位置および投稿者希少性に着目した 犯罪関連投稿の抽出手法

富山県立大学　工学部　情報システム工学科　　2015044　　平井　遥斗

1 はじめに

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．本研究では, Twitter に投稿される犯罪関連投稿の抽出 に焦点をあて、位置およ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．Twitter より投稿者が経験あるい は目撃した犯罪関連投稿を抽出することができれば, 警察 によるパトロールなど既存のセンシングでは顕在化困難で あった犯罪事象についての知見を得ることができる．

2 提案手法

提案手法では, Twitter の投稿情報より投稿者が犯罪を 経験あるいは目撃した投稿を抽出することで, 警察官のパ トロールなど既存のセンシングでは顕在化が困難であった 犯罪関連投稿を抽出する。 図 1 に提案手法における分析ステップを示す．

i) 犯罪語マッチ処理 Twitter より取得した投稿および, 犯罪関連語を含む 犯罪語辞書を用い, 投稿テキスト中に犯罪語辞書の単語を含む投稿を抽出する．

ii) 投稿者希少性を考慮した特徴量計算 犯罪語マッチより得られた投稿の投稿者が, 過去に同様の犯罪事象について投稿しているか否かを計算する．

iii) 位置希少性を考慮した特徴量計算

日本において, 同一人物が数ヶ月間に複数回犯罪を経験 あるいは目撃することは少ないそのため, あるユーザが 複数回犯罪に関する投稿をしている場合, それはユーザの 経験あるいは目撃した投稿ではなく, ニュースなどの伝聞 情報やノイズである可能性が高い。 例えば, 「泥棒」という 単語を高頻度で投稿している投稿者は, 泥棒を経験あるい は目撃しているのではなく, ゲームや映画など非現実世界 における事象について投稿している可能性が高い. また, 「警察」 という単語を高頻度で投稿している投稿者は, 警察 を目撃したのではなく, 警察関連組織における, 公式アカ ウントである可能性が高い. 一方で, 普段 「泥棒」という単語を投稿していない投稿者が 「泥棒」 という単語を投稿 した場合、投稿者が泥棒を経験あるいは目撃した可能性が 高いといえる．

文章分類問題において, テキストを Bag-of-Words で表現 し, 分類器を構築する手法が広く使われている．Dilrukshi ら [10] は Twitter の投稿テキストをラベリングし, Support Vector Machine (SVM) を利用することで, 投稿テキスト を分類する手法について検討している SVM を利用する ことで多量の特徴量を評価した分類器の構築が可能となる ことを示している．

3 数値実験

Twitter より提供されている, Public streams*3 を用い, Twitter の日本語投稿データを取得した。 データ取得期間 は 2015 年 1 月 1 日から 2015 年 9 月 31 日までの 9ヶ月間 であり, 日本語投稿および Twitter 公式のクライアント を利用して投稿されたツイートを対象とした．犯罪語辞書 を用いたキーワードマッチングにて, 犯罪語を含む日本語 ツイートのうち, 3600 件をランダムサンプリングし、犯罪 関連投稿である投稿を True, それ以外を False としラベ リングした．

4 おわりに

本研究では、 位置および投稿者希少性を用いることで、投稿者希少性を比較した際, 分布に差が出ることを示した。 評価対象データを bag-of-words で表現したのち, PISA を用い次元縮約を行い, SVM を用いて犯罪投稿および非 犯罪投稿を分類する分類器を構築した．SVM の特徴量に bag-of-words のみを用いた従来手法における分類精度と, 提案手法を取り入れた分類精度を比較することで, 提案手 法の有用性について評価した。