社会ネットワークにおける社会習慣の検出

一柳 有希　　V.V. Kryssanov　　小川 均
立命館大学　情報理工学部　情報コミュニケーション学科

1. はじめに

社会習慣とは，社会ネットワーク内での習慣，行動パターン，規則，禁制，昔からの習慣，暗黙の了解事等を指す．Bechtold は，企業内の社会習慣を分析することで，企業の営業への影響について研究を行っている[1]．しかし，現在の社会習慣を調査する際には主にアンケートが主流であるため，手間がかかる．そこで本研究では，社会ネットワークにおける社会習慣の検出を目的とし，多くの人が利用し，人と人の関係が容易に理解可能な SNS 内において公開されている情報の中から社会習慣を検出することに着目した．SNS では多様な表現がされているので，簡単なパターンのみでは情報収集が困難である．したがって，本稿では基本ルールを与え，機械学習で類似ルールや新しいルールを獲得する手法を用いる．得られたルールは新たなデータを用いることでシステム・ルールの有効性を確かめる．なお，今回検出する社会習慣は，研究室内 SNS における社会習慣を対象とする．本研究の応用例として，研究室に新しく配属となる学生を対象に研究室の特性やイベント等を知らせるアプリケーションが挙げられる．

2. ルールの分類と評価極性

本研究において，まず文の構成を基本ルールとし，それぞれどのような社会習慣であるかを大まかに分類し，学習する．しかし，品詞ごとに分類するだけでは社会習慣を判定することは困難である．よって，評価極性辞書を用いて入力データを更に分類する．以下のルールについて評価極性を説明する．

Rule1: 副詞＋名詞＋動詞
Rule2: 名詞＋助動詞

Rule1 を用いて日常的な社会習慣を検出する．なお，ここでの副詞は曜日等も含む．Rule2 においては，制限等を表す社会習慣を検出するためのルールである．

Detecting Socio-Cultural Habits in Social Network
Yuki Ichiyanagi
V.V. Kryssanov
Hitoshi Ogawa
Information Science and Engineering, Ritsumeikan University

例として，Rule1 にあてはまる「月曜日にゼミがある」を挙げる．「ある」という単語は評価極性辞書により好評極性と判定される．Rule2 において，「ゼミに遅刻は厳禁だ」という例を挙げる．ここでの「厳禁」という単語は不評極性である．これらの基本ルールと評価極性値を加味し，社会習慣の学習および判定を行う．

3. ベイジアンフィルタ

機械学習によって社会習慣を検出する際には，スパムメールの検出にも利用されるベイジアンフィルタを用いる．ベイジアンフィルタとは，ベイズの定理を用いて対象となるデータを解析・学習し分類する為のフィルタである．機械学習すればするほど判定結果の精度が向上する．

ベイズの定理とは，過去の結果をもとに未来に起こる事象の確率の予測を立てる定理である．現在，音声認識，画像認識，スパムメールフィルタなどに用いられている．

本研究における社会習慣か否かの判定にもベイジアンフィルタが有効であると考え，社会習慣を検出するベイジアンフィルタを設計・実装する．

4. 検出手順

本研究では，Paul Graham 方式[2]を社会習慣の検出に応用する．ベイジアンフィルタは主に学習アルゴリズムと，判定アルゴリズムで構成されている．学習用データとして，現在様々な研究室に所属している大学生 16 名に各々の研究室においてどのような社会習慣があるかアンケートを実績した．あらかじめ学習アルゴリズムで社会習慣である文章を，アンケートをもとに得た品詞の基本ルールごとに学習させ，単語ごとに社会習慣である確率を学習データベースに格納する．同様に社会習慣ではない文章も学習させ，単語ごとの確率を学習データベースに格納する．判定する文章が与えられた際に，学習データベース内の確率を呼び出し，各単語の社会習慣である確率の結合確率を計算し判定する．社会習慣とみなされなかった文章は，学習アルゴリズムにおいて学習する．学習後，自己判断で社会習慣であるが基本ルールに当てはまらなかった社会習慣である文章を，ル

ールベースに反映させ，新たなルールを追加する．

4. 学習アルゴリズム

入力された文章を単語の品詞ごとに分類する際に，日本語形態素解析システム Sen[3]を使用する．

形態素解析を用いて品詞ごとに分類した後，Paul Graham 方式を用いて各単語が社会習慣である確率を求める．単語が社会習慣である確率を $p(w_i)$ とし，(1)式を用いて計算する．

$$p(w_i)=\frac{\min(1.0, h/th)}{\min(1.0, 2*nh/tnh)+\min(1.0, h/th)} \quad (1)$$

h：社会習慣である文章内での単語 w_i の出現数

nh：社会習慣でない文章内での単語 w_i の出現数

th：社会習慣である文章の総数

tnh：社会習慣でない文章の総数

(1)式において，社会習慣ではないという判定にバイアスを書けるために社会習慣ではない文章における単語の出現回数を2倍している．社会習慣である文章にしか登場しない単語の社会習慣である確率は0.99，社会習慣でない文章にしか登場しない単語の確率は0.01，データベースに存在しない単語の社会習慣である確率は0.4とする．また，$2*h+nh<5$ である単語は省く．

5. 判定アルゴリズム

入力データを学習時と同様に Sen で文章を単語ごとに分解する．各単語の社会習慣である確率を学習データベースから呼び出し結合確率を，(2)式を用いて計算する．(2)式において，入力データにおける各単語の社会習慣である確率を $p(w_i)$ とし，入力データ全体の結合確率を $P(w)$ とする．

$$P(w)=\frac{\prod_{i=1}^{i=15} P(w_i)}{\prod_{i=1}^{i=15} P(w_i)+\prod_{i=1}^{i=15}(1-P(w_i))} \quad (2)$$

最も特徴的な単語の確率を用いて計算するために，社会習慣である確率が0.5から最も離れている確率から順に，上位15個の確率を用いて結合確率を計算する．結合確率 $P(w)$ が0.9以上であれば，入力データが社会習慣であると判定する．

6. 実験

入力データが社会習慣か否かを分類し，同時に新たな基本ルールを検出するための実験を行った．本実験では，第3章で述べたように，研究室に所属している大学生に実績したアンケートを学習データとして学習させる．その上で，基本ルールに基づいて入力データを判定し，基本ルールの有効性を確認する．当てはまらないものは，新たにルールを作成しルールベースに加える．また，新たに研究室に所属している大学生30名を対象に同様のアンケートを行った．収集したデータは有効性を確認するためのデータとして使用する．

実験結果では，基本ルールに当てはまる入力データは適切に分類された．しかし，「いつも」や「毎日」のような特徴的な単語が含まれない文章の判定が困難である．機械学習することで，社会習慣を検出するだけでなく，社会習慣を検出する社会ネットワーク特有の単語を抽出することが重要となってくることが確認できた．以下に実験結果の一部を示す．

- 基本ルールに適応した社会習慣である文章
  - 月曜日にゼミがある　(Rule1に適応)
  - ゼミに遅刻は厳禁だ　(Rule2に適応)
- 分類されなかったが，社会習慣である文章
  - 研究室でゲームはしない

上記の文章の「しない」は，動詞+助動詞で構成される．よって新たに Rule3「動詞+助動詞」を加えることで，「無断欠席をしない」という文章等にも対応することが可能となった．

7. おわりに

本稿では，基本ルールを用いて社会習慣を検出する手法を考案した．また実験によって，基本ルールの有効性を確認し，新たに基本ルールを検出することで，更に精度の高い社会習慣の検出が可能となった．

今後は検出したルールをもとに，研究室内 SNS において社会習慣を検出および利用したアプリケーションの作成，実装を行う．

参考文献

[1] Brigid L. Bechtold, "Toward a participative organizational culture: evolution or revolution?", Empowerment in Organization, Vol. 5, No.1, pp. 4 – 15, MCB University Press, 1997.

[2] 形態素解析 Sen http://www.mlab.im.dendai.ac.jp/~yamada/ir/MorphologicalAnalyzer/Sen.html

[3] Paul Graham, "A Plan for Spam" http://www.paulgraham.com/spam.html