論文データの Yahoo!知恵袋カテゴリへの自動分類

重田 識博† 難波 英嗣‡ 竹澤 寿幸‡

†広島市立大学 情報科学部 〒731-3194 広島県広島市安佐南区大塚東 3-4-1

‡広島市立大学大学院 情報科学研究科 〒731-3194 広島県広島市安佐南区大塚東 3-4-1

E-mail: †‡{shigeta, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

あらまし 専門知識を持たない一般人が、難解な学術情報に容易にアクセスできるシステムの開発を目指しており、その第一歩として、Q&A コミュニティ「Yahoo! 知恵袋」のカテゴリに論文を自動分類するシステムを構築する。本研究では、知恵袋の回答欄に論文データベース CiNii へのリンクを含むエントリに着目する。エントリに付与された知恵袋カテゴリをリンク先の論文のカテゴリと見なすことで、知恵袋カテゴリ付き論文データを作成した。このデータを用いて実験を行い、提案手法の有効性を確認した。

キーワード 文書分類, 学術論文, Yahoo! 知恵袋, ジャンル横断文書分類

1. はじめに

近年、テレビ、新聞、Web など、様々なメディア上で、科学技術に関する膨大な情報が流通している。その中には不確実あるいは誤った情報も少なからず含まれており、こうした情報が非専門家を混乱させたり不安に陥れたりするだけでなく、風評被害などで実害を与える場合すらある。こうした状況において、誰もが安心して安全な生活を営むために、非専門家でもより信頼性の高い学術情報に容易にアクセスできる技術が必要とされている。そこで、本研究では、学術論文を Yahoo! 知恵袋のカテゴリに自動的に分類する手法を提案する。

Yahoo! 知恵袋とは、日常生活上の様々な疑問を投稿すると、その疑問についてネット上の誰かが答えてくれる、というサービスである。例えば、「体力的に弱くて困っています。おすすめのサプリを教えてください。ビタミンとかいろいろありますが、なにかおすすめありますか？」という質問に対して、「ビール酵母とか凄く良いですよ。ビール酵母意外にも色々入ってるのを買った方が効果が望めます。」のように回答される。サプリのおすすめに関する質問が約 5000 件投稿されている。また、おすすめに関する質問は 14 万件以上存在する。

このサービスでは、過去に投稿された質問および回答事例を効率的に検索できるように、すべての質問−回答事例がカテゴリに分類されている。この分類体系(以下、知恵袋カテゴリ)は、第一階層 17 種類、第二階層 113 種類、第三階層 434 種類の三階層で構成され、全 564 カテゴリ存在する(2014 年 10 月現在)。図 1 に Yahoo! 知恵袋カテゴリの一例を示す。この分類体系(以下、知恵袋カテゴリ)に論文を分類すれば、知恵袋の質問−回答事例を探すのと同じように論文を探すことができるようになり、非専門家による学術情報への容易なアクセスが部分的に実現できると考えられる。また、論文を検索する学生に対しても、検索時間の軽減や目的の論文の検索が容易になることも期待される。そのため、Yahoo! 知恵袋のカテゴリに論文を分類することが必要であると考える。知恵袋カテゴリに論文を分類するためには、人手によるカテゴリ付与を行うことで実現可能であるが、これは、多くの時間を要する。この作業を機械学習によるカテゴリ自動分類で代用できると考える。また、自動分類を実現するためには、論文情報にカテゴリを付与したデータを準備し、正解データとして用いる必要がある。本研究では、学術論文データベース CiNii の論文を知恵袋カテゴリに自動分類し、実験により、その有効性を確認する。

これまでにも、既存の分類体系に文書集合を自動分類する研究は多く存在する。外部サイトの分類体系に基づいて文書集合を自動分類するアプローチとして、大量の文書集合を訓練用のデータセットとして用い、特徴ベクトルを素性に機械学習を行っている。「Support Vector Machine (SVM)」や、「k 近傍法(k-NN 法)」という分類手法を用いている。このような従来の手法に加えて、本研究では、「パーセプトロン」というニューロンをモデルとした古典的な機械学習アルゴリズムの一つを用いて分類を行う。

本論文の構成は以下のとおりである。次節では、関連研究について述べる。3 節では、カテゴリ毎における論文の自動分類手法について述べ、4 節では、有効性を調べるために行った実験について報告し、結果について考察する。最後に 5 節で本稿をまとめる。

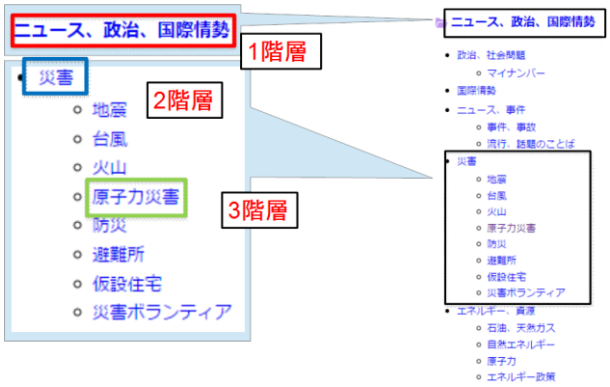

図 1：Yahoo! 知恵袋カテゴリの一例

2. 関連研究

2.1. Q&A サイト質問文の質問タイプへの自動分類

渡邊ら[1]は、Q&A サイトの質問を「事実」「根拠」「経験」「提案」「意見」の 5 つのタイプへ複数の判定者によって判定する分類実験を行っている。質問タイプの定義として、まず 2 種類に大別できると提案している。唯一の正解が存在する質問文と、そうでない質問文である。唯一の正解が存在する場合、事実を求める質問と、根拠も求める質問が存在するという。また、唯一の正解が存在しない場合、経験を求める質問と、提案を求める質問が存在するという。これらに属さない質問として意見を求めていると定義して、5 種類にタイプ分類を行っている。素性は、経験則によって選択した 70 語を用いて機械学習により自動分類をおこなっている。

さらに渡邊ら[2]は 5 つのタイプに加えて 6 つの質問要求属性を定義している。質問要求属性とは、質問者は質問の回答とは別に回答の速さや正確さといった期待をもって質問を投稿していると考えられる。確信性、多様性、独創性、客観性、保証性、要約性の 6 つを質問要求属性と定義して分類も行っている。本研究は、渡辺らの質問タイプの「事実」や「根拠」、「意見」にあたる回答に用いられた論文を使用する。Q&A サイトを使用すること、素性として単語を使用することが共通しているが、分類対象が質問ではなく、本研究では回答を用いて分類する。また、質問タイプへ分類するのではなく、知恵袋カテゴリへ分類するという点が異なる。

林ら[3]は渡邊らの 5 つの質問タイプの再定義をおこない、Q&A サイトから複数の質問が含まれていない質問文を人手で抽出後、キーワードによる手法と語の頻度を利用したスコア付けによる手法の 2 つを用いて自動分類をおこなっている。キーワードによる分類では、単語を人手で質問文から収集したキーワードに対して、5 つの質問タイプを対応させている。語の頻度による分類では、質問タイプごとに単語頻度を集計して、スコアを算出している。Q&A サイトからキーワード抽出や、語の頻度を用いて分類を行う部分では、似ているが本研究は、人手による単語抽出をしておらず、機械学習により単語を抽出している。

2.2. Q&A サイトの回答文の信頼性と回答補完

瀧ら[6]は、Q&A サイトの回答文を対象として、投稿した質問に対して信頼性の高い情報を効率的に入手するためのシステム構築を目的としている。そのため、どのような回答に対して信頼性を感じるかを様々な要素に対して、分析・実験を行っている。本研究と関連する部分として URL を用いた回答は、信頼度の高い回答という結果が得られている。しかし、URL を用いた回答文は全体の回答文に対しての割合が低いことが分かっている。瀧らの研究から、URL を用いた回答の方が高い信頼性を得られるが、用いたリンク先の信頼性までは、わからない。本研究では、用いるリンク先が論文であることから信頼度が高い回答が期待できる。通常のサイトを提示するよりさらに高い信頼度を持った回答ができると言える。

高田ら[7]は、Q&A サイトに投稿された回答文のうち、回答の信憑性や、情報欠落した回答に対して Web ページから回答に適した URL の提示するいわゆる回答補完を行っている。質問文の中でも正解を求める質問ではなく、回答者の意見や情報を求める質問文を対象とし、に質問文から単語を抽出し、検索クエリとして用いて回答として適したと判定される Web ページを提示する実験を行っている。本研究でも、Q&A サイトの単語を素性とし、URL の提示を行うことを目標としている。高田らの研究でも提示するサイトの信頼性は、わからない。

2.3. 学術論文情報を用いた自動再分類

柏木ら[8]は、論文のアブストラクトを用いて論文分類を試みている。彼らは、Learning Vector Quantization を用いて、原子分子物理学分野の論文について分類を行っている。あるカテゴリに分類されている論文に対して、アブストラクトの情報だけを用いて再分類を行っている。正しいカテゴリに分類することよりも多くの論文を収集することを重要視して実験を行っている。本研究でも、論文のアブストラクトを用いるが、論文分野は特定せず、知恵袋に引用された論文を対象として実験を行う。

榊ら[9]は、時系列的な変化を考慮した論文のカテゴリへの分類を行っている。過去の論文が分類されたカテゴリを正解として、現在使用されているカテゴリに属した論文に対して、再分類を行っている。従来の文書分類の技術と文書クラスタリングの技術を組み合わせて、新たな分類方法を提案している。論文を分類体系に分類する研究であり、本研究と類似するが、本研究では、一度も分類されていない論文を分類された

論文を基準に分類するため、時系列を考慮する必要がないと言える。

2.4. 他の分類体系に基づいた論文自動分類

福田ら[10]は、学術論文を効率的に検索するため、学術論文に対して特定の分類体系に基づく分類コードを自動付与する手法を提案している。科学研究費助成事業データベース(KAKEN)を対象に、KAKEN で定められている分類体系に基づき、国立情報学研究所が運営する学術論文情報データベースである CiNii article に収録されている学術論文を自動的に分類することを行っている。本研究でも福田らが使用した手法の一部を用いて実験を行う。先行研究と異なる部分として、学術論文を学術界の観点から分類を行っているため、使用用途が異なり、本研究は、学術論文を非専門家でも探しやすくするための分類を行う。

難波ら[11][12]は、学術論文の国際特許分類への自動分類を行っている。日本語論文や英語論文を、特許分類体系の「IPC」に自動分類している。IPC とは、「セクション」、「クラス」、「サブクラス」、「メイングループ」、「サブグループ」の 5 階層から構成・分類されている。この研究では、最下層の「サブグループ」レベルの IPC コードを論文に付与することを目的としている。本研究でも、他の分類体系を用いてデータセットを作成し論文を自動分類している。難波らは、学術論文を産業界の観点から分類を行っているため、本研究と使用用途が異なる。

3. 論文データの Yahoo!知恵袋カテゴリへの自動分類

3.1. データ収集

3.1.1. Yahoo! 知恵袋データ

本研究では、知恵袋の回答欄に論文データベース CiNii へのリンクを含むエントリに着目する。エントリに付与された知恵袋カテゴリをリンク先の論文のカテゴリと見なすことで、知恵袋カテゴリ付き論文データを作成した。このデータは、計 442 件あり、提案手法を評価するのには十分な数であると考えられるが、図 2 のように機械学習に基づく手法で分類器を構築する際の訓練用データとしては少なすぎる。そこで、図 3 のように Yahoo! 知恵袋の質問−回答対およびこれらに付与されたカテゴリを、論文の分類器を構築する上での擬似的な訓練用データとして用いる(約 1600 万件)。また、訓練用データから取得したカテゴリ数を表 1 にまとめる。その訓練データでは、一つの質問に対して CiNii へのリンクを複数用いた回答が存在する事例もある。その場合、すべての CiNii 論文に対して質問に付与されたカテゴリを付与する。

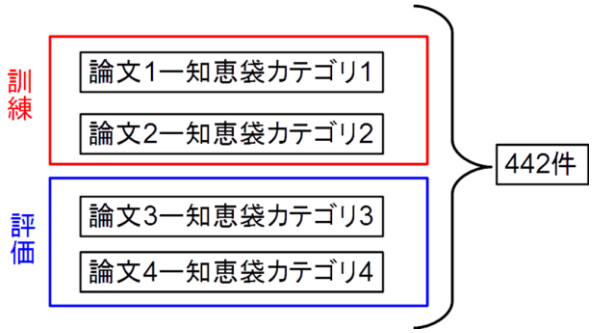

図 2：機械学習が困難な学習方針

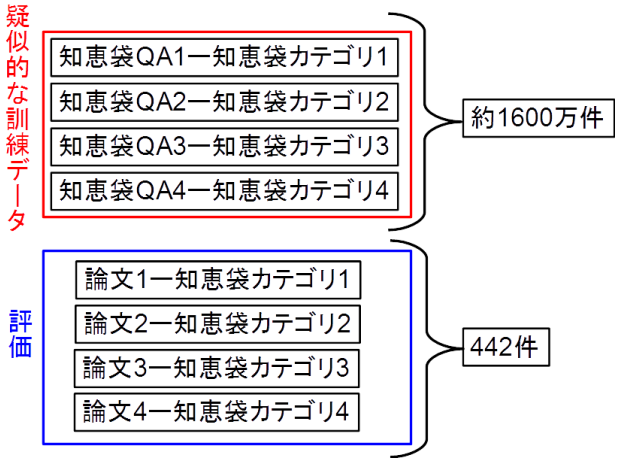

図 3：本研究の学習方針

表 1：訓練データのカテゴリ数

全カテゴリ	第三階層	第二階層	第一階層
453	348	105	16

3.1.2. 論文データの収集

3.1.1 で作成した知恵袋カテゴリ付き論文データについて説明する。論文を分類するため、論文データを収集する必要がある。CiNii のサイトから取得できる論文データとして、タイトル、著者名、キーワード、アブストラクトが存在する。本研究では、重要単語を多く含むタイトルとアブストラクトを使用する。キーワードも重要単語を多く含むが、使用しない理由として、形態素解析をする際に違う単語に分解してしまうことや、そのまま使用しても専門用語のピックアップであるため、分類精度に好影響を与えるとは言い難いためである。CiNii へのリンクを含む知恵袋エントリは 442 件であるが、論文の引用件数である 678 件を使用している。また、論文データのカテゴリ数を表 2 にまとめる。

表 2：論文データのカテゴリ数

全カテゴリ	第三階層	第二階層	第一階層
101	50	51	15

3.2. 提案手法

Yahoo! 知恵袋カテゴリを分類する手法として本研究では、Support Vector Machine (SVM) と Simple

perceptron (Perceptron) 、k-NN の三種類を用いる。SVM と Perceptron では、カテゴリと一致するかを判定する 2 値分類を行う。k-NN では、論文データと回答文の類似度を算出し、類似度の高い順にランキングする。CiNii へのリンク無の Yahoo! 知恵袋エントリの回答文の単語頻度を手がかりに、CiNii の論文データの単語頻度を与えることで、論文データを Yahoo! 知恵袋のカテゴリに分類する。

3.2.1. Support Vector Machine (SVM)

Support Vector Machine (SVM)は 2 値分類のための教師あり学習アルゴリズムである。高い汎化性能を持ち、様々な分野で広く用いられてきている。SVM を用いたカテゴリの分類では、知恵袋エントリの各カテゴリ数だけインデックスファイルから 2 値分類器を用意する。例えば、「病気、症状、ヘルスケア」という第三階層のカテゴリを持つエントリを対象にした場合、345 個の 2 値分類器を用意する。そして、入力クエリをそれぞれの分類器に与えた時に正例であると判断された場合、その分類器を示しているカテゴリ名に分類する。しかし、本研究では、入力クエリに対して少なくとも 1 つのカテゴリ名に分類しなければならないと定めている。しかし、SVM を用いた場合、次の 2 つのパターンが考えられる。まず、学習結果として正例が複数の学習器から出力された場合、候補となるカテゴリ名が複数存在することになる。また、学習結果として全ての分類器から負例が出力された場合、候補となるカテゴリ名が存在しないことになる。そこで本研究では、正例、負例に関わらず、最も計算結果が高かった学習器のカテゴリ名を入力クエリに分類するという手法を用いる。

3.2.2. Simple perceptron (Perceptron)

分類手法の Simple perceptron について説明する。Simple perceptron は、ニューラルネットワークの一種であり、入力層と出力層の 2 層から構成されている。学習能力を持つパターン識別器であり、線形分離不可能な問題は解けないという問題がある。n 次元の入力ベクトル x があったとき各成分をノードとして見て、これらを荷重ベクトル w で線形結合して出力 z を得る。従って、式(1)のように表され模式図では図 4 のようになる。本研究では、入力クエリに単語を使用し、知恵袋エントリごとに値を算出する。出力結果とエントリごとに定めた正解とする値の符号が違う場合に単語重みを更新していく。出力結果とエントリごとに定めた正解とする値の符号がすべて同じであれば学習を終了する。エントリごとに定める値は、対象としたカテゴリを正(+1)、それ以外のカテゴリを負(-1)と定める。

$$Z = \sum_{k=1}^{n} w_k x_k = w^T x \quad (1)$$

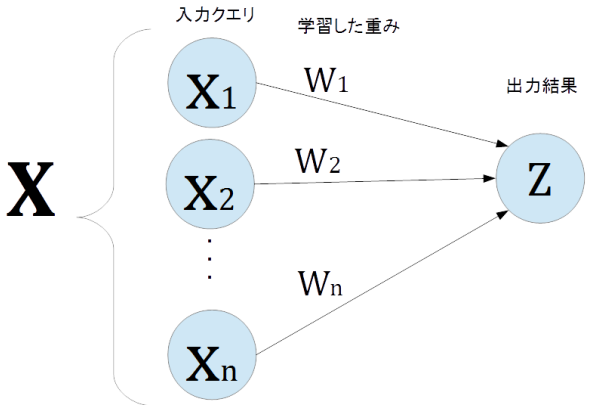

図 4：perceptron の模式図

線形分離できない場合に学習ができないため、収束条件を設けることが必要となる。今回の収束条件は知恵袋エントリを 200 回読み出したら終了する条件を設けている。また、以下の式により、精度(式 2)、再現率(式 3)を算出する。

$$\text{精度} = \frac{\textit{知恵袋カテゴリと学習結果の正である一致数}}{\textit{学習結果が正である数}} \quad (2)$$

$$\text{再現率} = \frac{\textit{知恵袋カテゴリと学習結果の正である一致数}}{\textit{対象としている知恵袋カテゴリ数}} \quad (3)$$

3.2.3. k-NN

k-NN 法では、知恵袋の回答文中の単語を素性として論文データとの類似度をランキングする。本研究では、福田らの研究で用いられた 3 種類のランキング手法をそれぞれ用いて、論文データに適合する上位 100 件(k=100) の知恵袋カテゴリのランク付けを行う。まず、上位 100 件の知恵袋エントリを抽出する。次に、抽出された知恵袋エントリに付与されたカテゴリの論文データに対するスコア値である $Score(c)$ を計算する。ここで、$Score(c)$とは、入力クエリ (本研究では論文データ) にラベル c (本研究ではカテゴリ名) が分類される可能性の尺度とする。そして、各カテゴリにおける $Score(c)$が高い順にソートし、上位 n 件までのカテゴリ名を入力論文データに分類する。本研究では、以下で述べる 3 種類のランキング手法を用いて、それぞれの精度を評価する。

(1) Naïve method

本手法におけるカテゴリ名のランキングは，抽出された知恵袋エントリにおいて、最も類似度が高かった回答文に付与されているカテゴリ名の順となる。(式 4)

$$Score_{Naive}(c) = \frac{1}{firstrank(c, \{d_1, d_2, \ldots, d_k\})} \quad (4)$$

◆ $firstrank(c, \{d_1, d_2, \ldots, d_k\})$: 抽出された知恵袋エント

リ$\{d_1, d_2, \ldots, d_k\}$において、カテゴリ名 c が最初に出現した順位を示す。もし、カテゴリ名 c が最初に出現した知恵袋エントリが d_i であれば、$firstrank(c, \{d_1, d_2, \ldots, d_k\}) = i$となる。

(2) Sum

Sum では、抽出された知恵袋エントリにおいて、カテゴリ名 c が付与されている全ての回答文の類似度の総和を、以下の式 5 を用いて算出する。

$$Score_{sum}(c) = \sum_{i=1}^{k} occur(c, d_i) \cdot sim(q, d_i) \quad (5)$$

- $occur(c, d_i)$: カテゴリ名 c が知恵袋エントリ d_iに付与されている$(occur(c, d_i) = 1)$かどうかを示す関数である。もし付与されていれば$occur(c, d_i) = 1$、付与れていなければ$occur(c, d_i) = 0$となる。
- $sim(q, d_i)$: 入力論文データ q と知恵袋エントリ d_i間の類似度。

(3) Listweak

上記で述べた Sum 手法に基づいたランキング手法であり、抽出された文書集合において、より類似度の高い文書を強調していく手法である。(式 6)

$$Score_{listweak}(c) = \sum_{i=1}^{k} occur(c, d_i) \cdot Sim(q, d_i) \cdot r_1^i \quad (6)$$

- r_1: 抽出された知恵袋エントリにおいて、より類似度の低い知恵袋エントリに対してペナルティを与えるパラメータ$(0<r_1<1)$。本研究でも福田らが採用した$r_1 = 0.95$と設定する。

4. 実験

4.1. 実験方法

Yahoo! 知恵袋の回答文を用いて論文を Yahoo! 知恵袋のカテゴリに分類する実験を行った。CiNii へのリンク無の Yahoo! 知恵袋エントリの回答文を分類対象の基準として、カテゴリ付き論文を与えることで、付与されたカテゴリに分類されるかの実験を行った。カテゴリ付き論文データのタイトルとアブストラクトに含まれる単語を素性とした。

前処理として、人手ですべてのカテゴリに共通して使用されている単語や記号などを除去した。例を挙げると、”半角数字”や、”、”、”/”、”ある”などがこれにあたる。しかし、すべての単語に対して行うことは不可能なため、出現頻度の上位に出てきた単語に対して人手で判定して除去を行った。

知恵袋回答文と論文データの単語類似度が高いカテゴリ名に論文を分類した。与える素性でアブストラクトの記載がない場合は、タイトルだけで行った。タイトル、アブストラクトと回答文の単語類似度の高い順にカテゴリが表示され、上位 n 件のカテゴリに論文のカテゴリが含まれている件数を集計し、精度を求めた。本研究では、n=1~5 の値を採用している。分類手法として SVM 、Perceptron、k-NN 法を用いた。

SVM では、345 カテゴリについて、各カテゴリに属するかどうかを判定する分類器を構築し、1 論文に複数のカテゴリが付与された場合は、分離平面からの距離が最も遠いカテゴリを、その論文のカテゴリとする。Simple perceptron では、単語重みの総和の出力結果が高い順に、k-NN 法では、スコアの高い順にランク付けを行った。

4.2. 実験結果

SVM での分離平面から距離順、k-NN 法の 3 種類でのランキング順、Perceptron の出力関数の結果順を上位 5 件までの結果を表 3 に示す。

表 3：論文の知恵袋カテゴリへの分類精度(3 階層)

	n=1	n=2	n=3	n=4	n=5
SVM	0.049	0.093	0.136	0.177	0.214
k-NN(Listweak)	0.149	0.215	0.251	0.282	0.310
k-NN(Sum)	0.140	0.189	0.236	0.280	0.310
k-NN(Naive)	0.111				
Perceptron	0.103	0.192	0.249	0.279	0.301

表 3 の結果から精度が高い手法は、k-NN 法であることから k-NN 法の各手法を用いてカテゴリ名の第二階層が一致している精度を表 4 に、また、第一階層が一致している精度に表 5 に示す。

表 4：論文の知恵袋カテゴリへの分類精度(2 階層)

	n=1	n=2	n=3	n=4	n=5
k-NN(Listweak)	0.237	0.320	0.357	0.400	0.432
k-NN(Sum)	0.223	0.285	0.332	0.378	0.414
k-NN(Naive)	0.205				

表 5：論文の知恵袋カテゴリへの分類精度(1 階層)

	n=1	n=2	n=3	n=4	n=5
k-NN(Listweak)	0.372	0.496	0.556	0.628	0.665
k-NN(Sum)	0.344	0.459	0.543	0.605	0.650
k-NN(Naive)	0.388				

表 3 において、k-NN 法 3 種類と Perceptron の上位 5 件すべての結果において perceptron と k-NN 法は、SVM よりも高い値を示した。また、SVM と比べて Perceptron や k-NN 法は、n の値が増加するごとに、値の増加値も大きいことから SVM よりも有効であることが分かる。Perceptron と k-NN 法を比較すると k-NN 法が高いため、3 手法で一番有効であることがわかる。この結果から k-NN 法を用いて階層を下げた実験を行った。

表 4 において、階層を一つ下げることで精度の向上が見られた。しかし、階層を下げたことで大きく精度

が向上すると予想していたが、n=1 では約 0.1 と低い向上であった。n=5 の結果であっても 0.5 未満の結果となった。

表 5 に、さらに階層を一つ下げて k-NN 法で実験を行った結果をまとめる。第二階層までの結果よりも精度が向上したが、上位 1 件では、0.1~0.18 程度の精度向上であった。上位 3 件より 0.5 を超える値が算出できた。しかし、一つのカテゴリに分類できる結果を目的としているため、n=1 の結果で 3 割ほどの正解カテゴリを含んでいるであることがわかった。

4.3. 考察

全体の 1 割程度しか分類できなかった原因の一つ目として、すべてのカテゴリに共通する単語の除去が不十分であることが挙げられる。具体的な単語として、「調べる」や「詳しい」、「思う」などは除くべきであると考える。しかし、どこまでが共通している単語であるかの線引きが非常に困難であるため、実現にはかなりの時間を要することが予測される。例えば、「%」のような単位を表す単語は、複数のカテゴリに出現する可能性があると考える。除去をすることで URL に含まれる「%」を除くことができるため精度向上すると考えられるが一方で、数学系のカテゴリでは、精度が低下すると考えられる。そのため、人手による判定を導入し、出現傾向を調査する必要があると考える。

次に挙げられる原因は、論文データは専門的用語を多く含んでいる。しかし、一般人のコミュニティでは、専門用語を用いることは少ない。例えば、「腎損癌」という単語は一般用語として用いられている。しかし、専門用語で「腎臓癌」は「腎癌」と表記される。また、「癌」をひらがなの「がん」と表記する事例も存在する。このことから、日常で頻度の低い漢字を使用しない傾向があることが分かる。そのため、専門用語を一般用語に変換する辞書を作成することで、精度向上が見込まれる。

他の誤り事例として、複数のカテゴリに分類可能である事例が存在した。例えば、カテゴリ付き論文のカテゴリが「トレーニング」であり、引用された論文タイトルが「ボクシングのパンチはどこから生まれるかストレートパンチ(右) の筋電図実験から」となっている例では、タイトルに「ボクシング」や、「パンチ」を含むことから知恵袋カテゴリの「ボクシング」というカテゴリにも分類します。このように、複数のカテゴリに分類される場合があるため、1 論文 1 カテゴリと決めずに、1 論文複数カテゴリを許すことを検討すべきだと考える。

5. おわりに

Yahoo! 知恵袋の回答文の単語を素性として論文を Yahoo! 知恵袋カテゴリに分類を行った。論文リンク無の知恵袋の回答文を分類基準として、知恵袋カテゴリ付き論文データを分類した。分類手法として SVM、Simple perceptron、k-NN 法を用いた。

実験で、3 つの手法によるカテゴリ分類を行い、上位 5 件で k-NN 法と Simple perceptron で 3 割程度の精度を得た。この実験で高い値を出した k-NN 法を用いて第一階層での分類、第二階層での分類を行い、第一階層の上位 5 件で 6 割程度の精度を得た。この結果から論文概要でのカテゴリ一致が可能であるかを検証できた。上位 3 件以上で 0.5 以上を示したが、カテゴリを一つに特定するいわゆる上位 1 件での精度向上を重要とするため、さらなる実験が必要である。単語の分析や、専門用語と一般用語の隔たりが原因であると考えられる。

参 考 文 献

[1] 渡邊 直人, 島田 諭, 関 洋平, 神門 典子, 佐藤哲司, “QA コミュニティにおける質問者の期待に基づく質問分類に関する一検討”, DEIM Forum 2011, B5-1, 2011.

[2] 渡邊 直人, 島田 諭, 関 洋平, 神門 典子, 佐藤哲司, “コミュニティ QA における質問の多面的評価法の検討”, 情報知識学会 第 19 回(2011 年度)年次大会, 情報知識学会誌, Vol.21, No.2, pp.163-168, 2011.

[3] 林 秀治, 山本 和英, “質問意図による QA サイト質問文の自動分類”, 電子情報通信学会言語理解とコミュニケーション研究会, NLC2013-10, pp.51-56, 2013.

[4] 西田 京介, 藤村 考, “階層的オートタギングによる Q&A コミュニティの知識整理”, 日本データベース学会論文誌, Vol.9, No.1, pp.1-6, 2010.

[5] 西田 京介, 星出 高秀, 藤村 考, 内山 匡, “階層的オートタギング技術とその応用”, 情報処理学会論文誌 データベース, Vol.6, No.1, pp.29‐40, 2013.

[6] 瀧寛文, 森崎修司, 大平雅雄, 松本健一, “Q&A コミュニティを対象とした回答の信頼性指標構築に向けた分析”, 情報社会学会誌, Vol.4, No.1, pp.49-58, 2009.

[7] 高田 夏希, 山本 祐輔, 小山 聡, 田中 克己, “質問応答コンテンツに対する Web による回答補完”, DEIM Forum 2009, C4-6, 2009.

[8] 柏木 裕恵, 高田 雅美, 佐々木 明, 城 和貴, “アブストラクトを用いた論文分類システムの設計と実装”, 情報処理学会研究報, pp33-36, 2006.

[9] 榊 剛史, 松尾 豊, 石塚 満, “制約付きクラスタリングを用いた論文分類”, 人工知能学会全国大会論文集, JSAI2006, pp.1-4, 2006.

[10] 福田 悟志, 難波 英嗣, 竹澤 寿幸, “要素技術とその効果を用いた学術論文の自動分類”, DEIM Forum 2015, F3-4, 2015.

[11] Nanba, H., Fujii, A., Iwayama, M., and Hashimoto, T. (2010) “Overview of the Patent Mining Task at the NTCIR-8 Workshop”. In Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 293-302.

[12] Nanba, H., Fujii, A., Iwayama, M., and Hashimoto, T. (2008) “Overview of the Patent Mining Task at the NTCIR-7 Workshop”. In Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of

Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 325-332.