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

EXTRACTION OF CRITICAL KNOWLEDGE CONCERNING SOCIAL PROBLEMS AND THEIR TECHNOLOGICAL SOLUTIONS

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

[1]工学修士　東京工業大学統合研究院　(E-mail:utsumi@iri.titech.ac.jp)
[2]博士（工学）　東京工業大学統合研究院　(E-mail:inui@iri.titech.ac.jp)
[3]博士（工学）　東京工業大学統合研究院　(E-mail:hashimoto@iri.titech.ac.jp)
[4]博士（工学）　奈良先端科学技術大学院大学　(E-mail: kmurakami@is.naist.jp)
[5]工学博士　東京工業大学統合研究院　(E-mail:ishikawa@iri.titech.ac.jp)

社会課題に関する情報を多く含む新聞記事を対象として，社会課題とそれを解決するために注目されている有用な技術的対策知識を抽出する新たなテキストマイニングの手法を開発した．この手法では，あらかじめ特定のキーワードを設定することなく，俯瞰的なアプローチによって同一課題を扱う記事クラスタを形成し，当該課題に解決をもたらす技術的な用語に対して新たに導入した課題関連度及び技術関連度によりスコアリングを行い，記事毎にキーワードを自動的に付与することを可能とした．本手法の応用として，医療課題（がん及び生活習慣病）に対して技術的対策用語を抽出し，Jaccard 指標，同等性指標及び近接指標を用いた共語分析を行うことによって，本手法によって抽出された情報の意義について検証した．

キーワード：俯瞰的アプローチ，共語分析，情報抽出，文書クラスタリング，言語パターン

1.　緒言

社会の問題が高度に複雑化するなかで，社会事象の情報を大量に含むテキストデータから，注目すべき社会課題や，その対策としての科学技術情報などを効率的に抽出・構造化し分析する手法の構築が必要である．

新聞記事は，社会課題に関する情報を多く含み，かつ社会一般が認知する視点での事象が多く掲載されているため，社会課題の分析に適している．実際の分析では，目的の社会課題に関する記事に含まれているようなキーワードを使って検索し，必要な記事を獲得し分析をする「検索的アプローチ」が一般的である．これに対して「俯瞰的アプローチ」，すなわち予め社会課題を設定せず，さまざまな社会課題や対策技術などの相互関係や全体的傾向について分析をするアプローチが注目されるようになった[1)]．俯瞰的アプローチにおいては，新聞記事から分析に必要な情報を自動的に抽出するテキストマイニング技術が重要である．新聞記事への俯瞰的アプローチの適用事例は，従来は記事の話題の効率的把握[2)]などを目的としたものが主であったが，俯瞰的に社会課題発見を試みる研究に発展しつつある[3)]．

新聞記事から社会課題やその対策技術などの情報を獲得し構造化する手法として共語分析が有効である．しかしながら，共語分析には予め分析に必要なキーワードを付与することが必要である．従来，共語分析は科学技術文献を中心に適用され，その有効性が検討されてきた[4)，5)]．学術文献のキーワードは通常人手により付与されるため，「付与者の認識に影響され，偏った分析結果が得られる」，「キーワードを付与するコストがかかる」という問題がある．新聞記事に対する共語分析についても報告されているが[6)]，分析のためのキーワードの一部は研究者により設定され，キーワード付与の問題は依然として解決していない．

我々は，俯瞰的アプローチに基づき，社会課題に関連した新聞記事から技術，制度，サービスに関する対策用語を自動的に抽出し構造化することにより，複雑に関係しあった社会課題と対策の関係を用語のネットワークとして視覚化し，鳥瞰できるようにする手法の開発を進めている．本論文では特に技術的対策用語に着目し，社会課題との関連性及び技術との関連性の２つの観点から，ある語が技術的対策用語であるかどうかを判定し抽出する手法について報告する．また，この抽出手法の有効性を検証するために，医療分野（「がん」及び「生活習慣病」）に関する新聞記事を用いて技術的対策用語の抽出を行った．さらに，自動抽出された技術的対策用語を用いて共語分析を行い，提案手法の意義と共語分析への適用可能

Table 1　対策クラスタと主要な自動形成クラスタとの対応関係

対策ｸﾗｽﾀ	ｸﾗｽﾀ番号	対策記事数	対策抽出率	要約キーワード
がん	23	6	27.3%	がん細胞，がん治療法，開発，エックス線，次世代，ミニ公募債
	25	5	21.7%	静岡県静岡がんセンター，開発，早期発見，同センター，ビジネス便，事前申込み
	41	9	14.8%	がん，胃がん，日本人，たばこ，病気，予防学
	54	13	10.3%	抗がん剤，肺がん治療薬，副作用，医薬品，臨床試験，一般名
	82	15	50.0%	がん，しゅよう，がん腫瘍，前立腺がん，患者，検査結果
	115	9	13.4%	コンピューター断層撮影装置，磁気共鳴画像装置，早期発見， 陽電子放射断層撮影装置，PET，医療関連ベンチャー
	116	4	11.4%	がん，がん治療，患者，医療専門誌，実力病院，全国調査
	138	4	16.7%	システム開発，ﾃﾞｰﾀﾍﾞｰｽ，ﾌﾟﾛﾌｧｲﾙ，開発，ネズミ，ベンチャー企業
生活 習慣病	58	7	14.6%	心臓病，心臓病治療，補助人工心臓，ｶﾃｰﾃﾙ，医療用細管，新規公開株
	145	29	29.9%	生活習慣病，生活習慣病予防，糖尿病，コレステロール，ダイエット，医食同源

性についても検証する．

2.　提案手法の内容

本研究で提案する手法は，大量の新聞記事から用語の類似性に基づいて記事クラスタを形成し，そこから記事クラスタを特徴付ける共通トピックとの関連性と，技術表現を含む文との位置関係からそれぞれ定義される課題関連度及び技術関連度により，共語分析に用いるキーワードを自動的に抽出する一連の手続きからなる．本提案手法では，課題用語と技術用語の両者の重要度をそれぞれの観点において独立に計量化する戦略をとった．これらの指標の積によって重み付けされたキーワードは，その課題及び技術関連度の度合いに応じて抽出することが可能となるため，情報抽出の精度が向上することが期待されるからである．また本手法は，一貫して予め特定のキーワードを設定することなく情報を得ることを可能とすることから，俯瞰的情報抽出と呼ぶこととする．

2.1.　分析・評価用クラスタ及び用語の作成

（1）記事クラスタの作成

a. 記事の自動クラスタリングとクラスタ選定

新聞記事データから社会課題の抽出を行うために，まず俯瞰的アプローチにより複数のトピックを含む記事群からトピック別のクラスタに記事を分類した．分析対象となる記事群は，日本経済新聞記事データベースより日経シソーラス[7]の中から医療に関わる検索語316を選定し，2005年の日経新聞本紙を検索することにより医療関連記事8,890件を得た．さらにこの記事群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・文字数等を指標化したものを要素にもつ記事ベクトルの類似性から200個の記事クラスタを形成した[3), 8)]．各クラスタには各単語の記事内での重要性を表すウェイトによってスコアリングした要約キーワード（スコアの上位6単語）を付与しているが，この要約キーワードを参照しながら，分析対象となるトピックを抽出した．この結果，要約キーワードに「がん」あるいは「生活習慣病」に関連する単語が含まれるクラスタが，それぞれ17個，6個あり，これらのトピックは注目度の高い医療課題と判断した．

b. 用語抽出用「対策クラスタ」の作成

提案手法による用語抽出には，クラスタリング法に起因する誤差と用語抽出法による誤差が存在するが，抽出結果から各誤差の程度を分離して評価することは難しい．本研究では，特に技術的対策用語の自動抽出法の提案とその評価に主眼を置いていることから，クラスタリング法による誤差を排除することを目的として，前項で作成されたクラスタから「がん」及び「生活習慣病」に関する技術的対策の内容のみを含む記事を人手にて抽出し，トピック単位で分析するために1つのクラスタにまとめた（記事数はそれぞれ83件，34件であった．以下「対策クラスタ」と呼ぶ）．Table 1には，これらの技術的対策記事が，自動形成されたクラスタにどの程度含まれているかを示した．表中の対策抽出率は，各クラスタの全記事数に対する技術的対策記事数の比率である．なお，対策クラスタ以外に対しても対策記事抽出を行ったが，医療関連記事全体での対策抽出率は10.1%であった．

（2）専門家による基準用語の作成

次に，専門家による対策クラスタからの技術的対策用語抽出を行い，自動抽出結果との比較に用いる参照データを作成した．専門家抽出では，原則として次のような基準を用いた．1) 対象記事の内容を特徴付ける技術用語を最優先で抽出する．2) 技術の適用対象に関する用語も

抽出する. 3) 上位概念と下位概念に位置づけられる技術用語が同じ記事に出ている場合は，より具体的な下位概念の用語を優先的に抽出する. 4) ２語以上がまとまって一つの意味をなしている用語はまとめて抽出する. 5) 従来技術と比較している場合は，従来技術は抽出しない. 6) 物質名，商品名のような固有名詞は抽出しない.

なお，専門家抽出用語の情報は，自動抽出用語との比較評価だけでなく，後述する技術的内容を含む文の抽出に使われる言語パターンの抽出や，このパターンと自動抽出用語との位置関係から用語を絞り込むためのルール作成にも反映した.

2.2. 技術的対策用語の自動抽出

抽出対象となる対策クラスタに含まれる各記事は，当該課題に関する内容が記述されており，これらの記事から抽出される技術用語は，課題に対応する技術であることが期待される. そこで, 課題関連度（problem relevancy : *pr*）と技術関連度（technical relevancy : *tr*）という２つの指標を定義し，この２つの指標に基づき，語の技術的対策用語らしさを測ることとする[9].

関連する先行研究として, 専門分野に関連の強い語（専門用語）を自動抽出する研究がある[10]. 技術的対策用語と専門用語は重複する部分もあるが, 異なる概念であり, 先行研究の手法をそのまま適用するだけでは不充分であるため，独自の指標を開発した.

（1）課題関連度

ある共通した社会課題に関連した記事で構成されるクラスタが与えられたとする. クラスタ内の記事に共通した社会課題との関連性が高い用語とは，クラスタを特徴付ける語であるといえる. ここで，クラスタを特徴付ける語とは，当該クラスタに含まれる記事にはよく出現するが，その他のクラスタにはあまり出現しない語と考える. ある文書集合において，語が特異的に偏って出現するかどうかを判定する指標として，χ^2 値が有効であることが知られている[11]. 語 t に関する課題関連度 は，Table 2 に示すような 4 種類の出現頻度を用いて，式（1）により計算される.

$$\mathrm{pr}(t) = \chi^2(t) = \frac{(a+b+c+d)(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} \qquad (1)$$

本研究の場合，表中の「クラスタ内」とは対策クラスタ内の記事,「クラスタ外」とは対策クラスタには含まれない記事を指す.

（2）技術関連度

a. 言語パターンマッチングによる技術表現文の抽出

新聞記事では，技術的内容を含む文は特徴的なパターンで表現されることが多い. 例えば，一般的な記事構成では，まず冒頭の文で「（人，企業）が〔主語（主体）〕，どのような技術を〔目的語〕，どうした〔述語〕.」という要点が簡潔に記述され，後段の文でその内容の詳細や背景などについて記述される. この時，技術用語は各文の目的語として記述され，さらに述語に特徴的な動詞をともないやすい（これを係り受け関係という）. 特に「開発する」やその変形（「開発を進める」,「開発に取り組む」など）が多く使われる. 本研究では，技術的内容を含む文で使用されることが多い約 50 種類の言語パターンを人手により選定・作成し，その言語パターンにマッチする文を技術表現文とした.

Table 3 に「がん」及び「生活習慣病」に関する対策クラスタからの技術表現文を抽出する際に用いた言語パターンの例を示す. パターン内の「→」は，係り受け関係を示す. 技術表現文に含まれ，その中の述語と係り受けの関係をもつ名詞は技術用語となる可能性が強く，技術

Table 2　課題関連度算出用の変数

区分	クラスタ内	クラスタ外
t の出現頻度	*a*	*b*
t 以外の語の出現頻度	*c*	*d*

Table 3　技術表現文の抽出に適用された言語パターン例

がん	
～を→開発する（264）	～の研究を→進める（28）
確認する（103）	開発を→進める（26）
組み合わせる（95）	技術を→使う（24）
導入する（74）	開発に→取り組む（18）
発見する（74）	技術を→確立する（18）
利用する（68）	研究を→始める（15）
突き止める（49）	開発に→つながる（14）
応用する（47）	開発を→目指す（13）
解明する（34）	開発に→乗り出す（10）
研究する（33）	研究に→乗り出す（10）
提供する（23）	
共同開発する（19）	
活用する（19）	
共同研究する（10）	
事業化する（10）	
生活習慣病	
～を→開発する（110）	～の開発に→つながる（18）
確認する（47）	開発を→目指す（15）
突き止める（19）	開発を→進める（10）
応用する（19）	実験に→使う（10）
解明する（19）	
研究する（14）	

〔カッコ内はパターンが含まれていた記事数（≧10）〕

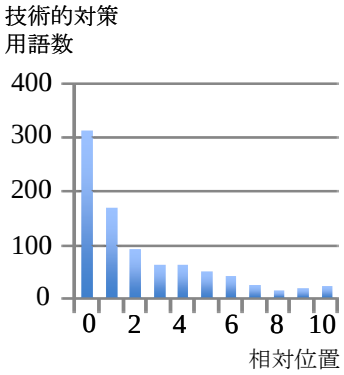

Fig. 1　相対位置と技術的対策用語数の関係

関連度が強いといえる．また，記事内において技術表現文の周辺の文脈では，主に技術に関する内容について述べることが多く，技術的対策用語を含む文が多い（Fig. 1 参照）．これから，技術表現文の近くに出現する用語についても技術関連度が強いと仮定した．

以上に基づき，ある用語 t が，上記の技術表現文を含む記事内で複数回出現する可能性があることを考慮して，記事 d 内で j 番目に出現する用語を t_j^d とすると，技術表現文との文間の位置関係に基づいた技術関連度 $\mathrm{tr}_{\mathrm{inter}}(t_j^d)$ と，技術表現文の文内位置に基づいた技術関連度 $\mathrm{tr}_{\mathrm{intra}}(t_j^d)$ を次のように定義することができる．

b. **文間の位置関係に基づいた技術関連度**

専門家により抽出した技術的対策用語を含む文と技術表現文との位置関係を調査したところ，相対位置ごとの技術的対策用語の度数分布は Fig. 1 のようになった．ここで，相対位置は，技術表現文が技術的対策用語を含んでいる場合を 0，両者が隣接している場合を 1，間に 1 つ文を挟む場合を 2，以降同様に定めた．この結果に基づいて指数回帰式を導出し，定数項を無視すると，$y = e^{-0.77x}$ を得る．この関数形にしたがい，文間の位置関係に基づいた技術関連度 $\mathrm{tr}_{\mathrm{inter}}(t_j^d)$ を以下で定義する．

$$\mathrm{tr}_{\mathrm{inter}}(t_j^d) = \max_{p_k^d \in P^d} \exp(-0.77 \cdot \mathrm{r\_pos}(p_k^d, t_j^d)) \quad (2)$$

ただし，P^d は，d に含まれる技術表現文の集合，$\mathrm{r\_pos}(p_k^d, d_j^d)$ は，p_k^d から見た t_j^d を含む文への相対位置である．技術表現文をもたない記事は常に $\mathrm{tr}_{\mathrm{inter}}(t_j^d) = 1$ とする．

c. **文内位置に基づいた技術関連度**

技術表現文では，言語パターンの述語に対する目的語及び目的語への修飾要素に位置する用語は他より技術関連度が強くなると考えられる．そこで，文内位置に基づいた技術関連度 $\mathrm{tr}_{\mathrm{intra}}(t_j^d)$ を以下で定義する．

$$\mathrm{tr}_{\mathrm{intra}}(t_j^d) = \begin{cases} 100 & (t_j^d \text{が言語パターン内の目的語} \\ & \quad \text{または，その修飾要素}) \\ 1 & (\text{otherwise}) \end{cases} \quad (3)$$

d. **技術関連度の定義**

記事 d における用語 t の技術関連度を次式のとおり，それぞれの技術関連度の積で定義する．

$$\mathrm{tr}(t,d) = \sum_{t_j^d \in \mathrm{ins}(t,d)} \mathrm{tr}_{\mathrm{intra}}(t_j^d) \cdot \mathrm{tr}_{\mathrm{inter}}(t_j^d) \quad (4)$$

ここで，$\mathrm{ins}(t,d)$ は d 内に出現する t の実体の集合を示す．

(3) 技術的対策用語の抽出

以上の準備のもとに，記事 d 内の各用語 t に対して，式(5)のような課題関連度及び技術関連度の積でスコアを計算し，スコア上位の n 個を技術的対策用語として出力した．

$$\mathrm{score}(t,d) = \mathrm{pr}(t) \cdot \mathrm{tr}(t,d) \quad (5)$$

ここで，技術的対策用語数は記事の長さに依存し，記事の文数が 10 以下の状況では，文数と用語数の間には緩やかな比例関係が観察され，それ以上の文数では用語数はあまり増加しないことが分かった．そこで，記事ごとに抽出用語数を変化させ，記事 d における抽出用語数を $n_d = \min(|d|, 10)$ とした．$|d|$ は記事 d の文数である．

自動抽出された技術的対策用語の例（「がん」及び「生活習慣病」の対策クラスタの任意の 5 つの記事から抽出された用語）を Table 4 に示す．表では，専門家のみ，専門家・自動の双方，自動抽出のみで抽出された用語を区別して示した．専門家のみで抽出された用語の多くは，自動抽出における指標のスコアリングで大きな値を獲得できずに選に漏れたものである．自動抽出のみによる用語は，技術的対策用語との関係は深いものの，技術用語ではない主体を表す語，あるいは専門家の興味の対象にはならなかった一般語，類義語，上位概念を表す語などが多い．

3.　提案手法の評価

「がん」及び「生活習慣病」の対策クラスタから専門家が抽出した技術的対策用語と，同クラスタから提案手法により自動抽出された技術的対策用語を比較することで，提案手法の有効性を定量的に評価した．すなわち，専門家が抽出した用語を，正解となる正しい用語であると仮定し，この用語集合からの自動抽出結果の乖離度を評価した．

Table 4　抽出された技術的対策用語例

	専門家のみ抽出	専門家・自動の双方が抽出	自動のみ抽出
「がん」に関する技術的対策用語	医工連携，医療福祉器具，細胞組織，分光器	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
	タンパク質，予測	副作用，診断法，血液検査，肺がん治療薬	患者，精度
	RNA（リボ核酸）干渉，医薬品開発，抗がん新薬	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
	抗がん剤，死滅，副作用，微小カプセル	がん治療法，薬物送達システム，DDS，中性子照射，ホウ素製剤，がん細胞	患部，中性子加速器，集中的
	腹腔（ふくくう）鏡	高度先進医療，膀胱がん，腹腔鏡下リンパ節摘出，精巣腫瘍	がん腫瘍，患者，開腹手術，泌尿・生殖器，転移部分
「生活習慣病」に関する技術的対策用語	受容体，食事指導，肥満治療	テーラーメード食事療法，遺伝子治療	体重，脂肪，糖尿病，脂肪細胞，ホルモン，基礎代謝量，合併症，人
	―	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
	ｺﾝﾋﾟｭｰﾀｰ断層撮影装置，新型 CT，早期発見	心臓病，診断精度向上，冠状動脈，動画	患者，放医研・東芝メディカル，血流，心筋梗塞
	IC カードシステム	健康管理	摂取カロリー，社員食堂，社員証，血圧，体重，試験導入，生活習慣，カード決済，体脂肪
	たんぱく質，代謝症候群，予防・治療	脂肪細胞，動脈硬化，ｱﾃﾞｨﾎﾟﾈｸﾁﾝ，ﾒﾀﾎﾞﾘｯｸｼﾝﾄﾞﾛｰﾑ，ｵｽﾓﾁﾝ，飽和脂肪酸	天然物質，薬，症状改善

$$\text{Precision} = \frac{R}{N} \qquad \text{Recall} = \frac{R}{C}$$

$$\text{F}-\text{measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2R}{N + C} \qquad (6)$$

評価尺度には，F 値（F-measure）[12] を採用した．F 値とは，式(6)で定義される適合率（Precision）と再現率（Recall）との調和平均であり，自動抽出結果が専門家の抽出結果と過不足なく一致した場合に限り 1 となり，重なり度合いが減少し，乖離度が高くなるにつれ 0 に近づく．式(6)の中で，R は自動抽出した正しい技術的対策用語の数，N は自動抽出した技術的対策用語の数，C は専門家によって抽出された技術的対策用語の数を表す．

文書からの特徴的な用語抽出における汎用的な指標として *tfidf*（用語 *t* の文書中出現頻度（*tf*）と全文書数に対する用語 *t* の出現文書数比率の逆数に基づく指標（*idf*）の積）が知られている [12]．本研究では，提案手法との比較のために *tfidf* をベースラインのスコア計算法として採用した．

実験結果を Table 5 に示す．課題関連度と技術関連度の組み合わせからなる提案手法（1 行目）は，両指標の単独使用（2 行目と 3 行目）あるいは，*tfidf*（4 行目）のいずれよりも高い F 値を達成した．また，課題関連度，技術関連度の単独使用も *tfidf* と同等あるいはそれ以上の F 値を得た．これより，新聞記事から技術的対策用語を抽出する場合，提案手法は *tfidf* よりも有効に働くことが確認できる．今回の実験結果を見る限り，課題関連度と技術関連度は抽出精度向上に同程度寄与していると言える．

自動抽出においては，技術的対策用語の範囲の明確化（上位概念を表す語の扱い，技術と関連性の高い専門用語でない語の扱い，従来技術の扱い等），及び用語の名寄せの高精度化（例えば，「たんぱく質」，「タンパク質」，「蛋白質」等の同一化）や類義語の同一化（例えば，「抗がん剤」，「抗がん薬」，「がん治療薬」等の同一化）などに取り組む必要がある．また，専門家抽出用語と自動抽出用語の乖離の低減は，技術関連度で用いられている重みや記事別の抽出用語数を調整したり，専門家抽出のみの用語で用いられている言語パターンを追加したりすることなどにより実現できるが，再現率と適合率はトレードオフの関係にあるので，抽出の目的に応じて重みや言語パターンを最適化するような方策を検討する必要がある．

Table 5　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
pr	0.473	0.558
tr	0.487	0.512
tfidf 法（ベースライン法）	0.470	0.509

4.　俯瞰的情報抽出された技術的対策用語による共語分析

従来，共語分析はシソーラスが整備されている科学技術文献に対して用いられてきた．これに対して新聞記事については，用語の統一が困難であり，かつ大量の記事へのキーワード付与には膨大な人手がかかることから，ほとんど共語分析の対象となることはなかった．前節に

述べたとおり，提案手法により新聞記事から自動的に技術的対策用語を抽出することが可能となった. ここでは，得られたキーワードを用いて，医療課題に対する共語分析を行ったのでその結果について報告し，本手法の意義について明らかにする.

手順は次の通りである．まず各用語間の共起頻度（２つの用語が共起している記事数）を求め，共語分析で用いられる代表的な３つの共起指標（Jaccard 指標，同等性指標，近接指標）[13] を算出した．さらにこれらの指標を用いて可視化し，技術，応用，対策，課題といった性格を有する各用語間の構造を分析した．分析に用いた各指標の算出式を以下に示す．ただし C_i は用語 t_i が出現する記事数，C_{ij} は用語 t_i 及び t_j の両方が出現する記事数，N は対策クラスタの記事数である.

Jaccard 指標　$$J_{ij} = \frac{C_{ij}}{C_i + C_j - C_{ij}} \quad (7)$$

同等性指標　$$E_{ij} = \frac{C_{ij}^2}{C_i \cdot C_j} \quad (8)$$

近接指標　$$P_{ij} = \frac{C_{ij}}{C_i \cdot C_j} \cdot N \quad (9)$$

4.1. 「がん」関連技術的対策用語に対する共語分析

俯瞰的情報抽出された「がん」に関連する技術的対策用語に対して３つの指標を算出し，後述するような円状のマップ（以下「共語マップ」と呼ぶ）を作成した．また，専門家抽出された用語に対しても同様に共語マップを作成し，２つのマップの比較を行った．なお，共語マップの半径方向の長さは用語の出現記事数を表しており，中心に向かうほど大きくなる．各用語の位置を角座標として見たときの角度成分には意味はないが，辺で連結されている用語はなるべく近傍に置き，多数の辺が集まる場合はできるだけ交差しないように配置した．また指標別のマップの作成に当たっては，Jaccard 指標の共語マップの用語位置を基準とし，同じ用語の場合は原則として他の指標のマップ上でも同じように配置した.

用語間を結ぶ辺の線種や用語の表記スタイルは，図の凡例に示す各指標の範囲に基づいて決められている．辺が実線となる範囲，及び用語が斜字となる範囲については，それぞれの範囲に含まれる用語数が全体の２割程度を占めるような値を設定した. 共起指標に対する閾値（下限値）の設定によって，マップ上に現れる用語や辺の数が決まるが，本研究では，特に Jaccard 指標の共語マップの中心部と周辺部の用語と辺がバランスよく出現するような閾値を設定した．同等性指標の閾値は，Jaccard 指標と比較するためにそれと同じ値を設定した．近接指標の閾値は，特に周辺部の用語がマップ上で強調されるような値を設定した.

（1）俯瞰的情報抽出された技術的対策用語の共語分析

俯瞰的情報抽出された技術的対策用語の共語関係に関する用語別データ，及びそれらから算出された Jaccard，同等性，近接の３指標を Table 6 に示す．表に掲載されている数値は共起頻度が２以上で，指標ごとに定めた閾値以上の用語についてのみ示した．なお，閾値未満のデータは共語マップには使用されないため，マップとの対応をとりやすくするため当該データの欄には「ー」を記載した.

a. Jaccard 指標及び同等性指標による共語分析

Table 6 のデータより作成した共語マップを Fig. 2, 3 に示した． Jaccard 指標と同等性指標に基づく共語マップには，共通している用語が多く，特に円の周辺部の用語の連結は類似している．Fig. 2, 3 に示されているように，中心に「がん」という最も基本的な用語が位置し，その周囲に連結している用語は大きく３つの集合を形成している. それぞれの円の中心方向に「治療法」,「抗がん剤」,「血液」,「たんぱく質」といった用語が存在することから，各集合はがんの治療，創薬，診断に関わる分野の用語が集まっていることがわかる．このことは，共語マップ上で関係のある用語のまとまりを階層構造的に見ることができ，上位（円の中心方向）にある用語（以下「上位語」と呼ぶ）がその下に位置付けられる用語群の共通する特徴を表現する可能性が高いことを示唆している.

各集合の上位語はほとんど斜字（用語にかかる辺数が５以上）となっており，この用語と連結している辺数が多く，その用語を介していくつもつながっている．例えば「たんぱく質」の場合,「血液」を経由してがん診断に関わる用語のネットワークと連結しているだけでなく,「抗がん剤」や「がん治療薬」のような薬剤関係のネットワーク,「細胞」関係のネットワークとも連結している．この場合は「たんぱく質」を利用したがん診断・治療に関わる各種の応用技術が発展していると見ることができる.

以上のように，自動抽出された技術的対策用語を共語分析した結果，共通する課題の下に，対策，原因，手段といった用語の組み合わせが，共語マップ上では並置され，連結関係を可視化できたが，それぞれの連結の意味は研究者が分析して把握する必要がある．なお，両指標ともに各語の出現記事数に対する共起数の比率が大きいほど大きな値となるが，出現記事数等が同じ条件のときには同等性指標のほうが Jaccard 指標よりも小さな値になる傾向があり，特に出現記事数が大きい用語ほどその傾向が強くなる．これはマップ上では，同等性指標のマップのほうが，中心付近の用語や辺が少なくなることに対応し, Fig. 2, 3 を比較してもそのような傾向を読み取ることができる.

Table 6　Jaccard 指標，同等性指標，近接指標の算出に用いた「がん」関連用語別データ

キーワード１	キーワード２	キーワード１を含む記事数	キーワード２を含む記事数	共起頻度	Jaccard指標 (≧0.15)	同等性指標 (≧0.15)	近接指標 (≧3.0)
がん	患者	41	31	16	0.2857	0.2014	―
	たんぱく質	41	20	10	0.1961	―	―
	治療法	41	12	10	0.2326	0.2033	―
	がん細胞	41	18	8	0.1569	―	―
	細胞	41	15	8	0.1667	―	―
患者	たんぱく質	31	20	8	0.1860	―	―
	血液	31	8	7	0.2188	0.1976	―
	治療法	31	12	7	0.1944	―	―
	副作用	31	17	7	0.1707	―	―
	高度先進医療	31	9	6	0.1765	―	―
たんぱく質	がん細胞	20	18	7	0.2258	―	―
	血液	20	8	6	0.2727	0.2250	3.1125
	抗がん剤	20	16	6	0.2000	―	―
	細胞	20	15	6	0.2069	―	―
	働き	20	9	5	0.2083	―	―
	がん治療薬	20	4	4	0.2000	0.2000	4.1500
	遺伝子	20	10	4	0.1538	―	―
	ＭＣＢＩ	20	2	2	―	―	4.1500
	がん患者	20	2	2	―	―	4.1500
	診断薬	20	2	2	―	―	4.1500
がん細胞	抗がん剤	18	16	7	0.2593	0.1701	―
	副作用	18	17	6	0.2069	―	―
	患部	18	5	3	0.1500	―	―
	血管	18	5	3	0.1500	―	―
	ＤＤＳ	18	2	2	―	―	4.6111
	がん治療法	18	2	2	―	―	4.6111
	がん放射線治療	18	2	2	―	―	4.6111
	コンピューター断層撮影装置	18	2	2	―	―	4.6111
	医療機関	18	3	2	―	―	3.0741
	動物実験段階	18	2	2	―	―	4.6111
	微小カプセル	18	2	2	―	―	4.6111
副作用	抗がん剤	17	16	10	0.4348	0.3676	3.0515
	細胞	17	15	5	0.1852	―	―
	遺伝子	17	10	4	0.1739	―	―
	治療効果	17	2	2	―	―	4.8824
	微小カプセル	17	2	2	―	―	4.8824
	薬剤	17	3	2	―	―	3.2549
抗がん剤	細胞	16	15	5	0.1923	―	―
	遺伝子	16	10	4	0.1818	―	―
	臨床試験	16	8	4	0.2000	―	―
	薬	16	5	3	0.1667	―	3.1125
	治療効果	16	2	2	―	―	5.1875
	微小カプセル	16	2	2	―	―	5.1875
細胞	遺伝子	15	10	5	0.2500	0.1667	―
	働き	15	9	5	0.2632	0.1852	3.0741
	酵素	15	4	3	0.1875	0.1500	4.1500
	体内	15	8	3	0.1500	―	―
	治療成績	15	3	2	―	―	3.6889
	正常	15	3	2	―	―	3.6889
	造血幹細胞	15	2	2	―	―	5.5333
治療法	高度先進医療	12	9	5	0.3125	0.2315	3.8426
	外科切除	12	2	2	0.1667	0.1667	6.9167
	治療技術	12	2	2	0.1667	0.1667	6.9167
	治療成績	12	3	2	0.1538	―	4.6111
遺伝子	働き	10	9	4	0.2667	0.1778	3.6889
	マウス	10	3	2	0.1818	―	5.5333
	酵素	10	4	2	0.1667	―	4.1500
	正常	10	3	2	0.1818	―	5.5333
体内	治療薬	8	4	2	0.2000	―	5.1875
働き	酵素	9	4	3	0.3000	0.2500	6.9167
	治療薬	9	4	2	0.1818	―	4.6111
	物質	9	6	2	0.1538	―	3.0741
腫瘍	高度先進医療	9	9	4	0.2857	0.1975	4.0988
	がん腫瘍	9	5	2	0.1667	―	3.6889
	胃	9	5	2	0.1667	―	3.6889
	外科切除	9	2	2	0.2222	0.2222	9.2222
	血管	9	5	2	0.1667	―	3.6889
	入院期間	9	2	2	0.2222	0.2222	9.2222
	量	9	3	2	0.2000	―	6.1481
	膵臓	9	2	2	0.2222	0.2222	9.2222
高度先進医療	がん腫瘍	9	5	5	0.5556	0.5556	9.2222
	胃	9	5	2	0.1667	―	3.6889
	外科切除	9	2	2	0.2222	0.2222	9.2222
	入院期間	9	2	2	0.2222	0.2222	9.2222
	腹腔	9	2	2	0.2222	0.2222	9.2222
臨床試験	がん組織	8	4	2	0.2000	―	5.1875
	治験	8	2	2	0.2500	0.2500	10.3750
	大腸がん	8	7	2	0.1538	―	―
血液	ＭＣＢＩ	8	2	2	0.2500	0.2500	10.3750
	がん患者	8	2	2	0.2500	0.2500	10.3750
	たんぱく質断片	8	3	2	0.2222	0.1667	6.9167
	研究グループ	8	2	2	0.2500	0.2500	10.3750
	診断法	8	4	2	0.2000	―	5.1875
	診断薬	8	2	2	0.2500	0.2500	10.3750
	早期がん	8	5	2	0.1818	―	4.1500
	量	8	3	2	0.2222	0.1667	6.9167
物質	陽電子放射断層撮影装置	6	4	2	0.2500	0.1667	6.9167
早期がん	コンピューター断層撮影装置	5	2	2	0.4000	0.4000	16.6000
	陽電子放射断層撮影装置	5	4	2	0.2857	0.2000	8.3000
がん腫瘍	腹腔	5	2	2	0.4000	0.4000	16.6000
脳腫瘍	悪性	4	2	2	0.5000	0.5000	20.7500
治療薬	がん治療薬	4	4	2	0.3333	0.2500	10.3750
たんぱく質断片	研究グループ	3	2	2	0.6667	0.6667	27.6667
	前立腺がん	3	3	2	0.5000	0.4444	18.4444
たばこ	肺がん	2	2	2	1.0000	1.0000	41.5000

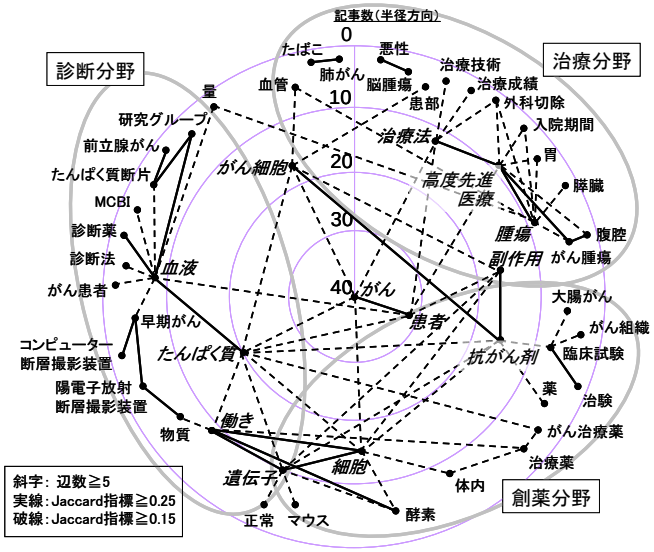

Fig. 2　Jaccard 指標に基づく「がん」関連技術的対策用語（自動抽出）の共語マップ

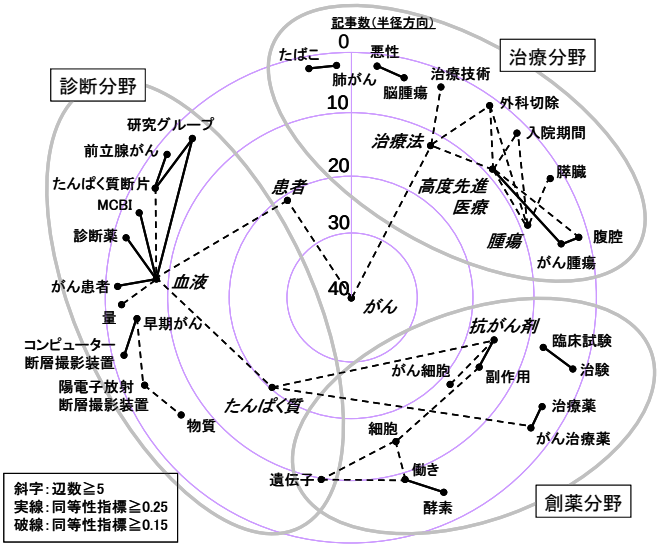

Fig. 3　同等性指標に基づく「がん」関連技術的対策用語（自動抽出）の共語マップ

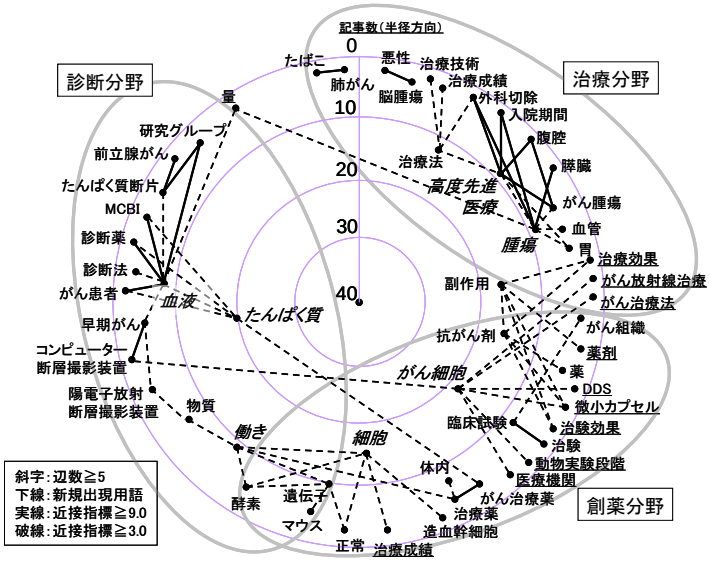

Fig. 4　近接指標に基づく「がん」関連技術的対策用語（自動抽出）の共語マップ

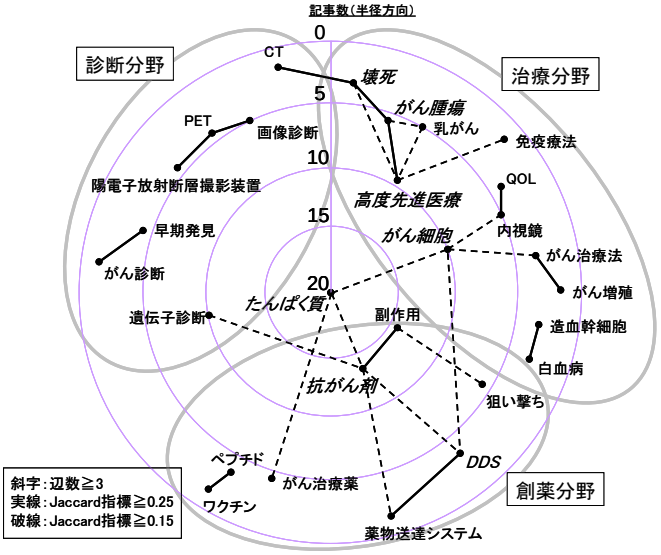

Fig. 5　Jaccard 指標に基づく「がん」関連技術的対策用語（専門家抽出）の共語マップ

b. 近接指標による共語分析

近接指標に基づく共語マップを Fig. 4 に示す．この指標は，低出現頻度の用語の中で関係の強いものを見いだすときに利用される．Fig. 4 では Fig. 2, 3 のマップの周辺部が強調された形になっており，両図には見られない新しい技術的対策用語も見られた（図中，下線のある用語）．傾向として技術用語は周辺部に多く，上位語の具体例，対策，手段といった位置づけになっている．近接指標は，このような出現頻度が少ない個別技術用語の位置付けの明確化に有用と考えられる．

(2) 専門家抽出された技術的対策用語の共語分析

専門家抽出された技術的対策用語の共語関係から算出した Jaccard 指標に基づく共語マップを Fig. 5 に示す．

専門家抽出では，技術的対策用語と関係の深い，専門用語ではない語（患者，働き，量など）や上位概念を表す語（がん，治療法など）は抽出されないことが多く，マップ上に位置づけられる用語数はかなり少なくなっている．また，全体的には専門家抽出によるマップは自動抽出によるマップの部分集合的な位置づけと見ることができるが，自動抽出では見られなかった用語が外周部に多く見られる（免疫療法，内視鏡，遺伝子診断など）．これらの用語は，新聞記事としては専門性の高い用語であり，基本的に用語の出現頻度は低くなる．本研究で用いられている指標は，用語の出現頻度がある程度確保され，設定された言語パターンとの関係が強い場合に高得点となることから，結果的に指標のスコアが低くなり自動抽出されなかった，もしくは自動抽出されても計量値は表示

Table 7　Jaccard 指標，同等性指標，近接指標の算出に用いた「生活習慣病」関連用語別データ

キーワード１	キーワード２	キーワード１を含む記事数	キーワード２を含む記事数	共起頻度	Jaccard指標（≧0.15）	同等性指標（≧0.15）	近接指標（≧3.0）
患者	心臓	8	8	5	0.4545	0.3906	―
	血管	8	7	4	0.3636	0.2857	―
	心筋梗塞	8	6	3	0.2727	0.1875	―
	バージャー病	8	2	2	0.2500	0.2500	4.2500
	心臓病	8	3	2	0.2222	0.1667	―
心臓	血管	8	7	3	0.2500	0.1607	―
	重症	8	5	3	0.3000	0.2250	―
	心筋梗塞	8	6	3	0.2727	0.1875	―
	幹細胞	8	2	2	0.2500	0.2500	4.2500
	血液	8	2	2	0.2500	0.2500	4.2500
	心臓病	8	3	2	0.2222	0.1667	―
生活習慣病	糖尿病	8	8	4	0.3333	0.2500	―
	血圧	8	5	3	0.3000	0.2250	―
糖尿病	合併症	8	3	3	0.3750	0.3750	4.2500
	インスリン	8	4	2	0.2000	―	―
	ホルモン	8	2	2	0.2500	0.2500	4.2500
	原因	8	2	2	0.2500	0.2500	4.2500
	酵素	8	2	2	0.2500	0.2500	4.2500
	人	8	2	2	0.2500	0.2500	4.2500
	働き	8	2	2	0.2500	0.2500	4.2500
血管	バージャー病	7	2	2	0.2857	0.2857	4.8571
	血管内	7	2	2	0.2857	0.2857	4.8571
	心筋梗塞	7	6	2	0.1818	―	―
	動脈硬化	7	4	2	0.2222	―	―
血糖値	インスリン	6	4	2	0.2500	0.1667	―
	血圧	6	5	2	0.2222	―	―
	細胞	6	5	2	0.2222	―	―
心筋梗塞	心臓病	6	3	2	0.2857	0.2222	3.7778
細胞	インスリン	5	4	2	0.2857	0.2000	3.4000
	ラット	5	2	2	0.4000	0.4000	6.8000
重症	インスリン	5	4	2	0.2857	0.2000	3.4000
	幹細胞	5	2	2	0.4000	0.4000	6.8000
インスリン	糖尿病患者	4	2	2	0.5000	0.5000	8.5000
健康管理	社員食堂	4	2	2	0.5000	0.5000	8.5000
	生活習慣	4	2	2	0.5000	0.5000	8.5000
	摂取カロリー	4	2	2	0.5000	0.5000	8.5000
合併症	ホルモン	3	2	2	0.6667	0.6667	11.3333
	働き	3	2	2	0.6667	0.6667	11.3333
最前線	治療研究	2	2	2	1.0000	1.0000	17.0000
生活習慣	社員食堂	2	2	2	1.0000	1.0000	17.0000
	摂取カロリー	2	2	2	1.0000	1.0000	17.0000
摂取カロリー	社員食堂	2	2	2	1.0000	1.0000	17.0000

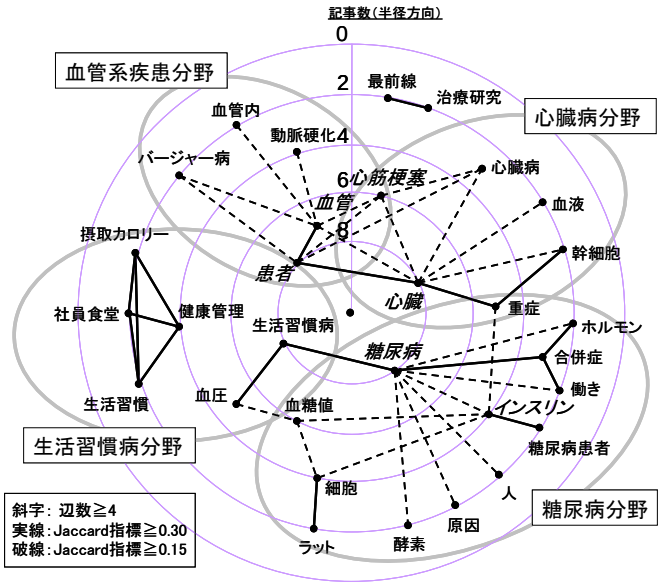

Fig. 6　Jaccard 指標に基づく「生活習慣病」関連技術的対策用語（自動抽出）の共語マップ

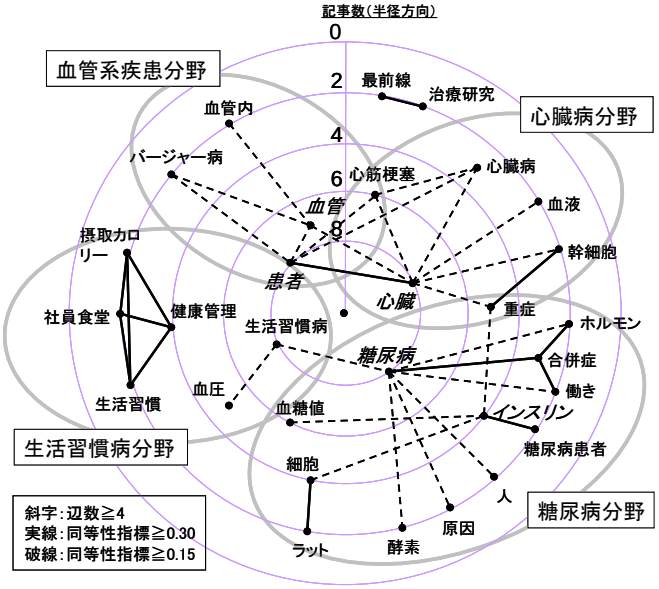

Fig. 7　同等性指標に基づく「生活習慣病」関連技術的対策用語（自動抽出）の共語マップ

上の閾値より低くなりマップから除外されたためである．一方、専門家は出現頻度に必ずしもとらわれず，技術的対策上の意味内容から適，不適を直感的に判断する傾向を無視できないことから，専門家抽出用語は実際の頻度と無関係に抽出されることが判明した．

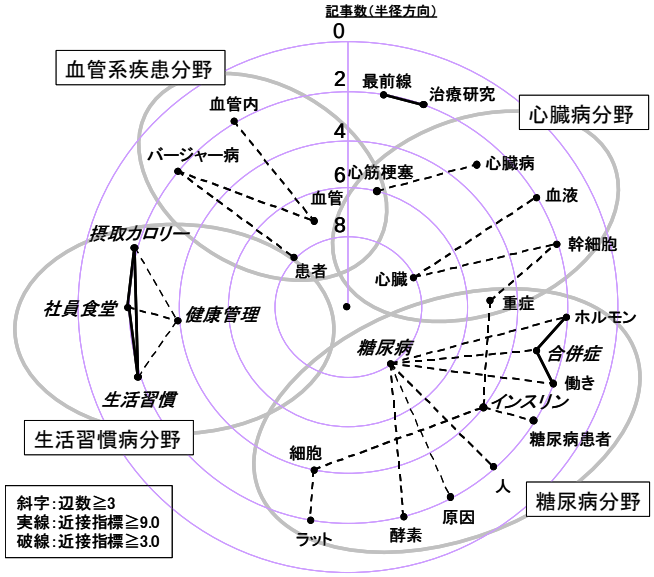

Fig. 8　近接指標に基づく「生活習慣病」関連技術的対策用語（自動抽出）の共語マップ

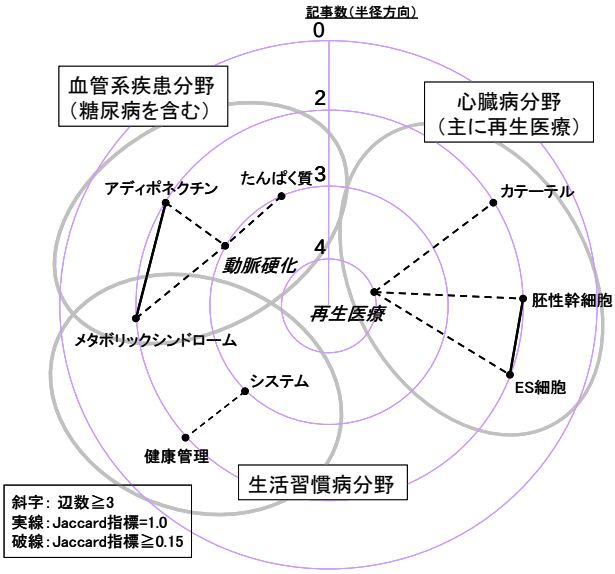

Fig. 9　Jaccard 指標に基づく「生活習慣病」関連対策用（専門家抽出）の共語マップ

4.2. 「生活習慣病」関連技術的対策用語に対する共語分析

（1）俯瞰的情報抽出された技術的対策用語の共語分析

a. Jaccard 指標及び同等性指標による共語分析

Table 7 に示すように，俯瞰的情報抽出された技術的対策用語の共語関係から Jaccard，同等性，近接の３指標を算出した．これらのうち，Jaccard 指標及び同等性指標のデータから作成した共語マップを Fig. 6, 7 に示す．

図に示すように，「生活習慣病」の場合は，同等性指標の共語マップのほうの辺が少なくなっているが，２つのマップはほぼ同等である．用語は４つの集合に分けられ，各集合の上位語が「心臓」，「糖尿病」，「生活習慣病」，「血管」であることから，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 分野の用語のまとまりと見ることができる．

各分野の上位語は，「生活習慣病」を除く３語が斜字になっているが，例えば辺数が最も多い「糖尿病」の場合，糖尿病の治療（ホルモン，インスリン），診断（酵素），予防（原因）などに関連する用語のネットワークが連結している．これは，糖尿病を克服するための各種の対策がマップ上で並置されていると見ることができる．

指標の値が大きい２つの用語の組み合わせも見られた．例えば「生活習慣病－血圧」，「細胞－ラット」，「インスリン－糖尿病患者」などが課題（病気）と手段の関係にある組み合わせである．また，「心臓－重症－幹細胞」や「健康管理－生活習慣－社員食堂－摂取カロリー」といった２語以上の組み合わせも見られ，これらにより表される社会事象を，より具体的に把握することができる．

b. 近接指標による共語分析

近接指標に基づく共語マップを Fig. 8 に示す．Fig. 7 と 8 を比べると，マップの中心部の辺が希薄になった．「がん」の場合と比べて，周辺部の強調度は小さく，新しい用語も出現しなかった．「がん」の場合には，出現記事数が中位（20 件前後）の用語と連結する周辺部の用語が新しく出現しており，これは指標の算出式の特徴から，全記事数（式(9)の N）が大きいときに上記のような用語の連結が現れやすくなることによる．「生活習慣病」の場合は，全記事数が相対的に小さいために「がん」に見られたような差が出なかったといえる．

（2）専門家抽出された技術的対策用語の共語分析

専門家抽出された技術的対策用語の共語関係から算出した Jaccard 指標に基づく共語マップを Fig. 9 に示す．Fig. 6 と比べて用語数は非常に少なく，共通している用語は「動脈硬化」と「健康管理」のみであった．これは，専門家抽出された用語は，自動抽出では他の抽出用語と比べて課題関連度及び技術関連度のスコアが相対的に低くなり，記事ごとの抽出用語数制限で落ちてしまうことが多いためである．理由としては，当該用語は設定された言語パターンと一緒に使われることが少なかったことなどが考えられるが，今後の検討課題である．

5.　結言

本研究では，俯瞰的情報抽出の手続きを提案し，抽出された技術的対策用語の共語分析を通じてその有用性を検討した．具体的には，「がん」及び「生活習慣病」に関する社会課題から技術的対策用語を自動的に抽出し，得られたキーワードを用いて記事内容の共語分析を行った．

その結果,「がん」や「生活習慣病」に関する課題や技術的対策の傾向を共語マップとして可視化することができ,その内容を直接に把握することが可能となった．共語マップは，概念的に関連性のある用語のまとまりを直観的に示しており，課題と技術的対策の関係（あるいは構造）を記事の内容のレベルで把握できることが確認された．

例えば「がん」に関しては，がんの克服のための対策として大きく治療，創薬，診断の３分野があり，それぞれ高度先進医療，抗がん剤，血中たんぱく質検出といった技術の位置付けが高かった.「生活習慣病」に関しては,特に心臓病，糖尿病，動脈硬化への対応が課題となっていることが把握され，例えば糖尿病は合併症への対応，ホルモン治療などが注目されていた．また，課題と技術的対策の用語間の関係から重要な情報も読み取れた．例えば,「たんぱく質」は血液診断関連，抗がん剤関連，遺伝子関連の用語と連結しており,たんぱく質関連技術が,広範ながん診断･治療に関与するキーテクノロジーであることが把握できた.

このように,共語マップ上で可視化された“まとまり”や“つながり”あるいは位置関係は，社会課題に対する技術的対策の重要性やプライオリティを判断するための有用な情報を提供してくれることを示している．これらの結果は，これまで適用が困難であった新聞記事に対しても共語分析が可能であることを示すものであり，テキストの内容にまで踏み込んだ共語分析の普及に強力な方法論を提供するものと考える．例えば，技術アナリストや科学技術関連政策立案者は，本研究で提案した手法を利用して,注目度の高い社会課題を解決する技術的対策,あるいは複数の社会課題の解決に関係する技術的対策などを把握し，さらなる有識者によるブラッシュアップにより，プライオリティの高い技術的対策の傾向分析やその結果に基づく科学技術政策立案に活用できる可能性がある.

また，自動抽出用語と専門家抽出用語にかなりの差が生じた点も興味深い．差が生じた理由の一つは，自動抽出の場合には技術的対策用語と関連の深い，専門用語でない語や，主体を表す語などが同時に抽出されることである．これらの用語の中には，共語マップ上で用語間の関係を表す構造の中で，その内容をわかりやすくする役割を果たすものがあるが，当然ノイズとなるものも存在する. 今回は, 精査した記事セットを使用したが, 今後,自然言語処理のメリットを活かして，人手による精査が困難なレベルの大量の記事から直接技術的対策用語を自動抽出する場合には，技術的対策用語抽出のための言語パターン設定の最適化や技術関連度の算出に用いた重みの調整など，ノイズを減らす方策を構築していくことが必要となる．また，我々は技術的対策だけでなく，制度的対策，サービス的対策の用語についてもテキストマイニングする研究を進めており，最終的には社会課題とこれらの対策用語との関係を見ることにより，社会課題の解決に有用な総合的知識を獲得することができると考える.

参考文献

1) 奥田英範, 川島晴美, 佐藤吉秀, 宮原伸二, 定方徹 (2006.05)「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』5, 22-25.
2) 佐藤吉秀，川島晴美，佐々木努，奥雅博 (2005)「時系列ニュース記事における最新話題語抽出方法」『電子情報通信学会技術研究報告』105(203)(NLC2005 1-12), 1-6.
3) 橋本泰一, 村上浩司, 乾孝司, 内海和夫, 石川正道 (2008.03)「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.
4) Callon, M., Courtial, J. P., and Laville, F. (1993). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics*, 22(1), 153-205.
5) Rip, A., and Courtial, J. P. (1983). Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics*, 6(6), 381-400.
6) Leydesdorff, L., and Hellsten, I., (2006). Measuring the meaning of words in contexts: An automated analysis of controversies about ‘Monarch butterflies,’ ‘Frankenfoods,’ and ‘stem cells’. *Scientometrics*, 67(2), 231-258.
7) 日本経済新聞･日経シソーラス, http://telecom21.nikkei.co.jp/
8) 内海和夫，乾孝司，村上浩司，橋本泰一，石川正道 (2008)「大規模テキストマイニングによる医療分野の社会課題・技術トレンド分析」『研究・技術計画学会 第 22 回年次学術大会 講演要旨集』684-687.
9) 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.
10) Nakagawa, H., (1999). Compound noun based system for automatic term recognition task. *Proceedings of the first NTCIR workshop on research in Japanese text retrieval and term recognition.*
11) Christopher, D., (2008). Introduction to Information Retrieval. *Cambridge University Press*. Chapter 17. 7.
12) 徳永健伸 (1999).「第 2 章 情報検索の基礎，第 3 章 情報検索システムの性能評価」『言語と計算５：情報検索と言語処理』東京大学出版会
13) 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

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.

EXTRACTION OF CRITICAL KNOWLEDGE CONCERNING SOCIAL PROBLEMS AND THEIR TECHNOLOGICAL SOLUTIONS

Kazuo UTSUMI[1], Takashi INUI[2], Taiichi HASHIMOTO[3], Koji MURAKAMI[4] and Masamichi ISHIKAWA[5]

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

The new procedure to extract technologically relevant words against societal issues was developed using newspaper articles. The originally introduced indices: 'problem relevancy (pr)' and 'technical relevancy (tr)' enabled to score their importance of the words appeared in the articles and to assign keywords to each article. The F-measure defined by precision and recall of the word extraction was evaluated and showed the higher values than those of the baseline indices (tfidf) . As its application, the keywords extraction of technological solutions against medical issues, i.e. the counter measures for cancer and lifestyle-related illness was executed using the articles from Nikkei newspaper. The obtained keywords enabled to construct the co-word maps using Jaccard index, equivalence index and proximity index. The results showed the usefulness of the procedure for co-word analysis.

Key Words: *comprehensive approach, co-word analysis, information extraction, document clustering, language pattern*