様式Ｃ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成 24 年 6 月 28 日現在

機関番号：32602
研究種目：若手研究(B)
研究期間：2008 年度～2011 年度
課題番号：20700224
研究課題名（和文）　様々な種類の文書に対応した汎用性の高い著者推定手法

研究課題名（英文）　Robust method for authorship attribution applicable to various document types

研究代表者
安形　輝　（AGATA TERU）
亜細亜大学・国際関係学部・准教授
研究者番号：80306505

研究成果の概要（和文）：
様々な種類の文書に対応した汎用性の高い著者推定手法の実現に向けて、圧縮プログラムを応用した著者推定手法を提案した。様々な圧縮プログラムを適用した著者推定実験を通じて、その有効性を検証した。また、圧縮率と著者推定精度の関係を分析した。また、著者推定実験を通じて古い文書や未解読文書など様々なタイプの文書に対してもこの手法が適用可能かを試みた。

研究成果の概要（英文）：
The robust method for authorship attribution applicable to various document types was proposed. Authorship attribution with combination of this method and various compression programs proved the effectiveness of the method. The experiment showed high performance compression programs gave high precision ratio.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	700,000	210,000	910,000
2009 年度	500,000	150,000	650,000
2010 年度	500,000	150,000	650,000
2011 年度	600,000	180,000	780,000
年度			
総　計	2,300,000	690,000	2,990,000

研究分野：　図書館情報学
科研費の分科・細目：情報学　・　図書館情報学・人文社会情報学
キーワード：　圧縮プログラム、著者推定、圧縮アルゴリズム

１．研究開始当初の背景

真贋や著者の真正性が疑われる歴史的文書や著名な文学作品に対しては、コンピュータの登場以前から筆跡鑑定を含め様々な著者推定研究が行われてきた。人手で著者推定を行う場合には、信頼性・客観性の問題、大規模文書データへの適用が難しいという規模の問題がある。そのため、近年はコンピュータを用いた計量書誌学、計量文献学の手法が用いられる。

この領域での研究の多くは、単一の文書タイプを対象とした著者推定実験という形で行われ、文学作品と学術論文といった複数の文書タイプの集合を用いた実験は少ない。さらに、大半は文学作品を実験対象としている。文学作品は、一定以上の長さを持つものが多く、文体や語彙に著者の特徴が多く現れるため、著者推定対象としては、比較的容易なタ

イプの文書と言える。しかし、真贋が問題となるのは文学作品ばかりではない。新聞記事、学術論文、電子メールといった、より形式が統制されている、あるいは、長さが非常に短いタイプの文書についても、電子データが急増し盗用が容易になったことで、著者推定に対する需要は高まっている。特に、論文盗作事件が近年、国内外で注目を集めている 1) にも関わらず、学術論文に関する著者推定研究はほとんど行われてこなかった。また、既存の著者推定手法は品詞情報や単語長、文の長さなどの言語的な特徴に大きく依存するため、ある言語で開発された手法は他の言語の文書に原則的には適用できない。

２．研究の目的

様々な文書タイプ、文書サイズに対応し、言語に依存しない著者推定手法は汎用性の高い、非常に有用なものだと考えられる。本研究で提案した圧縮プログラムを応用した著者推定手法はいずれの条件も満たす汎用性の高い手法といえる。

そこで本手法を用いた著者推定実験を行うことで手法の有効性を検証する。また、組み合わせる圧縮プログラムを変更した場合に、圧縮性能と著者推定精度がどのような関係にあるのかも分析する。さらに、様々なタイプの文書に適用した実験を通じて、その汎用性についても検証を行う。

３．研究の方法

基本的に著者推定実験は、ある一文書に対する試行を以下の手順で行う
①テスト集合からある文書を除き、対象文書とする
②著者推定手法を用い、対象文書とテスト集合の他の文書間の類似度を算出する
③対象文書を除く全文書に対する類似度に基づき類似度順出力を行う
④順位 1 位に出力された文書の著者が対象文書と同一ならば成功、異なれば失敗とする
この手順をテスト集合中の全文書に対して試行し、著者推定精度を算出する。失敗した試行については類似度順出力上位に出力された文書の出現語彙や品詞情報を用いてなぜ失敗したかについて詳細な分析を行う。

日本語近代文学テキストを始め、様々なタイプの文書を対象として、このような著者推定実験を行なった。また、組み合わせる圧縮プログラムを変更した時にどのように著者推定精度が変わるかを実験から分析した。

４．研究成果

本研究で提案した著者推定を用いた実験からは、様々な圧縮プログラムと組み合わせた場合に、実験結果からは以下の 3 点が明らかとなった。
①圧縮性能が高い圧縮プログラムはどのようなアルゴリズムであってもほぼ 100%に近い高い著者推定の平均成功率を示す
②データ長を 20,000 バイトまで短くした場合にも圧縮性能の高いプログラムでは 9 割以上の成功率であった。
③平均圧縮率と著者推定の平均成功率には高い相関がみられた。

また、未解読文書や他のタイプの文書に対する実験からはその有効性が明らかとなった。

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安形輝；安形麻理．文書クラスタリングによる未解読文書の解読可能性の判定：ヴォイニッチ写本の事例．Library and Information Science. no.61, 2009/6, p.1-23.
http://lis.mslis.jp/pdf/LIS061001.pdf

②安形輝．複数の圧縮プログラムを用いた近代日本文学の著者推定．亜細亜大学学術文化紀要．no.20, 2012/1, p.15-33.

〔学会発表〕（計 6 件）
①安形輝；安形麻理．部分文書出現位置からの未解読文書の真正性の判定．2009 年日本図書館情報学会春季研究集会．2009 年 5 月 23 日．駿河台大学（埼玉県）

②宮田洋輔；安形輝ほか 3 名．学術論文 PDF の自動判定：学習用集合が判定性能に与える影響．2010 年日本図書館情報学会春季研究集会．2010 年 5 月 29 日．同志社大学(京都府)

③Emi Ishita, Teru Agata, Atsushi Ikeuchi, Nozue Michiko, Miyata Yosuke, and Shuichi Ueda. 2010. A search engine for Japanese academic papers. In Proceedings of the 10th annual joint conference on Digital libraries (JCDL '10). ACM, 2010/06/22. New York, NY, USA, 379-380. DOI=10.1145/1816123.1816189

④ Teru Agata, Yosuke Miyata, Atsushi Ikeuchi, and Shuichi Ueda. 2010. The deep web in institutional repositories in Japan. In Proceedings of the 73rd ASIS¥&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47 (ASIS&T '10), Vol. 47. American Society for Information Science, Silver Springs, MD, USA, 2010/10/25, Article 138 , 2 pages.

⑤安形輝ほか４名．学術情報に特化した検索エンジンの開発：機械学習による英語論文の自動判定．2010年度日本図書館情報学会研究大会．2010年10月9日．藤大学（北海道）

⑥Emi Ishita, Teru Agata, Atsushi Ikeuchi, Miyata Yosuke, and Shuichi Ueda. 2011. Detecting academic papers on the web. In Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries (JCDL '11). 2011/07/15
DOI=10.1145/1998076.1998161

〔その他〕
ホームページ等
http://gyoseki.asia-u.ac.jp/aauhp/KgApp?kyoinId=ymdygyysggy
http://itasan.mydns.jp/

６．研究組織
(1)研究代表者
安形　輝　（AGATA TERU）
亜細亜大学・国際関係学部・准教授
研究者番号：80306505

(2)研究分担者
なし

(3)連携研究者
なし