1D-3

地域に関する意見収集のための Web 上の記事を用いたイベント追跡システム

平田 紀史 † 白松 俊 † 大囿 忠親 † 新谷 虎松 †
名古屋工業大学大学院工学研究科情報工学専攻 †

1 はじめに

特定の地域に関する意見を収集できれば，その地域の住民や行政による意見活用が可能となる．例えば，同じ地域で類似する意見を持った人の発見や，政策立案の根拠としての活用ができる．本稿では意見収集の補助として，特定の地域に限定してイベント追跡を行うシステムを提案する．提案システムでは，Web 上において，地域に関する意見収集を行う状況を想定する．ユーザは自分の意見を入力する前後に，関連情報を得ることで，自分の意見を修正，補強することができる．意見収集の対象となる事象に関して，理された情報の提示が，意見収集の補助となる．

意見収集の補助のために関連情報を提示すると，入力される意見がその情報によって影響を受ける．ここで，既に入力された他人の意見を参考にすると，入力するユーザの意見が偏る可能性がある．特に，根拠のない情報に基づく意見が多い場合には問題となる．そのため，ユーザには，事実に基づいた情報を提示することが望ましい．提案システムでは，複数のニュースサイトから配信されている記事を基に，イベントとして関連情報を提示する．

本稿では，市区町村程度の規模以上の地域を対象に意見収集を行う場合を想定する．地域の規模が小さいと，関連するニュース記事が十分に得られないためである．また，意見収集自体の問題設定はその管理者が行うものとする．意見収集の管理者は，提案システムに対し，問題に関連の深い地域と記事を設定する．提案システムは，その情報を基に記事をイベントとしてまとめて，ユーザに提示する．

2 地域性のある記事とイベントの抽出

2.1 記事に関連する地域の抽出

ニュースサイトには，“地域” というカテゴリやジャンルが存在することがあり，地域に関係した記事を取得できる．しかし，同一な異様の記事でも分類はサイトによって異なり，“国内” や “環境” というカテゴリに分類されることもある．複数のニュースサイトからの記事を利用するため，一定の基準で地域性を評価する必要がある．

記事と関連のある地名をラベルとして記事に付与する．記事中に地名が明示されている場合は，その地名をラベルとして，関連度を表す評価値を高くする．地名の判定は，形態素解析器 MeCab[1] によって “地域” と判定された単語とする．ニュース記事には，5W(1H) を明示して書くという慣例があり，解析結果だけでも十分な判定が可能である．

本研究ではさらに，地名ではないが地域に関連する単語を考慮するために，Wikipedia のデータを利用する．記事中の出現単語 w が関連する地名を w_l で表し，単語 w が地名 w_l を表す評価値 $locality(w, w_l)$ を式 (1) に示す．w が形態素解析器によって地名と判定された場合は W_l は 1 要素のみで，$w_l = w$ である．それ以外の場合は，上述の Wikipedia 記事を用いて単語 w が地名 w_l に分類される確率 $p(w_l|w)$ を計算する．145,668 記事から確率 $p(w_l|w)$ を計算するために，スケーラブルに動作する分散環境 Hadoop 上のマイニングライブラリ Mahout を用いた．α は $p(w_l|w)$ の影響を調整するための変数であり，$0 \leq \alpha \leq 1$ としている．

$$locality(w, w_l) = \begin{cases} 1, & if \quad w_l = w \\ \alpha \cdot p(w_l|w), & otherwise \end{cases} \quad (1)$$

$$W_{l,w} = \{w_l | locality(w, w_l) \geq \theta\} \quad (2)$$

$$W_{l,a} = \bigcup_{w \in a} W_{l,w} \quad (3)$$

式 (2) に示すように，$locality(w, w_l)$ が閾値 θ 以上の値を示す w_l の集合 $W_{l,w}$ を求め，記事 a に w_l を地名ラベルとして付与する．記事 a の地名ラベルの集合 $W_{l,a}$ は式 (3) で表される．

2.2 特定の地域に関するイベントの抽出

本稿でのイベントは類似する記事の集合とする．意見収集の管理者が設定した地域と記事によって，関連するイベントを抽出する．

記事間の類似度として，地域の内容的な類似性と，空間的な類似性についてそれぞれ評価する．内容的な類似性は，式 (4) に示すように，各単語の tf·idf の値をベクトルの大きさとしたコサイン類似度によって表わす．また，$\lambda^{|t_1 - t_2|}$ によって近い時刻に配信された記事の類似度を高くする．t_1 は記事 a_1 が配信された時刻であり，$0 \leq \lambda \leq 1$ である．

$$sim_{cont}(a_1, a_2) = \lambda^{|t_1 - t_2|} \cdot cos(a_1, a_2) \quad (4)$$

$$sim_{space}(w_{base}, a) = \min_{w_l \in W_{l,a}} d(w_{base}, w_l) \quad (5)$$

空間的な類似性は，式 (5) に示すように，設定された地名 w_{base} と記事 a の地名ラベルの集合 $W_{l,a}$ が示す座標の最短距離とする．$W_{l,a}$ は記事 a に付与された地名ラベルの集合である．$d(w_1, w_2)$ は地名 w_1, w_2 が示す座標の距離を表す．

イベント抽出の対象は，式 (5) に示す距離が，一定の範囲となるように制約を設ける．この範囲は意見収集の管理者が決定する．イベントはクラスタリングにより抽出するが，地域によって関連する記事数が異なり，抽出されるイベント数を予め決められない．また，ニュース記事は次々に配信されるため，容易にデータが追加可能である必要がある．データが追加，整理されるため，Text Analysis Conference の Update Summarization[1] というタスクに類似するが，提案システムでのイベントの抽出では，単純にクラスタリング手法の leader-follower 法を用いる．記事とイベントとの類似度は，イベントに属する各記事との類似度 sim_{cont} の平均とする．

ユーザには，はじめに，管理者が設定した地名と記事を含むイベントが提示される．そして，同一地域において，類似するイベントを抽出し，時系列に並べて提示することで，イベントを追跡することが可能となる．類似するイベントとは，管理者の設定した記事と各イベントとの類似度 sim_{cont} が閾値以上のイベントとする．

3 イベント追跡システム

3.1 システムの設定と動作

意見を入力するユーザが提案システムを使用する前に，意見収集の管理者が関連のある地名とその範囲，記事を設定す

An Event Tracking System using Local News on the Web
Norifumi HIRATA, Shun SHIRAMATSU, Tadachika OZONO and Toramatsu SHINTANI
† Dept. of Computer Science and Engineering, Graduate School of Engineering Nagoya Institute of Technology, 466-8555, Nagoya, Japan

[1] http://mecab.sourceforge.net/

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.

図 1: 関連イベントの一覧

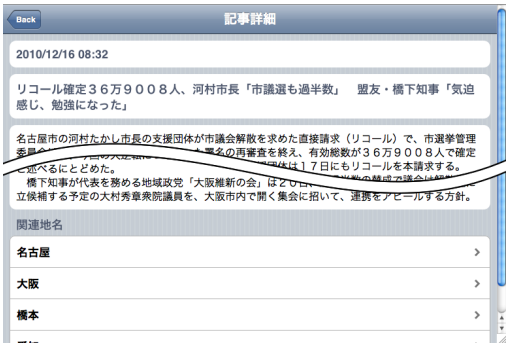

図 2: 記事詳細と関連地名の提示

る必要がある．範囲とは地名の表す座標からの半径を表し，大きく設定すれば，より広範囲の地域のイベントを追跡可能になる．地名の表す座標は Yahoo!が提供するローカルサーチ API[2]を利用した．システムは，関連のある地域と記事から，前処理としてイベントを抽出し，イベント間の類似度を計算しておく．また，一定時間ごとに新着記事の有無を確認し，イベントの新規作成，統合を行う．

3.2　インターフェースと実行例

本稿で示す実行例は，2010 年 10 月 1 日から 12 月 31 日に配信された 33,193 記事を対象に，システムを利用した結果である．対象としたニュースサイトは，asahi.com，毎日 jp，MSN 産経ニュース，YOMIURI ONLINE である．

ユーザには，管理者によって設定された地名と記事に関連するイベント集合が提示される．図 1 は，地名を名古屋市，範囲を 10km，名古屋市のリコールに関する記事を設定した例であり，12 イベント，51 記事が提示された．

図 1 のリストの各項目がイベントを表している．イベントを表す各項目は，イベントを構成する記事の tf·idf の平均ベクトルに最も類似する記事のタイトルである．イベントの選択によりイベントを構成する記事のタイトル一覧が表示される．さらに，記事のタイトルを選択することで，図 2 に示すように，記事の本文を含めた詳細が閲覧可能になる．リストではなく，グラフ構造での表現 [2] も可能である．しかし，地域に関するイベントと限定しているため，リスト表現でもイベント追跡が可能であると判断した．

記事詳細には，記事にラベル付けされた地名が表示される．地名を選択すると，閲覧中の記事と地名を基にした関連イベントが提示され，他地域の関連情報を得ることが可能となる．システムの使用前に管理者が地名と記事を設定したが，この場合と同じ動作により実現されている．地名の選択のみで，ユーザは他地域の関連情報を得ることが可能となる．

また，図 1 の右上部の “名古屋市” の最近のイベントというボタンを押すと，“名古屋市” に関するイベントが時系列に提示される．

3.3　考察

図 1 から，提案システムによって，特定の地域のみに関連したイベントを抽出可能であると言える．本稿の例では，無関係の地域に関する記事は，1 記事のみであり，イラン関係の記事であった．“中東” の座標を求めた場合に，日本国内の座標になっていたことが原因であると考えられる．しかし，51 記事中 1 記事のみという点を考えれば，高い精度である．

また，図 2 の関連地名に示すように人名が地名と判断される場合があった．“橋下” が地名であるか人名であるのかは式 (1) の性能次第である．これに対し，記事が属するイベント内の別記事の利用が考えられる．例えば，別記事の地名ラベルとの比較を行い，距離の遠い場合は関係が薄いとして，関連地名として除外する手法が考えられる．

本稿の例では，“名古屋市” という関連する記事が多く存在する地名を用いている．関連記事の少ない地名に関して提案システムを使用すると，提示されるイベントの数が少なくなると考えられる．この問題の対策として，提案システムでは，複数情報源からの記事を対象としている．しかし，地域の偏りは存在し，ブログやマイクロブログ等のニュースサイト以外の複数情報源を利用したとしても，限界がある．そこで，他地域の類似するイベントを提示することでの問題の解消が期待できる．例えば，ある町にリコール問題が発生した場合，より多くの記事が配信された都市でのリコール関連のイベントを提示すれば，参考にできる．他地域のイベントであることを明示して提示すれば，地域に関する意見入力の補助として利用可能である．また，図 2 の関連地名の選択部分を拡張し，他地域でも関連する記事が存在する場合は，記事の地名ラベルに含まれてない地名も提示することが望ましい．

4　おわりに

本稿では，記事をまとめたイベントを追跡することで，関連情報の理解と意見収集の補助となることを目的としたシステムを提案した．実際に配信された記事を利用することで，提案システムによって，地域を限定した状態でのイベント追跡が可能であることが確認できた．また，地名の誤判定や関連記事の少ない場合の課題などが明らかになった．

今後は，提案システムが意見入力に対して有用であるかどうかを，実際の運用によって調査する必要がある．

参考文献

[1] Dan Gillick, Benoit Favre, Dilek Hakkani-Tur1, Berndt Bohnet, Yang Liu and Shasha Xie: The ICSI/UTD Summarization System at TAC 2009, Proceedings of Text Analysis Conference 2009, 2009.

[2] 平田紀史, 白松俊, 大囿忠親, 新谷虎松: ユーザの観点に基づくイベント系列化を用いた Web ニュース記事閲覧支援システムの実装, 人工知能学会論文誌, vol.26, no.1, pp.228-236, 2011.

[2]http://developer.yahoo.co.jp/webapi/map/

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.