Vol.2018-CSEC-81 No.4
Vol.2018-IOT-41 No.4
2018/5/17

ハンドメイド作品を対象としたECサイトにおける大量生産品の検出

財津 大夏[1,a)]　三宅 悠介[2]　松本 亮介[2]

概要：日本ホビー協会によると，ハンドメイド作品を対象としたECサイトの流通は年間177億円が見込まれている．ハンドメイド作品はその希少性や独創性が価値の一端を担っているが，市場の拡大に伴い，これらの性質を満たさない大量生産品の出品が問題となっている．大量生産品の増加は，ハンドメイド作品を期待する購入者のECサイトへの不信感や流通低下に繋がり，長期的には市場の衰退を招きかねない．このためハンドメイド作品を対象としたECサイトでは，大量生産品の削除や出品者のアカウント停止などの対応が行われるが，全商品の目視確認は困難であるため，大量生産品を自動的かつ継続的に検出する仕組みが必要である．本報告ではこの課題を解決するため，商品の削除やアカウント停止を回避する振る舞いを出品者の異常な振る舞いとして検出する手法を提案する．大量生産品の出品者は，アカウント停止を回避するために多数のアカウントを作成し，アカウント名に無意味な文字列を設定する傾向がある．また，商品の削除に備えて出品数を増やすために国外ECサイトの商品を翻訳しており，商品説明文などの日本語が不自然であるという特徴がある．これらの特徴を異常と定義し，単純ベイズ分類器で分類することで大量生産品を検出する．本報告では実際のハンドメイド作品を対象としたECサイトのプロダクション環境のデータに対して提案手法を適用し，検出率を計測して有効性を検証した．

Detection of Mass-produced Goods at EC Site to Trade Handmade Goods

Hiroka Zaitsu[1,a)]　Yusuke Miyake[2]　Ryosuke Matsumoto[2]

Abstract:
Japan Hobby Association reports that the market size of EC sites to trade handmade goods reaches to 17.7 billion in Japanese Yen. The value of handmade goods include the rarity and originality, but as the market expands, the exhibition of mass-produced goods which are not handmade becomes a problem. The increase of mass-produced products leads to the customer distrust and a long-term decline of the market. Electronic Commerce (EC) site administrators have a hard time to eliminate mass-produced items and the accounts of the sellers, but the manual operation does not scale, so the automated and continuous detection is essential. In this report, we propose a method to detect anomalous behavior which the seller tries to avoid product and account deletion. The example cases include the two cases of that the seller sets meaningless account names to create many accounts, and that the product information consists of unnatural Japanese words because of the use of machine translation from the explanations of the overseas EC sites. We define these features as anomalous and detect the anomaly of mass-produced products by classifying the features of these texts by simple Bayes classifiers that shows high performance in document classification. We present and evaluate the effectiveness of the proposed system by introducing the system to a handmade EC site and measuring the detection rate.

[1] GMO ペパボ株式会社 技術部 デザイン戦略チーム
Design Strategy Team, Engineering Department, GMO Pepabo, Inc., Tenjin, Chuo ku, Fukuoka 810-0001 Japan
[2] GMO ペパボ株式会社 ペパボ研究所
Pepabo Research and Development Institute, GMO Pepabo, Inc.
[a)] zaitsu@pepabo.com

1. はじめに

成長を続けるホビークラフト関連の国内市場は 2016 年時点で 8,900 億円規模に達しており，CtoC の EC サイトの発展と相まって，ハンドメイド作品を対象とした EC サイトの 2017 年の流通総額は年間 177 億円が見込まれている [15]．ハンドメイド作品はその希少性や独創性が価値の一端を担っているが，市場の拡大によって多くの購入者が EC サイトに集まるようになり，取引の機会が増加するに伴って，ハンドメイド作品としての価値を持たない大量生産品をハンドメイド作品と偽って出品するユーザーの出現が問題となっている．ハンドメイド作品を期待する購入者にとって，大量生産品は商品を探したり購入したりする際の障壁となり，ハンドメイド作品を対象とした EC サイトとしてのブランドイメージが毀損されることで，サイトへの不信感や流通低下に繋がる．そのため，ハンドメイド作品を対象とした EC サイトの運営者は，大量生産品の削除や，その出品者のアカウントを停止する対応を行っている．

大量生産品の削除や，その出品者のアカウント停止を行うためには，大量生産品を判断した上で検出する必要があるが，ハンドメイド作品であると偽った大量生産品を商品名・商品説明文・商品画像などの属性情報のみによって大量生産品であると判断することは難しいため，GMO ペパボ株式会社の運営するハンドメイド作品を対象とした EC サイト「minne」[4] では，大量生産品が出品されている一般的な EC サイトで一致する商品を確認することによって判断を行っている．商品の一致を判断する仕組みは，人間の目視によって手動で行う方法や，画像の一致によって機械的に行う方法が考えられるが，それぞれに課題がある．まず目視によって行う方法について，商品や出品者の登録・更新が頻繁に発生し得る EC サイトにおいては総検出数や 1 件あたりの検出に掛かる時間の短縮に限界があり，継続的に大量生産品の検出を行うことができない．ただし，ハンドメイド作品やその出品者に対して商品の削除やアカウントの停止を誤って行った場合，大量生産品の削除によって守ろうとしている EC サイトへの信頼感やブランドイメージを却って損なうことに繋がるため，minne では大量生産品が出品されている EC サイトを目視によって確認し，明らかに同一であると判断できる商品のみを検出する方法を採っている．次に機械的に行う方法について，minne の商品数は 676 万点を超えている [11] ため，その規模から機械的に大量生産品の検出が行われることが望ましいが，全ての商品について一致不一致の判断を行うには，大量生産品が出品されている EC サイトから商品画像を取得するために膨大な数のリクエストが必要となり，該当の EC サイトに与える負荷や通信量の点から現実的ではない．加えて，大量生産品が出品されている EC サイトはハンドメイド作品を対象とした EC サイトの運営者の管理外にあり，比較対象とするサイト自体の網羅や常に更新される商品情報の大規模なクローリングは容易ではない [14][12]．

そこで本報告では，大量生産品を扱う出品者に共通事項があると仮定し，出品者を同定することによって外部情報に依存せず継続的かつ機械的に大量生産品の出品者を検出するシステムを提案する．minne において大量生産品の出品者は，アカウントのうち一部が停止されても販売を継続できるように多数のアカウントを作成する傾向があり，これら多数のアカウントを作成するために，アカウント名には無意味な文字列を設定する特徴を見出した．この特徴を用いて，N-Gram トークンの分布を特徴量として抽出し，単純ベイズ分類器で分類する．人間が生成する意味のある文字列と対称的な，無意味な文字列に固有のパターンを分類することで，外部情報を用いず機械的に新たな大量生産品やその出品者を検出することが可能となる．更に，大量生産品の出品者は，出品数を増やすために国外 EC サイトの商品を翻訳しており，商品名や商品説明文の日本語が不自然である傾向を見出した．これらの文章中に出現する単語を用いた単純ベイズ分類器により，高い精度で大量生産品の出品者とハンドメイド作品の出品者を分類する．また，これら 2 つの手法を組み合わせることで，minne の内部情報であるアカウント名のみを用いて，機械的に大量生産品の疑いがある商品の出品者を検出した上で，更に同一の出品者が作成した他のアカウントを検出することが可能となる．

本論文の構成を述べる．2 章では従来 minne で行われてきた大量生産品の検出手法とその課題を整理する．3 章では既存の検出手法における課題を解決するための提案手法について，大量生産品の出品者に見出した 2 つの特徴と，それらを検出するための特徴量抽出と分類器の実装について述べる．4 章では提案手法の有効性について実験と考察を行い，5 章でまとめとする．

2. 従来の大量生産品の検出とその課題

ハンドメイド作品は，その作り手によってデザインや製作が手作業で行われたものである．これらは一点物であることも多く，その希少性や独創性が価値の一部となっている．「minne」[4] のようなハンドメイド作品を対象とした EC サイトは，ハンドメイド作品を求める人々に売買の場を提供する目的で運営されており，出品物は全てハンドメイド作品であることが前提となっている．工業的に大量生産された商品がハンドメイド作品と偽って出品され，それが大量生産品であることが利用者の目から明らかになった場合，EC サイトのブランドイメージの毀損や利用者からの信頼を失うことに繋がるため，EC サイトの運営者は大量生産品の削除やその出品者のアカウント停止などの対応を行っている．このような対応を行うためには，まず商品

が大量生産品であるかどうかを判断する必要があるが，従来の大量生産品の検出手法には課題がある．

ハンドメイド作品と偽った大量生産品の商品名と商品説明文の多くは，当然にハンドメイド作品であるかのように謳っている．また，商品画像についても，商品が手作りされたものか，大量生産されたものかを，該当商品の画像のみから断定することは困難である．このように，EC サイトにおいて商品画像・商品名・商品説明文などの属性情報のみから商品が大量生産品であるかどうかを判断することは難しい．そこで minne では，大量生産品が出品されている一般的な EC サイトで商品の検索を行い，画像の比較によって商品が同一であることを確認した場合，minne に出品されている商品が大量生産品であると判断することで回避しているが，判断の根拠を minne の外部情報に求めることになっている．

2.1 目視によって大量生産品を検出することの課題

minne における大量生産品の判断は実際には，大量生産品が出品されている EC サイトを人間が目視によって確認し，商品画像などの商品情報から明らかに同一であると判断できる商品を検出することによって行われている．これは，ハンドメイド作品やその出品者について商品の削除やアカウントの停止を誤って行うことで，大量生産品の削除によって守ろうとしている EC サイトへの信頼感やブランドイメージを却って損なうことを回避するためである．

また，人間が目視を行うことで，同一の商品を撮影しているにもかかわらず構図が全く異なる商品画像に対しても検出を可能にしている．しかし，目視による検出の手順は，まず minne の商品ページに Web ブラウザでアクセスし，過去の経験に基づいて疑わしい商品を挙げ，それらについて大量生産品が出品されている一般的な EC サイトで検索を行い，商品が一致するかどうか判断するといった段階を踏むため，1 件あたりの検出に数分から十数分程度の時間が掛かる．minne の商品数は 676 万点を超えている [11] ため，これら全てについて目視で検査を行うには数千時間が必要となり現実的ではない．検出を行う運営者の人数は限られるが，商品やアカウントの登録を自由に行うことができる EC サイトにおいては，大量生産品について商品の削除やアカウントの停止を行っても同一の出品者によって新規アカウントの登録と再出品が容易に行われるため，網羅的・継続的に検出を行うことができない．

2.2 画像比較によって機械的に大量生産品を検出することの課題

2.1 節で述べたような規模の問題から大量生産品の検出は機械的に行われることが望ましいが，画像の比較を機械によって行う際，ハンドメイド作品を対象とした EC サイトに出品される大量生産品の商品画像は，撮り直しや画像加工が行われることで，同一の商品であっても大量生産品が出品されている EC サイトの商品画像と完全一致しない場合がある．このような完全一致しない画像についても，深層畳み込みニューラルネットワーク [6] を用いて機械的に分類を行う手法が提案されており [1]，minne でも商品画像の特徴量変換を行うことで商品同士の類似画像検索を行っている [16] が，全く異なる構図の商品画像から人間の目視のような曖昧な基準によって商品の同一性を判断することは難しい．また，商品画像を用いて機械的に一致不一致の判断を行うには，大量生産品が出品されている EC サイトから商品画像を取得するために膨大な数のリクエストが必要となるが，比較対象とするサイト自体の網羅や常に更新される商品情報の大規模なクローリングは容易ではなく [14][12]，運営者の管理下にないサイトに対して大規模なクローリングを行うことは倫理的にも好ましくない．

2.3 キーワード検索によって機械的に大量生産品を検出することの課題

商品名や商品説明文のようなテキスト情報による大量生産品の検出について，minne では既知の大量生産品に含まれる語を用いて，単純なキーワード検索による検出も行っている．しかしながら，EC サイトのように商品種別が常に追加される状況では，追加される未知の語彙に対して追従しつづけることが必要となり，現実的には既知のブランド名のような語彙以外での検出は困難である．また，2 章の序文でも述べた通り，大量生産品であってもその商品名と商品説明文ではハンドメイド作品であるかのように謳っている場合や，ノーブランド品のように既知の語彙を含まない場合は，キーワード検索による検出は難しい．

ここまで述べたことから，従来の大量生産品の検出における課題は以下のようにまとめることができる．

(1) 目視の画像の比較では検出可能な数に限界があるため継続的に大量生産品の検出を行うことができない
(2) 画像比較をそのまま機械化することは基準の曖昧さと規模の問題が存在するため困難である
(3) キーワード検索によって機械的に検出を行うことは未知語に対して頑健でない問題から困難である

3. 提案手法

2 章で述べた課題を解決するためには，以下を満たす必要がある．

(1) 外部情報との比較を必要としない判断基準で大量生産品を検出する
(2) 機械化可能な方法によって大量生産品を検出する

本報告では，上記の 2 つの要件を満たすシステムとして，出品者単位の情報によって大量生産品の出品者を同定することで，外部情報との比較を必要とせず，継続的かつ機械的に大量生産品を検出する手法を提案する．手法の提案に

```
import ngram

def tokenizer(text, n):
    index = ngram.NGram(N=n)
    return [token for token in index.ngrams(index.pad(text))]

tokenizer('handmade', 2)
# ['$h', 'ha', 'an', 'nd', 'dm', 'ma', 'ad', 'de', 'e$']
```

図 1　テキストから 2-Gram トークンの抽出を行う実装例
Fig. 1　Implementation examples of extracting 2-Gram tokens from text.

あたり，大量生産品の出品者は，アカウント停止を回避するために多数のアカウントを作成しアカウント名に無意味な文字列を設定する傾向があることと，商品の削除に備えて出品数を増やすために国外 EC サイトの商品を翻訳しており商品名や商品説明文の日本語が不自然であるという 2 つの特徴を見出した．これらの特徴を用いて大量生産品の出品者を同定することを考えると，上記の 2 つの要件に対してそれぞれ (1) ハンドメイド作品を対象とした EC サイトの内部情報であるため，外部情報との比較によらず判断を行うことができ，(2) テキスト処理による特徴量抽出と分類器を実装することで機械化が可能である．

3.1　アカウント名から大量生産品の出品者を検出する

大量生産品の出品者のアカウント名には無意味な文字列が多く見られる一方で，ハンドメイド作品の出品者のアカウント名には屋号などの名前を示す文字列が多く見られるため，文字列中の文字の組み合わせを特徴量として扱うことで分類を行う．このような文字の組み合わせを特徴量として利用する際は，任意の n 文字の連続した文字列である N-Gram トークンによって，文字単位の共起関係を扱うことができる [5]．python-ngram ライブラリ [2] を用いて，テキストから 2-Gram トークンの抽出を行う Python コードを図 **1** に示す．

また，本報告では文字単位の共起関係に加えて，トークンの同時発生確率を分類器に組み込むため，単純ベイズ分類器を採用する．機械学習のライブラリである scikit-learn[8] を用いて，N-Gram トークンを特徴量として抽出して単純ベイズ分類器を訓練する Python コードを図 **2** に示す．

人間が生成する意味のある文字列に対して無意味な文字列を分類できることは，人間が生成する意味のあるドメイン名に対して，ボットネットによってアルゴリズム的に生成されるドメイン名の N-Gram トークンに固有のパターンがあることで示されている [7] が，本研究におけるハンドメイド作品と大量生産品の出品者の分類においては，文字単位の出現頻度である 1-Gram トークンと，隣り合う文字の共起関係を表す 2-Gram トークンの両方を特徴量として用いることでモデルがもつ情報量を増やし精度の向上を図る．

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB

vectorizer = CountVectorizer(analyzer='char_wb',
                             ngram_range=(1, 2))
vectorized_text = vectorizer.fit_transform(X)

X_train, _, y_train, _ = train_test_split(vectorized_text, y)
classifier = MultinomialNB()
classifier.fit(X_train, y_train)
```

図 2　N-Gram トークンを特徴量に用いて単純ベイズ分類器を訓練する実装例
Fig. 2　Implementation examples of training naive bayes classifier with N-Gram tokens.

3.2　商品名と商品説明文から大量生産品の出品者を検出する

提案手法では，EC サイトの内部情報である商品名と商品説明文を出品者単位で扱い，ハンドメイド作品の出品者と大量生産品の出品者を分類する．文章の書き手の同定にはしばしば単語や文節を特徴量として利用する [9][10][13]．本研究では，商品名と商品説明文を連結した文字列を出品者ごとにまとめて 1 つの文章と見なし，単語の特徴量変換を行った．3.1 節で述べた通り，文章の単語分割と特徴量変換は scikit-learn ライブラリの CountVectorizer クラスによって行うことができるが，商品名と商品説明文はほぼ全てが日本語で記述されているため，日本語の単語分割を行う解析器を実装する必要がある．また，品詞によって大量生産品の出品者による商品名と商品説明文の特徴を表す程度に差があると考えられるため，解析器は特定の品詞の語を取り出すようにする．日本語の形態素解析エンジン MeCab[3] を用いて名詞を取り出す解析器の Python コードを図 **3** に，実装した解析器を CountVectorizer で用いて日本語の文章から単語分割と特徴量抽出を行い単純ベイズ分類器を訓練する Python コードを図 **4** に示す．

ここまでに示した手法によって機械的に大量生産品を検出することで，従来の目視による検出に比べて時間的なコストの削減が期待できる．

4.　実験と考察

本報告では，minne の運営者によって目視で検出された大量生産品の出品者のアカウントと，同様に運営者によってハンドメイド作品であることが確認されている「ピックアップ作品[*1]」に掲載された出品者のアカウントを使用して分類器の構築を行った．提案手法の有効性を検証するため，3 章で実装した 2 つの手法について，それぞれテスト

[*1] 運営者の推薦商品を紹介する EC サイト内のコンテンツ

```
class WordDividor:
    def __init__(self):
        self.dic = '/path/to/dic/mecab-ipadic-neologd'
        self.tag = '-Ochasen -d ' + self.dic
        self.tagger = MeCab.Tagger(self.tag)

    def extract_words(self, document):
        if not document:
            return []

        words = []
        pattern = re.compile('^[0-9]')

        self.tagger.parse("")
        node = self.tagger.parseToNode(str(document))
        while node:
            features = node.feature.split(',')
            if features[0] == "名詞":
                word = node.surface
                if word != '':
                    words.append(word)

            node = node.next
        return words
```

図 3 日本語の単語分割を行い名詞を取り出す解析器の実装例

Fig. 3 Implementation example of analyzer which performs Japanese word segmentation and extracts nouns.

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB

wd = WordDividor()
vectorizer = CountVectorizer(analyzer=wd.extract_words)
vectorized_text = vectorizer.fit_transform(X)

X_train, _, y_train, _ = train_test_split(vectorized_text, y)
classifier = MultinomialNB()
classifier.fit(X_train, y_train)
```

図 4 日本語の文章から単語分割と特徴量抽出を行い単純ベイズ分類器を訓練する実装例

Fig. 4 Implementation examples of training naive bayes classifier with features from Japanese word segmentation.

データにおける精度を示す．また，分類器の構築に用いていない出品者のデータを分類した上で，大量生産品の出品者であると分類されたアカウントを運営者が目視で確認することにより，実際に EC サイトから大量生産品の出品者を検出することができるか評価を行った．

4.1 アカウント名から大量生産品の出品者を検出する分類器の評価と考察

分類器の構築にあたり，大量生産品の出品者のアカウント名とハンドメイド作品の出品者のアカウント名における 1-Gram トークンの特徴を確認するため，それぞれの度数分布を**図 5** と**図 6** に示す．ハンドメイド作品の出品者のアカウント名では日本語の母音である a,i,e,o の度数が特に

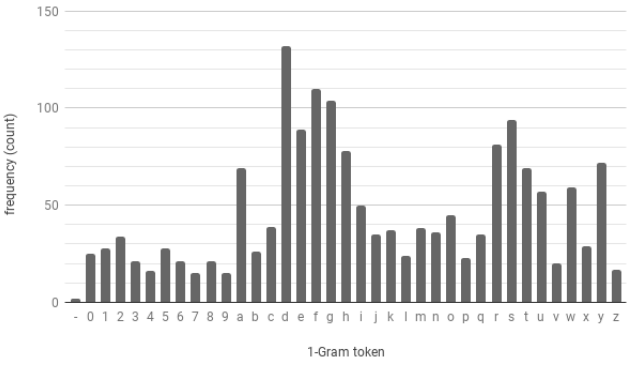

図 5 大量生産品の出品者のアカウント名における 1-Gram トークンの度数分布

Fig. 5 Frequency distribution of 1-Gram tokens in account names of sellers of mass-produced goods.

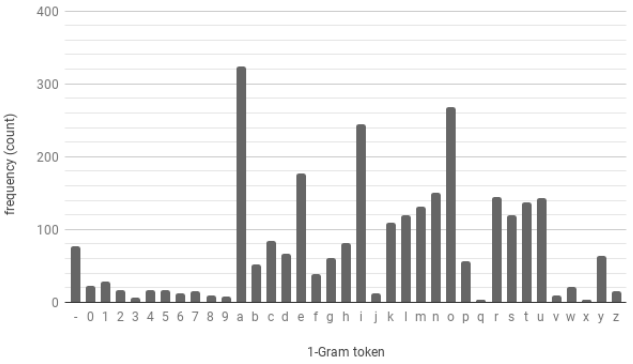

図 6 ハンドメイド作品の出品者のアカウント名における 1-Gram トークンの度数分布

Fig. 6 Frequency distribution of 1-Gram tokens in account names of sellers of handmade goods.

高く，数字や日本語の子音に現れない q,v,x,z の度数は低い．一方で，大量生産品の出品者のアカウント名ではの通りこのような特徴は見られず，-の度数が著しく低い．

このように 1-Gram トークンの分布に特徴が見られることから，分類器による大量生産品の検出が可能であると考え，minne の運営者によって目視で検出された大量生産品の出品者のアカウント名 181 個と，同様に運営者によってハンドメイド作品であることが確認されている「ピックアップ作品」に 2018 年 1 月 1 日以降に掲載された出品者のアカウント名 326 個を使用して分類器の構築を行った．特徴量にはアカウント名の 1-Gram トークンと 2-Gram トークンの両方を用いる．訓練データ 456 個，テストデータ 51 個に分割して，大量生産品の出品者のアカウント名とハンドメイド作品の出品者のアカウント名を分類した際のテストデータにおける混同行列を**表 1** に示す．

大量生産品の出品者のアカウント名について 21 件中 18 件を正しく大量生産品の出品者のアカウント名であると分類できた．また，ハンドメイド作品の出品者のアカウント

表 1　テストデータにおける混同行列
Table 1　Confusion matrix in the test data.

		正解ラベル	
		大量生産品	ハンドメイド作品
予測ラベル	大量生産品	18	1
	ハンドメイド作品	3	29

名について 30 件中 29 件を正しくハンドメイド作品の出品者のアカウント名であると分類することができ，精度は 92.16%であった．一方，誤検知について，ハンドメイド作品の出品者についてアカウント停止等の対応を誤って行った場合，大量生産品の削除によって守ろうとしている EC サイトへの信頼感やブランドイメージを却って損なうことに繋がるため，特にハンドメイド作品の出品者を大量生産品の出品者であると誤って検出する率が低いことが好ましい．今回の評価では該当の件数は 51 件中 1 件に留まっており，実際の運用ではシステムによる検出結果を人間が目視した上でアカウント停止等の対応が行われるため，誤ってアカウント停止などの対応が行われることは十分に防ぐことが可能であると考えられる．

また，特徴量として用いる N-Gram トークンの次数について，文字単位の出現頻度である 1-Gram トークンのみを特徴量として利用した場合も大量生産品の出品者のアカウント名について 21 件中 18 件を正しく検出することができたが，ハンドメイド作品の出品者のアカウント名を正しく分類できた件数は 30 件中 27 件であり，精度は 88.24%であった．隣り合う文字の共起関係を表す 2-Gram トークンのみを特徴量として利用した場合は大量生産品の出品者のアカウント名を正しく分類出来た件数が 21 件中 16 件，ハンドメイド作品の出品者のアカウント名を正しく分類できた件数が 30 件中 29 件であり，精度は 1-Gram トークンのみを用いた場合と同じく 88.24%であった．このことから 3.1 節で述べた通り，特徴量にアカウント名の 1-Gram トークンと 2-Gram トークンの両方を用いることで，文字単位の出現頻度と文字同士の共起関係がモデルに組み込まれ，精度が向上したと考えられる．

一方，分類器の構築に用いていない未知の出品者のデータとして，ある 1 週間にアカウントを登録して商品の出品を行ったアカウント名 544 個を分類器に適用したところ，12.9%にあたる 70 個が大量生産品の出品者のアカウント名として検出された．これらを運営者によって目視で確認したところ，少なくとも検出されたアカウントのうち 24.3%にあたる 17 個が大量生産品の出品者のアカウントであった．EC サイトの利用者の情報にあたるためアカウント名を示すことはできないが，ハンドメイド作品の出品者のうち大量生産品の出品者と分類されたアカウント名の多くは英字と数字を組み合わせたアカウント名であり，子音が連続しているなど人間の視点でランダムな文字列のように見えるものも多い．これらは分類器を用いてアカウント名のみから正しく分類することができなかったが，大量生産品の出品者は短期間に多くの商品を登録する傾向が確認されているため，このような基準を追加することで更に絞込を行うことができると考えられる．

表 2　学習データに用いる品詞ごとのテストデータに対する分類器の精度
Table 2　Accuracy of classifier for test data for each part of speech used for learning data.

品詞	精度
感動詞	70.59%
記号	70.59%
形容詞	94.12%
助詞	72.55%
助動詞	64.71%
接頭詞	74.51%
動詞	86.27%
副詞	70.59%
名詞	98.04%

時間的なコストの観点から従来の目視による検知と比較すると，未知の出品者のデータ 544 個について，すべて目視で検査する場合は 40 分程度の時間が掛かるが，分類器を用いる場合は 1 ミリ秒程度で 70 個の検出が完了し，これらを目視によって検査する時間は 5 分程度に短縮される．また，実運用において 1 日分の登録作品に対して 1 日 1 回検査を行う場合の分類器による検出は 0.1 秒以下で完了する．検出結果を目視によって検査する時間も 90 分程度となり，時間的なコストを大幅に削減することができる．

4.2　商品名と商品説明文から大量生産品の出品者を検出する分類器の評価と考察

minne の運営者によって目視で検出された大量生産品の出品者のアカウント 179 個と，minne の運営者によりハンドメイド作品であることが確認される「ピックアップ作品」に 2018 年 1 月 1 日以降に掲載された出品者のアカウント 330 個[*2]がもつ商品の商品名と商品説明文を用いて分類器の構築を行った．ただし，ハンドメイド作品の出品者については 1 アカウントあたりの商品数が多い場合があり EC サイトのデータベースからのデータ抽出に長時間を要するため，「ピックアップ作品」に掲載された商品のみを扱う．訓練データ 458 個，テストデータ 51 個に分割し，分類に有用な品詞を求めるため単語の解析器が取り出す品詞ごとに構築を行う．テストデータに対して大量生産品の出品者とハンドメイド作品の出品者を分類した際の分類器の精度を，学習データに用いる品詞ごとに**表 2** に示す．

テストデータを 90%以上の精度で分類できた品詞は名詞

[*2] EC サイトでは利用者によるデータの変更がリアルタイムに行われるため，学習データとしてデータベースから抽出されたアカウント数が，アカウント名から大量生産品の出品者を検出する分類器と異なる

表 3　テストデータにおける混同行列
Table 3　Confusion matrix in the test data.

		正解ラベル	
		大量生産品	ハンドメイド作品
予測ラベル	大量生産品	21	1
	ハンドメイド作品	0	29

であった．これは，大量生産品の商品名や商品説明文について，新品か中古品かを示していると思われる「新旧」，アナログ時計を示していると思われる「指針」などの独特の言葉遣いがされていることを捉えていると考えられる．更に，最もよくテストデータを分類できた名詞を用いる分類器の混同行列を**表 3** に示す．大量生産品の出品者のアカウントについては 21 件すべてを大量生産品の出品者のアカウントであると正しく分類できた．ハンドメイド作品の出品者のアカウントについては 30 件中 29 件をハンドメイド作品の出品者のアカウントであると分類することができ，精度は 98.04%であった．

一方，分類器の構築に用いていない未知の出品者のデータとして，ある 1 ヶ月間に登録されたアカウントで商品を登録しているもののうち，分類器の構築に使用していない 720 個を抽出し分類器に適用したところ，19.7%にあたる 142 個が大量生産品の出品者であると分類された．これらを運営者によって目視で確認したところ，少なくとも分類されたアカウントのうち 7.7%にあたる 11 個が大量生産品の出品者のアカウントであった．テストデータに対しては高精度に分類できている一方で，未知のデータに対してはハンドメイド作品の出品者を大量生産品の出品者であると誤って分類しており，過学習していることが考えられる．

ただし，未知の出品者のデータ 720 個について，すべて従来通り目視で検査する場合は 50 分程度の時間が掛かるが，分類器を用いる場合 0.3 秒程度で 142 個の検出が完了し，これらを目視によって検査する時間は 10 分程度に短縮される．また，実運用において 1 日分の登録作品に対して 1 日 1 回検査を行う場合，分類器による検出は十数秒程度で完了し，検出結果を目視によって検査する時間も 2 時間余りとなるため，現実的な時間で全数検査を行うことが可能となる．

5. まとめ

本報告では，ハンドメイド作品を対象とした EC サイトにおいて外部情報によって大量生産品を検出せざるを得ない状況において，外部情報に依存せず，継続的かつ機械的に大量生産品の出品者を検出するシステムを提案した．そして提案手法の有効性を示すために，運営者によって検出された大量生産品の出品者と，同様に運営者によってハンドメイド作品であることが確認されている商品の出品者について分類器を作成し，テストデータに対して高精度に分類を行えることを検証した．しかし，未知のデータに対しては，ハンドメイド作品の出品者について大量生産品の出品者であると誤って検出する傾向にある．一方で，従来の目視のみによる検出に対して分類器の検出結果を目視することで，検出に掛かる時間的なコストを大幅に減少させることが可能であり，現実的な時間で全数検査を行えることが示された．2 つの分類器を運用する際は，どちらも EC サイトのデータベースからデータの抽出を行う Python スクリプトと，モデルによる予測を行う Python スクリプトが動作するためのサーバーが必要だが，既に述べた通りモデルによる計算量は最大でも数十秒で完了する程度に収まるため，大量の計算資源を必要とすることはない．今後の課題としては，未知データに対しても頑健となる特徴量をモデルに組み込むことで，より効率よく分類器による検出を行う仕組みを検証する必要がある．

参考文献

[1] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich: Going Deeper with Convolutions, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015
[2] gpoulter/python-ngram, `https://github.com/gpoulter/python-ngram/`.
[3] MeCab, `http://taku910.github.io/mecab/`.
[4] minne, `https://minne.com/`.
[5] Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, Jenifer C. Lai: Class-based n-gram models of natural language, Computational linguistics, Vol. 18, No. 4, pp. 467-479, 1992
[6] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner: Gradient-based learning applied to document recognition, Proceedings of the IEEE, Vol. 86, No. 11, pp. 2278-2324, 1998
[7] Sandeep Yadav, Ashwath Kumar Krishna Reddy, A.L. Narasimha Reddy, Supranamaya Ranjan: Detecting algorithmically generated malicious domain names, Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, pp. 48-61, 2010
[8] scikit-learn, `http://scikit-learn.org/stable/`.
[9] 金 明哲: 文節パターンに基づいた文章の書き手の識別, 日本行動計量学会行動計量学, Vol. 40, No. 1, pp. 17-28, 2013
[10] 金 明哲, 村上 征勝: ランダムフォレスト法による文章の書き手の同定, 統計数理研究所統計数理, Vol. 55, No. 2, pp. 255-268, 2007
[11] GMO ペパボ株式会社: 国内最大のハンドメイドマーケット「minne(ミンネ)」3 年で約 10 倍に急増！ 2017 年の年間流通総額が 100 億円突破！～“ものづくり”のスキルの「シェアリングエコノミーサービス」として注目～, 2017 `https://pepabo.com/news/press/201712251500`.
[12] 田村 孝之, 喜連川 優: 大規模 Web アーカイブのための更新クローラの設計と実装, 電子情報通信学会 第 18 回データ工学ワークショップ論文集, B9-5, 2007
[13] 中島 泰, 山名 早人: 品詞と助詞の出現パターンを用いた類似著者の推定とコミュニティ抽出, 第 3 回データ工学と情報マネジメントに関するフォーラム, B6-5, 2011
[14] 廣瀬 信己: 国立国会図書館におけるウェブ・アーカイビングの実践と課題　－インターネットを安定的な知的社会資本とするために, 情報処理学会研究報告, Vol.

2003-DBS-130, No. 51, pp. 95-112, 2003
[15] 一般社団法人日本ホビー協会: ホビー白書 2017 年版, 2017
[16] 三宅 悠介, 松本 亮介, 力武 健次, 栗林 健太郎: 特徴抽出器の学習と購買履歴を必要としない類似画像による関連商品検索システム, 情報処理学会研究報告, Vol. 2017-IOT-37, No. 4, pp. 1-8, 2017