様式 C-19

科学研究費補助金研究成果報告書

平成　23 年　　5 月　　18 日現在

機関番号：32663
研究種目：研究活動スタート支援
研究期間：2009～2011
課題番号：21800087
研究課題名（和文）　オンライン学習コミュニティログからのソーシャル・アティテュードの抽出と分析
研究課題名（英文）　Extracting and analyzing social attitudes from the logs in online learning communities
研究代表者
鈴木　崇史（SUZUKI TAKAFUMI）
東洋大学・社会学部・講師
研究者番号：70507037

研究成果の概要（和文）：

現在、Web の発達により、大量かつ多種の情報資源が蓄積されている。テキストデータは、そのような情報資源の中でも、重要な位置を占め、これを新たな応用へと役立てることが、学術的、社会的に急務の課題となっている。本研究では、とりわけ、学習、知識交換コミュニティに焦点をあて、テキストから人々のコミュニケーションの特徴を明らかにするとともに、有効な分析手法の検討を行った。

研究成果の概要（英文）：

Along with the development of the Web, a large amount of, and various types of information resources are available now. Among them, textual data is so important that we are urged to make better use of it. Against this background, this study, focusing on the texts in learning and knowledge sharing communities, analyzed the character of ones' communication styles from the texts, as well as developing the methods for the analysis.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	560,000	168,000	728,000
2010 年度	430,000	129,000	559,000
年度			
年度			
年度			
総　計	990,000	297,000	1287,000

研究分野：総合領域
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：計算文体論、計量情報学、情報図書館学、人文情報学、テキスト分析

１．研究開始当初の背景

現在、データベースの蓄積と Web の発達により、大量かつ多種の情報資源が蓄積されている。テキストデータは、そのような情報資源の中で重要な位置を占め、これを新たな応用へと役立てることが、 学術的、社会的に急務の課題となっている。とりわけ、ブログや掲示板をはじめとする CGM（消費者生成メディア）の発達は、新しいコミュニケーション様態を生みだし、テキストから、流行主題や重要トピックのみならず、コミュニケー

ションのあり方を抽出・分析することへの期待が高まっている。
従来から、研究代表者は、大量かつ多種の情報資源の蓄積の中で、テキストによるコミュニケーションと社会との関わりに注目し、感情、評価、価値観、論理性、対話力、表現力など、書き手(話し手)の情報発信行為を動機づける要因をソーシャル・アティテュードとよび、新たに研究対象としてきた。特に、テキストがもつ文体的特徴が、これをあらわすと考え、計算文体論手法を改良するとともに、新たな応用分野を模索してきた。
オンライン上の学習、知識交換コミュニティは、このような、ソーシャル・アティテュードが重要な役割をもつ対象であるにも関わらず、その特徴は未だ十分明らかにされていない。そこで、これを主要な分析対象として、ソーシャル・アティテュードの抽出、分析を行うとともに、他のテキストをも利用して、ソーシャル・アティテュード抽出のための分析手法の検討を行うこととした。
オンライン学習、知識交換コミュニティのテキストは、多様かつ断片的な、ダイアログテキストであり、新たな特徴をもつものである。従って、これを分析することは、他の様々なテキストデータの分析にも有効な知見を提出するものである。

２．研究の目的

本研究により、オンライン上の、教育、知識交換コミュニティにおける、人々のコミュニケーションの特徴が明らかになる。また、これをもとに、感情、評価、価値観、論理性、対話力、表現力などソーシャル・アティテュードをあらわす指標を提案することができる。
同時に、広く、他のテキストをも含めて、ソーシャル・アティテュード抽出、分析のための、分析手法を提案することができる。

３．研究の方法

(1)テキストデータの整備

オンライン学習、知識交換コミュニティのテキストデータを作成する。データファイルを作成、整形、ノイズを除去し、それぞれのデータセットを作成する。

(2)基本的な特徴量の計量

①作成成したテキストファイル（本文部分）に対して、MeCab（mecab.sourceforge.net）による形態素解析、CaboCha（chasen.org/~taku/software/cabocha）による係り受け解析を適用する。
②基本的なテキスト統計量（文章長、パラグラフ長、文長、文字（n-gram）頻度、形態素(n-gram）頻度、共起頻度、 係り受け頻度、要約特徴量）を計量する。

(3)探索的分析

探索的分析を行うことで、教育、知識交換コミュニティテキストに影響を与える主要な要因を明らかにし、その後の分析の統制条件（日本語一般の変化の影響、諸属性の影響等）を確認する。
特徴量は、②に記載したものを用い、統計手法としては、主成分分析、探索的因子分析、階層的クラスター分析を適用する。

(4)機械学習による特徴分析

注目するカテゴリーごとの分類問題として機械学習を適用する。分類に際しての変数の重要度を計算することで、重要な特徴量を明らかにする。機械学習手法としては、分類性能の頑強さ、重要度計算アルゴリズムの適切さなどから、ランダムフォレスト機械学習法を利用する。

(5)新たな特徴量、分析手法の検討

他のテキストも利用して、分析に有効な特徴量、分析手法の検討を行う。共起を利用した特徴量、とりわけ、ネットワーク特徴量や分布特徴量の有効性を検討する。

(6)全体の統括

研究全体を総括し、実証的知見を整理すると共に、分析手法の提案を行う。

４．研究成果

以下では、3.研究の方法で記した項目ごとに研究成果を記述する。研究全体の成果、今後の課題については、(6)に記載する。

(1)テキストデータの整備

オンライン学習、知識交換コミュニティのテキストデータを作成した。前者に関しては、手作業で収集することとし、後者に関しては、Yahoo!株式会社が国立情報学研究所を通じて提供しているYahoo!知恵袋データを利用することとした。それぞれ、整形、ノイズ除去を行い、テキストファイル、データセットを作成した。

(2)基本的な特徴量の計量

①作成したテキストファイルに対して、MeCab による形態素解析、CaboCha による係り受け解析を適用した。

②それぞれのテキストについて、基本的なテキスト特徴量を計量した。

(3)探索的分析、および、(4)機械学習による特徴分析

体系的なデータセットが提供されていることから、Yahoo!知恵袋データに関しては、(1)(2)の作業が短期に終了した。そこで、まずこれを中心として分析を行うこととし、主成分分析、ランダムフォレスト機械学習法を適用し、テキスト特徴量に影響を与える主要な影響要因を検討し、また、カテゴリーごとの特徴を明らかにした。

具体的には、Yahoo!知恵袋のテキストを PC、恋愛相談等、質問項目のカテゴリーおよび、質問（questions）、優れた回答（best answers）、通常の回答（normal answers）にわけ、機能語 bag of words モデルと主成分分析、ランダムフォレスト機械学習法によって、それぞれのカテゴリーおよび質問、優れた回答、通常の回答の差異を明らかにした。以上の成果を、*JADT2010*（5.[学会発表]②）で発表した。

Yahoo!知恵袋に関しては、質問者、回答者の分布構造が、カテゴリーの特徴に影響を与えていると考えられる。そこで、上記の分析結果をさらに深く検討するために、カテゴリーごとに質問者―回答者グラフを作成、集中度指標を計算した。その結果、カテゴリーのコミュニケーションタイプ（知識交換、相談、議論）別に、コミュニティの質問者・回答者の分布構造が異なるという知見を得た。これは、Adamic らが Yahoo! Answers について、異なる指標を用いて示した知見と整合的なものであり、コミュニケーションタイプ別のコミュニティ構造の差異を明らかにするものである。以上の成果を情報処理学会全国大会（5.学会発表③）で発表した。

(5)新たな特徴量、分析手法の検討

他のデータも利用して、ソーシャル・アティテュードの抽出、分析に有効な分析手法の検討、とりわけ、新たな特徴量の提案を行った。まず、青空文庫コーパスを利用し、共起に基づく特徴量が著者の性格、心理等を分析する、著者プロファイリング、計算社会言語学等の著者分析のタスクに有効であることを示した。同時に、政治テキストを利用して、手作業によるカテゴリー化や特定の品詞に関する分布特徴量が、著者の特徴や時代ごとの変化を明らかにするために有効であることを示した。上記の成果を、*International Relations of the Asia-Pacific*、行動計量学（5.雑誌論文①②）、*JADT2010* において発表した（5.学会発表①）。

(6)全体の総括

以上、本研究では、大量かつ多種の情報資源蓄積の中で、新たなテキストデータの利用を念頭に、学習、知識交換コミュニティに焦点をあて、そのテキストの特徴を分析し、分析手法を検討することで、ソーシャル・アティテュードの抽出、分析を目指した。

当初の研究計画のうち、(1)テキストデータの整備、(2)基本的な特徴量の計量について、当初予定通りの成果を得た。(3)探索的分析、(4)機械学習による特徴分析については、Yahoo!知恵袋データを先行して分析し、そのテキスト特徴量に影響を与える主要因やカテゴリーごとの特徴、質問、優れた回答、通常の回答の特徴を明らかにすることができた。(5)新たな特徴量、分析手法の検討については、他のテキストをも利用することで、ソーシャル・アティテュード抽出、分析に有効な特徴量、とりわけ、共起にもとづく特徴量、分布特徴量の有効性を明らかにすることができた。

一部データについては、未だ、研究が(1)(2)の段階にとどまっており、(3)(4)の分析をさらに進める余地がある。今後、この点を進め、研究発表につなげていきたい。さらに、今後、本研究で得られた成果をもとに、ソーシャル・アティテュードを計量するための指標の提案や、任意のテキストからソーシャル・アティテュードを自動的に抽出、分析するシステムの構築を目指したい。

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 SUZUKI, Takafumi, Investigating macroscopic transitions in Japanese foreign policy using quantitative text analysis, *International Relations of the Asia-Pacific*, 査読有，2011 年，採録決定

②鈴木崇史、影浦峡、名詞の分布特徴量を用いた政治テキスト分析、行動計量学 、査読有、2011 年、採録決定

〔学会発表〕（計 3 件）

① SUZUKI Takafumi, KAWAMURA Shuntaro, YOSHIKANE Fuyuki, KAGEURA Kyo, AIZAWA Akiko, Co-occurrence-based indicators for investigating authors' styles, Bolasco, S., Chiari, I. and Giuliano, L. (ed.) *Statistical Analysis of Textual Data, Proceedings of*

10th International Conference Journées d'Analyse statistique des Données Textuelles 9-11 June 2010 - Sapienza University of Rome, Edizioni Universitarie di Lettere Economia Diritto, Milano, 363-373,
② SUZUKI Takafumi, KAWAMURA Shuntaro, AIZAWA Akiko, Exploratory analysis of stylistic characteristics in Japanese Q&A communities, Bolasco, S., Chiari, I. and Giuliano, L. (ed.) *Statistical Analysis of Textual Data, Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles 9-11 June 2010 - Sapienza University of Rome*, Edizioni Universitarie di Lettere Economia Diritto, Milano, 355-362,
③浅石卓真・村山遼・河村俊太郎・芳鐘冬樹・鈴木崇史・相澤彰子、ネットワーク構造からみた Q&A コミュニティの分析、情報処理学会創立 50 周年記念（第 72 回）全国大会発表論文集，4.825-4.826 ，2011/3/8～12，東京大学

〔その他〕
Web **サイト**
http://researchmap.jp/stratovarius

６．研究組織

(1)研究代表者
鈴木　崇史
（SUZUKI TAKAHUMI）
東洋大学・社会学部・講師
研究者番号：70507037

（2）研究分担者（0）

（3）連携協力者（0）

(4)研究協力者
河村　俊太郎
(KAWAMURA SHUNTARO）
東京大学大学院・教育学研究科・大学院博士課程

浅石　卓真
（ASAISHI TAKUMA）
東京大学大学院・教育学研究科・大学院博士課程

村山　遼
（MURAYAMA RYO）
東京大学大学院・教育学研究科・大学院修士課程