2008－NL－186（8）
2008/7/18

社会課題発見のための文書クラスタリングとクラスタ評価指標

橋本 泰一† 村上 浩司†† 乾 孝司† 内海 和夫† 石川 正道†

† 東京工業大学 統合研究院
〒226-8503 神奈川県横浜市緑区長津田町 4259
†† 奈良先端科学技術大学院大学 情報科学研究科
〒630-0192 奈良県生駒市高山町 8916-5
E-mail: †{hashimoto,inui,utsumi,ishikawa}@iri.titech.ac.jp, ††kmurakami@is.naist.jp

あらまし 文書クラスタリングは，大量の文書を俯瞰的に分析するために非常に有効的な技術である．しかし，文書クラスタリングによって分類するだけでは，クラスタのまとまりの良さやクラスタが表す話題の関係性を読み取ることは難しい．本論文では，大量の新聞記事から社会課題を発見するというタスクにおいて，共語分析のアプローチを階層的文書クラスタリングに適用した新しい分析手法を提案する．また，文書クラスタリングによって得られたデンドログラムを利用し，クラスタの話題の類似性を表す密度，クラスタ内の記事の話題の関係性を表す中心度という2つの指標を定義する．この2つの指標をもとにクラスタを選別することにより，分析者が有用な話題や情報を含む記事を効率的に取得できることを評価実験により検証した．

キーワード 文書クラスタリング，階層型クラスタリング，クラスタ評価

Document Clustering for Social Problem Detection and Cluster Evaluation Measures

Taiichi HASHIMOTO†, Koji MURAKAMI††, Takashi INUI†, Kazuo UTSUMI†, and Masamichi ISHIKAWA†

† Tokyo Institute of Technology
†† Nara Institute of Science and Technology
E-mail: †{hashimoto,inui,utsumi,ishikawa}@iri.titech.ac.jp, ††kmurakami@is.naist.jp

Abstract Document clustering that is one of core technology of text mining is useful for macro analysis of large scale of documents. However it is difficult that analyst efficiently knows which clusters include important information from the result of document clustering. This paper presents a method to support a detection of social problems using newspaper articles. The proposed method is based on a hierarchical clustering algorithm. The hierarchical clustering algorithm is able to generate a dendrogram of clusters according to the similarity of document vectors. The document vector is calculated on length and position of term in the document. And we define two new measures to detect important clusters from the dendrogram. One is called density which is a measure of relevancy of documents in the cluster. The density is calculated from rate of term that documents in cluster shared. The other is called centrality which is a measure of relevancy of clusters. The centrality is calculated from depth of shared ancestor of clusters in the dendrogram and the number of documents in the cluster. We conducted experiments to evaluate the proposed method using NIKKEI newspaper articles which describe to the organizational hazards caused by Japanese industries and found that the proposed method is able to detect important clusters from the dendrogram genetrated by the hierarchical clustering.

Key words document clustering, hierarchical clustering algorithm, cluster evaluation

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]．このような社会課題に関する情報は，新聞記事に記述されていることが多い．なぜならば，新聞記事は社会の様々な出来事の発生から関連する事柄について日々提供しており，ある出来事を発端とした社会課題や社会問題について述べられているだけでなく，その出来事が時間経過とともに社会に与えた影響に関する情報も豊富に記載されているためである．そのため，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 [2]．しかし，大量の新聞記事を人手により収集し，分類，分析することは多大な労力が必要になる．

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途を辿っており，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，大量の文書から新しい情報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の要素技術の一つである文書クラスタリングは，大量の文書に対して，自動的に内容が類似する文書群（クラスタ）に分類できるため，文書の俯瞰的な分析に適している．文書クラスタリング技術を用いて，新聞記事をベースに社会の動向分析を行う研究が行われている [3], [4]．しかし，分類されたクラスタのうち，どのクラスタが重要なクラスタであるのか判別することが困難である．

語の共起情報をもとに科学技術の重要性や影響力を分析する科学計量学の分析手法である共語分析において，重要なクラスタを選定するための指標が Callon らにより提案されている [5]．Callon らは，語の共起をもとに形成したネットワーク構造から得られる語のクラスタの解釈として，クラスタ内の語の関連度（密度）とクラスタ間の語の関連度（中心度）を定義し，その2つの指標をもとにクラスタの重要性や影響力を解釈可能であると報告している．

現在，我々は大量の新聞記事から社会が抱えている課題や問題についての分析と分析を支援するためのシステムの開発を行っている [4]．本論文では，共語分析手法を拡張し，文書クラスタリングをベースに分析者が新聞記事を効率的に分析するための手法を提案する．提案手法では，次に述べる手順により新聞記事における課題発見を行う．

（1） 新聞記事を全文検索し，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（デンドログラム）．

（3） 個々のクラスタについて，分析において重要なクラスタを判別．(密度，中心度)

（4） 分析者が重要なクラスタ内の記事の内容から社会課題について考察，発見．

まず新聞記事を語の文字列長，記事中の出現位置を考慮した文書ベクトルで表現し，階層的クラスタリングにより記事が取り上げる話題の類似度をもとに構造化を行う．次に，Callon が提案した指標を階層的クラスタリングに適用できるように拡張し，密度と中心度により重要クラスタを選定可能にした．密度は，クラスタ内の記事の話題のまとまりの良さを測る指標である．中心度は，クラスタ内の記事の話題が他の記事の話題との関連性を測る指標である．産業活動に伴う事件・事故・災害に関する新聞記事を対象として，新聞記事分析における文書クラスタリングの有効性について評価実験により検証した．さらに，密度，中心度をもとにクラスタを分析した場合，効果的に新聞記事を選定可能であるかどうかについて検証した．

2. 共語分析

科学技術活動を定量的に評価しようとする研究を科学計量学と呼ぶ [6]．科学計量学では，論文や特許を用いて行われる分析手法として，論文数分析，引用分析，共著分析，謝辞分析，共語分析などがある．

共語分析は，同じ文書に出現する語の頻度から語と語の関係を見いだし分析を行う．Callon らは，論文に付与されたキーワードの共起関係に注目することで，科学技術の動向に関する分析を行うことが可能であると報告した [5]．Callon らが提案した分析手法は，最初に論文の内容を表すキーワードを複数付与し，キーワードをノードに見立て，共起関係にあるキーワード間を繋ぎ，語のネットワークを構成する．次に，共語関係の強い語をまとめて，語のクラスタを形成する．最後に「密度」と「中心度」と呼ばれる2つの指標を基に語のクラスタの意味付けを行う．

「密度」とは，クラスタ内部の語の共起関係の強さを示す指標である．密度が高いクラスタは，ある程度確立された研究テーマを表し，密度が低いものは研究が発展途上の新しい研究テーマを表すと見なす．一方，「中心度」とは，当該クラスタと他のクラスタの共起関係の強さを示す指標である．中心度が高いクラスタは，多くの研究テーマと関連している必須の研究を表すクラスタであると見なす．個別クラスタの密度及び中心度は、これら指標を座標軸とする2次元図上にマップされ、科学技術の戦略性を順序づける分析手法として強力な役割を果たしてきた．

3. 文書クラスタリングへの共語分析的アプローチの適用

共語分析には，分析に用いるキーワードの統一が難しいという問題がある．大量な文書に対して統一されたキーワードを付与することは人手でも機械的でも非常に困難である．また，分野や時代によってキーワードが変化することも統一を困難にする原因の一つである．異なる分野の研究者が同じキーワードを付与してもその意味は異なる場合が多い．反対に，異なるキーワードであっても同じ内容を意味する場合もある．時代によっても同じである．そのため，様々な分野や異なる時代の論文を同時に分析しても，語のクラスタが誤った解釈を与えてしまう可能性がある．

社会課題の発見を目的とした新聞記事の分析として共語分析を用いる場合，この問題が顕著になる．社会課題は，様々な主体，要因が複雑に絡み合って構成される．そのため，新聞記事には，社会課題についての多種多様な分野の記事を同時に分析をする必要があり，共語分析に適していない．また，大量な新聞記事に対して統一的にキーワードを付与することも困難である．加えて，実際に社会課題について分析を行う際には，個々の新聞記事の内容に踏み込んで分析を行う必要があり，語の分析だけでは十分ではない．

本論文では，共語分析のように語のクラスタに注目するのではなく，記事の類似性に基づいた文書クラスタに注目した分析手法を提案する．提案手法では，単語ベクトルをベースとした階層型文書クラスタリングにより，新聞記事のクラスタを形成する．そして，文書クラスタリングにより得られるデンドログラムから求める新たなクラスタの密度と中心度を定義し，その指標により社会課題について分析者が想起できるようなクラスタを識別することを可能にする．

3.1 文書クラスタリング

分析対象の文書集合（D）には，一般に，(関連のある) 複数の話題が含まれる．これら文書集合に対して文書クラスタリングを実施し，話題が共通する文書をまとめる．

文書クラスタリングには，階層的ハードクラスタリング・アルゴリズム [7] を採用する．階層的クラスタリングでは，クラスタリング結果をデンドログラム（系統樹）として可視化でき，文書内容の類似性に基づいてクラスタ間の関係を俯瞰できるため，分析者が分析しやすいという利点がある．

文書クラスタリングにおいて各文書は，文書に含まれる語の情報から構成されるベクトル（文書ベクトル）として扱われる．ある文書（$d \in D$）に対応する文書ベクトル $\mathbf{d}$ は次のようにして作成する．まず，d の先頭から n 文を抽出する．そして，これらの文を形態素解析器 ChaSen [8] で解析し，語と品詞の情報を獲得する．そして，解析結果の中から，名詞および名詞が連続する語（名詞連続）のみを抽出することによって文書ベクトルを作成する．ただし，D 中における頻度が閾値 th_f（本論文では $th_f = 5$ とした） 未満となる名詞や名詞連続は考慮しない．また，文書ベクトルとして考慮される名詞，名詞連続を e で表すとし，文書ベクトル $\mathbf{d}$ の i 番目の要素の値 $w_d(e_i)$ を式 (1) で定義する．

$$w_d(e_i) = tf_d(e_i) \times \log_2 \frac{|D|}{df_D(e_i)} \times |e_i| \times \frac{1}{1 + \log(first_d(e_i))} \quad (1)$$

$tf_d(e_i)$ は文書 d 内での語 e_i の出現頻度，$|D|$ は総文書数，$df_D(e_i)$ は語 e_i の出現する文書数，$|e_i|$ は語 e_i を構成する文字数，$first_d(e_i)$ は文書 d 内で初めて語 e_i が出現した文の位置 ($1 \leqq first_d(e_i) \leqq n$) を表す．この重み付けは，TF-IDF をベースとして，分析において重要な情報となる固有表現は文字数の長い語で構成されること，および，新聞記事では記事先頭に近い文ほど記事の概要的な内容が記述されやすいという特性を考慮している．

3.2 クラスタの密度と中心度

文書クラスタリングにより文書を分類できたとしても，クラスタ内の文書の共通の話題（トピック）が識別しにくいまとまりの悪いクラスタからは，分析者は有益な情報を得ることができない．そのため，分析に有効なクラスタがどれであるのかを判別することができれば，効率的にクラスタの分析を行うことができる．

Callon らの提案した密度，中心度を階層的クラスタリングにより得られるデンドログラムをもとに新たに定義する．分析者は密度，中心度をもとに分析の際に重要となる可能性の高いクラスタを選別することができる．

階層的文書クラスタリングにおける密度と中心度を次のように定義する．

- **密度**：類似した話題の文書があつまっているか
- **中心度**：クラスタ内の文書と関連があるクラスタが多いかどうか

3.2.1 密　　度

密度をクラスタに含まれる記事に共通して出現する語により定義する（式 (2)）．これは，クラスタ内にある 2 つの文書間において共に出現する語（文書ベクトル作成の際に考慮された名詞および名詞連続）の数の割合に基づいて定義されており，共に出現する語の数が多いほど密度が高くなり，類似した話題の文書を多く含んでいると考えられる．そのため，分析者にとって，密度が高いクラスタは話題の識別が比較的容易なクラスタであることが期待できる．

$$density(c) = \frac{\text{クラスタ } c \text{ 内の 2 つ以上の文書に出現する語の数}}{\text{クラスタ } c \text{ 内の文書に出現する語の数}} \quad (2)$$

3.2.2 中　心　度

デンドログラムの構造情報をもとにした中心度を定義する (式 (3))．あるクラスタに対して，デンドログラム上で深さが深いノードで結合するクラスタが多いかどうか，結合したクラスタの文書数が多いかどうかを中心度の軸として考慮する．なぜならば，デンドログラム上での深いノードで結合するクラスタは，文書クラスタリングにおいて強い関連性を見いだせることを意味しており，クラスタに含まれる文書数は話題の波及性や影響力の高さを意味していると考えられるためである．具体的には，クラスタ c_i から見た c_j の関連度をデンドログラム上で c_i と c_j の共通する最初の祖先 $share(c_i, c_j)$ の深さ $depth(share(c_i, c_j))$ とクラスタ c_j の文書数 $|c_j|$ をかけた値で定義する．そして，その平均値をクラスタ c_i の中心度として定義する．

$$centrality(c_i) = \frac{1}{|C|} \sum_{c_j \in C} |c_j| \times depth(share(c_i, c_j)) \quad (3)$$

ただし，デンドログラムの根の深さは 0 とする．また，$|C|$ はクラスタの個数を表す．

4. 評価実験

4.1 評価実験データ

評価実験データは，キーワードを用いて「事件，事故，災害」

表 1　評価実験データの記事数

年	全体	企業災害	自然災害	その他	企業災害					自然災害		
					爆発	品質	安全	原発	その他	地震	台風	その他
2000	454	329	47	78	74	192	27	29	7	23	1	23
2003	340	241	36	63	147	41	29	12	12	23	0	13

に関連した新聞記事を検索し，その検索結果から得られた新聞記事に対して人手により 11 のカテゴリに分類した．

検索クエリは，日経シソーラスの中分類「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 結合を，検索対象は日経新聞本紙を用いた．検索クエリが記事の見出しと本文の第一段落のいづれかで条件を満たす記事のみを評価データとして用いる．

2000 年と 2003 年の日経新聞新聞記事に対し検索を行い，検索結果から得られた記事を企業に多大な影響を与える事件・事故・災害という観点のもと分類した．大きなカテゴリとして「企業災害」「自然災害」「その他」の 3 つに分類した．さらに，「企業災害」は，「爆発・火災」「品質管理」「安全管理」「原子力発電所」「その他」の 5 つのカテゴリにさらに分類した．「自然災害」は，「地震」「台風」「その他」の 3 つのカテゴリに分類した．それぞれのカテゴリの判定基準と記事内容の例を次に示し，新聞記事の分類結果を表 1 に示す．

- **企業災害**
  企業が起因する災害
  - **爆発・火災**
    工場など企業施設の爆発，火災に関する記事
    代表例）ブリヂストン栃木工場火災（2003 年）
  - **品質管理**
    リコールなど製品の品質管理に関する記事
    代表例）雪印食中毒事件（2001 年），三菱自動車リコール隠し（2002 年）
  - **安全管理**
    工場からの有毒物質流出などの安全管理に関する記事
    代表例）アスベスト問題（2005 年）
  - **原子力発電所**
    原子力発電所で起きた事故に関する記事
    代表例）美浜原子力発電所の蒸気漏れ（2004 年）
  - **その他**
    その他の企業で起きた事故に関する記事
    代表例）転落事故，落下事故など
- **自然災害**
  自然が起因する災害
  - **地震**
    地震に関する記事
    代表例）宮崎県沖地震（2003 年）
  - **台風**
    台風やハリケーンに関する記事
    代表例）ハリケーン・カトリーナ（2005 年）
  - **その他**
    その他の自然災害に関する記事
    代表例）落雷，大雪など
- **その他**
  その他の記事
  代表例）殺人事件，汚職事件，紛争など

4.2 文書クラスタリングの評価実験

評価実験として，本研究で提案する語の文字列長，文書内の出現位置を考慮した文書ベクトルの有効性を評価した．評価実験では，各年の新聞記事に対して，2 種類の文書クラスタリングを行い，評価指標としてエントロピーと純度で評価を行った．

4.2.1 文書ベクトル

評価実験では，次の 4 つの語の重み付けを比較した．

- **TI**
  語の TF-IDF による重み付け
- **TI+P**
  語の TF-IDF，文書内の出現位置のみ
- **TI+L**
  語の TF-IDF，文字列長のみ
- **TI+L+P**
  語の TF-IDF，文字列長，文書内の出現位置

4.2.2 クラスタリング・アルゴリズム

評価実験に用いた階層型ハードクラスタリング・アルゴリズムとして，文書をまとめながらクラスタリングを行う集約型階層的クラスタリング（Agglomerative Hierachical Clustering; AHC）[7] と文書を分割しながらクラスタリングを行う分割型階層的クラスタリング（Patitional Hierachical Clustering; PHC）[7] の 2 種類のアルゴリズムにより評価実験を行った．

集約型階層的クラスタリングは，非加重結合法（Unweighted Pair Group Method with Arithmetic mean, UPGMA）を採用した．文書の類似度はコサイン類似度（式 (4)）により計算した．

$$sim(d_i, d_j) = cos(d_i, d_j) = \frac{d_i \cdot d_j}{|d_i||d_j|} \quad (4)$$

分割型階層的クラスタリングは，k–平均法（k–Means）を再帰的に行い階層的クラスタリングを行う手法 [9] を採用した．基準関数式 (5) [10] を最大化するようにクラスタリングを行い，文書の類似度はコサイン類似度（式 (4)）により計算した．

$$\sum_{c \in C} \sqrt{\sum_{d_i, d_j \in c} sim(d_i, d_j)} \quad (5)$$

ここで，c はクラスタ，$sim(d_i, d_j)$ はクラスタ c 内の文書 d_i と d_j の類似度を表す．

評価実験には，上記のアルゴリズムを実装するクラスタリング・ソフトウェア CLUTO（Version 2.1.2）[11] を用いて実施した．

4.2.3 クラスタ数

上記のクラスタリング・アルゴリズムは，事前にクラスタ数を決定しなければならない．本実験では，各年の記事数を考慮して，1 クラスタに含まれる記事数の平均が約 20 から 30 に分類されるようクラスタ数を決定した．(表 2)

表 3 文書クラスタリングの評価実験結果

2000	AHC		PHC	
	Entropy	Purity	Entropy	Purity
TI	0.294	0.756	0.262	0.776
TI+P	0.313	0.734	0.269	0.780
TI+L	**0.228**	**0.807**	**0.184**	**0.840**
TI+L+P	0.284	0.774	0.233	0.796

2003	AHC		PHC	
	Entropy	Purity	Entropy	Purity
TI	0.470	0.606	0.431	0.659
TI+P	0.467	0.635	0.420	0.659
TI+L	**0.409**	0.653	0.415	0.653
TI+L+P	0.423	**0.679**	**0.413**	**0.668**

表 2 評価実験で用いたクラスタ数

年	記事数	クラスタ数
2000	454	25
2003	340	20

4.2.4 評価指標

文書クラスタリングの評価指標として，エントロピーと純度を用いた．エントロピー（*Entropy*）と純度（*Purity*）の定義[10]は次のとおりである．

$$Entropy = \sum_{r=1}^{p} \frac{n_r}{n}(-\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r}) \quad (6)$$

$$Purity = \sum_{r=1}^{p} \frac{n_r}{n}(\frac{1}{n_r}\max_i(n_r^i)) \quad (7)$$

ここで，c_r は r 番目のクラスタ，p はクラスタ数，q はカテゴリ数を表す．また，n は総文書数，n_r はクラスタ c_r に含まれる文書数，n_r^i はクラスタ c_r に含まれるカテゴリ i の文書数を表す．

4.2.5 評価実験結果と考察

2000年と2003年の各年の新聞記事に対する評価実験結果を表3に示す．

表3からクラスタリング・アルゴリズムに問わず，文書ベクトルは，TF-IDFにより作成されたもの（TI）に比べ，提案手法（TI+L，TI+L+P）がエントロピーと純度ともに良い結果を示している．しかし，出現位置のみを考慮したもの（TI+P）は，TF-IDFと同等の精度であり，TI+LとTI+L+Pの結果には大きな違いがない．文書内の出現位置よりも語の文字列長に関する情報がクラスタリングの精度向上に貢献しており，出現位置の重み付けについては改良の必要性があることがわかった．

また，クラスタリング・アルゴリズムを比べた場合，大きな差は見られないが，わずかながら集約型階層的クラスタリングに比べ，分割型階層的クラスタリングの方が良い結果を示している．

4.3 密度・中心度によるクラスタ選定の評価実験

密度はクラスタ内の話題のまとまりの良さ，中心度はクラスタ間での話題の近接性・中心性を表す指標である．分析者がこの指標をもとに順にクラスタを選択した場合，分析に有効な記事を効果的に獲得できるかどうかにより，密度と中心度の評価実験を行った．

密度や中心度によりクラスタを順序づけし，分析者はその順番に記事を分析すると仮定する．このとき，分析者が獲得できる分析に有効な記事の数の推移により評価を行う．この評価実験における分析に有効な記事とは，「企業災害」もしくは「自然災害」のカテゴリに分類される記事とした．

4.3.1 評価実験環境

前述の文書クラスタリングに用いた2000年と2003年の新聞記事に対して評価実験を行った．文書ベクトルは，語の頻度，長さ，出現位置を考慮して作成した（TI+L+P）．クラスタリング・アルゴリズムは，集約型階層的クラスタリングと分割型階層的クラスタリングの2種類を用いた．比較したクラスタ選択方針は，以下の通りである．

- **baseline**（ベースライン）
  ランダムに選択した場合の期待値
- **density**
  密度が高いクラスタを優先的に選択
- **centrality**
  中心度が高いクラスタを優先的に選択
- **combination**
  密度と中心度を掛けた値の高いクラスタを優先的に選択
- **upper bound**（上限）
  有効な記事のみを選択

4.3.2 評価実験結果と考察

評価実験結果を図1から図4に示す．2つのクラスタリング・アルゴリズムどちらにおいても，密度と中心度の性能は各年のデータによってばらつきがある．例えば，図1においては，密度に比べ中心度を考慮した方が分析に有効な記事を獲得できているが，図3においては，逆に密度の方が有利である．一方，密度と中心度をともに考慮した（combination）場合，安定して分析に有効な記事を獲得可能であった．

クラスタリング・アルゴリズムを比較してみると，集約型階層的クラスタリングが，分割型階層的クラスタリングよりも良い結果を示している．集約型階層的クラスタリングは，文書やクラスタの共通性をもとにクラスタリングを行う．一方，分割型階層的クラスタリングは，相違性をもとにクラスタリングを行う．本論文で提案した密度や中心度は文書やクラスタの相違性よりも共通性を測る指標として機能しており，結果として，分割型よりも集約型の方が良い効果を表したのではないかと考えられる．このことについては，他の分野を含め様々なデータについて評価実験を行い，検証する必要がある．

5. ま と め

本論文では，新聞記事から社会課題の発見に向けた共語分析を文書クラスタリングへ適用した分析手法を提案した．具体的には，分析対象となる新聞記事に対して，語の頻度，文字列長，文書内の出現位置をもとに文書ベクトルを作成し，階層的クラスタリングを施す．共語分析に用いられる2つの指標（密度，中心度）を階層的クラスタリングから得られるデンドログラムに適用するために新たに定義した．密度は，クラスタ内の文書の話題のまとまりの良さを表し，中心度は，クラスタ間の話題

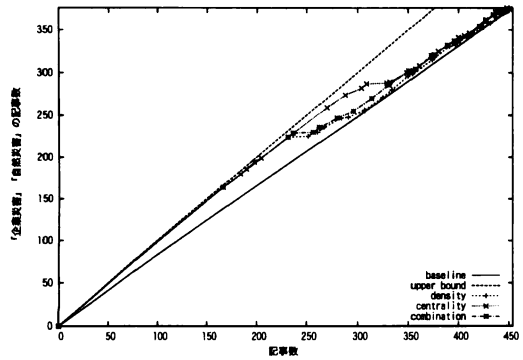

図 1　クラスタ選定の評価実験（2000 年, AHC）

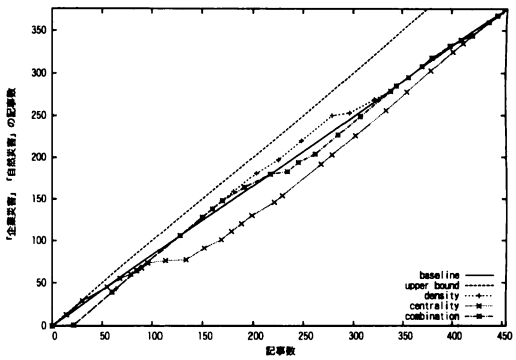

図 2　クラスタ選定の評価実験結果（2000 年, PHC）

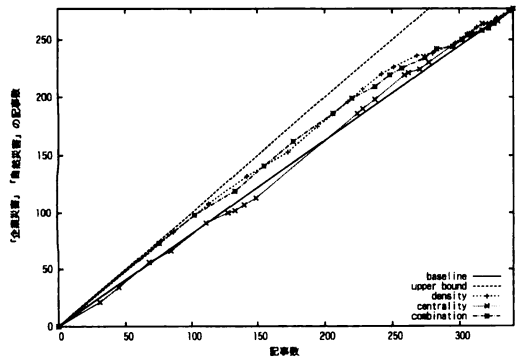

図 3　クラスタ選定の評価実験結果（2003 年, AHC）

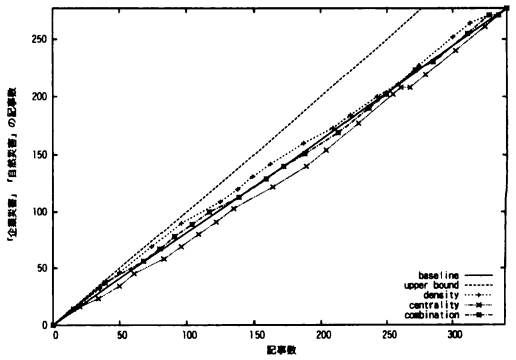

図 4　クラスタ選定の評価実験結果（2003 年, PHC）

の関連性を表す．この指標を用いることで話題性の高いクラスタを効率的に選別可能である．

2000 年と 2003 年の産業活動に伴う事件・事故・災害に関する新聞記事を対象として評価実験を行った．この実験結果より，TFIDF を利用した一般的な文書ベクトルに比べ，語の頻度，文字列長，出現位置を考慮した文書ベクトルが階層的クラスタリングの性能を向上させることが確認できた．さらに，密度と中心度をもとにクラスタの重要性を評価し，この重要度をもとに分析を行うことで分析に有効な文書を効率的に獲得可能であることを示した．

今後は，人手による分析結果と密度・中心度の関係性について深く分析，考察を行う必要がある．また，異なる分野の文書に対しても，提案手法が有効であるのかについても検証も行いたい．さらに，文書クラスタリング結果を分析者に効果的に示すためには，情報抽出や要約といった技術を用いて，クラスタが表す話題をクラスタ内の文書から抽出する手法についての検討も必要である．

文　献

[1] 堀井：“安全安心のための社会技術”, 東京大学出版会 (2006).

[2] 奥田, 川島, 佐藤, 宮原, 定方：“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, pp. 22–25 (2006).

[3] 佐藤, 川島, 佐々木, 奥：“時系列ニュース記事における最新話題語抽出方法”, 情報処理学会自然言語処理研究会 (2005-NL-168), pp. 1–6 (2005).

[4] 橋本, 村上, 乾, 内海, 石川：“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”, 社会技術研究論文集, **5**, pp. 216–226 (2008).

[5] M. Callon, J. P. Courtial and F. Laville: “Co-word analysis as a tool for describing the netword of interactions between basic and technological research: The case of polymer chemistry”, Sientometrics, **22**, pp. 155–205 (1991).

[6] 藤垣, 平川, 富澤, 調, 林, 牧野：“研究評価・科学論のための科学計量学入門”, 丸善株式会社 (2004).

[7] Y. Zhao and G. Karypis: “Hierarchical clustering algorithms for document datasets”, Data Mining and Knowledge Discovery, **10**, 2, pp. 141–168 (2005).

[8] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda and M. Asahara: “Japanese morphological analyzer chasen users manual version 2.0.”, Technical Report Technical Report NAIST-IS-TR990123, Nara Institute of Science and Technology (1999).

[9] S. M. Savaresi and D. L. Boley: “On the performance of bisecting k-means and PDDP”, First SIAM International Conference on Data Mining, pp. 1–14 (2001).

[10] Y. Zhao and G. Karypis: “Criterion function for document clustering”, Technical report, Department of Computer Science, University of Minnesota, Minneapolis, MN 55455 (2003).

[11] G. Karypis: “CLUTO A Clustering Toolkit Release2.1.1”, University of Minnesota, Department of Computer Sciense (2003). http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview.