図書に言及するツイートのクラスタリング分析による類型化

◯矢田 竣太郎, 影浦 峡（東大 教育学研究科）
言語理解とコミュニケーション研究会（NLC）

2015年6月5日

本発表の流れ

1. 研究目的

2. 実験方法

3. 分析結果

4. 課題と展望

1. 研究目的

1. 研究目的

(1) 図書に言及するツイートとしての「書名を含むツイート」を集め、図書の言及様式を調べたい

(2) 図書への言及から得られる付加情報（文脈・状況）に何があるか知りたい

なぜ図書に言及するツイートを扱うのか①

- 前読書家の読書を触発する図書推薦システム ***Serendy*** の開発

➡ **前読書家：「何を読めばいいかわからない」人**
（能動的探索へと踏み出す適切な外圧・刺激を得られていない）

➡ **触発の属性：日常性・近接性・非強迫性・誘引性**

なぜ図書に言及するツイートを扱うのか②

そもそも図書に言及するツイートとは？

どんな特徴があって何種類あるのか？

具体的な言語表現として類型化したい

類型化のために

- まず図書に言及している可能性の高い「書名を含むツイート」を収集し、個別に吟味しながら類型化への第一歩を踏み出す
- 書名を固有表現と捉えれば、将来的には機械学習によるアプローチが有効
    - 学習のための正解データとしても類型化したい

類型化の観点

- **観点1) 書名(を含むツイート)抽出**
  - ➡ 差異化表現（かぎかっこ）
  - ➡ Twitter 固有の表現
  - ➡ その他の書誌情報
  - ➡ ツイート方法
- **観点2) 言及の触発性**
  - ➡ 図書言及の文脈・状況

2. 実験方法

固有表現としての書名の特徴

- 一般的な文章表現との差異に乏しい
  - ➡ 『あしたもね』『ご冗談でしょう、ファインマンさん』『朝バナナダイエット』
- メタな差異化表現がルール化されている
  - ➡ 『』、斜体
- 全てのエントリを網羅することができる
  - ➡ 日本では国立国会図書館が網羅

2. 実験方法

1）形態素解析用の書名辞書を構築

2）日本語ツイートのStreamingから書名を含むツイートをフィルタ

3）形態素を素性としてK-meansでクラスタリング

4）結果を目視で精査し、図書に言及するツイートを類型化

1) 形態素解析用の書名辞書を構築

- Webcat Plus* に収録された書名を固有名詞として辞書をコンパイル（MeCab向け）

* NII 阿辺川准教授の協力による

- 本として言及されにくい書名を除く1421556件

** S. Toshinori https://github.com/neologd/mecab-ipadic-neologd

➡ Neologd**（日本語の固有表現で比較的新しいものを意欲的に収録した辞書）に含まれる書名

➡ 記号、年月日、1形態素のみの書名

2) 日本語ツイートの Streaming から書名を含むツイートをフィルタ

- 書名辞書を用いてツイートを形態素解析

- Streaming から得る理由

➡ いま現在話題になっている図書が言及されやすい：Serendy の実際の挙動に沿う

➡ 検索と違ってアクセス制限がない

3) 形態素を素性として K-means でクラスタリング

- 書名の周囲の表現が似ているツイート同士をまとめる

- クラスタリング実行対象の制限

  ➡ 書名を1つだけ含むとされたもの

  ➡ 出現頻度2以下の書名

素性付けの方法

- ✦ 形態素を書名の前後で区別するため「+」「-」を付与
- ✦ 前後3つ分をさらに区別するため距離を付与

今は中学生の時に買った、『ボーイズラブ小説の書き方』とか言う教材を読んでる。

```
[
  "今-", "は-", "中学生-", "の-", "時-", "に-", "買う-",
  "た-3", "、-2", "『-1",
  "』+1", "とか+2", "言う+3",
  "教材+", "を+", "読む+", "でる+", "。+"
]
```

4) 結果を目視で精査し、図書に言及するツイートを類型化

- 観点1) 書名(を含むツイート)抽出
  - ➡ 差異化表現（かぎかっこ）
  - ➡ Twitter 固有の表現
  - ➡ その他の書誌情報
  - ➡ ツイート方法
- 観点2) 言及の触発性
  - ➡ 図書言及の文脈・状況

3. 分析結果

統計的な情報

- 2015年4月30日12:00–5月5日14:45の期間　74330 件
- 図書が1件だけ含まれるとされたツイート　70844 件
- 出現頻度2以下の書名を含むツイート　10791 件

- k = 25 で K-means を実行

クラスタに含まれるツイート数

平均値	標準偏差	最小値	下部四分位数	中央値	上部四分位数	最大値
431.6	386.2	103.0	196.0	298.0	550.0	1870.0

- 目視の結果得られた図書に言及するツイート　211 件

観点1) 書名(を含むツイート)抽出

(1)差異化表現（かぎかっこ）

(2)Twitter 固有の表現

(3)その他の書誌情報

(4)ツイート方法

観点1) 書名(を含むツイート)抽出(1)

- 差異化表現（かぎかっこ）

記号	出現数
「」	33
『』	47
【】	50
なし	81
	211

「もし僕がヒットラーの子供だったら、戦争を止められただろうか?」をテーマとして展開した児童書「ヒットラーのむすめ」はすげー名作だと思う。 http://t.co/0SbdRKI8PS

【速読の基本が面白いほど身につく本 (ポイント図解)/呉 真由美】を読んだ本に追加
→http://t.co/AbuOp2804Z
#bookmeter

ダニッチの怪読み終わったー

観点1) 書名(を含むツイート)抽出(2)-1

- Twitter に固有の表現

種類	出現ツイート数
URL	101
Hashtag	66
Reply	27

【目隠し姫と鉄仮面〈1〉 (レジーナ文庫)/草野瀬津璃】【再読】やさぐれた心を落ち着かせたくて(笑)再読。
のほほんとしていて、やっぱり好きなお話でした(^w^) →http://t.co/zad2bHWPck #bookmeter

@Haiena0303 ありがとうただいま！！！
商業用なら私は桜日梯子さんの『年下彼氏の恋愛管理癖』をめっっっっちゃオススメする！！！！！

観点1) 書名(を含むツイート)抽出(2)-2

- URL リンク先ドメイン（出現頻度2以上）

リンク先	出現頻度
bookmeter.com	52
twitter.com	18
www.amazon.co.jp	10
booklog.jp	3
www.ohtabooks.com	2
www.google.com	2

twitter.com は写真・画像

気になってた「ずかん文字」を買ってきた。読んでいてわくわくする内容で、かなり読み応えもある。全ページカラーなのに高くないがイイね。

写真を含むツイートの例

観点1) 書名(を含むツイート)抽出(2)-3

- URL リンク先ドメイン（出現頻度2以上）

リンク先	出現頻度
bookmeter.com	52
twitter.com	18
www.amazon.co.jp	10
booklog.jp	3
www.ohtabooks.com	2
www.google.com	2

twitter.com は写真・画像

書誌情報のスクレイピングが可能

観点1) 書名(を含むツイート)抽出(3)-1

- その他の書誌情報（全ツイートの総計）

書誌情報	出現頻度	bookmeter除く
著者	84	36
叢書	43	3
副書名	19	7
著者略	15	15
巻数	12	7
出版社	5	5

最も多いのは著者

おおむね連携アプリ経由

マンガへの言及に多く見られた

観点1) 書名(を含むツイート)抽出(3)-2

- 著者の例

『ここを過ぎて悦楽の都』平山瑞穂著、読了。なんか唐突に終わってしまって、おいてけぼりを食った気分。ちゃんと説明してほしい。。

- 著者の省略例

田亀先生の「君よ知るや南の獄」読破！

ラストにねー、思わず涙が…
たった一言を言え無かった二人の関係にグヌヌってなります；＿；

- 巻数を併記する例

フラッと本屋寄って、クズの本懐1～3、ダメな私に恋してください6、ナナとカオル15買った。クズの本懐は前に友達に借りて読んですきになったから4巻だけ家にあるよ。

観点1) 書名(を含むツイート)抽出(4)

- ツイート方法（クライアント）

クライアント	出現頻度
読書メーター	50
Twitter for iPhone	40
Twitter Web Client	38
Twitter for Android	23
iOS	6
ついっぷる	4
ブクログ(booklog.jp)	3
Twitter for iPad	3
TweetDeck	3
Mobile Web (M2)	3

Not tweets mentioning books

client	count
Twitter for iPhone	2256
Twitter for Android	1189
Twitter Web Client	978
IFTTT	534
dlvr.it	375
autotweety.net	121
twiroboJP	107
篠山莉沙system	94
iOS	82
twicca	79

スパムの排除に有効な指標となる

観点2) 言及の触発性

➡ 図書言及の文脈・状況

観点2) 言及の触発性

- 図書言及の文脈・状況

複数に含まれうる場合は第一義的な方に振り分けた

	感想共有	行動報告	推薦・同調	期待表明	参照
ツイート数	57	80	35	11	28
総トークン数	2686	1670	1176	283	1227
平均トークン数	47.1	20.9	33.6	25.7	43.8

森見登美彦「四畳半王国見聞録」読了。なんてバカなんだ。なのになんで最後ちょっと感動するんだ。電車で読むとニヤニヤしたりウルウルしたりで大変だった。面白かった。

ナイス >> 10の短編で出来ている「六月の夜と昼のあわいに」独特な世界が味わえる恩田ワールド満載の作品です。とても不思議な話しばかりだけど、その中でも窯変・田久保順子」と「Y字路の事件」が気に入りました。H27.191 http://t.co/egtMLD3MAo

新幹線で読んでた残留思念捜査、さくっと読めたしおもしろかった

感想共有のツイート例

比較的長文で、形容詞のタイプが豊富

(形容詞: 34タイプ, 54トークン/他は10タイプ, 20トークン程度)

【働くことがイヤな人のための本 (新潮文庫)/中島 義道】を読んだ本に追加 →http://t.co/DAEtGuAw7V #bookmeter

「気分障害ハンドブック」という本をポチってみた。

あかね色シンフォニア届きました

これからラヴクラフト全集4読む

行動報告のツイート例

動詞「読む」の出現頻度が高い

(58トークン/次点は感想共有の20トークン)

@amacro_36 和訳が難解すぎてなかなか頭に入らんかったような・・読めだすと面白いんだけどねー。
ドラゴンランス伝説がなんか好き。覚えてないけど^_^

@sanagi_iganas そんなさぁちゃんに立原えりかさんの「小さな花物語」をおすすめしませう(お疲れ様ー)

レヴィの『ヴァーチャルとは何か？』マストバイ。

推薦・同調のツイート例

指示語（連体詞）が多い（7タイプ, 10トークン/

感想共有で5タイプ, 9トークン)

大好きな作家、ミュリエル・バルベリが新作を出したらしい。でもフランス語版しかないようで、日本語訳が出ないかと昨夜から悶々としている。「優雅なハリネズミ」の河村真紀子さんが訳してくれないかなと、これまた悶々と考えている。この河村さんの訳がとてもいい！朝からモンモンする。

「僕の好きな人が、よく眠れますように」
「星に願いを、月に祈りを」
も欲しい！！

わああん花は咲くか終わっちゃったよう(´・ω・`)

期待表明のツイート例

助動詞のタイプ組成が他と異なり、「たい」「う」など未来志向のものが上位に含まれる

岡崎久彦さんの『国家と情報』にはっきり書いてありますけれど、日本の地理的条件は、米中露の結節点であり、そのどこかと同盟するしか選択肢はなく、そのうち一番合理的なのは米国なのは衆目の一致するところ。まず、一番のジャイアンである、そういう話です。

日比谷公園西側の緑化道路でカラタネオガタマ（唐種招霊：モクレン科オガタマノキ属）が咲いていました。『樹に咲く花』によると中国原産で江戸時代に渡来し庭や神社に植栽されたそうです。花からバナナのような甘い香りが漂ってきました。美味しそう！ http://t.co/RNb1t5d8yC

参照のツイート例

感想共有と似た形態素組成だが、動詞「読む」のトークン数が少なく、一般的な動詞（する、いる、ある）が多い

4. 課題と展望

分析によって得られた知見

- 観点1) 書名(を含むツイート)抽出
  - ➡ 差異化表現（かぎかっこ）　　差異化されずに言及されうる
  - ➡ Twitter 固有の表現　　URLリンク先の解析が図書の同定に有効
  - ➡ その他の書誌情報　　著者(省略含め)が同時に言及されやすい
  - ➡ ツイート方法　　スパム排除に有効
- 観点2) *Serendy* への応用
  - ➡ 図書言及の文脈・状況　　大きく5種類に分けられる

課題と展望

- 課題
  - ➡ 言及頻度の高い図書ではどうか
  - ➡ 図書への言及が複数ツイートに亘る場合はどうか
- 展望
  - ➡ 簡素な分類器の実装によるツイート収集の効率化
  - ➡ 5種類の文脈における非強迫性と誘引性のタグ付け