関連語の分散表現に基づく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　〒214–8571 神奈川県川崎市多摩区東三田 1-1-1
†† 明治大学理工学部　〒214–8571 神奈川県川崎市多摩区東三田 1-1-1
E-mail: †{jamxzzz,t_haya,hikita}@cs.meiji.ac.jp

あらまし　Web 検索結果を整理して検索効率を向上させるための手法の 1 つとして，自動タギングがある．本稿では，Web 検索結果を対象とした自動タギングにおいて，タギングの人気のもとであるフォークソノミの特徴の観点から既存手法の課題を改善する手法を提案する．提案手法では，検索クエリログから動的に得た関連語と Web ページとの類似度を分散表現に基づいて計算し，類似度が閾値を超えた関連語をその Web ページのタグとすることで自動でタギングをおこなう．提案手法の評価を既存手法の課題改善およびタギング性能の観点からおこない，提案手法が Web 検索結果に対する自動タギング手法として有用たり得ることを確認した．

キーワード　Web 検索，自動タギング，クラスタリング，検索クエリログ，分散表現

1. はじめに

1.1 背　　景

Web 検索エンジンを用いた Web ページ検索（以下，Web 検索）が広く一般に普及する (注1) につれて，Web 検索結果を自動で整理する研究が盛んになっている．この理由の 1 つとして，Web 検索結果を整理することによってユーザの Web 検索結果選別の負担が軽減され，検索効率の向上が期待できることが挙げられる．一般に，Web 検索結果にはユーザにとって不要な Web ページが混入しやすいため，ユーザは，Web 検索結果の選別を，タイトルとスニペット（2・3 文程度の概要）とを読んで（場合によっては本文を読んで）おこなう必要があることが多い．そのため，検索効率向上の観点から，Web 検索結果を整理してユーザに提供することで，ユーザの Web 検索結果選別の負担を軽減することが望ましい．

Web 検索結果をはじめとする文書集合を整理する手段の 1 つとして，タギングがある．タギングとは，情報資源に対して，タグと呼ばれるキーワードを複数付与して整理する手法のことである．タギングは，フォークソノミ（Web 上の情報資源を多数のユーザからのタギングにより分類する手法）を実装したサービスの普及に伴って，近年人気を集めている [5] ．各 Web ページに対してタグを複数付与することで，次の 2 つの観点からユーザの選別の負担を軽減することが可能である：

（1） タグが Web ページの要約になり，タグを見るだけで——長い文章を読むことなく——Web ページの選別が可能になり得るため，選別のために読む文字数が削減される；
（2） タグにより Web ページを分類することが可能なため，興味のあるグループを選ぶことで複数の Web ページが取得でき，Web ページの選別回数が削減される．

タギングにより文書集合を整理する際に，タギングの特徴を考慮することは重要である．ここでのタギングの特徴とは，タギングの人気のもとになったフォークソノミの特徴を指す．タギングの特徴については様々な議論がされており [1, 3–5, 7, 11–13]，要約や分類以外に次の利点がある：

（1） 多数のユーザにより情報資源が管理されているため，新しく生まれた単語や既存の単語の意味の変化に即時に対応できる；
（2） 複数のタグにより広範な話題が示されるため，ロングテールのトピックの発見が可能である．

また，次の欠点がある：

（1） 多数のユーザによりタギングされるため，表記ゆれや同義語のタグが多くなり，タグによる分類がうまく作用しないことがある；
（2） 専門知識のない一般のユーザがタギングするため，誤ったタグや悪意のあるタグが付与されることがある；
（3） タグが短い語で表現されるため，複数の意味をもつ単語をタグとする際に曖昧さが残ることがある．

文書集合に対する既存の自動タギング手法としては，(1) 文中からの重要語抽出に基づくものや，(2) あらかじめタギングされた文書を使用するもの——があるが，これらはそれぞれに，タギングの特徴の観点で課題が残っている．既存手法 (1) の課題は，(1) 文中の重要語以外の語をタグとして使用できないことや，(2) タグの表現力が弱いこと——である．既存手法 (1) は，タグが文中の単語に依存するため，表記ゆれや同義語タグが多くなる．また，文中の重要語以外のタグが付与できないため，表現力が弱い（たとえば，Google の Web 検索エンジンサイト (注2) には，「検索エンジン」というタグが付与できない）ということも指摘されている [5]．既存手法 (2) の課題は，(1) 事前の準備に時間やコストがかかることや，(2) 新語への対応が

(注1)：2015 年の総務省の調査 [18] によると，最も頻繁に利用する情報収集手段を尋ねた際に Web 検索エンジンと回答した人数は約 7 割であり，他の手段を利用する人数と比べて圧倒的に多い．

(注2)：https://www.google.co.jp/

困難なこと――である．既存手法（2）は，事前にタギングされた文書を必要とするため，その準備に時間が掛かる．また，事前に準備した文書に付与されたタグ以外のタグが付与できないため，新語への対応が困難である．

1.2 研究概要

本稿では，タギングの特徴の観点から既存手法の課題を考慮したうえで，Web検索結果に対してタギングすることを目的とする．具体的には，

（1） 表記ゆれや同義語タグをできる限り少なくし，
（2） 文中の単語以外のタグを付与できるようにし，
（3） 事前にタギング済みの文書を使用することなく，
（4） 新語のタグを付与できるようにし，
（5） 誤ったタグが付与されることを防ぐためにタギングの性能を考慮し，
（6） タグの曖昧性をできる限り少なくする

――ことを考慮してタギングをおこなう．

なお，悪意のあるタグやロングテールのトピックについては考慮しない．なぜならば，自動のタギングであれば，悪意のあるタグは付与されず，ロングテールのトピックのタグも自然に付与されるためである．

提案手法によるタギングの手順は次のとおりである．まず，(1) ユーザから入力されたクエリ（以下，入力クエリ）と関わりの深い単語（以下，関連語）のリスト（以下，関連語リスト）を動的に作成する．次に，(2) 関連語リストと Web 検索結果との類似度を計算する．最後に，(3) 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．なお，関連語リストの作成には検索クエリログ（3.3 項）を使用し，関連語と Web ページとの類似度計算のための指標としては分散表現（3.4 項）を用いた．

1.3 本稿の構成

本稿の構成は次のとおりである．2 節では，関連研究および関連技術について述べる．3 節では，提案手法について詳説する．4 節では，提案手法の評価結果を報告する．5 節では，本稿のまとめと今後の展望について言及する．

2. 関連研究・関連技術

2.1 重要語抽出に基づくタギング

Brooks ら [3] は，ブログ記事から抽出した tf-idf 値の高い重要語をタグとして用いる自動タギング手法を考案し，その手法の有効性を，人手によるタギングとの比較から示した．Brooks らの手法をはじめとする重要語抽出に基づくタギング手法は，タグの表記ゆれや同義語への対応が困難なことに加え，タグの表現力が乏しい場合がある．我々の手法は，表記ゆれや同義語を削減することや，文中の単語に依存せずにタグを付与することを考慮している点でこれらの手法とは異なる．

2.2 タギング済み文書を使用するタギング

Mishne [9] は，協調フィルタリングをタギングに適用し，大量のブログ記事を基にして未知のブログ記事に付与するべきタグを自動で推薦する “Auto Tag” というツールを考案した．Auto Tag では，ユーザをブログ記事，製品をタグとして協調フィルタリングが適用される．Mishne の手法は，大量のタギング済み文書が必要なことに加え，それらの文書に付与されたタグ以外のタグが付与できず，新語への対応が困難である．我々の手法は，タギング済み文書ではなく生の文書を用いる(注3)点や，新語への対応ができる点で Mishne の手法とは異なる．

Ohkura ら [10] は，Web から収集した，カテゴリタグが付与されているブログ記事を教師データとしてタグごとに SVM（Support Vector Machine）で 2 クラス分類器を作成し，それらをタグ付与器として用いるタギング手法を提案した．また，タグ付与器の学習を繰り返しおこなうことで，新語などの語彙の変化に対応した．Ohkura らの手法は，大量のタギング済み文書が必要なことに加え，頻繁な再学習とタギング済み文書の管理とが必要である．我々の手法は，タギング済み文書ではなく生の文書を用いる点や，再学習と文書の管理とを基本的に必要とせずに新語に対応できる(注4)点で Ohkura らの手法とは異なる．

2.3 その他のタギング

加藤ら [15] は，LDA（Latent Dirichlet Allocation）により，大量に入力された文書を基にして各文書にトピックを割り当て，トピック中の term-score（tf-idf における文書をトピックに置き換えた手法）上位の重要語をその文書のタグとする手法を提案した．加藤らの手法では，タグとして用いる重要語の抽出範囲を単一の文書からトピック中の複数の文書に拡大することで，単一の文書からの重要語抽出手法と比較して，表記ゆれや同義語が少なく，タグの表現力の低さも軽減される．一方で，加藤らの手法は，文中の重要語をタグとすることによる表記ゆれや同義語の課題およびタグの表現力の課題は残っているように思われる．我々の手法は，文中の単語と完全に独立してタグが付与される点で加藤らの手法とは異なる．

2.4 文書分類

文書の整理・分類手法としての非排他的な文書分類と提案手法のタギングとは，あらかじめ設定したカテゴリ（提案手法での関連語）に文書を割り当てる点で類似性がある．一方で，文書分類は分類先のカテゴリが静的に決められているが，提案手法のタギングは分類先のカテゴリが動的に決まるという点で両者は異なる．

2.5 文書クラスタリング

文書の整理・分類手法としての非排他的な文書クラスタリング（とくに，クラスタラベルを先んじて決定するクラスタリング）と提案手法のタギングとは，分類先のクラス（提案手法での関連語）が動的に決まる点で類似性がある．一方で，提案手法のタギングは，分類先クラスの粒度が細かいという点で一般の文書クラスタリングとは異なる．

我々の研究と最も類似する文書クラスタリングの研究として，安川ら [17] の研究がある．安川らは，Web 検索結果を対象に，

(注3)：生の文書を使用する利点としては，タギング済み文書と比べて文書に対する手動のタギングが必要ないため準備コストがあまり掛からず，使用可能である文書も多いことが挙げられる．
(注4)：提案手法では再学習をしなくても新語などの語彙の変化に対応できるが，タギング性能を考慮する場合，低い頻度で随時再学習することが好ましい．

検索クエリログから得た関連語をクラスタラベルとして使用するクラスタリング手法を提案した．安川らの研究と我々の研究とは，分類先のクラスとして検索クエリログから得た関連語を用いる点で類似性がある．一方で，安川らの研究は，クラスの要素としてクラスタラベルが含まれている Web ページを割り当てるが，我々の研究は，関連語と Web ページとの意味的類似度から要素を決定するという点で両者は異なる．

2.6 AND 検索

ユーザクエリ q での検索結果をタグ t で絞り込んだ Web ページ群と，q と t との AND 検索で得られる Web 検索結果とは，q と t とに関する Web ページ群が得られる点で類似する．一方で，まったく同じ結果は得られないという点で両者は異なる．AND 検索による絞り込みは，q と t とを含む Web ページのみを提供する(注5)ため厳格であるが，タグによる絞り込みは，t が含まれていないが t を意図する Web ページが得られる場合がある．

3. 提 案 手 法

3.1 タギング対象

本稿では，Web 検索結果を対象としてタギングをおこなう手法を提案する．これは，提案手法が入力クエリに依存するためである．そのため，ふつうの文書集合は本稿の対象外となる．ただし，文中の重要語を擬似的な入力クエリとすることで，ふつうの文書集合にも提案手法を適用することは可能である．

また本稿では，入力クエリとして，informational query [2] を主な対象としている．informational query とは，ある事柄についての調査のためのクエリのことで，入力クエリの約 5 割を占めるとされている．informational query を主な対象とする理由は，informational query のように複数の Web ページを必要とする入力クエリでこそタギングの利点が活きるためである．

3.2 提案手法の概要

提案手法では，タギング処理の前に入力クエリの関連語リストを作成し，関連語リストと Web 検索結果との類似度を計算し，類似度が閾値を超えた関連語をその Web ページのタグとすることで Web 検索結果に対してタギングをする．

提案手法の概要を図 1 に示す．提案手法によるタギングの流れは次のとおりである．まず，(1) ユーザから入力クエリを受け取り，(2) タギング対象である，入力クエリによる Web 検索結果を取得し，(3) 検索クエリログを利用して，入力クエリから関連語リストを作成する．次に，(0) あらかじめ学習された分散表現に基づき，(4) 各 Web ページの分散表現ベクトルおよび (5) 各関連語の分散表現ベクトルを作成する．さらに，(6) 関連語リスト中の表記ゆれ・同義語タグを結合する．そして，(7) 各 Web ページと各関連語との類似度を計算し，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．最後に，(8) (表記ゆれ・同義語タグとは別に) 類似したタグをまとめ，(9) タギング結果をユーザに返す．

(注5)：正確には，検索エンジンによるクエリの修正により，入力クエリに含まれていない Web ページが得られる場合がある．

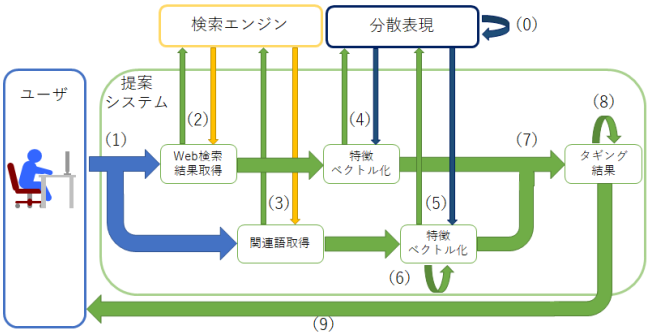

図 1 提案手法の概要

3.3 関連語リストの作成

提案手法は，最初に関連語リストを作成し，リスト中の関連語をタグの候補とする．つまり，Web ページに付与されるタグは，必ず関連語リスト中に存在し，関連語リストにない単語がタグとして付与されることはない．このような関連語リストを作成する利点としては，次のことが挙げられる：

(1) 関連語リスト作成とタギング処理とが独立していて関連語の追加や削除が容易であるため，表記ゆれや同義語，新語タグに対応しやすい；
(2) タグが文中の単語やあらかじめ用意した文書に付与されたタグに依存しないため，文中に登場しない単語をタグとして付与することや新語へ対応することが可能である．

提案手法では，関連語リストを自動で作成するために，検索クエリログを利用する．検索クエリログとは，検索エンジンに蓄積される，多数のユーザからの入力クエリのログのことである．検索クエリログを利用することで，ある単語と頻繁に検索されている別の単語群を得ることができる．

関連語リストの自動作成に検索クエリログを利用することで，次の利点がある：

(1) 多数のユーザからの検索要求が得られるため，関連語の話題の網羅率が高い；
(2) 入力クエリごとに動的に粒度の細かい関連語リストが作成できる；
(3) 日々大量のクエリが入力されるため，新語を関連語として取得することができる；
(4) ユーザから多数入力されたクエリを参照できるため，表記ゆれや同義語が少ないことが期待できる．

3.4 分 散 表 現

提案手法では，関連語と Web ページとの類似度計算のための指標として分散表現を使用する．分散表現とは，単語や文書のベクトル表現を，従来の疎な離散値のベクトルとは異なる，任意次元の密な連続値のベクトルで表現する技術のことである．分散表現は，Mikolov ら [8] の word2vec をきっかけとして，近年，多くの自然言語処理タスクに用いられている．

分散表現を用いることにより，次の利点がある：

(1) 分散表現の加法構成性（単語の足し引きができる性質）を利用することで，関連語と Web ページとの類似度計算が簡単にできる；

（2） 分散表現の連続値の密なベクトルから単語の細かな意味が表現できるため，タギング性能の向上が期待できる．

分散表現は，コーパスを教師データとして，教師データ中に含まれる単語の分散表現を学習する．本稿では，分散表現の学習はシステム稼働前に一度だけおこなうことを想定している．なお，既存手法による学習とは異なり，タギング済み文書ではなく，生の文書から学習をおこなうことに注意されたい．

分散表現の学習のために，gensim 版の word2vec を用いた．学習のモデルには Skip-gram モデルを使用し，パラメータとしては実験的に，次元数を 200 に設定した．学習のためのコーパスとしては，汎用性を考慮して日本語版 wikipedia コーパスを使用し，その分かち書きのために，新語や固有名詞を考慮して MeCab [6] と mecab-ipadic-NEologd [16] とを用いた．

3.5 Web ページの分散表現

Web ページの分散表現を求める前に，Web ページの特徴として使用する文章を決定する必要がある．Web ページの特徴として使用する文章としては，(1) Web ページの本文すべてや，(2) Web ページ本文から抽出した重要文，(3) Web ページのスニペットやタイトル——などが考えられる．

本稿では，Web ページのスニペットとタイトルとを Web ページの文章として用いる．この理由としては，(1) スニペットとタイトルとは，本文を使用する場合に比べて Web ページ間のサイズに差異がでにくいことや，(2) 処理が簡単なこと——が挙げられる．また，スニペットとタイトルとを文章として使用することは，本文を使用することと比べても性能的に遜色なく，処理時間の削減になることが，類似タスクのクラスタリングの研究で報告されている [14]．

Web ページの特徴として使用する文章を決めた後に，Web ページを分散表現ベクトル化する．Web ページの分散表現ベクトルの構築手順は次のとおりである．まず，(1) MeCab と mecab-ipadic-NEologd とを用いてタイトルとスニペットとを形態素解析し，入力クエリを除く名詞および動詞を抽出する．次に，(2) 抽出したそれぞれの単語の tf-idf 値を求める．最後に，(3) 抽出した単語ごとに分散表現と tf-idf 値を掛け合わせた新たなベクトルを求め，そのベクトルの平均値を Web ページの分散表現ベクトルとする．

3.6 関連語の分散表現

関連語の分散表現としては，学習した単語の分散表現をそのまま用いることが考えられるが，それには問題がある．具体的には，新語や固有名詞など，学習時点でコーパスに登場しない単語の分散表現ベクトルが得られないことである．そのため，分散表現をそのまま用いる場合は，新語への対応が困難である (注6)．

提案手法では，関連語の分散表現ベクトルを求めるために，関連文書により関連語を拡張する．具体的には，入力クエリと関連語とで AND 検索して得られる Web 検索結果上位数件を関連語の拡張に用いた．これにより，新語や固有名詞など，学習時にコーパスに含まれない単語であっても擬似的に分散表現ベクトルが得られる．

関連語の分散表現ベクトルの構築手順は次のとおりである．まず，(1) 入力クエリと関連語との AND 検索により関連語の関連 Web ページを取得する．次に，(2) 得られた 30 件の Web ページのタイトル群とスニペット群とを 1 つの文書とし，3.5 項と同様の方法でその文書を分散表現ベクトル化する．

3.7 表記ゆれ・同義語の削減

検索クエリログによって自動で作成した関連語リストには表記ゆれや同義語が混入しているため，クラスタリングによりそれらを削減する．具体的には，類似度の計算にコサイン類似度を用い，閾値を実験的に設定し，最短距離法にて凝集型クラスタリングをおこなった．

クラスタリングによる表記ゆれ・同義語削減の手順は次のとおりである．まず，(1) 関連語リスト中の関連語どうしの類似度を計算し，最も類似度が高い関連語の組を求める．次に，(2) 両関連語の平均のベクトルを新たなベクトルとしてリストに追加し，両関連語をリストから削除する．最後に，(3) 類似度の最高値が閾値を下回るまで (1) と (2) とを繰り返す．

3.8 タギング

関連語リストと Web 検索結果とを突き合わせて類似度を計算し，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，タグの見やすさと曖昧さの軽減とを考慮して，表記ゆれや同義語のクラスタリング時に定めた値よりも低い閾値を設定し，類似度が閾値を超えた類似タグどうしをクラスタリングによりまとめた．

4. 評　　価

1.2 項で示した目的を基にして，提案手法の評価をおこなった．表記ゆれや同義語については 4.3 項で，文中の単語以外のタグの付与，新語タグの付与，タグの曖昧さの軽減については 4.2 項で，タギングの性能については 4.4 項でそれぞれ説明する．なお，タギング済みの文書を使用しないことは 3 節で示しているためここでは説明しない．

4.1 実験データ

提案タギング手法の評価のためのクエリ（以下，評価クエリ）として，2016 年および 2017 年の Google Trends ランキング (注7) の急上昇ワード上位 5 件の計 10 クエリを用いた．

Web ページの取得には Google Custom Search API (注8) を用い，タギング対象として上位 50 件，関連語の拡張のために上位 30 件の Web ページをそれぞれ取得した．なお，定めた数の Web ページを得られないクエリについては，得られるすべての Web ページを使用した．

関連語の取得には Google のサジェストを利用した．具体的には，評価クエリと a から z までのアルファベット 1 文字とをそれぞれ組み合わせることで，評価クエリごとに 200 程度の関連語を取得した．

(注6)：新語に対応するためには，コーパスを変更して再学習する必要がある．

(注7)：https://trends.google.co.jp/trends/topcharts

(注8)：https://developers.google.com/custom-search/

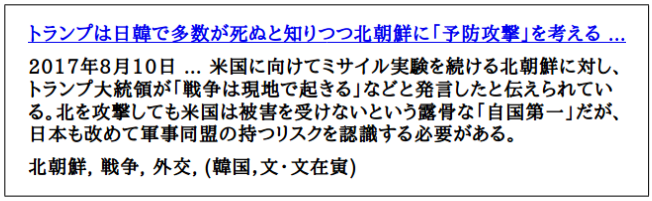

図 2　提案手法によるタギング結果例 1

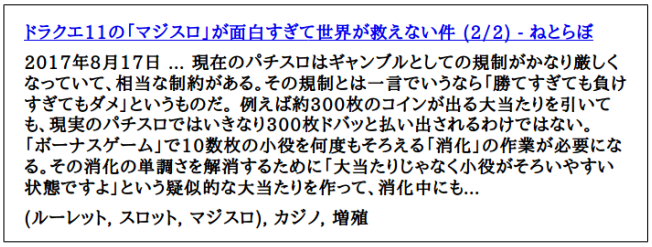

図 3　提案手法によるタギング結果例 2

4.2　タギング結果例

図 2 と図 3 とに提案手法によるタギングの結果例を示す．図の上段と中段とが，それぞれタイトルとスニペットとを表し，下段が提案手法により付与されたタグを示す．タグは，タグごとにカンマで区切られ，表記ゆれ・同義語タグは中黒で並べられ，類似タグは丸カッコで囲まれて表示される．

図 2 は，評価クエリ「トランプ」から得られる検索結果に対するタギング結果を抜粋したものである．この Web ページ (注9) は，現アメリカ大統領のトランプ氏が北朝鮮との戦争について発言した内容を解説した記事である．これに対して，「北朝鮮」や「戦争」，他国との関係を表す「外交」，関係国の「韓国」などのタグが付与されている．

図 2 で特筆するべき点は，(1) 文中に登場しない「外交」や「韓国」，「文・文在寅」というタグが付与されていること，(2) 現韓国大統領の文在寅氏を表す「文」と「文在寅」とが表記ゆれ・同義語として処理されていること，(3)「韓国」タグと「文・文在寅」タグが類似するタグとして処理されていること――である．(1) は，提案手法では文中に登場しない単語をタグとして付与することができることを示している．(2) では，「文」タグと「文在寅」タグとが表記ゆれ・同義語としてまとめられ，表記ゆれや同義語の削減につながっていることが分かる．(2) と (3) とから，「文」という曖昧な単語が「文在寅」や「韓国」とまとめられることで，曖昧さが軽減されていることが確認できる．これらのことから，提案手法は，文中にないタグの付与や，ある程度のタグの曖昧さの軽減が可能であるといえる．

図 3 は，評価クエリ「ドラクエ 11」から得られる検索結果に対するタギング結果を抜粋したものである．この Web ページ (注10) は，ドラクエ 11 というゲーム中のスロットを模したミニゲームである「マジスロ」について説明・評価した Web ページである．これに対して，「マジスロ」や「スロット」，マジスロで遊ぶ施設を示した「カジノ」，コインが増えることを意図しているであろう「増殖」などのタグが付与されている．

図 3 で特筆するべき点は，文中からの抜き出しではないにも関わらず，「マジスロ」という新語のタグが付与できていることである．タギングを実行した時期が 2018 年 2 月であるのに対し，ドラクエ 11 が発売された時期が 2017 年 7 月であるため，「マジスロ」という単語は比較的新しいものであるといえる．このことから，提案手法は新語に対応可能であることが分かる．

4.3　表記ゆれ・同義語

関連語の表記ゆれ・同義語の評価のために，自動で作成した関連語リスト中の表記ゆれまたは同義語の組数を人手で調べた．また，表記ゆれ・同義語以外の関連語が結合されないように閾値を定めてクラスタリングをおこなった後の，関連語リスト中の表記ゆれ・同義語組数を求めた．

表 1　関連語リスト中の表記ゆれ・同義語組数

	平均値	最大値	最小値	中央値
関連語数	204.6	220	193	206
表記ゆれ・同義関連語組数	8.2	16	5	7
クラスタリング後の表記ゆれ・同義関連語組数	3.2	5	2	3

評価結果を表 1 に示す．表から，検索クエリログにより作成した関連語リスト中には約 200 個の関連語があり，その内で表記ゆれや同義語は約 8 組あり，クラスタリング後に約 3 組になることが分かった．このことから，関連語リスト中の表記ゆれ・同義語数は少なく，表記ゆれ・同義語タグが付与されることはあまりないといえる．

4.4　タギングの性能

提案手法のタギングの性能は，(1) タグとして付与するべき単語が関連語リスト中に含まれていること（関連語の話題網羅率）と，(2) 関連語リスト中に含まれるタグを正しく付与できること（タギングの適合率・再現率・F 値）と――により決定される．そのため，両者の評価によりタギング性能を確認する．

4.4.1　関連語の話題網羅率

関連語リストにない単語をタグとして付与することはできないため，タグとして付与するべき単語はリストの中に含まれていることが必須である．そのため，関連語リスト中の単語が話題を網羅しているか否かを調べる必要がある．

評価のために，5 名の評価者が人手で正解データを作成した．評価手順は次のとおりである．まず，(1) 検索結果の上位 50 件の Web ページに対して人手で自由にタグを付与する．次に，(2) 付与したタグのリスト（以下，正解リスト）を作成し，表記ゆれ・同義語を排除する．最後に，(3) 正解リストと関連語リストとを突き合わせ，正解リスト中の単語と類似する単語が関連語リストに含まれているか否かを確認し，その割合を計算する．

評価結果を表 2 に示す．表から，検索クエリログより得られた 200 の単語によって，上位 50 件の Web ページの話題を，平均約 8 割網羅することが分かった．網羅できなかった単語としては，固有名詞や，「Amazon」，「Instagram」，「pixiv」などが挙げられる．固有名詞を除くこれらの単語は，入力クエリに依らずにあらかじめ関連語リストに追加しておくことで，網羅率は現状より改良されると考えられる．

(注9)：http://diamond.jp/articles/-/138120
(注10)：http://nlab.itmedia.co.jp/nl/articles/1708/17/news031_2.html

表 2　関連語リストの話題網羅率

	平均値	最大値	最小値	中央値
正解リストのタグ数	66.5	71	63	66
関連語リストおよび正解リスト両方に含まれるタグ数	54.3	57	51	54
タグの話題網羅率	0.82	0.86	0.76	0.81

表 3　タギングの適合率・再現率・F 値

	平均値	最大値	最小値	中央値
適合率のマイクロ平均	0.70	0.74	0.65	0.70
再現率のマイクロ平均	0.74	0.77	0.69	0.72
F 値のマイクロ平均	0.72	0.77	0.68	0.72

4.4.2　タギングの適合率・再現率・F 値

正しいタグを付与するためには，関連語リスト中に付与するべきタグが含まれていてかつシステムがそのタグを付与できることが求められる．そのため，タギングの再現率，適合率，F 値を基にして評価をおこなった．

評価のために，人手で最大 5 個のタグを付与し，正解データを作成した．ここで，タグ数を最大 5 個とした理由は，関連語の話題の網羅率評価のために作成した正解データのタグ数が概ね 5 個を下回ったためである．評価手順は次のとおりである．まず，(1) 検索結果の上位 50 件の Web ページに対して，関連語リスト中から人手で最大 5 個のタグを付与する．次に，(2) システムで類似度の高い順に最大 5 個のタグを付与し，人手で作成した正解データと比較して適合率，再現率，F 値を求める．

評価結果を表 3 に示す．表から提案手法によるタギングにおいて，上位 5 件のタグを付与した場合に，適合率・再現率・F 値すべてにおいて平均約 7 割の値が得られた．また，「ドラクエ 11」や「ポケモン go」など，関連語に固有名詞が多い評価クエリは，他の評価クエリと比較して適合率が低くなった．これは分散表現が分布仮説に基づいている影響だと考えられる．以上から，提案手法によるタギングは，関連語として固有名詞の多いクエリについては性能が下がるものの，F 値として約 7 割の値が得られるため，実用に耐える性能があることが分かった．

5. おわりに

本稿では，Web 検索結果の整理・分類手法としての自動タギングにおいて，タギングの利点や欠点の観点から既存手法の課題を指摘し，既存手法とは異なるタギング手法を提案した．提案手法では，検索クエリログから得た関連語と Web 検索結果との類似度を，分散表現を用いて比較することでタギングをおこなった．また，提案手法をタギングの利点や欠点の観点から評価し，提案手法が Web 検索結果に対する有効な自動タギング手法たり得るという結論を得た．

今後の展望としては，(1) 分散表現学習時の単語の曖昧性解消や，(2) 関連語リスト中の関連語の洗練――が考えられる．近年，単語の曖昧性を考慮して分散表現ベクトルを学習することによる性能向上が報告されており，これを提案手法に適用することで，タギング性能の向上が期待できる．また，本稿では，類似度の高いタグを付与してしているが，そのタグがユーザにとって必ずしも有用であるとは限らない．そのため，関連語リストの洗練により，タグの有用性を考慮することが好ましい．

文　献

[1] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su, “Optimizing web search using social annotations,” In Proceedings of the 16th international conference on World Wide Web, pp. 501-510, 2007.

[2] A. Broder, “A taxonomy of web search,” ACM Sigir forum, Vol. 36, No. 2, pp. 3–10, 2002.

[3] C. H. Brooks, and N. Montanez, “Improved annotation of the blogosphere via autotagging and hierarchical clustering,” In Proceedings of the 15th international conference on World Wide Web, pp. 625–632, 2006.

[4] S. A. Golder, and B. A. Huberman, “Usage patterns of collaborative tagging systems,” Journal of information science, Vol. 32, No. 2, pp. 198–208, 2006.

[5] M. Gupta, R. Li, Z. Yin, and J. Han, “Survey on social tagging techniques,” ACM Sigkdd Explorations Newsletter, Vol. 12, No. 1, pp. 58–72, 2010.

[6] T. Kudo, K. Yamamoto, and Y. Matsumoto, “Applying conditional random fields to Japanese morphological analysis,” In Proceedings of the 2004 conference on emprical methods in natural language processing, 2004.

[7] A. Mathes, “Folksonomies—cooperative classification and communication through shared metadata,” <http://adammathes.com/academic/computer-mediated-communication/folksonomies.html>, 2004.

[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” In Advances in neural information processing systems, pp. 3111–3119, 2013.

[9] G. Mishne, “Autotag: a collaborative approach to automated tag assignment for weblog posts,” In Proceedings of the 15th international conference on World Wide Web, pp. 953–954, 2006.

[10] T. Ohkura, Y. Kiyota, and H. Nakagawa, “Browsing system for weblog articles based on automated folksonomy,” In Proceedings of the WWW 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, at WWW, 2006.

[11] E. Quintarelli, “Folksonomies: power to the people,” Paper presented at the ISKO Italy-UniMIB meeting, <http://www.iskoi.org/doc/folksonomies.htm>, 2005.

[12] J. Trant, “Studying social tagging and folksonomy: A review and framework,” Journal of Digital Information, Vol. 10, No. 1, 2009.

[13] X. Wu, L. Zhang, and Y. Yu, “Exploring social annotations for the semantic web,” In Proceedings of the 15th international conference on World Wide Web, pp. 417–426, 2006.

[14] O. Zamir, and O. Etzioni, “Web document clustering: A feasibility demonstration,” In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 46–54, 1998.

[15] 加藤亮, 吉川大弘, and 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,” 人工知能学会全国大会論文集, Vol. 28, pp. 1–4, 2014.

[16] 佐藤敏紀, 橋本泰一, and 奥村学, “単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討,” 言語処理学会第 23 回年次大会発表論文集, 2017.

[17] 安川美智子, and 横尾英俊, “クエリログから獲得した関連語のクラスタリングに基づく Web 検索,” 電子情報通信学会論文誌 D, Vol. 90, No. 2, pp. 269–280, 2007.

[18] 総務省 | 平成 27 年版 情報通信白書 | 情報収集, 総務省, <http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h27/html/nc122310.html> (参照 2017-12-28).