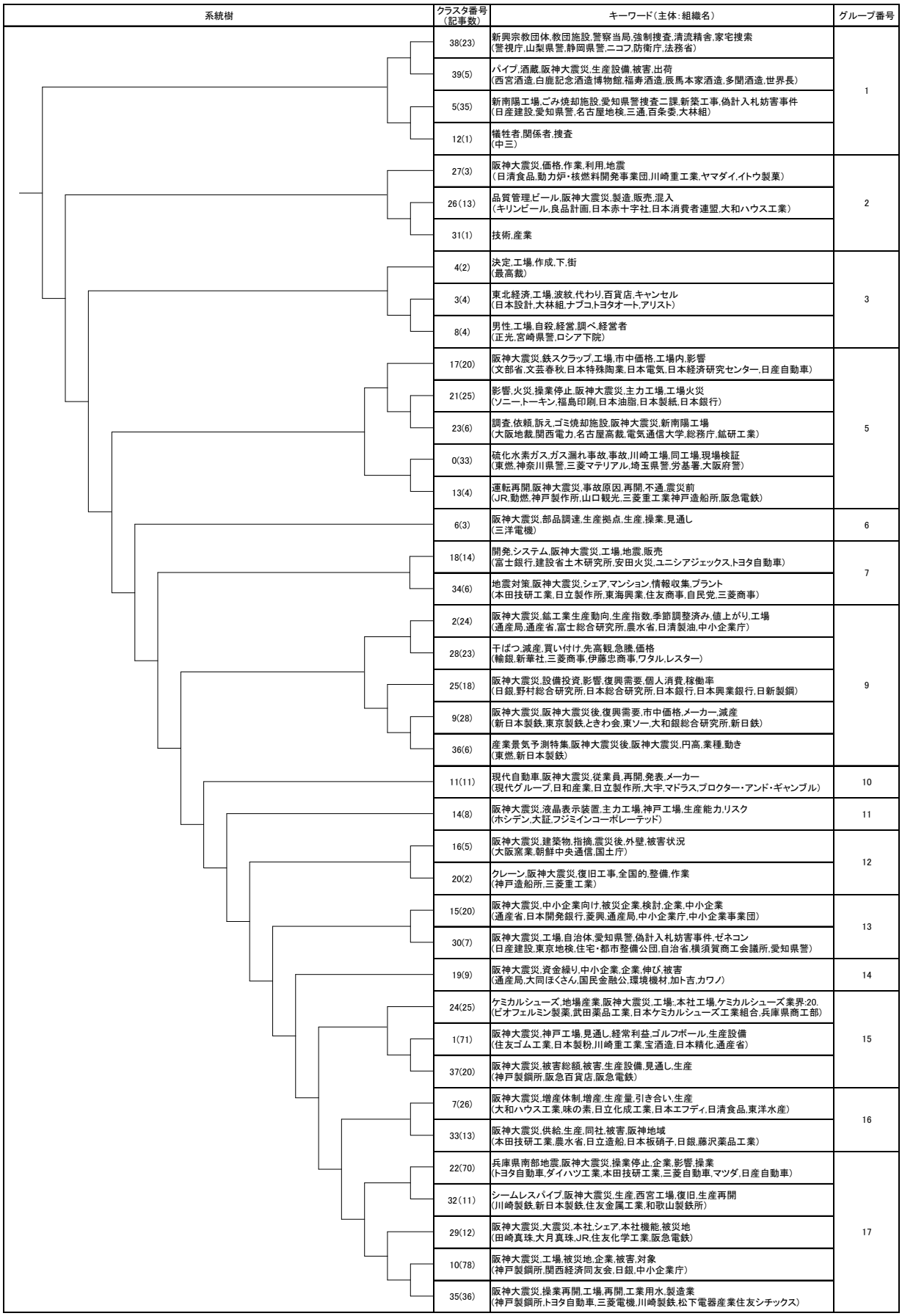
（1）文書クラスタリング

1995 年の「事件・事故・災害」に関する記事数は，707 件であった．これを 40 個のクラスタに分類した．この 40 という数字は，1 クラスタ当たり 20 から 30 件となるように新聞記事を分類したとき，経験的に記事内容の把握が容易になるためである．クラスタの数を増やすことは，系統樹において子孫の方向に向かって枝と葉の数が増すことを意味する．このため，俯瞰的に分析したい場合にはクラスタ数を少なめにとり，より細かい分類で記事をまとめたい場合にはクラスタ数を多くとればよく，必要に応じてクラスタ数を変えればよい．

文書クラスタリングによる分類結果を Fig. 4 に示す．図には，系統樹の枝分かれとクラスタに分類された記事数およびそれぞれのクラスタの記事を要約するキーワード（スコアの上位 6 件まで）および主体（組織名）について出現記事数をもとに多い順に併記した．クラスタをグループ化した結果をグループ番号として示した．Fig. 4 の作成にあたっては，分析者の特別な操作は必要なく，自動的に生成した．

クラスタ内に含まれる記事数は1件から78件の間で分布した．記事数の少ないクラスタはその内容が他のクラスタと比べて類似性に乏しいかあるいはクラスタ内で内容が散漫な傾向にあった．要約キーワードは，スコア計算に基づき抽出されているため，互いに意味をなさない語（ノイズ）が含まれた．俯瞰的な記事の分析においては，キーワード列全体からクラスタの内容を推測することとし，必要に応じて最大 20 語まで参照することによって記事の概要を把握できるようにした．

1995 年の記事は，グループ 1 から 17 まで孤立的なクラスタを形成することなく逐次的に構造化されている．さらに，キーワードを通覧することによりかなりのクラスタが阪神大震災に関わる内容を含むことが見て取れる．



系統樹	クラスタ番号（記事数）	キーワード（主体：組織名）	グループ番号
	38(23)	新興宗教団体,教団施設,警察当局,強制捜査,清流精舎,家宅捜索 (警視庁,山梨県警,静岡県警,ニコフ,防衛庁,法務省)	1
	39(5)	パイプ,酒蔵,阪神大震災,生産設備,被害,出荷 (西宮酒造,白鹿記念酒造博物館,福寿酒造,辰馬本家酒造,多聞酒造,世界長)	
	5(35)	新南陽工場,ごみ焼却施設,愛知県警捜査二課,新築工事,偽計入札妨害事件 (日産建設,愛知県警,名古屋地検,三通,百条委,大林組)	
	12(1)	犠牲者,関係者,捜査 (中三)	
	27(3)	阪神大震災,価格,作業,利用,地震 (日清食品,動力炉・核燃料開発事業団,川崎重工業,ヤマダイ,イトウ製菓)	2
	26(13)	品質管理,ビール,阪神大震災,製造,販売,混入 (キリンビール,良品計画,日本赤十字社,日本消費者連盟,大和ハウス工業)	
	31(1)	技術,産業	
	4(2)	決定,工場,作成,下,街 (最高裁)	3
	3(4)	東北経済,工場,波紋,代わり,百貨店,キャンセル (日本設計,大林組,ナブコ,トヨタオート,アリスト)	
	8(4)	男性,工場,自殺,経営,調べ,経営者 (正光,宮崎県警,ロシア下院)	
	17(20)	阪神大震災,鉄スクラップ,工場,市中価格,工場内,影響 (文部省,文芸春秋,日本特殊陶業,日本電気,日本経済研究センター,日産自動車)	5
	21(25)	影響,火災,操業停止,阪神大震災,主力工場,工場火災 (ソニー,トーキン,福島印刷,日本油脂,日本製紙,日本銀行)	
	23(6)	調査,依頼,訴え,ゴミ焼却施設,阪神大震災,新南陽工場 (大阪地裁,関西電力,名古屋高裁,電気通信大学,総務庁,鉱研工業)	
	0(33)	硫化水素ガス,ガス漏れ事故,事故,川崎工場,同工場,現場検証 (東燃,神奈川県警,三菱マテリアル,埼玉県警,労基署,大阪府警)	
	13(4)	運転再開,阪神大震災,事故原因,再開,不通,震災前 (JR,動燃,神戸製作所,山口観光,三菱重工業神戸造船所,阪急電鉄)	
	6(3)	阪神大震災,部品調達,生産拠点,生産,操業,見通し (三洋電機)	6
	18(14)	開発,システム,阪神大震災,工場,地震,販売 (富士銀行,建設省土木研究所,安田火災,ユニシアジェックス,トヨタ自動車)	7
	34(6)	地震対策,阪神大震災,シェア,マンション,情報収集,プラント (本田技研工業,日立製作所,東海興業,住友商事,自民党,三菱商事)	
	2(24)	阪神大震災,鉱工業生産動向,生産指数,季節調整済み,値上がり,工場 (通産局,通産省,富士総合研究所,農水省,日清製油,中小企業庁)	9
	28(23)	干ばつ,減産,買い付け,先高観,急騰,価格 (輸銀,新華社,三菱商事,伊藤忠商事,ワタル,レスター)	
	25(18)	阪神大震災,設備投資,影響,復興需要,個人消費,稼働率 (日銀,野村総合研究所,日本総合研究所,日本銀行,日本興業銀行,日新製鋼)	
	9(28)	阪神大震災,阪神大震災後,復興需要,市中価格,メーカー,減産 (新日本製鉄,東京製鉄,ときわ会,東ソー,大和銀総合研究所,新日鉄)	
	36(6)	産業景気予測特集,阪神大震災後,阪神大震災,円高,業種,動き (東燃,新日本製鉄)	
	11(11)	現代自動車,阪神大震災,従業員,再開,発表,メーカー (現代グループ,日和産業,日立製作所,大宇,マドラス,プロクター・アンド・ギャンブル)	10
	14(8)	阪神大震災,液晶表示装置,主力工場,神戸工場,生産能力,リスク (ホシデン,大証,フジミインコーポレーテッド)	11
	16(5)	阪神大震災,建築物,指摘,震災後,外壁,被害状況 (大阪窯業,朝鮮中央通信,国土庁)	12
	20(2)	クレーン,阪神大震災,復旧工事,全国的,整備,作業 (神戸造船所,三菱重工業)	
	15(20)	阪神大震災,中小企業向け,被災企業,検討,企業,中小企業 (通産省,日本開発銀行,菱興,通産局,中小企業庁,中小企業事業団)	13
	30(7)	阪神大震災,工場,自治体,愛知県警,偽計入札妨害事件,ゼネコン (日産建設,東京地検,住宅・都市整備公団,自治省,横須賀商工会議所,愛知県警)	
	19(9)	阪神大震災,資金繰り,中小企業,企業,伸び,被害 (通産局,大同ほくさん,国民金融公,環境機材,加ト吉,カワノ)	14
	24(25)	ケミカルシューズ,地場産業,阪神大震災,工場;,本社工場,ケミカルシューズ業界:20, (ビオフェルミン製薬,武田薬品工業,日本ケミカルシューズ工業組合,兵庫県商工部)	15
	1(71)	阪神大震災,神戸工場,見通し,経常利益,ゴルフボール,生産設備 (住友ゴム工業,日本製粉,川崎重工業,宝酒造,日本精化,通産省)	
	37(20)	阪神大震災,被害総額,被害,生産設備,見通し,生産 (神戸製鋼所,阪急百貨店,阪急電鉄)	
	7(26)	阪神大震災,増産体制,増産,生産量,引き合い,生産 (大和ハウス工業,味の素,日立化成工業,日本エフディ,日清食品,東洋水産)	16
	33(13)	阪神大震災,供給,生産,同社,被害,阪神地域 (本田技研工業,農水省,日立造船,日本板硝子,日銀,藤沢薬品工業)	
	22(70)	兵庫県南部地震,阪神大震災,操業停止,企業,影響,操業 (トヨタ自動車,ダイハツ工業,本田技研工業,三菱自動車,マツダ,日産自動車)	17
	32(11)	シームレスパイプ,阪神大震災,生産,西宮工場,復旧,生産再開 (川崎製鉄,新日本製鉄,住友金属工業,和歌山製鉄所)	
	29(12)	阪神大震災,大震災,本社,シェア,本社機能,被災地 (田崎真珠,大月真珠,JR,住友化学工業,阪急電鉄)	
	10(78)	阪神大震災,工場,被災地,企業,被害,対象 (神戸製鋼所,関西経済同友会,日銀,中小企業庁)	
	35(36)	阪神大震災,操業再開,工場,再開,工業用水,製造業 (神戸製鋼所,トヨタ自動車,三菱電機,川崎製鉄,松下電器産業住友シチックス)	

Fig. 4　1995 年の記事より生成した系統樹及び要約キーワード上位 6 語，主体抽出結果

Table 1　トピックを形成するクラスタからの情報抽出（1995 年）

グループ番号（中心度）	クラスタ番号（記事数）	密度 中心度	該当率（%）	記事内容	トピック
17 (0.235)	22 (70)	0.605 0.234	94	・兵庫県南部地震発生（1/17），広範囲な企業が操業停止 ・英ロイズ，地震保険規制により損失規模は国際再保険の枠組みで解消可能と発表 ・トヨタ，本田技研，部品調達が困難となり，一部のラインが操業停止 ・震災死者が4000人を超えて，戦後最悪の被害 ・JR貨物の復旧が長引き，物流の海運シフトが進む ・工場再開へ向け企業連携 ・阪神大震災の大阪府被害は工場など14%34社（アンケート結果）	被害実態把握，被災後復興への対応，操業再開への取り組み，課題
	10 (78)	0.568 0.236	96	・地場産業であるケミカルシューズ工業組合の半数以上の工場が焼失（中小企業庁調べ） ・関西経済同友会が被災地の被害報告と復興対策について復興会議 ・神戸市被災企業の移転先に産業団地前倒し分譲，優先枠設定 ・兵庫県仮設賃貸工場が400社分110億円で被災中小企業を支援 ・労働省緊急対策として，雇用調整助成制度適用 ・被災企業向けに神奈川県が融資 ・兵庫県と神戸市中小向けに特別融資，一定期間無利子 ・中国天津市中小企業復興支援に工場リースや受託加工 ・企業を悩ます3難題：雇用，流通，自宅待機 ・震災から1ヶ月で被害額公表企業が90社超す ・企業影響：金属製品，酒造，家具，陶磁器，百貨店，ゴム靴，空調機器，一般機器，建機リース ・日本鉄鋼連盟，税制，雇用などで企業支援策要望 ・関西経済連合会，大阪商工会議所，「復興に関する緊急提言」を発表，税制特別措置，特別法制定，規制緩和 ・仮設工場入居希望，12.7倍 ・震災で倒産22件（2月） ・阪神大震災被害状況調査で減収4ヶ月2.6兆円となり，資産被害と上回ったことが判明	
	35 (36)	0.448 0.236	94	・震災後復旧活動の取り組み：新日鉄操業再開，自動車，電機など大手メーカー一部操業再開 ・工場用水停止で水確保急ぐ ・アジアの製造業日本から部品供給困難，日系メーカー新規調達先を開拓 ・中小企業の操業再開に向けて仮工場確保難航する ・震災，円高で産業の空洞化が懸念	
	29 (12)	0.383 0.235	92	・神戸の顔ファッション業界情報発信機能が低下，田崎真珠，大月真珠	
	32 (11)	0.346 0.234	91	・川鉄，住金に生産委託要請 ・新日鉄，川鉄，シームレスパイプ部品ですみ分け生産，コスト削減 ・川鉄，カラー鋼板撤退前倒し ・川鉄の被害額107億円	
16 (0.231)	7 (26)	0.317 0.231	96	・保存食，乾電池などの必需品を震災地向けに大幅増産，東洋水産，日清食品，味の素，松下電池工業 ・ミネラルウォーターを大幅増産，ハウス食品，サントリーなど飲料各社 ・無菌パック米飯引き合い殺到，全国農協食品 ・松下電工，建築関連資材の増産 ・灘被災で福岡酒造会社が倒産 ・プレハブ，仮設店舗増産 ・厨房機器メーカー，仮設住宅向けキッチンを増産，サンウェーブ，クリナップ ・金庫，耐火庫を増産，熊平製作所 ・廃材リサイクルに向けコンクリート破砕機増産 ・ポリエチレン製ガス管を増産，積水化学	復興需要に向けた増産活動
	33 (13)	0.239 0.231	85	・医療メーカー，生産，物流の復旧に全力 ・米支給支援に向け近畿の卸各社フル稼働 ・住友ゴム，工業被災のため本多二輪生産休止 ・三井東圧，復興需要に向け塩化ビニルを緊急輸入 ・日本板硝子増産 ・梁瀬産業，「点滴キット」設備増強	
15 (0.228)	1 (71)	0.570 0.229	99	・阪神大震災による製造業被害発表，川重80億円，新明和60億円，ナフコ45億円，昭和産業42億円，富士通40億円，日本製粉20億円，日清製粉20億円 ・神戸製鋼への支援策固まる，自動車産業にも配慮 ・日本製粉神戸工場を閉鎖，住友ゴム神戸工場閉鎖 ・外国損保保険金支払い150億円 ・阪神大震災の建築物，都市基盤被害額は9兆6千億円（国土庁）	産業活動への影響および被害見積もり
	24 (25)	0.373 0.228	95	・ケミカルシューズに関わる零細企業など地場産業に打撃 ・ケミカルシューズ，輸入急増に危機感高まる ・主力工場が破綻したビオフェルミン製薬は武田薬品工業に生産委託	
13 (0.214)	15 (20)	0.327 0.214	75	・兵庫県，神戸市など近畿自治体は中小向けに緊急融資を開始 ・通産省，「復興対策チーム」を省内に設置，貸工場建設支援 ・工業用水道の復旧は上水道優先のため，最低でも1ヶ月かかる（通産省） ・高シェアをもつ半導体，合成皮革など被災企業の製品，素材の出荷停止は当面の代替措置で影響を回避，長期的には不透明（通産省） ・通産省，投資減税，低金利融資など被災企業支援 ・通産省，工業用水道の防災対策強化およびマニュアル作成 ・大阪府・市，被災企業の積極的受け入れ	復興特別立法，復興計画策定など国・自治体による被災企業支援対策
9 (0.177)	9 (28)	0.561 0.177	96	・阪神大震災復興に建設用鋼材（H型鋼）の需要が高まるが供給に余力 ・新日鉄など鉄鋼各社は神戸製鋼所の生産肩代わりに緊急増産 ・復旧需要，中部経済にプラス ・復旧需要に鉄鋼，空調機器，セメント，合板などで増産の動き（日銀福岡） ・H型鋼復興需要は低迷，当て込み増産，在庫増で安売り激化 ・震災で自粛の中，製油工場再稼動で食用油の供給量が上回り安値販売 ・土木工事は進む一方で建築向け復興需要は盛り上がりに欠く（7月） ・近畿の素材産業緩やかに在庫調整進み，需要が本格化（12月）	増産・在庫調整，生産地のシフトなど震災復興の経済波及
	2 (24)	0.423 0.176	83	・食用油メーカー操業停止で肥料用菜種かす高値，農水省備蓄飼料穀物放出 ・生産停止と交通網遮断で中小企業の生産指数回復鈍化，親企業は海外調達，生産移転 ・1月の鉱工業生産指数，震災で1.4%低下 ・非被災地への生産シフト，中部鉱工業生産3.9%上昇，四国5.5%上昇，中国1.0%上昇，兵庫県1月11%減 ・復興需要は3年間で36兆円（関西産業活性化センター）	
	25 (18)	0.388 0.177	56	・震災による収益減と復興需要で鉄鋼，化学，など収益計画の大幅な修正を迫られる ・山陽新幹線不通で新幹線利用率18.9%減の一方で航空旅客が増える，大阪ホテルの稼働率は上昇 ・幕張都心ホテルは，修学旅行客の関西から関東への誘致に攻勢 ・復興需要は当て込み増産が裏目に出て，下期以降に本格化する見込み ・震災復興対策で経済成長率押し上げ，今後3年間0.2-0.4%の寄与（IMF）	
7 (0.158)	18 (14)	0.320 0.158	57	・地震に強い耐震木製家具の開発 ・冷却に油を使用しない空冷式の防災型変圧器の輸入，販売 ・揺れを検地し点灯する非常灯の開発，販売 ・空中から消火活動が行えるヘリ用消火装置開発 ・住宅および工場向けの耐震性の高い壁面用下地材の新工法の開発	住宅，工場の耐震，防火対策技術の開発

しかし，グループ 1 から 5 に含まれるクラスタは，阪神淡路大震災に一部関連する記事を含むものの，震災とは関係のない事件に関するものが多い．ここで震災関連記事とそうでない記事が同程度に含まれるグループ 5 を分析の対象とするかどうかがこの段階では不明確であるので，次の密度 - 中心度の評価によって判定する．

(2) クラスタの密度 - 中心度指標による評価

系統樹の構造とキーワードおよび主体からおおよその記事の傾向を読み取ることができるが，これから 1995 年の中心的なトピックを定量的に抽出するために密度 - 中心度の関係を検討した．中心度が高いクラスタには，系統樹上でそのクラスタの近くに類似した他のクラスタが密集しており，その話題が近傍のクラスタに広く及んでいる．また，密度は，クラスタ内に含まれる記事が類似した内容の場合に高くなり，類似した内容が度々報道されたことを意味する. 得られた 40 個のクラスタ全てについて中心度および密度を算出し，プロットした結果を Fig. 5 に示す. グループ毎に図中の記号を揃えて示した．また，点線は密度および中心度の平均値を表す．中心度の平均値（0.163）近くおよびこの値を超えるクラスタについてみると，グループ 7～グループ 17 までのクラスタが該当し, 523 件（記事全体の 74%）の記事が含まれた．この時の CBD 因子は 24 となった．系統樹上で近接するノードにあるグループ 5 および 6 を加味すると 614 件（87%），CBD 因子は 33 に増加した．グループ 5 および 6 について記事の内容を吟味したところ，阪神淡路大震災に関する記事の該当率を調べると 50%以下と低く，「事件・事故・災害」という話題とは無関係の記事が多かった．CBD 因子の増大は，内容の相補性にギャップがあることが原因だと考えられる．

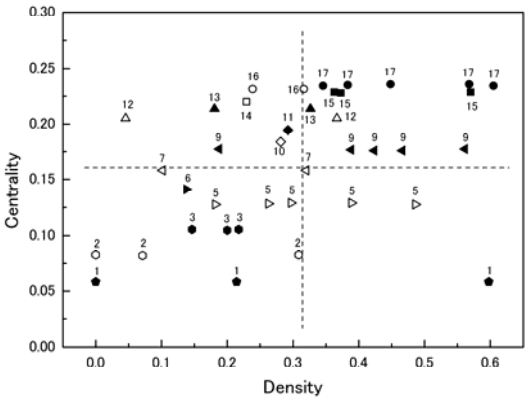


Fig. 5　密度 - 中心度図（1995 年）
図中の番号は Fig. 4 のグループ番号に対応

(3) 課題発見

これまでの分析によって，1995 年の記事から阪神淡路大震災に関係する記事を含む重要クラスタおよびグループを抽出した．これらクラスタに含まれる記事を要約キーワードについて KWIC 検索し，個々の記事から情報抽出した結果を Table 1 に示す．表には，中心度の高いグループの順に密度が平均値程度以上のクラスタを記載した．中心度の高いグループに所属するクラスタは記事の該当率（「事件・事故・災害」に関係する記事の割合）は 90%以上と高い．中心度が平均値程度となると 50%程度と低くなったが，全体として 86%の高水準の記事分類が達成されていた．

各クラスタ群のトピックを中心度の高い順に記す．

①被害実態把握，被災後復興への対応，操業再開への取り組みおよび課題.
②復興需要に向けた増産活動.
③産業活動への影響および被害見積もり.
④復興特別立法，復興計画策定など国・自治体による被災企業支援対策.
⑤増産，在庫調整，生産地シフトなどによる震災復興の経済波及.
⑥家屋，工場の地震，防火対策技術の開発.

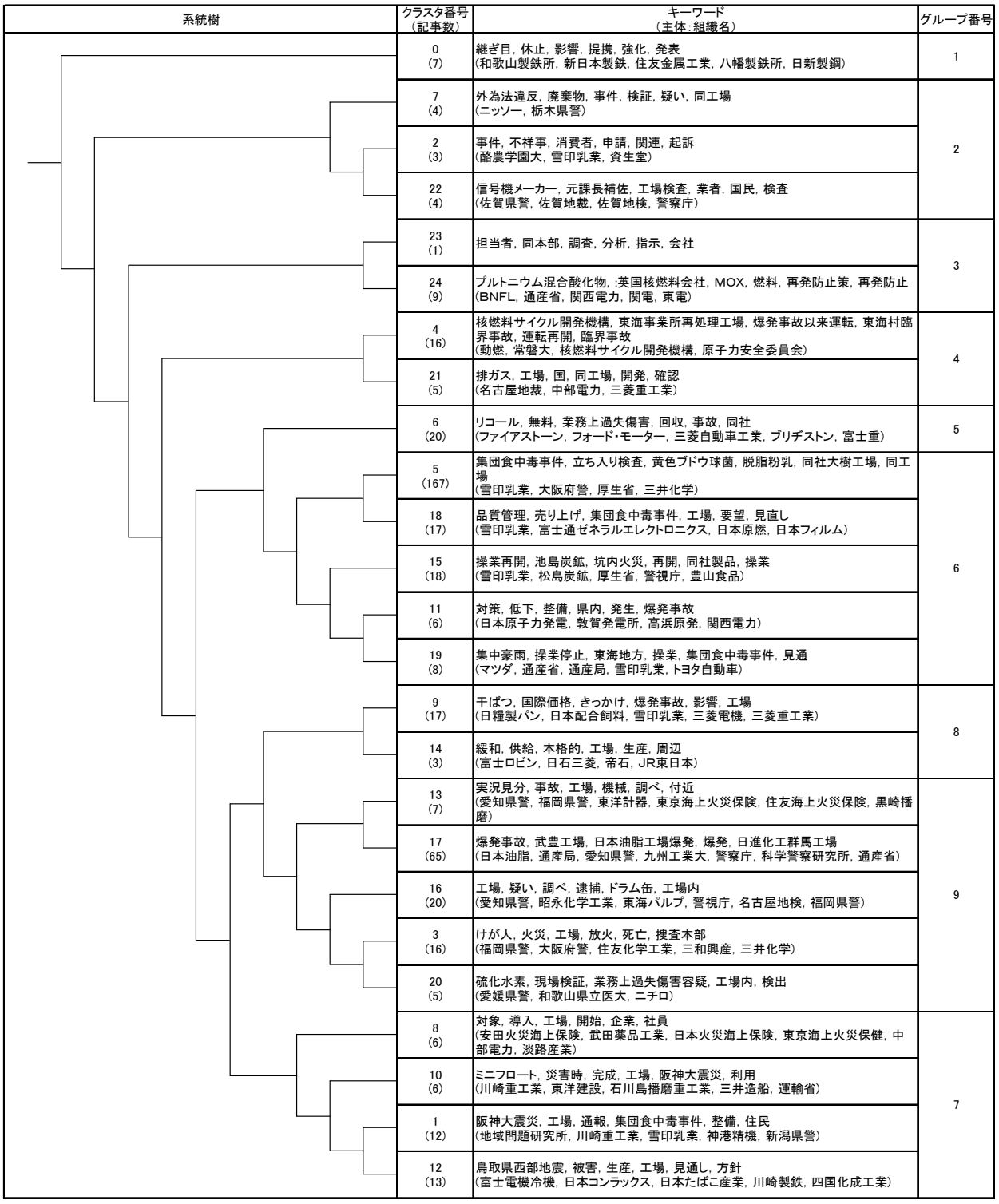
阪神淡路大震災は，わが国経済が拡大発展して以降，初めて経験した大規模な自然災害となった．1995 年 1 月 17 日に発生したことから，1 年間の記事解析は，地震災害による大規模操業停止とその後短期間に生じる緩和過程を見出すこととなった．①は，②以降の内容を含む網羅的なものとなった．これは中心度が高いことを裏付ける結果である．①に含まれるクラスタ毎に個々の記事を時系列的に追うと，産業社会の動きと戸惑いが極めて詳細に把握される．例えば，地場産業の壊滅と中国への生産シフト，企業間提携，国・自治体の復興対策など，急激に産業活動の建て直しに向けた協力と再編が次々と連鎖した点などである．②項以下の内容は①項の出来事を取り出し抽象化したような関係にあり，内容が構造化されていることが理解できる．

以上のように，利用者は対象とする社会課題の知識を前提とすることなく，どのようなイベントが具体的にどのように経時的に発生し，推移していくかを，記事の語彙類似性に基づいて客観的に分類することができる．また，抽出された要約キーワードをもとにトピックを抽出することによってトピック間の関連性を広範囲に把握することができる．

3.2.　2000 年

(1) 文書クラスタリング

クラスタの数は，25 個として分類を行なった．Fig. 6 に生成された系統樹を示す．クラスタに分類された記事数は，1 件から 167 件に分布した．記事数が 167 件と最も多かったクラスタは，雪印乳業による集団食中毒事件に関するもの，2 番目に多かったクラスタは記事数 65 件で日本油脂武豊工場の爆発事故に関するものであり，いずれも話題性が高い事件であった．2000 年の記事はいくつかのまとまったクラスタの束なりが存在する．例えば，



系統樹	クラスタ番号（記事数）	キーワード（主体:組織名）	グループ番号
	0 (7)	継ぎ目, 休止, 影響, 提携, 強化, 発表 (和歌山製鉄所, 新日本製鉄, 住友金属工業, 八幡製鉄所, 日新製鋼)	1
	7 (4)	外為法違反, 廃棄物, 事件, 検証, 疑い, 同工場 (ニッソー, 栃木県警)	2
	2 (3)	事件, 不祥事, 消費者, 申請, 関連, 起訴 (酪農学園大, 雪印乳業, 資生堂)	
	22 (4)	信号機メーカー, 元課長補佐, 工場検査, 業者, 国民, 検査 (佐賀県警, 佐賀地裁, 佐賀地検, 警察庁)	
	23 (1)	担当者, 同本部, 調査, 分析, 指示, 会社	3
	24 (9)	プルトニウム混合酸化物, :英国核燃料会社, MOX, 燃料, 再発防止策, 再発防止 (BNFL, 通産省, 関西電力, 関電, 東電)	
	4 (16)	核燃料サイクル開発機構, 東海事業所再処理工場, 爆発事故以来運転, 東海村臨界事故, 運転再開, 臨界事故 (動燃, 常磐大, 核燃料サイクル開発機構, 原子力安全委員会)	4
	21 (5)	排ガス, 工場, 国, 同工場, 開発, 確認 (名古屋地裁, 中部電力, 三菱重工業)	
	6 (20)	リコール, 無料, 業務上過失傷害, 回収, 事故, 同社 (ファイアストーン, フォード・モーター, 三菱自動車工業, ブリヂストン, 富士重)	5
	5 (167)	集団食中毒事件, 立ち入り検査, 黄色ブドウ球菌, 脱脂粉乳, 同社大樹工場, 同工場 (雪印乳業, 大阪府警, 厚生省, 三井化学)	6
	18 (17)	品質管理, 売り上げ, 集団食中毒事件, 工場, 要望, 見直し (雪印乳業, 富士通ゼネラルエレクトロニクス, 日本原燃, 日本フィルム)	
	15 (18)	操業再開, 池島炭鉱, 坑内火災, 再開, 同社製品, 操業 (雪印乳業, 松島炭鉱, 厚生省, 警視庁, 豊山食品)	
	11 (6)	対策, 低下, 整備, 県内, 発生, 爆発事故 (日本原子力発電, 敦賀発電所, 高浜原発, 関西電力)	
	19 (8)	集中豪雨, 操業停止, 東海地方, 操業, 集団食中毒事件, 見通 (マツダ, 通産省, 通産局, 雪印乳業, トヨタ自動車)	
	9 (17)	干ばつ, 国際価格, きっかけ, 爆発事故, 影響, 工場 (日糧製パン, 日本配合飼料, 雪印乳業, 三菱電機, 三菱重工業)	8
	14 (3)	緩和, 供給, 本格的, 工場, 生産, 周辺 (富士ロビン, 日石三菱, 帝石, JR東日本)	
	13 (7)	実況見分, 事故, 工場, 機械, 調べ, 付近 (愛知県警, 福岡県警, 東洋計器, 東京海上火災保険, 住友海上火災保険, 黒崎播磨)	9
	17 (65)	爆発事故, 武豊工場, 日本油脂工場爆発, 爆発, 日進化工群馬工場 (日本油脂, 通産局, 愛知県警, 九州工業大, 警察庁, 科学警察研究所, 通産省)	
	16 (20)	工場, 疑い, 調べ, 逮捕, ドラム缶, 工場内 (愛知県警, 昭永化学工業, 東海バルブ, 警視庁, 名古屋地検, 福岡県警)	
	3 (16)	けが人, 火災, 工場, 放火, 死亡, 捜査本部 (福岡県警, 大阪府警, 住友化学工業, 三和興産, 三井化学)	
	20 (5)	硫化水素, 現場検証, 業務上過失傷害容疑, 工場内, 検出 (愛媛県警, 和歌山県立医大, ニチロ)	
	8 (6)	対象, 導入, 工場, 開始, 企業, 社員 (安田火災海上保険, 武田薬品工業, 日本火災海上保険, 東京海上火災保健, 中部電力, 淡路産業)	7
	10 (6)	ミニフロート, 災害時, 完成, 工場, 阪神大震災, 利用 (川崎重工業, 東洋建設, 石川島播磨重工業, 三井造船, 運輸省)	
	1 (12)	阪神大震災, 工場, 通報, 集団食中毒事件, 整備, 住民 (地域問題研究所, 川崎重工業, 雪印乳業, 神港精機, 新潟県警)	
	12 (13)	鳥取県西部地震, 被害, 生産, 工場, 見通し, 方針 (富士電機冷機, 日本コンラックス, 日本たばこ産業, 川崎製鉄, 四国化成工業)	

Fig. 6　2000 年の記事より生成した系統樹及び要約キーワード上位 6 語，主体抽出結果

Table 2　トピックを形成するクラスタからの情報抽出（2000 年）

グループ番号（中心度）	クラスタ番号（記事数）	密度 中心度	該当率（%）	記事内容	トピック
9 (0.279)	17 (65)	0.586 0.277	88	・化学薬品製造会社，日進加工群馬工場，フリーヒドロキシルアミン，蒸留塔爆発，4人死亡27人けが250世帯停電，県は被害を受けた事業所に中小企業復旧資金制度の利用を呼びかけた ・耐火れんがメーカー，クロサキ（北九州）で工場爆発，1人死亡 ・日本油脂武豊工場（愛知県）で火薬庫（無煙火薬）爆発，住民ら51人が負傷，民家被害800戸，住民15人にPTSD症状，事故爆発損害額は13億円，管理体制不備を認める，住民に30億円賠償 ・電子部品製造会社，鈴木製作所（町田）工場内で集塵機の修理中に爆発，5人けが ・ノリタケ砥石製造工場で受注増に向け短時間乾燥試験運転中に爆発事故	工場の大規模爆発・火災事故および放火・殺人事件
	16 (20)	0.515 0.278	70	・共和レーザー新城工場（愛知県）で燃料油1000リットルが側溝を通じ川に流出 ・化学薬品メーカー，昭永化学工業で強盗殺人，遺体切断で社員逮捕 ・羽田整備工場で日航ジャンボ貨物室のワイヤーが切断される ・東海パルプ（静岡県）で火炎瓶，放火未遂事件 ・トヨタ工場で命綱外し，作業中転落	
	20 (5)	0.514 0.280	100	・宮山製肥工場（和歌山県）で羊毛から肥料を作る実験中硫化水素発生，1人死亡2人重体 ・紡績会社，フジボウ愛媛，排水槽で清掃作業中硫化水素中毒，3人死亡	
	3 (16)	0.382 0.280	81	・発泡スチロール粉砕工場，三和興産から出火，放火 ・プラスチック加工業，クラウン化学工業所（大阪），工場火災，2人死亡2人重体，放火 ・冷凍食品，ニチロ大江工場，火災 ・あじかん静岡工場火災，操業停止 ・三井化学大阪工場，塗料原料製造プラントで火災 ・住友化学千葉工場，合成ゴムプラントから出火 ・協和発酵堺工場，廃液タンク爆発，民家の窓ガラスが割れる	
8 (0.269)	9 (17)	0.406 0.269	76	・北米カイザー・アルミニウム社工場の爆発事故でアルミナの国際価格2.5倍の高止まり ・ブラジルの干ばつでコーヒー，砂糖の国際価格がこう着状態 ・米国，寒波で暖房油が上昇し，農産物市況上昇 ・ルーマニア，精錬工場の過失で深刻な環境汚染	海外企業の事故および干ばつによる原材料の海外市況
7 (0.266)	10 (6)	0.450 0.265	83	・災害時に復旧活動を支援する浮体式防災基地（ミニフロート）設置，大阪港および名古屋港 ・有珠山噴火活動	地震，噴火災害への対応
	12 (13)	0.255 0.267	69	・鳥取県西部地震，工場への被害は軽微，半導体，自動車は早々に操業再開，食品は一部操業停止 ・岸壁および漁港の施設が損壊，漁業，観光に影響，地震被害455億円（鳥取県復興災害本部） ・有珠山噴火により周辺住宅，工場に被害	
6 (0.257)	5 (167)	0.914 0.256	97	・雪印乳業大阪工場，低脂肪乳から黄色ブドウ球菌を検出（7月3日） ・大阪府警，発症者が増加の一途をたどり，雪印乳業大阪工場を業務上過失傷害容疑で現場検証（7月3日） ・雪印乳業乳製品撤去進む ・雪印大阪工場閉鎖，発症者1万人を超す ・厚生省，大阪工場「総合衛星管理製造過程」の承認（HACCP）を取り消し，大阪工場は閉鎖，厚生省対策本部，雪印の衛生管理の審査強化 ・雪印他社工場も順次調査開始，雪印静岡工場，洗浄記録がないことにより製造ライン停止 ・雪印食品中毒，1ヶ月で名古屋工場生産再開，再発防止に向けて1年間監視，再発防止策は1)雪印乳業に対する監視の継続，2)自治体と連携した企業に対する指導の強化，3)「危険度分析による衛生管理」（HACCP）の運用を改善するための検討会の設置，4)加工乳の再利用のあり方を協議する有識者懇談会の設置，5)食品監視の重点化 ・雪印大樹工場で製造された脱脂粉乳から黄色ブドウ球菌毒素検出（8月18日） ・雪印大樹工場，停電放置で菌増殖，ずさんな管理確認で営業停止，安全宣言の信頼が崩れる ・厚生省，雪印の集団食中毒事件を受けてHACCP承認制度の見直しのための評価検討会初会合を開催 ・雪印営業赤字420億円，21工場を半減へ，生産量80%減 ・雪印，食中毒の原因が大樹工場の脱脂粉乳の製造過程にあると最終報告（12月22日） ・荏原藤沢工場，全国総量に匹敵するダイオキシンを流出 ・福岡県大牟田川から高濃度ダイオキシン（基準の350倍）検出（8月26日） ・三井化学大牟田工場内で基準値を超えるダイオキシンを検出していたことを報告（9月27日） ・福岡県，三井化学大牟田工場粗製品から高濃度ダイオキシンを検出（10月26日）	集団食中毒，ダイオキシン汚染など生活を脅かす重大な過失事件と加害企業の操業再開
	15 (18)	0.371 0.258	72	・池島炭鉱（九州）坑内火災および操業再開 ・厚相，雪印20工場に安全宣言，操業再開したが生産量は自粛状態 ・静岡県長田酪農業協同組合，事故再発防止の全作業が完了したとして操業再開	
5 (0.246)	6 (20)	0.330 0.246	75	・富士重工「レガシー」暴走事故に対してリコール隠しで元幹部を略式起訴 ・三菱重工で51万台分リコール隠し ・米ブリヂストン・ファイアストーン，フォード車死亡事故多発問題でリコール隠し，デザイン・製造過程に原因 ・GMブラジル「コルサ」，シートベルト不具合により25件事故，2人死亡	大型リコールおよびリコール隠し

グループ 5 および 6，グループ 8 および 9，グループ 7 に属するクラスタの塊である．これらのグループのトピックは互い相補的な内容をもつにもかかわらず異なった事件を含むことが想定される．このことは密度 - 中心度図による評価でさらに明らかとなった．

⑵ クラスタの密度 - 中心度指標による評価

25 個のクラスタについて密度および中心度をマップした結果を Fig. 7 に示す．点線は密度および中心度の平均値を表す．中心度の平均値 0.232 を超えるクラスタについてみると，グループ 5～グループ 9 までのクラスタが該当し，396 件（87%）の記事が含まれた．この時の CBD 因子は 12 となった．系統樹上でこれらのグループに最も近いグループ 4 を加えると記事数は 417 件となり 92%の記事をカバーする．同様にこの時の CBD 因子は 25 となった．グループ 4 は，1997 年に発生した核燃料サイクル機構再処理工場における臨界事故後の運転再開要請に関する記事が多く含まれる．CBD 因子でみるとその値は 2 倍となり，内容の相補性に大きなギャップがあると判断できる．値自体は，1995 年の阪神淡路大震災に関する記事とほぼ同程度ではあったが，グループ 4 の

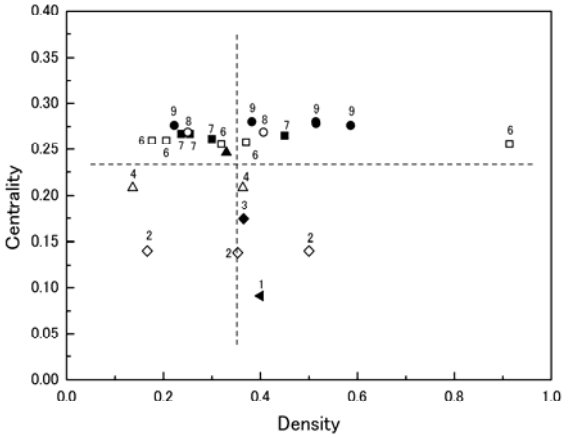


Fig. 7　密度 - 中心度（2000 年）
図中の番号は Fig. 6 のグループ番号に対応

中心度が 0.208 であり，平均値を下回ることからこれを除くこととした．このように内容の広がりを CBD 因子として定量的に把握できることが CBD 因子導入のメリットである．

(3) 課題発見

Table 2 は，中心度の高いグループの順に密度が高いクラスタを記載した．中心度の高いグループに所属するクラスタは記事の該当率が 70～100%であり，全体として 81%の高水準の記事分類が達成された．クラスタからの情報抽出結果に基づいてグループ毎にトピックを要約すると次の様である．

①工場の大規模爆発，火災事故および放火，殺人事件.
②海外企業の事故および干ばつによる海外市況変化.
③地震，噴火災害への対応.
④集団食中毒，ダイオキシン汚染など不特定多数の生活を脅かす重大な過失事件と操業再開.
⑤大型リコールおよびリコール隠し.

2000 年は 1995 年に次いで記事数が多い年であった．この年が提起する課題は，化学工場の大爆発，集団食中毒事件，自然災害など，一見異なったジャンルのトピックともみられるが，製造企業の過失あるいは対応の誤りが社会に大きな不安を引き起こしたことに共通点がある．企業による組織災害を社会全体の監視によって歯止めを設けることの必要性を強く印象づけた. 実際に 2000 年以降，マスコミも類似した事件を積極的に取り扱い，安全安心に関する企業の管理意識も高まった．系統樹上では集団食中毒や爆発事故は異なった枝に分割されたが，中心度および密度指標ではいずれも高い値を示しており，一見異なるトピックをもつ記事についても相対的重要性を客観的に把握することが可能であった．

また，2000 年の分析では集団食中毒事件および化学工場の大爆発など，重大事故に関する記事が高い該当率でクラスタリングされた．これらは科学技術振興機構が提供する失敗知識データベースに収録されている事例の中でも注目度の高いケースとなっており [17]，提案手法による分析と組み合わせれば失敗事故の発生に伴う社会変化についても発生当時の状況を時系列で追跡することが可能となる．また，データベースに収録されていない関連事例も抽出することができるなど，今後の安全安心に関する社会技術研究に寄与することが期待される．

以上二つの分析事例について本手法の有効性について評価した．これまで新聞記事を対象としたテキストマイニングの事例として Uramoto らの報告 [9] がある．これと比較すると本手法の特徴は次のように述べることができる．Uramoto らの手法は，個々の記事同士の関連性，相違性を評価し，単に文書クラスタを構造化するに留まっている．この範囲では，類似した記事の集合が脈絡なく大量に生成され，結果的に分析者は全てのクラスタ内の記事に目を通すことが必要となり，小規模な文書に対しては有効であるが，文書量が増えるに従って，分析作業は困難となる．我々は，共語分析の方法として知られる密度及び中心度指標 [6) 7) 14)] に注目し，これを系統樹上でのクラスタの関係構造を反映する指標として再定義し，記事クラスタを定量的にランキングする方法を導入した．本手法で用いた個々のパラメータのチューニングは，実験的にパラメータを振り，最も良かったパラメータを使用した．これらによって，従来の方法と比べて格段に分析の精度を高め，作業を効率化するシステムが実現できたと考えている．

4. まとめ

本研究では，新聞記事に対してテキストマイニングの手法を応用し，解析の対象となる記事文書集合を取得して自動的にこれを分類して情報を要約, トピックを抽出, さらには課題発見に至る新しい定量的な解析システムを構築した．具体的には，得られた文書集合に対して階層的クラスタリングを施し，クラスタ間の語彙使用の類似性に基づく構造化を行い，クラスタを要約するキーワードおよび関係主体を抽出することによって内容を把握する俯瞰的アプローチを可能とし，さらに CBD 法の概念を発展させて課題（仮説）発見に至る手続きを示した．実際にこの手法を産業活動に伴う事故・災害に関する社会の課題発見に適用し，80%程度の記事該当率およびカバー率で，記事集合の内容を反映した課題発見を導く事例を示した．

俯瞰的アプローチの利点は，文書集合に対して特定のキーワードあるいは観点を設けることなく，語彙使用の類似性という定量的な尺度でトピックを分類することにある．これによって社会事象の変化をキーワードに左右されることなく連続的にモニタリングすることができる．

しかしながら，俯瞰的アプローチでは記事数の多いトピックを中心にグループ化を行うために，まだ中心的な位置づけを持っていない新規な記事はノイズとして漏れてしまうことになる．これを補う機能として，文書クラスタから技術，サービス，制度，評価などに係る情報を自動抽出する技術についても開発を進めている．このような情報抽出機能を付加することによって，俯瞰的な分類だけでは見逃してしまう社会のトピックおよびトレンドを分析することができる．

今後，本手法の適用事例を増やし，ノウハウの蓄積を図ることも必要と考えている．特に安全安心に関する社会変化のトレンド分析は社会の関心が高く，政策分析へのインプットとしても有益である．科学技術振興機構が提供する失敗知識データベースが安全安心の解析に貢献していることは周知であるが，本研究で提案する手法と組み合わせれば，大事故，大事件に対して社会がいかに応答し，変化していくかを動的に解析することも可能となろう．

参考文献

1) 市川惇信(2005)『ristexNEWS』1, 1.
2) 堀井秀之(2006)『安全安心のための社会技術』東京大学出版会.
3) 奥田英範, 川島晴美, 佐藤吉秀, 宮原信二, 定方徹 (2006)「俯瞰的アプローチに基づく情報場ナビゲーション技術」『NTT 技術ジャーナル』18(5), 22-25.
4) Rajman, M., and Besanceon, R. (1997) Text Mining: Natural Language techniques and Text Mining applications, *Proceedings of the seventh IFIP 2.6 Working Conference on Database Semantics*, (DS-7).
5) Hearst, M.A. (1998) Untangling Text Data Mining, *ACL'98*, 3-10.
6) Callon, M., Law, J., and Rip, A. (Ed.) (1986) *Mapping the Dynamics of Science and Technology*. Macmillan Press;
7) Callon, M., Coourtial, J.P., and Laville, F. (1991) Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry, *Scientometrics*, 22, 155-205.
8) L. ライデスドルフ(2001), 『科学計量学の挑戦』, 玉川大学出版会.
9) Uramoto, N., and Takeda, K. (1998) A Method for Relating Multiple Newspaper Articles by Using Graphs, and Its Application to Webcasting, *COLING-ACL'98*, 1307-1313.
10) B.C.M. Fung, K. Wang, and M. Ester (2003), Hierarchical document clustering using frequent item sets, *Proceedings of the SIAM International Conference on Data Mining*.
11) George Karypis (2006), 『CLUTO - Software for Clustering High-Dimensional Datasets 』, http://glaros.dtc.umn.edu/gkhome/views/cluto/
12) Matsumoto, M., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., and Asahara, M. (1999) Japanese Morphological Analyzer ChaSen Users Manual version 2.0. *Technical Report NAIST-IS-TR990123*, Nara Institute of Science and Technology Technical Report.
13) 山田寛康, 工藤 拓, 松本裕治 (2004)「Support Vector Machine を用いた日本語固有表現抽出」『情報処理学会論文誌』 43(1), 44-53.
14) Stegmann, J. and Grohmann, G.. (2003) Hypothesis generation guided by co-word clustering. *Scientometrics*, *56*, 111-135.
15) Swanson, D.R., and Smalheiser, N.R. (1997) An interactive system for finding complementary literatures: a stimulus to scientific discovery, *Artificial Intelligence*, 91, 183-203.
16) Neil R. Smalheiser(2002) 『 Arrowsmith 』 http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html, [2007 年 10 月 31 日].
17) 科学技術振興機構(2001),『失敗知識データベース』, http://shippai.jst.go.jp/fkd/Search, [2007 年 10 月 31 日]
18) 日本経済新聞デジタルメディア(1982),『日経シソーラス』, http://telecom21.nikkei.co.jp/help/contract/price/00/help_KIJI_thes.html, [2007 年 10 月 31 日]

謝辞

本研究は, 文部科学省科学技術振興調整費「戦略的研究拠点育成プログラム」の支援の下に実施した. 本研究に遂行にあたって有益なご助言をいただいた東京工業大学統合研究院下田隆二教授および大熊和彦教授に感謝いたします.

TOPIC EXTRACTION AND SOCIAL PROBLEM DETECTION BASED ON DOCUMENT CLUSTERING

Taiichi Hashimoto [1], Koji Murakami [2], Takashi Inui [3], Kazuo Utsumi [4], and Masamichi Ishikawa [5]

[1]Ph.D. (Engineering) Associate professor, Tokyo Institute of Technology (E-mail: hashimoto@iri.titech.ac.jp)
[2]Ph.D. (Engineering) Tokyo Institute of Technology, (E-mail: murakami@iri.titech.ac.jp)
[3]Ph.D. (Engineering) Tokyo Institute of Technology, (E-mail: inui @iri.titech.ac.jp)
[4]M.E. Tokyo Institute of Technology, (E-mail: utsumi@iri.titech.ac.jp)
[5]Ph.D. (Engineering) Professor, Tokyo Institute of Technology (E-mail: ishikawa@iri.titech.ac.jp)

The method that enabled to extract important topics from document clusters containing text documents of many subjects retrieved from Nikkei newspaper was developed. The hierarchical clustering algorithm, UPGMA was used to generate the tree structure of clusters according to the similarity of document vectors defined by noun words appeared in the documents. The document clustering revealed the intimate relationship with the process of the societal problem detection, classifying similar documents in each topical group and structuring the groups according to their contents. The method was evaluated by applying to the subject of the organizational hazards caused by Japanese industries during 1990-2005.

Key Words: *Text mining, Document clustering, Text summarization, Information extraction, Social problem detection*