様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　６　月　　５　日現在

機関番号：　１７１０２
研究種目：　基盤研究(C)
研究期間：　2012～2014
課題番号：　２４５００１７６

研究課題名（和文）手掛語と内容語の双対ブートストラップ・マイニング

研究課題名（英文）Dual Bootstrap Mining with Feature Words and Contents Words

研究代表者
廣川　佐千男（Hirokawa, Sachio）

九州大学・学内共同利用施設等・教授

研究者番号：４０１２６７８５

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：検索条件で限定される分析対象文書集合について、SVM(support vector machine)を適用して特徴語を抽出するテキストマイニングの手法の開発と研究を行った。少数の手掛語で検索結果の文書群を特徴付けることができた。特徴語の一般性と特殊性を定量化としてブートストラップ法を開発した。学術論文概要、授業への学生のコメント、英語学習作文例、有価証券報告書、医療情報文書、Web文書を対象とする実験により有用性が確認できた。

研究成果の概要（英文）：We developed a text mining method to extract feature words of search result using SVM (support vector machine). We succeeded to find small number of feature words that characterize the documents. We extended the bootstrap method for a measurement of generality and specificity of feature words.
We confirmed the effectiveness of the methods by applying them to analyze the real data of scientific articles, students' free comments, English writing errors, security reports, medical records and Web documents.

研究分野：　テキストマイニング

キーワード：　SVM　属性選択　ブートストラップ　可視化　特徴語　機械学習

１．研究開始当初の背景

検索エンジンは、条件に合致する文書を求めるだけなく、検索結果全体の分かりやすい解釈という内容まで求められるようになっている。例えば、ブログ記事中の商品評判情報、新聞記事の中の因果関係、特許情報から企業の研究動向の抽出を試みる研究などがある。これらの検索では、分析目的に応じて、どのような単語に着目すべきか、文書を評価するための単語の出現文脈をどう評価するかが重要な課題となっている。

本研究では、分析目的を手掛語集合として定式化し、内容語と手掛語の共起情報を使って、両方を双対的に抽出する手法を研究することとした。

２．研究の目的

検索結果の概要を理解するためには、特徴語抽出は必須である。しかし、同じ文書群を対象としても、分析目的によって抽出すべき特徴語は異なる。車の検索でも、知りたいのが機能か人気かで、また、特許の検索でも、知りたいのがその企業の研究目的か技術かで、特徴語は異なる。

本研究では、文書群に現れる単語を、分析目的に合致する文を抽出するための手掛語と、分析内容そのものを解釈するための内容語という二つの観点で捉え、手掛語と内容語の共起情報を利用することで、少数の事例から、内容語と手掛語を双対的に抽出する手法を研究する。また、内容語については、その単語が一般的であるか、専門的あるいは限定的であるか表す尺度を求める。

３．研究の方法

分析対象の文書群、分析目的、分析内容を表す典型的な内容語が与えられたとき、分析の要点となる重要文を抽出するための手掛語と、より多くの内容語とを双対的に求める手法を開発し、その抽出性能を評価した。具体的な分析対象として、企業情報、学術論文、ブログ記事、医療情報文書、自由記述アンケートなどの文書データについて、企業の経営状況、学術論文が対象とする問題、ブログ記事の主観客観性、長期入院となった理由、成績がよくない理由のなどの内容語を抽出するための手掛語を求める。

企業情報については、まず、これまでの研究成果である倒産理由を表す内容語と手掛語を用いて健全企業と倒産企業の識別性能を評価した。また、業績データに基づき、健全企業と優良企業を識別するための内容語と手掛語を求めた。抽出手法としては、SVM と属性撰択を用いて、判別が最適となる特徴語集合を求めた。

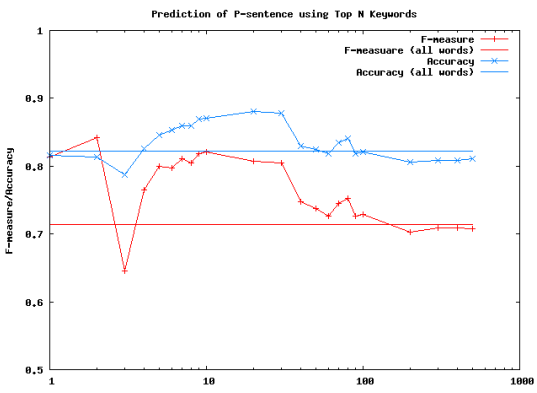
４．研究成果

検索条件で限定される分析対象文書集合について、SVM(support vectormachine) を適用して特徴語を抽出するテキストマイニングの手法の開発と研究を行った。特に、少数の手掛り語で検索結果の文書群を特徴付けるため、新な属性撰択の方法を検討し、性能評価を行った。

評価実験の文書群として、学術論文概要、授業への学生のコメント、英語学習作文例、有価証券報告書、医療情報文書、および Web 文書を対象として、提案手法を適用し、属性撰択で得られた特徴語集合の妥当性と識別性能の評価を行った。学術論文概要については、問題点、手法、結果などの観点を表す文の手掛り語を求めた。全ての単語を使う方法より判定性能が向上することを示した。

学生の自由記述アンケートから成績を推定するため、授業前(Pre)、授業中(Current)、次回への準備(Next)という三つの観点に着目し、それらの観点を表す手掛り語を SVM と属性撰択で抽出し、各文が三つの観点を表す度合(PCN スコア)を機械的に推定する手法を構築した。

例えば下の図で赤い線は、学生のコメントが授業前の文かどうかを機械的に判定したときの判定性能（適合度）を表す。

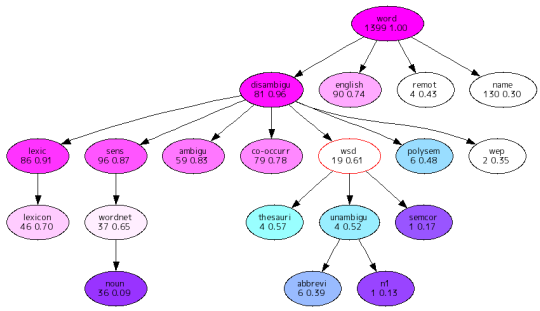


全ての単語を使った場合（水平線）７割なのが、属性選択の結果１０個の単語で８割に向上している。推定性能はよくないが、PCN スコアと学生の成績推定性能に相関があることができた。すなわち、PCN の観点についてきちんとコメントを書いている学生については、その学生の成績を推定でき、学生の指導への活用という画期的な成果が得られた。

英文の典型的な誤りパターンに対する手掛り語を求め、機械的に推定した間違いパターンを使って、母国語を推定することができることを示した。
医療情報への応用として、股関節手術の手術

記録を対象に、術後入院日数の長い患者を特定するための手掛り語を求め、全ての単語受かって SVM を適用した場合より、識別性能が向上することが確認できた。この結果に基づき、他の様々な医療情報への応用が見えてきた。

また、特徴語の一般性と特殊性を定量化するブートストラップ法を検討した。情報関連文献の論文概要を対象としてプロトタイプの分析システムを構築した。具体的には検索やマイニングに関する 16 個の国際会議予稿集と 5 つのジャーナルに掲載された 45719 件の論文概要を対象とする検索システムを構築した。
次の図は、wsd を検索語としたときの、関連語をマインドマップとして表示したものである。



単語の下に表示している数値は、その単語を含む文書数ならびに、関連語リストにおける単語の一般度を表す。
図において、wsd の下に表示される部分が、wsd による検索結果の特徴語の関連を表している。いずれも、wsd よりも下位の単語となっているので、青で表示されている。wsd を上に辿ると disambigu と word が並んでいる。この二つは wsd の意味を表す上位語といえる。赤で色がついている単語が他に 7 個あるが、それらは wsd を根とする部分木には含まれていない。wsd から始めて、関連語を求めるブートストラッピングの過程で、より一般的な単語である disambigu や word が得られるが、それらの関連語は必ずしも最初に与えた単語 wsd とは関連していないことに対応している。つまり、ブートストラップにおいてトピックドリフトが起っていることを表す。

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３件）

① 合田和正、峯恒憲、廣川佐千男,
学習態度に関する自己評価記述の正確さと成績推定性能の相関,電子情報通信学会情報・システムソサイエティ論文誌,採択決定(査読有)
② Shaymaa E. Sorour, Teunenori Mine, Kazumasa Goda, Sachio Hirokawa, A Predictive Model to Evaluate Student Performance, Journal of Information Processing, Vol.23, No.2, pp.192-201, 2015（査読有）
③Takanori Yamashita, Yoshifumi Wakata, Satoshi Hamai, Yasuharu Nakashima, Yukihide Iwamoto, Brendan Flanagan, Naoki Nakashima, Sachio Hirokawa, Extraction of Key Factors from Operation Records by Support Vector Machine and Feature Selection, Indian Journal of Medical Informatics, Vol.8, pp.70-71, 2014（査読有）

〔学会発表〕（計２６件）

① Shaymaa Sorour, Tsunenori Mine, Kazumasa Goda, Sachio Hirokawa, Predicting students' grade by measuring semantic similarity between free style comments data, Proc. ICWL2014, pp.142-151, 2014(査読有)
② Sachio Hirokawa, Emi Ishita, Non-Topical Classification of Healthcare Information on the Web, Frontiers in Artificial Intelligence and Applications, Volume 262: Smart Digital Futures 2014, pp. 237-247, 2014(査読有)
③Yamashita, T., Wakata, Y., Nakashima, N., Hirokawa, S., Hamai, S., Nakashima, Y., Iwamoto, Y., Extraction of Determinants of Postoperative Length of Stay from Operation Records, Proceedings of 2014 IEEE Workshop on Electronics, Computer and Applications, pp.822-827, 2014(査読有)
④ Sachio Hirokawa, Brendan Flanagan, Chengjiu Yin, Hiroto Nakae, Vizualization of Relation and Generality of Words in Search Result, Proceedings of the Third Asian Conference on Information Systems,pp.90-95, 2014(査読有)
⑤ Bren Flanagan, Chengjiu Yin, TakahikoSuzuki, Sachio Hirokawa Intellegent Computer Classification of English Writing Errors, Proc. KES-IIMS2013, pp.174-183, 2013(査読有)
⑥ Toshihiko Sakai, Sachio Hirokawa, Feature Words that Classify Problem Sentence in Scientific Article, Proc. iiWAS2012, pp.360-367, 2012(査読有)

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）
○取得状況（計　2　件）
名称：情報処理装置、情報処理方法及びプログラム
発明者：廣川佐千男、御手洗秀一
権利者：廣川佐千男
種類：特許
番号：2011-104418

出願年月日：平成２３年５月９日
取得年月日：平成２７年１月２７日
国内外の別：国内

名称：検索方法、検索装置及びプログラム
発明者：廣川佐千男
権利者：廣川佐千男
種類：特許
番号：特許 2011-104414
出願年月日：平成２３年５月９日
取得年月日：平成２７年４月８日
国内外の別：国内

〔その他〕
ホームページ等
http://tml.cc.kyushu-u.ac.jp/
http://catalog.lib.kyushu-u.ac.jp/ja/xc/search/sachio%20hirokawa

６．研究組織
(1)研究代表者
廣川 佐千男（HIROKAWA, Sachio）九州大学情報基盤研究開発センター・教授
研究者番号：40126785

(2)研究分担者
中藤 哲也（NAKATOH, Tetsuya）九州大学情報基盤研究開発センター・助教・
研究者番号： 20253502

(3)連携研究者
（ ）

研究者番号：