マイクロブログマイニングの現在

奥村 学
東京工業大学精密工学研究所

2012 年 1 月

1 はじめに

Web2.0，集合知，CGM などという用語が用いられるようになった数年前以後も，現在ではソーシャルメディアと称されるようになったメディアが WWW 上には数多く新しく出現している．ブログ (blog; Weblog) は急速に普及し，もはや十分定着した感があるメディアの 1 つと言える．一過性のものだろうという予測も当初はあったが，現在でも依然注目を集めている Web 上の情報源の 1 つであることは間違いないであろう．

このブログを凌駕する勢いで，近年急速に普及し，注目されている情報源としてマイクロブログ (microblog) がある．ブログよりもさらに，速報性，リアルタイム性のある，まさに「今」現在の実世界を表現する新鮮な情報が発信されることから，ブログ同様有用な情報源と考えられるようになってきている．

本稿では, マイクロブログから情報を抽出，発掘する，いくつかのマイクロブログマイニング技術について，ブログマイニング技術と対照しながら，概説する．

2 マイクロブログとは?

マイクロブログを提供する Web 上のサービスの代表は，2006 年に開始された Twitter であり，この Twitter の出現により，マイクロブログは急速に普及したと言っても間違いではないであろう．以後，Twitter を元に，マイクロブログの機能を概説する．

マイクロブログは，tweet と呼ばれる，140 文字以内という長さの制約のある要素の集合で構成されている．マイクロブログのユーザは，他のユーザのマイクロブログを follow することで購読する．ユーザは，他のユーザの tweet を自分の tweet 内で retweet [1] することもできるし, 自分の tweet に hashtag (‘# ’記号を先頭に伴う) と呼ばれるタグを付与することもできる．また，他のユーザを参照 (mention)(‘@ ’記号を先頭に伴う) することで, そのユーザへ reply することもできる [2] .

ブログと同様，マイクロブログの書き手の多くが一般の個人であり, その内容から一般の人々が何をし, 何を思っているかを抽出できる可能性があることが, マイクロブログを情報源として魅力的にしていると言える. さらに, ブログと同様，マイクロブログの場合, follow などの付加的な機能により, 人と人の間のつながりに関する情報を入手し易いという特徴もあり, follow 関係を元にしたコミュニティ抽出などのように, 個人間のつながりに関する分析に向いているという特徴もある [3].

[1] E-mail の forward に対応する．

[2] Honeycutt と Herring[21] は mention の使われ方を，danah boyd ら [14] は retweet の使われ方を，それぞれ分析している．

[3] この特徴は, マイクロブログの SNS(Social Network Service) としての特徴とも言える．

マイクロブログでは，ブログよりもさらに，速報性，リアルタイム性のある，まさに「今」現在の実世界に関する情報が発信されており，マスメディアよりも時には早く情報が伝達され，それを後にマスメディアが取り上げるということも起きているということが，マイクロブログを情報源として魅力的にしている第二の理由と言える．

マイクロブログでは，この他，mention 機能を用いて，チャットのように，複数の人間で非同期で対話的なコミュニケーションが行えるという特徴もある．また，140 文字以内という長さの制約等により，これまでのテキストとはかなり異なった言語使用が行われており，そのため，後述するように，マイクロブログに特化した言語処理技術の開発が不可欠であるという要請も生まれることになる．

3 マイクロブログマイニングの概要

マイクロブログを対象にした分析にはどのようなものが考えられるだろうか. この問いに答える前にまず，類似するコンテンツであるブログを対象としたマイニング技術を振り返ってみることにしよう．[58, 59] では，ブログマイニングと呼ぶことができる分析技術として以下のようなものを挙げている．

- Authority 分析 (被リンク数によるランキング)
- トレンド分析
- 評判分析
- コミュニティ抽出
- ブログの書き手の属性推定
- 実世界の動向 (たとえば, 株価, 売上) との相関分析
- スパムフィルタリング
- 自動要約
- 情報の重要性, 信頼性評価
- ブログエントリの自動分類, トピック同定
- マスメディア (たとえば, 新聞記事) とブログの自動対応づけ

マイクロブログでも同様に，以下のような分析技術が研究開発されてきており，次節以後この順序で概説する．

- Authority 分析 (follower 数によるランキング)
- 評判分析
- 実世界の動向 (たとえば, 株価, 売上) の予測
- マイクロブログの書き手の属性推定
- マイクロブログのトピック同定

- トレンド分析
- 自動要約
- 情報の信頼性評価

マイクロブログは，前節で述べたように，ブログとは異なる 2 つの特徴を有する:

- まさに「今」現在の実世界に関する情報を発信している，
- 長さの制約等により，これまでのテキストとはかなり異なった言語使用が行われている．

このため，以下のようなマイクロブログに特徴的な分析技術開発も行われることになる．

- social sensor としてのマイクロブログ分析技術
  「今」実世界で起きている出来事を検出し，速報できれば有用であることから，(今起きている) 出来事を検出するイベント検出や，緊急時のコミュニケーション手段としてのマイクロブログ活用を前提とした分析技術，(病気などの) 流行の予測を目指す研究開発が行われている．
- tweets 用のテキスト処理ツール
  単語への品詞付与 (Part-of-Speech Tagging)，固有名抽出 (Named Entity Recognition)，意味役割付与 (Semantic Role Labeling) など，これまで研究開発されてきている言語処理技術の多くがそのままでは tweets に対しては性能劣化が激しく適用不可能であるため，tweets 用の技術開発が必須となっている．

4 Authority, Influencer 分析

マイクロブログにおける検索では, 従来の検索システムと同様に, tweets を検索対象として検索し, ランキングして表示できるのは言うまでもないが, 個人 (マイクロブログの書き手) なども, ランキングの対象となりうる．Tweets を検索対象とした場合, 従来の検索システムと同様に, 検索要求との照合の度合 (relevance) の順にランキングできるのは言うまでもないが, マイクロブログならではのランキングとして, Twitter では時間順，retweet 数を用いたランキングが行われている．

マイクロブログのユーザは, Twitter では follower 数に基づいてランキングされている[4]．いずれの考え方でも，それぞれユーザがどの程度参照されているかが指標化される. PageRank, HITS アルゴリズム等で有名になったように, 多くのユーザから参照される対象は, Authority と考えられ, 重要度を高く見積もるわけである．

Authority，influencer の同定は，ランキング以外に，(商品に関する情報を優先的に配信するなどのように) マーケティングに利用することが想定されている．

Duan ら [17] では，tweets のランキングに関して，ランキング学習を用いている． 素性としては，内容の適合性，アカウントの authority 度 (follower 数，list された回数等を元にしている)，URL リンクを含むかどうか，tweet の長さを用いており，authority 度， URL リンクを含むかどうか，tweet の長さの 3 つの素性の組み合わせが最良だったと報告している．

Weng ら [54] は，影響力のある Twitter ユーザの検出のために PageRank の拡張である TwitterRank を提案している．ユーザのトピックをまず同定し，トピックごとにユーザのネットワークを構築し．そのネットワークに対してランキングアルゴリズムを適用する．トピックの同定には LDA(Latent Dirichlet Allocation) を用いている．

[4]retweet，reply, mention の割合を利用するという考え方も存在する．

後述するように，tweet 中に書かれている評判を集約することで，株価の上昇を予測する研究がこれまで行われてきている．各 tweet を肯定的/否定的/中立的 (positive/negative/neutral) に分類し，tweets 集合に対する分類結果を元に，投資の意思決定を行う (たとえば，最も肯定的な株式を購入し，最も否定的な株式を売却する)．しかし，すべてのユーザの tweet が同様に信頼できるわけではないのは明らかである．Bar-Haim ら [3] は，専門家を非専門家と区別し，その情報を予測に利用する手法を提案している．

Bar-Haim らは，tweet の内容と株価の関係を直接学習し，株価の上昇/下降を tweet の内容を元に判別する分類器を訓練する．この判別を正確に行える tweet(株価の変化を的確に予測していると言える) を書いているユーザを専門家として同定する．そして，同定された専門家の tweets 集合のみで株価との関係を学習した分類器を作成する．

5 評判分析

ブログ同様，マイクロブログが個人の発信するメディアであることから, 現在評判分析がマイクロブログマイニングで最も関心を持たれている技術の 1 つと言って良いだろう. 評判分析では, tweet 中の特定の対象への評判を元に，tweet を肯定的/否定的/中立的に分類することが目的となる．

教師あり学習を元にした評判分類器では，ラベルつきの訓練データをどのように入手するかが問題になる．そこで，Davidov ら [15] は，特定の hashtag や顔文字集合をラベルと対応付けて利用することで，人手で訓練データにラベル付けする手間を省いている．Barbosa と Feng[4] も，tweets を対象にした既存の 3 つの評判分類器の出力をラベルとして利用することで，人手で訓練データにラベル付けする手間を省けることを示している．

同様に，教師あり学習を元にした評判分類器では，tweets 集合が時系列データであり，時々刻々内容が変化するため，訓練データ自体も最新のものに更新していく必要がある．そこで，Silva ら [44] は，少量の訓練データからの self training により評判分類器を学習する手法を示している．

Tweet 中の特定の対象への評判を元に，tweet を分類するタスクには通常，レビュー等を対象に開発された評判分類器は適用が困難である．対象に依存した評判を評判分類器が捉えられないからである．そこで，Jiang ら [22] は，評判が対象に対するものかどうかを，対象と tweet 中で構文的に関係している単語を素性として判定する手法を示している．Tweet は短く曖昧なので，周辺の関連する tweets の情報も考慮している．同一人物の tweets，reply-to 関係にある tweets，retweet している tweets 等が利用可能であるとしている．

音声における韻律情報と同様に，テキストにおいては，capitalization，アスタリスクで単語を囲む，単語の長音化等が重要語の強調等の役割を果たしていると考えられる．Brody と Diakopoulos [8] は，単語の長音化が極性 (positive/negative) を含む単語の検出に寄与することを示している．

個々の tweet ではなく，hashtag についてその評判を分類する研究も存在する．Wang ら [52] は，その hashtag を含む tweets 集合が分類に有用なのは言うまでもないが，hashtag 間の共起関係や hashtag の (字義通りの) 意味が有用であることを示している．最初の 2 つの情報を利用するため，グラフに基づく評判分類モデルを提案している．

6 実世界の動向の予測

マイクロブログマイニングの結果, たとえば，トレンドや評判の推移 (時間変化) に関する情報が得られるようになると, 次は, このマイクロブログ中での動向が, 実世界での動向とどのように相関

があるのかを分析したいという関心も当然高まってくる．この典型例が, マイクロブログの中での記述が選挙結果とどのように関連したかを分析するものと言える. 同様に, 株価や商品の売上の推移がマイクロブログ中の記述とどのような相関にあるかを分析するという研究も当然ありうる.

Bollen ら [7] は，tweet の mood(感情) を，POMS(Profile of Mood States) を元にした 6 つの mood(tension, depression, anger, vigor, fatigue, confusion) について分析した結果と，株式市場，原油市況，主要な出来事との関係を調査している．そして，様々な出来事が感情の傾向に影響を与えていることを明らかにしている．

Tumasjan ら [49] は， LIWC(Linguistic Inquiry and Word Count) 分析ツールを用いて，政党や政治家を参照している tweets を分析し，参照している tweet 数が選挙結果に反映していることを報告している．また，O ’Connor ら [31] は，tweets 中の評判が世論調査と相関することを示している．

Diakopoulos と Shamma[16] は，テレビ番組とマイクロブログを結びつける試みについて報告している．Twitter 中の評判を元に集約した rating を表示するシステムを示している．

Asur と Huberman[2] は，tweets が映画の興行収入を予測するのに利用できることを示している．(容易に想像できることだが，) よく語られている映画はよく見られている，評判情報は予測性能の向上に寄与することを報告している．

7　マイクロブログの書き手の属性推定

マイクロブログにおける分析で，たとえば，性別, 年齢等の書き手の属性がわかれば,「男性の中で話題になっているトピック」とか,「20 代の女性に好評のレストラン」といったように, 他の分析技術との組合せにより, 属性による分類ごとに分析結果を示すことができるようになり, より深みのある分析を実現できる. また, それ以外にも, たとえば, 居住地域がわかれば, その居住地域に関して記述されている内容は,「地元」の人の記述として, 遠方の人のものより信頼できると見なせる. また，マーケティングにおいて，特定の宣伝を特定のユーザに配信する (ある都市に住んでいる人にだけ宣伝を配信する) 際などにも，書き手の属性の情報は利用可能である [5].

書き手のデモグラフィックな属性としては, 性別, 年齢, 居住地域等が, 現在推定の対象となっている. 性別, 年齢の推定に関する研究はいずれも, テキストのスタイルの情報を利用するなどして, いわゆるテキスト分類の問題として解いている (テキストの書き手の性別, 年齢をいくつかのクラスに分け，そのうちどれであるかを推定する).

Cheng ら [11] は，tweet 中の各単語の位置との相関 (条件付き確率) を元にした確率モデルを用いて，市レベルでユーザの位置を推定する手法を提案している．Cheng らによれば，Twitter ユーザのうち 26%しか市レベルの情報を発信しておらず，また，0.42%の tweets にしか位置情報が付与されていない．

Eisenstein ら [18] も，ユーザの地理的位置を推定する手法を提案している．Eisenstein らも，特定の位置との結びつきの強い単語が存在するというアイデアを元にしているが，潜在トピックと地理的地域を一緒に推論する多レベルの生成モデルを用いた手法を提案している．

Rao ら [39] は，性別，年齢，宗教，政治的指向の 4 つの属性を分類する研究である．単語 n-gram, follower 数，retweet の頻度などを素性とした教師あり学習による分類器を用いている．

Burger ら [9] は，性別の推定を行う研究である．Tweet の量，profile を記述したメタデータの有無が性能に影響することを報告している．単語/文字 n-gram を素性とした教師あり学習による分類器を用いている．

[5]Twitter の場合，ユーザのプロファイルが存在するが，不完全だったり不確かだったりすることから，推定が必要になる．

Pennacchiotti と Popescu[35] は，政治的指向，民族，特定のビジネスへの親近感を推定する手法を示している．Tweets の情報と，ユーザ間のリンクで構成されたグラフ中のクラスラベルの分布の情報を利用した分類モデルを提案している．LDA を元にした Topic model を Tweets の情報によるユーザ分類に利用している．

8　マイクロブログのトピック同定

マイクロブログのユーザに，記述している内容に応じて，タグを付与したり，そのトピック (ユーザの関心) を同定したりする研究も近年数を増している．内容に応じたタグが付与できれば，それを利用した検索，推薦が可能になるという利点がある．また，ユーザの関心はそれ自体，前述したユーザの属性の一種とも言うことができるので，ターゲット広告での利用も可能であるし，また，前述したように，ユーザの属性推定にも利用できる．

Pennacchiotti と Gurumurthy[34] は，LDA を用いて，ユーザをトピックの混合物として表現し，類似のユーザを提示する推薦の手法を示している．類似度は，KL(Kullback Leibler) divergence やコサイン類似度を用いて計算できる．

Ramage ら [38] は，LDA を，tweet についていることのあるラベル (hashtag など) を教師情報として利用できるように拡張した Labeled LDA を用いることで，推薦の性能が向上することを示している．Labeled LDA では，tweets 集合は，ラベルと潜在トピックの混合物としてモデル化される．

通常 LDA を tweets 集合に対して適用する場合，1tweet を 1 文書と考えるか，Author-topic model[46] の考えを元に，1 ユーザの全 tweets を 1 文書と考えるかしている．Zhao ら [57] は，1tweet は 1 トピックについてであるという仮説を元に Twitter-LDA モデルを提案し，前者 2 つと比べて優れていることを示している．

Wu ら [55] は，単語に対する tfidf 重みと，単語をノードとし，同一 tweet 中での共起を元にしたリンクで構成されるグラフ上で TextRank を適用した結果を用いて，ユーザをタグ付けるキーワードを tweet から (教師なしで) 抽出する手法を示している． Zhao ら [56] も，Zhao らの提案している Twitter-LDA を適用し，トピック集合を発見した後，発見したトピックごとに，topical PageRank を適用してキーワードを抽出する手法を提案している．

First Story Detection (FSD) は，TDT(topic detection and tracking) [1] の 1 サブタスクであり，あるイベントを記述する最初の文書を同定するタスクである．通常は，文書中の単語の重みを値とするベクトルで表現された文書について，過去のものと類似度を計算し，最大の類似度が閾値を越えない場合，first story と判断される．しかし，マイクロブログにおける FSD は，データ量の問題があり，通常のアルゴリズムをそのまま適用しようとすると，計算量的に問題が生じる．

Petrovic ら [36] は，最近傍の文書を見つける高速なアルゴリズムを locality-sensitive hashing (LSH) を用いて開発し，性能は維持したまま 1 桁スピードアップが図れることを報告している．

9　マイクロブログにおけるトレンド分析

ブログの場合，ある程度の規模のデータを利用することが可能なら，それらのエントリの中で，あるキーワードの出現頻度がどのように推移するかを測ることで，そのキーワードが「いつ」「どの程度」注目されていたのかを知ることが可能である．Kleinberg[23] は，流行の話題を，キーワードの急激な出現頻度の増加により検出できることを示しており，この現象を ‘burst ’と呼んでいる．

マイクロブログの場合，この現象は非常に顕著な形で出現する．注目するべき出来事が発生すると,「分」,「秒」の単位で時間的に近接して，大量の類似した tweets が書き込まれることになるからである．このことから，注目するべき出来事の発生を，tweets 集合をクラスタリングするなどして検出しようという試みも当然行われている．

Weng と Lee[53] は，通常のクラスタリングではクラスタ数 (イベント数) を事前に決めておく必要があるが，実際にはイベント数は未知であり決定できないことから，Wavelet 分析を用いたイベント検出の手法を提案している．

Becker ら [5] は，tweets 系列をクラスタリングして得られた tweets のクラスタから素性集合を得た上で，得られた素性集合を元に，教師あり学習で構築した分類器を用いて，クラスタが実際に起きた出来事かどうかを online で判別する手法を示している．

Benson ら [6] は，イベントを n 項組 (record) として抽出する手法を提案している．Message(Tweet) からの要素の抽出には CRF を用い，潜在 record 集合と record-message の対応関係を同時に学習する graphical モデル (factor-graph モデル) としてモデル化している．結果として，イベントに対応する tweets 集合と，その中から抽出された record が出力されることになる．複数文書からの情報抽出の一種と考えることができる．

Lanagan と Smeaton[25] は，スポーツ中継の最中の注目するべき場面を検出し，代表的な単語でタグ付けする手法を示している．

10　マイクロブログの自動要約

ブログの場合，個々のエントリは一定の長さがあり，ある程度の量のエントリが存在する場合には，それらをまとめて要約することにはそれなりに意義があると思われる．しかし，マイクロブログの場合，1 tweet は最長でも 140 文字しかなく，果たして要約する意義はあるのだろうか．

マイクロブログの要約は，ブログの場合と同様，関連する tweets 集合をまとめて要約することになるが，前節で述べたように，(クラスタを構成すると考えられる) 関連する tweets 集合は，特定の出来事に対応すると考えられるので，前節と同様，イベント検出を行っていると考えても良い．また，前節で紹介したトレンド分析では，単に流行のキーワードを特定するだけであるが，キーワードだけでは何が流行っているのか分からないことも多い．そこで，キーワードに「肉付け」を行い，何が起こったのかを分かりやすく表示するための技術と考えることもできる [6]．

Sharifi ら [43] は，tweets 集合の要約を行う手法を示している．trending phrase(流行っている句) を含む tweets 集合から，その句を包含する最頻出の句を抽出している．前節で述べたように，ユーザは近接する時間内で類似の単語を用いて tweets を書きやすいという性質を利用している．

Takamura ら [48] は，時系列に並んだ文書としての tweets 系列から，重要な tweet を選択する要約モデルを提案している．Takamura と Okumura[47] が提案している施設配置問題としての要約モデルを応用している．時系列に並んだ tweets 集合を要約対象にしていることから，時間の情報 (近接性) をモデルに取り込む拡張が加えられている．

Liu ら [26] も，ILP(Integer Linear Programming) による，概念に基づく最適化手法を用いた要約モデルを提案している．Liu らは，要約に，tweet からリンクされた Web コンテンツも利用している．

[6] キーワード抽出技術と自動要約技術の関連と同様である．

11　マイクロブログ中の情報の信頼性評価

ブログ同様，マイクロブログも一般の人間が日常的に記述しているものなので，すべての情報が正しいという保証はない．そこで，情報の信頼性評価という，重要な課題が生まれることになる．

Castillo ら [10] は，tweet の内容が信頼できるかどうかを判別する分類器を決定木学習を用いて構築している．素性には，tweet に関する情報 (長さ，'!','?' を含むか，肯定的/否定的な単語の数，retweet かどうか)，ユーザに関する情報 (年齢，follower/followee の数，過去の tweet 数)，外部の情報源の引用の有無 (URL を含む tweet 数) を用いている．

Qazvinian ら [37] は，噂検出を行うベイズ分類器を構築している．素性には，単語，その品詞，URL を含むかどうかなどを用いている．

12　Social sensor としてのマイクロブログ

Social sensor としてマイクロブログが機能することを示した先駆的な研究は，Sakaki ら [42] のものであろう．教師あり学習により得られた分類器を用いて，tweet の内容が地震，台風に関するイベントかどうかを判定し，地震の震源，台風の軌道 (進路) を特定する手法を示している．

インフルエンザなどの病気の流行の予測を，マイクロブログを用いて行う研究も数多い．その最初は，Culotta[13] の研究と思われる．Tweets 中の，インフルエンザに関連する少数のキーワードを用いて線形回帰モデルにより，ILI (Influenza-like-illness) 統計と 95%以上の相関で予測できることを示している．Lampos ら [24] も同様にインフルエンザの流行を，マイクロブログを用いて予測している．

Paul と Dredze[33] は，(インフルエンザに限定しない) 病気全般を扱っている．Tweet が病気に関係する/しないを判別する教師あり学習に基づく分類器と，(病気がトピックに対応する)Topic model を用いている．

13　緊急時のコミュニケーション手段としてのマイクロブログ

日本では，昨年 3 月 11 日の東日本大震災の際に強く認識されたように，マイクロブログは，自然災害 (火事，台風，洪水，地震) などの緊急時のコミュニケーション手段として注目されているし，また有望でもある [45, 51]．このような場合，tweets には災害の状況報告などが記述されることになる．

このような tweets 集合を分析し，緊急時のコミュニケーションの支援が行えないかという試みも始まっている．Corvey ら [12]，Verma ら [50] は，固有名抽出と，教師あり学習により得られた分類器を用いて，災害の状況判断に用いることができる tweet かどうかを判別する手法を示している．

日本の若手言語処理研究者の有志が 3 月の震災の直後に立ち上げた anpi NLP というプロジェクトの報告は，[30] にある．固有名抽出と，安否情報を含むかどうかの判別を行い，抽出された安否情報は，Google person finder に追加される形で利用された．

14　Tweets 用のテキスト処理ツール

長さの制約等により，マイクロブログでは，これまでのテキストとはかなり異なった言語使用が行われている．そのため，これまで研究開発されてきた言語処理技術の多くがそのままでは tweets

に対しては適用不可能であるため，tweets 用のテキスト処理ツールが開発されるようになってきている．

Gimpel ら [19] は，Twitter 特有の素性を用いた POS tagger を開発し，公開している (`http://www.ark.cs.cmu.edu/TweetNLP`).

Liu ら [29] は，系列ラベリング問題としての固有名抽出において，K-NN 分類器と線形 CRF モデルを組み合わせた，半教師あり学習 (具体的には，self training) を元にした枠組みを示している．K-NN 分類器のラベルを CRF において素性として用いている．

Tweet 中には，様々なクラスの固有名が出現するが，多くは出現頻度が非常に小さい．このようなクラスについて十分な訓練データを入手するのは困難と考えられる．そこで，Ritter ら [40] は，固有名抽出を，(固有名の) 境界認定と，固有名クラス分類の 2 つの問題として分けて解いている．境界認定のみを系列ラベリング問題として解き，CRF を用いている．クラス分類では distant supervision を用いている．Ritter らは Labeled LDA を用いた distant supervision において，open-domain オントロジー (Freebase) から収集した大規模な entity 辞書を利用している．Labeled LDA を用いることで，各 entity(固有名候補個所) はクラスの混合物としてモデル化される．最終的に，固有名のクラスは，tweets 集合中での entity のクラスに対する分布から決定される．各 entity は，その entity 中および文脈中の BoW として表現されている．開発した NLP ツール (POS tagger, chunker, NER) は公開されている (`http://github.com/aritter/twitter_nlp`).

Liu ら [27] は，述語としての動詞に対する引数を tweet 中から同定し，それらに意味役割ラベルを付与する意味役割付与の問題を解いている．新聞記事を訓練データとして学習した意味役割付与システムをそのまま tweets に適用すると，新聞記事では 90%の性能が得られるのに対し，F1 値が 43.3%の性能しか得られないことを報告している．Liu らは，新聞記事で発信された内容は News tweets としても発信される傾向にあり，その場合，それらは内容的に類似しているはずであり，さらに，それらの述語項構造も類似しているはずであるという仮説を元に，新聞記事を訓練データとして学習した意味役割付与システムを用いて，News tweets 集合に述語項構造を自動的に付与し，このデータを tweets 用の意味役割付与システムを学習する際の訓練データに利用している．

Liu ら [28] は，類似 (あるいは同一) の tweets が多数存在することに着目し，この冗長性を利用して意味役割付与システムを構築する手法を示している．Tweets 集合をクラスタリングし，1 段目の意味役割付与システムの適用では，確信度の高い結果から，頻出する述語/項/役割の 3 つ組のような統計情報を収集する．2 段目の意味役割付与システムは，収集した統計情報を元に，意味役割付与の結果を洗練する．

Tweets を対象とした品詞付与，固有名抽出，意味役割付与に関する研究を紹介してきた．これらの処理は，tweets 中では通常のテキストと同様の単語が使われていることを前提にしている．だが，この前提は果たして本当に成り立っているのだろうか? Tweets 中では,「わあああい」,「やったーーー」などのような，辞書に登録されている単語を何らかの形で変形した単語が多く使われており，これらをそのままにしておくと，数多くのこれらの未知語により処理が精度よく行われる保証がないというのが実情である．

‘u’(you)，‘b4’(before)，‘goooood’ などの省略，音声置換，単語の長音化などによる未知語を，前処理として，本来の単語に戻すことができれば，品詞付与など，その先の処理の性能を向上させることができると考えられる．Han と Baldwin[20] は，これらの単語をまず分類器を用いて検出し，形態音声的類似度を元に訂正候補単語を生成し，類似度と文脈を用いて最も妥当な候補を選択することで, 辞書に登録されている単語を復元する手法を示している．文字ベースあるいは音声ベースで，未知語から閾値以下の編集距離の単語を候補単語としている．分類器は SVM を元にしているが，訓練には正例として未知語を含まない tweets 集合の部分集合，負例には正例の特定の単語を

未知語に置換したものを用いている．

15 おわりに

マイクロブログから情報を抽出，発掘する，いくつかのマイクロブログマイニング技術について，ブログマイニング技術と対照しながら，概説してきた．今後もこのサーベイは定期的に更新していきたいと考えている．

最後に，いくつか研究の紹介以外で関連する情報提供をしておきたい．まず,TREC では 2011 年から Microblog Track が開始されている．

Twitter Mining に関する書籍としては,[32, 41] が参考になるかもしれない．

2012 年 1 月号の人工知能学会誌 (Vol.27, No.1) には「Twitter とソーシャルメディア」と題する特集が組まれており，5 編の解説が収録されている．

マイクロブログマイニングに関するサービスとしては以下のようなものが存在する．

- Tweettronics(`http://www.tweettronics.com`)
  商品等に関する tweets を分析し，tweets を肯定的/否定的に分類．また, 影響力のあるユーザを同定
- Web2express Digest(`http://web2express.org`)
  面白い話題の発見
- sentiment search
  - Tweetfeel(`http://www.tweetfeel.com/`)
  - Twendz(`http://twendz.waggeneredstrom.com/`)
  - TwitterSentiment(`http://twittersentiment.appspot.com/`)
- recommendation
  - Twitter ‘Who To Follow’
  - Google Follow Finder’(`http://www.followfinder.googlelabs.com/`)
- directories
  - WeFollow(`http://wefollow.com/`)

参考文献

[1] James Allan. *Topic Detection and Tracking: event-based information organization*. Kluwer Academic Publishers, 2002.

[2] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. In *WI2010*, 2010.

[3] Roy Bar-Haim, Elad Dinur, Ronen Feldman, Moshe Fresko, and Guy Goldstein. Identifying and following expert investors in stock microblogs. In *EMNLP 2011*, pp. 1310–1319, 2011.

[4] Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *Coling 2010*, 2010.

[5] Hila Becker, Mor Naaman, and Luis Gravano. Beyond trending topics: Real-world event identification on twitter. In *ICWSM 2011*, 2011.

[6] Edward Benson, Aria Haghighi, and Regina Barzilay. Event discovery in social media feeds. In *ACL 2011*, pp. 389–398, 2011.

[7] Johan Bollen, Alberto Pepe, and Huina Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *WWW2010*, 2010.

[8] Samuel Brody and Nicholas Diakopoulos. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs. In *EMNLP2011*, pp. 562–570, 2011.

[9] John D. Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *EMNLP2011*, pp. 1301–1309, 2011.

[10] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *WWW 2011*, pp. 675–684, 2011.

[11] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: A content-based approach to geo-locating twitter users. In *CIKM'10*, 2010.

[12] William J. Corvey, Sarah Vieweg, Travis Rood, and Martha Palmer. Twitter in mass emergency: What nlp techniques can contribute. In *NAACL 2010 Workshop on Computational Linguistics in a World of Social Media*, pp. 23–24, 2010.

[13] Aron Culotta. Detecting influenza outbreaks by analyzing twitter messages. `http://arxiv.org/abs/1007.4748`, 2010.

[14] danah boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *HICSS-43*, 2010.

[15] Dmitry Davidov, Oren Tsur, and Ari Pappoport. Enhanced sentiment learning using twiiter hashtags and smileys. In *Coling 2010*, pp. 241–249, 2010.

[16] Nicholas A. Diakopoulos and David A. Shamma. Characterizing debate performance via agrregated twitter sentiment. In *CHI2010*, 2010.

[17] Yajuan Duan, Long Jiang, Tao Qin, Ming Zhou, and Heung-Yeung Shum. An emprical study on learning to rank of tweets. In *Coling 2010*, pp. 295–303, 2010.

[18] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *EMNLP 2010*, 2010.

[19] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *ACL 2011*, pp. 42–47, 2011.

[20] Bo Han and Timothy Baldwin. Lexical normalisation of short text messages: Makn sens a #twitter. In *ACL 2011*, pp. 368–378, 2011.

[21] Courtenay Honeycutt and Susan C. Herring. Beyond microblogging: Conversation and collaboration via twitter. In *HICSS-42*, 2009.

[22] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-depedent twitter sentiment classification. In *ACL2011*, 2011.

[23] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1–25, 2002.

[24] Vasileios Lampos, Tijl De Bie, and Nello Cristianini. Flu detector – tracking epidemics on twitter. In *Proc. of ECML-PKDD'10 (demo track)*, 2010.

[25] James Lanagan and Alan F. Smeation. Using twitter to detect and tag important events in live sports. In *ICWSM 2011*, pp. 542–545, 2011.

[26] Fei Liu, Yang Liu, and Fuliang Weng. Why is "sxsw" trending? exploring multiple text sources for twitter topic summarization. In *ACL Workshop on Language in Social Media*, pp. 66–75, 2011.

[27] Xiaohua Liu, Kuan Li, Bo Han, Ming Zhou, Long Jiang, Zhongyang Xiong, and Changning Huang. Semantic role labeling for news tweets. In *Coling 2010*, pp. 698–706, 2010.

[28] Xiaohua Liu, Kuan Li, Ming Zhou, and Zhongyang Xiong. Collective semantic role labeling for tweets with clustering. In *IJCAI 2011*, pp. 1832–1837, 2011.

[29] Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. Recognizing named entities in tweets. In *ACL 2011*, pp. 359–367, 2011.

[30] Graham Neubig, Yuichiroh Matsubayashi, Masato Hagiwara, and Koji Murakami. Safety information mining – what can nlp do in a disaster –. In *IJCNLP2011*, 2011.

[31] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *ICWSM2010*, pp. 122–129, 2010.

[32] T. O'Reilly and S. Milstein. *The Twitter Book*. O'Reilly, 2009.

[33] Michael J. Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. In *ICWSM 2011*, 2011.

[34] Marco Pennacchiotti and Siva Gurumurthy. Investigating topic models for social media user recommendation. In *WWW2011*, pp. 101–102, 2011.

[35] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, republicans and starbucks afficionados: User classification in twitter. In *KDD'11*, pp. 430–438, 2011.

[36] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to twitter. In *The Anuual Conference of the North American Chapter of the ACL 2010*, pp. 181–189, 2010.

[37] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *EMNLP*, pp. 1589–1599, 2011.

[38] Daniel Ramage, Susan Dumais, and Dan Liebling. Characterizing microblogs with topic models. In *ICWSM 2010*, 2010.

[39] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pp. 37–44, 2010.

[40] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named entity recognition in tweets: An experimental study. In *EMNLP 2011*, pp. 1524–1534, 2011.

[41] Matthew A. Russell. *21 Recipes for Mining Twitter Distilling Rich Information from Messy Data*. O'Reilly Media, 2011.

[42] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *WWW 2010*, 2010.

[43] Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. Summarizing microblogs automatically. In *The Anuual Conference of the North American Chapter of the ACL 2010*, pp. 685–688, 2010.

[44] Ismael S. Silva, Janaina Gomide, Adriano Veloso, Wagner Meira Jr., and Renato Ferreira. Effective sentiment stream analysis with self-augmenting training and demand-driven projection. In *SIGIR'2011*, pp. 475–484, 2011.

[45] Kate Starbird, Leysia Palen, Amanda L. Hughes, and Sarah Vieweg. Chatter on the red: What hazards threat reveals about the social life of microblogged information. In *CSCW 2010*, 2010.

[46] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *SIGKDD 2004*, 2004.

[47] Hiroya Takamura and Manabu Okumura. Text summarization model based on the budgeted median problem. In *CIKM 2009*, pp. 1589–1592, 2009.

[48] Hiroya Takamura, Hikaru Yokono, and Manabu Okumura. Summarizing a document stream. In *ECIR 2011, 33rd European Conference on Information Retrieval*, pp. 177–188, 2011.

[49] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *ICWSM2010*, pp. 178–185, 2010.

[50] Sudha Verma, Sarah Vieweg, William J. Corvey, Leysia Palen, James H. Martin, Martha Palmer, Aaron Schram, and Kenneth M. Anderson. Natural language processing to the rescue?: Extracting "situational awareness" tweets during mass emergency. In *ICWSM 2011*, 2011.

[51] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events: What twitter may contribute to situational awareness. In *CHI 2010*, 2010.

[52] Xiolong Wang, Furu Wei, Xiohua Liu, Ming Zhou, and Ming Zhang. Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach. In *CIKM'11*, pp. 1031–1040, 2011.

[53] Jianshu Weng and Bu-Sung Lee. Event detection in twitter. In *ICWSM 2011*, pp. 401–408, 2011.

[54] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM 2010*, 2010.

[55] Wei Wu, Bin Zhang, and Mari Ostendorf. Automatic generation of personalized annotation tags for twitter users. In *NAACL 2010*, pp. 689–692, 2010.

[56] Wayne Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achananuparp, Ee-Peng Lim, and Xiaoming Li. Topical keyphrase extraction from twitter. In *The Anuual Meeting of the Association for Computational Linguistics 2011*, pp. 379–388, 2011.

[57] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *ECIR 2011*, 2011.

[58] 奥村 学. blog マイニング–インターネット上のトレンド, 意見分析を目指して–. 人工知能学会誌, Vol. 21, No. 4, pp. 424–429, 2006.

[59] 奥村 学. ブログマイニング技術の最新動向. 電子情報通信学会誌, Vol. 91, No. 12, pp. 1054–1059, 2008.