コミュニティ知識を用いた機械学習によるイベント情報の構造化

森近 憲行 † 濱崎 雅弘 ††　亀田 尭宙 † 大向 一輝 ‡　武田 英明 †

† 東京大学　†† 産業技術総合研究所　‡ 国立情報学研究所

1 はじめに

今日，インターネット上でさまざまな情報が発信，公開されている．これらの情報には構造化されRSSなどのフォーマットをもって発信されるものもあるが，ほとんどのものは構造化されておらず，そのままで情報を機械で処理したり，再利用することは困難である．このような構造化されるべきデータのひとつに，イベント情報がある．イベント情報をカレンダーソフトなどで扱うためには，必要な情報を抽出して，フォーマットに沿った形に構造化する必要がある．

イベント情報抽出に関する研究としては，機械学習を用いるもの [1] や，抽出規則を人手で作成しそれを適用するもの [2] がある．しかし，機械学習による抽出は人手による抽出規則を適用する方法に比べて精度が低い傾向があり，人手による抽出規則を適用する方法には，抽出規則を作成するためにかかる負荷が大きい，という欠点がそれぞれあった．さらに，同じイベント情報といっても，その表記のされ方はウェブサイトによって異なるのが一般的である．（さらにいえば，同一ウェブサイト内でも，さまざまな表記が存在しうるし，それらは時間経過によって変化することも考えられる．上で述べる同一ウェブサイトのように，ある程度情報の記述のされ方が決まっている単位を以下コミュニティと呼ぶことにする．）これらの手法をさまざまなウェブサイトに適用して，精度良くイベント情報を抽出するのは困難である．

一方でWikipediaに代表されるように，システムの利用者が情報を発信することが一般的になってきている．このような集合知を利用するシステムにおいては，いかに多くの信頼性の高い情報を収集するか，つまり，利用者の情報発信に対するハードルを下げることと，情報の信頼性を高めることが重要となる．

本研究ではこれらを踏まえ，機械学習とシステム利用者の集合知を組み合わせることで，人手で作る規則の精度の良さを担保しながら，保守の負荷を軽減することが可能となる手法を提案する．ユーザの入力する情報を機械学習に組み込むことで，ノイズデータを取り除くことができ，また機械学習によって多様なユーザ参加の形態を用意することできるのでユーザの情報発信に対するハードルを下げることができる．一方，ユーザに情報を入力してもらうことで，情報抽出の精度を担保でき，コミュニティ毎にあった学習器のチューニングを行うことができる．以下，2章で提案手法について，3章で実装例について，最後に4章で今後の課題について述べる．

2 提案手法

本研究では，教師あり機械学習にユーザが投稿する情報や知識を反映させる手法を提案する．ユーザが投稿する情報とは，正例や負例といった教師データだけでなく，データから入力ベクトルを作成するための抽出規則や特徴ベクトルを作るための手掛かり語なども含まれる．また，入力方法も単なる正誤判定（二択問題）から，複数候補からの正解の選択，文章からの正解情報の抽出，さらには抽出規則の入力と多岐にわたる．このようにユーザ知識を機械学習を挟んでシステムに反映することで，ユーザの知識やモチベーションに応じた多様なユーザ参加を可能にし，また入力されたユーザ知識の信頼性も担保できるようにする．ユーザが入力しうる情報を列挙すると、下のようになる．

1. 教師データ（入力ベクトルと目標ベクトルのペア）
   (a) 抽出結果の正誤判定（二択問題）
   (b) 正しい情報の選択（複数からの選択）
2. 特徴ベクトル
   (a) 抽出ルールの記述（プログラムの記述）

図1に，以上で説明した手法の全体図を示す．

3 実験システム

提案手法を用いて，人工知能学会のメーリングリスト（以下，ML）のイベント情報構造化システムを実装した．イベントに関連する情報としてはタイトルや

Extraction of Structured Information by Machine Learning using Community Information

† Noriyuki Morichika
†† Masahiro Hamasaki
† Akihiro Kameda
‡ Ikki Ohmukai
† Hideaki Takeda
the University of Tokyo (†)
National Institute of Advanced Industrial Science and Technology (††)
National Institute of Informatics (‡)
Faculty of Engineering, University of Tokyo, Bunkyo-ku, Tokyo, Japan

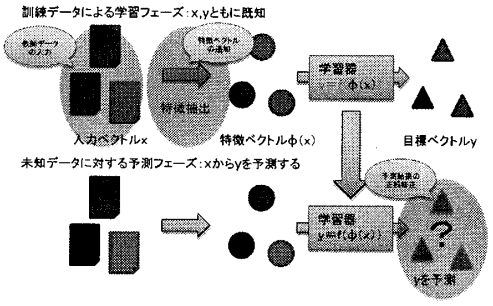

図 1: 教師あり機械学習とユーザ参加.

概要，開催地，開催日，参加申込締切日，講演者名などが考えられる．今回は特にイベントのタイトル，開催日，開催地を抽出した．タイトル抽出については，対象 ML のルールとしてメールのタイトルがイベントのタイトルとなっていたため，そのまま用いた．開催日と開催地の抽出についてユーザ参加を実装した．

図 2 がシステムのトップページである．左側にある検索フォームから，キーワードや開催日，開催地などでイベントを検索できる．図 3 はユーザ参加を行う（情報を入力する）フォームの例である．ユーザがシステムが抽出した開催日や開催地情報に間違いがあると気付いた場合は，このようなフォームから正解情報を入力する．

図 3 のフォームでは，システムが開催日と思われる候補を複数提示し，ユーザはその中から正解を選択するという，簡単なタスクを行う．メール本文から開催日を記載している行を発見し，日付を入力するという，図 3 のタスクと比べると負荷が大きいものや，日付情報を抽出するための正規表現を入力するタスクもある．これはより負荷が大きく、参加できるユーザも限られるが，ごく少数のデータで精度向上に大きく貢献する可能性を持つユーザ参加であると考えられる．

3.1 開催地情報

開催地情報は住所表記の定型パターンもあるが，大学名，建物名など固有の名称が多く，正規表現での抽出は困難である．そこで，山田ら [3] の手法を参考にし，各行内に現れる単語などの特徴ベクトルを用いて，SVM で各行が開催地情報を含むか否かを判定した．

3.2 開催日情報

開催日情報は，年月日の 3 つ組で構成されることなどの構造パターンが決まっているため，正規表現を利用することで精度よく抽出できると考えられる．よって，正規表現にマッチするか否か，特定キーワードを行内に含むか否かなどの特徴ベクトルを用いて，SVM

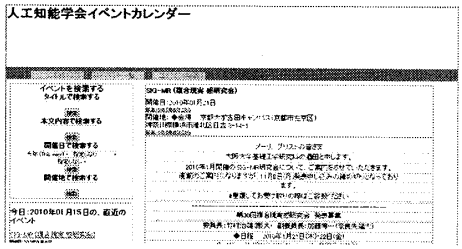

図 2: システムのスクリーンショット.

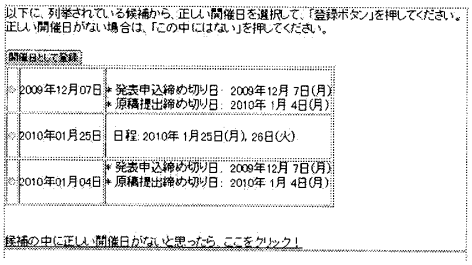

図 3: ユーザ参加（開催日候補からの選択）の例.

で各行が開催日情報を含むか否かを判定した．

4 今後の課題

今後の課題として，実装したシステムをユーザに利用してもらい，ユーザはどのように振舞うのか，ユーザの入力する情報によって情報抽出の精度はどのように変化するのか，などについて検証する必要がある．具体的には，各ユーザ参加について，修正に要する時間や間違い投稿の発生率，ユーザアンケートの結果などから，負荷がどのようにかかっており，ユーザの振る舞いの傾向はどうなっているのか，教師データと特徴ベクトルの追加によって精度の変化にどのような差が生じるのか，などについて調べる必要がある．

参考文献

[1] Xin Xin, Li Juanzi, Tang Jie, and Luo Qiong. Academic conference homepage understanding using constrained hierarchical conditional random fields. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pp. 1301–1310, New York, NY, USA, 2008. ACM.

[2] 佐藤円, 佐藤理史, 篠田陽一. 電子ニュースのダイジェスト自動生成. 情報処理学会論文誌, Vol. 36, No. 10, pp. 2371–2379, 19951015.

[3] 山田寛康, 工藤拓, 松本裕治. Support vector machine を用いた日本語固有表現抽出. 情報処理学会論文誌, Vol. 43, No. 1, pp. 44–53, 20020115.