1N3-OS-39b-1

DBpediaを用いた社会課題タグ自動付与APIの試作

Prototyping Web API for Automatic Annotation of Social Problem Tags using DBpedia

渡辺 賢 *1　　白松 俊 *1
Masaru Watanabe　Shun Shiramatsu

*1名古屋工業大学 大学院工学研究科 情報工学専攻
Department of Computer Science, Graduate School of Engineering, Nagoya Institute of Technology

Civic tech, which refers to activities for addressing social problems using information technologies through collaboration between citizens and IT engineers, is recently practiced in various localities in Japan. In the civic tech activities, sharing background information behind social problems enables people to invent innovative solutions. In this study, we assume that tagging hierarchical social problems on articles or council minutes contributes to sharing such background information. In this paper, we develop a Web API for automatic annotation of social problem tags and deal with two technical issues: (1) extracting ontologies of social problems from the category "Social Problem" in DBpedia Japanese and (2) selecting social problem tags suitable for an article. Support vector machine is used for (1) and Paragraph Vector is used for (2).

1. はじめに

近年，市民とITエンジニアが連携し，情報技術を活用して社会問題へ取り組むシビックテックと呼ばれる活動が日本の様々な地域で実施されている．シビックテックでは社会問題の背後にある背景知識を共有することで，その問題に関する革新的な解決アイデアの発想に繋がる可能性がある．そのために地域の社会問題を共有するサイトや，社会問題を解決するためのアイディアを共有するサイトが作られている．しかし，それらのサイトのほとんどはタグ付けが手動で行われており，事前に用意されているタグが無かったりと十分な量であるとは言い難い．

本研究では，ある社会問題に関する文章を与えた時，それを表すのにふさわしいタグを自動で付与するシステムの構築を目指し，適切な社会問題のタグの自動付与方法を分析する．そのために，社会問題のタグの自動付与方法を以下の2つの観点から分析した．

(1). DBpediaを用いた社会問題のタグ候補の抽出

(2). 社会問題に関する記事に対して最適なタグを社会問題のタグ候補から選択

(1)では，DBpediaからSPARQLクエリによって社会問題に関するタグ候補を取得した．更に，フィルタを用いた手法とSVMを用いた手法の2手法を用いて，DBpediaから取得したタグ候補から社会問題ではないと判断されるものを省いた．これらの手法に対して評価実験を行い，再現率，適合率，F値を評価した．

(2)では，予め用意した理想的な社会問題タグの候補102個のうち，1つのタグ候補と関連性が見受けられる記事を10個用意した．これと102個のタグ候補に対し2つの手法で類似度を計算した．類似度が閾値を超えたものを記事に対する付与タグ，そうでないものを非付与タグと定め，これに対する正解率を被験者から得たアンケート結果を用いて評価した．また，同時に相関係数も求めた．

連絡先: 渡辺賢, 〒466-8555 名古屋市昭和区御器所町 名古屋工業大学つくり領域 白松研究室.

2. 社会問題オントロジーの構築

本研究ではオントロジー [溝口 99] を用いた社会問題タグ候補の自動生成を取り扱う．タグ候補をオントロジーから抽出することで，階層構造を持ったタグ候補を抽出することができる．また，階層構造を持ったタグ候補は，持たないタグ候補に比べて探索的閲覧がしやすくなる．

抽出元のオントロジーには日本語版DBpediaを用いる．日本語版DBpediaには「Category:社会問題」という社会問題に関するカテゴリをまとめたページが存在する．本研究では「Category:社会問題」の下位カテゴリを辿り，辿ったカテゴリの中のいずれかに属するページのページタイトルを取得した．具体的には，あるカテゴリの上位カテゴリを示している「skos:broader」という述語と，あるページがどのカテゴリに属しているかを示している「dcterms:subject」という述語を用いた．これを基にしたSPARQLクエリをDBpediaのエンドポイントに対して実行することによって社会問題のリストを取得した．用いたSPARQLクエリの一例を以下に示す．

```
select distinct ?pageo where {
 ?s1 skos:broader?
  <http://ja.dbpedia.org/resource/Category:社会問題> .
 ?s2 skos:broader? ?s1 .
 ?pages dcterms:subject ?s2 .
 ?pages rdfs:label ?pageo .
}
```

この方法で取得したページタイトルの中には，明らかに社会問題ではないものも含まれてしまっている．そこで，本研究では他のある特定のカテゴリの下位カテゴリに属しているページを除外する独自のフィルタを作成する手法，SVM (Support Vector Machine) を用いることで社会問題といえるタグを抽出する手法の二つの手法を検討した．

2.1 フィルタを用いた手法

フィルタリングに使用するカテゴリが一部異なる2種類のフィルタを作成した．フィルタAは取得したページタイトルの内，「Category:スタブカテゴリ」「Category:計算機科学」「Category:裁判」「Category:作品」「Category:社会運動団体」「Category:人物」「Category:生物学の分野」「Category:犯罪学」「Category:犯罪類型」「Category:平和学」「Category:論理学」というカテゴリから下位カテゴリを辿り，3階層以内のカテゴ

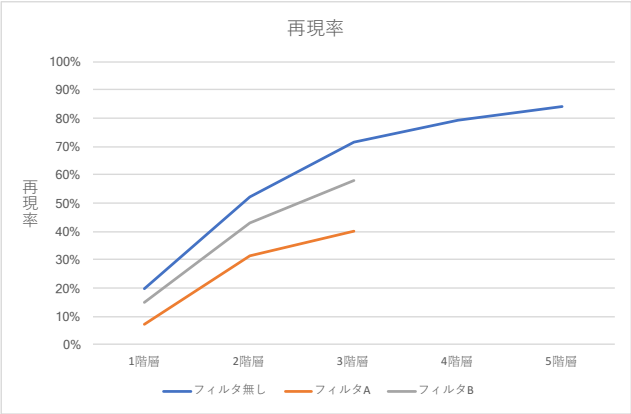

図 1: 再現率

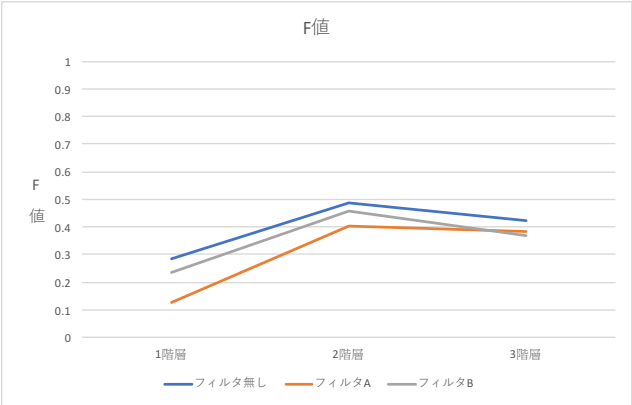

図 3: F 値

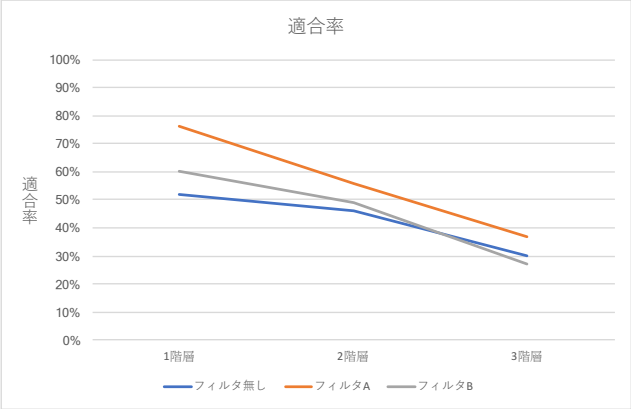

図 2: 適合率

リページの中のいずれかに属するページタイトルを除外する．フィルタ B はフィルタ A の除外条件から「Category:生物学の分野」のみを外した．

「Category:社会問題」から辿った階層数毎にフィルタを掛けなかったもの，フィルタ A を掛けたもの，フィルタ B を掛けたものを用意し，それらによって取得されたタグ候補のリストに対し，再現率，適合率の調査を行なった．再現率は予め用意した理想的な社会問題タグの候補 102 個が，上記の方法で取得したリストの中にどの程度含まれているかによって判定した．これを図 1 に示す．

また，これらのリストからランダムに 100 件のデータを取り出し，それが社会問題であるかを 5 段階評価で被験者に判定していただいた．本研究では 100 件のうち被験者が 3 以上と判定したものの比率を適合率とした．これを図 2 に示す．

さらに，再現率と適合率を元に計算した F 値を図 3 に示す．

評価の結果，同階層の物に関して「フィルタ B」の方が「フィルタ A」よりも高い再現率を示しており，「フィルタ無し」の低階層で理想的なタグがほとんど含まれていないことが分かった．また，フィルタを用いた手法よりフィルタを用いない手法の方が F 値が高くなってしまった．これはフィルタを用いたことで社会問題ではないタグ候補を除去する段階で，社会問題であるものまで多く除去してしまったことを意味する．フィルタの設計方法によって多少の改善の余地はあるが，除去されたものを逐一確認して精度を高めるのは現実的ではないと考えた．そこで，本研究ではフィルタを用いる以外の方法を検討することとし，次に SVM を用いた手法での実験を行なった．

2.2 SVM を用いた手法

SVM のベクトルを作成するため，「Category:社会問題」の 3 階層以内の下位カテゴリに属するページに関して，各ページから 5 階層以内に辿ることができるカテゴリページを取得した．カテゴリページの中で出現頻度が 9 以上となっているものを SVM の一つのベクトルとして用い，各ページから何階層以内にそのカテゴリページへたどり着けるかを各ベクトルの数値とした．ただし，5 階層以内にたどり着けなかった場合のベクトルの数値は 6 とした．

SVM の学習に用いるデータとして，フィルタを用いた手法の適合率を判定する際取得した 5 段階評価で 4，5 と判定されたものを正例，1，2 と判定されたものを負例として，ランダムに 120 件抽出した．

SVM のライブラリとしては scikit-learn[Pedregosa 11] の SVC クラスを用いた．パラメータは全て初期値で，事前に与えられた正例の中で正に分類されたものの割合を再現率，正と判定したものの中で事前に与えられた正例に分類されていたものの割合を適合率とした．評価には 10 分割のクロスバリデーションテストを用いた．

評価の結果，再現率は 79.2%，適合率は 68.8%，F 値は 0.74 となった．分母が揃っていないため正しく比較できているとは言い難いが，フィルタを用いた手法と比べて大きく改善された．

しかし，用いたベクトルの数が 10000 を超えており，次元の呪いによって精度が落ちている可能性も考えられる．最適なベクトルの設定方法に関しては今後の研究課題としたい．

3. 社会問題に関する文章に対するタグ付け

本研究ではある社会問題に関する記事に対するタグの自動付与を取り扱う．社会問題に関する記事と比較対象となる記事のベクトルを作成し，社会問題に関する記事のベクトルと各比較対象のベクトルとの Cos 類似度を計測し，類似度の大きさが閾値以上かつ上位 10 個以内となったものを付与タグとした．記事のベクトルは TF-IDF とパラグラフベクター [Le 14] を用いることによって作成した．また，付与するタグの候補は予め用意した理想的な社会問題タグ 102 個を用い，比較対象となる記事には候補タグと同名のタイトルを持つ Wikipedia のページを用いた．

本システムの性能を評価するために，用意したタグ候補 102 個の内最低 1 つが付与されるであろう社会問題に関する記事を 10 個用意した．この記事に対してそれぞれの手法でタグを付与し，付与されたタグ全てと付与されなかったタグ 3 個を被験者に見せ，そのタグ候補が記事に対するタグとしてふさわしいかを 7 段階で評価していただいた．

3.1 TF-IDF を用いた手法のタグ付け評価

TF-IDF を用いたタグ付けに対するアンケート評価の散布図を図 4 に，正解率を図 5 に示す．

アンケート評価値とシステム評価値の相関係数は 0.7320 を示し，強い相関を示した．システムの評価値で 0.3 以上を出したものは全てアンケート評価値で 4 以上の値を示している．このことから，システム評価値で 0.3 以上を出した値の殆どはタグとしてふさわしいと言える．また，アンケート評価値が 7 となったタグ候補のシステム評価値は 0.5890 から 0.1538 の範囲で大きくばらけている．これは TF-IDF アルゴリズムの性質上，単語の意味を考慮せずに単語の数を数え上げるものであるため，類義語や同義で異なる漢字を用いた単語を違う単語と判定しているためだと考えられる．事実，今回の実験で飢饉に関する社会問題の記事には「飢饉」という単語の代わりに「飢餓」という単語が多用されていた．そのため，「飢饉」という単語が頻出ではないと判定され，飢饉というタグ候補の類似度は低く判定されていた．

正解率は閾値を 0.2 に設定した時に最も高い正解率 0.8118 を示し，0.25 と 0.3 の時では 0.7882 となり，差が見られなかった．類似度 0.2 以上と判定されたタグは全タグ候補 85 個中 37 個存在し，1 つの記事に対して平均 3.7 個のタグが付与される結果になった．

3.2 パラグラフベクターを用いた手法のタグ付け評価

パラグラフベクターを用いたタグ付けに対するアンケート評価の散布図を図 6 に，正解率を図 7 に示す．

アンケート評価値とシステム評価値の相関係数は 0.3462 を示し，弱い相関を示した．

本研究では評価基準となる Wikipedia の記事を 102 個しか用いていなかった．Quoc Le ら [Le 14] の研究では評価を行う際，11855 文もの文書を用いていることから，文書数が不足していたために精度が出なかったのではないかと考えられる．また，パラグラフベクターのアルゴリズムは語順を考慮し，用いている単語が同じでも違う文体であれば違うものであると判定する．そのため，コーパスとして利用している Wikipedia のページと集めた社会問題記事の文体が違うので精度が出なかったのではないかと考えられる．

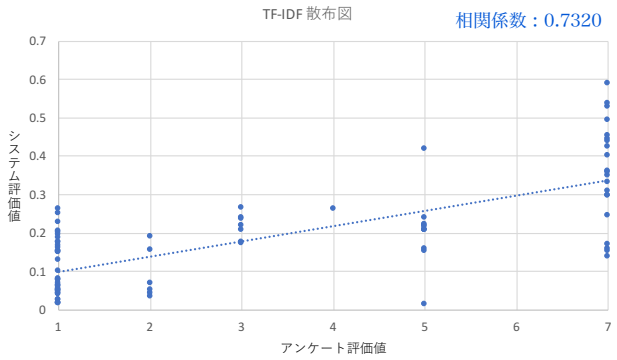

図 4: TF-IDF 散布図

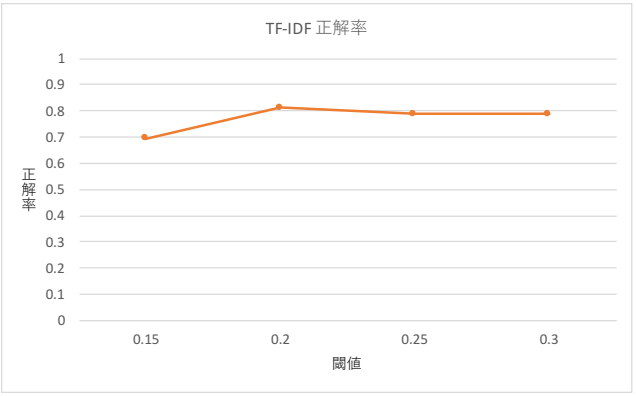

図 5: TF-IDF 正解率

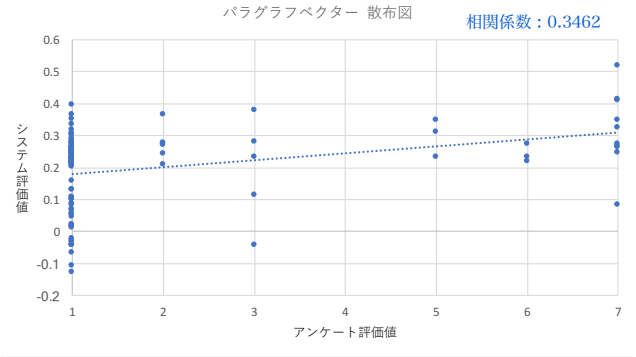

図 6: パラグラフベクター散布図

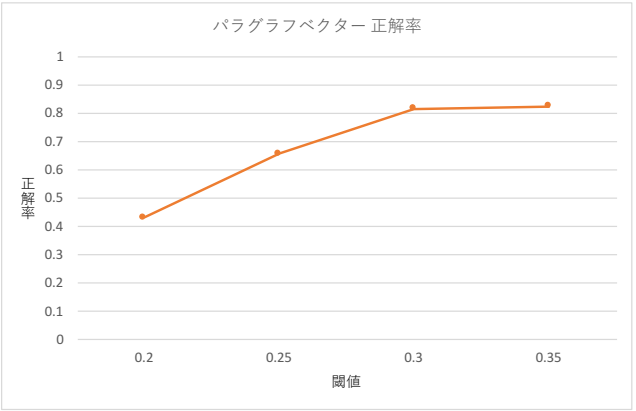

図 7: パラグラフベクター正解率

具体的には，本研究で「いじめ」に関する記事に対して「モンスターペアレント」や「虐待」がタグの候補として上位に上がってきた．この際，用いた「いじめ」に関する記事はいじめられないような子供を育てるため，どのように親が子供を育てるべきかという内容であった．Wikipedia の記事が「いじめとは何か」という説明をする文章なのに対し，使用した記事は「子供がいじめられないよう親ができることは何か」という文章であったため，親が主語となる「モンスターペアレント」や「虐待」が文体の比較で上位に上がってきたのではないかと考

える．これらの解決策としては，一つのタグに対して文体の異なる複数の記事を集め，それらと社会問題記事との類似度が最大となるものをシステム評価値として用いるというものが考えられる．

システム評価値で 0.2 以下と判定されたものでアンケート評価値が 4 以上のものは 30 個中 1 個しかなかった．また，唯一システム評価値が 0.2 以下でアンケート評価値が 7 であった「ホームレス」という記事に対して行われた「経済的不平等」の評価も，記事自体に「経済的不平等」に関する記述があるわけではなく，むしろタイトルの「ホームレス」との関連語として評価されたものだと判断できる．従って，文体から明らかにタグ付与には不適切といったタグ候補の足切りのように用いるのならば，十分実用性はあると考える．

正解率は 0.3 になるまで上がり続け，0.3 では 0.8137，0.35 では 0.8235 とあまり差が見られなかった．ただし，全タグ候補 102 個のうち，類似度 0.3 以上では 17 個，類似度 0.35 以上では 8 個しか該当するタグがない．このような閾値に設定すると，1 つの記事に対して付与される平均的なタグ数は閾値 0.3 で 1.7 個，閾値 0.35 で 0.8 個になってしまう．8 割以上の正解率になるこれらの閾値は実用上適さない恐れがある．しかし，閾値を 0.25 にしてしまうと正解率は 0.6569 となってしまい，これは十分な値とは言えない．

4. Web API の試作

前章の結果を元に，本研究では TF-IDF を用いて，閾値を 0.2 に設定したタグ付けシステムを Web API として実装する．本 API は REST (REpresentational State Transfer) の構造をしており，GET メソッドで送信されたタグを付与する対象となる文書に対し，システム側で用意されたタグに関する文書との類似度を計算し，閾値以上となった上位 10 個以内のタグ名とその類似度を JSON 形式で返す．本 API が出力する JSON を以下に示す．

```
[
    {
        "name": "いじめ",
        "similarity": 0.8737471577348052
    },
    {
        "name": "モンスターペアレント",
        "similarity": 0.43941567811546905
    },
    {
        "name": "格差",
        "similarity": 0.42751699621958034
    }
]
```

本研究では単純な形式の JSON を出力するにとどまったが，将来的には白松ら [Shiramatsu 12] が提案した AnnotationInfo クラスや，Wikidata[Denny 14] で用いられている Qualifiers を参考にし，W3C[W3C 17] が 2017 年 2 月に勧告した Web Annotation Data Model に則ってより汎用性が高い形式での提供を行うことを目指したい．

5. おわりに

本研究では (1)DBpedia を用いた社会問題のタグ候補の抽出と，(2) 社会問題に関する記事に対して最適なタグを社会問題のタグ候補から選択するという二つの研究を行った．(1) では SVM を用いたことである程度の精度が得られ，(2) では TF-IDF を用いた際，高い相関性を得ることが出来たが，一般的に IT-IDF より精度が高い手法であるパラグラフベクターを用いた際，TF-IDF より劣る結果となった．今後の課題としては，タグ候補の抽出における SVM で用いるベクトルの再検討，SVM 以外の方法によるタグ候補の抽出方法の検討，パラグラフベクターの精度向上，TF-IDF やパラグラフベクター以外の文書間類似度測定手法の活用，汎用性が高い形での API の公開などが挙げられる．

なお，本研究の成果として得られる社会問題タグの階層は，社会問題に関連する背景情報の探索的な閲覧に利用することも可能である．例えば，地方議会の議事録に含まれる発言に社会問題タグを付与しておけば，社会問題の階層構造を起点として体系的に探索的閲覧ができるユーザインタフェースを提供できる．

そのような事例として成瀬ら [成瀬 17] は，名古屋市の Web ページにある分野体系をタグとして用い，名古屋市会議事録の探索的閲覧が可能なユーザインタフェースを開発中である．今後は，本研究で得られた社会問題の体系を議事録の探索的閲覧に利用できるよう，システムの統合を検討する予定である．

謝辞

本研究は，JSPS 科研費 (25870321)，JICE 研究開発助成，および JST CREST の支援を受けた．

参考文献

[Denny 14] Denny Vrandeč ić and Markus Kr ö tzsch: “Wikidata: a free collaborative knowledgebase.” In *Magazine Communications of the ACM CACM Homepage archive*, Vol. 57, Issue 10, pp. 78–85, 2014.

[Le 14] Le Quoc V, and Tomas Mikolov: “Distributed Representations of Sentences and Documents.” In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[Pedregosa 11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E.: “Scikit-learn: Machine Learning in Python” *Journal of Machine Learning Research*. Vol. 12. 2011.

[Shiramatsu 12] Shun Shiramatsu, Robin M. E. Swezey, Hiroyuki Sano, Norifumi Hirata, Tadachika Ozono and Toramatsu Shintani: “Structuring Japanese Regional Information Gathered from the Web as Linked Open Data for Use in Concern Assessment.” In *Electronic Participation – 4th IFIP WG 8.5 International Conference, ePart 2012, Proceedings*, Springer LNCS, Vol. 7444, pp. 73–84, 2012.

[W3C 17] W3C : “Web Annotation Data Model.” `https://www.w3.org/TR/annotation-model/`, 2017.

[溝口 99] 溝口理一郎: “オントロジー研究の基礎と応用” 人工知能学会誌, Vol.14, No.6, pp.977–988, 1999.

[成瀬 17] 成瀬雅人，白松俊，松島格也: 探索的閲覧のための地方議会議事録の構造化手法の検討. 情報処理学会第 79 回全国大会 講演論文集, 6ZB-08, 2017.