マイクロブログ・ストリームからの社会的知識の獲得

藤山 健一郎　　是津 耕司　　木俵 豊

独立行政法人 情報通信研究機構 知識創成コミュニケーション研究センター

1. はじめに

Twitter[1]を初めとするマイクロブログ(MB)サービスでは、多くのユーザから様々な事象に対する無数のメッセージが投稿され続けている。近年ではイラン大統領選挙のような事象に対しても、従来メディアではカバーできない情報のリアルタイム・メディアとして活用された。よって、この連続したメッセージ群…MB ストリームを解析することで、特定事象に関する詳細な情報等、社会でリアルタイムに生成されている知識を得られると考えられる[2][3]。本稿ではMB ストリームから、これら社会的知識を獲得する手法について論じる。

2. 課題検討

本項では Twitter の特徴を説明し、それ故に発生する知識獲得の際の課題を述べる。

2.1. Twitter

Twitter は、ユーザが短いテキストメッセージを投稿することで互いに繋がるコミュニケーションサービスである。メッセージは tweet と呼ばれ、140 文字以下という制限がある。Twitter の特徴として、以下に挙げるような、他ユーザにもメッセージが伝えられる関係構造がある。

- **フォロー/フォロワ：** ユーザ間の関係であり、フォローユーザの投稿したメッセージはフォロワユーザにも伝えられる。
- **リプライ：** 特定ユーザの特定メッセージに対して返信すること。
- **RT(ReTweet)：** 特定ユーザの特定メッセージを引用し、再投稿すること。
- **DM(DirectMesssage)：** 特定のユーザに非公開のメッセージを送ること。

2.2. 知識獲得時の課題

知識を獲得するには、まずは大量の MB ストリームから、特定事象に関する有用なメッセージを見つける必要がある。ある文章の内容が特定事象に関連するかを判定するには、特定事象を表現した特長ベクトルを用いた類似度判定が一般的である。しかし、Twitter のメッセージは140 文字以下と短いため、文章判定を行うのに十分な情報が無いという問題がある。そのため、有用メッセージの取りこぼしが発生しやすい。

そこで、メッセージを個別に判定するのではなく、なんらかの関係のある一連のメッセージを集約した複合メッセージを生成し、情報量の多いそれに対して文書判定を行うという方式が考えられる。複合メッセージは内容的に関係のあるメッセージから成ることが理想だが、そもそも単体メッセージを内容判定できないため、それはできない。よって、内容に基づかず関係のあるメッセージを特定し、適切な複合メッセージを生成し、判定する方式を提案する。

3. 提案方式

提案方式の概略図を図1に示す。

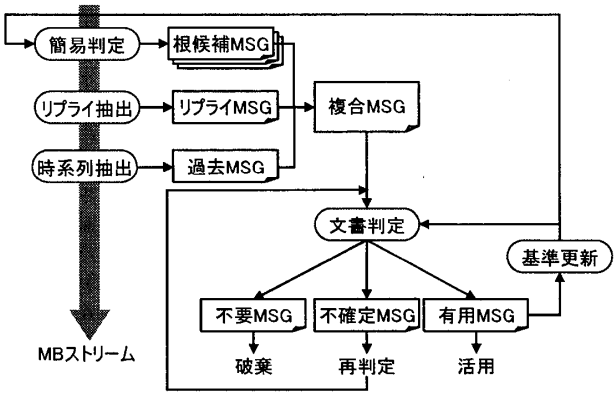

図1： 提案手法概略図

内容に基づかず関係のあるメッセージを特定するために、我々は Twitter の関係構造に着目した。すなわち、関係構造的に繋がりがあるメッセージは、内容的にも関係があると仮定し、それら繋がりのあるメッセージを用いて複合メッセージを生成する。用いる関係構造は、以下の理由からリプライを採用する。

- **×フォロー/フォロワ：** メッセージ内容に依存した構造でないため有効でない。
- **×RT：** RT 元のメッセージの内容を自メッセージ内に内包するため、既に複合メッセージであり、特別な処理は必要ない。
- **○リプライ：** Tweet の約 40%はリプライによる会話であるという報告もあり有効。
- **×DM：** 非公開のため利用しにくい。

Harvesting Social Knowledge from Micro-blogging Streams.
Ken-ichoro FUJIYAMA, Koji Zettsu and Yutaka Kidawara.
Knowledge Creating Communication Research Center, NICT.

さらに、リプライ構造に加え、メッセージの投稿時間、投稿ユーザという属性も利用する。これは同じユーザが近い時間に投稿したメッセージも内容的に近いという仮定に基づく。すなわち、時間的に近いメッセージも用いて複合メッセージを生成する。

各複合メッセージの生成方法の詳細は以下のとおり。

リプライ： 各メッセージはユニークな ID を属性として持ち、リプライしたメッセージは、リプライ元の ID も持つ。よって、リプライが発生したら、元 ID を辿ることで複合メッセージを生成する。ただし、全てのメッセージがリプライされ複合メッセージの根となる可能性があるため、それらを記憶しておく必要がある。しかし、MB ストリームは永続的に発生するため、全てのメッセージを記憶しておくことはできない。よって、根候補となるメッセージは、基準を緩やかにした簡易な文書判定…例えば特徴ベクトルが 1 つでも一致すれば有用、ただし false positive が多い…によってある程度の選抜を行う。また、一定期間リプライされなかった場合は破棄する。

投稿時間： 現在のメッセージを基点とし、同一ユーザの過去一定時間内のメッセージをまとめて複合メッセージとする。複合メッセージ生成には過去一定時間内のメッセージのみをバッファしておき、古いメッセージは破棄して良いため、永続的な MB ストリームを対象としても無限の記憶量が必要とならない。ただし、バッファしたメッセージに対し、リプライと同様の簡易文書判定を行い、まったく関係のないメッセージは事前に破棄する。なぜなら、時間に基づく関係は、リプライ関係に比べ、関係のないメッセージが含まれる可能性が高いからである。

上記のようにすることで、より文章判定に向いた複合メッセージを生成することが出来る。

なお、複合メッセージを生成したら、そのメッセージ全体に特徴ベクトルを用いた文書判定を行う。判定結果は類似度が一定値以上であれば特定事象に関係したものであるため有用、一定値以下であれば不要、その間であれば不確定とする。有用と判定された場合、複合メッセージを構成する元メッセージ全てを有用なメッセージとする。その際、その複合メッセージから新たな特徴ベクトルを抽出し、文書判定および簡易判定の基準を更新する。不要判定の場合はメッセージを破棄する。不確定の場合は基準が更新された状態で再度文書判定を行う。

4. 評価

提案方式を評価するため、サンプル MB ストリームに対し、どれだけ特定事象に関連するメッセージを抽出できるか実験した。特定事象はある海難事故であり、サンプルはそれに関するメッセージを含む 22 ユーザによる 1273 メッセージである。上記サンプルに対し、以下の 3 つの方法で関連メッセージの抽出を行い、比較する。

1. 人手で関連するメッセージを特定
2. 各メッセージを個別に文書判定
3. 複合メッセージを生成した上で文書判定

なお、2、3 に関しては、特定事象を表す特徴ベクトルは事前に人手で設定し、有用メッセージによる判定基準の更新は行っていない。実験結果を表 1 に示す。

表 1 ：実験結果

	抽出数	正解数	再現率	適合率
1. 人手	78	78	100%	100%
2. 個別	23	22	28%	96%
3. 複合	47	36	46%	76%

抽出数は特定事象に関連する有効メッセージと判定した数、正解数はそのうち、本当に関連するメッセージの数である。なお、ここでは 1. 人手の抽出結果である 78 メッセージを本当に関連するメッセージとしている。

以上の結果から提案方式は適合率は落ちるが、再現率が上昇しており、関連メッセージの取りこぼしが少ないということが分かる。

5. おわりに

本稿では MB ストリームから特定事象に関するメッセージの抽出、特に取りこぼしを少なくする手法を提案した。提案手法により再現率の上昇が確認された。一方適合率は下がるため、提案手法は不要なメッセージを大雑把に除去するフィルタとして向いていると考えられる。

今後は、未評価である判定基準の更新等の実装、評価を行う予定である。

謝辞

本研究の一部は文科省科研費補助金特定領域研究 A01--24、課題番号 2101305（代表：木俵豊）による。ここに記して謝意を表す。

参考文献

[1] twitter, http://twitter.com
[2] 藤山 他, “Twitter を利用した実世界アノテーションの検討”, 第 2 回 UC 研究会, 2010.
[3] Jagan S. et al, “TwitterStand: News in Tweets”, 17th ACM SIGSPATIAL, 2009.