Twitterにおける被検索ユーザの分類による ワードスコープの同定に関する基礎検討

山西 良典[†,a]　　奥 健太[†,b]

† 立命館大学情報理工学部

a) *ryama@media.ritsumei.ac.jp*　b) *oku@fc.ritsumei.ac.jp*

概要　代表的なソーシャルメディアである Twitter は，評判情報の取得源として有効的に活用されている．しかしながら，任意のクエリについての tweet を検索機能によって獲得すると，検索者の意図しない tweet が多く得られることがある．これは，クエリの焦点（ワードスコープ）がユーザ毎に異なることが原因と考えられる．本稿では，被検索ユーザの分類によるワードスコープの同定について基礎的検討を行う．分類されたユーザ情報を参照することで，ワードスコープの同定のみならず，クエリ拡張に応用可能な関連情報も取得可能になると考える．

キーワード　Twitter，ワードスコープの同定，ユーザの分類，クエリ拡張

1　はじめに

Twitter に代表される Social Network Service（SNS）は，評判情報の取得源として近年注目を集めている．スマートフォンなどによって場所を問わず容易に自分の経験や意見を投稿可能なことから若年層を中心として多くのユーザがおり，一般大多数の意見取得のための Web サービスとして大いに利用されている．例えば，位置情報付きの tweet を用いた観光スポットの推薦 [1] や，tweet 中の単語の極性を参照した株価の予測 [2] などが報告されている．このような Twitter を情報源とする研究の多くは，Twitter の検索機能を用いて目的のクエリが含まれる tweet を取得している．

Twitter には 140 文字という字数制限が設けられている．Twitter の字数制限はユーザの直感的な感想の投稿を演出する一方で，単文や一言のみの tweet が多くなり，tweet 中の文脈が省略されてしまう問題がある．例えば，「プロ演レポートつらい」という tweet からは，tweet 中の「プロ演」が「“立命館大学”のプロ演」なのか「“明治大学”のプロ演」なのかを同定することは不可能である．

本稿では，クエリ検索によって得られた tweet 中の単語の焦点（**ワードスコープ**）の同定を図る手法を検討する．語義曖昧性の解消を試みた研究は多く報告されているが，単語の焦点を対象とした研究は少ない．また，将来的にクエリを tweet する可能性が高い潜在的なユーザを発見する手法についても議論する．

2　本研究のねらい

本研究のねらいは大きく，1）tweet 中でのクエリのワードスコープの同定，2）クエリを tweet する可能性が高いユーザを取得するための二次クエリの生成，の 2 点として整理できる．

2.1　tweet 中でのクエリのワードスコープの同定

人間が tweet の閲覧時にワードスコープを同定する方法を考えてみる．例えば，上述の「プロ演レポートつらい」という例では，プロフィール情報に「立命館」と記述されていたり，過去の tweet に「南草津」や「BKC」などが存在すれば，「“立命館大学”のプロ演」であると判断する．つまり，プロフィール情報や過去の tweet 履歴などの情報を参照してユーザを分類し，ワードスコープの同定を図っている．

本稿では，検索によって得られた情報のみではなく情報の発信源である被検索ユーザを分類する．被検索ユーザを分類可能となれば，分類された被検索ユーザ集合でのクエリのワードスコープが推定可能になると考えられる．これにより，文脈が読み取れない tweet 中のワードスコープの同定を試みる．

2.2　クエリを tweet する可能性が高いユーザを取得するための二次クエリの生成

多くの意見・評判情報を獲得するために，クエリ拡張に関する研究が数多く報告されている．クエリ拡張は，検索ログによる拡張，クエリ間の類似性による拡張，クエリの多様化に大別される．本研究で目指す二次クエリの生成は，このうちクエリの多様化 [3] に位置付けられる．

2.1 節において述べたユーザの分類において高い寄与率を示した単語は，ワードスコープ毎のユーザ集合において特徴的な単語とみなすことができる．これらの単語は，クエリについて tweet する可能性が高い潜在的なユーザを獲得するための二次クエリとして転用可能であると考えられる．

表 1　ユーザが各潜在的グループ（LG）である確率 (%) の例.

ユーザ ID	LG1	LG2	LG3	LG4
36	7	0	93	0
40	0	100	0	0
46	1	7	0	92

3　被検索ユーザの分類によるワードスコープの同定

被検索ユーザの分類によるワードスコープの同定について検討する．本稿では，1) クエリ検索によって得られたユーザの「ユーザの過去の tweet」などの情報を獲得，2）ユーザごとに，得られた情報に出現する自立語の出現頻度を算出，3）各自立語の出現頻度をユーザを行，自立語を列とした行列として表現する，4）上記で得られた行列に対して，非負値行列因子分解（NMF）[4] を適用，の手順での被検索ユーザの分類について議論する．

3.1　NMF によるユーザ集合の分類と特徴語抽出

クエリ検索で得られたユーザ集合の潜在的なユーザ集合への分類には，NMF を適用する．NMF では，与えられた $I \times J$ サイズの非負値行列を，$I \times K$ の非負値行列 W と $K \times J$ の非負値行列 F の積として近似する．ここで，I は上記手順 1）におけるユーザ数，J は全ての被検索ユーザの過去の tweet に出現した自立語の総異なり数となる．また，K は潜在的な因子数を示し，本研究においてはユーザの潜在的なグループ（LG）の数を指す．

NMF によって得られる 2 つ行列 W と F からはそれぞれ，ユーザが各 LG である確率と各 LG を構成する上での自立語の寄与率を見ることが可能となる．このとき，各 LG 構成において高い寄与率を示す自立語が，二次クエリとして応用可能でないかと考える．

3.2　デモンストレーション

「遅延」を検索クエリとして，47 名の被検索ユーザを得た．この被検索ユーザのそれぞれの過去の tweet を取得し，総異なり数 1294 の自立語を得た．この 47×1294 の非負値行列に対して，上述の手順に従ってユーザの分類を行った．ここで，因子数 $K = 4$，行列の最適化のための更新回数は 150,000 回とした．

表 1 に，分解された行列 W から算出されるユーザが各 LG である確率を示す．表中の各ユーザの「遅延」が含まれた tweet はそれぞれ，ユーザ 36 は「岐阜ベンチャーサミットに向かうために何年ぶりかで東海道本線に乗りに来たけど、事故かなにかでダイヤ遅延中。」，ユーザ 40 は「濃霧のため磐越西線が 10 分弱遅延。郡山はかなり涼しい。というか寒い (@ ＪＲ郡山駅 13 番線ホーム)」，ユーザ 46 は「頭良いの頑張れ w 遅延証明書って駅でもらえるの？」であった．ユーザ 36 と 40 では遅延の対象としている路線が異なることがわかる．これは，過去の tweet 内容を考慮した上でユーザ分類が実現され，結果的に異なる路線の「遅延」情報を分類できたと考えられる．

つぎに，分解された行列 F を参照することで，各 LG を構成する上で高い寄与率を示した自立語について考察する．一例として，LG2 の寄与率の上位の単語を見てみると，「福島県」「宇都宮市」「郡山市」「郡山駅」「宇都宮駅」「会津若松駅」などの単語が存在した．これらの単語を二次クエリとして用いることで，LG2 と同様のスコープで「遅延」を tweet する可能性が高いユーザを事前に検索できる可能性が期待される．本デモンストレーションにおいて，福島県近辺の電車の遅延情報を想定して「遅延」をクエリとして検索したとすれば，「郡山駅」「宇都宮駅」「会津若松駅」などを含む tweet をしているユーザを獲得しておくことで，該当地域の電車の「遅延」について tweet する可能性が高いユーザを事前に察知可能になると考える．

4　おわりに

本稿では，Twitter におけるワードスコープの同定を被検索ユーザの分類によって実現するための手法について基礎的な検討を行った．また，被検索ユーザの分類によって得られる二次クエリへの転用が高い単語について考察を行った．

今後は，NMF における初期値依存の問題や因子数の動的な決定手法についても議論しつつ，より多くの実験考察を経て，ワードスコープの同定および二次クエリ生成手法の確立を目指す．

謝辞

本稿の執筆にあたり，立命館大学情報理工学部，福本淳一教授の助言を得た．記して謝意を表す．

参考文献

[1] 奥健太，橋本拓也，上野弘毅，服部文夫，“位置情報付きツイート対応付けに基づく観光スポット推薦システムの開発,” Web インテリジェンスとインタラクション研究会，no.2，pp.7–12，2013.

[2] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” Journal of Computational Science, vol.2, no.1, pp.1-8, 2011.

[3] Q. Mei, D. Zhou, and K. Church, “Query suggestion using hitting time,” Proc. of the 17th ACM conference on Information and knowledge management, pp.469–478, 2008.

[4] S. Sra, and I.S. Dhillon, “Generalized nonnegative matrix approximations with bregman divergences,” in Advances in Neural Information Processing Systems 18, eds. Y. Weiss, B. Schölkopf, and J. Platt, pp.283–290, MIT Press, 2006.