

ユーザ興味推定に基づく Twitter タイムラインの速度制御手法

Controlling Speed for the Twitter Timeline based on User's Preference

河野 四季 Shiki Kouno

コンピュータ科学科 インタラクティブメディア研究室

指導教員 北山 大輔 助教

1 はじめに

近年、手軽に情報を発信することができるマイクロブログが普及し、代表的なものの1つに Twitter がある。Twitter では、『フォロー』したユーザの投稿である『ツイート』を時系列に表示することできるタイムラインを持つ。従って、多くのユーザをフォローするとタイムラインに表示されるツイート数が増大し、結果としてタイムラインの速度も速くなるため、読みたいツイートを見落としがちになる。興味のないものがタイムラインに増えてしまう例として、スポーツ番組などの際にその内容をツイートする実況があげられる。この事例に対しては青島ら [1] が単語の共起度とタイムスタンプを用いた方法で、ツイートの文脈からつながりを考慮したツイートの抽出を行い、同一の話題に対するまとめ処理によるタイムラインの速度調節を行なっている。それに対し、本研究ではタイムラインにおける単語の出現頻度からユーザの興味のあるツイートを抽出し、それ以外の興味のないツイートの数をユーザごとに最適な数に調節することでユーザの読みたいツイートを見落とさずに読める速度に調節する手法を提案する。

2 ユーザの興味抽出

2.1 ユーザの興味に関する語の抽出

我々は、フォローしたユーザが共通して投稿する内容に、ユーザのフォローした理由である興味のある内容が含まれていると考えた。そのため興味のないツイートはおおよその場合、複数のユーザが呟くことは少なく、「興味のあるツイート」は複数のユーザが呟くことが多いと仮定した。このことから「興味のあるツイート」の抽出にはタイムライン中のツイートに含まれる単語を解析し、その単語についてツイートしているユーザ数を出すことで求めることが出来ると考え

た。Twitter ではユーザは自分の気に入ったツイートをお気に入りに登録することが出来る。よってお気に入りに登録されたツイートにはユーザが興味を持つ内容が含まれているはずである。ここで、興味を表す内容は単語によって表現されるものとする。またユーザは興味のあるツイートをしている人をフォローするのでフォローした時点より前のツイートで、フォローするきっかけとなった興味のあるツイートをしていると思われる。よってお気に入り内のツイートおよびきっかけとなったと思われるツイートに対して形態素解析を行い、名詞を抽出し、それをお気に入り内のツイートに関しては各単語を含むツイート数およびツイートしたユーザ数の2点でスコアリングを行う。その後ツイート数のスコアでの上位語をツイートしているユーザのツイートを検索し、ユーザ数でのスコアの上位語をツイートしていればそのまま、ツイートしていなければそのユーザのツイートをスコアの算出から除外する。きっかけとなったツイート中でツイートしているユーザ数で単語のスコアリングを行う。お気に入りに関してはすべてのツイート、きっかけとなったツイートはフォローした時のタイムスタンプをもとに、その前に行われたツイート 20 件に対して行う。ユーザ数でのスコアリングは以下の式で行う。

$$U(F \cup K, x) \quad (1)$$

とする。 x は計算対象のキーワードであり、 U は入力されたツイート集合の中で対象単語について呟いているユーザ数を返す関数である。 F はお気に入り内のツイート集合、 K はきっかけとなったツイート集合である。 $U(F \cup K, x)$ が高い単語を過去にツイートしたことがあるユーザに限定して、ツイート数を数えた。ツイート数に関するスコアは次の式で求める。

$$Tw(F, x) \quad (2)$$

Tw は x のツイート頻度を返す変数である。以上の方法で求めたスコアをユーザの興味度とする。

2.2 関連語の抽出

前項でユーザの興味のある単語についてのスコアを求めた。しかしお気に入りと多くのユーザにツイートされている単語だけではユーザの興味のある単語として十分であるとはいえない。よって求めた単語の共起語を抽出し、ユーザの興味語に加える。手法としては、スコア付けされた単語をフォローしているユーザのツイートから検索し、出てきた共起語に関して何割のユーザがツイートしているかを求める。その後出てきた割合を検索語のスコアに掛け合わせその興味度とする。

3 速度調節

前節でユーザの興味語に関してのスコアリングを行った。このスコアと、利用時のタイムラインの各単語に関してツイートしているユーザ数をもとにスコアリングを行い、ユーザが設定した速度になるように確率的に削除し速度を制御する。ここで、利用時に多く出ている単語が必ずしも残す対象とは限らず、興味語の上位にあるが、タイムラインではツイートしているユーザが少ないものは残しておきたいものであると考えた。よって利用時の各単語についてのユーザ数を取り、それで興味語のスコアをわることで利用時における最終スコアとする。以下の式で算出を行う。

$$\frac{Tw(F, x)}{Tw(TL, x)} \quad (3)$$

$Tw(TL, x)$ は利用時点での単語 x のツイート頻度である。ツイート頻度は、タイムライン中でどれだけのツイートに含まれているかを表す指標であり、この値が高ければ、タイムライン上でありふれた単語であることを表す。そのため、この値が低いもので興味度が高い興味語を高確率でタイムライン上に残した速度調節が可能となる。

4 実験結果

行った実験の結果を載せる。今回はフォローしているユーザのツイートは最新のものから 200 件を用いた。式 (1) によるユーザ数のスコアによってお気に入りから 20 件の単語を抽出し、式 (2) で求めたツイート数スコア 20 件に関連語も含めて上位 20 件を抽出した。図 1 は上記手法で求めたスコアの表である。ツイート数のスコアによる結果を見ると、おおよそ興味のある単語が上位に来ることが確認できたが、『人』や

ユーザ数スコア		ツイート数スコア		関連語を追加したスコア	
モバ	25	雪美	40	雪美	40
マス	24	ゆきみ	37	ゆきみ	37
課金	18	モバ	30	モバ	30
人	16	マス	29	マス	29
フェス	15	こと	21	こと	21
cg	13	ディアス	21	ディアス	21
imas	13	ノーマル	21	ノーマル	21
pixiv	12	課金	20	課金	20
プロ	10	人	17	人	17
ドラ	8	フェス	16	ガチャ	16
バズ	8	明日	16	フェス	16
アイドル	8	佐城	15	明日	16
シンデレラ	7	cg	14	佐城	15
上位	7	imas	14	cg	14
FLASH	6	pixiv	14	imas	14
プロデューサー	6	蘭子	12	pixiv	14
ロリコン	6	pt	7	イベント	13
ゲーム	5	ロリコン	3	蘭子	12
ホモ	5	グッド	2	仁菜	11
SR	4	バーフェクト	2	上位	10

図 1: スコア表

『こと』、『明日』といった興味に関係があると思われる単語や『pixiv』といったサイト名などが上位に来てしまうことも確認された。関連語を追加したスコアを見ると、『人』と『明日』が残ってしまっているがほぼ興味のある単語が上位に来る結果となった。これは検索した関連語を追加したことがより興味に関する内容であったためだと考えられる。興味のある内容の単語が 12 個、興味がある内容に関しても使われるが他のことに関しても使われる単語が 5 個、興味のない単語が 3 個で適合率は 85% だった。実験結果からお気に入り内にある興味のある単語がスコアの上位に来ることが確認できた。しかし汎用的な語が上位に来てしまうなどの問題も発見された。また関連語に関しては十分な数、スコアが取れなかった。この原因は元となったフォローしている人のツイートの取得数が少なかったことと、同一の内容に関するツイートを複数に分割されていることが原因だと考えられる。

5 まとめと今後の課題

本手法ではお気に入り内の単語から興味度を抽出したが、画像の内容に興味がありお気に入りに登録することもあり、画像解析も併用することで本手法よりも正確な興味度を算出できると考えられるので画像データの利用が今後の課題である。

参考文献

- [1] 青島 傳隼, 坂本 翼, 横山 昌平, 福田 直樹, 石川 博文脈的なつながりを考慮したツイート群の効果的な抽出・提示手法の実現, 情報処理学会論文誌: データベース, Vol.6, No.2, PP.61-84, 2013