The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

サイト内の話題分布素性を用いた分類器学習による ノウハウサイト同定

Identifying Know-How Sites
by Classifier Learning with Features of Topic Distribution within a Site

大川 遥平 *1*2 Yohei Ohkawa　前田 竜治 *1 Tatsuya Maeda　陳 騰揚 *1 Tengyang Chen　宇津呂 武仁 *3 Takehito Utsuro　河田 容英 *4 Yasuhide Kawada

*1筑波大学大学院システム情報工学研究科 Grad. Sch. Sys. & Inf. Eng, Univ. of Tsukuba
*2(株)AVILEN AVILEN Inc.
*3筑波大学システム情報系 Fclty. Eng, Inf. & Sys, Univ. of Tsukuba
*4(株) ログワークス Logworks Co., Ltd.

This paper proposes techniques of automatically discovering tips Web sites from a large collection of Web pages using a topic model and support vector machine (SVM). Tips refer to practical knowledge or expertise that is used to help accomplish certain tasks in a particular field. We designed several approaches of extracting features with respect to domain names based on their distribution among Web pages and candidate tips Web sites. In addition, search engine suggests, the query keywords used to fetch Web pages from the search engine are also considered to present patterns that can be potential features.It was discovered from our dataset that domain names of tips Web sites (Web sites containing tips on a certain specific theme) are more likely dispersed among topics and Web pages. These domain names also tend to correspond to a larger number of search engine suggests. This paper verifies such observed patterns by training an SVM using those extracted features. Evaluation is performed in precision and recall to measure correctness of classifying whether or not a domain name belongs to a tips Web site.

1. はじめに

インターネットの発展により，日々，膨大な量のデータがウェブ上に増え続けている．自身の行動の助けとなるようなノウハウ・知識などの有用な情報がどこに掲載されているかを特定することが難しくなってきている．その背景には，近年，自社サイトのページ閲覧を目的としたウェブサイト，SEO (Search Engine Optimization) 対策を施し，ウェブ検索結果で注目されることを目的としたウェブサイトが増加傾向にあることが挙げられる．そのため，検索エンジンが果たすべき重要な責務として，有益な情報を掲載するページと，有益な情報は掲載しないが，SEO 対策の結果ウェブ検索上位に順位付けされるページをいかにして識別するか，という課題が挙げられる．例えば，(a) 花粉症，(b) 結婚，(c) 就活，(d) マンション，(e) 虫歯，(f) 食虫毒，の各クエリを Google 検索エンジンで検索した際の上位 50 件のウェブページのうち，ノウハウ知識を掲載するページの割合を図 1 に示す．どの検索クエリの検索結果においても，検索上位 50 件においては，ノウハウ知識掲載ページの割合は 50%かそれ以下となっている．これより，検索エンジンのユーザーは有益な情報が掲載されているウェブページを判別するのが困難なことが予想される．

この結果をふまえて，本論文では，ノウハウ知識を含むウェブサイトをドメイン単位で自動的に判定することを目的とし，その分類方法を提案する．提案する分類方法は，ウェブページの収集，トピックモデルを用いたウェブページのクラスタリング，各ドメインの特徴抽出，SVM のトレーニングからなり，それらの手法について記述する．また，評価実験の結果において，一定以上の再現率・適合率が達成できたことを示す．

連絡先: 大川 遥平，筑波大学大学院システム情報工学研究科，〒 305-8573 茨城県つくば市天王台 1-1-1，029-853-5427

2. 検索エンジン・サジェストを用いたウェブページの収集

本論文では，分析の対象であるクエリ・フォーカス「就活」,「花粉症」,「結婚」,「マンション」,「虫歯」，および,「食中毒」について，検索エンジン・サジェストを利用してウェブページを収集する．そのためにまず，Google 検索エンジンを用いて，一つのクエリ・フォーカスにおいて，約 100 通りの文字列を指定し，最大約 1,000 語の検索エンジン・サジェストを収集する．クエリ・フォーカス「花粉症」,「結婚」,「就活」,「マンション」,「虫歯」,「食中毒」において，収集されたサジェストの数を表 1 に示す．ここで，クエリ・フォーカスに対して収集されたサジェストの集合を $\mathbb{S}$ として，あるサジェスト $s \in \mathbb{S}$ に対して，クエリ・フォーカスとの AND 検索の検索結果上位 N 件以内に検索されるウェブページの集合を $S(q)$ とする．本論文では，$N = 20$ とした．このとき，各クエリ・フォーカスに対して収集されるウェブページの集合 $S(q)$ を次式で定義する．

$$S(q) \quad = \quad \left\{ s \in S \middle| q \in Q(s, N) \right\}$$

また，収集可能なサジェストとクエリフォーカスとの AND 検索により収集可能なウェブページを Q とする．

$$Q \quad = \quad \bigcup_{s \in S} Q(s, N)$$

3. ノウハウサイトの候補集合

3.1 トピックモデル

本論文では LDA(Latent Dirichlet Allocation) [Blei03] を用いる．LDA はトピックモデルのもっともメジャーなモデルのひとつである．トピック数 K を与えると，各トピック z_n における語の生成確率 $P(w \mid z_n)$ $(w \in V)$，および各トピック P_z

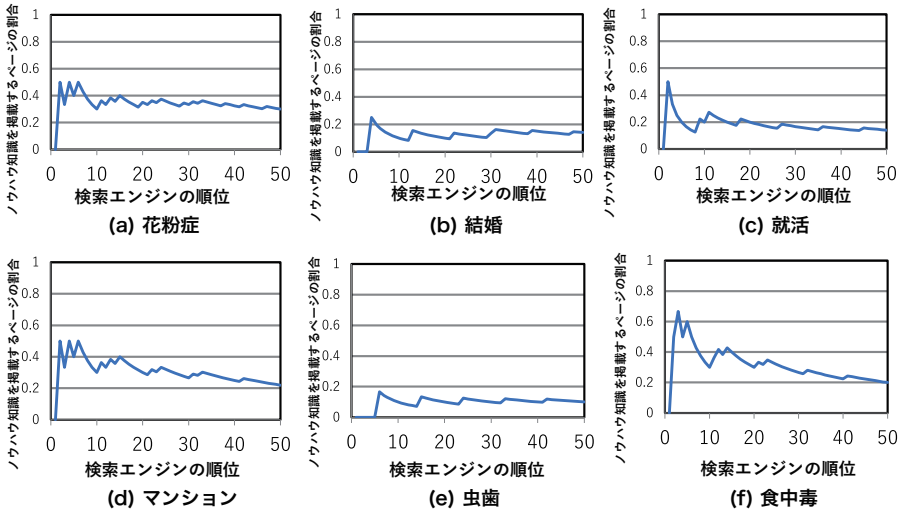

図 1: (a) 花粉症，(b) 結婚，(c) 就活，(d) マンション，(e) 虫歯，(f) 食中毒，の各クエリを Google 検索エンジンで検索した際の上位 50 件のウェブページのうち，ノウハウ知識を掲載するページの割合．

表 1: 各クエリのデータ概要

クエリ フォーカス	サジェスト 数	ウェブ ページ数	ノウハウサイト 候補の数 (正例/負例)
花粉症	849	9,738	101 (45/56)
結婚	959	13,256	85 (40/45)
就活	926	12,073	82 (51/31)
マンション	958	13,734	89 (57/32)
虫歯	835	9,554	92 (45/47)
食中毒	806	7,668	93 (28/65)
合計	5,333	66,023	542 (266/276)

表 2: ノウハウサイト候補のドメインに対する評価基準

<table>
<tr><td colspan="3">ドメインそのものがノウハウ知識を提示する個別ページへのリンクを一覧するページである</td><td>A 群</td></tr>
<tr><td rowspan="4">ドメインそのものがノウハウ知識を提示する個別ページへのリンクを一覧するページではない</td><td rowspan="2">ノウハウ知識を提示する個別ページへのリンクを一覧するページが存在する</td><td>ドメインのトップからノウハウ知識を提示する個別ページへのリンクを一覧するページに容易に辿り着ける</td><td>B 群</td></tr>
<tr><td>ドメインのトップからノウハウ知識を提示する個別ページへのリンクを一覧するページに容易には辿り着けない</td><td>C 群</td></tr>
<tr><td rowspan="2">ノウハウ知識を提示する個別ページへのリンクを一覧するページが存在しない</td><td>ノウハウ知識を提示する個別ページが存在する</td><td>D 群</td></tr>
<tr><td>ノウハウ知識を提示する個別ページが存在しない</td><td>E 群</td></tr>
</table>

における各文書 (本論文ではウェブページ q) におけるトピック Z_n の確率分布 $P(z_n|d)(n = 1, \ldots, K)$ を推定する．本論文は，全部の 6 つのクエリのウェブページ集合に対し，トピック数 K を 50 に設定し，実験を行った．

3.2 ウェブページ集合のトピックへの分類

各ウェブページ q におけるトピックの確率が最大となるトピックをそのウェブページ q のトピックとした．したがって，全てのウェブページにトピックを付与することができる．トピック z_n が付与されたウェブページの集合 Q は次式で定義される．

$$Q(z_n) = \left\{ q \in Q \,\middle|\, z_n = \underset{z_u\ (u=1,\ldots,\ K)}{\operatorname{argmax}} P(z_u|q) \right\}$$

3.3 ノウハウサイトの候補集合

各トピックにおける確率が低いウェブページは，そのトピックに関して詳しく記述されていない，一つのトピックにしか出現しないドメインのウェブサイトがノウハウサイトである可能性は低いという仮説をもとに，ノウハウサイトの候補集合とするサイトの基準を設けた．これらの仮説をもとに，トピックモデルを適用した結果において，各トピック上位 30 位のウェブページのみに出現するウェブページを対象とし，2 個以上のトピックにまたがって出現するドメインをノウハウサイトの候補集合とした．

4. データセット

ドメインごとにウェブサイトをアノテーションする際に，5 つのタイプのウェブサイトを観測した．その 5 つのタイプを A，B，C，D，E 群とし，その基準を表 2 に記述した．表 2 の基準に従い，A 群，B 群，C 群のサイトをノウハウサイトとし正例，D 群と E 群を負例としてデータセットを作成した．合計 542 のノウハウサイトの候補集合に対し，266 が正例で 276 が負例となった．6 つのクエリからそれぞれ 20 サイトを無作為に選択し，データセット作成者とは別に表 2 の基準に基づいてアノテーションを行った．その結果，正例/負例の分類における kappa 係数は 0.96 であった．

5. 素性

本節では，各ドメインに対して用いた素性について述べる．

5.1 ページ内の話題分布に関する素性

ウェブサイト内のウェブページの話題分布を測定する素性を利用する動機は，ノウハウサイトの候補のアノテーション作業を通じて，ノウハウサイトはノウハウサイトではないサイトと比べて，話題が一定の広がりを見せる傾向を発見したことに由来する．ウェブサイト内の話題分布を測定する手法については，本論文では次の 2 種類の素性を用いた．一つは LDA トピックモデルに基づくもの，もう一つは doc2vec モデルに基づくものである．これらの二つのアプローチはまったく異なる背景を持つ．LDA トピックモデルは大きな文書集合内のトピックをモデル化することを目的としたものだが，doc2vec は文書のベクトル表現を得ることを目的としている．本論文では，これらの二種類の素性の性能比較，および，それらを併用した場合の性能評価を行う．

5.1.1 LDA トピック数素性

トピックモデルを適用した結果においてウェブページ $q(\in P(t))$ に対して，トピックの確率分布が最大になるトピック $z(q)$ をウェブページ q に割り当てる．ここでドメイン t における各トピック $z(q)$ の集合を次式で定義する．

$$Z(t) \quad = \quad \bigcup_{q \in Q(t)} \{z(q)\}$$

そして，サイト t におけるトピックの異なり観測数を $f_z(t)$ で定義する．

$$f_z(t) \quad = \quad \big|Z(t)\big|$$

5.1.2 ウェブサイト内の doc2vec の距離の平均

doc2vec[*1] モデル [Le14] は word2vec モデルと同様の訓練方法を用いて文書をベクトル化するモデルである．単語のベクトルの学習に加えて文書 (本論文ではウェブページ q) のベクトルも同時に学習する．doc2vec モデルを訓練するうえで，クエリフォーカスに関して収集されたすべてのウェブページをトレーニングデータとして使用した．トレーニングに使用された各クエリフォーカス内のページ数は表 1 のウェブページ数にあたる．例えば，doc2vec が結婚というクエリに属する文書のベクトル $v(q)$ を学習するとき，13,256 の入力文書を利用した．本論文で使用した doc2vec の訓練パラメータは，過去の doc2vec に関する実験で利用されたもので，出力するベクトルの次元数は 300，ウィンドウサイズは 5 を用いた．低頻度語の影響を考慮せず，トレーニングテキスト内のすべての単語を訓練に利用した．doc2vec によって文書ベクトルを訓練した後，ドメイン t の $Q(t)$ 内の任意の 2 つのウェブページ q と q' のユークリッド距離 $\mathrm{Edist}(q, q')$ の平均を次式で求め，これをドメイン t の doc2vec 距離平均素性とする．

$$f_{doc2vec}(t) \quad = \quad \underset{q,\ q' \in Q(t)}{\mathrm{average}} \Big(\mathrm{Edist}(v(q),\, v(q'))\Big)$$

5.2 その他の素性

ドメイン内のページの話題分布を定量化する素性とは別に，以下 3 つの素性を定義した．

*1 https://radimrehurek.com/gensim/models/doc2vec.html

5.2.1 ウェブサイト内のページ数

ノウハウサイトにおいては，収集されたウェブページの数が多い傾向にある．このことをふまえて，ノウハウサイトの候補 t のウェブページ集合を $Q(t)$ として，ウェブサイト内のページ数素性を次式で定義する．

$$f_q(t) \quad = \quad \big|Q(t)\big|$$

5.2.2 検索ボリューム

検索エンジンサジェストとは，あるクエリが 1ヶ月間に検索エンジンにおける検索数を表す．検索数は Google AdSense[*2] を用いた．ノウハウサイトは，比較的検索量の多い特定のクエリキーワードを介してアクセスされることが予想される．まず，ドメイン t に割り当てられたサジェスト集合 $S(t)$ について考える．ドメイン t に対応するウェブページ q が検索されたサジェストの和集合として下式で表せる．

$$S(t) \quad = \quad \bigcup_{q \in Q(t)} S(q)$$

検索エンジンサジェスト s の，検索ボリュームを $sv(s)$ とすると，ドメイン t に対応する検索ボリュームの最大値 $fv\text{-}max(t)$ と平均値 $fv\text{-}ave(t)$ は以下のように定義できる．

$$fv\text{-}max(t) \quad = \quad \max_{s \in S(t)} sv(s)$$

$$fv\text{-}ave(t) \quad = \quad \frac{1}{\mid S(t) \mid} \sum_{s \in S(t)} sv(s)$$

5.2.3 URL 素性

ノウハウサイトの候補集合内のドメイン t の URL に関して，2 種類の素性を提案する．一つ目は，ドメイン t が “com”，“org”，“jp”，“net”，または “co.jp” を含むか否かを示す，URL の地域コードに関する素性 $f_{d1}, \ldots, f_{d5}$ である．二つ目は，ウェブサイトが暗号化通信を行えるか否か，すなわち URL の先頭が “https” あるいは “http” のいずれであるかを示す素性 f_h である．

6. 評価

6.1 評価手順

本論文の訓練用データは，542(表 1 内のウェブサイトの候補集合の数) のドメインすべてから構成される．doc2vec によるウェブページのユークリッド距離の平均の素性以外の素性は訓練前に平均が 0，分散が 1 となるように正規化した [*3]．本論文では，scikit-learn パッケージにおける SVM (sklearn.SVM.SVC ツール) を用いて評価実験を行った．グリッドサーチの結果より，カーネル関数とガンマパラメータを設定した．全 6 個のクエリ・フォーカスのうち，5 個分のクエリ・フォーカスのサイトで訓練をし，残りの 1 個のクエリ・フォーカスで評価をするという 6 分割交差検定を行い，その平均値を結果として出力した．また，判定結果の推定信頼度を出力し，信頼度に対して下限値を設け，高信頼度な判定結果に限定した範囲での適合率，再現率の評価を行った．参照用ノウハウサイト集合 (正例集合) を R，信頼度 $conf$ が下限値 c 以上でノウハウサイト (正例) と判定されたサイト集合を $S(conf \geq C)$ として，再現率 $(conf \geq C)$，および，適合率 $(conf \geq C)$ を下式で定義

*2 https://ads.google.com/home/tools/keyword-planner

*3 予備実験において，doc2vec によるウェブページの平均距離素性の素性は，正規化なしの方が正規化しない場合よりも高いパフォーマンスを観測した．

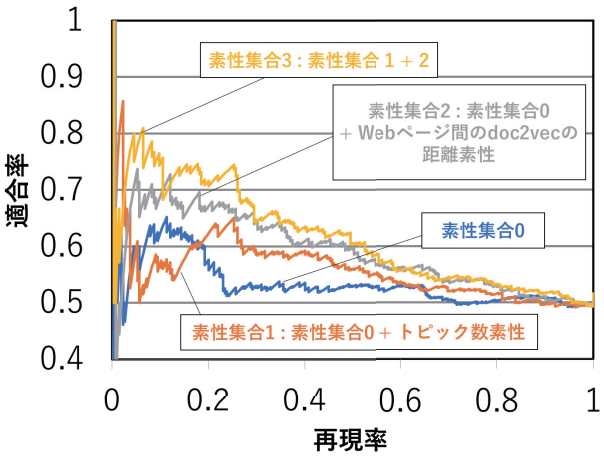

図 2: 評価結果 (素性集合 0:ウェブページ数素性 + 検索エンジンボリューム素性 + URL の素性集合，素性集合 1: 素性集合 1 + LDA のトピック数，素性集合 2:素性集合 0 + ページ内のウェブページの doc2vec ベクトルのユークリッド距離の平均，素性集合 3: 素性集合 1 と素性集合 2 の和集合)

した．

$$\text{再現率}\ (conf \geq c) = \frac{|R \cap S(conf \geq c)|}{|R|}$$

$$\text{適合率}\ (conf \geq c) = \frac{|R \cap S(conf \geq c)|}{|S(conf \geq c)|}$$

評価では，下記で定義する 4 つの素性集合ごとに，適合率と再現率をグラフにプロットすることにより評価を行った．また，これに加えて ROC 曲線を描き，AUC (ROC 曲線の下の面積) も測定した．

(i) 素性集合 0: ウェブページ数素性 + 検索エンジンボリューム素性 + URL の素性集合

(ii) 素性集合 1: 素性集合 1 + LDA のトピック数

(iii) 素性集合 2: 素性集合 0 + ページ内のウェブページの doc2vec ベクトルのユークリッド距離の平均

(iv) 素性集合 3: 素性集合 1 と素性集合 2 の和集合

6.2 評価結果

図 2 では，異なる 4 つの素性集合を用いて作成した分類器の評価を示した．信頼度のしきい値 c が，素性集合 (i) から (iv) についてそれぞれが 0 から 1 まで変化するため，各曲線は各素性集合の適合率/再現率のペアを表している．図 2 のグラフの形状と AUC による評価から，各素性が他の素性と比べて有効か否かについて述べる．doc2vec によるドメイン内のウェブページの文書ベクトルの平均距離素性は，LDA のトピック数素性よりもわずかに優れている．AUC による評価でも，素性集合 2 (doc2vec の距離素性を含む) は 0.58，素性集合 1 (LDA トピック素性を含む) とは 0.55 で，これは図 2 による評価と一致する．また，素性集合 0 の AUC は 0.54 で，素性集合 1 より低く，これも図 2 における比較と一致する．次に，図 2 において素性集合 1 および 2 の和集合を素性集合 1 および 2 と比較すると，和集合の方が素性集合 1 あるいは 2 よりもわずかに優れている．ただし，AUC に関しては，素性集合 1 および 2 の和集合は 0.58 で，素性集合 2 (doc2vec の距離素性を含む) と同程度であった．

7. 関連研究

関連研究として，[井上 16] では，検索エンジン，クエリ・フォーカスを用いて，サジェストおよびウェブページを収集，トピックモデルを用いることで話題を集約，ユーザの入力したクエリ・フォーカスに対し網羅的に話題とそのウェブページを示すためのインタフェースを作成した．[守谷 15] では，[井上 16] に対し，Yahoo 知恵袋などの質問回答サイトおよびウェブページから収集した文書集合を一つの文書集合とし，この混合文書集合に LDA トピックモデルを適用することで，より有用なノウハウ知識を得られることを示した．また，[李 17] では，収集したウェブページにトピックモデルを適用した結果に対し，複数のトピックにウェブページがまたがるドメインは，ノウハウサイトである可能生が高いと仮定し，いくつかのクエリ・フォーカスにおいて実際に評価を行い，候補サイトの半数以上がノウハウサイトであることを確認した．本研究は，[井上 16, 守谷 15] におけるトピックモデルの適用に加え，ページの話題の広がりを定量化する尺度として doc2vec を用いたという点で大きく異なっている．また，素性集合ごとに評価を行い，どの素性がノウハウサイトか否かを判定するのに重要かを検証した点においても大きく異なる．

8. おわりに

本論文では，ドメイン単位でそのウェブサイトがノウハウ知識を提示するページを有するか否かを判定する手法を提案した．分類に使用した SVM においては，ウェブサイト内のウェブページの話題分布に基づく素性，およびウェブページ数，検索エンジンの検索ボリューム数，URL に関するその他の素性を用いた．評価においては，LDA のトピック数素性と比べて，ウェブページ間の doc2vec ベクトルの平均距離素性が有効であることを示した．今後の課題として，大規模評価実験を通してこれらの知見の信頼性を検証することが挙げられる．

参考文献

[Blei03] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003).

[井上 16] 井上祐輔, 今田貴和, 陳磊, 徐凌寒, 宇津呂武仁, 河田容英：検索エンジン・サジェストおよびトピックモデルを用いたウェブ検索結果の集約, 第 8 回 DEIM フォーラム論文集 (2016).

[Le14] Le, Q. and Mikolov, T.: Distributed Representations of Sentences and Documents, *Proc. 31st ICML*, pp. 1188–1196 (2014).

[李 17] 李佳奇, 趙辰, 林友超, 馬場瑞穂, 宇津呂武仁, 河田容英, 神門典子：トピックモデルにおける話題分布特性を手がかりとするノウハウサイトの収集, 第 9 回 DEIM フォーラム論文集 (2017).

[守谷 15] 守谷一朗, 井上祐輔, 今田貴和, 聶添, 宇津呂武仁, 河田容英, 神門典子：質問回答事例および検索エンジン・サジェストを用いたノウハウ知識の相補的収集, 第 7 回 DEIM フォーラム論文集 (2015).