様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１５　日現在

機関番号：　２３８０３
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００４２９

研究課題名（和文）凝縮性に基づく有用単語群の検出と構造化技術の構築

研究課題名（英文）Detection of Annotation Words and Construction of Structured Infomations based on Cohesiveness

研究代表者

大久保　誠也（Seiya, OKUBO）

静岡県立大学・経営情報学部・講師

研究者番号：９０４２２５７６

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：本研究の目的は、文章ビックデータを分類し、各分類に適切な注釈語（アノテーションワード）を付与することである。本目的を達成するため、クラスタ係数の概念を発展させた凝縮性という指標を導入し、それに基づいた分類や注釈語付与手法を提案した。提案手法の妥当性を評価するために、計算機実験や実証評価を行った。実証評価では、近代新聞検索サイトに適用し、広く公開することにより評価を行った。計算機実験では、Web上に存在するさまざまなデータに適用することにより、妥当性の評価を行った。また、文書データ以外のデータにも適用範囲を広げ、農業環境データ等への応用も行った。以上のことから、提案手法の有用性が明らかとなった。

研究成果の概要（英文）：The purpose of this study is to classify document data and give annotation words for each classification. In order to achieve this purpose, we introduce an index called "Cohesiveness" that developed the concept of cluster coefficients, and proposed a classification and annotation word assignment method based on it. In order to evaluate the validity of the proposed method, we performed computer experiments and empirical evaluations. In the empirical evaluation, we applied it to the data set of KindaiShinbun articles. In computer experiments, validity was evaluated by applying it to various data existing on the Web. In addition, the scope of application was extended to data other than document data, and it was also applied to agricultural environmental data etc.From the above, the usefulness of the proposed method became clear.

研究分野：　量子計算

キーワード：　凝縮性　情報構造化　アノテーションワード　文書クラスタリング

１．研究開始当初の背景

インターネットサービスの急速な発展により、日々、大量の文章データがやりとりされるようになった。これらのデータは、従来の手法では全体像を把握することは困難である。そのため、この文章ビックデータから、重要かつ特徴的な話題（トピックス）の文章集合を自動検出し、適切な解釈後（アノテーション）を付与する技術が必要である。この基本課題は、文章クラスタリングとその解釈の本質的な難しさ（M. Hearst, "Search user interfaces," Cambridge University Press, 2009）を根本問題として包含しており、トピック検出と追跡（J. Allan, “Topic detection and tracking: event-based information organization," Kluwer Academic Publishers, 2002）や、ホットトピック抽出（J. Kleinberg, "Bursty and hierarchical structure in streams," Data Mining and Knowledge Discovery, 7: 373-397, 2003）などを端緒に、多様な研究が行われている。

２．研究の目的

本研究では、複雑ネットワーク分析アプローチを土台に、この根本問題の解決に向けた新たな方法論を確立する。具体的には、与えられた文章データに対して、文章や出現単語をそれぞれノードとする二部グラフを構築することで可視化するとともに、重要な単語をアノテーションワードとして提示する手法を確立する。

重要な単語を抽出する際には、クラスタ係数（D. Watts, S. Strogatz, “Collective dynamics of 'small-world' networks, ” nature 393: 440-442, 1998）の概念を発展させた凝縮性（cohesiveness）と呼ぶ指標を新たに導入することで、各単語の有用性を定量化する。クラスタ係数は、各ノードが隣接する任意のノードペア間にリンクが存在する期待（平均）値で定義されるのに対し、単語の凝縮性は、文章全体での平均類似度と比較し、その単語を含む（文章-単語の二部グラフで隣接する）文章ペア間の平均類似度が有意に大きいかのｚスコアで定義する。直感的には、各単語と隣接する文章集合の任意のペア間に類似度で重み付けされたリンクが付与され、その文章集合の重み付き完全結合ネットワークの平均密度により凝縮性は定義される。一般に、凝縮性の高い単語は、類似した内容で共通する話題を持つ文章集合と隣接する傾向となり、非隣接文章には出現しない識別的性質より、その単語自体が文章集合の適切な解釈語になることが大いに期待できる。ただし、凝縮性の高い単語は一般に複数存在し得る。よって本研究では、凝縮性に基づいた有用な単語群を検出する技術の確立とともに、それら単語と隣接する文章集合の類似度などによる、検出した単語を適切に構造化する技術の確立を課題とする。

以上を踏まえた本研究の具体的な目的は、

(1). 凝縮性の高い単語群の検出技術の確立とその特性や有効性の評価
(2). 凝縮性の高い単語群の構造化技術の確立とその結果の有用性評価
(3). 確立した検出技術と構造化技術の近代新聞検索サイト等での実証評価

の３つである。

３．研究の方法

各目的を達成するため、多数の対象に対して提案手法を適用することにより、その評価と改善を行った。具体的には、各目的に対し、以下のような方法をとった。

(1). 観光情報サイトやSNSのデータに対して提案手法を適用することにより、妥当なアノテーションワードが抽出できるか否かの検討を行った。また、単語や文章の重要度は、その発信者の属性や影響度にも強く影響を受けると考え、ネットワーク上で重要な役割を果たすユーザを検出する手法について検討を行った。
(2). (1)と同様に、観光情報サイトやSNSのデータに対して提案手法を適用することにより、その評価と改善を行った。また、いくつかのネットワーク可視化手法を比較し、各手法にどのような特長があるかの評価を行った。
(3). 上記の研究で得られた成果に基づいた検索サイトを構築し、広く公開することによる実証実験を行った。その際、アンケート調査を行うと共に、ユーザどのような行動を取るのかの情報の収集も行うことで、従来手法との差を評価した。

加えて、以下の方法により、提案手法を文章以外のデータに対して適用する手法について検討を行った。

(4). 様々な非文章データに対して提案手法を適用することにより、有効性の評価を行った。適用対象は非文章データであるため、単語としてのアノテーションワードは無い。しかしながら、時系列データにおける急激な変化点やユーザー行動の特長が、その分類の特長であると考え、それらの特長により分類の説明ができるかを検討した。

４．研究成果

(1)については、複数のデータに対して提案手法を適用することにより、その評価を行った。まず、観光デビューサイトに対して適用を行った。その後、アニメや電化製品、映画、レストラン等、さまざまなレビューサイトのデータを用いて、提案手法の妥当性を検証した。これらの結果から、提案手法は妥当なアノテーションワードを抽出しているという結果が示された。

また、重要な単語を選ぶ際、すべてのノード（ユーザーや記事）の影響度は一定ではないと考えた。つまり、より影響度の大きいノードは、より強い影響力を持つため、集団等

の特長となりやすいと考えた。そこで、ユーザーの特長や影響度を評価する手法について検討を行った。まず、各ユーザーがどのような特長を持つか分析するため、自己中心トライアドという指標を提案するとともに、レビューサイトのデータに適用することによる評価を行った。また、レビューサイトにおいて、より強い影響力をもつユーザーを抽出する方法を考案し、SNS データに対して適用することによる評価を行った。

(2)については、(1)と同様のデータに対して適用することにより、提案手法の評価を行った。特に、(1)のアノテーションワードの付与は、分類後の集合に対して行われるため、(2)における分類結果に強い結果を受ける。そこで、さまざまな類型化法を用いることにより、どのような特長を持つかについての検討を行った。例として、同じ観光情報データを、MST を用いた手法・PAC を用いた手法・KNN を用いた手法の 3 つを用いて可視化した結果を、図 1 から図 3 に示す。各種法で、可視化結果がかなり異なっていることがわかる。それにともない、アノテーション結果も変化している。これらの研究により、各可視化方法の特長が明らかになるとともに、RNG を用いると良いという結論を得た。

図 1 MST を用いた可視化

図 2 RNG を用いた可視化

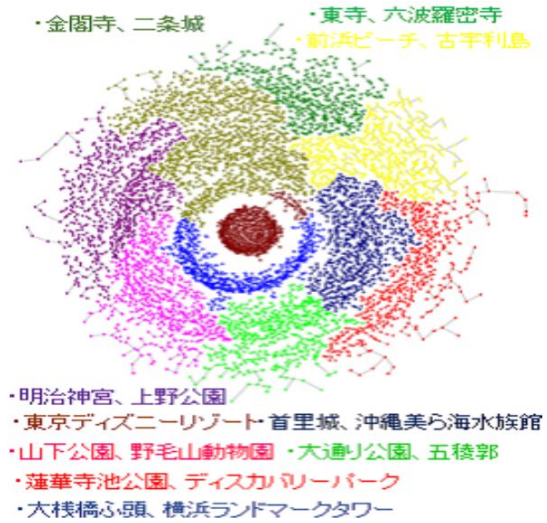

図 3 KNN を用いた可視化

(3)については、静岡県における近代の新聞記事を検索できるサイトである "近代新聞検索"（https://www.npoabc.jp/）と連携し、近代新聞検索の記事を検索できるサイトを構築した。構築したサイトは、本研究により得られた成果に基づいており、幾つかの特長を備えている。キーワードを入力して検索を行うと、関係の高い記事の一覧が表示されるところは、従来の検索手法と同様である。ただし、記事を選択した際、本サイトでは、各記事の関連性がグラフとして表示される（図 4 参照）。中央に選択した記事のノードが表示され、そこから文章の関連度にしたがって、記事の繋がりが可視化される。また、グラフは彩色されることによって分類されており、各分類の特長となるアノテーションワードも表示される。グラフは端に行くほど、元の記事とは離れた情報となり、関連する情報のみならず、関連する情報に関連する情報までを一望することができる。ユーザーが各ノードを選択すると、その記事が表示され、希望すれば新しいノードを中心としたグラフを表示することができる（図 5 参照）。この機能により、ユーザーは検索したキーワードに関する記事だけにとどまらず、提示されたグラフから得た新しい気づきに基づき、知識を得ることができる。このため、セレンディピティ性が高い検索手法であるとして、気づき型検索システムと呼んでいる。

このようにして構築されたサイトは、実証実験として、2015 年 11 月から 2016 年 2 月末にかけて広く公開された。その際、各ユーザーから感想を収集するとともに、ユーザーがどのように検索し、どのような行動を行うかのログも収集した。これらの結果から、提案手法はセレンディピティ性が高いという結果が得られた。一方で、ユーザーインターフェース等の改善点も明らかとなった。

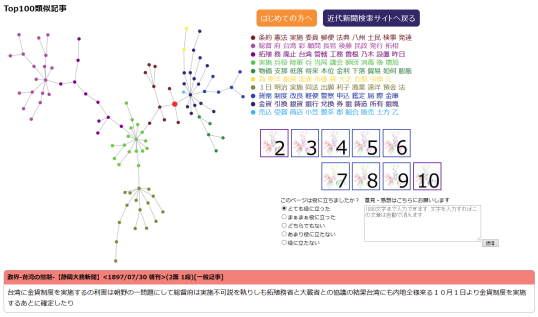

図 4 構築したサイトで記事を選択した場面

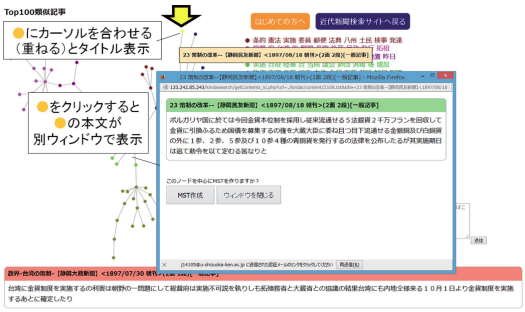

図 5 記事の選択を行った場面

(4)については、非文章データとして、特に時系列データに対する検討を行った。具体的には、レビューサイトや環境データに対して適用することにより、対象となるオブジェクトの分類や、各分類の特徴付けを行った。時系列データは、変化の類似による類似度の評価と分類はできるものの、非文章データであるため、分類結果を直接説明する単語は存在しない。しかしながら、時系列においては、データの変化の仕方、特に急激な変化点が特長であると考えた。そこで、急激な変化点を自動抽出する手法を構築した。

時系列データの分類・可視化の例として、野菜の月別市場データのうち、取引数量に基づいて評価を行った結果を、図 6 に示す。図では、上の方に季節野菜である八つ頭と、それに類似した出荷が行われるグループが存在している。また、変化点抽出の例として、平均湿度の変化点検出について、図 7 に浜松市の結果を、図 8 に静岡市の結果を示す。図中、青い線が湿度の変化、赤と緑の線が提案手法による変化点の検出である。平均湿度を示す青い線を見ただけでは、どのような差があるのか分かりづらいグラフであるが、変化点を検出することにより、どのような特長を持つのかがわかりやすくなっていることがわかる。

これらの結果より、文章データだけではなく、時系列データに対しても分類や特徴付けができるようになった。

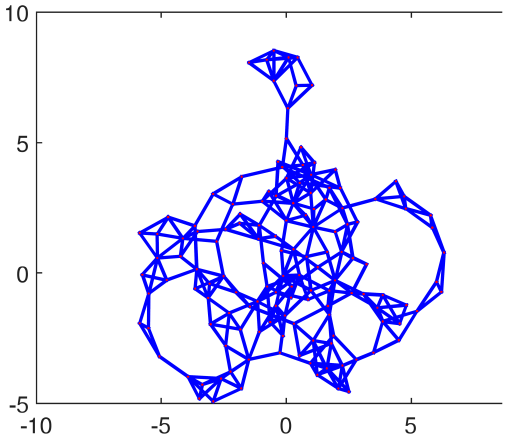

図 6 野菜の月別市場データを可視化結果

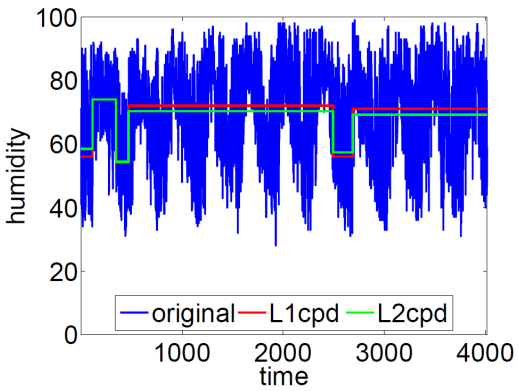

図 7 浜松の平均湿度の変化点抽出

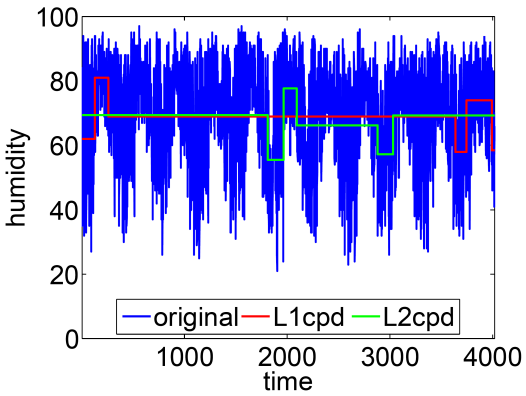

図 8 静岡の平均湿度の変化点抽出

また、本研究で得られたグラフの解析手法を、地図データに応用する方法についても検討を行った。地図データを考える際、道路網はグラフ、交差点はノードと捉えることができる。また、地図データにおける重要地点は、本研究で注目していた重要なノードと考えることができる。そこで、各都市の特性の分析や、重要地点が閉塞した時の影響度評価などを行った。他にも、産業分析への適用についての検討も行った。

以上をすべてまとめると、本研究により、各種文章データを分類・可視化するとともに、各文章集合に対して妥当なアノテーションワードを付与することが可能となった。それに加え、時系列データ等の文章データ以外のデータに対しても分類・可視化するとともに、そのデータの特長を捉えることが可能となった。今後、更なる研究により、適用範囲の拡大が期待できる。

５．主な発表論文等

〔雑誌論文〕（計 2 件）

① 山岸祐己，岩崎清斗，斉藤 和巳，"多群出現順位統計量に基づく時系列データの変換，" 情報処理学会論文誌，数理モデル化と応用，Vol.11，No.1，pp.45-52,May．2018．

② 小林えり，斉藤 和巳，大久保 誠也，池田 哲夫，渡邉秀博，"アノテーション付き可視化法を用いた気づき型検索システム，" 経営と情報，Vol.28，No.2，pp.1-10，Mar．2016．

〔学会発表〕（計 17 件）

① 田中雄大，斉藤 和巳，岩崎清斗，大久保誠也，"類似ネットワークによる市場データの分析，"第 16 回情報科学技術フォーラム（FIT2017），Sep.2017.

② 我妻勇貴，斉藤 和巳，岩崎清斗，大久保誠也，"経済物理アプローチによる市場データの分析，"第 16 回情報科学技術フォーラム（FIT2017），Sep.2017.

③ 小野拓海，斉藤 和巳，岩崎清斗，大久保誠也，"変化点検出法による農業環境データの分析，"第 16 回情報科学技術フォーラム（FIT2017），Sep.2017.

④ 鈴木一矢，斉藤 和巳，岩崎清斗，大久保誠也，"K-medoids 法による農業環境データの分析，"第 16 回情報科学技術フォーラム（FIT2017），Sep.2017.

⑤ 楊小龍，斉藤 和巳，"上流度の流れ図可視化法による産業分析,"情報処理学会第 79 回全国大会(IPSJ2017), Mar.2017.

⑥ 大畑圭佑，斉藤 和巳，"埋め込み手法の違いによるアノテーション付き可視化の特性評価,"情報処理学会第 79 回全国大会(IPSJ2017), Mar.2017.

⑦ 白澤穂香，斉藤 和巳，"実距離とステップ距離に基づく近接・媒介中心性による都市特性の分類,"情報処理学会第 79 回全国大会(IPSJ2017), Mar.2017.

⑧ 鈴木優人，斉藤 和巳，"投票者モデルに基づくレビュー回数ユーザ重み影響度分析,"情報処理学会第 79 回全国大会(IPSJ2017), Mar.2017.

⑨ 塚本 竜太郎，斉藤 和巳，"道路閉塞による迂回影響分析，"第 15 回情報科学技術フォーラム（FIT2016），Sep.2016.

⑩ 鈴木 優人，斉藤 和巳，"投票者モデルに基づく複数レビューサイトでの影響度分析，"第 15 回情報科学技術フォーラム（FIT2016），Sep.2016.

⑪ 大畑 圭佑，斉藤 和巳，"アノテーション付き可視化によるユーザ行動分析，"第 15 回情報科学技術フォーラム（FIT2016），Sep.2016.

⑫ 楊 小龍，斉藤 和巳，"産業連関表を用いた業種の上流度分析，"第 15 回情報科学技術フォーラム（FIT2016），Sep.2016.

⑬ 永倉 卓弥，小林 えり，斉藤 和巳，"投票者モデルに基づくレビュー順序影響度分析，"情報処理学会第 78 回全国大会(IPSJ2016), Mar.2016.

⑭ 後藤 裕，小林 えり，斉藤 和巳，大久保誠也，池田 哲夫，"多種情報を利用したアノテーション付き可視化法，"情報処理学会第 78 回全国大会(IPSJ2016), Mar.2016.

⑮ 永倉 卓弥，小林 えり，斉藤 和巳，"投票者モデルに基づくユーザ影響度分析，"第 14 回情報科学技術フォーラム（FIT2015），Sep.2015.

⑯ 西 可南子，斉藤 和巳，池田 哲夫，大久保 誠也，"自己中心トライアド変化曲線によるユーザ分析，"第 12 回 ネットワーク生態学シンポジウム（NETECO2015），Aug.2015.

⑰ 樋岡 真菜美，斉藤 和巳，大久保 誠也，池田 哲夫，"凝縮性に基づく単語検出法による観光レビュー記事の分析，"第 12 回観光情報学会全国大会（STI2015），Jun.2015.

６．研究組織

(1)研究代表者

大久保 誠也（OKUBO，Seiya）
静岡県立大学・経営情報学部・講師
研究者番号：90422576

(2)研究分担者

池田 哲夫（IKEDA，Tetsuo）
静岡県立大学・経営情報学部・教授
研究者番号： 60363727

風間 一洋（KAZAMA，Kazuhiro）
和歌山大学・システム工学部・教授
研究者番号： 60647204

斉藤 和巳（SAITO，Kazumi）
静岡県立大学・経営情報学部・教授
研究者番号： 80379544

湯瀬 裕昭（YUZE，Hiroaki）
静岡県立大学・経営情報学部・教授
研究者番号： 30240162