1K-02

分散表現を用いたWeb検索結果の自動タギング

細川涼平† 早川智一‡ 疋田輝雄‡
明治大学大学院理工学研究科† 明治大学理工学部‡

1. はじめに

近年，検索エンジンを用いたWeb検索結果のWebページ群（以下，Web検索結果）を整理する研究が報告されている．これは，Web検索結果のすべてが利用者の求めるWebページであるとは限らず，要求を満たすWebページを探す手間が利用者の負担となることに起因する．

Web検索結果のような，文書集合を整理する手段の1つにタギングがある．タギングとは，文書に対してその文書の特徴を表す語（タグ）を付与する手法のことである．タギングの利点は，文書内容を素早く理解できることおよびタグから類似文書を絞り込み検索できることである．

既存のタギング手法としては，文書中の重要語をタグとして用いる手法（以下，既存手法 1）や，タギング済みの文書を参考にして付与するタグを決定する手法（以下，既存手法 2）が一般的だが，それぞれに課題がある．既存手法1では，タグが文書中の単語に依存し，統一性がなくなるため，絞り込み検索が非効率的になることがある．また，既存手法2では，事前のタギングにコストが掛かることに加え，新たなタグを付与することが困難である．

本稿では，既存手法の課題を解決した上でWeb検索結果に対してタギングすることを目的とし，文書分類タスクのカテゴリ分類を参考にした手法を提案する．具体的には，サジェスト（3.2 節）から取得した，利用者の入力キーワード（以下，キーワード）と関係が深い単語（以下，関連語）をカテゴリとし，その特徴ベクトルの取得を単語の分散表現または検索エンジンで実現する．また，タギングにおいて付与したタグの正確性は重要であるため，タギングの精度も考慮し，既存手法との比較から有効性を確認した．

2. 関連研究

井上ら[2]は，Web検索において，サジェストをクラスタリングすることでキーワードの話題を集約し，話題に関する非冗長なWebページを選択的に提示可能にした．井上らと我々の研究とは，サジェストを用いてWebページを整理する点で類似する．一方で，彼らは，キーワードの話題とその話題に関するWebページを提供するが，我々は，あるWebページが示す話題を提供する点で異なる．

加藤ら[3]は，文書のタギングにおいて，文書中に出現する単語が，文書に割り当てたトピック中の重要単語である場合に，その単語をタグとする手法を用いた．加藤らと我々の研究とは，既存手法の課題解決を目的とする点で類似している．一方で，我々の研究は，文書中の単語を直接タグとして使用しない点で彼らの研究と異なる．

3. 提案手法

提案手法では，関連語を事前に用意しておき，関連語集合の中から付与するタグを選択することで，既存手法1の課題および既存手法2の新たなタグの付与が困難な課題を解決する．また，この手法では一般的に，既存手法2と同様に，関連語に関するタギング済み文書が必要となる．そのため，（1）単語の分散表現の利用や，（2）検索エンジンを用いた，関連語に関する文書の取得――によって既存手法2の事前のタギングが必要な課題を解決する．

3.1 提案手法の概要

提案手法の概要を図1に示す．提案手法の流れは以下のとおりである．まず，（1）利用者からキーワードの入力を受け取り，(2) キーワードをクエリとしてWeb検索を行う．このWeb検索結果がタギング対象となる．また，（3）関連語を複数取得しておく．次に，（4）タギング対象の各Webページおよび各関連語を特徴ベクトル化し，（5）タギング対象の各Webページと各関連語の類似度を求め，（6）類似度が閾値を超える関連語をそのWebページのタグとする．最後に，（7）類似するタグをまとめ，（8）タギング後のWeb検索結果を利用者に提供する．

3.2 関連語の取得

サジェストを利用して関連語を用意する．サジェストとは，検索エンジンに蓄積されたクエリログを参考にし，キーワードに関連する語句を抽出する機能のことである．

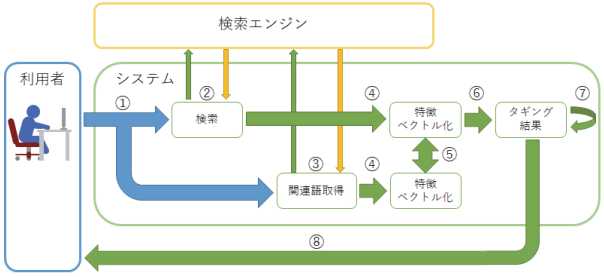


図1 提案手法の概要

Fig. 1 Overview of the proposed method.

Automatic tagging for web search results by using distributed representation
†Ryohei Hosokawa, ‡Tomokazu Hayakawa and Teruo Hikita
†Graduate School of Science and Technology, Meiji University
‡School of Science and Technology, Meiji University

3.3 特徴ベクトル化

本稿では，単語および文書の特徴ベクトル化のために，単語の分散表現を用いた．単語の分散表現とは，単語を密なベクトルで表す手法のことである．文書を単語集合とすると，単語の分散表現により特徴ベクトル化が可能である．

単語の分散表現の学習には，Mikolov ら[1]が提案した word2vec の Skip-gram モデルを用いた．パラメータは実験的に，分散表現の次元数 200，ウィンドウ幅 5 とした．また，学習用コーパスとして日本語版 Wikipedia を用いた．

3.4 関連語の特徴ベクトル化

関連語の特徴ベクトル化には，単語の分散表現のみを用いる手法（以下，提案手法 1）および単語の分散表現と検索エンジンとを用いる手法（以下，提案手法 2）を検討した．図 2 に関連語の特徴ベクトル化の手順を示す．

提案手法 1 の手順は以下のとおりである．(1) キーワードおよび各関連語の分散表現ベクトルを抽出し，(2) それらの平均値を関連語の特徴ベクトルとする．提案手法 1 は，関連語の特徴ベクトル化が容易である一方，キーワードと関連語間の関係性を表現できていない可能性があるため，タギング精度を考慮し，提案手法 2 を採用した．

提案手法 2 の手順は以下のとおりである．まず，(1) キーワードと関連語の AND 検索を行う．検索で得た Web ページ上位 50 件のスニペットおよびタイトルを，関連語についての 1 文書とした．また，特徴語として名詞のみを用いた．次に，(2) 文書を形態素解析し，文書中の各単語の tf-idf 値を求める．最後に，(3) 分散表現ベクトルと tf-idf 値を掛けた値を各単語の特徴ベクトルとし，それらの和を関連語の特徴ベクトルとする．

3.5 タギング対象 Web ページの特徴ベクトル化

タギング対象 Web ページの特徴ベクトル化は，提案手法 2 の（2）から（4）までの手順と同様である．各 Web ページの文書としては，そのタイトルとスニペットを用いた．

3.6 タギング

特徴ベクトルを基にして，各タギング対象 Web ページと各関連語との類似度を求め，類似度が閾値を超えた関連語を，その Web ページのタグとして付与する．類似度の計算にはコサイン類似度を用いた．また，見やすさのために，付与するタグの中で互いに類似するものも同様に，コサイン類似度と閾値とを参考にし，1 つの集合にまとめた．

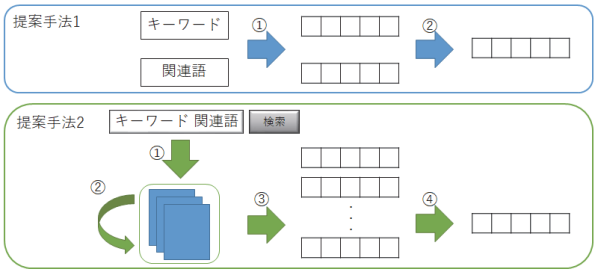
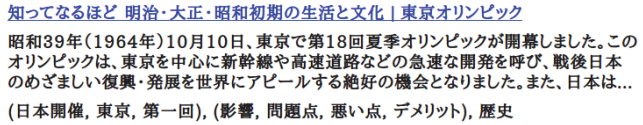


図 2　関連語の特徴ベクトル化の手順

Fig. 2　Converting related words to feature vectors.


知ってなるほど 明治・大正・昭和初期の生活と文化 | 東京オリンピック

昭和39年（1964年）10月10日、東京で第18回夏季オリンピックが開幕しました。このオリンピックは、東京を中心に新幹線や高速道路などの急速な開発を呼び、戦後日本のめざましい復興・発展を世界にアピールする絶好の機会となりました。また、日本は...

(日本開催, 東京, 第一回), (影響, 問題点, 悪い点, デメリット), 歴史

図 3　タギング結果例

Fig.3　Example of tagging.

表 1　タギング結果の評価

Table 1　Evaluation of tagging.

手法	正解率	有用率
提案手法 1	0.64	0.46
提案手法 2	0.76	0.57
既存手法 1（tf-idf）	0.52	0.43

4. タギングの実行結果

タギングの実行結果例を図 3 に示す．図 3 は，提案手法 2，キーワード「オリンピック」での結果である．結果画面では，タイトルが上段，スニペットが中段，タグが下段に表示される．互いに類似するタグは括弧でまとめられる．

図 3 の Web ページは，第 1 回およびそれ以前に計画された東京オリンピックや，その時の事件に関する内容である．それに対し，「東京」や「第一回」，「事件」，「歴史」などのタグが付与されており，結果が適切であることが分かる．

5. 評価

提案手法によるタギングの精度を評価した．具体的には，キーワード「オリンピック」および「錦織圭」で，Web 検索結果上位 100 件に最大 3 種類のタグを付与し，付与したタグの中で，Web ページの内容と一致しているタグの割合を「正解率」，十分参考になるタグの割合を「有用率」として値を算出した．評価のための比較対象として，既存手法 1 において，重要語を tf-idf 値から求めたものを用いた．

評価結果を表 1 に示す．表より，両提案手法の正解率が既存手法に比べ，大きく上回った．一方，有用率はわずかな向上にとどまった．これは，サジェストによる関連語において，「話題」などの抽象的な単語が多いことに起因する．

6. おわりに

本稿では，既存手法の課題を解決した上で Web ページに対するタギングを行う手法を提案した．今後の課題は，正解率と有用率の向上や，既存手法との更なる比較である．

参考文献

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. “Distributed Representations of Words and Phrases and their Compositionality”. In *Advances in Neural Information Processing Systems 26*, pp. 3111-3119, 2013.

[2] 井上祐輔，今田貴和，陳磊，徐凌寒，宇津呂武仁，“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.

[3] 加藤亮，吉川大弘，古橋武．“潜在的ディリクレ配分法を利用した文書への自動タグ付与に関する検討”．第 28 回人工知能学会全国大会論文集, 2014.