

マイクロブログにおけるハッシュタグ自動付与システムの提案

野呂 勇太[†] 廣田 雅春^{††} 横山 昌平^{†††} 石川 博^{†††}

[†] 静岡大学情報学部 〒432-8011 静岡県浜松市中区城北 3-5-1

^{††} 静岡大学創造科学技術大学院 〒432-8011 静岡県浜松市中区城北 3-5-1

^{†††} 静岡大学情報学部 〒432-8011 静岡県浜松市中区城北 3-5-1

E-mail: [†]cs09065@s.inf.shizuoka.ac.jp, ^{††}dgs11538@s.inf.shizuoka.ac.jp ,

^{†††}{yokoyama,ishikawa}@inf.shizuoka.ac.jp

あらまし マイクロブログの代表的なサービスのひとつである Twitter は、ハッシュタグと呼ばれる機能がある。ツイートの検索機能のひとつとして、ハッシュタグによる検索があり、興味のある話題について情報収集する際に利用される。これは、ユーザにより投稿されたハッシュタグが付与されているツイートを検索する機能である。しかし、一部のツイートにのみハッシュタグが付与されているため、検索によって得られるツイートの数は限られる。ハッシュタグが付与されたツイートを増加させるため、ハッシュタグの利用を促進することを目的とする。そこで、本研究では、ユーザの入力したツイートに対して、ハッシュタグの意味と時間を考慮して適切なハッシュタグを推定し、ユーザに推薦するシステムを提案する。

キーワード ハッシュタグ推定, ハッシュタグ推薦, Twitter

Proposal of a System for Automatically Hashtagging in Microblog

Yuta NORO[†], Masaharu HIROTA^{††}, Shouhei YOKOYAMA^{†††}, and Hiroshi ISHIKAWA^{†††}

[†] Faculty of Informatics, Shizuoka University

3-5-1 Jouhoku, Naka-ku, Hamamatsu, Shizuoka, 432-8011 Japan

^{††} Graduate School of Science and Technology, Shizuoka University

3-5-1 Jouhoku, Naka-ku, Hamamatsu, Shizuoka, 432-8011 Japan

^{†††} Faculty of Informatics, Shizuoka University

3-5-1 Jouhoku, Naka-ku, Hamamatsu, Shizuoka, 432-8011 Japan

E-mail: [†]cs09065@s.inf.shizuoka.ac.jp, ^{††}dgs11538@s.inf.shizuoka.ac.jp ,

^{†††}{yokoyama,ishikawa}@inf.shizuoka.ac.jp

Abstract In twitter, we can use a hashtag as query to search tweets annotated with hashtag. However, we can not obtain tweets to be retrieved, since all tweets do not have a hashtag. Therefore, we promote that users annotate tweets with a hashtag to increase the number of tweets annotated with hashtags. In this paper, we propose a system for estimating a hashtag for tweets and recommend this to a user, considering meaning and time of hashtag.

Key words Hashtag Estimation, Hashtag Recommendation, Twitter

1. はじめに

近年, Twitter [1] に代表されるマイクロブログは, 急速に普及し, 多くの人々に利用されている。マイクロブログの記事は, 短いテキストであり, Twitter の場合それぞれのユーザがツイートと呼ばれる 140 字以内の短文を投稿する。そのため, 更新が容易であり, 即時性が高いことが特徴である。ユーザが投稿する記事の話題は, 日常生活に関するものやニュース, ユーザが視聴しているテレビ番組, 参加している行事など多岐に渡

る。従来のメディアと比較してこれらの記事は即時性が高く, マイクロブログの記事はユーザにとって有益な情報源となっている。また, Twitter のツイートの数は, 2012 年 11 月 16 日の発表において, 2.5 日間に 10 億件, 月間アクティブユーザ数は 2012 年 12 月 18 日の発表において, 2 億人を越えたと発表されている^(注1)。そのため, ユーザが大量のツイートの中から興味のあるツイートを探すことは困難である。

(注1) : <http://www.itmedia.co.jp/enterprise/articles/1212/20/news101.html>



図 1 #iPod の検索結果

表 1 ツイートに付与されるハッシュタグの数の割合

ハッシュタグ数	ツイート数	割合 (%)
1	545,984	79.53
2	75,178	10.95
3	37,654	5.48
4 and more	27,737	4.04

Twitter において、大量のツイートの中からユーザが求めるツイートを発見する主な手段として、ユーザのフォロー、キーワード検索、およびハッシュタグ検索が挙げられる。ユーザをフォローした場合、フォローしたユーザが発信した情報を得ることができる。ユーザは、興味のある話題について多くツイートしているユーザをフォローすることで、求める情報を得ている。キーワード検索は、入力したキーワードを含むツイートのタイムラインを取得できる。興味のある話題に関連するキーワードを入力することで、その話題に関連するツイートを取得できる。しかし、キーワードを含まないツイートである場合、話題に関連していても取得できない。そこで、ハッシュタグを用いた検索が利用される。ハッシュタグは、# 記号とそれに続く文字列で構成され、ツイートがどのような話題に関連するものであるかを表すために付与される場合が多い [2]。ハッシュタグを用いて検索を行うことで、そのハッシュタグが付与されたツイートを取得することができる。図 1 に、ハッシュタグ “#iPod” で Twitter のハッシュタグ検索を行った検索結果の一部を示す。ユーザがある話題に関連するツイートを検索する際、ハッシュタグを用いた検索がよく用いられている。Teevan ら [3] は、Twitter における検索クエリを収集し、全検索クエリ中の約 21% がハッシュタグを用いた検索であるということを示し、ハッシュタグを用いた検索はユーザがツイートを発見する一般的な手段として利用されていることを述べている。

しかし、ハッシュタグは、それぞれのユーザが自由に定義するため、表記揺れや、類義語による同一の概念を表すハッシュタグが複数存在する場合がある。そのような場合、ユーザは、ツイートを検索する際に複数のハッシュタグから適切なハッシュタグを選択しなければならない。また、一部のツイートのみ

ハッシュタグが付与されているため、そのようなハッシュタグのひとつを用いて、検索されるべきツイートの全てを検索することは困難である。ここで、2012 年 8 月 20 日に Twitter からランダムに収集した 13,550,415 件の日本語のツイートを用いて、ハッシュタグの利用状況の調査を行った。収集したツイートの 686,553 件は、ツイートにハッシュタグが付与されていた。これは、収集したツイートの約 5% である。また、ハッシュタグが付与されている 686,553 件のツイートについて、それぞれのツイートに付与されているハッシュタグの数とその割合を表 1 に示す。ツイートに付与されているハッシュタグが 1 つである場合が最も多く、約 79% である。そして、残りの約 21% のツイートは、複数のハッシュタグが付与されている。

ハッシュタグが付与されたツイートの割合を増加させることで、ハッシュタグを用いてツイートを検索した際に、より多くのツイートが検索結果として取得可能になると考えられる。そのため、本研究は、ツイートに対して、適切なハッシュタグを推定し、ユーザに推薦することで、ユーザにハッシュタグの利用を促すシステムを提案する。ユーザにハッシュタグの利用を推薦することにより、ハッシュタグが付与されたツイートが増加することに加えて、システムが推薦するハッシュタグを統制することによって、ハッシュタグの表記揺れなどを防ぐ事が可能になると考えられる。

また、単一のハッシュタグのみでは、ツイートの内容を十分に表せないような場合がある。そのため、ツイートのハッシュタグを推定する際に、そのツイートに複数の適切なハッシュタグを推定する必要があると考えられる。

本論文の構成を以下に示す。2 章では、関連研究について述べる。3 章では、ユーザのツイートに対してハッシュタグを推定し、ユーザにその結果を推薦する手法と、それを用いたシステムについて述べる。4 章では、システムの実用例と実行結果の考察について述べる。最後に 5 章では、本論文のまとめと今後の課題について述べる。

2. 関連研究

ハッシュタグは、ユーザが自由に定義することができるため、Folksonomy の一種と考えられる。Folksonomy は、多くのユーザがタグ付けを行うことによって分類がなされる。丹波らは、従来の Taxonomy による分類と比較し、Folksonomy による分類は、多くのユーザが分類を行うため分類速度が速い、多くのユーザにとって分類が実用的である、時代のパラダイムやユーザの意識に適応した分類構造が自動的になされるといった特長を挙げている [4]。これらの特長は Twitter のハッシュタグも同様に持ち合わせていると考えられる。一方、Twitter のハッシュタグに関しても、Folksonomy と同様に同義語、タグの抽象度に関する課題 [5] がある。

タグの自動付与に関する研究として、藤村らの研究 [6] と廣田らの研究 [7] がある。藤村らは、ブログ記事に対して、複数の適切なタグを推定し、自動的に複数のタグを付与する手法を提案した。廣田らは、Web 上の画像に対して適切なタグを推定する手法を提案した。これらの研究では、適切なタグを推定

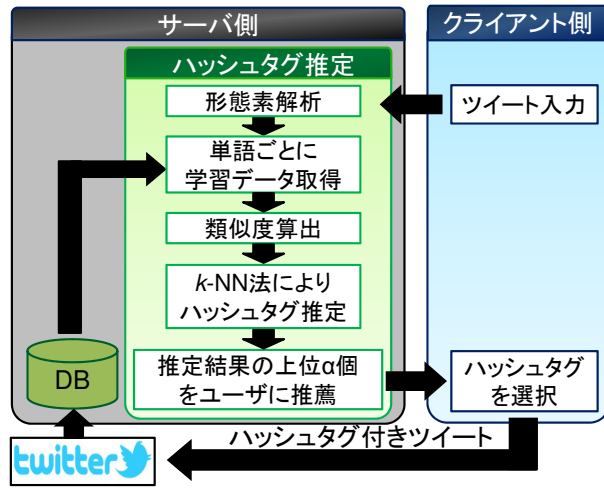


図 2 システム構成図

する手法として、 k -NN 法を用いている。本研究では、Twitter に投稿される記事に対して、適切なタグを推定する。ブログなどと比較して、Twitter は、リアルタイム性が高く、学習データに含まれないような新しいハッシュタグが生成される可能性があると考えられる。そのため、本研究では、学習データとして利用するハッシュタグが付与されているツイートを定期的に取得する。

ハッシュタグの推薦に関する研究として、Zangerle らの研究 [8] がある。Zangerle らは、Twitter に投稿されている、ハッシュタグが付与されたツイートを学習データとして用いて、ユーザが入力したツイートに対して適切なハッシュタグを推薦する手法を提案した。学習データとユーザが入力したツイートの類似度によって、学習データ中のツイートをランク付けし、上位 k 個のツイートに付与されたハッシュタグを推薦している。実験により、上位 5 個のツイートに付与されたハッシュタグを推薦することで最も良い結果が得られた。

マイクロブログの記事検索に関して、ユーザの入力したクエリを拡張することで、ツイートの検索精度を向上させる研究が行われている。Efron らは、トピックに関するマイクロブログ記事の検索システムについて、ユーザが入力したトピックのキーワードに対して、トピックに関連するハッシュタグのリストをユーザに提示するという課題について述べている。[2] また、トピックに関連するハッシュタグをランク付けしたリストを作成してフィードバックすることで、ハッシュタグのリストによるクエリの拡張に取り組んでいる。本研究では、マイクロブログの検索において、クエリの拡張を行うのではなく、ユーザにハッシュタグの利用を促進し、ツイートに付与されるハッシュタグを増加させることで、ツイートの検索精度の向上を目指している。また、本研究を適用する事で、ユーザの投稿するハッシュタグを統制し、クエリの拡張を適用することで、より検索精度を向上させることが可能になると考えられる。

3. 提案システム

本システムは、Twitter から収集したハッシュタグが付与さ

れたツイートを用いて、ツイートに対する適切なハッシュタグを推定し、その推薦結果をユーザに推薦する。そして、推薦されたハッシュタグの中からユーザが選択したハッシュタグをツイートに付与する。本論文で提案するシステムの構成図を図 2 に示す。

ハッシュタグの推定を行うため、まずハッシュタグが付与されたツイートの取得を行い、取得したツイートに対して形態素解析などを行うことで学習データを生成する。そして、ユーザが入力したツイートに対して、学習データを用いてハッシュタグの推定を行う。

3.1 ツイートデータの取得

Twitter から、ハッシュタグが付与されたツイートを取得する。ツイートの取得には、Twitter API を用いる。取得したそれぞれのツイートから URL や記号などを削除し、形態素解析を行う。形態素解析には、Kuromoji [9] を用い、形態素解析辞書として、UniDic [10] を用いる。それぞれのツイートの形態素解析の結果から名詞・動詞・形容詞と判定された単語のみを抽出し、それぞれのツイートの単語集合とする。形態素解析によって、1 つの熟語が複数の単語として抽出される場合があるため、隣接する名詞の連結を行った。名詞の連結ルールとして、前の形態素が「普通名詞」もしくは「固有名詞」であり、「副詞可能」でない、かつ「形状詞可能」でないとき、後ろの形態素が「普通名詞」もしくは「固有名詞」であり、「副詞可能」でない、かつ「形状詞可能」でない、かつ「サ変可能」でない場合に連結を行った。形状詞は、形容動詞の語幹を表す。生成した単語集合と、ツイートに含まれているハッシュタグをデータベースに格納する。このデータベースに格納されているデータを学習データとする。また、定期的にツイートを取得し、データベースに追加し、学習データを増加させる。

3.2 ハッシュタグの推定

ユーザが入力したツイートに対して適切なハッシュタグを、学習データに基づいて推定する。3.1 節と同様の手順で、ユーザが入力したツイートに含まれる URL、記号を削除し、形態素解析を行う。形態素解析の結果から名詞・動詞・形容詞を抽出し、名詞の連結を行い、ユーザが入力したツイートの単語集合を生成する。ユーザが入力したツイートと学習データのそれぞれのツイート間の類似度を算出するために、それぞれのツイートの単語特微量として、IDF を利用する。全ツイートに含まれるすべての単語 $W = \{w_1, w_2, \dots, w_N\}$ のある単語 w_i の IDF を以下のように求める。

$$idf_{w_i} = \log \frac{D}{n_{w_i} + 1} \quad (1)$$

ここで、 w_i は W に含まれる i 番目の単語、 D は全ツイート数、 n_{w_i} は単語 w_i が含まれるツイート数とする。そして、IDF に基づいて、ツイート a の特徴量 $\vec{a} = \{t_{1a}, t_{2a}, \dots, t_{Na}\}$ を以下の式を用いて求める。

$$t_{ia} = \begin{cases} idf_{w_i} & w_i \in W_a \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

ここで t_{ia} は、ツイート a における単語 w_i の特徴量、 W_a は、

ツイート a に含まれる単語の集合である。そして、それぞれのツイート a と b の類似度をコサイン類似度に基づいて、以下のよう求める。

$$\cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| |\vec{b}|} \quad (3)$$

ハッシュタグの推定には、コサイン類似度に基づいた k -NN 法を用いる。 k -NN 法では、ユーザが入力したツイートに類似する k 個のツイートを学習データから選出する。提案システムでは、まず、ユーザが入力したツイートから抽出したそれぞれの単語を含んでいるツイートを学習データから取得する。そして、それぞれの単語について取得したツイート集合に対して、 k -NN によりそれぞれ k 個のツイートを選出する。選出された k 個のツイートの中で、それぞれのハッシュタグが付与されたツイートの数に基づいて順位づけし、上位 α 個のハッシュタグを推定結果とする。入力されたツイートから抽出したそれぞれの単語について、同様にハッシュタグの推定を行い、その結果をユーザに推薦する。

4. 本システムの実行結果

Twitter から収集した 2012 年 8 月 20 日中のハッシュタグが付与されたツイート、262,804 件を学習データとして用いてハッシュタグの推定を行う。また、 k -NN の k を 100 とし、順位付けたハッシュタグを最大で上位 10 個、ユーザに推薦する。

4.1 実行例

“今日は巨人と阪神の試合だ！楽しみ”というツイートを入力した結果を図 3 に示す。赤色の点線で囲われた部分は、ツイートの入力エリアとなっており、緑色の点線で囲われた部分は、各単語に対してハッシュタグを推定した結果である。推定結果として表示されたハッシュタグを選択すると、入力したツイートに選択されたハッシュタグが付与される。選択状態を解除すると、付与されたハッシュタグはツイートから除去される。図 3 の結果から、“巨人”や“阪神”、“試合”という具体的な内容を示す単語に対しては、“#giants”や“#kyojin”、“#hanshin”というツイートの内容に適したハッシュタグやツイートの話題に関連するハッシュタグが結果として得られた。一方、“今日”や“楽しみ”というどのような話題のツイートにも出現する可能性がある一般的な単語に対しては、“#onitama795”や“#cmsn”というラジオ番組やカミスンという TV 番組のハッシュタグなど野球に直接的な関連が見受けられないハッシュタグが結果に含まれた。

図 4 は、“[アメブロ更新]お楽しみに(*^o^*) http://t.co/P6A5Nigm”というツイートを入力して実行した結果となっている。“更新”という単語に対する推定の結果として、“#ametwi”というアメブロに関連するハッシュタグが得られた。しかし、その他の入力したツイートの内容に適するハッシュタグは推定の結果として得られなかった。

4.2 考察

“今日”や“楽しみ”という一般的な単語に対して、良い推定結果が得られなかった。ハッシュタグ推定の処理において、一般的な単語を含むツイートを学習データから取得した後、さ



図 3 野球に関連するツイートを入力した実行結果

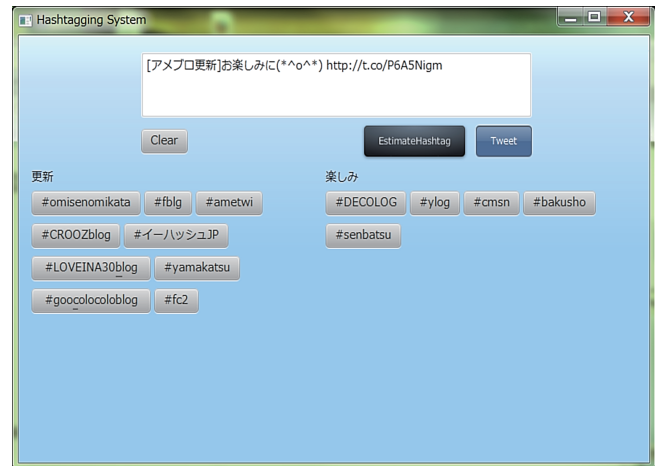


図 4 ブログに関連するツイートを入力した実行結果

らに学習データを入力したツイートに含まれる具体的な単語によって絞り込むことで推定の精度向上を目指す。また、Twitter において、ユーザが注目するような出来事が発生すると、その話題に関連したツイートが急激に増加する場合があります。この状態をバースト状態と呼ぶ。推薦時に、バースト状態にある話題のツイートをを用いて、それらのツイートに付与されているハッシュタグを推定結果の上位にする手法を検討することで、そのような話題に関連するユーザのツイートに対して、適切なハッシュタグを推薦できると考えられる。

5. おわりに

本論文では、ユーザのハッシュタグの利用を促進することでハッシュタグが付与されたツイートを増加させるため、ユーザが入力したツイートに対して適切なハッシュタグを推定し、推定したハッシュタグの中でユーザが選択したハッシュタグをツイートに自動で付与するシステムを提案した。また、ユーザが入力したツイートに類似するツイートからハッシュタグを推定する手法を提案した。

今後の課題として、時間を考慮したハッシュタグ推定手法として、4.2 節で述べた、バーストを考慮してハッシュタグを推定する手法を検討する

文 献

- [1] Twitter. <https://twitter.com/>.
- [2] M. Efron. Hashtag retrieval in a microblogging environment. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788. ACM, 2010.
- [3] J. Teevan, D. Ramage, and M.R. Morris. #twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 35–44. ACM, 2011.
- [4] 丹羽智史, 土肥拓生, 本位田真一. Folksonomy の 3 部グラフ構造を利用したタグクラスタリング. 人工知能学会. セマンティックウェブとオントロジー研究会 (第 14 回), SIG-SWO-A602, 2006.
- [5] S.A. Golder and B.A. Huberman. The structure of collaborative tagging systems. 2009.
- [6] 藤村 滋, 藤村 考, 奥田 英範. ブログ記事への自動マルチタグ付与. The 21st Annual Conference of the Japanese Society for Artificial Intelligence, 2007.
- [7] M. Hirota, N. Fukuta, S. Yokoyama, and H. Ishikawa. A robust clustering method for missing metadata in image search results. *Journal of Information Processing*, 20(3), 2012.
- [8] E. Zangerle, W. Gassler, and G. Specht. Recommending #tags in twitter. In *Proceedings of CEUR Workshop*, volume 730, pages 67–78, 2011.
- [9] Kuromoji. <http://www.atilika.org/>.
- [10] UniDic. <http://www.tokuteicorpus.jp/dist/>.