様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　６月１４日現在

機関番号：３２７０６
研究種目：基盤研究（C）
研究期間：2010～2012
課題番号：２２５００１４０
研究課題名（和文）　古今東西の全言語を対象にしたテキストマイニングに関する研究

研究課題名（英文）　Text Mining for Languages of All Ages and Countries

研究代表者
鈴木　誠　（SUZUKI MAKOTO）
湘南工科大学・工学部・准教授
研究者番号：80339796

研究成果の概要（和文）：文字 N-グラムに基づく言語独立なテキスト分類手法である蓄積手法を提案した。蓄積手法は、索引語を形成する際に文字 N-グラムを使用するので、言語固有の文法構造に依存しない。テキスト文書が Unicode で表現されてさえいれば、蓄積手法は異なる言語に対しても同一のプログラムを用いて文書を分類することができる。そこで、この蓄積手法を用いて英語と日本語と韓国語と中国語のテキスト文書の分類実験をした。その結果、英語の Reuters-21578 は 94.5%、日本語の毎日新聞の実験データは 88.5%、韓国語のハンギョレー新聞の実験データは 90.2%、中国語の人民日報の実験データに対しても 92.6%の精度で分類することができた。このように、蓄積手法が様々な言語で比較的高い精度で分類できることを確認した。さらに、蓄積手法の数理モデルを構築し、その数理的な意味を解明することができた。

研究成果の概要（英文）：We proposed the accumulation method, which is a language-independent text classification method that is based on the character N-gram. The accumulation method does not depend on the language structure, because this method uses the character N-gram to form index terms. If text documents are expressed in Unicode, then the accumulation method can classify documents using the same algorithm. Therefore, we classified English, Japanese, Korean, and Chinese text documents. As a result, the highest macro-averaged F-measures of the proposed method were 94.5% for the English Reuters-21578, 88.5% for the Japanese CD-Mainichi 2002 data set, 90.2% for the Korean Hankyoreh 2008 data set, and 92.6% for the People's Daily 2009-2010 data set. Thus, we obtained good results for these languages. Moreover, we were able to construct a mathematical model of the accumulation method and were able to clarify the mathematical meaning.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,100,000	330,000	1,430,000
2011 年度	1,000,000	300,000	1,300,000
2012 年度	1,100,000	330,000	1,430,000
総　計	3,200,000	960,000	4,160,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学・知識発見とデータマイニング
キーワード：多言語処理・機械学習・モデル化・文書自動分類・*N*-gram

１．研究開始当初の背景

近年、IT 技術の進歩により計算機の性能が向上し、あらゆる文書が電子化され、計算機上で利用可能なテキストデータの量が年々増加している。また、多言語の文字を単一の文字コードで取り扱うため Unicode が 1980 年代に提唱され、1993 年に国際標準化機構で ISO/IEC 10646 の一部として標準化された。その後、中国語・日本語・朝鮮語で使用される漢字が統合された領域、いわゆる CJK 統合漢字などの様々な問題が生じたが、バージョンアップを重ねて、一般的に普及してきている。このような状況の中で、大量に存在するテキストデータを自動的に処理するテキストマイニング技術に対する需要も高まってきている。このテキストマイニングの中にも、文書検索・文書要約・文書分類・文書クラスタリングなどの様々な課題が存在する。この中で、本研究は文書分類の問題に焦点をあてる。コンピュータによる自動分類は古典的な機械学習の問題でもあり、今までは、決定木・ニューラルネットワーク・Naive Bayes・k 近傍法・ブースティング・サポートベクタマシンなどの数多くの機械学習アルゴリズムが利用されてきた。

２．研究の目的

本研究の目的は、主に、(1)処理対象となる言語の拡張、(2)基礎理論の構築、の二点である。

具体的には：

(1) 言語固有の文法知識を一切使わずに、古今東西の Unicode で表現された全言語を分類・クラスタリングできるプログラムを開発する。

(2) 本研究の方式を確固たる数理モデルに基づいて考察し、基礎理論面でも研究を発展させる。

について検討した。

３．研究の方法

「２．研究の目的」に沿って説明する：

(1) 処理対象となる言語の拡張

現在提案しているテキスト自動分類手法は、特徴語に N-gram を利用しているので、言語の意味によらず、言語独立であるという特徴をもっている。一般的に、N-gram は文字 N-gram と単語 N-gram に大別される。例えば「私は鈴木誠です」という文章を文字 N-gram を使って区切ると、N=2 の場合は「私は/は鈴/鈴木/木誠/誠で/です」となる。N=3 の場合は「私は鈴/は鈴木/鈴木誠/木誠で/誠です」となる。一方、「My name is Makoto Suzuki」という文章を単語 N-gram で区切ると、N=2 の場合は「My name/name is/is Makoto/Makoto Suzuki」となり、N=3 の場合は「My name is/name is Makoto/is Makoto Suzuki」となる。このように、1 文字もしくは 1 単語ずつスライドさせながら N 文字もしくは N 個の単語の塊を作り、字面だけで処理を進めることができるため、言語固有の単語知識や文法知識を一切必要としない。すなわち、抽出された特徴語自体には基本的に意味はなく、言語が異なっていても常に同一の手法で処理ができる。そこで、本手法を用いることにより、英語と日本語の他にも韓国語や中国語でも良い分類結果が得られることを示した。

(2) 基礎理論の構築

近年の本研究の成果により、文書分類の数理モデルは従来の文書検索のモデルとは異なった特殊性を持っていることが分かってきた。例えば、文書検索においては「どの文書にどのキーワードが出現したか」は重要な情報であるが、文書分類の学習段階においては上記の情報よりも「どのカテゴリにどのキーワードが出現したか」の方が重要な情報となりうる。この状況を踏まえ、本方式の数理モデルを考察し、文書分類の基礎理論を構築した。

４．研究成果

「２．研究の目的」に記載した「(1)処理対象となる言語の拡張」と「(2)基礎理論の構築」の二点について主に研究を進めた。

はじめに(1)について説明する。実験に用いたデータは、英語が Reuters-21578 である。これはベンチマークデータであり、インターネットからダウンロード可能である。日本語・韓国語・中国語に関しては、毎日新聞・ハンギョレ一新聞・人民日報から新聞記事を購入し、独自に実験データセットを作成した。これらの実験データを用いて分類実験を行った結果、英語は 94.5%、日本語は 88.5%、韓国語は 90.2%、中国語は 92.6%の精度で分類することができ、様々な言語で比較的高い精度で分類できることが確認できた（この数値は、すべてマクロ F 値である)。これらの結果を図１から図５にまとめた。

図２から図５が、英語・日本語・韓国語・中国語の分類結果である。各々の図の横軸は文字 N-gram の N であり、左側の縦軸が分類精度（%）を、右側の縦軸が記憶した文字 N-gram の数（千個）を示している。各文字 N-gram における棒グラフは、図１に示す通り左から、適合率のマイクロ平均（miP)、再現率のマイクロ平均（miR)、適合率のマクロ平

均（maP）、再現率のマクロ平均（maR）を表しており、棒グラフの上の太字の数値は左からマイクロ F 値、マクロ F 値である。また、折れ線グラフ（赤線）は、各文字 N-gram において記憶しておく必要がある文字 N-gram の数を示している。例えば、図２の英語の分類結果では、N が 6 の場合（すなわち文字 6-gram の場合）が最も分類精度が高く、マクロ F 値は 94.5%であり、その際に記憶しておかなければならない文字 6-gram の数は 691,500 個である。図２から図５に示す通り、言語により分類精度が最高となる文字 N-gram の長さ（N の値）に差があることが分かる。分類精度が最高となるのは、中国語の場合は文字 3-gram であり、韓国語の場合は文字 4-gram、日本語の場合は文字 6-gram であった。もちろん、これは実験データによっても多少の差があり、分類精度も実験データセットによって多少のバラツキがある。

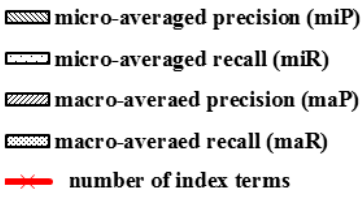

図１：各分類結果の凡例

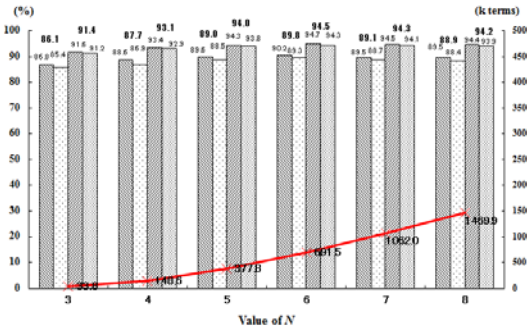

図２：英語の分類結果

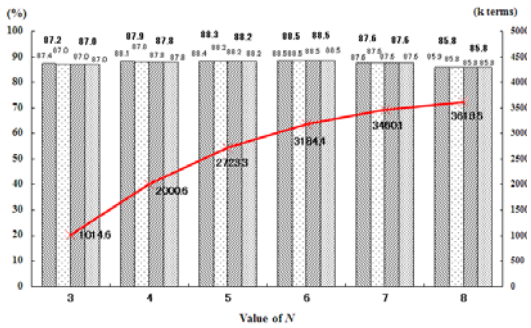

図３：日本語の分類結果

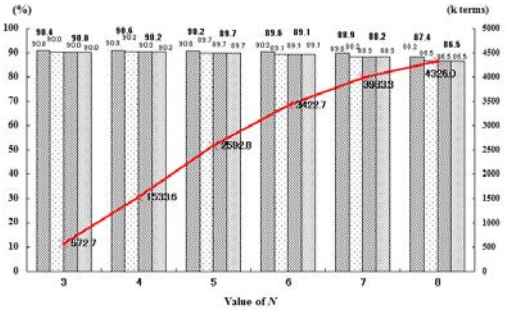

図４：韓国語の分類結果

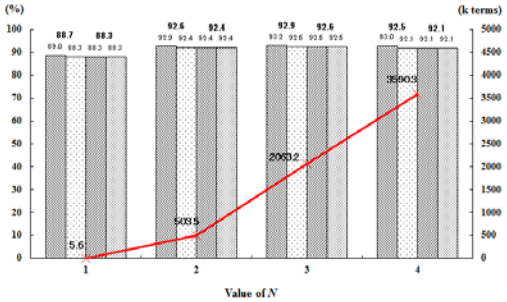

図５：中国語の分類結果

さらに、蓄積手法を用いてコンピュータが各文書の分類の難しさを自動的に識別することができる枠組みを構築した。例えば、図５の中国語の分類を例にとり説明する。図５の中国語の分類においては、文字 N-gram の N を１から４で動かしているが、1・2・3・4 と動かした場合に各記事の分類先を記憶して整理した結果が表１である。

表１：可変 N-gram の結果（中国語の場合）

Support Number	Number of Correct Documents	Number of Target Documents	Ratio of Correct Answers (%)
1	0	0	-
2	39	81	48.15
3	217	285	76.14
4	2531	2634	96.09

すなわち、表１においては、N を 1・2・3・4 と動かしても分類先のカテゴリが変化しない記事は支持数（Support Number）が 4 となり、96.09%というかなり高い確率で正解となる。これに該当する記事は 2634 件あり、この分類実験におけるテストデータ全体の記事数が 3000 件なので、全体の 87.8%を占めている（2634/3000=0.878 である）。一方で N が 1・2・3・4 と変化する間に一つでも異なるカテゴリに分類されてしまった場合（支持数は 3）、その分類結果は 76.14%となり 24.86%が不正解となってしまう。さらに、支持数が 2 となった場合の正解率は 48.15%であり、約半数は誤分類となってしまい、その記事は分類が非常に難しいという判断をコンピュータ

は自動的に下すことができる（この詳細は、以下の[学会発表]の文献③を参照されたい)。
次に(2)については、蓄積手法の数理モデルを構築し、その数理的な意味を解明することができた。ここでは、蓄積手法による分類手順について簡単に記す（この詳細は、以下の[学会発表]の文献①を参照されたい)。
[Step1] 学習用文書をタームに切り分ける。
[Step2] 各学習用文書における各タームの存在を確認する。このとき，必要に応じて分類に使用するタームを選択する。
[Step3] 各学習用文書における各タームの出現情報に基づき、タームカテゴリ行列を生成する。
[Step4] 各カテゴリのカテゴリ代表ベクトルをカテゴリごとに計算する。
[Step5] 分類すべき新規文書を新規文書ベクトルで表現し、新規文書ベクトルと各カテゴリ代表ベクトルとの類似度をカテゴリごとに計算し、類似度の高いカテゴリへ分類する。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）
〔雑誌論文〕（計２件）

① 山岸直秀，鈴木誠，渡辺重佳 ： 任意の外部記憶容量で動作するマージソート，電子情報通信学会論文誌，Vol.J96-D,No.3，pp.441-451，2013
② 後藤正幸，石田崇，鈴木誠，平澤茂一 ： 高次元ベクトル空間モデルによるテキスト分類問題について - 分類性能と距離構造の漸近解析 -，日本経営工学会論文誌，Vol.61，No.3，pp.97-106，2010

〔学会発表〕（計７件）

① 鈴木誠，山岸直秀 ： 単語N-gramと文字N-gram を用いた日本語の文書分類に関する一考察，第 35 回情報理論とその応用シンポジウム予稿集，pp.660-665，2012，大分
② M.Suzuki，N.Yamagishi，Y.C.Tsai and M.Goto ： English and Japanese Text Categorization Using Word and Character N-grams，Proc. of Asia Pacific Industrial Engineering and Management Systems Conference (APIEMS 2012)，pp.715-722，2012，タイ
③ M.Suzuki，N.Yamagishi and Y.C.Tsai ： Chinese Text Categorization Using the Character N-gram，Proc. of International Symposium on Information Theory and its Applications (ISITA 2012)，pp.722-726，2012，アメリカ合衆国
④ 山岸直秀，鈴木誠，渡辺重佳 ： 状態遷移モデルへの適応による教師なし単語分割手法の提案，日本経営工学会 西関東支部 第 12 回学生論文発表会予稿集，pp.57-58，2012，早稲田大学
⑤ M.Suzuki，N.Yamagishi and M.Goto ： Korean Text Categorization Using the Character N-gram，Proc. of International Conference on Information Technology and Applications (ICITA 2011)，pp.197-202，2011，オーストラリア
⑥ M.Suzuki，N.Yamagishi，Y.C.Tsai，T.Ishida and M.Goto ： English And Taiwanese Text Categorization Using N-gram Based on Vector Space Model，Proc. of International Symposium on Information Theory and its Applications (ISITA 2010)，pp.106-111，2010，台湾
⑦ M.Suzuki，N.Yamagishi，T.Ishida，M.Goto and S.Hirasawa：On a New Model for Automatic Text Categorization Based on Vector Space Model，Proc. of IEEE International Conference on Systems，Man，and Cybernetics 2010 (SMC 2010)，pp.3152-3159，2010，トルコ

〔図書〕（計１件）

① 須子統太，鈴木誠，浮田善文，小林学，後藤正幸 ： 確率統計学，オーム社，2010

〔その他〕
ホームページ等
http://www.info.shonan-it.ac.jp/suzuki-lab/profile.html

６．研究組織
(1)研究代表者

鈴木　誠　（SUZUKI MAKOTO）
湘南工科大学・工学部・准教授
研究者番号：80339796

(2)連携研究者

大須賀　昭彦　（OHSUGA AKIHIKO）
電気通信大学・大学院情報システム学研究科・教授
研究者番号：90393842
後藤 正幸　（GOTO MASAYUKI）
早稲田大学・創造理工学部・経営システム工学科・教授
研究者番号：40287967
須子　統太　（SUKO TOTA）
早稲田大学・メディアネットワークセンター・助教
研究者番号：40409660