社会課題発見のためのテキストマイニングシステム: RiverStone

橋本泰一 乾孝司 村上浩司 内海和夫 石川正道

東京工業大学 統合研究院

{hashimoto, inui, murakami, utsumi, ishikawa}@iri.titech.ac.jp

1 はじめに

社会の複雑化によって社会に不安と不信を引き起す要因が増加している．しかも，一つの問題に多くの主体が関与し，事件が起こるとその波及範囲が想定外の分野にも波及している [11,14]．このような社会課題の時系列変化についは，新聞記事に詳細に記載されている．新聞記事には出来事の発生から関連する具体的な事柄が日々蓄積され，時間の経過とともにこれが多方面に波及して行く過程の情報が豊富に含まれる．しかし，膨大な量の新聞記事を収集し，分類，分析することは大きな労力を要する．また，新聞記事から情報収集するには，情報検索技術を利用して，キーワードにより関連した個々の記事を発見する方法が一般的である．この場合，個々の記事からだけでは，課題全体の多様性を把握することが難しい．多くの新聞記事から関連する情報を多面的に獲得し，これをもとに変化する社会課題の発見に導く俯瞰的な分析手法の確立が望まれている [7]．

一方，コンピュータやインターネットの普及に伴い，電子化テキストが増加の一歩を辿っており，新聞記事を含め，手軽に大量の文書を入手することが可能になった．そのため，大量の文書から欲しい情報を獲得し，何らかの傾向を発見したいというニーズが高まっている．このニーズを満たすためにテキストマイニングに関する研究・開発が盛んに行われている．

テキストマイニングでは，文書中の語彙の出現分布により文書を表現する．2 つの文書が互いに類似した語彙の出現分布を持つ場合には，同一の話題（トピック）を扱っていると考えられ，大量の文書を自動的に類似した話題の文書群に分類したり，文書群の関係を構造化したりすることができる．

Uramoto らは，大量の新聞記事に対して，語彙の使用分布が類似した記事を関連づけることにより新しい発見を支援するシステムを提案した [4]．このシムテムでは，新聞記事を単語ベクトルとして表現し，ベクトルの類似性に基づき記事を関連づけ，グラフ構造として表示する．

本論文で提案するテキストマイニングシステムは，大量の記事集合から互いに類似した内容をもつ記事を自動的に処理し，分析者による課題発見の作業を容易とすることを目的とする．提案システムは，Uramoto らのシステムと同様に記事を単語ベクトルとして表現する．そして，そのベクトルの類似度を基にいくつかの記事集合（クラスタ）に分類し，構造化を行う（階層型文書クラスタリング）．さらに，重要なクラスタを特定するための技術として，クラスタのグループ化，クラスタの指標（密度・中心性）を実装した．

2 社会課題発見のためのテキストマイニングシステム: RiverStone

提案システムは，次に述べる手順により新聞記事における課題発見を行う．(図 1)

1. 新聞記事を全文検索し，分析の対象となる記事文書集合を取得．
2. 得られた文書集合に階層型クラスタリングを施し，文書を記事群（クラスタ）へ分類，構造化（デンドログラム）．
3. 個々のクラスタについて話題性の強い重要クラスタを判別．(密度，中心度)
4. クラスタを特徴づけるキーワードおよび主体（組織名）の抽出．
5. 俯瞰的に社会課題を分析者の視点を加えて発見．

本システムは，Ruby on Rails により実装を行っている．システムの解析結果の出力画面を図 2 に示す．

2.1 文書検索

本システムは，新聞記事データとして 1975 年から 2006 年までの日本経済新聞紙本紙を収録している．総記事数は約 480 万記事である．分析対象となる記事集合を選択するためキーワードにより新聞記事を検索す

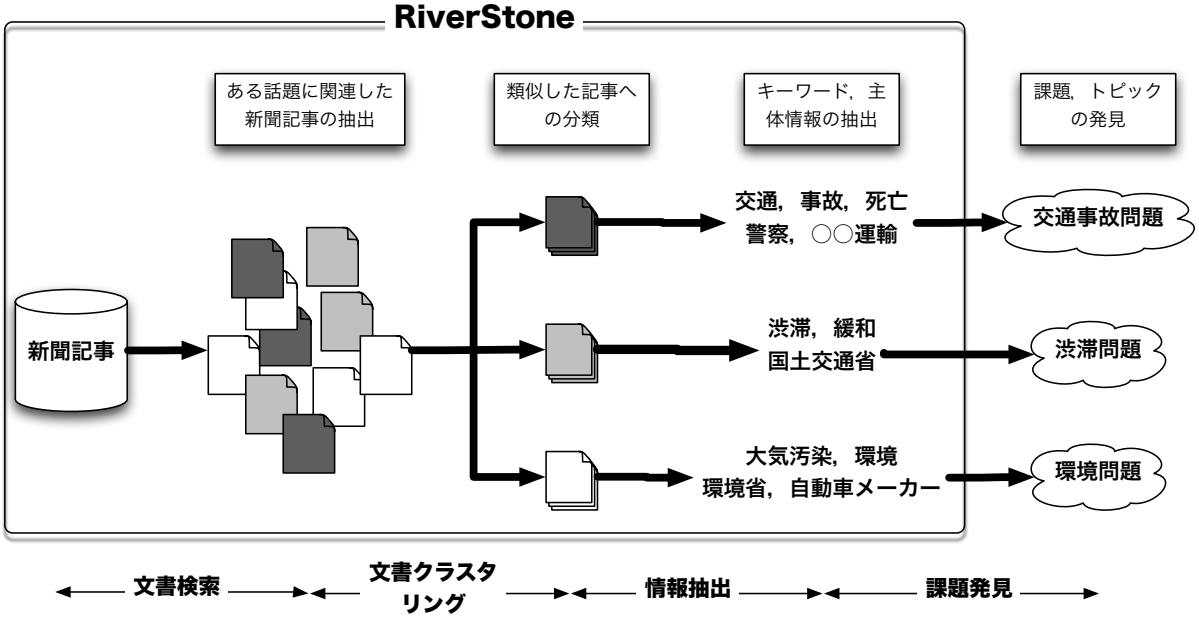

図 1: RiverStone における解析フロー

る機能を保有している．この検索機能は，全文検索システム Hyper Estraier [13] を利用し，N-gram 方式の文書検索を実現した．

2.2 文書クラスタリング

文書検索によって選択された分析対象の文書集合には，一般に，関連のある複数の話題が含まれる．これら文書集合に対して文書クラスタリングを実施し，話題が共通する文書をまとめる．

文書クラスタリングには，階層的ハードクラスタリング・アルゴリズムの一つである非加重結合法（Unweighted Pair Group Method with Arithmetic mean, UPGMA）[5,6] を採用した．文書の類似度はコサイン類似度により計算する．階層的クラスタリングでは，クラスタリング結果をデンドログラム（系統樹）として可視化でき，文書内容の類似性に基づいてクラスタ間の関係を俯瞰できるため，分析者が分析しやすいという利点がある．クラスタリングソフトウェア CLUTO [2] を用いて実現した．

文書クラスタリングにおいて各文書は，文書に含まれる語の情報から構成されるベクトル（文書ベクトル）として扱われる．記事の先頭から 5 文を形態素解析器 ChaSen [3] で解析し，解析結果の中から名詞および名詞が連続する語（名詞連続）のみを抽出することによって文書ベクトルを作成する．ただし，記事集合における頻度が閾値（本システムでは 5 とした） 未満となる名詞や名詞連続は考慮しない．また，文書ベクトルとして考慮される名詞，名詞連続を e で表すとし，文書ベクトル $\mathbf{d}$ の i 番目の要素の値 $w_d(e_i)$ を式 (1) で定義する．

$$w_d(e_i) = tf_d(e_i) \times \log_2 \frac{|D|}{df_D(e_i)} \times |e_i| \times \frac{1}{1+log(first_d(e_i))} \quad (1)$$

$tf_d(e_i)$ は文書 d 内での語 e_i の出現頻度，$|D|$ は総文書数，$df_D(e_i)$ は語 e_i の出現する文書数，$|e_i|$ は語 e_i を構成する文字数，$first_d(e_i)$ は文書 d 内で初めて語 e_i が出現した文の位置（$1 \leq first_d(e_i) \leq n$）を表す．この重み付けは，TF-IDF をベースとして，分析において重要な情報となる固有表現は文字数の長い語で構成されること，および，新聞記事では記事先頭に近い文ほど記事の概要的な内容が記述されやすいという特性を考慮している．

2.3 重要クラスタの選別

文書クラスタリングにより文書を分類できたとしても，分析したい内容と無関係な文書やまとまりの悪いクラスタからは，分析者は有益な情報を得ることができない．そのため，分析に有用なクラスタがどれであるのかを判別することが重要である．

Callon らは，共語分析において，語の共起によるネットワークをもとに語を人手によりクラスタリングし，クラスタ内の語の結びつきの強さ（密度）と他のクラスタと連結の強さ（中心度）を計算することによりクラ

図 2: RiverStone の出力画面

スタを評価し，分析する手法を提案している [1].

本システムでは，Callon らの提案した密度，中心度を階層的クラスタリングにより得られるデンドログラムをもとに新たに定義する．分析者は密度，中心度をもとに重要なクラスタを選別することで効率的に分析を試みることができる．

2.3.1 密度

密度 $density(c)$ をクラスタに含まれる記事に共通して出現する語により定義する（式 (2)）．これは，クラスタ内にある 2 つの文書間において共に出現する語（文書ベクトル作成の際に考慮された名詞および名詞連続）の数の割合に基づいて定義されており，共に出現する語の数が多いほど密度が高くなり，同一の話題を表す文書を多く含んでいると考えられる．

$$density(c) = \frac{\text{クラスタ } c \text{ 内の 2 つ以上の文書に出現する語の数}}{\text{クラスタ } c \text{ 内の文書に出現する語の数}} \tag{2}$$

2.3.2 中心度

クラスタがもつ話題の波及性や影響力を測る指標として，デンドログラムの構造情報をもとにした中心度 $centrality(c_l)$ を定義する．あるクラスタに対して，デンドログラム上で深さが深いノードで結合するクラスタが多いかどうか，結合したクラスタの文書数が多いかどうかを中心度の軸として考慮する．なぜならば，デンドログラム上での深いノードで結合するクラスタは，文書クラスタリングにおいて強い関連性を見いだせることを意味しており，クラスタに含まれる文書数は話題の波及性や影響力の高さを意味していると考えられるためである．具体的には，クラスタ c_i と c_j の関連度をデンドログラム上で c_i と c_j の共通する最初の祖先 $share(c_i, c_j)$ の深さ $depth(share(c_i, c_j))$ とクラスタ c_j の文書数 $|c_j|$ をかけた値で定義する．そして，その平均値をクラスタ c_i の中心度として定義する．

$$centrality(c_i) = \frac{1}{|C|} \sum_{c_j \in C} |c_j| \times depth(share(c_i, c_j)) \tag{3}$$

ただし，デンドログラムの根の深さは 0 とする．また，$|C|$ はクラスタの個数を表す．

2.4 情報抽出

クラスタがどのような記事の集合であるか把握するためにクラスタを特徴づけるキーワードと記事内に出現する組織名を抽出する．キーワードは，クラスタ内の記事に含まれる語ごとに式 (1) のスコアの和を計算し，最大上位 20 語をそのクラスタのキーワードとし

た．組織名抽出は，山田らが提案した教師あり機械学習に基づく固有表現抽出の手法を用いて抽出する [9]．

3 まとめ

本論文では，新聞記事に対してテキストマイニングの手法を応用し，解析の対象となる記事文書集合を取得して自動的にこれを分類して情報を抽出し，課題発見に至るテキストマイニングシステムを提案した．具体的には，大量な文書に対して階層的クラスタリングを施し，クラスタ間の語彙使用の類似性に基づく構造化を行い，クラスタを要約するキーワードおよび関係主体を抽出することによって内容を俯瞰することを可能とした．

我々は，本システムを用い，社会課題についての分析を行っている [10,12]．特に安全安心に関する社会変化のトレンド分析は社会の関心が高く，政策分析へのインプットとしても有益である．科学技術振興機構が提供する失敗知識データベース [8] が安全安心の解析に貢献していることは周知であるが，本システムを利用することで，大事故，大事件に対して社会がいかに応答し，変化していくかを動的に解析することも可能になる．この他，文書クラスタからの技術用語の自動抽出についても技術開発を進めており，課題と技術との相関，さらには科学技術研究の動向との関連性を評価するなど，トレンドを分析する技術の開発を目指す．

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．本研究に遂行にあたって有益なご助言をいただいた東京工業大学統合研究院下田隆二教授および大熊和彦教授に感謝いたします．

参考文献

[1] M. Callon, J. P. Courtial, and F. Laville. Coword analysis as a tool for describing the netword of interactions between basic and technological research: The case of polymer chemistry. *Sientometrics*, Vol. 22, pp. 155–205, 1991.

[2] Geoge Karypis. *CLUTO A Clustering Toolkit Release2.1.1.* University of Minnesota, Department of Computer Sciense, 2003.

[3] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, and M. Asahara. Japanese morphological analyzer chasen users manual version 2.0. Technical Report Technical Report NAIST-IS-TR990123, Nara Institute of Science and Technology, 1999.

[4] Naohiko Uramoto and Koichi Takeda. A method for relating multiple newspaper articles by using graphs, and its application to webcasting. In *Proceedings of the 36th annual meeting on Association for Computational Linguistics*, 1998.

[5] Ying Zhao and Geoge Karypis. Hierachical clustering algorithms for document datasets. Technical Report MN 55455, Department of Computer Science, University of Minnesota, Minneapolis, 2003.

[6] Ying Zhao and George Karypis. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, Vol. 10, No. 2, pp. 141–168, 2005.

[7] 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.

[8] 科学技術振興機構. 失敗知識データベース. http://shippai.jst.go.jp/fkd/Search.

[9] 乾孝司, 村上浩司, 橋本泰一, 内海和夫, 石川正道. 文書からの組織名抽出における辞書利用. 情報処理学会自然言語処理研究会, 第 NL182 巻, 2007.

[10] 橋本泰一, 村上浩司, 乾孝司, 内海和夫, 石川正道. モビリティ社会の安全安心に関わる社会課題の俯瞰的評価. 科学技術社会論学会第 6 回年次研究大会, 2007.

[11] 市川惇信. 発刊のことば. *ristexNEWS*, Vol. 1, , 2005.

[12] 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.

[13] 平林幹雄. Hyper Estraier. http://hyperestraier.sourceforge.net/index.ja.html.

[14] 堀井秀之. 安全安心のための社会技術. 東京大学出版会, 2006.