アスキーアート自動抽出法の提案と評価

中澤　昌美† 　松本 一則† 　柳原　正†
池田　和史† 滝嶋　康弘† 帆足　啓一郎†
†KDDI 研究所 〒356-8502 埼玉県ふじみ野市大原 2-1-15
E-mail: †{ms-nakazawa, matsu, td-yanagihara, kz-ikeda, takisima, hoashi}@kddilabs.jp

あらまし　電子掲示板では，文字や記号の組み合わせにより視覚的表現が可能なアスキーアートが発展してきた．また，形態素解析や構文解析をはじめとする自然言語処理技術の発展により，文章の解析とその意味理解が進められている．これらの自然言語処理の解析においては，日本語表現のみが対象とされている．このため，顔文字やアスキーアートを含む文章の解析は困難となっている．そこで，掲示板への投稿文章を解析する前処理として，文章中の各行に現れる文字種の頻度からアスキーアート部位を推定・抽出する手法を提案し，その評価を行う．

キーワード　アスキーアート，自動検出，SVM

Proposal and its Evaluation of ASCII-Art Extraction

Masami NAKAZAWA† 　Kazunori MATSUMOTO† 　Tadashi YANAGIHARA†
Kazushi IKEDA† 　Yasuhiro TAKISHIMA† 　and 　Keiichiro HOASHI†
†KDDI R&D Laboratories　2-1-5 Ohara, Fujimino-shi, Saitama, 356-8502 Japan
E-mail: †{ms-nakazawa, matsu, td-yanagihara, kz-ikeda, takisima, hoashi}@kddilabs.jp

Abstract　In bulletin boards of the internet, ASCII-Arts have been developed. ASCII-Arts enable visual presentation by combine characters and symbols. In the other hand, these expressions degrade the quality of natural language analysis, such as morphological processing and structure analysis. For the preprocessing for the analysis of articles posted to bulletin boards, this paper proposes the ASCII-Art detector.

Keyword　ASCII-Art, Automatic detection, Support Vector Machine

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いる．その種類は，様々なテーマを取り扱う総合掲示板，特定のテーマのみに絞った専門掲示板，画像を扱う画像掲示板，利用者を制限した掲示板など多岐にわたる．これらの各電子掲示板には毎日多数の書き込みが行われており，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．

電子掲示板ではプレーンテキストを用いるのが一般的である．このような制約があるため，電子掲示板では，文字や記号などのテキストのみで絵の表現が可能なアスキーアート（テキストアート）が発展してきた．アスキーアート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，画像を投稿することができない電子掲示板において，有効な手段となっている．巨大掲示板では，アスキーアートを用いた独自のキャラクターが生み出されたり，そのキャラクターを登場させたストーリーが作成されたりと，電子掲示板利用者に親しまれている．

図 1．アスキーアート例

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．複数行アスキーアートの利用

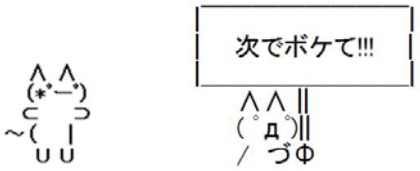

図 2.（左）デフォルメ・アスキーアート例
（右）日本語表現を含むアスキーアート例

は電子掲示板やメールマガジンなどがメインで，その他ではあまり用いられない．一方，1 行アスキーアートはそれらの他にも，ブログ，電子メールなど，幅広く用いられている.アスキーアートの例を図 1 に示す．

文章においては，形態素解析や構文解析による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．そこで，自然言語処理の解析をスムーズに行うためには，文章中から現在テキスト解析ができないアスキーアートを抽出する必要がある．アスキーアートが抽出できれば，文章とアスキーアートが分離できる．アスキーアートが除かれた文章に対し，テキスト解析を行うことで文章の意味理解が可能となる．

本稿では，文章を入力すると，自動的にアスキーアートを検出し，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章とに分離する手法を提案する．

2. 既存手法と問題点

文章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, SVM を用いて判定する手法を提案している． SVM の特徴量として，バイトパターンと形態素解析による品詞情報の 2 つが用いられている．バイトパターンとは，データを UTF-8 で表現したときのバイトストリームをバイト単位に切り分けたデータの出現頻度を表す．この手法により，文書中のアスキーアートの有無が判定可能となったが，アスキーアートを抽出するためには，文書中でのアスキーアートの位置を特定する必要があるため，この手法を本稿の目的に利用することはできない．

上記の手法は形態素解析を用いるため，日本語のテキスト以外には対応していないのに対し，多言語にも対応するアスキーアートの抽出法が提案された[2]．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，この手法では機械学習の際，同じ記号が 2 回連続で現れる回数を特徴として用いている.「２ちゃんねる」をはじめとする電子掲示板では，アスキーアートの一種である図 2（左）のようなデフォルメ・アスキーアートが用いられる．デフォルメ・アスキーアートは，2 回連続で同じ文字が並ぶことが少ないため，この手法では抽出できないアスキーアートが存在する．

また，機械学習を用いずにアスキーアートを抽出する手法が提案されている．この手法は,「テキスト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，抽出するものである．テキストアートらしさを特徴付ける条件としては,「ひらがな，カタカナ，漢字，アルファベットなどの通常の文字を除く記号から構成されている」,「一部に通常の文字を含み，大部分が記号からなる」,「同じ文字が繰り返される」などが挙げられている．この手法では，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．一例として，図 2（右）のような日本語表現を含むアスキーアートの場合，日本語の部分だけが抜け落ちて抽出されてしまう．

上記の 3 つの手法のうち，文献[2]の手法では，2 行以上の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．

以上をまとめると，既存手法ではアスキーアート部位を判定できない，アスキーアートがある位置の特定が可能となっても，抽出できないアスキーアートが存在する，1 行アスキーアートが対象外であるといった問題点がある．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．提案手法は次のような特徴をもつ．

機械学習を用いることで，アスキーアートらしさの特徴をとらえ，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意味のあるテキスト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，前後の行の確率をもとに，平滑処理を行うことによって，1 つのアスキーアートを分割することなく抽出する．

提案手法は図 3 のように,「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とは，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を表す．

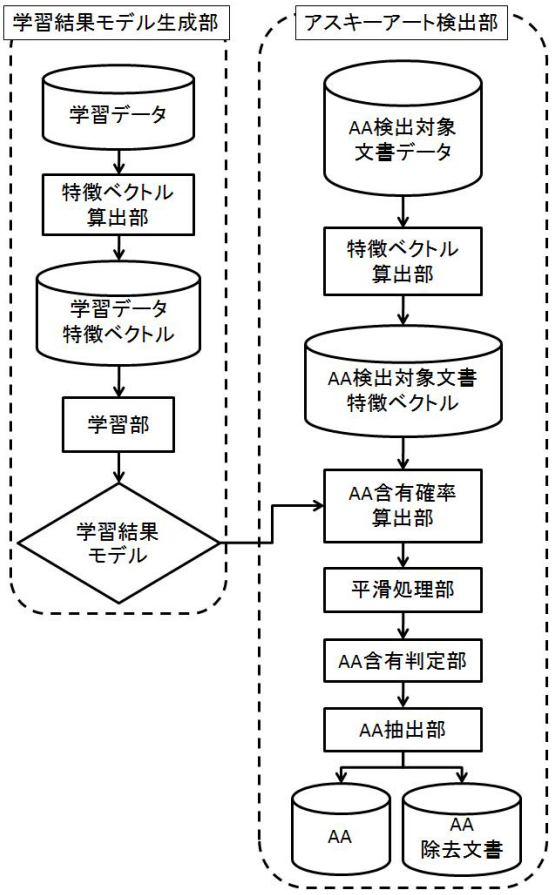

図 3. アスキーアート検出構成図

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，ブログ文書を収集し，負例データとして用いる．

次に，集めたデータから特徴ベクトルを算出する．文字を UTF-8 で表現したとき,バイト単位に切り分け，10 進数にし，各行に現れる出現頻度を特徴量とする．このような変換を行うことで，図 4 のように，テキストの特徴を 256 次元という，現実的に処理が可能な次元で表現することが可能である．UTF-8 では文字種により文字コードが固まって存在しないため，上記の操作による特徴ベクトル作成は，ルールベース手法とは異なるものである．

ある行にアスキーアートを含む行があることが判明した場合，それに連続する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操作を行う．この操作では，ある行の特徴量と

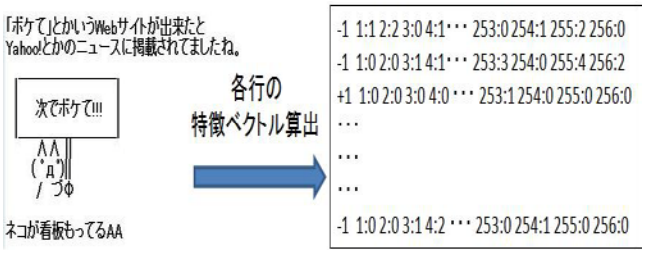

図 4. 各行の特徴ベクトル算出

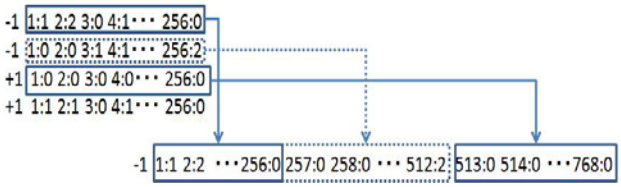

図 5. 前後 1 行を考慮した特徴ベクトル作成

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．例えば，前後 1 行ずつを用いて特徴ベクトルを作成する場合，該当行は，256 次元の 3 倍の 768 次元の特徴ベクトルとして表現する．このとき，N 行目の特徴ベクトルの前に,N－1 行目の特徴ベクトルを追加し，さらに，N 行目の特徴ベクトルの後に，N＋1 行目の特徴を追加する．N 行目の 768 次元の特徴ベクトルは，N－1 行目の 256 次元特徴ベクトル，N 行目の 256 次元特徴ベクトル,N＋1 行目の 256 次元特徴ベクトルとして構成される．図 5 は，2 行目の 256 次元特徴ベクトルを,前後 1 行の各 256 次元特徴ベクトルを加え，768 次元で表現する手法を示している．ここの例では，前後の 1 行を対象としているが，実験では対象行数を変化させて比較する．

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と考えられる．

上記の手順で作成した全特徴ベクトルを一つのファイルに出力し,学習データ特徴ベクトルを作成する．このファイルは,正例データの各行の特徴ベクトルと，各負例データの各行の特徴ベクトルから構成される．

生成された学習データ特徴ベクトルを Support Vector Machine(SVM)[3][4]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とを用い，SVM により，

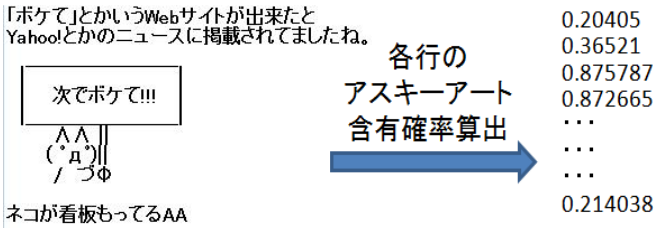

図 6. 行毎のアスキーアート含有確率推定

図 6 のような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と考えられる. 例えば，図 2（右）のように，意味のある文章「次でボケて！！！」を含む行のアスキーアート含有確率は，前後の行のアスキーアート含有確率に対して，大幅に低くなることが多い.そこで,その前後の行に対して，アスキーアート含有確率が大幅に低い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．本稿では加重平均を用いる．平滑処理により，1 つのアスキーアートを分割して検出することを防ぐことが可能となる．平滑処理の単位の行数は，任意に設定できる.

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，アスキーアートの検出ができる.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.

これにより，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

実験に用いる学習データは，正例として 1,000 個のアスキーアート（13,423 行），負例として 5,000 行のブログ文章を用いる．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, 顔文字を含む文章 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 行のブログ文章を用いる．正例・負例合計 18,423 行のデータを学習に用いる．また，評価データは，正例として 1,000 個のアスキーアート（9,025 行），負例として 6,000 行のブログ文章を用いる.

表 1. AA 行抽出実験結果

	平滑処理	適合率	再現率	F 値
実験 1-1	なし	0.87	0.84	0.85
実験 1-2	あり	0.95	0.88	0.92
実験 2-1	なし	0.96	0.86	0.91
実験 2-2	あり	0.98	0.97	0.98

各実験に用いるデータ（アスキーアート検出対象文書データ）は，複数の非アスキーアート行（文章）の間に，1 つの未学習アスキーアートを挿入して作成する.

学習と予測には SVM を用いる．SVM は，教師あり学習を用いる識別手法の一種で，マージン最大化を用いることにより，未学習データに対し高い汎化性能をもつ．本稿では，SVM のライブラリの一種である LIBSVM[5]を用いる.

実験では，提案した 2 種類の特徴ベクトル作成手法を用いる．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，学習データの前後各 1 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.

以下に提案する各実験の詳細を示す．各実験の適合率，再現率，F 値を表 1 に示す．また，単純手法を用い，提案手法との比較を行う.

4.1. 実験 1

実験 1 では，文書の各行を図 4 のように，1 文章から 256 次元の 1 つの特徴ベクトルを作成する.

はじめに 10-fold Cross Validation により，性能評価を行う．18,423 データを 10 分割し，9 組で学習，1 組で予測を行うことを 10 回繰り返し,その平均値を求める．この結果，c=32768.0，g=0.001953125 の場合，rate=97.0 となった.

次に，このパラメータ c,g の値を用い，18,423 の学習データで学習し，評価データを用いて適合率と再現率を求める.

実験 1 において,平滑処理を行わないものを実験 1-1, 平滑処理を行うものを実験 1-2 とする.

4.2. 実験 2

実験 2 では，図 5 のように，実験 1 で作成した特徴ベクトルを基に，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.

はじめに 10-fold Cross Validation により，性能評価を行う．18,423 データを 10 分割し，9 組で学習，1 組で予測を行うことを 10 回繰り返し,その平均値を求める．この結果，c=8, g=0.0078125 の場合，rate=95.8%となった.

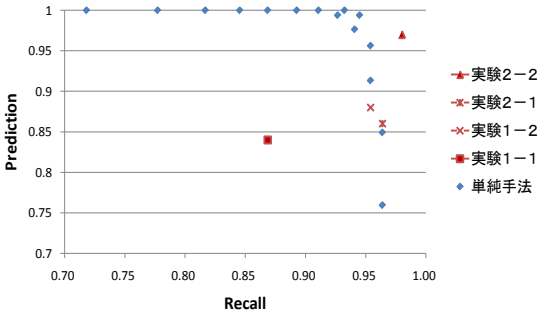

図 7. 提案手法・単純手法の Recall,Precision 結果

次に，このパラメータ c,g の値を用い，18,423 データを全て学習に用いて実験 1 と同様の評価を行う．

4.3. 考察

結果から，特徴ベクトル作成において，前後行を考慮すると，Recall, Precision ともに上がったことから，前後行を考慮することは，行単位でのアスキーアート判定に有効であるといえる．また，平滑処理を行うと，Recall，Precision ともに向上した．これより，セリフや看板などの日本語表現を含むアスキーアートにおいて，日本語を含む行はアスキーアート含有行として判定されやすくなった．

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章行は，平滑処理により正解が増えたことから，平滑処理はアスキーアート抽出に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．

実験において，日本語表現を含むアスキーアートは，日本語表現部分がアスキーアートの一部として抽出できたことから，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．

これら提案手法の比較実験として，単純手法を用いた評価を行う．単純手法とは，文章の各行の記号・空白率がある一定値以上であると，アスキーアート含有行であると判定する手法である．この閾値 0.3～1.0 を刻み幅 0.05 で変化させてアスキーアート抽出 Recall, Precision を求める．結果を図 7 に示す．このグラフから単純手法は，記号・空白率の閾値の変化により，Recall と Precision がトレード・オフの関係になっていることが確認できる．単純手法では，閾値が 0.3～0.7 までの場合，Recall はそれほど良くないが，Precision がほぼ 1.0 となり，アスキーアート行の抽出精度はよい．しかし，閾値が 0.75 以上になると急激に Precision が下がる．提案手法の一つである，前後行を考慮した特徴ベクトルを作り，さらに平滑化を行った実験 2-2 は，Recall, Precision, F 値が全て最も高い値となった．この手法は特に Recall が最も高く，アスキーアート行抽出の際の取りこぼしが少ないことから，アスキーアート行抽出に最も適した手法であるといえる．

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，自動的にアスキーアートを含む行を推定・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．提案手法と単純手法を比較し，評価行った．提案手法においては，約 98 %の精度でアスキーアート行と文章行を分離することができた．

今後の課題として，本稿では行単位でアスキーアートの抽出を行ったが，文章からアスキーアートのみを抽出できるように改良したい．また，アスキーアートを取り除く目的で行ったが，文章中におけるアスキーアートはそれぞれに意味があると考えられる．各アスキーアートの意味解析が可能になれば，アスキーアートを含めた文章の意味解析が可能になる。

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] N.Cristianini, J. Shawe-Taylor, “*An Introduction to Support Vector Machines*”, Cambridge University Press, 2000.

[4] C.W Hsu, C.C. Chang, C.J Lin, “A practical Guide to Support Vector Classification”

[5] 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/