4M1-1

Web 上の情報を用いた政治資金収支報告書の分析

Analysis of Reports on Political Funds using Information on the Web

渡辺 哲朗 *1　　松尾 豊 *1
Tetsuroh Watanabe　Yutaka Matsuo

*1東京大学大学院 工学系研究科 技術経営戦略学専攻
Department of Technology Management for Innovation, School of Engineering, the University of Tokyo

Inappropriate incomes and expenditures of political funds by politicians are found by searching the Web. However, there are too much records of political funds to inspect all of them by humans. In this paper, we analyzed recent Japanese reports on political funds, and tried to not only collect information related records of political funds from the Web, but also determine whether these incomes and expenditures seem to be inappropriate or not automatically.

1. はじめに

政治家による政治資金に係る不正の問題は，俗に「政治とカネ」の問題とも呼ばれ，政治における長年の課題とされてきた．その状況は今日も続いており，重責を担う政治家による政治資金の不適切な支出の発覚などの報道が後を絶たない．政治資金の収支については，政治資金収支報告書の作成と提出が政治資金規正法 [1] により義務付けられている．各政党内部や報道機関においては，主に Web 上の情報を検索する方法によって，政治資金収支報告書に記載のある収支の相手（個人・店舗・法人）に関する情報収集が人力によって実施され，政治資金の不適切な収支の有無が調査点検されているという．

しかし，政治資金収支報告書の総記載量は膨大であり，人手で全ての収支の関連情報をくまなく調査することは困難である．こうした調査が不十分であることは，大物政治家の過去の政治資金の不正が長期間経過後に芋づる式に発覚する，という事例が多発しているという実情からも窺い知ることができる．社会通念上不適格な政治家の存在を防止する観点からも，政治資金の不正有無を効果的に検知できることは重要である．

2. 外部情報としての Web 上の情報の活用

政治資金収支報告書の点検において手動で行われているような，Web 上の情報を外部情報として活用することで手元の情報を補完する手法は，その効率向上・自動化に向けて様々な研究が進められている．ニュース記事の理解に必要な背景知識を Web から取得して自動的に補足説明を実現する研究 [2]，与えたキーワードをより詳細に説明する語を Web 上から取得する研究 [3]，特定の事象に関連するニュース記事を Web 上から収集し一つの記事に要約する研究 [4] などが行われている．こうした研究により，Web 上のリソースによって手元の限られた情報を充実させることが可能となるが，他方，得られた情報の解釈や，それを用いた判断についてはあくまで人間に委ねられる．

政治資金収支報告書における不適切な収支記録の有無の判定という観点に即して言えば，記録されている収支相手の名称と住所を基に，関連する情報を Web 上から自動的に取得できたとしても，そうして充実した情報をチェックして不適切か否かを判断するのはあくまで人手となる．政治資金収支報告書は，国会議員の毎年一回の定期公表だけでも数千団体分の提出があり，一団体あたりの分量は数十ページに亘る．更にそこに追加団体および解散団体の分が適時加わるという，膨大な分量の記録である．Web 上の情報を外部情報として効率よく取得することに加えて，取得した情報を効果的に自動判定することが可能であれば，Web からの情報取得とその活用である判定とが有機的に結合することによる相乗効果が期待できる．

連絡先: 渡辺 哲朗，東京大学大学院 工学系研究科 技術経営戦略学専攻，東京都文京区本郷 7-3-1 東京大学工学部 2 号館 92C1 号室，03–5841–7718，watanabe@weblab.t.u-tokyo.ac.jp

3. 研究の目的と方針

本研究では，Web 上の情報を用いて，政治資金として不適切な収支と疑われる記録を政治資金収支報告書から自動検出するための手法を提案する．特に本研究においては，既存の諸研究には見られない新規性ある試みとして，Web 上から取得した情報を用いての不適切性の判定をも自動で行うべく，効果的な判定モデルの構築を目指す．

ここで問題となるのは，一般的な判定モデルに見られるような十分な「教師データ」の準備が事実上不可能な点である．報道によってその不適切性が明らかになる収支データは確かに存在するが，収支データ全体と比べると，その数は僅少である．また，不適切なものとして一度でも報道された収支については，Web 上に存在する情報の大半が当該報道のものになってしまい，報道に依らないニュートラルな情報源が埋没してしまうため，事後的に教師データとなり得る情報を得ることが困難となってしまう．こうした課題に対して，本研究のアイデアは，政治資金に関連する過去の様々な報道記事や文献の文書資源を活用することによって，直接的な教師データを用いることなく政治資金収支の不適切性の有無を判定できるモデルを構築しようと試みるものである．

本研究の概観を，図 1 に示す．

4. 利用データと分析手法

本研究では，分析対象として，各国会議員の政治資金管理団体によって総務省に提出された，過去 4 年分（2011 年〜2014 年 公表分）の政治資金収支報告書のデータを用いる．これらは，政治資金規正法に基いて総務省 Web サイトに公開 [5] されており，自由に閲覧が可能である．本研究では，一般的な Web スクレイピングの方法によりこれらを全て取得した．取

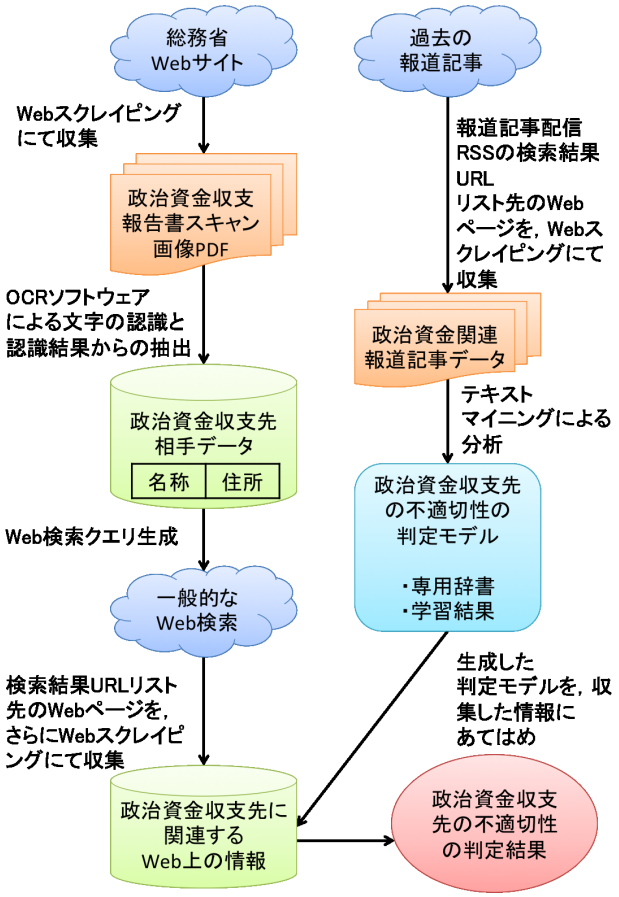

図 1: 本研究の概観

得した PDF ファイル総数は 14,662 個であり，そのページ総数は 140,287 ページである．

こうして取得できるデータは全て PDF 形式のファイルであり，その中身は全て政治資金収支報告書の現物のスキャン画像である．そのままでは分析可能な文字情報としては利用できないため，本研究では OCR ソフトウェア [6] を用いて，これらのスキャン画像を可能な限り日本語文字データ化した．

文字データ化された収支データのうち，OCR ソフトウェアによる文字認識に破綻がなく，正常な日本語としてデータ化されているもののみを抽出した．そこから，今回は関連する過去の新聞記事が豊富に存在する「支出」のデータに限定し，さらに支出先相手の「名称」と「住所」の双方が重複するデータを可能な限り除くことにより，総計 107,891 件のユニークな支出先相手データを得た．これらの各支出データに記載されている支出先相手の「名称」と「住所」を用いて，

「名称　AND　住所」

という検索クエリを生成し，検索エンジンに送信することで，検索結果上位（最大 50 件）の URL リストを得た．そして，これらの URL が示す Web ページの情報を，さらに Web スクレイピングの手法により取得した．検索クエリに基づいた Web 検索の実現には Microsoft Bing Search API [7] を用いた．なお，当該 API の利用量制限の都合上，現時点までに検索結果とその中の URL が示す Web ページの情報を取得した収支データは，全収支データからランダムに選択した 1,989 件であり，これらの検索結果により情報が得られた Web ページの総数は 82,878 ページである．

本研究においては，分析対象である収支データそのものを説明する情報ではなく，政治資金に関連する過去の報道記事などの周辺データに基いて，分析対象データの不適切性についての判定を試みる．そこで本研究では，次のような手法の適用を計画している．

- 過去の報道記事の文書を用いて，不適切性の判定に利用する専用の単語辞書を生成し，その辞書を用いて各収支データを説明する Web 上の情報を分析する方法
- 過去の報道記事の文書分析における学習から，各収支データを説明する Web 上の情報の分析のための規則を獲得する，転移学習 (transfer learning) [8] [9] による方法

このうち，本稿においては，過去の報道記事の文書を用いての，不適切性の判定に利用する専用の単語辞書生成に関する実験について述べる．

5. 実験

5.1 概要

本実験，政治資金の支出相手としての不適切性を，専用の単語辞書を生成し利用することによって判定する際に，その辞書生成の基礎となりうる，政治資金に関する過去の報道記事の文書データ群を取得する．それらを分析することにより，

- 政治資金関連の過去の報道記事に登場する単語のリストから，政治カテゴリ全体の過去の報道記事の要素を除外することにより，政治資金関連の報道記事に特化した単語リストの効果的な抽出が可能であろう

という，専用の単語辞書の生成にあたっての本研究における我々の仮説を検証する．

本実験の概観を図 2 に示す．本実験では，Google ニュースの RSS フィード [10] 配信システムを用いて関連する直近の報道記事の掲載 URL リストを取得し，それらの掲載 URL が示す Web ページをさらに取得することで，直近の報道記事の文書データ群を獲得した．政治資金に関連する直近の報道記事を Google ニュースから取得する際に，RSS フィード配信システムに今回指定した検索単語は表 1 の通りである．それぞれの

表 1: 過去の関連報道記事の収集に用いた検索語

政治資金
政治資金規正法
政治資金管理団体
政治資金収支報告書
政務活動費
政務調査費
政治とカネ
政治と金

検索単語につき最新 100 件ずつの報道記事掲載 URL リストを取得し，これらを結合し重複を除いた URL リストのそれぞれが指し示す Web ページのデータを取得した．こうして，政治資金に関する過去の報道記事，計 547 件の Web ページのデータを獲得した．また，政治カテゴリ全体の直近の報道記事については，Google ニュースにの「政治」カテゴリに掲載された

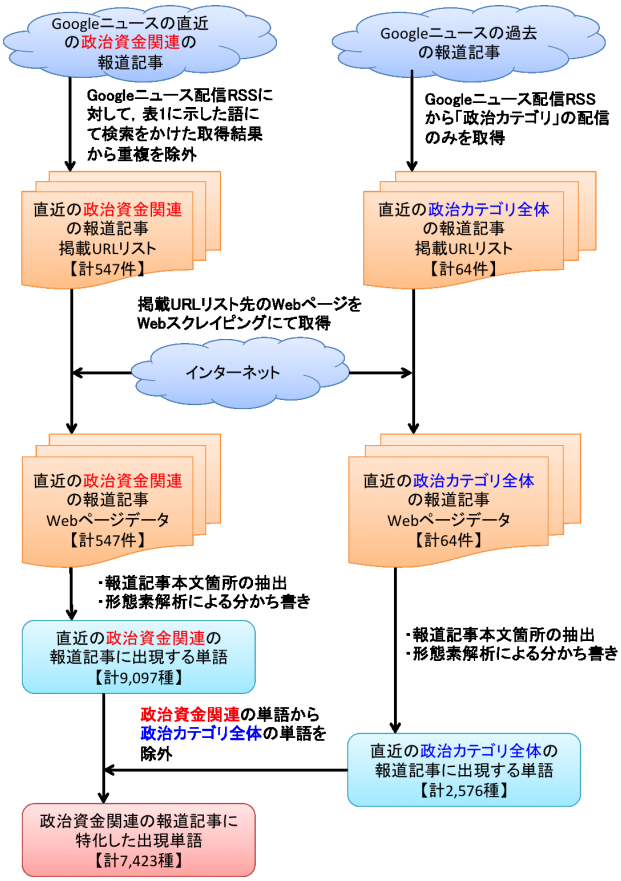

図 2: 本実験の概観

最新 64 件の報道記事掲載 URL リストを同様に取得し，それらの Web ページのデータを同様に取得した．

こうして得られた直近の報道記事の Web ページのデータに対して，形態素解析による分かち書きを行い，以下の条件にあてはまる単語のみを抽出した．

- 名詞であること
- 名詞のうち，非自立語，接尾語，数，代名詞，人名，組織名のいずれにも該当しないこと

これらの報道記事に出現する単語の出現回数の総計を単語別にカウントし，順位付けを実施した．なお，Web ページのデータ中には報道記事本文と関係のない記述も多分に含まれるため，Web ページの本文抽出を実行するオープンソースライブラリ ExtractContent [11] を用いることにより，記事本文のみを抽出し，これを分析対象とする．形態素解析にはオープンソース形態素解析エンジン MeCab [12] を用いた．また，形態素解析を行う際に用いる単語辞書として，Web 上の言語資源に基いて新語を随時追加するシステム辞書である mecab-ipadic-NEologd [13] の最新版を適用した．

政治資金関連の報道記事に出現した単語から，政治カテゴリ全体の報道記事に出現する単語を除外することで，政治資金関連の報道記事に特化した単語のリストを生成する．なお，政治資金関連の報道記事は，概念的には政治カテゴリ全体の報道記事に包含されるものではあるが，今回は政治カテゴリ全体の報道記事の取得件数を相対的に少なくすることによって，一般的な政治カテゴリ要素の色のより強い単語のみの除外を実現し，以って政治資金要素の色のより強い単語のみを残すことを試みる．

5.2 結果

こうして得られた，政治資金関連の報道記事に特化した単語の頻出上位 20 件を，その順位および出現回数とともに表 2 に示す．なお，直近の政治資金関連の報道記事に出現した単語は計 9,097 種，直近の政治カテゴリ全体の報道記事に出現する単語は計 2,576 種であり，表 2 に示すもの含めた最終的な集計に出現した単語は計 7,423 種類であった．

表 2: 政治資金関連の報道記事に特化した出現単語の頻出上位

順位	出現回数	単語
1 位	698 回	議員
2 位	630 回	支出
3 位	414 回	政務活動費
4 位	352 回	政治資金
5 位	301 回	収支
6 位	229 回	政治とカネ
7 位	201 回	会派
7 位	201 回	返還
9 位	198 回	政治家
10 位	158 回	政治資金収支報告書
11 位	157 回	収入
12 位	154 回	購入
13 位	138 回	政治活動
14 位	128 回	報告書
15 位	121 回	使途
16 位	113 回	資金管理団体
17 位	112 回	違法
18 位	92 回	政
19 位	90 回	後援会
20 位	89 回	報道
21 位	85 回	政党支部
22 位	82 回	パーティー
23 位	78 回	金額
24 位	75 回	作成
25 位	72 回	会費
25 位	72 回	公職

5.3 考察

前述の本実験結果によれば，政治資金にまつわる事象を代表するような単語群が確かに表 2 に示す頻出上位に登場していることが見て取れる．第 7 位の「会派」のような，治資金に直接は関連しない一般的な政治カテゴリの事象に関する単語も一部は見て取れるものの，上位は政治資金の報道の色が強い単語が大半を占めた．この結果から，政治資金関連の報道記事に特化した単語リストの効果的な抽出に関する本実験の仮説は支持された．

他方，本実験結果の全体を目視で確認してゆくと，支出先相手の政治資金的な不適切性判定のための辞書生成にあたり，重要な単語として効率よく抽出されることが望ましいと思われる単語が，下位方向の領域にも遍在していることも，本実験の結果によって明らかになった．こうした単語には，一例として「キャバクラ（1,060 位/出現回数 11 回）」などが挙げられ

る．本実験の結果から，政治資金に特化した報道記事資源の抽出は有効に機能することが明らかになったため，引き続き辞書生成のアプローチを取るにあたっては，支出先の不適切性を判定するという本研究の趣旨により合致するような，最適な分析方法を更に整備することが有効であると言える．なお，こうした文書資源を用いる単語辞書の自動生成に関連する研究には，EC サイトの商品レビューコメントから商品評価の表現に特化した辞書を自動生成する研究 [14] などがある．

6. おわりに

本稿では，政治資金収支報告書を分析することによって，政治資金として不適切になるような収支の自動判定を，Web 上の情報を活用することによって実現することを目指すという本研究の目的とその方針について述べ，それを実現するための利用データの概要と実際に行ったデータの準備，およびデータの分析手法について述べた．また，実際に Web から取得した情報を分析する実験を行うことによって，政治資金関連の報道記事に特化した出現単語の効果的な抽出を実現した．

本稿では，原則として政治資金というテーマを題材とした研究方針や実験に関する研究の内容について述べたが，Web 上の膨大な情報を用いて何かしらの自動判定モデルを構築するという本研究の意図は，政治資金にとどまるものではないと考える．我々は，まずは政治資金という社会的意義の大きい領域を対象とし，それを足がかりとして，より汎用的・抽象的な研究の確立を目指してゆくものである．

参考文献

[1] 政治資金規正法，法令データ提供システムにて閲覧可，http://law.e-gov.go.jp/htmldata/S23/S23HO194.html (1948 制定 / 2014 年最終改正)

[2] 田中祥太郎, ヤトフトアダム，田中克己：ニュース記事の理解支援のための背景知識抽出と補完，電子情報通信学会技術研究報告，Vol. 114，No. 173，pp. 95–100 (2014)

[3] S.Oyama, K.Tanaka : Query modification by discovering topics from Web page structures, Proc. 6th Asia Pacific Web Conference (2004)

[4] 金田晃征，野村浩郷：情報検索と情報集約による情報取得システム，情報処理学会研究報告，Vol. 2007，No. 47，pp. 31-36 (2007)

[5] 総務省，政治資金収支報告書及び政党交付金使途等報告書，http://www.soumu.go.jp/senkyo/seiji_s/seijishikin/

[6] ABBYY FineReader 12，http://finereader.add-soft.jp/

[7] Microsoft Bing Search API，https://datamarket.azure.com/dataset/bing/search

[8] NIPS 2005 Workshop – Inductive Transfer: 10 Years Later, http://iitrl.acadiau.ca/itws05/ (2005)

[9] 神嶌敏弘：転移学習，人工知能学会誌，Vol. 25，No. 4，pp. 572–580 (2010)

[10] Google ニュース RSS フィード，https://support.google.com/news/answer/59255?hl=ja

[11] 本文抽出モジュール ExtractContent，Cybozu Labs，http://labs.cybozu.co.jp/blog/nakatani/2007/09/web_1.html

[12] mecab – Japanese morphological analyzer，https://code.google.com/p/mecab/

[13] mecab-ipadic-NEologd : Neologism dictionary for MeCab，https://github.com/neologd/mecab-ipadic-neologd

[14] 谷本融紀，太田学：特定評価属性の関連属性自動抽出による評価表現辞書の生成，情報処理学会研究報告，Vol. 2012-DBS-155，No. 12，pp. 1–9 (2012)