

メタデータ付与行為による情報フィルタリング

山本直史†

高崎尚人‡

小林亜樹†

† 工学院大学工学部情報通信工学科

‡ 工学院大学大学院工学研究科電気・電子工学専攻

1 はじめに

マイクロブログの投稿数が膨大なものとなり、複数のユーザが何らかのトピックに言及を行う中、情報フィルタリングは、ユーザや記事中に含まれる語を指定する単純な絞り込み程度しか提供されず、横断的に興味トピックを抽出するのには不十分である。

マイクロブログにおけるトピック抽出では Twitter-LDA モデル [1], Author-topic モデル [2], Twitter-TTM モデル [3] などが提案されている。本稿では、トピックのとらえ方に個人差があることを考慮し、他利用者のトピック分類を選択的に利用できるようにするため、メタデータストリームの概念を導入する。複雑化するインタフェースについても述べる。具体的には、記事へのキーワード付与行動ストリームを提案する。

2 SNS の記事閲覧モデル

ある SNS に投稿された記事の投稿総数 M として、記事 a_i ($1 \leq i \leq M$)、全記事集合を A と表す。記事には、本文のほか投稿者 ID、キーワード集合 $*$ が属性として存在する。これを、式 (1) のように表す。

$$a = \begin{cases} \text{user : author} \\ \text{text : 記事本文} \end{cases} \quad (1)$$

また、記事から投稿者、キーワード集合を得る関数を、それぞれ $\text{id}(a)$, $\text{KWs}(a)$ とすると、投稿者 ID が j である記事集合 A_j 、キーワード w を含む記事集合 T_w はそれぞれ、

$$\begin{aligned} A_j &= \{a_i \mid a_i \in A, \text{id}(a_i) = j\} \\ T_w &= \{a_i \mid a_i \in A, w \in \text{KWs}(a_i)\} \end{aligned} \quad (2)$$

となる。これらは、それぞれ投稿者毎、キーワード検索時のタイムラインに現れる記事集合に相当する。

一方、キーワード付与という情報を考えることができ、利用者 ID u が付与したキーワード w を $pK(u, w)$ と書くこととする。投稿者以外によるキーワード付与を許していないサービスにあっては、 u は常に投稿者に一致する。

既存サービスでは、 A_j , T_w によるフィルタリングを提供しているが、両者の組み合わせについては提供している例は少ない。また、キーワード付与 $pK(u, w)$ によるフィルタリング機能は提供されていない。これらは、概念の複雑さとそれらを利用者に提供するインタフェース上の制約と考えられるが、多数の情報を扱う利用者からすると、これらを自由に組み合わせたフィルタリングが求められる。

3 キーワード付与行為によるフィルタリング

3.1 キーワード付与行為

利用者 ID m がキーワード付与した記事集合を M_m とする。 $pK(u, w)$ は拡張されて、 $pK'(a, t, u, w)$ となる。 a はキーワード付与が行われた記事を表し、 t はキーワード付与が行われた時刻を表す。なお、投稿者以外がキーワード付与を行えるとき、 $pK(u, w)$ の u は必ずしも投稿者と一致しない。

一般に、キーワードを付与する人物はあるトピックに関連する多くの記事にキーワードを付与すると予想される。一方で、付与者にとって興味の無い記事および専門外の記事、またはトピックに関連するが有用でない記事、についてはキーワード付与を行わないと予想される。

3.2 フィルタリング

キーワード付与 pK は、記事に対するメタデータの時系列ストリームとなる。これを利用して、 $pK'(a, t, u, w)$ の各パラメータやその範囲などによるフィルタリングを行い、その結果メタデータストリームが指し示す記事を結果として得るフィルタリングを提案する。例えば、「指定利用者が特定のキーワード付与を当該記事の投稿後一定時間以内に行った記事」といった複雑なフィルタリングを設計できる。従来の A_j, T_w などとの演算も当然定義される。

4 提案手法

本提案の実装環境として、Twitter を選択した。Twitter では、フォローおよびリストと検索機能それぞれの和集合を取ることは可能であるが、その組み合わせについては提供されていない。本章では、追加する機能とその実装手法について述べる。

Information filtering by metadata creation behavior

†Naofumi Yamamoto ‡Naoto Takasaki †Aki Kobayashi

†Department of Information and Communications Engineering, Faculty of Engineering, Kogakuin University

‡Electrical Engineering and Electronics, Kogakuin University Graduate School

*一部の SNS に限る。また、本文中に含む実装もある。

4.1 タグ付与

本手法では、投稿者以外の人物も Tweet にタグとしてキーワードを付与できるような環境を試作する。利用者が Tweet にタグを付与するとき、そのタグ付与に紐付いて、付与者がタグを付与したという行動がメタデータとして付与される。よって、閲覧者はタグによって Tweet を検索することも可能であるが、タグ付与行動を基にして Tweet を求めることも可能である。

タイムライン上に表示されているタグ付与 Tweet は、下部にキーワードとしてタグが表示されるようになる。

タグ付与行動は、テキストボックスによる入力か、付与者が今迄に付与したことの有るタグ履歴群から選択する形で行われる。

4.2 タイムライン

フィルタリングの段階において重要な意味を持つ時間軸であるが、それは提示段階でも同様である。閲覧者がタイムラインを閲覧するとき、複数の興味トピックを持っていることは想像に難くない。しかし、トピック毎にタイムラインを切り分けて閲覧することは、時系列が崩壊し、それに伴ってトピックの関連性が薄れる可能性が有る。また、複数同一 Tweet がタイムライン毎に表示される無駄は排除したい。

そこで、フィルタリングによる複数タイムラインを一つに統合する。このとき、Tweet にフィルタリングの条件によって色付けを行う。これによって一つのタイムラインであっても、閲覧者は直感的に複数のトピックが存在することを理解可能である。また、一つのタイムラインにすることによって、同一の Tweet が複数回出現するのを回避し、Tweet 単位で相対的な時系列を確保する。その模式図を図 1 で示す。

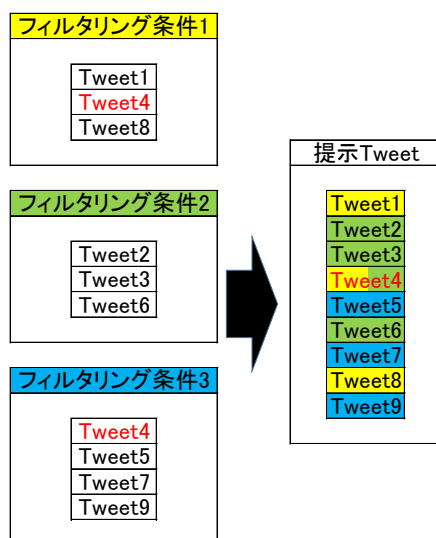


図 1: タイムライン統合化模式図

5 実装

本提案における試作環境のブロック図を図 2 で示す。StreamingAPI によって Tweet を収集し、MySQL によって管理する。Tweet の基本情報は TweetDB に格納される。キーワードや付与者といったメタデータは他の DB によって取り扱う。クライアントが DB から Tweet を求める時は、中継器であるサーバへフィルタリング要求を送信する。サーバは受け取った要求を DB へ送信する。DB は格納されている Tweet 群から、要求に沿った Tweet をサーバ経由でクライアントへと返す。クライアントが Tweet にタグ付与を行うときも同様である。

6 おわりに

本研究では、SNS における新たなフィルタリング方法を提案した。ユーザの欲求を容易に実現するためのインタフェースなどは今後の課題となる。

謝辞

本研究の一部は科学研究費（基盤 (A)）No. 26242013 による。

参考文献

- [1] W.X. Zhao, J.jiang, J.He, Y.Song, P.Achananuparp, E.-P. Lim, and X.Li, "Topical keyphrase extraction from twitter," The Annual Meeting of the Association for Computational Linguistics 2011, pp. 379-388, 2011.
- [2] M.Steyvers, P.Smyth, M.Rosen-Zvi, and T.Griffiths, "Probabilistic author-topic models for information discovery," SIGKDD2004, 2004.
- [3] 佐々木謙太郎, 吉川大弘, 古橋武, 「Twitter におけるユーザの興味と話題の時間発展を考慮したオンライン学習可能なトピックモデルの提案」, 情報処理学会論文誌. 数理モデル化と応用 7(1), 53-60, 2014-03-28.

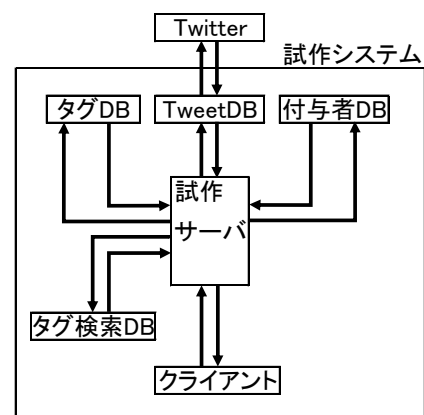


図 2: 試作環境ブロック図