１Ｂ１７

テキストマイニングによる社会課題及びその解決に結びつく科学技術に関する有用知識の抽出

○内海和夫，乾孝司，橋本泰一（東京工業大学）
村上浩司（奈良先端科学技術大学院大学），石川正道（東京工業大学）

1. はじめに

社会の問題が複雑化するなかで，大量のテキスト情報から科学技術や社会制度による社会課題への対策，あるいは対策により新たに生じる課題等を構造化して分析する手法の構築は，合理的な科学技術の研究開発推進や政策立案のために重要な方法論的課題である．分析の対象となる文書を特徴づけるキーワードの共出現（共起）の関係を一般化したアクター・ネットワーク理論[1)]は，このような分析手法の考え方として有用と考えられ，言語の共起関係の分析（共語分析）により，文書からの連続性のある課題概念の抽出や，課題とそれを解決する対策・手段等の関係性等を把握できる可能性があるが，文書へのキーワードの付与が人手によること，適用対象が科学技術文献に限られることなど，制約が多く，その適用例は限られたものとなっている．

科学技術情報をはじめとするさまざまな社会事象情報を含む新聞記事情報に対して共語分析を適用し，社会課題とその対策としての科学技術との関係を明らかにしようとする場合，新聞記事には科学技術文献のような関連文献情報（引用情報）やキーワードが付与されていないため，①特に社会課題と対策の関係等に対する俯瞰的な分析（＝予めキーワードが想定されていない場合の分析）に必要な共通トピックを持った記事集合を自動的に形成することが難しい，②キーワードそのものがないので共語分析が困難である，といった問題点が存在する．①の問題点については，我々は既に大規模な新聞記事情報から，階層的クラスタリング等の手法により共通トピックを持つ記事集合（クラスタ）を自動的にまとめあげる手法を確立しており，社会課題と関連を持つような特定のトピックで特徴付けられる記事集合を形成することができる[2)]．

本報告では，②の問題点を解決するためのテキストマイニング手法を提示し，新聞記事情報に対する共語分析を実現する．具体的には，上記のような共通トピック（今回は「がん」と「生活習慣病」）で特徴付けられる記事集合を形成し，そこに含まれる社会課題とそれを解決する技術的対策を表現するキーワード（以下，技術的対策用語）を自動抽出する手法を提案する．さらにその技術的対策用語を用いて行った共語分析結果についても述べ，大量の新聞記事情報から社会課題関連トピックで特徴付けられるクラスタの自動形成，及びそのクラスタから共語分析に活用できる技術的対策用語の自動抽出を一貫して行えるテキストマイニング手法の有用性を検証する．

2. 提案手法

2.1 分析用記事集合の作成

2.1.1 俯瞰的分析によるクラスタの特定

社会課題の解決に関連する技術的対策用語を抽出するために，まず俯瞰的分析により社会課題情報を多く含むクラスタを形成する．分析対象となるデータセットは，日本経済新聞記事データベースより日経シソーラスの中から医療に関わる検索語 316 を選定して検索することにより得られた（日経新聞本紙 2005 年，8,890 記事）．さらに，このデータセットに対してクラスタリングを行い，記事に含まれる単語の出現頻度・位置・文字数等を数値化したものを要素にもつ記事ベクトルの類似性から 200 の記事クラスタを形成した[2), 3)]．各クラスタには各単語の要素に基づくスコアリングにより要約キーワード（スコアの上位 6 単語）が付されているが，要約キーワードに「がん」あるいは「生活習慣病」に関連する用語が含まれるクラスタ数は，それぞれ 17，6 であった．このように，俯瞰的分析により分析対象となるトピック情報を含むクラスタを絞り込み，技術的対策用語の抽出を行う．

2.1.2 自動抽出の基準となる用語のマニュアル抽出

本研究では，特に技術的対策用語の抽出に関する評価実験を行うことから，前項で絞り込まれたクラスタから，技術的対策の内容のみを含む「がん」及び「生活習慣病」に関するクラスタをマニュアルで作成した（記事数はそれぞれ 83 件，34 件）．また，同クラスタから専門家による技術的対策用語の抽出を行い（専門家が抽出する用語には，純粋に技術的な用語だけでなく，その適用先に関する用語も含

まれる），自動抽出結果との比較を行うとともに，抽出された用語を含む文から後述する言語パターンの抽出を行い，用語抽出に活用した．

2.2 技術的対策用語の自動抽出

抽出しようとする技術的対策用語は，次の２つの要件を満たすことが望ましい．

①社会課題と強い関係性を有する

②技術と強い関係性を有する

そこで，①，②に対してそれぞれ課題関連度と技術関連度という指標を導入し，これら指標の積算値にしたがって用語候補を順序付け，順序の上位に位置する用語候補を技術的対策用語として抽出する．各記事における抽出用語数は，記事の文数（最大 10）とした．

2.2.1 課題関連度

社会課題と関連するトピックで特徴付けられる記事集合（クラスタ）から，課題関連度の強い用語を抽出する処理は，クラスタから特徴的な用語を抽出するクラスタ・ラベリングとほぼ等価な処理であると考えて差し支えない．そこで，クラスタ・ラベリングで適用される基本的な指標であるカイ２乗値[4)]を，課題関連度を測る指標として採用する．

2.2.2 技術関連度

技術関連度の指標を設定するに当たり，新聞記事の表現法の特徴を踏まえ，技術的内容を含む文（「技術表現文」と呼ぶ）の抽出法を構築した．例えば，新技術の開発や研究についての話題記事の場合，まず冒頭に標準的な文型として「(誰，どこ）が，(どのような技術）を開発した」という簡潔な文があり，そのあとに内容説明や背景の文が続いていく．新聞記事の技術表現文では固有の述語パターンが用いられることが多く（表 1 参照），そのパターンに着目することにより技術以外の文との識別を行うことができる．そこで，前述したクラスタより，技術表現文で使用されることが多い約 50 種類の言語パターンを抽出・設定し，分析対象の記事集合からその言語パターンとのマッチングによる技術表現文の抽出を行い，この文との位置関係に基づいた指標を定義する．この指標は，基本的には次式のように用語候補 t が技術表現文中のパターンから離れるにしたがい小さな値をとる[5)]．

〔技術関連度の相対距離関数部分〕：$\exp^{(-0.77\times \mathrm{relative\_position(t)})}$

〔relative_position(t)：用語候補 t が出現する文から，t の最も近くにあるパターンを含む文へ移動する際にまたぐ文の数（t とパターンとの文間相対距離)〕

表１　「がん」の記事集合で設定された言語パターン例（出現頻度上位 20 パターン）

・～を開発する，組み合わせる，確認する，利用する，発見する，導入する，応用する，研究する，突き止める，解明する，活用する，共同開発する，提供する
・～の開発を進める，開発に取り組む，開発につながる，研究を進める，研究を始める，技術を確立する，技術を使う

2.3 用語抽出手法の評価

2.1 で述べた「がん」及び「生活習慣病」の技術的対策の内容を含むクラスタ対して，上述の手法により抽出した技術的対策用語と，同一データから専門家が抽出した用語との重なり度合いを求めた[5)]．

表２　用語抽出手法の評価結果

手法	がん	生活習慣病
提案手法	0.532	0.608
tfidf 法（ベースライン法）	0.470	0.509

表２の値は，次式で定義される再現率と適合率の調和平均であり，抽出結果が専門家の結果と過不足なく一致した場合に限り１となり，重なり度合いが減るにしたがい０に近づく．

再現率＝| A ∩ B |／| A |　，　適合率＝| A ∩ B |／| B |

Ａ＝専門家によって抽出された用語の集合

Ｂ＝提案手法によって抽出された用語の集合

また，表中の tfidf 法とは，テキストから特徴的な語を抽出する際に汎用的に用いられる指標であり，提案手法との性能を比較するために結果を掲載した．表の結果から，新聞記事から技術的対策用語を抽

出する場合，本研究の提案手法は tfidf 法よりも有効に働くことが確認された．

自動抽出された技術的対策用語の例を表３に示す．赤字は専門家により抽出された用語であるが，黒字の用語が存在するのは，自動抽出では技術的用語と関係の深い主体用語や非技術的用語なども抽出されるためである．

表３　自動抽出された技術的対策用語例

がん		生活習慣病	
記事番号	抽出された用語	記事番号	抽出された用語
1	遺伝子科学，がん，生命科学研究センター，境界領域分野，学部間，共同研究，境界領域	1	テーラーメード食事療法，遺伝子診断，体重，脂肪，糖尿病，脂肪細胞，ホルモン，基礎代謝量，合併症，人
2	副作用，診断法，血液検査，肺がん治療薬，患者，精度	2	生活習慣病，自動判定，健康診断，危険度，数値，血糖値，健康科学センター，血圧
3	固形がん，細胞，酵素，治療薬開発，たんぱく質，ケア研，働き，抗がん薬	3	心臓病，動画，診断精度向上，冠状動脈，患者，放医研･東芝メディカル，血流，心筋梗塞
4	がん治療法，薬物送達システム，DDS，中性子線照射，ホウ素製剤，がん細胞，患部，中性子用加速器，集中的	4	膵島細胞，糖尿病，島移植手術，血糖値，脳死者，生体移植，すい臓，ｲﾝｽﾘﾝ，生体，健康
5	局所療法，ラジオ波，がん細胞，患者，治療法，実力病院，肝がん治療法，治療成績，普及，技術水準確保，がん治療	5	脂肪細胞，動脈硬化，アディポネクチン，メタボリックシンドローム，オスモチン，飽和脂肪酸，天然物質，薬，症状改善

（注）赤字：専門家により抽出された技術的対策用語

3. 自動抽出された技術的対策用語の共語分析結果

本研究では，自動抽出された技術的対策用語に対し，共語分析で用いられる代表的な４つの指標（Jaccard 指標，同等性指標，包含指標，近接指標）[6] を計算して用語間の関係を分析し，後述するようなマップを作成したが，このうち図１，２は次式で示される同等性指標を用いて「がん」及び「生活習慣病」に関する技術的対策用語を連結しマッピングしたものである．

同等性指標　$E_{ij} = C_{ij}^2 / (C_i * C_j)$

〔C_i：語 t_i を有する記事数，C_{ij}：語 t_i と t_j の両方を有する記事数〕

半径方向の軸は用語の出現記事数を示しており，エッジの黒線，赤線は図の凡例に示すような基準で決められた．また，用語に連結するエッジ数が５または４以上のものを赤字にした．なお，各用語の相対的な位置関係には意味がなく，見やすくなるように配置されている．

同等性指標は，共起している２語の共起率（各語の出現頻度に対する共起数の比率）が共に大きい場合に大きな値となるが，マップ上では適当な閾値以上になったときにエッジで連結される．図に示すように，連結されている用語は大きな集合を形成しており，例えば図１の場合は３つのまとまりがあり，それぞれの集合の上位（円の中心方向）に「治療法」，「抗がん剤」，「血液，たんぱく質」といった用語があることから，各集合はがんの治療，創薬，診断に関わる３分野と見ることができる．同様に図２では，心臓病，糖尿病，生活習慣病，血管系疾患といった病気に関わる４分野のまとまりを見ることができる．これは，共起語マップ上で類似性（連続性）をもつ概念を表す用語のまとまりを階層構造的に見ることができることを示唆している．また，各分野の上位語のほとんどが赤字となっているが，エッジ数が多いということはその用語と連結した関連語のネットワークがいくつも重なっていることを意味する．例えば，図２の「糖尿病」の場合，糖尿病の治療，診断，予防，患者等に関連する用語のネットワークが連結し，「糖尿病」という用語がそれらのノード（結節点）となっていることを示しており，糖尿病を克服するための各種の対策が同レベルで位置づけられていると見ることができる．

同等性指標が大きい場合（図１の場合は 0.25 以上，図２の場合は 0.3 以上）にエッジを赤線で示したが，これらの用語の組み合わせは関係性が深く，特に周辺部にある語の組み合わせは複合語的な見方で扱ってもよいと考えられる．以上のほかに，指標の閾値の条件が同程度の場合に Jaccard 指標のマップよりも同等性指標のマップのほうが中心部のエッジが少なくなること，近接指標のマップは低出現頻度の用語の関係性を見るときに有用で，マップ上では円の周辺部に位置づけられる用語が強調されること，専門家による抽出結果と自動抽出結果と比較した場合，前者のほうが抽出用語数が少なく後者の部分集合的になること，などが把握された．

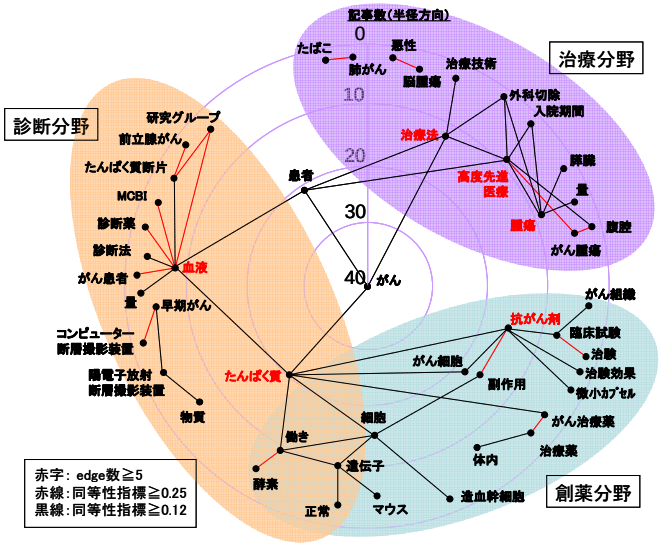

図１　自動抽出された「がん」に関する
技術的対策用語の共起語マップ
（同等性指標による連結）

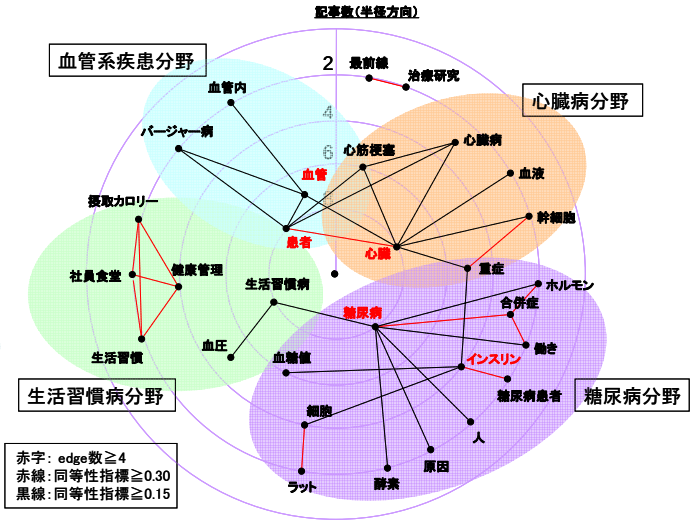

図２　自動抽出された「生活習慣病」に関する
技術的対策用語の共起語マップ
（同等性指標による連結）

4. 結語

本研究では，社会事象に関する情報を大量に含む新聞記事から社会課題とそれを解決するための技術的対策情報を抽出し構造化するために，科学技術分野でよく用いられる共語分析を新聞記事に適用することを目的として，特に共通的な社会課題関連情報を多く含むクラスタを自動形成し，そこから共語分析に必要なキーワード（今回は特に技術的対策用語）を自動抽出するためのテキストマイニング手法を開発し，その適用可能性を検討した．新聞記事情報への共語分析の適用研究としてはまだ端緒についたばかりであり，実用に供するためには，特に抽出精度向上のための問題点（例えば言語パターンにおけるパターンの設定方法，技術的対策用語の名寄せの方法等）を解決していく必要があるが，今後は，クラスタリングにより形成されたクラスタに用語抽出を適用してさらに適用可能性を検証するとともに，医療以外の分野への適用や技術以外の対策用語抽出の適用などを通して適用事例を増しつつ問題点の解消を図っていく必要がある．

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」の支援のもとに実施した．

参考文献

1) Callon, M. 1983: “From translations to problematic networks: an introduction to co-word analysis,” *Social Science Information*, 22(2), 191-235

2) 橋本泰一，村上浩司，乾孝司，内海和夫，石川正道 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」『社会技術研究論文集』5, 216-226

3) 内海和夫，乾孝司，村上浩司，橋本泰一，石川正道 2007：「大規模テキストマイニングによる医療分野の社会課題・技術トレンド抽出」『研究・技術計画学会 第 22 回年次学術大会 講演要旨集』684-687

4) Christopher, D. 2008: “Introduction to Information Retrieval,” *Cambridge University Press*, chapter 17.7

5) 乾孝司，内海和夫，橋本泰一，村上浩司，石川正道 2008：「新聞記事からの社会課題に対する技術的対策情報の抽出」『第 7 回情報科学技術フォーラム 講演論文集第 2 分冊』169-170

6) 藤垣裕子，平川秀幸，富澤宏之，調麻佐志，林隆之，牧野淳一郎 2004：「第 10 章 語の分析，共語分析，共分類分析」『科学計量学入門』丸善