

位置及び投稿者希少性に着目した 犯罪関連投稿の抽出方法

高橋 柊 菊池 悠 深澤 佑介

平井 遥斗

富山県立大学 情報システム工学科

2022 年 11 月 1 日

背景

近年，ソーシャルメディア上のデータを用いた社会センシング技術の研究 Human Probe(HPB) が活発化している。



<https://human-probe.info/>

本研究の目的

Twitter に投稿されている犯罪関連の投稿の抽出に焦点を当て，位置及び投稿者希少性を用いた投稿抽出手法について検討する。

ステップ1

■ 犯罪語マッチ処理

表 1: 犯罪語辞書 (一部抜粋)

犯罪語	ルビ
痴漢	チカン
ナンパ	ナンパ
覗き	ノゾキ
置き引き	オキビキ
キャッチ	キャッチ
盗撮	トウサツ

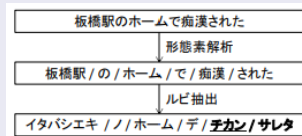


図 1: キーワードマッチングの実例

ステップ2

■ 投稿者希少性を考慮した特徴量計算 ($tf \cdot idf$ 法)

$$user_score(T, d) = tf(d, T) \cdot \log \frac{N}{df(T)} \quad (1)$$

ステップ3

- 位置希少性を考慮した特徴量計算

P 犯罪投稿群における単語の出現分布

Q 過去投稿群における単語の出現分布

$$poi_score(P, Q) = D_{JS}(P \parallel Q) \quad (2)$$

ステップ4

- 投稿フィルタリング

SVM を用いて分類器を構築する. (Bag-of-words)

→ その際, 特徴量に位置及び投稿者希少性を考慮した特徴量を追加することで, 犯罪関連投稿抽出を行う.

提案手法の有用性について

- 犯罪語を含む投稿をランダムにサンプリング

True 犯罪関連投稿である投稿

False それ以外

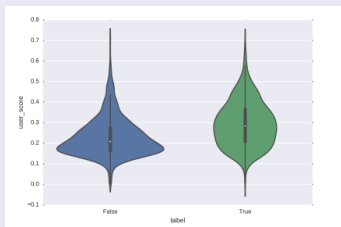


図 2: *user_score* の分布
[False/True]

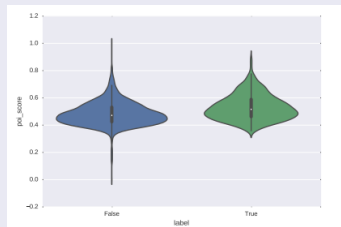


図 3: *poi_score* の分布
[False/True]

既存手法 (Bag-of-words のみ) との比較

- 既存手法と提案手法を用いた分類器の精度を比較する

表 2: 学習器のスコア

	label	適合率	再現率	F
既存	False	0.89	0.70	0.79
	True	0.37	0.67	0.47
提案	False	0.90	0.71	0.80
	True	0.39	0.71	0.50

まとめ

位置及び投稿者希少性を用いることで分類精度を向上させることができた。