Twitter 上の画像転載の検出と検索サービスの作成

0811021　小貫 翔　　岩崎研究室

1　背景

Twitter は優れた速報性を持つソーシャルメディアとして広く利用されている。一方で Twitter 上ではツイートや画像の転載が横行している。転載ツイートは信憑性という点ではほとんど価値がなく、すなわち、ツイートが転載か否かを確認することは、信憑性の検証として有用である。さらに、転載ツイートから元ツイートを発見できれば、ユーザーはより信憑性が期待できる情報を入手できる。

ツイート本文の転載であれば既存の検索方法で容易に検証可能である。しかし、画像転載時に本文が変更されると、転載元ツイートを見つけることは困難である。例えば Twitter 公式の検索では画像を用いてツイートを検索できず、Google 検索はキーワードや画像を用いて類似画像を含むページを検索できる一方で、Twitter 内の各ツイートはほとんどヒットせず、Twitter 内の検索は困難である。画像付きツイートは影響力が大きい傾向にあり、また、画像は説明文によって解釈が大きく変わりうることから、パクリやデマといった悪意ある行為に利用されやすい。

2　目的と方針

本研究では画像付きツイートに対して、類似画像を含むツイートを検索できるサービスを作成することを目的とする。これにより、画像転載に対する元ツイートを容易に発見できるようになり、画像付きツイートの信憑性を検証しやすくなることが期待される。

本サービスは不特定多数による利用を想定し、また、ツイートは常時取得する必要があるため、Web サービスとして提供する方針である。Twitter 全体を検索できるのが理想だが、ツイート数が極めて多い上、そもそも全ツイートを取得する方法は一般公開されていない。転載元ツイートの発見という目的と負荷の兼ね合いから、検索対象は登録ユーザーのタイムラインに表示されるものと、お気に入りやリツイート数の多いものとする。

3　予備実験

3.1　実験内容

画像転載の実態把握および、転載元ツイートの発見の有用性を検証するために実験を行った。自分の Twitter アカウントのタイムラインに流れてくる画像を収集し、類似画像をグループ化した。収集期間は 9/6 から 9/11 である。類似画像の検出には DCT Hash を用いた。DCT Hash の特徴を以下に示す。

- 局所性鋭敏型ハッシュ (似た画像には似た値を返す)
- ハッシュ値は 64bit と短い
- 計算が速い

3.2　実験結果

画像付きツイートは 9966 個、総画像数は 13386 枚であった。また、類似画像のグループは 199 個、グループ内のツイートは延べ 497 個であった。各グループは以下のように分類できた。

- パクリと見なせるもの: 147 グループ
- 関連アカウントによる再投稿: 24 グループ
- 画像に対する議論: 12 グループ
- 画像ジェネレーターによるもの: 6 グループ
- Tweet Button 等による同一画像: 5 グループ
- 似ているが無関係な画像: 6 グループ

類似画像の閾値は、見かけ上同一の画像を取りこぼさない程度としたため、画像ジェネレータのように画像の一部を任意に改変する性質の物は取りこぼしが多い。また「似ているが無関係な画像」はそれぞれ同一アプリのスクリーンショットであった。

類似画像のグループ化によって、パクリツイートから元ツイートを発見したり、怪しい画像に対する議論を確認したりできることが確認できた。ただし、この実験では収集対象が限られているため、転載元ツイートを正確に提示できているわけではない。

4　関連研究

pHash[1] は、画像用の局所性鋭敏型ハッシュを生成するアルゴリズムを複数実装したライブラリである。本研究で使用した DCT Hash はこれに含まれている。また、DCT Hash のペアから画像の類似度を判定するには、ハッシュ値のハミング距離を計算すればよい。本研究ではハミング距離が一定以下の全ペアを高速で列挙するアルゴリズムである、複合ソート法 [2] を採用した。

5　現状と今後

現時点で完成しているものは予備実験が行える程度の内容である。具体的には

- 1 ユーザーのタイムラインを取得し、画像を収集
- 類似画像を検出し、全ペアを列挙
- 類似画像を含むツイートをグループとして表示

上のような処理が可能である。今後は Web サービスとして利用可能にするべく、複数ユーザー対応化、表示方法の改善、検索機能の実装を行う予定である。ユーザーによる評価も行う。

参考文献

[1] pHash http://www.phash.org/
[2] 宇野毅明, An Efficient Algorithm for Finding Similar Short Substrings from Large Scale String Data. In Lecture Notes in Artificial Intelligence 5012, pp. 345-356, 2008