アスキーアート自動抽出法の提案

中澤　昌美†　松本　一則†　柳原　正†　池田　和史†　滝嶋　康弘†
株式会社ＫＤＤＩ研究所†

1. はじめに

現在，インターネット上には多くの電子掲示板が開設されており，インターネット上におけるコミュニケーションの場の一つとなっている．電子掲示板では一般に，プレーンテキストを用いる．この制約のため，電子掲示板では，文字や記号などのテキストのみを用いることで絵を表現することができるアスキーアート（テキストアートとも呼ばれる）が発展してきた．これは，テキストのみで視覚的な効果を得ることができるため，画像を投稿できない電子掲示板において，有効な手段となっている．一般的に，アスキーアートは複数行のものを指すことが多いが，本稿では，1行で表現される顔文字もアスキーアートに含める．アスキーアートの例を図 1 に示す．

形態素解析や構文解析による文章の意味理解が盛んになってきた．しかし，アスキーアートはそれ自体では言語的な意味を持たないため，既存の解析手法ではアスキーアートを含むテキスト解析は困難である．このため，自然言語処理の解析をスムーズに行うには，文書からアスキーアートを抽出する必要がある．本稿では，テキスト解析に不要なアスキーアートを抽出するため，入力文書から自動的にアスキーアート位置を特定し，アスキーアートの集合とアスキーアートが除去された文書とに分離する手法を提案する．

2. 既存手法と問題点

文書からアスキーアートを抽出する手法は以前から提案されている[1]，[2]．文献[1]では，文書中のアスキーアートの有無を，Support Vector Machine(SVM)を用いて判定する手法を提案しているが，アスキーアートを抽出するためには，文書中におけるアスキーアート位置を知る必要があるため，この手法を本稿の目的に利用することはできない．

文献[2]は，ウインドウサイズを設定し，文書を走査することで，アスキーアートの境界判定を行うものである．この手法により，アスキーアートの位置判定が可能となるが，この手法は機械学習の際，同じ記号が2回連続で現れる回数を特徴として用いる．このため，2回連続で同じ文字が並ぶことが少ない顔文字やキャラクターをデフォルメしたアスキーアートは抽出できない．

またこれらの他に，機械学習を用いない，ルールベースによる手法が提案されており，この手法は，「アスキーアートらしさを特徴付ける条件」を設定し，その条件を満たす部分をアスキーアートと特定し，抽出する．しかし，アスキーアートらしさを特徴付けるルールを一つずつ設定しなければならないため手間がかかる．また，必要な条件設定が不足してしまう可能性がある．

えっΣ(ﾟДﾟ;)

(ﾟ･ω･ﾟ)ｼｮﾎﾞｰﾝ

(･ω･`;)テレビのリモコンが行方不明に。

図 1．アスキーアート例

文献[2]とルールベース手法は，複数行アスキーアートのみを抽出の対象としており，1行のアスキーアートを抽出することはできない．

本稿では，これらの問題を考慮したアスキーアート抽出手法を提案する．

3. アスキーアート抽出法

与えられたテキスト文書の中からアスキーアートを含む行を推定し，抽出する手法を提案する．提案手法は以下の特徴をもつ．

- 機械学習によりアスキーアートの特徴を捉え，自動的にアスキーアートを抽出する．
- アスキーアートに含まれる日本語表現をアスキーアートの一部として抽出する．
- 予測結果に対し，前後の行のアスキーアートを分割することなく抽出する．

提案手法は，「学習結果モデル生成部」と「アスキーアート検出部」の2つの部分から構成される．学習結果モデル生成部では，学習データを収集し，特徴抽出を行い，機械学習を用いてモデルを生成するまでの操作を行う．アスキーアート検出部では，アスキーアート検出文書データを入力してから，アスキーアートとアスキーアート除去文書に分離するまでの操作を行う．

以下に，提案するアスキーアート自動抽出手法の詳細を示す．

3.1. 学習結果モデル生成部

まず，学習に用いるデータを収集する．アスキーアートが登録されているサイトから，アスキーアートのデータを収集し，正例データとして学習に用いる．また，文書を収集し，負例データとして用いる．

次に，集めたデータから図2のように，特徴ベクトルを算出する．文字をUTF-8で表現し，バイト単位に切り分け，10進数で表現し，各行に現れる出現頻度を特徴量とする．ここで，ある行にアスキーアートが含まれることが判明した場合，その前後の行にアスキーアートが存在する可能性が高いと言える．このような性質から，前

Proposal of ASCII-Art Extraction

Masami NAKAZAWA†, Kazunori MATSUMOTO†, Tadashi YANAGIHARA†, Kazushi IKEDA† and Yasuhiro TAKISHIMA†
†KDDI R&D Laboratories
2-1-15 OHARA FUJIMINO-SHI SAITAMA, 356-8502, JAPAN
{ms-nakazawa, matsu, td-yanagihara, kz-ikeda, takisima}@kddilabs.jp

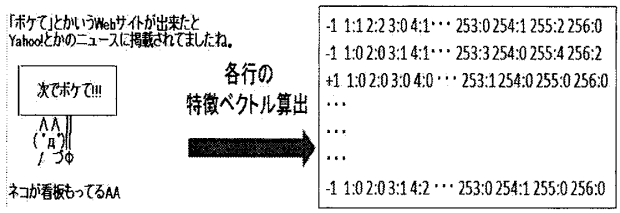

図2．各行の特徴ベクトル作成

後の行の特徴を得るため，特徴ベクトルの次元数を増やす．この操作では，ある行の特徴量と該特徴ベクトル算出対象行に連続する行の特徴量とから，該特徴ベクトル算出対象行の特徴ベクトルを作成する．このように，連続する行の特徴を含めて特徴ベクトルを作成し，次元数を増やすことにより，複数行から構成されるアスキーアートの検出精度を向上させることができる．

生成された学習データ特徴ベクトルファイルに対し，SVMを用いて学習を行い，モデルを作成する．

3.2. アスキーアート検出部

アスキーアート検出対象文書データに対し，3.1節と同様の方法で特徴ベクトルを作成し，アスキーアート検出対象文書特徴ベクトルファイルを生成する．

3.1節で生成した学習結果モデルと，アスキーアート検出対象文書特徴ベクトルファイルとを用いて，SVMにより，アスキーアート検出対象文書データの各行のアスキーアートを含む確率(アスキーアート含有確率)を算出する．

アスキーアート含有確率において，ある行のアスキーアート含有確率が前後の行のアスキーアート含有確率に対して大幅に低い場合，アスキーアートが存在する行に文字が含まれている可能性が高い．その行でアスキーアートが分離されることがないように，線形平滑化を行う．本稿では平滑処理として加重平均を用い，1つのアスキーアートを分割して検出することを防ぐ．

平滑処理を行った結果の各行のアスキーアート含有確率が50%以上である行を，アスキーアートを含む行（アスキーアート含有行）であると判定し，その行を，入力データであるアスキーアート検出対象文書データから抽出することで，アスキーアートが検出できる．また，アスキーアート検出対象文書データから，アスキーアート含有行を除去した残りの文書データを，アスキーアート除去文書として出力する．

以上の操作により，アスキーアート検出対象文書データからアスキーアート含有行が除去されたアスキーアート除去文書と，アスキーアート含有行のみが集まったデータが得られる．

4. 実験

提案手法を用いて，文書からアスキーアートを抽出する実験を行う．実験に用いる学習データは，正例18,423行，負例5,000行のデータを用いる．正例データは，複数行アスキーアート10,156行，顔文字1,000行，顔文字を含む文書2,267行の3部分から構成される．負例データは，顔文字などのアスキーアートを含まない5,000行の文書を用いる．正例・負例合計18,423行のデータを学習に用いる．

各実験に用いるテストデータ（アスキーアート検出対象文書データ）は，非アスキーアート行の間に，1つのアスキーアートを挿入して作成する．複数行アスキーアートを10個，（1行アスキーアートを10個，日本語表現を含むアスキーアートを10個）用いて，テストファイルを作成する．学習と予測には，SVMのライブラリの一種であるLIBSVM[3]を用いる．

特徴ベクトルの作成手法を変えて実験を行う．実験1では，学習データの各行を256次元の特徴ベクトルに表して学習を行い，入力した文書からアスキーアート含有行を抽出する．実験2では，学習データの前後の行の特徴を考慮して特徴ベクトルを作成，学習を行い，入力した文書からアスキーアート含有行を抽出する．

以下に各実験の詳細を示す．

4.1. 実験

実験1では，文書の各行を256次元の特徴ベクトルに変換する．10-fold Cross Validationにより性能評価を行った結果，精度は約97.0%となった．また，平滑処理を施すと，複数行アスキーアートの抽出精度が，93.2%から95.5%に上がった．

実験2では，前後1行ずつの特徴を加えて768次元の特徴ベクトルを作成する．平滑処理を行った結果，複数行アスキーアートの抽出精度は，98.6%となった．

4.2. 考察

実験1において，アスキーアート含有行と誤判定されていた文書行は，平滑処理により正解が増えたことから，平滑処理は有効であるといえる．実験2において，次元数を増やすことにより，前後の行の特徴を含めることで，アスキーアートの抽出精度が向上することが分かった．

日本語表現を含むアスキーアートは，日本語表現の部分がアスキーアートの一部として抽出できていることから，SVMによる学習はアスキーアート抽出に有効であるといえる．

5. まとめ

アスキーアートが入った電子掲示板の記事から，アスキーアートを含む行を自動的に推定・抽出し，テキスト行とアスキーアート行に分離する手法を提案した．

特徴ベクトルの算出には，文字の種類とその頻度に着目した．また，アスキーアートは複数行のものが多い点を考慮し，連続する行の特徴を含めて特徴ベクトルを作成した．さらに，一つのアスキーアートが分割して抽出されることがないように，平滑処理を行った．提案手法の実験を行った結果，約98%の精度でアスキーアート含有行が抽出できた．

参考文献

[1] 谷岡広樹，丸山稔，“形態素解析に基づくSVMを用いたアスキーアートの識別，”2005 電子情報通信学会技術研究報告．PRMU，パターン認識・メディア理解，104(670),pp.25-30, 20050218

[2] 林和幸，小熊光，鈴木徹也，“テキストアートの言語に依存しない抽出法，”2009 情処全大，巻71st号1, pp.627-628, 20090310

[3] Chih-Chung Chang and Chih-Jen Lin, “LIBSVM: A Library for Support Vector Machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/