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　２３日現在

機関番号：１２１０２
研究種目：基盤研究（Ｂ）
研究期間：2008～2010
課題番号：20300032
研究課題名（和文）　トピックの特性を言語間で比較・対照分析する多言語ウェブテキストマイニングの研究
研究課題名（英文）　Research on Cross-Lingual Web Text Mining for Comparative Analysis of Topics
研究代表者
宇津呂　武仁　　(UTSURO　TAKEHITO)
筑波大学・大学院システム情報工学研究科・准教授
研究者番号：　90263433

研究成果の概要（和文）：　本研究では，ウェブ上で収集可能な多言語文書を情報源として，多言語での関心動向や，意見の分布を分析し，国・文化・言語の間にどのような違いがあるのかを発見する過程を支援するテキストマイニング技術について研究を行った．Wikipedia の概念体系を利用して多言語文書空間を索引付けするとともに，同一のトピックに関する多言語文書間において，文化間の差異を発見する過程を支援する技術を実現した．

研究成果の概要（英文）：　In this project, given a collection of multilingual documents available on the Web, we cross-lingually and cross-culturally compare less well known facts and opinions that are observed in the collected documents. We conceptually index the space of multilingual documents based on conceptual hierarchy of Wikipedia. We have realized techniques for assisting the process of discovering cross-cultural differences among multilingual documents that are closely related to a given topic.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	5,500,000	1,650,000	7,150,000
２００９年度	4,700,000	1,410,000	6,110,000
２０１０年度	4,500,000	1,350,000	5,850,000
年度			
年度			
総　計	14,700,000	4,410,000	19,110,000

研究分野：　情報工学
科研費の分科・細目：　情報学・メディア情報学・データベース
キーワード：　ディレクトリ・情報検索，多言語処理，テキストマイニング，トピック分析，ブログ，ニュース，Wikipedia，スパムブログ

１．研究開始当初の背景

(1) ウェブの世界では情報爆発の問題が指摘されている．商用検索エンジンなども普及してはいるが，爆発的に増加する情報の中で，自分が必要とする情報が検索結果のランキング下位のロングテールに埋もれてしまうことも多い．学術分野においては，この問題への取り組みとして，自然言語処理技術を用いた高機能な次世代型検索エンジンの実現が重要視され，主に単言語において，事実や意見情報を収集・集約する技術について研究成果があがっている．しかしこれらの成果はまだ研究開発段階であり，商用検索エンジン等で採用されて実用規模で普及するには至っておらず，依然として情報爆発の壁は高く厚い．

(2) 一方，ウェブ上の外国語文書から言語を横断して情報を検索する場合には，上記の

情報爆発の問題に加えて，外国語文書を日本語に翻訳してから情報にアクセスするという言語の壁の問題が立ちはだかる．廉価な翻訳ソフトも利用可能ではあるが，韓日翻訳等の構文構造の近い言語間の翻訳を除いては，依然として誤訳が多く性能面での問題が多い．

(3) 研究代表者，分担者，連携研究者のグループでは，多言語ニュース・ブログの傾向と国籍による違いについての予備調査結果を報告している．その中で特に，ウェブ上の外国語文書における情報爆発の問題を克服するためには，国・文化・言語の違いを考慮して，情報の収集・集約・比較対照分析を行う必要があるという知見を得た．

２．研究の目的

本研究では，ウェブ上で収集可能な多言語文書を情報源として，多言語での関心動向や，意見の分布を分析し，国・文化・言語の間にどのような違いがあるのかを発見する過程を支援するテキストマイニング技術について研究を行う．特に，以下の観点において，研究目的を設定する．

(1) 情報の単位としてトピックに注目し，ウェブ上の多言語ニュース・ブログ等の各ジャンルにおいて，トピックの持つ特性，および，各トピックがどのような観点で強く関心を持たれているか，どのような意見を持たれているかを特定する．

(2) 各トピックの持つ特性，観点・意見の分布といった多様なアスペクトに対して，国・文化・言語の間の差異に着目する．ウェブから収集される多言語文書を対象として，この差異を発見する過程を支援する技術を実現する．

３．研究の方法

(1) 本研究では，多言語に渡る網羅的なトピック体系として，新規の流行的トピックの掲載が最も早いことが期待できる Wikipedia を利用する．Wikipedia には，対訳関係にある項目間リンクが一定数存在する．この Wikipedia を知識源として用いて，多言語文書の概念的な索引付けを実現する．

(2) 本研究では，多言語・多ジャンルの情報源から発信される情報の間の差異を発見する過程を支援することを目的とする．そのために，百科事典的事項を記載した Wikipedia，詳細な事実情報を報道するニュース，個人の主観的意見や経験などを豊富に記載したブログ，という多ジャンルの情報源の間で，記述内容や意見の差異を発見する過程を支援する．また，同一のトピックについて記述された多言語のブログにおける文化間の差異を発見する過程を支援する．

(3) 本研究においては，各トピックの特性，および，関心動向を推定し，それを言語間で比較・対照分析することが研究の根幹をなすが，スパムは，特に一般利用者の関心動向に便乗する傾向が強いため，スパムを除去した上で一般利用者の関心動向を正確に把握することが不可欠である．そこで，スパムにおけるトピックの特性を把握するとともに，的確にスパムを除去する技術を実現する．

４．研究成果

(1) 本研究の基盤となる技術として，同一のトピックについて書かれた多言語文書を高精度で収集することにより，文化間差異の発見を効果的に支援することが必須となる．本研究では，このことを保証するため，Wikipedia エントリを知識源として多言語文書を索引付けする手法を開発した．この研究においては，Wikipedia エントリから抽出した関連語の頻度をスコアとして用いる手法，および，それらの頻度を機械学習の素性として用いる手法の評価を行い，既存の検索エンジン API の性能を大幅に改善できることを実証した．

(2) 同一のトピックに関する多言語文書間において，文化間の差異を発見する過程を支援する技術を実現するために，以下の研究を行った．

① ある同一のトピックについてまとまった規模の記述が書かれたブログサイトを，日英各言語について検索し，その記述内容を二言語間で対照分析する方式を実現した．これによって，同一のトピックが対象の場合でも，ブログ特有の個人レベルの関心が日英ブログの間で異なっている様子や，個人が持つ意見の分布が日英ブログの間で異なっている様子が容易に観測可能となった．

② 文化間差異の発見過程を支援するインタフェースを用いて，多数のトピックを対象として，文化間差異発見過程の定性的評価を行った．さらに，サンプルトピックを対象として，日英間の差異が大きいもの，中程度のもの，小さいものに大別できることを示した．

③ 特定のトピックに関して詳細な記述を含

むブログ記事集合に対して，特定トピックにおける詳細な話題・関心事項をファセット(観点)とみなして，各ファセットごとにブログ記事を分類し，トピック空間・ブログ空間の集約を実現した．この枠組みにおける基本的な知識源として，Wikipedia を用いた．さらに，この方式を多言語化し，多言語間でトピック空間の集約結果，および，ブログ空間の集約結果の差異の分析を行った．トピック空間の集約結果におけるファセット(観点)一覧を比較し，異なる言語の間で共通するファセット(観点)，および，各言語特有のファセット(観点)の両方を観測した．また，それぞれのファセットに分類されるブログ記事集合の集約結果を比較し，異なる言語の間で共通する関心事項，および，各言語特有の関心事項の両方を観測した．

(3) ウェブ上の文書のジャンルとして，主に事実を報道するニュースと，主として一般利用者の意見や経験を伝えるブログとを対比的にとりあげ，ニュース，ブログ間で関連する項目や記述を相補的に検索する方式を実現した．これによって，ニュース・ブログという異ジャンル間で，観点の差異や意見の有無を発見する過程の支援が可能となった．

(4) ブログ等のウェブ文書を対象とした情報分析においては，スパムの除去が必要不可欠である．本研究においては，ウェブ情報分析の正確さ，頑健さを保持することを目的として，スパムにおけるトピックの特性を把握し，的確にスパムを除去することを実現した．

① 日本語スパムブログに含まれるトピックの特性を分析した．また，他者の文書を盗用してスパムを自動生成する際のスパム作成者の嗜好を分析した．分析の結果，一人のスパム作成者が大量のスパムブログを自動生成していることが確認できた．

② 機械学習の枠組みにより，ブログにおけるトピック分析の障害となるスパムブログ除去方式を実現した．また，能動学習の枠組みにより，年とともに変貌するスパムブログのうちの重要変化分を効率よく同定する方式を実現した．

③ 教師なし学習によるスパムブログ検出の枠組みにおいて，各ブログサイトの間の HTML 構造の類似性を利用する方式を考案し，主要 10 ブログホストのうちの半数以上を対象として，その有効性を検証した．この方式により，年とともに変化するスパムブログの傾向に追随して，HTML 構造の類似するスパムブログを発見することが容易になった．

④ ブログにおいてアフィリエイト収入を得ることを目的とするスパムブログについて，HTML 構造の類似性およびアフィリエイト ID という異なる二種類の手がかりの特性の分析を行った．特に，既知のスパムブログに対して HTML 構造が類似するブログサイトを大規模に収集することにより，既知のスパムブログに類似するスパムブログが高密度で自動収集できることを示した．また，これらの二種類の手がかりを単独で用いた場合には，それぞれの適用範囲が十分ではなく，両者の手がかりを併用する必要があることを示した．さらに，両者いずれの手がかりによっても検出できないスパムブログに対して，機械学習を適用し，高適合率の検出を実現した．

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 6 件）

① Yuki Sato, Daisuke Yokomoto, Hiroyuki Nakasaki, Mariko Kawaba, Takehito Utsuro and Tomohiro Fukuhara, Linking Topics of News and Blogs with Wikipedia for Complementary Navigation, Lecture Notes in Computer Science, 査読有, 6045巻, 2010, 75-87

② Hiroyuki Nakasaki, Yusuke Abe, Takehito Utsuro, Yasuhide Kawada, Tomohiro Fukuhara, Noriko Kando, Masaharu Yoshioka, Hiroshi Nakagawa and Yoji Kiyota, Cross-Lingual Analysis of Concerns and Reports on Crimes in Blogs, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 査読有, 40巻, 2010, 315-320

③ Yusuke Abe, Takehito Utsuro, Yasuhide Kawada, Tomohiro Fukuhara, Noriko Kando, Masaharu Yoshioka, Hiroshi Nakagawa, Yoji Kiyota and Masatoshi Tsuchiya, Extracting Concerns and Reports on Crimes in Blogs, Lecture Notes in Computer Science, 査読有, 6335巻, 2010, 498-509

④ 中崎寛之，川場真理子，横本大輔，宇津呂武仁，福原知宏，多言語Wikipediaエントリを知識源とする特定トピックの日英ブログサイト検索と日英対照ブログ分析，人工知能学会論文誌， 査読有， 25巻， 2010，613-622

⑤ 川場真理子，中崎寛之，横本大輔，宇津

呂武仁, 福原知宏, Wikipedia概念体系とブログ空間の間のトピック対応の推定, 査読有, 日本データベース学会論文誌, 8巻, 2009, 17-22

⑥ Hiroyuki Nakasaki, Mariko Kawaba, Takehito Utsuro, and Tomohiro Fukuhara, Mining Cross-Lingual/Cross-Cultural Differences in Concerns and Opinions in Blogs, Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence, 査読有, 5459, 2009, 213-224

〔学会発表〕(計49件)

① Daisuke Yokomoto, Kensaku Makita, Takehito Utsuro, Yasuhide Kawada, and Tomohiro Fukuhara, Utilizing Wikipedia in Categorizing Topic related Blogs into Facets, 12th Conference of the Pacific Association for Computational Lingustics, 2011年7月21日, クアラルンプール(マレーシア)

② Taichi Katayama, Akihito Morijiri, Soichi Ishii, Takehito Utsuro, Yasuhide Kawada, and Tomohiro Fukuhara, Comparing Similarity of HTML Structures and Affiliate IDs in Splog Analysis, Database Systems for Adanced Applications: 16th International Conference, DASFAA 2011, International Workshops: SNSMW, 2011年4月22日, 香港(中国)

③ 片山太一, 森尻惇宜史, 石井聡一, 宇津呂武仁, 河田容英, 福原知宏, HTML構造の類似性およびアフィリエイトを用いたスプログの分析, Webとデータベースに関するフォーラム(WebDB2010), 2010年11月11日, 早稲田大学(東京都)

④ Taichi Katayama, Takayuki Yoshinaka, Takehito Utsuro, Yasuhide Kawada, and Tomohiro Fukuhara, Detecting Splogs using Similarities of Splog HTML Structures, 4th International Conference on Ubiquitous Information Management and Communication, 2010年1月14日, 水原(韓国)

⑤ Mariko Kawaba, Daisuke Yokomoto, Hiroyuki Nakasaki, Takehito Utsuro, and Tomohiro Fukuhara, Towards Conceptual Indexing of the Blogosphere through Wikipedia Topic Hierarchy, 23rd Pacific Asia Conference on Language, Information and Computation, 2009年12月5日, 香港(中国)

⑥ Hiroyuki Nakasaki, Mariko Kawaba, Sayuri Yamazaki, Takehito Utsuro and Tomohiro Fukuhara, Visualizing Cross-Lingual/Cross-Cultural Differences in Concerns in Multilingual Blogs, 3rd International AAAI Conference on Weblogs and Social Media, 2009年5月19日, サンノゼ(アメリカ)

６．研究組織

(1) 研究代表者

宇津呂　武仁 (UTSURO　TAKEHITO)
筑波大学・大学院システム情報工学研究科・准教授
研究者番号： 90263433

(2)研究分担者

藤井　敦 (FUJII　ATSUSHI)
東京工業大学・大学院情報理工学研究科・准教授
研究者番号： 30302433

(3)連携研究者

中川　裕志 (NAKAGAWA　HIROSHI)
東京大学・情報基盤センター・教授
研究者番号： 20134893

清田　陽司 (KIYOTA　YOJI)
東京大学・情報基盤センター・特任講師
研究者番号： 10401316

福原　知宏 (FUKUHARA　TOMOHIRO)
(H20-21) 東京大学・人工物工学研究センター・寄付研究部門教員
(H22) 独立行政法人産業技術総合研究所サービス工学研究センター・特別研究員
研究者番号： 10401316