2E07

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

○内海和夫、乾孝司、村上浩司、橋本泰一、石川正道（東京工業大学）

１．はじめに

膨大なテキスト情報から俯瞰的アプローチにより社会課題・トレンドを抽出するニーズが高まっている。我々は近年急速に発展してきたテキストマイニング技術を応用して、大量の新聞記事を語彙分布に基づいて階層的にクラスタリングし、個々のクラスタ（記事集合）を自動要約することによって、クラスタから中心的なトピックを抽出し、効率的かつ客観的な記事情報分析を可能とするマイニングツール“RiverStone”の開発を目指している。最近報告されている新聞記事へのクラスタリングの適用研究では、特定の内容を象徴するキーワードの集合を抽出してから、そのキーワードに関連する記事集合（背景記事集合）をクラスタリングの手法を用いて形成する研究が多いが [1),2)]、ビブリオメトリクスを含む従来のテキストマイニングの事例では、ほとんどの場合分析の目的に沿ったキーワード抽出から出発するため、データセット全体の傾向を俯瞰的かつ効率的に把握する手段としては不向きであった。

今回、我々が開発したツールでは、それとは逆に大規模な語彙セットを用いて記事全体をクラスタリングすることを可能とした。さらに各クラスタを特徴付ける重要キーワードを抽出するとともに、クラスタ内あるいはクラスタ間の類似性を示す指標を導入することにより、トピックやトレンドを俯瞰的かつ効率的に分析することに成功した。ここでは、そのような俯瞰的なアプローチによる分析を医療分野のデータセットに適用した結果について報告する。

２．分析方法

本研究では、テキストマイニングツールを利用した医療分野の社会課題抽出及びトレンド分析を、次に述べる手順に従って行った [3)]。

１）日本経済新聞記事データベースより、日経シソーラス [4)] の中から医療に関わる検索語 316 を選定して検索し、解析の対象となる記事文書のデータセット（日経新聞本紙 2000 年～2006 年：約 1 万件／年）を作成。

２）データセットに対し階層的クラスタリングを行い、文書に出現する単語の分布の類似性からクラスタを形成（年次ごとに 200 クラスタを形成）。

３）クラスタの形成過程に基づいた樹状図（デンドログラム）を形成しクラスタを構造化するとともに、ノード(枝分かれ箇所)にぶらさがるクラスタをそれらの語彙使用の類似性からグループ化。

４）自動要約によりクラスタを特徴付ける重要キーワードを抽出し、それらを含む文（重要文）のリストを作成。

５）クラスタ間の類似性を表す中心性（Centrality）とクラスタ内の記事文書の類似性を表す密度（Density）の２つの指標をクラスタごとに算出。

６）密度を用いてクラスタをスクリーニング（密度が 0.2 未満のクラスタを削除）。

７）分析者の視点を加えつつ、重要文と樹状図を見ながら社会課題とそれに対する対策を抽出し構造化。

８）課題別・年次別に記事文書数、中心性、密度等を見ながらトレンドを分析。

３．テキストマイニングによる医療分野の社会トレンド分析

ここでは、2000 年から 2006 年のデータセットに対するクラスタリング結果に基づき、まず年次ごとに中心性の値がトップのクラスタ、及びそれを含むグループを特定し、そのグループ内で重要度が最も高いキーワード（記事の冒頭部に近いところでの出現頻度が高い複合語）を抽出したところ、2000 年が「インターネット・ホームページ」、2001 年が「遺伝子（遺伝子治療、遺伝子組み換え等を含む）」、2002 年が「薬害エイズ・訴訟」、2003 年が「世界保健機構（SARS）」、2004 年が「がん（各種がん、がん細胞等を含む）」、2005 年が「医療事故」、2006 年が「障害者（視覚・聴覚・知覚・身体障害者を含む）」となった。これらは、各年次を特徴付ける社会トレンドを表現する重要キーワードと見ることができる。次に、これらの重要キーワードを多く含むクラスタを選出し、各クラスタの記事数の合計を年次ごとにプロットしたのが図１である。この図から次のようなトレンドが読み取れる。

・「世界保健機構（SARS）」は 2003 年に最も話題となり記事数も特異的に多いが、それ以外はほとんどなく、単発的な社会課題である。
・SARS ほど極端ではなく注目度も低いが、「薬害エイズ・訴訟」はこれと類似した傾向を示している、
・「がん」に対する注目度は増加傾向にある。
・「遺伝子」に対する注目度は減少傾向にあったが、2005 年以降また増加の兆しがある。
・「インターネット・ホームページ」や「医療事故」に対する注目度は、2003 年をピークに減少傾向にある。
・「障害者」に対する注目度は、毎年一定のレベルを保っている。

以上は、特定のキーワードに着目したトレンドであるが、それぞれが複数のクラスタに分布しており、その内容を見ることによりさらに詳細な分析が可能である。例えば、図の中の 2006 年の「がん」のプロットは 10 のクラスタにまたがる集合の記事数の合計であり、それぞれのクラスタを特徴付ける「がん」関連キーワードは、「抗がん剤」、「細胞に関するがん（がん細胞を攻撃する免疫療法・DDS、がん幹細胞等）」、「がん保険」、「大腸がんリスク（疫学調査等による）」、「有名人のがん（王監督、三笠宮殿下等）」、「アスベスト（中皮腫）」、「乳がん」、「がん対策基本法」、「静岡がんセンター」、「公立病院のがんセンター」であった。このうち、“がん保険を組み込んだ住宅ローンの発売”、“がん対策基本法の策定”、“王監督の入院”が、2006 年の「がん」の記事数を増加させた主な原因であることが把握された。

４．密度と中心性から見た社会課題の特徴分析

図２は、上記と同様に抽出された「インフルエンザ」（2003 年において「世界保健機構（SARS）」に次いで重要度が高かったキーワード）を含むクラスタ群の密度と中心性の平均値が、重要キーワードを含むクラスタ全体の密度と中心性の平均値とどのくらいずれているかを、年次ごとにプロットしたものである。

密度は、クラスタ内での記事の類似性を示しており、同じような内容・書き方の記事が集まっているほど密度は大きな値となる。一方、中心性はクラスタ間での類似性（⇔独立性）を示しており、重要キーワードが重複する度合いが大きいほど中心性は大きくなる。「インフルエンザ」の場合、2000～2002 年は、その年のインフルエンザの流行の状況や、それに対する対策等についての記事が主体であり、記事数も少なく密度や中心性はほぼ平均並みであるが、2003 年にまず中国で鳥インフルエンザが発生し、次いで 2004 年には日本で発生し、徐々に拡散・感染していったため、比較的長期にわたり事件的な記事が掲載された。2003 年では、鳥インフルエンザだけでなく、抗体医薬やワクチンに関する記事も多く、それが中心性を若干高めた原因と考えられる。一方、2004 年は国内での鳥インフルエンザの流行

の記事が主体で、同じような内容・書き方の記事が集まったため、密度が非常に大きくなったと考えられる。2005 年になると中心性がマイナスに転じているが、これは鳥インフルエンザに関してはタイ人への感染や北朝鮮での流行等、事件性の高い記事が多かったのと、ローマ法王がインフルエンザで入院し死去した記事が加わったため、クラスタ間での類似性の低い記事が増えたことが原因と考えられる。

このように、図２に示すように各年次の密度と中心性の平均値の動きを見ることにより、波及性や関連性の高い記事が多いかどうか（中心性が大きくなる）、あるいは事件性の高い（構文的な類似性は高いが、内容的には独立性が高い）記事が多いかどうか（密度が大きくなり、中心性が小さくなる）、といった分析ができると考えられる。

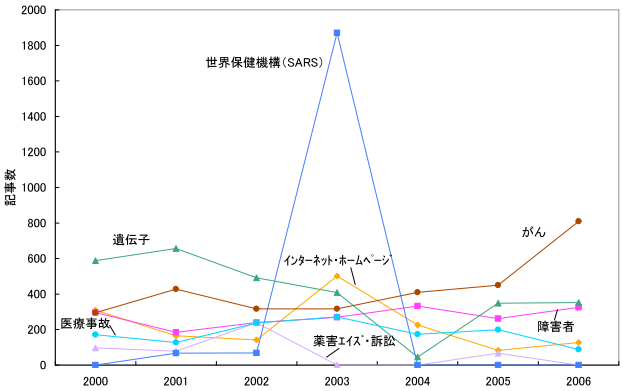


図１　重要キーワードを含むクラスタ群の記事数の推移（2000 年～2006 年）

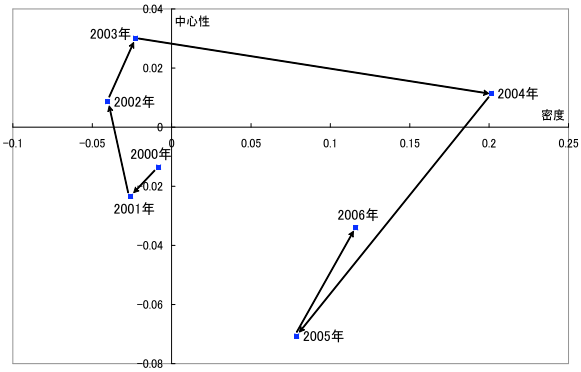


図２「インフルエンザ」に関する記事クラスタ群の密度／中心性ダイヤグラム

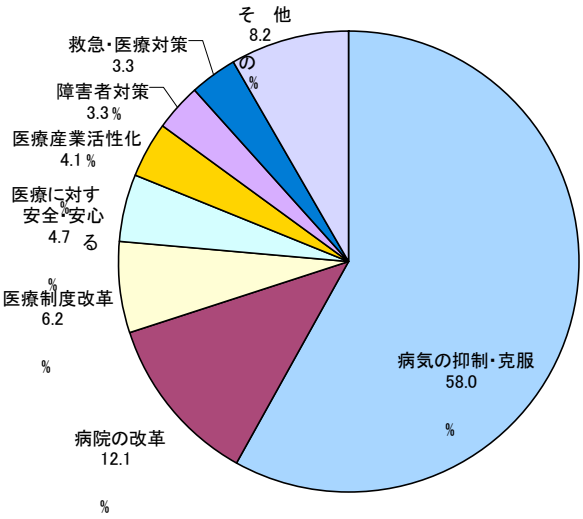
５．医療分野の社会課題と対策に関する情報抽出

2000 年から 2006 年の医療分野のデータセットのうち、2005 年について詳しく分析を行った。2005 年には 8,890 件の記事が含まれ、これらをクラスタリングして 200 のクラスタに分類し、密度によるクラスタのスクリーニング等を経て、自動要約により抽出された各クラスタを特徴付ける重要キーワードとそれを含む文を読みながら、医療分野の社会課題と対策に関する情報抽出を行うことにより、課題の構造化を行った。

分析者により構造化された課題の数は高位の分類レベルで 12、中位レベルで 30、低位レベルで 131 であり、それらに対応する対策数は合計で 900 であった。課題別の対策数は、その課題の注目度を示す一つの指標として考えることができる。例えば、高位レベルの課題別の対策数比率は、図３に示すように「病気の抑制・克服」が最も多く 58%であり、次いで「病院の改革」(12.1%)、「医療制度改革」(6.2%)、「医療に対する安全・安心」（4.7%）等が続いた。また、「病気の抑制・克服」のなかでは、「がんの克服」（12.0%)、「先端技術の医療への適用」（11.0%)、「生活習慣病対策の充実」（8.1%)、「感染症の撲滅」（8.1%）などの比率が高く、がん、生活習慣病、感染症への対策に関する注目度が高いことが把握された。

図４は、中位レベルの課題別の対策数に関し上位６課題をとり上げ、その対策の内訳（技術的対策、製品・サービス的対策、制度的対策、その他の数）を見たものである。注目度の高い病気の中でも、がんについては技術的対策の件数が多い一方で、生活習慣病は製品・サービス的対策、感染症は制度的対策の件数もかなり多いという特徴が見られた。なお、「がんの克服」に関する低位レベルの課題については、「低侵襲・非侵襲治療」や「早期発見」に関する対策数が多く、できるだけ早期にがんを発見するとともに、患者の負担を軽減し QOL（Quality of Life）を維持・向上するニーズが高いことが把握さ

れた。また、技術的対策の内容は、低侵襲治療装置・システム、遺伝子・たんぱく質・細胞といった生体物質、あるいは生体物質の計測情報に基づいた解析を活用した対策が多かった。



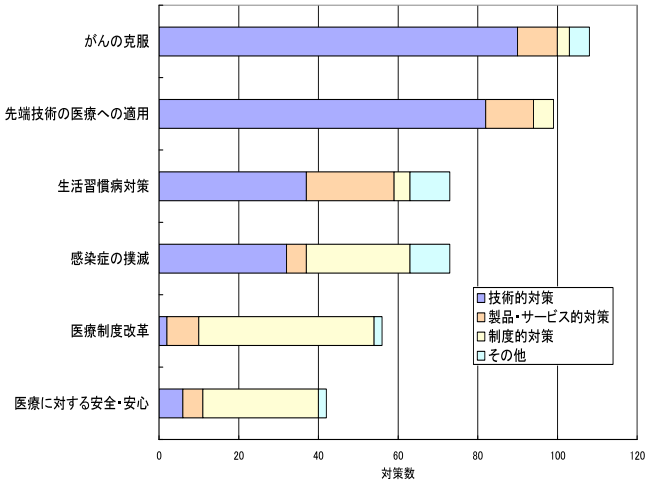


図３　医療分野における社会課題の注目度（2005 年）　図４　各課題に対する対策数と対策内容

６．おわりに

以上のように、一般的な情報検索では分析が不可能な大規模テキスト情報のデータセットから、階層的クラスタリング等の手法により分類・抽出されたクラスタあるいは重要キーワード等から、社会課題やそれらに対する技術的対策等の俯瞰的トレンドを効率的に把握する例を示した。また、新しく提示された密度や中心性といった指標を用いることにより、クラスタのスクリーニングや類型化に有効であることを検討した。今後は、さらに分析例を増やし、実効性の検証を行っていく予定である。

謝辞

本研究は、文部科学省科学技術振興調整費による「戦略的研究拠点育成プログラム」の支援の下に実施された。

参考文献

[1] 広瀬千夏、岩沼宏治、鍋島英知：背景記事集合の類似度に基づく新聞記事クラスタリング、電子情報通信学会技術研究報告、Vol.105、No.594 (NLC2005 106-113)、Page25-30 (2006)

[2] 福原知宏、中川裕志、西田豊明：感情表現と用語のクラスタリングを用いた時系列テキスト集合からの話題検出、人工知能学会全国大会論文集、Vol.20th、Page2E1-2 (2006)

[3] 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（投稿準備中）

[4] 日本経済新聞、日経シソーラス、http://telecom21.nikkei.co.jp/