特集：分類をみつめなおす part2

人工知能による文書分類

難波　英嗣*

近年，人工知能はコンピュータ囲碁や将棋，自動車の自動運転など，様々な分野で目覚ましい発展を遂げており，その成果をインターネット，新聞，テレビなどで目にする機会も少なくない。自然言語処理（NLP）は人工知能の一分野であり，人間が日常的に使っている言葉（自然言語）をコンピュータに処理させる技術のことを指す。人間が文書を分類する作業を，コンピュータで自動化することは，自然言語処理における代表的な研究課題のひとつである。本稿では，コンピュータによる文書分類に焦点を当て，様々な研究事例やその仕組みを紹介する。

キーワード：自然言語処理，文書分類，機械学習，フィルタリング，k 近傍法，ナイーブベイズ分類器

1. はじめに

近年，人工知能はコンピュータ囲碁や将棋，自動車の自動運転など，様々な分野で目覚ましい発展を遂げており，その成果をニュース等で耳にする機会も少なくない。自然言語処理は人工知能の一分野であり，人間が日常的に使っている言葉（自然言語）をコンピュータに処理させる技術のことを指す。人間が文書を分類する作業を，コンピュータで自動化することは，自然言語処理における代表的な研究課題のひとつであり，我々の日常生活の中でも広く使われている。例えば，受信した電子メールがスパム（迷惑）メールかどうかを判定するスパムフィルタリングは，我々の日常生活の中でも広く使われている，コンピュータを用いた文書分類技術の代表例である。また，Google 社が提供する Gmail というメールサービスには，電子メールの重要度を自動推定し，重要度が高いメールを優先トレイに振り分ける機能があるが，この機能を使うことで，ユーザが受信トレイを見る時間が 15%短縮された，という報告もある。

本稿では，コンピュータによる文書分類に焦点を当て，文書分類に馴染みのない読者を対象に，様々な研究事例やその仕組みを紹介する。また，文書分類の将来の可能性や課題について述べる。

2. コンピュータを用いた文書分類

2.1　様々な文書分類

自然言語処理における文書分類とは，一般に文書を特定の分類体系に自動的に割り当てる処理のことを指し，これまでに数多くの研究が行われている。以下に，その一部を挙げる。

1. トピック分類：学術論文を国際特許分類や科研費カテゴリなどの分類体系に分類する [1),2),3)]。
2. 著者推定：著者不詳の文書の著者を推定する [4)]。
3. 属性推定：ブログ著者の属性（居住域，性別，年齢）を推定する [5)]。
4. 言語識別：文書の言語を識別する [6)]。
5. 評判分析：商品，映画，サービスなどのレビュー文書が，肯定的なのか否定的なのかを判定する [5)]。
6. 品質評価：人間が書いた文書の品質を自動評価する [7)]。
7. フィルタリング：送付されてきた電子メールがスパムメールかそうでないか，あるいは優先的に読むべきかそうでないかを判定する。

「6. 品質評価」は，自然言語処理分野では文書分類として捉えられることはあまりないが，本稿では文書分類として扱う。1 節で述べた Gmail の優先トレイ機能は，受信したメールを重要度順にならべ，その上位を優先トレイに表示するが，入学試験や入社試験で実施される小論文試験において，受験者が書いた文書を質の高い順にならべ，一定水準以上のものを合格とする，といった状況を想定すれば，「6. 品質評価」も Gmail の優先トレイと同じ文書分類と考えられるであろう。

2.2　文書分類の種類

2.1 節で述べた文書分類は，以下に述べる 2 値分類または多値分類のいずれかに分けることができる。2 値分類とは，分類するカテゴリが 2 種類しかないもので，例えば，ブログの著者の性別の判定やスパムメール判定が該当する。これに対し多値分類とは，分類するカテゴリが 3 種類以上存在するものである。例えば，ある文書が何語で書かれているかを判定する言語識別の場合，日本語，英語，フランス語，ドイツ語などがカテゴリとなるため多値分類となる。

カテゴリが 3 種類以上存在するが，2 値分類と考えられる場合もある。例えば，カテゴリが A，B，C の 3 種類あ

＊なんば　ひでつぐ　広島市立大学大学院情報科学研究科
〒731-3194　広島市安佐南区大塚東 3-4-1
Tel. 082-830-1584　E-Mail: nanba@hiroshima-cu.ac.jp
（原稿受領　2016.3.28）

り，ある文書が A と B 両方に分類できる場合である。この場合は，文書がカテゴリ A に属するかどうか，B に属するかどうか，C に属するかどうかという 3 種類の 2 値分類を行うことになる。

分類問題が上記のどのケースに属するかによって，文書分類システムの構築方法が変わる。次節では，文書分類の仕組みについて説明する。

3. 文書分類の仕組み

本節では，まず，文書分類の基本的な考え方について説明する。次に，コンピュータによる文書分類の事例をいくつか取り上げ，その仕組みを紹介する。

3.1 文書分類の基本的な考え方

今，ある文書 X を，カテゴリ A，B，C のいずれかに分類する図 1 のような場合を考えてみる。なお，カテゴリ A，B，C には，すでに人手で複数の文書が分類済であるものとする。

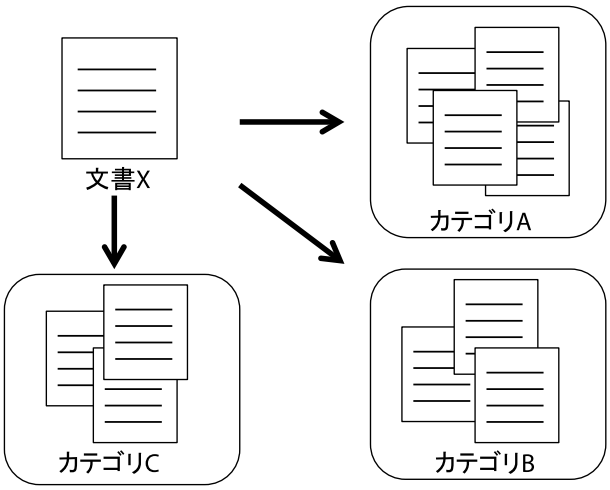

図 1 文書分類の例

手順は，以下のとおりである。

1. カテゴリ A，B，C に含まれるすべての文書から何らかの特徴をあらかじめ抽出しておく。
2. 同様に，文書 X からも特徴を抽出しておく。
3. 文書 X の特徴と似た特徴を数多く含む文書のカテゴリに文書 X を分類する。

ここで，(1)文書の特徴とは何か？(2)手順 3 において，どうやって文書 X と各カテゴリ中の文書の特徴が似ていると判断するのか？という 2 点について検討する必要がある。(1)文書の特徴としてもっともよく使われるのは，文書中の単語である。あるニュース記事が，金融，スポーツ，科学技術，エンターテインメントの中のどのカテゴリに分類されるかを判断する時，例えば，その記事にサッカーやゴールといった語が含まれていれば，スポーツカテゴリに分類される。これは，スポーツカテゴリ中の文書には，他のカテゴリ中の文書と比べると，サッカー，選手，野球，代表など，スポーツに関する用語が相対的に数多く含まれるからである。ただし，文書の特徴は，単語以外にも様々なものが利用される。これは，文書をどのようなカテゴリに分類しようとするのかに深く関係する。これについては，次節以降で，いくつかの事例を用いて説明する。

次に，(2)の文書 X と各カテゴリ中の文書の特徴の類似性を測る方法であるが，近年では，機械学習と呼ばれる手法が考案され，この技術を用いることが一般的になっている。本稿では，機械学習の中でも特に教師有り学習と呼ばれる手法を中心に紹介する。これは，人間であれば文書をどのように分類するのかという教師データと，個々の文書の特徴をコンピュータに与えると，その特徴を組み合わせて人間が分類するのと出来る限り同じように分類する，という枠組みである。機械学習には様々な方法が考案されており，文書分類でも SVM（Support Vector Machine），k 近傍法，ロジスティック回帰などがよく利用されている。次節以降でその一部を紹介する。

3.2. トピック分類

著者らは，学術論文を以下の 3 種類のカテゴリに分類する研究を行っている。

- 科研費カテゴリ：科学研究費助成事業の研究課題を分類するための体系。4 階層からなり，2015 年時点で，最下層で 319 カテゴリ存在している [1]。
- 国際特許分類：特許を分類するための体系で，5 階層からなり，2006 年時点で，最下層で 68,711 カテゴリ存在している [2]。
- Yahoo! 知恵袋カテゴリ：Yahoo! 知恵袋の質問–回答事例を分類するためのもの。3 階層からなり，2014 年時点で，最下層で 564 カテゴリ存在している [3]。

3 種類のカテゴリに分類する理由は，科研費カテゴリは研究者，国際特許分類は企業などの開発者，知恵袋カテゴリは非専門家と，それぞれ，異なる利用者を想定しているためである。このように，学術論文を様々なカテゴリに分類しておけば，論文を色んな側面から探しやすくなるだけでなく，例えば，同じカテゴリに分類された論文と特許を用いて，学術界と産業界の技術の関係性を分析するといったことも可能になる。

上記の 3 種類のカテゴリのうち，本節では，国際特許分類をとりあげ，説明する。学術論文を国際特許分類に分類するという課題は，国立情報学研究所が主催する評価ワークショップ NTCIR（NII Testbeds and Community for Information access Research）において，特許マイニングタスクとして著者らがオーガナイザとなって実施した [8),9)]。このタスクの最終目標は，論文と特許を対象とした技術動向分析を実現することで，サブタスクのひとつとして，論文の国際特許分類への自動分類を実施した。このタスクには，国内外の 12 の研究機関からの参加があった。ここで

は，多くの参加研究機関が採用した k 近傍法と呼ばれる手法について述べる。

まず，タスクの概要について述べる。このタスクでは，論文をひとつ以上の国際特許分類カテゴリに分類する。言い換えれば，論文が国際特許分類の各カテゴリに属するかどうか，カテゴリごとにひとつひとつ判断していくことになるため，この課題は，2.2 節で述べた文書分類の種類における 2 値分類に該当する。2 値分類は，様々な課題でその有効性が確認されている SVM という機械学習手法を使うのが一般的であるが，この場合はカテゴリの数が数万件もあるため，機械学習に膨大な時間がかかるという点が問題となり，上述のとおり多くの参加研究機関は k 近傍法を採用した。なお，ロジスティック回帰と呼ばれる別の機械学習手法を用いた参加研究機関もあった [10]。

次に，k 近傍法について述べる。k 近傍法とは，分類対象となる文書と内容が類似する文書を集め，収集された多くの文書が属するカテゴリを，分類対象文書のカテゴリと考える手法である。図 2 において，文書 X（学術論文）は分類対象となる文書である。文書 X 中に含まれる単語を特許検索システムの入力とし，文書 X と類似した特許を検索する。ここで，検索結果の各特許には，あらかじめ人手でカテゴリ（国際特許分類）が付与されているものとする。上位 4 件 [11]の結果を見ると，1 件目と 3 件目の特許がカテゴリ A に分類されているため，文書 X もカテゴリ A に分類される可能性が高そうであるが，それが実際にどの程度カテゴリ A に分類されそうか数値として表すために，文書 X と各検索結果との類似度を用いる。この類似度は，特許検索システムの出力結果として得られるもので，非常に大雑把に言えば，文書 X と各文書中の出現傾向が似ていれば値が大きくなる。図 2 の場合，カテゴリ A に分類されている 1 件目と 3 件目の類似度がそれぞれ 0.8 と 0.5 であるため，その和である 1.3 を，文書 X のカテゴリ A への分類されやすさと考える。検索結果上位 4 件には，この他にカテゴリ C に属する 2 件目の文書とカテゴリ B に属する 4 件目の文書が存在するため，カテゴリ B およびカテゴリ C への分類されやすさをそれぞれ 0.4 と 0.6 と考える。この分類されやすさのスコアの大きい順にカテゴリを並べ替え，分類結果として出力する。

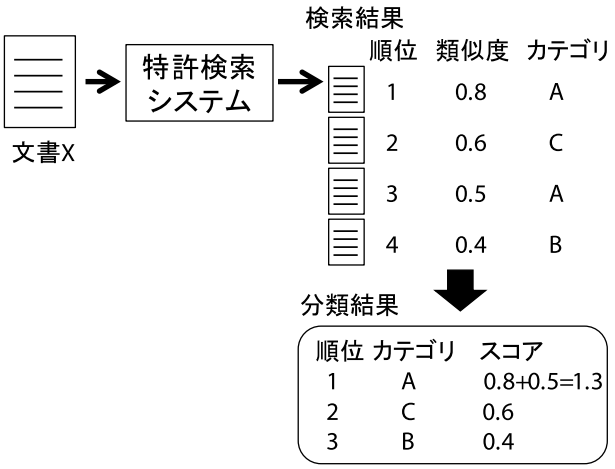

図 2　k 近傍法を用いた文書分類

図 2 の例では，検索結果の上位 4 件を用いて文書 X を分類したが，上位何件までを使うかは，文書集合の性質によって異なるため，調整が必要となる。なお，本節では，論文の国際特許分類への分類について説明したが，科研費カテゴリや Yahoo! カテゴリへの分類についても，全く同様に k 近傍法で実現できる。

3.3　フィルタリング

本節では，古典的ではあるが現在でも使われているナイーブベイズ分類器を，スパムフィルタリングを例に説明する。ある単語が含まれていれば必ずスパムメールであると判断できるのであれば，スパムフィルタリングは簡単に実現できる。しかし，多くの場合はある単語がスパムメールに出現しやすい，あるいはしにくいという傾向はあるものの，スパムメールとハムメール（スパムではないメール）両方に出現する。今，新しいメールを受け取ったとする。このメールの文面に，スパムメールでよく見かける単語がもしひとつしか見つからなければ，それだけで，そのメールをスパムであると断定はできないかもしれない。しかし，スパムメールでよく見かける単語がいくつも見つかると，そのメールはスパムメールである可能性が高い。ナイーブベイズ分類器とは，このような考え方に基づいた分類方法である。

図 3 を用いて，ナイーブベイズ分類器の仕組みを説明する。図 3 では，計 10 件のメール（①ー⑩）があり，このうち①ー④の 4 件がスパムメールである。また，10 件のメールには，A，B，C，D の 4 種類の単語が含まれているとする。

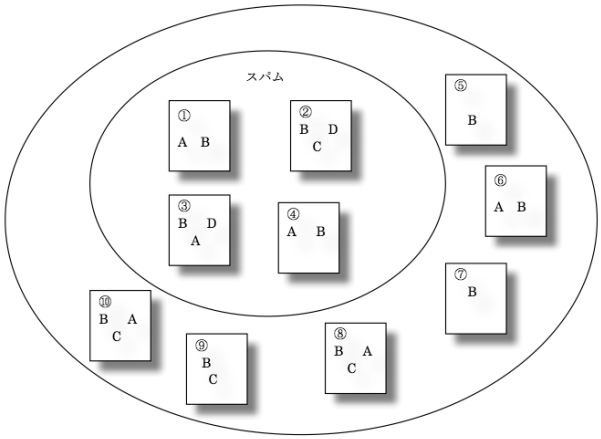

図 3　スパムメールとハムメール

図 3 から，A，B，C，D に，スパムメールかどうかを判定する能力がそれぞれどのくらいあるのか調べてみる。この作業では，各単語のスパムメール中の出現確率と，メール全体の中での出現確率を求め，それらの比を求める。

表 1 において，単語 A は 4 件のスパムメール中，3 件に出現しているので，確率 X が 3/4 となる。一方，メール全

体では 10 件中 6 件に出現しているので，確率 Y は 6/10 となる。X を Y で割った値が 1.25 となり，これが単語 A のスパムメール識別能力となる。スパムメールかどうかを識別する能力が低い単語とは，スパムメールにもメール全体にも同じ確率で出現する場合で，この時 X/Y＝1 となる。一方で，ある単語がスパムによく出現するのであれば，X/Y の値は大きくなり，スパムメールの識別能力も高くなる。逆に，スパムメールには滅多に出現しないがハムメールにはよく出現する単語もスパムメールの識別能力が高いと言える。この場合，X/Y の値が 1 よりも小さくなる。表 1 では，単語 C が該当する。

表 1　各単語のスパム識別能力

単語	スパムメール（X）	メール全体（Y）	X/Y
A	3/4	6/10	1.25
B	4/4	10/10	1
C	1/4	4/10	0.625
D	2/4	2/10	2.5

表 1 に示す各単語の識別能力を用いて，あるメールがスパムメールかどうかを判定する。今，図 4 のようなメール⑪が届いたとする。

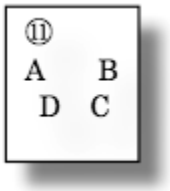

図 4　判定対象のメール

メール⑪には A, B, C, D すべての単語が含まれている。この時，メール⑪のスパムらしさは，A，B，C，D のスパム識別能力（X/Y）を掛けあわせたもの，つまり，1.25×1×0.625×2.5＝1.953125 となる。この値が大きければ大きいほどスパムらしいと判断される。

それでは，具体的にいくつ以上の値であればスパムと判断すべきであろうか？これは，すでにスパムかハムかが分かっている①～⑩について，それぞれスパムらしさを計算すれば，スパムとハムの境界値を決定できる。①～⑩のスパムらしさを計算した結果を表 2 に示す。

①～④がスパムメールであり，⑤～⑩と比べると，全体的に大きな値になっていることがわかる。この表を用いてスパムかハムかの境界値を決定する。スパムメールの中でスパムらしさの値が最も小さなものは①と④の 1.25 であるが，仮に「1.25 以上の値をとればスパムメールである」と決めると，メール⑥もスパムメールであると判定されてしまう。スパムメールと判定されたものは，通常はゴミ箱フォルダに自動分類されてしまい，ユーザの目にふれることはない。しかし，これでは重要なメールを見逃してしまう危険性があるため，一般には，スパムメールが多少混じっても，ハムメールを誤って破棄してしまわないような境界値を設定する。この場合は，1.25 より大きければスパムメールであると判定すれば，①と④はハムメールと判定されてしまうが，⑥は破棄されなくなる。この境界値を用いて，改めてメール⑪を見ると，スパムらしさは 1.953125 と，1.25 より大きいため，スパムメールと判定される。

表 2　メール①～⑩のスパムらしさ

① 1.25	⑥ 1.25
② 1.5625	⑦ 1
③ 3.125	⑧ 0.78125
④ 1.25	⑨ 0.625
⑤ 1	⑩ 0.78125

3.4　この他の文書分類の手法

3.2 節，3.3 節で紹介した手法は，いずれも文書中の単語をその文書の特徴と考えるものであったが，単語以外の特徴を使った文書分類もある。そのひとつが著者推定である。著者推定では，何について書いているかというよりも，どのような文体で書いているのかが，分類する上で重要となる。石田ら [4]は計量文体学の分野で使われてきた様々な特徴，例えば，文の長さ，単語の長さ，漢字や名詞や接続詞などの出現頻度，文中の読点の位置などをまとめ，5 種類の指標「量」,「構文」,「位置」,「表現」,「内容」に分類している。実は，文体を用いた文書分類は，著者推定だけでなく，前節で述べたスパムフィルタリングで利用されることもある。

小論文やエッセイの自動評価にもふれておく。石岡 [7]は，小論やエッセイを以下の 3 つの観点から自動評価する手法を提案している。

- 修辞：文章としてよく書けているか。
- 論理構成:アイディアが理路整然と表現されているか。議論が深められているか。
- 内容：出題文に適切に応えているか。

これらの評価を行う上で，上述の著者推定と類似した特徴が用いられている。例えば，文章の読みやすさに関する評価では，文の長さ，句の長さ，句の数，埋め込み文の存在などが使われている。なお，この成果は日本語小論文評価採点システム Jess[12]として Web 上で公開されている。

4. おわりに

本稿では，コンピュータを用いた文書分類について，いくつか事例をとりあげ，その仕組みを紹介した。コンピュータによる分類の詳しい理論が知りたい方は，書籍 [13]が参考になる。

コンピュータによる文書分類の現状と将来の課題について述べる。現在の自然言語処理では，文書中の単語について，単なる文字列ではなく，その背後にある意味を扱おうとする研究が増えつつある。例えば，「日光」という言葉に

は太陽光線という意味と，栃木県の地名という意味がある。両者を区別しなければ，文書を誤って分類してしまう可能性がある。この問題に対し，トピックモデルと呼ばれる統計モデルを用いれば，各文書に潜在的に含まれるトピックを検出することができ，例えば「日光　参拝　鬼怒川　旅　紅葉」と「日光　光合成　植物」のように，単語のまとまりとして意味を捉えることができるようになる。この統計モデルを用いた文書分類も研究レベルでは行われるようになってきており[14]，今後は実サービスシステムとしての運用への展開が望まれる。

この他にも，さらに難しい文書分類の問題が数多く残されている。例えば，ある学術分野の論文のリーディングリスト（読むべき論文のリスト）を作成するという課題がある。ある分野の論文を漏れなく集めることは，現在の文書分類の技術でもある程度は可能であるが，その中で読むべき論文を選定するのは非常に難しい。古典的な論文であれば，被引用数などの指標を利用することもできるが，発表されて間もない重要論文については別の指標に頼らざるをえないが，現時点では効果的な指標が見つかっていない。

最後に，今後は，文書分類という課題そのものを改めて見なおしてみる必要があるかもしれない。情報検索の分野では，従来，ユーザが入力したクエリに適合する文書をどれだけ過不足なく検索できるか，ということが中心的な研究課題であったが，近年では，ユーザが試行錯誤を繰り返して目的の情報を得る，という一連の流れの中において情報検索の有用性（usefulness）を意識する必要性が認識されつつある[15]。文書分類においても，文書分類という課題の中で閉じて考えるだけではなく，文書分類が情報を探すという行為の中でどのような位置づけにあり，その有効性とは何なのか，改めて考える時期にきているのかもしれない。

註・参考文献

1) Fukuda, S., Nanba, H., Takezawa, T., and Aizawa, A. Classification of Research Papers Focusing on Elemental Technologies and Their Effects. Proceedings of the 6th Language&Technology Conference (LTC'3), 2013.
2) 難波英嗣，竹澤寿幸．2 種類の翻訳システムを用いた学術論文の特許分類体系への自動分類，情報処理学会論文誌データベース．2009，vol.2，no.3，pp.76-86.
3) 重田識博，難波英嗣，竹澤寿幸．論文データの Yahoo! 知恵袋カテゴリへの自動分類，第 7 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2015）．2015.
4) 石田栄美，安形輝，野末道子．文体からみた学術的文献の特徴分析．三田図書館・情報学会研究大会論文集．2004，pp.33-36.
5) 大塚裕子，乾孝司，奥村学．意見分析エンジン―計算言語学と社会学の接点，コロナ社．2007.
6) http://blog.cybozu.io/entry/2158 [accessed 2016-03-21]
7) 石岡恒憲．小論文およびエッセイの自動評価採点における研究動向．人工知能学会誌．2008．vol.23, no.1, pp.17-24.
8) Nanba, H., Fujii, A., Iwayama, M., and Hashimoto, T. Overview of the Patent Mining Task at the NTCIR-8 Workshop. Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access. 2010.
9) Nanba, H., Fujii, A., Iwayama, M., and Hashimoto, T. Overview of the Patent Mining Task at the NTCIR-7 Workshop. Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access. 2008.
10) Fujino, A. and Isozaki, H. Multi-label Classification using Logistic Regression Models for NTCIR-7 Patent Mining Task. Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access. 2008.
11) k 近傍法の k とは，検索結果の上位 k 件までを用いるということを意味している。この場合は k＝4 となる。
12) http://tk2-203-11024.vs.sakura.ne.jp/jess/ [accessed 2016-03-21]
13) 高村大也．言語処理のための機械学習入門，コロナ社．2010.
14) Ramage, D., Hall, D., Nallapati, R., and Manningj, C.D. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 2009, pp.248-256.
15) Belkin, N.J. Salton Award Lecture: People, Interacting with Information. Proceedings of the 38th Annual SIGIR Conference. 2015.

Special feature: □□□□主査より後で入れます□□□□□□□□□. Document Classification by Artificial Intelligence.

Hidetsugu Nanba (Graduate School of Information Sciences, Hiroshima City University, 3-4-1 Ozukahigashi, Asaminamiku, Hiroshima 731-3194 JAPAN)

Abstract: Recently, artificial intelligence (AI) has made remarkable progress in various fields including research, such as computer Go, computer Shogi, and autonomous car. We can often find the news about these on the internet, newspapers, and TV. Natural language processing (NLP) is a field of AI, where in we make computer programs to process human language (natural language). Classifying documents through the use of computers on behalf of humans is a typical research field of NLP. In this paper, we focus on automatic document classification, and introduce some researches with their mechanisms.

Keywords: natural language processing / document classification / machine learning / filtering / k-Nearest Neighbor method / naïve Bayes classifier