マイクロブログにおける有用な記事の発見支援

岩木　祐輔†　　アダム　ヤトフト†　　田中　克己†

†京都大学大学院情報学研究科社会情報学専攻　〒606-0123 京都市左京区吉田本町

E-mail:　†{iwaki, adam, tanaka}@dl.kuis.kyoto-u.ac.jp

概要　本論文では，ブロガーの行動およびブロガー同士のリンク構造に着目して，マイクロブログから有用な情報を効率よく発見するための手法を提案する．その方針としては，有用さの指標としてユーザーと記事の近接度を考え，マイクロブログで多く語られる独り言や生活記録など，万人にとって有用とはいえない情報によってユーザーの求める体験記事が埋もれないように情報提示を行う．ユーザーと記事との近接関係が，ブロガーの行動や感性の類似度あるいは単純にブロガー同士のリンク構造によく表れると本研究では仮定し，特徴語の共起をもとにした感性辞書の作成や，ブロガー同士の様々なリンク構造の分析行って，それぞれ指標がどれほどユーザーと記事の近接関係を表すことができるものか実験を行った．そして，検索システムのプロトタイプを作成し，本手法の評価を行った．今後の課題としては，検索速度の向上および，プロファイルを持たないブロガーに対する検索手法を検討する．

キーワード　マイクロブログ，ブログ検索，リンク構造

Supporting finding read-valuable articles in micro-blogs

Yusuke IWAKI†　　Adam JATOWT†　　and　Katsumi TANAKA†

† Department of Social Informatics, Graduate School of Informatics, Kyoto University Yoshida-honmachi, Sakyo, Kyoto, 606-8501 Japan

E-mail:　†{iwaki, adam, tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract　In this article, we suggest how to efficiently discover information that is valuable for reading from micro-blogs based on the activity of bloggers and their link structure. Micro-blogs are an emerging type of blogs which are characterized by the high frequency and short length of posts, which are usually made in real time in everyday life situations. We regard the proximity degree between a user and an article as an index of read-valuability. In this study, we assume that the similarity degree of a user and an article is affected by the link structure between bloggers and their activities. We have created a sentiment dictionary, analyzed the link structure, and finally tested which index mainly affects the proximity relations between posts and bloggers.　Finally, we created the prototype of our system and evaluated it. The future problems are the improvement of the search speed and the search technique for bloggers who do not have any profiles.

Keyword　Micro-blog, Blog-search, link-structure

1. はじめに

近年，ブログや SNS，ソーシャルブックマークなど様々な Web サービスの普及により，多くの人が Web 上に知識や体験を残すようになってきた．また，マイクロブログというチャットとライフログの中間的なサービスも登場し，そのコンテンツは爆発的に増え続けている．一方，コンテンツの内容は，従来は情報提供を目的としたものが大半であったが，最近では単なる生活記録や忘備録，独り言など，いわゆる「個人的」なコンテンツの割合が増加してきた．

このように，爆発的に増えた投稿の中からユーザーが体験を検索する際には，現在のところ主だった手段としては，キーワード検索あるいはタグ検索などしか提供されていない．そのため，たとえば最新の携帯電話を購入した感想などを見る際に，よい評価と悪い評価が混在していて直観的にわかりにくかったり，話題のニュースへの意見を見る際に，感想や文句などの様々なジャンルの投稿があり，一見して判断がつきにくかったりすることが多い．

そこで本研究では，マイクロブログからユーザーにとって読む価値のある記事を効率よく発見する方法を考案する．読む価値のある記事には 2 通りあり，1 つはユーザーの興味と近接した記事である場合，そしてもうひとつは記事の内容自体が万人向けである場合である．これをまとめると，以下の表 1 のようになる．

表 1 読む価値のある/ない記事

	public な投稿	private な投稿
author に興味がある	*valuable*	*valuable*
author に興味がない	*valuable*	*not valuable*

本研究のアプローチとしては,「author に興味がある/ないの判別」のみによって記事の valuable/not valuable を決めることにする．記事の内容の public/private という側面を無視することで「auhor に興味がないが public である投稿」の false dismiss が生じてしまうが，マイクロブログではこのような性質を持つ書き込みはボット(Web 上のニュースなどの情報を機械的に横流しするシステム)による投稿であることが多いため，本研究ではこの誤りまでは考慮しないことにする．記事の内容に対する public/private の判別問題については，今後の研究課題とする．

記事の author に対する興味の要素として，本研究では「ブロガー同士のつながり」および「過去のユーザーインタラクションの内容」の 2 つを考える．マイクロブログに見られる「友達」という概念のリンク構造の解析や，記事に対するコメントが繰り返し行われた部分に対する特徴語抽出によって，ユーザーの興味のある author の抽出を試みた．

最後に，以上のシステムのプロトタイプシステムの実装を行った．本手法の有効性については今後，ブロガーのユーザー評価を行って解析する．

2. 関連研究

ブログのマイニングに関する研究は多数行われている．体験のマイニングについては，ある場所における体験をブログから抽出し可視化する体験ブログマップ [1]や，ブログに語られた経験の再利用を目的として抽出する経験マイニング [2] [3]などがある．これらは，いずれも単語の出現頻度のみによってマイニングを行っている．本研究では，言語処理だけでなく，人のネットワークにも着目してマイニングを行っている．

マイクロブログの研究はまだ歴史が浅い． [4]では，マイクロブログにおける時系列の話題語の調査，HITS アルゴリズム [5]を用いたコミュニティ発見などを行っているが，それぞれの記事の有用さなどについては一切触れられていない．

以上より，本研究の特徴を整理すると，人のつながりを考慮してマイニングを行っている点，さらにそれを記事の有用さへと帰着させる点，が挙げられる．

3. マイクロブログ

本章では，最も有名なマイクロブログのサービスで，本研究でも用いた twitter[1]を例に基本的なシステムを説明する．

マイクロブログは大雑把には「ライフログとチャットの中間的な存在」である．ライフログのように機械的な記録ではなく人の手によって記録され，ブログのように書きたい内容を集約して書く必要がなく，チャットのようにコミュニティ内のレスポンスを強要されることがなく,自由な投稿ができることが特徴である．また，SNS に見られる「友達」(friend/follower)という概念も存在する．

図 1 ユーザーABC の投稿一覧ページ

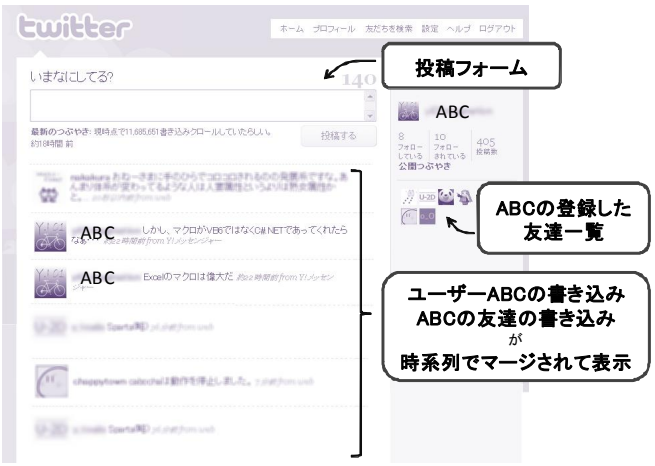

図 2 ユーザーABC の専用ページ

マイクロブログのユーザーには図 1 のような一覧ページと図 2 のようなユーザー専用ページが与えられる.一覧ページはユーザーの過去の投稿が羅列され，誰でも見ることができる．一方，ユーザー専用ページは，自分のみが見ることができるページで，過去の投

[1] http://twitter.com./

稿に加えて登録した友達の投稿もあわせて見ることができるようになっている．これによって，ユーザーはただ自分の思うことを書くだけでなく，友達の投稿に対して「返信」という形でも投稿をすることができる．

マイクロブログからの記事の発見において問題となるのは，単純なキーワード検索で返される結果に

- 様々なコンテキストが含まれる
- 投稿者の属性がわからない

ことである．たとえば最新の携帯電話について調べようとして「最新　ケータイ」などとキーワード検索して「最新のケータイ X-01Y を買った」や「最新の X-01Y 使いやすい!!」といった投稿にたどり着けることは少なく，発見できたとしても，そのユーザーの属性がわからず，どれがためになる情報なのか判断がつかない．

図 3 twitter search での検索例

そこで，以下ではこれらの情報を補完しながら，読む価値のある記事を抽出する方法を述べる．

4. 提案手法

まず，分析手法の概観を図 4 に示す．

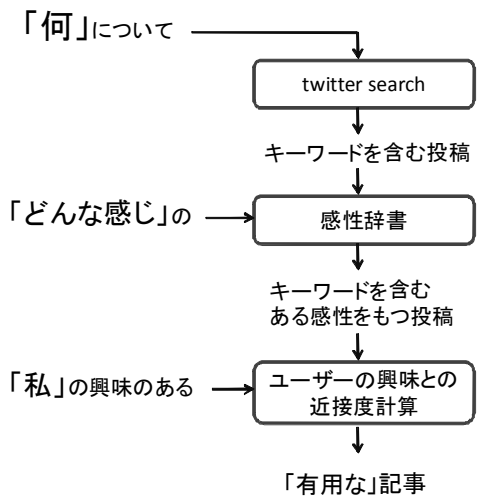

図 4 本システムの概要

まず，ユーザーの知りたいことをキーワード検索し，出力された様々なジャンルの投稿の中で，ユーザーがどのような情報を求めているのかコンテキストを限定し，最終的にユーザーのプロフィールや過去の投稿をシステムが収集し，ユーザに近いものからランキングして，結果を出力する．なお本研究では，サーチするユーザー自身がアクティブなブロガーであることを仮定してすすめることにする．

4.1. 感性辞書によるジャンル特定

検索の適合率を高めるために，まずはユーザーの求める記事のジャンルを絞ることを考える．そこで，感性辞書による記事の分類を行う．本節では，感性辞書の作成方法および，その評価について述べる．

一般に，辞書を作成する際には語の共起をもとに相関ルールを抽出する。たとえば語集合 T について，

$$T \Rightarrow \text{"happy"}$$

というルールを導出する際には，支持度の最小値 α と確信度の最小値 β を決めて，

$$\frac{N(\text{"happy"},T)}{N(*)} > \alpha$$

$$\frac{N(\text{"happy"},T)}{N(T)} > \beta$$

を満たすかどうかを調べる．

しかし、マイクロブログでは TF 値が概して低いため上のような方法で感性辞書作成を行うと、不都合が生じる。α や β の閾値を高くとると、語彙（抽出されるルール）が非常に少ない辞書となってしまう。逆に、これらの閾値を低く設定すると、語彙が増える一方でめったに現れない特殊なルールが無駄に大量に生成されることとなる。また、ノイズも増加する。

また、マイクロブログには即時性という大きな特徴があり、感性辞書を静的に作成してしまうとこれも不都合が生じる。たとえば、「X-01Y, タッチパネル⇒happy」のようなルールがひとたび生成されると、静的な手法では永遠に X-01Y のタッチパネルがよいものという認識がされて、時系列的に変化しうる評判などでは不適合となる。

以上の 2 点を考慮に入れて、本研究ではマイクロブログ以外のソース（ブログ検索・ブログタグ）も用いて、動的に感性辞書を作成し、その辞書をマイクロブログ検索結果に適用する、というアプローチを取った。これによって不必要なルールの計算を押さえ、さらに on-the-fly で感性による検索が可能となった。

感性による検索のシステムの流れを図 5 に示す。大まかな流れとしては、ユーザーの入力したキーワードの検索結果に含まれる特徴語の中から「感性を変化させる」語を抽出し感性辞書に登録する。

本研究では語 w がもつ感性を(”うれしい”,”悲しい”,”つらい”,”怒り”,”独り言”)の 5 要素を規定とする Sentiment-Vector(w)としてあらわす。Sentiment-Vector

のそれぞれの要素の値は「ブログ検索における、語 w とうれしい/悲しい/つらい/怒り/独り言、それぞれとの共起頻度」と定義した。

に示したように「キーワード単体での感性のヒストグラム」とキーワード，5 語とそのシソーラスをベースに，上のような導出ルールで辞書を作成した．

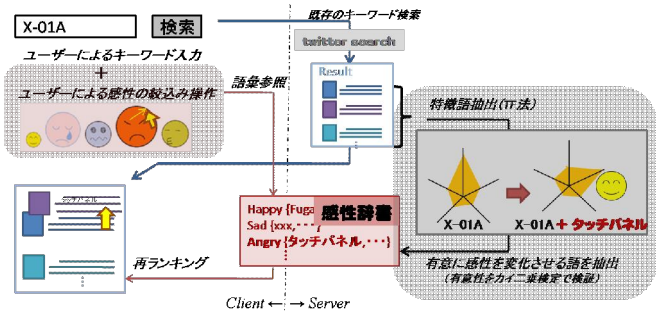

図 5 感性辞書の作成・適用

4.2. ブロガーの近接度の計算

検索結果に現れるブロガーX と検索ユーザーA の近接度は，「X と興味が似ている人が A のどの程度(人のつながり的に)近傍にいるか」という考え方に基づき，図 6 のように定義する．

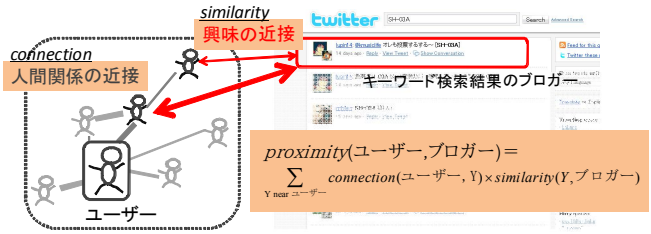

図 6 ブロガーの近接度の計算

人のつながりを表すものとして twitter には友達(friend/follower)の概念があるが，これはつながりの度合いに関する情報はもたない．あるユーザーにとって，単純に他人の投稿を流し読みする程度の弱いつながりであるのか，情報を交わすことを目的として強いつながりを持っているのかを知るには friend/follower 以外の情報も必要となる．

そこで，本研究では「ユーザーの書き込みに対する返信」に着目して，つながりの度合いを考えた．たとえば，ユーザーA が B の書き込みに対して何か返信をして，さらに B が A に対してまた返信をして，というインタラクションがしばしばある場合には A と B は強いつながりを持っており，逆に，ただフォローしているだけでそのようなインタラクションが見られない場合には A と B はつながりを実質的にもたないものと推測できる．人のつながり(connection)はユーザーの返信回数を区間[0,1]にマッピングする関数 $f(n)=1-e^{-\frac{n}{K}}$ (K は定数)によって計算した．

ユーザーとブロガーの興味の近接度については，ブロガーがそれまでに返信行動をおこすきっかけとなった記事周辺に含まれる特徴語の類似度と定義した．

特徴語抽出については言選 Web[2]を利用した．言選 Web では，単名詞バイグラムを用いて得られる複合名詞の連接情報・候補語の頻度情報をもとに文章から重要語抽出を行っている．以下に抽出例を示す．

文章例

そもそも人間の感覚情報の 80%は視覚だ．つまりホームページのよしあしの第一印象は「ぱっと見」だ．特徴語をこねこねしても見た目には関係ないし，，，ああ，現実逃避，，，PageRank はリンク数を頼りにしているけど，それが見やすいかどうかまで考慮して PageRank を補正できないものか，，

表 2 言選 Web によるキーワード抽出例

特徴語	重要度
PageRank	2
特徴語	1.41
感覚情報	1.41
現実逃避	1.41
リンク数	1.41
関係	1
ホームページ	1
よしあし	1
人間	1
視覚	1
印象	1
見た目	1

この例の文章に返信をするようなブロガーはおそらく情報技術に詳しい人間であり，抽出されたキーワードも情報技術に関するものが多いことがわかる．

興味の近接度(similarity)はこのようにして抽出された特徴語のキーワード検索結果のコサイン類似度によって計算した．

以上のように計算した connection と similarity をもとに，ブロガーとユーザーの総合的な近接度(proximity)を算出し，その値の大きい順にキーワード検索結果のブロガーを再ランキングした．

5. プロトタイプアプリケーション

4 で構成したシステムをもとに，プロトタイプのアプリケーションを作成した．ブラウザとの親和性から UI は HTML と JavaScript で作成し，バックエンドのデータ処理は PHP で実装した．

ユーザーが本システムで仮定している感性の絞込み操作を直感的に行えるよう，顔アイコンの大きさを操作する UI を作成した．これにより，例えばユーザーは笑顔のアイコンを大きくすることで，検索結果を

[2] http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb.html

「うれしい」記事優先に変化させることができる．

また，検索結果のブロガーが自分にとってどういうつながりを持つ人であるかを視覚的に示すため，コミュニティの表現に用いられるばねグラフモデルを用いてユーザーの近傍の人間関係の可視化を行った．

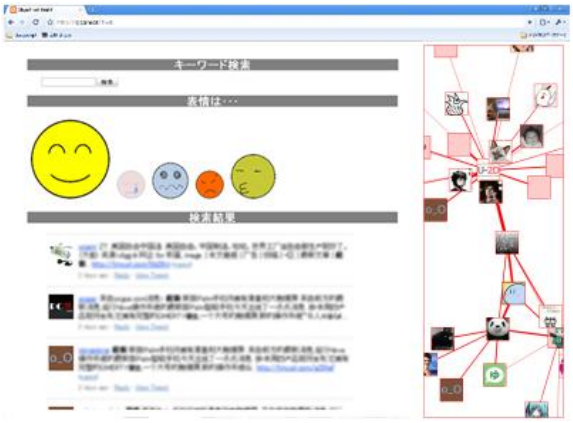

図 7 プロトタイプアプリケーション

また、ユーザーそのものの特徴はここでは可視化されないため、図 8 のような返信行動に基づく特徴をタグクラウドとして可視化する機能も付した。

図 8 ユーザーの特徴タグクラウド

6. まとめと今後の課題

本論文では，マイクロブログから読む価値のある記事を効率よく抽出するため，ブロガー同士のつながりやブロガーのインタラクションを分析してその効果を検証した．ブロガー同士のつながりの強さを返信回数によって計算し，返信内容の特徴語を抽出することで，ユーザーの興味に近い記事を判別できることを示した．

今後の課題としては，ブロガーとしてのプロフィールを持たないユーザーが検索を可能にする方法の考案，および，検索に要する時間の短縮が挙げられる．

謝辞 本研究は一部，グローバル COE 拠点形成プログラム「知識循環社会のための情報学教育研究拠点」，科研費：計画研究「情報爆発時代に対応するコンテンツ融合と操作環境融合に関する研究」(課題番号 18049041)，若手研究 (B)「情報検索とウェブアーカイブにおけるマイニング」(課題番号：18700111) によるものです．ここに記して謝意を表すものとします．

文 献

[1] 倉島健，手塚太郎，田中克己．“街 Blog からの体験抽出とその空間的提示手法の提案”，情報処理学会研究報告，DBS-137, pp47-53, 2005

[2] Nozomi Kobayashi, Kentaro Inui and Yuji Matsu-moto. Extracting Aspect-Evaluation and Aspect-Of Relations in Opinion Mining. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp 1065-1074, 2007.

[3] 乾健太郎，原一夫．経験マイニング： Web テキストからの個人の経験の抽出と分類．言語処理学会第 14 回年次大会，2008.

[4] Akshay Java, Xiaodan Song, Tim Finin, Belle Tseng. Why we twitter: understanding microblogging usage and communities, International Conference on Knowledge Discovery and Data Mining. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis,San Jose, California, pp 56-65, 2007.

[5] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. Proceedings of the 9th ACM Conference on Hypertext and Hypermedia, 1998.