社会課題と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（東京工業大学），乾孝司（東京工業大学），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，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，合理的な科学技術社会論の実装にむけて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るが，文書へのキーワードの付与が人手によること，対象が科学技術文献に限られるなど，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，言語の共出現（共起）関係を分析する共語分析を適用する場合，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俯瞰的に社会課題と対策の関係等を把握したい場合（＝予めキーワードが想定されていない場合）に，分析対象となる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とともに，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る．

2．技術的対策用語の抽出

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．

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乗値を，課題関連度を測る指標として採用する．

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，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[3)]．

$$技術関連度＝\exp^{(-0.77\times relative\_position(t))}$$

〔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

日本経済新聞記事データ（本紙，2005 年）より形成したクラスタ[4)]から，「がん」及び「生活習慣病」の技術的対策の内容を含む記事集合を作成し（記事数はそれぞれ 83 件，34 件），これらに対して，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人手で抽出した用語との重なり度合いを求めた[3)]．

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．

$$再現率＝|A\cap B|／|A|\quad，\quad 適合率＝|A\cap B|／|B|$$

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．

3. 自動抽出された技術的対策用語の共語分析結果

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指標のうち，次の式で示される同等性指標及び近接指標を用いて分析を行った[5)]．

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

$$近接指標\quad P_{ij}＝(C_{ij}*N)／(C_i*C_j)$$

〔C_i：語M_iを有する全記事数，C_{ij}：語M_iとM_jの両方を有する記事数，N：全記事数〕

図１，２はそれぞれ同等性指標及び近接指標を用いて連結した「がん」に関する技術的対策用語をマッピングしたものである[8)]．半径方向は各用語の出現頻度（記事ベース）を示し

ており，エッジの黒線，赤線は各図の凡例に示すような基準で設定された．また，用語に連結するエッジ数が５を超えるものを赤字にした．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，共出現率（各語の出現頻度に対する共出現〔＝共起〕数の比率）が大きいほど大きな値となり，マップ上では適当な閾値以上になるとエッジで連結される．図１に示すように，関係性の強い語が連結されて大きな集合を形成しており，大きく治療，診断，創薬の 3 分野に分けることができる．また，各分野の中で，統合的な意味をもつ用語は上位（中心方向）に位置づけられており，赤字となっている「たんぱく質」,「血液」,「抗がん剤」,「高度先進医療」などは，がんの技術的対策における統合的用語である．また，この記事集合では，がん対策としてたんぱく質を利用した診断や抗がん剤開発，あるいはがん治療用の高度先進医療が注目されていることが把握できる．

近接指標は，低出現頻度の用語の中で関係性の強いものを見るときに利用されるが，図 2 は図 1 のマップの周辺部が強調された形となっており，図 1 にはない用語も新たに加わっている（図中斜字の用語）．純粋に技術的な用語は周辺部に多く，統合的な用語の具体例あるいは手段といった位置付けになっている．この結果から，近接指標は出現頻度が一般的に少ない個別技術の構造化に有用であると考えられる．

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）を抽出するためのテキストマイニング手法を開発し，その有用性を確認した．今後は，技術以外の対策用語，あるいは医療以外の分野への展開などにより適用事例を増やすとともに，対策用語と社会課題あるいは主体との関係性の分析を深耕し，社会と科学技術の関係の構造化の精度を高めていく必要がある．

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) 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

4) 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

5) 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

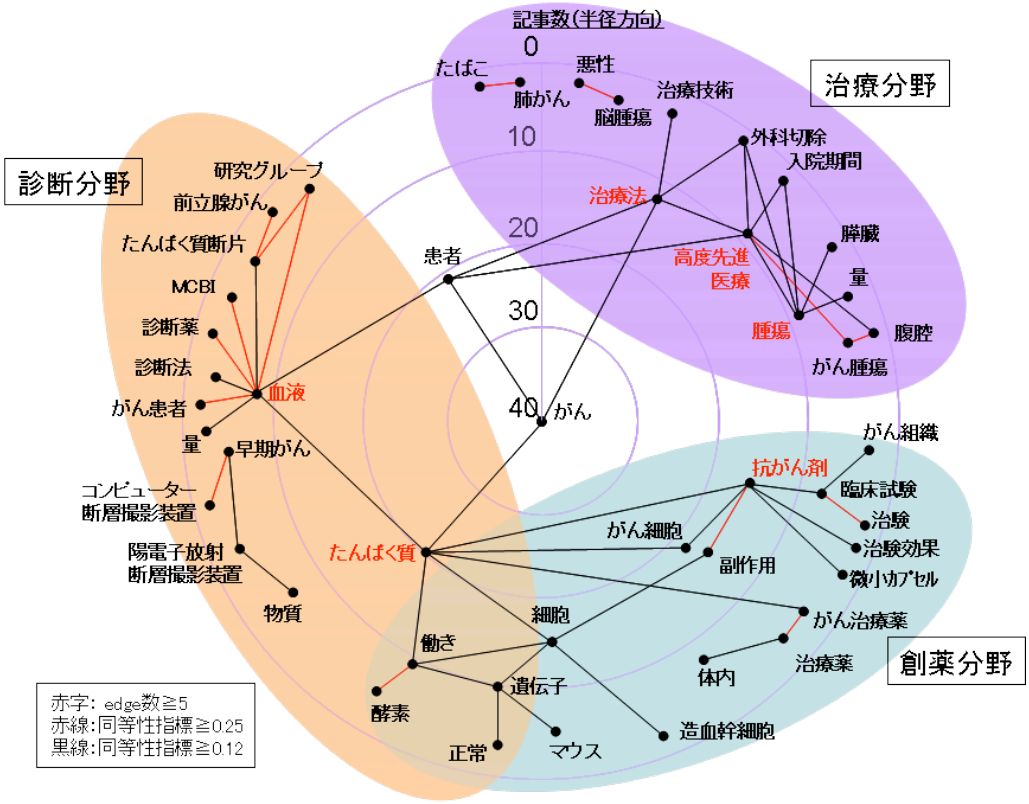

図１　自動抽出された技術的対策用語の共起語マップ（同等性指標による連結）

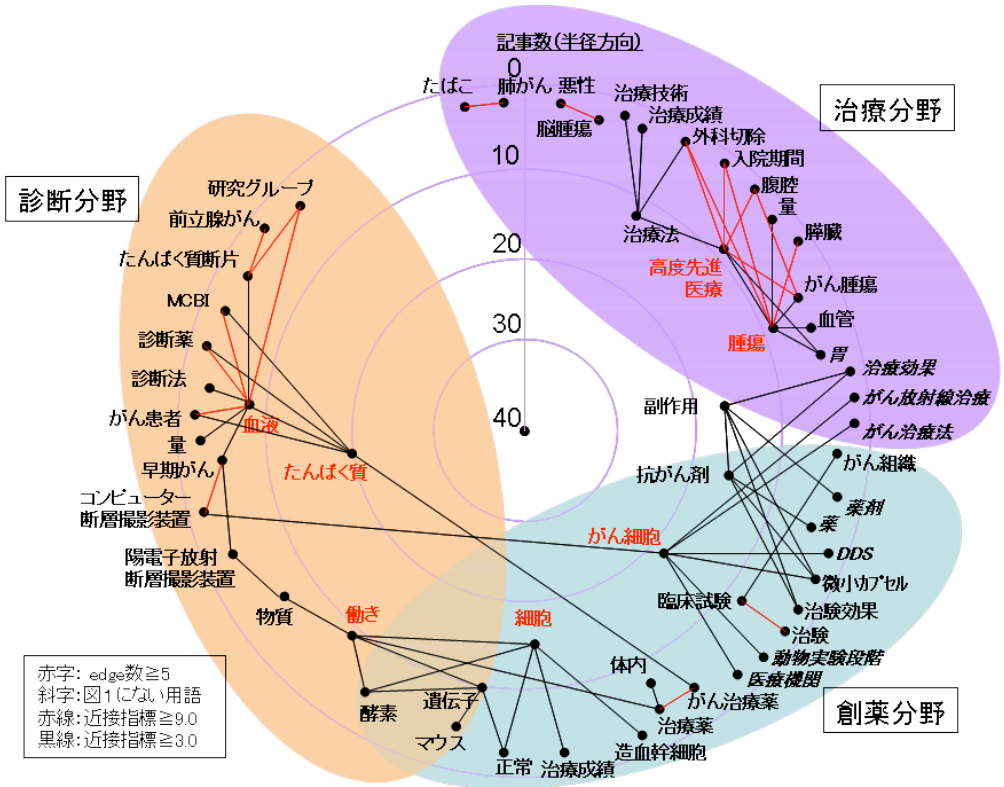

図２　自動抽出された技術的対策用語の共起語マップ（近接指標による連結）