ニュース記事の地域度分析に基づく地図への自動分類

藤崎　日和†　　荘司　慶行††　　田中　克己†

† 福知山公立大学 情報学部　〒 620 – 0886 京都府 福知山市 字堀
†† 青山学院大学 理工学部　〒 252 – 5258 神奈川県 相模原市 中央区 淵野辺
E-mail: †{32045083,tanaka-katsumi}@fukuchiyama.ac.jp, ††shoji@it.aoyama.ac.jp

あらまし　ニュース記事の「地域度」を，記事に出現する住所や緯度経度情報から計算した言及地域と記事内容の言及地域への関連性（話題の「ユビキタス度」（どこにでも出ている度合い））を用いて計算し，この地域度を基にしてニュース記事を地図上にマッピングするシステムを提案する．

キーワード　ニュース記事，地域度，自動分類，地理情報システム

1　はじめに

ニュース記事の閲覧環境は，現在，時代と共に急速に変化しつつある．初期のニュース記事の閲覧環境は紙媒体の新聞（全国紙/地方紙）や TV（キー局/ローカル局）であり，全国/地域に関するニュース記事は，それぞれ，全国紙/地方紙（キー局/ローカル局）を閲覧することで地域の情報を取得できていた．その後，各新聞社の Web サイトや Yahoo!ニュース等のニュース閲覧サイトでの閲覧など，ニュース記事の閲覧環境は急速に変化しつつある．ニュース閲覧サイトから発信されるニュース記事は，量的に記事数が少ないことに加えて，その殆どが，ジャンル毎に分類されて表示され地域毎に分類されていないため，ある地域に関するニュース記事を見ることは容易ではなくなってきている．

本論文では，ニュース記事の閲覧・検索環境として，地図をベースとしたインタフェースを提案する．ニュース記事がどの程度地域の情報であるかを示す「ニュース記事の地域度」の概念を導入し，記事に出現する住所や緯度経度情報と記事の話題の「ユビキタス度」（どこにでも出ている度合い）を用いて地域度を計算し，この地域度に基づいてニュース記事を地図上のある地域にマッピングすることを試みる．ニュース記事の地域度分析に基づく地図への自動分類のプロトタイプを，地理情報システム QGIS 上で開発中であり，地図をベースとしたニュース記事の閲覧が可能となりつつあるので報告する．ニュース記事の収集については，手動で収集するほか，Octoparse を用いたニュース記事のテキスト抽出（https://www.octoparse.jp）などの利用を予定している．

本論文の構成は以下のようになっている．第 2 節では，関連研究を紹介し，本研究の位置づけを示す．第 3 節では，本論文で提案する「地域度」の定義や計算方法をを示す．第 4 節では，計算された地域度を基にニュース記事を地図にマッピングし閲覧する手法やインタフェースについて述べる．第 5 節では，まとめと今後の課題について述べる．

2　関連研究

Web ニュース記事の地域度分析については，松本・馬らの Web ニュース記事の地域度計算に関する研究 [1] [2] がある．松本・馬らは Web ニュース記事の地域度を，記事内容の地域度（Localness of Contents) と，記事の話題の遍在性（Ubiquitousness of Topics）を用いて計算する手法を提案し，実際に ASAHI.COM に掲載されたニュース記事 500 件に対して地域度分析を行っている．

著者らは，災害報道が全国紙/地方紙によってどのように異なっているかを比較分析（「朝日新聞/神戸新聞の災害報道の比較」[3]）するとともに，本研究の先行研究 [4] も開始している．

ニュース記事のトピック分析・感情分析による地図マッピングについては，河合・熊本らの先行研究 [5] [6] [7] がある．

3　ニュース記事の分析

3.1　ニュース記事の収集とテキスト抽出

本研究では，yahoo!JAPAN ニュース (https://news.yahoo.co.jp) において,「熱海　土砂」という検索をかけ「熱海市伊豆山土石流災害」に関するニュース記事を収集したものを実験に使用した．この意味で，本研究で使用するニュース記事テキストの収集とテキスト抽出は，手作業で収集・抽出を行ったものである．

今後のニュース記事テキストの収集とテキスト抽出は Web スクレイピングツール Octoparse（https://www.octoparse.jp）の使用を予定している．Octoparse にニュース記事の URL を入力すると，記事のタイトル，内容，投稿日時などが抽出され，Excel や csv ファイルに変換できる．

3.2　ニュース記事の内容分析

ニュース記事の地域度を算出するために、ニュース記事の内容分析を，ニュース記事集合から次の T 値，MI 値，及び，TF-IDF 値を計算により行う．

- **T 値**

T 値は単語の共起関係の有無を調べる簡便な指標であり，以下の式で求められる．T 値の絶対値が 2 以上であると有意である

図 1　Octoparse によるテキスト抽出

とされる．

T 値＝ (実測値－期待値)/実測値の平方根

T 値とはある単語に対して，共起関係について調べるものであり，今回は,「熱海」,「土砂」という単語との共起関係を調べる．

Term	Before	After	Span	Total	T	MI
市	0	137	137	176	11.0827442	4.2341322
静岡	50	2	52	86	6.7178108	3.8697067
県	51	0	51	156	6.2378900	2.9825549
で	7	62	69	504	5.7969743	1.7267763

図 2　熱海という単語において T 値の高いもの

- **MI 値**

MI 値は，ある記号が出現することが、別の特定の記号の出現を予測させる度合いで，以下の式で求められる．

MI 値＝ log2(共起回数/共起語の期待値)

MI 値は低頻度の単語であっても共起関係を抽出することができるという特徴を持つ．MI 値も T 値と同様に,「熱海」,「土砂」という単語との共起関係を調べる．

25	おきる	0	1	1	1	0.9586375	4.5955317
32	ごとし	1	0	1	1	0.9586375	4.5955317
69	アーケード	1	0	1	1	0.9586375	4.5955317
71	マリンスパ	0	1	1	1	0.9586375	4.5955317
85	伊東	0	1	1	1	0.9586375	4.5955317

図 3　熱海という単語において MI 値の高いもの

- **TF-IDF 値**

TF 値とは局所的重みと呼ばれ，ある文書内である単語がどれくらい多い頻度で登場するのかを表し，以下の式で求められる．

TF 値＝文書内の単語の出現回数/文書全てでの単語の出現回数

IDF 値とは大域的重みと呼ばれ，分析対象とする文書全体のうち，いくつの文書に出現しているのかを識別する指標であり，以下の式で求められる．

IDF 値＝ log(総文書数/ある単語が出現する文書数)

それぞれのニュース記事に対し，記事内に含まれる単語がその記事内でどれだけ重要であるのかということを示す値である．これによりニュース記事の報道内容について考察することができる．

そして，多数のニュース記事を TF-IDF 値を利用して，ニュース記事の内容によりクラスタリングを行う．これによりニュース記事の分類ができる．今回は，階層的クラスタリングを用いクラスタ間の距離はウォード法を利用している．

	data1.csv
駅	0.47356646
運転	0.41364501
線	0.29597904
成田	0.27624710
上下	0.19731936
線路	0.18461483
間	0.14515096
頃	0.13972454
伊東	0.13812355
倉見	0.13812355
全線	0.13812355

ニュース記事（data1）におけるTF-IDF値の高い語

図 4　ニュース記事に出現する tf/idf 値が高い語

3.3　ニュース記事の地域度分析

本研究では，各ニュース記事の「地域度」を計算し，計算された地域度に基づいてニュース記事の地図へのマッピングを行っている．

ニュース記事の「地域度」とは，直観的には，ニュース記事の「地域」への関連度を表す指標である．ここで,「地域」については，市町村，県，圏，国などと言った階層性があることに注意が必要である．

例えば，ニュース記事 N が与えられた場合，ニュース記事 N の「地域度」は，記事 N が言及している地域を推定すると共に，記事 N の言及地域への関連度の両方を考慮して決定する．

– **記事の言及地域の同定**

対象とするニュース記事中に出現する地域名（住所等）を抽出し，頻出する地域名を発見する．発見した地域名が複数ある場合は，最も多く登場している地名を選択する．今回はニュース記事全てに含まれている固有名詞の頻出回数を求め，最も多く報道されている地点はどこなのかを特定する．

– **記事の言及地域への関連度**

記事の言及地域への関連度とは，その記事内容が言及地域にどの程度関連しているのかを表す度合いである．この関連度を計算するために，記事の「ユビキタス性」を考慮する．記事の「ユビキタス性」とは，その記事内容が日常の生活情報になっているかを表す指標であり，記事のユビキタス性が高ければ，この記事の「言及地域への関連度」は大きいものと判断する．これは，日常生活情報など日常性の高い情報は何時でも何処でもありうる情報なので，特定の地域の人々にしか興味を持たれないため，これらの記事はその言及地域への関連度が高いと判断する．これらを判断するために，T 値や MI 値，TF-IDF 値などを利用してニュース記事の内容を考慮する．

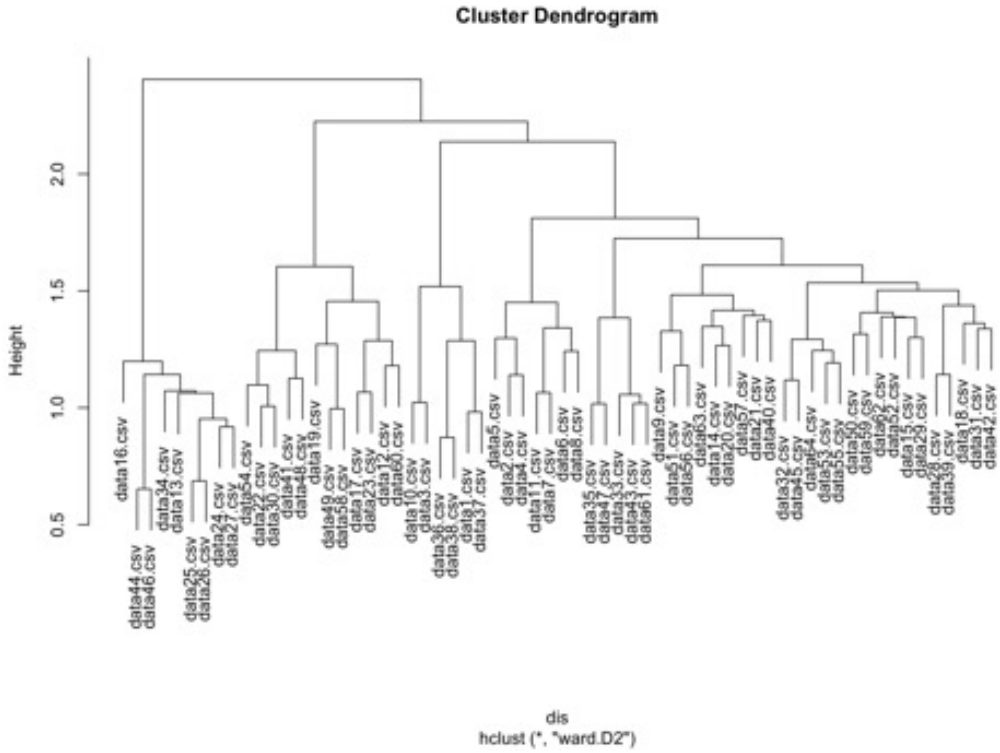

図 5　ニュース記事の階層的クラスタリング

```
> freq31 %>% filter(POS1 == "名詞" , POS2 %in% c("固有名詞"))
                  TERM POS1      POS2 data31.csv
33  ニューフジヤホテル 名詞 固有名詞           1
46                光広 名詞 固有名詞           2
69                宮城 名詞 固有名詞           1
104             東日本 名詞 固有名詞           2
107               沼津 名詞 固有名詞           1
110                 渚 名詞 固有名詞           1
117               熱海 名詞 固有名詞          16
155               銀座 名詞 固有名詞           1
157               静岡 名詞 固有名詞           3
```

記事データ（data31)に含まれる固有名詞と出現回数

図 6　記事データからの固有名詞の抽出

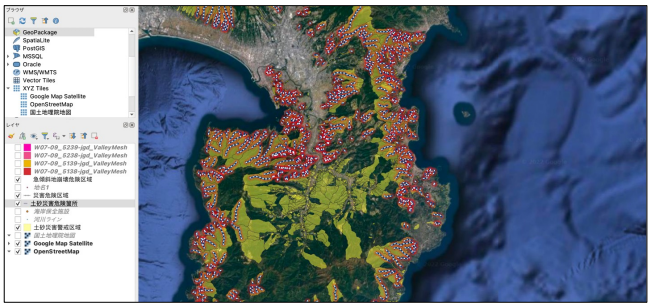

図 7　災害危険区域の地図マッピング

4　ニュース記事の地図マッピング

各ニュース記事を地理情報システム（QGIS）上にマッピングするために，まずは，言及地域の地理情報について調べる．国土交通省の国土数値情報を利用し、災害危険区域・土砂災害危険箇所・土砂災害警戒区域・急傾斜地崩壊危険区域を Google Map Satellite 上にマッピングする．

そして、地理情報が出力された QGIS 上にニュース記事をその地域度に基づいてマッピングする．具体的には，以下の手順でニュース記事のマッピングを行う．

（1） 記事の言及地域の同定を行う．具体的には，記事に出現する住所や地名・駅名などの位置を表す文字列を，緯度経度の座標値に変換するアドレスマッチングを用いて言及地域の同定を行う．

（2） 各ニュース記事をクラスタリングされたもので色分けをし，地図上の点で表示配置する．同一地域にマッピングされた記事は，サークルで記事数を表現する．

（3） 地図上にマッピングされた点 (ニュース記事) のタイトルを地図上に重畳表示する．

5　ま　と　め

本論文では，ニュース記事の「地域度」を提案した．ここで「地域度」とは，記事が言及している地域と，記事の言及地域への関連度の両方を考慮して計算するものである．具体的には，記事に出現する住所や緯度経度情報から言及地域の同定を行い．さらに，記事内容の話題の「ユビキタス度」（どこにでも出ている度合い）を用いて記事の言及地域への関連度を計算するものである．さらに，本論文では．計算された地域度を基にしてニュース記事を地図上にマッピングするシステムを構築した．これらの研究により，災害により報道された地点は，あらかじめ災害が起きやすいと考えられていた地域なのか，地理的特徴があるのかを考察することができる．この考察から，災害

図 8　ニュース記事の地図マッピング

図 9　ニュース記事の地図マッピング（記事タイトル付き）（OpenStreetMAP 上に表示）

防止策や災害予測につながると考えている．今後の課題としては，地域度の計算をわかりやすく効果的に行うこと，同じ地点にニュース記事が存在した場合などの地図マッピング方法の改善，マッピングの結果から，災害が起きやすい地形・特徴を分析・予測，災害ニュース記事だけではなく，交通事故のような他のトピックニュースについて分析，が挙げられる．

謝　　辞

本研究は，令和 4 年度科研費基盤研究 (B)（一般）「機械学習による情報の意味獲得と意味類似情報の検索・生成」（代表：田中克己，課題番号 22H03905）によるものである．

文　　献

[1] Chiyako Matsumoto, Qiang Ma, and Katsumi Tanaka. Web information rewtrieval based on the localness degree. In *Proceedings of DEXA2002, Lecture Notes in Computer Science 2453*, pp. 172–181, 2002.

[2] Qiang Ma, Chiyako Matsumoto, and Katsumi Tanaka. A localness-filter for searched web pages. In *Proceedings of 5th Asian-Pacific Web Conference(APWeb) , Lecture Notes in Computer Science 2642*, pp. 525–536, 2003.

[3] 藤崎日和. 新聞記事の比較分析　地方紙/全国紙，経済紙/一般紙の比較分析 ，2021 年度福知山公立大学地域情報 pbl 基礎最終報告書, 2021.

[4] 藤崎日和, 田中克己. ニュース記事の地域性分析に基づく地図への自動分類. 東海関西データベースワークショップ 2022, p. https://sites.google.com/mil.doshisha.ac.jp/dbws2022, 2022 年 9 月.

[5] 張建偉, 河合由起子, 熊本忠彦, 田中克己. 地域性に基づく発信者の観点差異を可視化するセンチメントマップシステムの提案. 情報処理学会論文誌 データベース, Vol. 3, No. 1, pp. 38–48, 2010.

[6] 張建偉, 河合由起子, 熊本忠彦, 白石優旗, 田中克己. 多様な印象に基づくニュースサイト報道傾向分析システム. 知能と情報（日本知能情報ファジイ学会誌）, Vol. 25, No. 1, pp. 568–582, 2013.

[7] 熊本忠彦, 河合由起子, 田中克己. 回帰分析を応用したテキスト印象マイニング手法の設計と評価. In *The 24th Annual Conference of the Japanese Society for Artificial Intelligence*, pp. 1B1–1, 2010.