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

Discovering Technical Solutions to the Social Problems from Newspaper Articles

乾 孝司[†] 内海 和夫[†] 橋本 泰一[†] 村上 浩司[‡] 石川 正道[†]
Takashi Inui Kazuo Utsumi Taiichi Hashimoto Koji Murakami Masamichi Ishikawa

1 はじめに

社会の複雑化と共に，社会に不安や不信を引き起こす要因が増加しており，社会的課題群を俯瞰的に把握する手段が切望されている [6]．このような背景を鑑み，我々はこれまで，新聞記事集合から社会課題に関する情報を自動的に抽出する手法を提案し，その有効性を検証してきた [1, 5]．次なる発展的話題として，社会課題に対する対策活動に関する情報の抽出手法を検討している．本稿では，このうち，技術に基づく対策活動*を表す用語の自動抽出手法について述べる．

2 提案手法

提案手法の入出力例を図 1 に示す．入力は，先行技術 [5, 1] によって獲得された記事クラスタであり，各記事クラスタは，「がん」や「交通事故」等，ある特定の社会課題に関する記事群で構成されている．提案手法では，入力された課題記事クラスタ内の各記事ごとに対策用語を抽出し，出力する．

入力となる課題記事クラスタに含まれる各記事は，いずれもある特定の社会課題について記述されている．抽出したい対策用語は，このような特定課題と関連が強く，かつ，技術との関連が強い用語であると考えられる．そこで，用語に対して，課題関連度（problem relevancy：pr），技術関連度（technical relevancy：tr）の 2 つの指標を定義し，これらの指標に基づくスコア計算によって対策用語を抽出する．

2.1 課題関連度

社会課題は記事クラスタとして表現されている．この状況において，課題関連度の強い用語を抽出することは，記事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の問題となる．そこで，用語 t の課題関連度を，クラスタ・ラベリングで適用される基本的な指標である χ^2 値 [2] で定義する（$pr(t) = \chi^2(t)$）．具体的な $\chi^2(t)$ は，一年分の記事集合と課題クラスタ内の記事集合における t の出現数に基づいて計算する．

2.2 技術関連度

技術関連度は言語パタンに基づいて定義する．図 2 のような，技術的事柄を表す際に使われやすい言い回しを言語パタンとして用意し，言語パタンと照合する文を技術関連文と呼ぶ．この時，技術関連文の近くに現れる用語ほど技術関連度が強いと仮定する．

ある用語 t は一記事内で複数回現れる可能性がある．

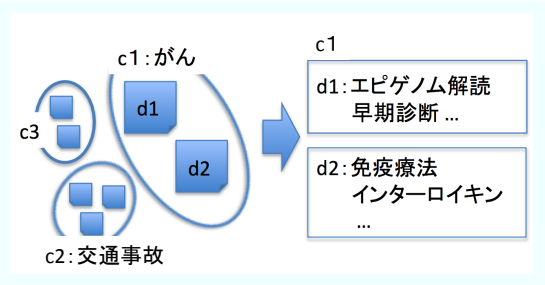

図 1 課題記事クラスタからの対策用語抽出

X を → 開発する
X の研究を → 手掛ける （「→」は係り受け関係）

図 2 言語パタン

そこで，記事 d 内の用語 t に対し，j 番目に現れるものを t_j^d とする．その上で先の仮定に従い，技術関連文との文間の位置関係に基づいた技術関連度 $tr_{inter}(t_j^d)$ と，技術関連文の文内位置に基づいた技術関連度 $tr_{intra}(t_j^d)$ を定義する．

2.2.1 文間の位置関係に基づいた技術関連度

予め人手で準備した対策用語に対し，対策用語を含む文と技術関連文との位置関係を調査した．技術関連文が対策用語を含んでいる場合を相対位置 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，以降同様に技術関連文から対策用語を含む文への相対位置を定めると，相対位置ごとの対策用語の度数分布は図 3 のようになる．このデータに基づいて指数回帰式を導出し，定数項を無視すると，$y = \mathrm{e}^{-0.77x}$ を得る．この関数形に従い，$tr_{inter}(t_j^d)$ を以下で定義する：

$$tr_{inter}(t_j^d) = \max_{p_k^d \in P^d} \exp(-0.77 \times r\_pos(p_k^d, t_j^d)). \quad (1)$$

ただし，P^d は，d に含まれる技術関連文の集合，$r\_pos(p_k^d, t_j^d)$ は，p_k^d からみた t_j^d を含む文への相対位置である．技術関連文をもたない記事は常に $tr_{inter}(t_j^d) = 1$ とする．

2.2.2 文内位置に基づいた技術関連度

技術関連文では，言語パタンの述語に対する項（図 2 の「X」）および項への修飾要素に位置する用語は他より関連度が強くなると考えられる．そこで，$tr_{intra}(t_j^d)$ を以下で定義する：

$$tr_{intra}(t_j^d) = \begin{cases} 100 & (t_j^d \text{ が言語パタン内の項} \\ & \quad or \text{ その修飾要素}) \\ 1 & (otherwise). \end{cases} \quad (2)$$

---

† 東京工業大学統合研究院.
‡ 奈良先端科学技術大学院大学情報科学研究科.

* 他の対策活動として，国家主導でおこなう，制度に基づく対策活動などがある．例えば，鳥インフルエンザ・ウィルスの問題では，予防ワクチンの開発は技術的対策であるが，生きた鳥類の輸入制限は制度的対策であると言える．

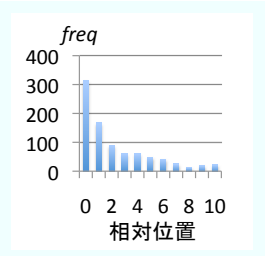

図 3　相対位置と対策用語数の関係

2.2.3　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で定義する：

$$tr(t,d) = \sum_{t_j^d \in ins(t,d)} tr_{intra}(t_j^d) tr_{inter}(t_j^d). \quad (3)$$

ここで，$ins(t,d)$ は d 内に現れる t の実体の集合を示す．

2.3　対策用語の抽出

以上の準備のもと，記事 d 内の各用語 t に対して，$pr(t)tr(t,d)$ でスコアを計算し，スコア上位の n 個を対策用語として出力する．対策用語数は記事の長さに依存する．記事の文数が 10 以下の状況では，文数と対策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対策用語数はあまり増加しない．そこで，記事毎に抽出用語数を変化させ，記事 d における抽出用語数を $n_d = min(|d|, 10)$ とする．$|d|$ は記事 d の文数である．

3　評価実験

日本経済新聞 2005 年版のうち医療分野の記事から得られた課題記事クラスタ [5] から，「がん」（84 記事）と「生活習慣病」（35 記事）のクラスタを評価実験に用いた．評価のために，このデータに対して人手であらかじめ正解となる対策用語を抽出しておき，提案手法で抽出された用語との間で F 値を求めた．

スコアを付与する用語の候補は，まず，対象記事の各文を CaboCha[†] で文節にまとめあげ，各文節から助詞等の付属語を取り除くことで生成した．固有物[‡]以外の固有表現は，スコアが高くなるが抽出すべき対策用語でないことが多いため，固有物以外の固有表現は用語候補から削除した．

課題関連度として用いる χ^2 はイエーツ補正値を用いた．技術関連度で用いる言語パタンは図 2 のような表現を約 50 表現用意した．また，文書からの用語抽出における汎用的な指標として $tfidf$ がある [3, 4]．評価実験では，$tfidf$ をベースラインのスコア計算法として採用し，提案手法と比較する．

実験結果を表 1 に示す．提案手法（1 行目）は，$pr(t)$, $tr(t,d)$ の単独使用あるいは $tfidf$ のいずれよりも高い F 値を達成した．また，$pr(t)$, $tr(t,d)$ 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な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，$pr(t)$ と $tr(t,d)$ は，抽出精度向上に同程度寄与していると言える．

正しく抽出できた対策用語の例を表 2 に示す．

表 1　実験結果（F 値）

スコア	がん	生活習慣病
$pr(t)tr(t,d)$	0.532	0.608
$pr(t)$	0.473	0.558
$tr(t,d)$	0.487	0.512
$tfidf$	0.470	0.509

表 2　正しく抽出できた例（一行が一記事分の出力に対応）

がん	テーラメード医療 分子標的薬 遺伝子
	抗がん剤 薬物送達システム 中性子線照射
	腫瘍脊椎骨全摘術 ワイヤ状 のこぎり
	食道がん アセトアルデヒド 呼気
	がん早期発見 分光推定 血管 光照射
生活習慣病	携帯型心電計 防水性能
	心臓病 冠状動脈 動画
	生活習慣病予防 トイレ 血圧

4　おわりに

本稿では，新聞記事集合から社会課題に対する技術的な対策情報を自動抽出する手法を提案した．提案手法では，課題との関連度および技術との関連度を考慮したスコアに基づき対策用語を抽出する．今回は 2 つの記事クラスタによって提案手法を評価したが，今後，より大規模なデータに基づく評価を実施したい．

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．

参考文献

[1] 橋本泰一, 村上浩司, 乾孝司, 内海和夫, 石川正道. 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, Vol. 5, , 2008.

[2] Christopher D. Manning. *Introduction to Information Retrieval*, chapter 17.7. Cambridge University Press, 2008.

[3] Y. Matsuo and M. Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, Vol. 13, No. 1, pp. 157–169, 2004.

[4] G. Salton. Syntactic approaches to automatic book indexing. In *Proceedings of the 26th annual meeting on Association for Computational Linguistics*, pp. 204–210, 1988.

[5] 内海和夫, 乾孝司, 村上浩司, 橋本泰一, 石川正道. 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, 2007.

[6] 奥田英範, 川島晴美, 佐藤吉秀, 宮原信二, 定方徹. 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, Vol. 18, No. 5, pp. 22–25, 2006.

[†] http://chasen.org/~taku/software/cabocha/

[‡] IREX（http://nlp.cs.nyu.edu/irex/index-j.html）の定義に従う．