E-037

blog 記事からの不審者情報の抽出と分類

Extraction and Categorization of Suspicious Behavior Information from the Blogspace

石井 基一† Motokazu Ishii　**山田 剛一**‡ Koichi Yamada　**絹川 博之**† Hiroshi Kinukawa

１．はじめに

近年，blog が個人の情報発信源として注目されている．blog は作者の個人的な体験や日記，時事問題などについて書かれている，時系列で記録される Web サイトである．blog 記事から，記述された実世界の出来事を抽出することで，ある特定の場所についての情報が得られる．このような場所の情報はその地域の周辺住民や，その地域へ行く人にとって有意義な情報になると考えられる．

本研究では，実世界情報として，近年，問題視されて対応策の検討が活発化している不審者に関する情報を対象とする．そして blog 記事から情報を抽出および分類を行い，最終的に地図上へマッピングし，検索も行えるようにすることを目的としている．

公的機関等が不審者情報を地図で提供していることがあるが，更新が遅いことが多い．blog 等へのテキストでの情報提供は比較的早く行われるので，その情報をマッピングできれば，地図での情報提供が早くできるようになると考えられる．

本研究では，「不審者」という表現を含む blog 記事を一記事でも含む blog サイトの全 blog 記事を対象としている．blog 記事から不審者情報を抽出し，抽出された場所情報をもとに地図へマッピングを行い不審者情報の提供を行う．

２．不審者情報を含む blog 記事

「不審者」で検索を行った場合，不審者情報が書いてある記事とそうでない記事とが見つかる．不審者情報でない記事には不審者情報についての意見や感想を書いた記事や，不審者対策や訓練について書かれているものがある．また，blog 著者やその身近な人のことを不審者といって嘲笑的に使われている記事もある．実際の不審者について書かれている記事でも，場所の情報が書かれていないこともある．何が起きたかがわかっても，どこで起こったのかがわからなければ，その情報を見る人にとって有益な情報とはならないと考えられる．

2.1 不審者情報の発信者

有益な不審者情報を含んでいる記事には警察や学校，教育委員会等の公的機関により発信されているものと，個人の体験・見聞を書いた日記のような記事とがある．公的機関の発信する情報は詳細で定型的であることが多い．また，表にまとめられていることもある．個人の情報では，実際に体験したことであっても，自身の個人情報となる場所を特定できる表現が含まれず，地域特定が不可能なことがほとんどである．

† 東京電機大学大学院
‡ 東京電機大学/JST-CREST

2.2 不審者情報の項目

不審者情報を含む blog 記事の例を図 1 に示す．不審者情報の記事に書かれている内容を項目ごとに分類すると，次のようになる．

- 場所
- 日時
- 不審者の特徴
- 分類
- 状況

不審者情報を項目に分けて抽出することで，情報として利用しやすくなる．なお，この中で「場所」と「状況」の情報が含まれない記事は対象としない．

場所情報としては地名，ランドマーク名を抽出し，一意に特定する．ランドマークとは，施設や建物等の場所を特定する際に地名と同等の意味を持つものとする．日時は記事中に記載があればそれを抽出し，無ければエントリの更新日時とする．不審者の特徴は不審者の性別や体格，服装などについて書かれた情報を抽出する．分類は，「声掛け」「つきまとい」等の種別を特定し分類とする．状況では実際に起きたことについての情報を抽出する．

> ○月○日午後○時○分ころ、○区○3 丁目 3 番付近路上で、女子小学生が白っぽいバイクに乗った○～○歳くらいの男に、手を握られ、手を振りほどいたところ、わいせつな声をかけられた。

図 1. 不審者情報を含む blog 記事の例

３．blog 記事に出現する場所の表現

blog 記事において，場所の表現は地名やランドマーク名の形で出現する．場所の情報として扱うには，それら地名やランドマーク名から場所を特定する必要がある [1]. ランドマーク名には郵便局，小学校，マンション等の一般名詞が含まれていることも多いため，それらを含めた複合語としてランドマーク名であると認識する必要がある．また，ランドマークと，緯度経度による位置情報との対応付けが必要である．

3.1 ランドマークの認識

ランドマークを場所の情報として特定するためには，まず，記事中の表現がランドマークを指す表現であると認識しなければならない．記事中のランドマーク名は形態素解析辞書に登録されているものとそうでないものがある．登録されているものは固有名詞として認識できる．登録されていないランドマーク名は，一般名詞やカタカナ語や未知語から複合語が構成されることが多く，また，これらの語と地名や組織名などの固有名詞とともに複合

語が構成されることも多い．この固有名詞との共起を利用し，ランドマーク名を構成する一般名詞を，ランドマーク構成語として辞書登録を行う．これにより，固有名詞を伴わないランドマーク名でも，ランドマーク構成語と複合語を構成していれば認識できるようになる．また，これらの辞書に登録されていないカタカナ語・未知語もランドマーク名の候補とする．

・ 形態素解析辞書に固有名詞として登録されている語
・ ランドマーク構成語
・ 辞書登録されていないカタカナ語・未知語

これらの語とその複合語を Yahoo!地図情報[2]のローカルサーチ API 等に投げ，ランドマークとして位置情報が返ってくれば場所情報として抽出する．

3.2 記事に書かれている出来事の場所の同定

対象としている blog 記事はある特定の地域を想定して書かれていることが多く，想定された地名は省略することが多い．また，地名には都道府県，市町村といった階層があり，記事に出現する地名の表現レベルは様々で粒度の揺れが存在する．また，日本全国には，同名の地名，ランドマーク名は多数存在し，それだけで場所を判別することはできないことがある．このような同名の存在する地名については，同定をして曖昧性を解消する必要がある．複数存在する地名は，その地名，ランドマークの属する上位階層の地名を取得すれば一意に定められることが多い．なお，ランドマークは同じ地域内に複数存在することもあるが，blog 記事で想定された範囲内に複数存在するものは，あらかじめ曖昧にならないような区別された表現を用いると考えられる．

同名の曖昧性解消には，blog サイト内に出現する地名，ランドマークの緯度経度を用いる．blog サイト内の地名，ランドマーク名から Yahoo!地図情報 ローカルサーチ API で緯度経度を取得する．この際，曖昧性のあるものは複数の結果が返ってくる．blog サイト内の地名，ランドマーク名の複数についての緯度経度の候補を取得し，その緯度経度が集中している所を blog で想定している地域とする．そこに最も近い緯度経度の地名，ランドマークを正解として同定する．

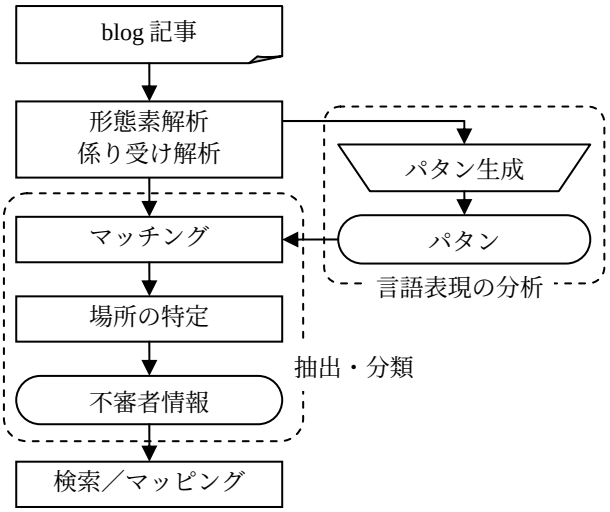

図 2. 不審者情報抽出・分類・提供システムの流れ

4．不審者情報抽出・分類・提供システム

本手法では， blog 記事から不審者情報を抽出し，抽出された場所情報をもとに地図へマッピングを行い不審者情報の提供を行う．不審者情報抽出・分類・提供システムの流れを図 2 に示す．

4.1 言語表現の分析

項目別の表現の特徴を知るため，あらかじめ抽出する項目の情報を含んだ文の集合に対して，形態素解析による，形態素の基本形，品詞，活用形等の情報と，係り受け解析による係り受け関係を手がかりに，出現する文章の表現の分析を行う．係り受け解析には CaboCha[3]を使い，形態素解析結果としては CaboCha の解析結果に付く ChaSen[4]のものを利用した．その結果に対して意味的タグ付けを行い，木構造の XML 文書にした．これには GDA [5]の SDK が生成する構造に並列，同格のタグを追加したものを利用した．

抽出する項目ごとの抽出パタンの生成には，形態素の出現数およびその係り受け関係などを利用する．

4.2 抽出・分類

「不審者」という表現を含む blog 記事を一記事でも含む blog サイトの全 blog 記事を対象として，形態素解析及び係り受け解析をかける．その結果に対して，パタンを基にマッチングを行い，抽出を行う．

分類は記事に明示的に書かれているものがあればそれを抽出し決定する．書かれていなければ状況の記述から分類を決定する．

4.3 検索・地図へのマッピング

抽出した不審者情報は各項目をクエリとして検索し，結果を地図上へマッピングする．地図上の場所情報の位置にマーカを置き不審者情報を提供する．マッピングは抽出で得られた場所情報を元に，Google Maps API[6]を用いて Google Maps に行う．

5．おわりに

blog 記事から不審者情報を抽出し，抽出された場所情報をもとに地図へマッピングを行い不審者情報の提供を行う不審者情報抽出・分類・提供システムの提案をした．今後，実装を進め評価を行う．

参考文献

[1] 金木雄太, 山田剛一, 絹川博之, 中川裕志: 地名辞書を利用した地名の曖昧性解消と文書の場所分類, 第 19 回人工知能学会全国大会論文集,2E1-03 (2005).
[2] Yahoo!デベロッパーネットワーク Yahoo! 地図情報 http://developer.yahoo.co.jp/map/
[3] 日本語係り受け解析器 CaboCha http://chasen.org/~taku/software/cabocha/
[4] 形態素解析システム ChaSen http://chasen.naist.jp/hiki/ChaSen/
[5] 大域文書修飾 Global Document Annotation (GDA) http://i-content.org/gda/
[6] Google Maps API http://www.google.com/apis/maps/