様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１９　日現在

機関番号：　１２１０２
研究種目：　基盤研究(B)（一般）
研究期間：　2014～2016
課題番号：　２６２８０１１１

研究課題名（和文）ウェブ検索における情報要求観点の言語間比較・対照分析システムの研究

研究課題名（英文）A Study on a System for Comparative Analysis of Web Search Information Needs between Languages

研究代表者

宇津呂　武仁（UTSURO, Takehito）

筑波大学・システム情報系・教授

研究者番号：９０２６３４３３

交付決定額（研究期間全体）：（直接経費）　　12,200,000円

研究成果の概要（和文）：本研究では，ウェブ検索者の情報要求観点を情報源とする「言語間対照分析支援システム」を作成した．具体的には，各言語の検索エンジン・サジェストおよび質問回答サイトを情報源として，特定のクエリ・フォーカスに対する情報要求観点，および，その観点に関して得られる情報を収集した．そして，収集された情報を言語間で比較・対照分析することにより，他国と自国との間の文化・関心・意見の違いを発見する過程を支援するシステムを作成した．

研究成果の概要（英文）：In this project, we developed a system for comparative analysis of Web search information needs between languages. As for the source of collecting information in multiple languages, we focus on search engine suggests as well as question-answer sites in multiple languages. Given a specific query focus, we collect search information needs as well as Web pages and question-answer pairs for those search information needs. Then, developed a system for assisting the process of comparatively analyzing those collected search information needs and the specified Web pages / question-answer pairs between multiple languages and that of discovering differences of culture, concern, and opinions between multiple countries.

研究分野：　情報工学

キーワード：　ディレクトリ・情報検索　多言語処理　情報要求観点　検索エンジン・サジェスト　言語間比較・対照分析

１．研究開始当初の背景

(1) 宇津呂，吉岡は，本研究課題の前身として，基盤(B)「トピックの特性の多観点把握に基づく多言語ウェブテキストの言語間対照分析システム」(代表：宇津呂，H23～25)等において，多言語ニュース・ブログを情報源として，詳細な話題・関心，時系列特性，書き手の実体験，賛否・主観といった多様な観点のもとで，国・文化・言語間の差異発見過程を支援する方式を開発した．

(2) (1)の研究では，ブログ等における言及数の多少によって，関心の度合いや関心の動向を把握していた．しかし，時事的話題のように，時間的変遷が急激な場合には，ブログ等における言及数の動向が収束し関心の動向や度合いが把握できるまでの間に遅延が生じ，関心動向の迅速な把握が困難であった．この遅延を克服するために，発想を転換し，ブロガー等のウェブ執筆者の対極に位置するウェブ検索者が，報道等の一次情報に対して行う検索行動に着目した．そして，ウェブ検索者の情報要求観点を直接収集することによって，(1)の研究における遅延を克服でき，関心動向の迅速な把握が可能となると考えた．

(3) 一方，時間的変遷が緩やかな文化・慣習に関する話題の場合も，関心の動向や度合いが，ウェブ検索者の情報要求観点からしか収集できないものが多数存在することが判明した．

(4) 以上の知見を得て，本研究課題を提案するに至った．

２．研究の目的

本研究では，ウェブ検索者の情報要求観点を情報源とする「言語間対照分析支援システム」を作成する．特に，各言語の検索エンジン・サジェストおよび質問回答サイトを情報源として，特定の検索対象に対する情報要求観点を収集し，言語間で比較・対照分析することにより，他国と自国との間の文化・関心・意見の違いを発見する過程を支援するシステムを作成する．

３．研究の方法

(1) 検索エンジン･サジェストを対象として，ウェブ検索者の情報要求観点を収集する．検索エンジンに対して，一検索対象キーワード当り約 100 通りの文字列を指定し，最大約 1,000 語のサジェストを収集する．ここで，収集されるサジェストの多くは話題が重複し冗長なため，冗長性集約方式を開発する．

(2) (1)の情報要求観点のもとで収集されるウェブ検索結果を集約・俯瞰する方式を開発する．この方式においては，(1)で集約した情報要求観点を活用することによって，効率的かつ効果的なウェブ検索結果の集約・俯瞰を実現する．

(3) ウェブ検索者の情報要求観点，および，その観点に関して得られる情報を収集するための情報源として，質問回答サイトを対象とする．日中二言語の各質問回答サイトから質問回答事例を収集する．次に，(1)で収集したサジェストを情報要求観点として，各サジェストを含む質問回答事例を収集した後，トピックモデルを適用して質問回答事例集合を集約・俯瞰する．

(4) (2)および(3)において収集・集約された情報に対して，ウェブ検索結果および質問回答事例の間での差異の有無を検証する．

(5) 特定の検索クエリに対して，ウェブ検索者が検索時に指定した情報要求観点，および，その観点に関して得られる情報に対して，利用者が比較･対照分析を行い，言語間の違いを発見する過程を支援する方式を実現する．

４．研究成果

(1) 検索エンジン・サジェストを対象として，ウェブ検索者の情報要求観点を収集した．ここで，収集されるサジェ ストの多くは話題が重複し冗長なため，冗長性集約方式を二種類開発した．方式 1 では，各サジェストを情報要求観点として指定して収集したウェブページ集合に対して LDA 等のトピックモデルを適用して，ウェブページ集合の話題集約を行った後，情報要求観点として指定したサジェストそのものを集約した．方式 2 では，各サジェストを情報要求観点として指定して収集した検索結果のスニペット間の類似度を用いてサジェスト間の類似度を定義し，この類似度を用いてサジェストの集約を行った．

(2)ウェブ検索者の情報要求観点，および，その観点に関して得られる情報を収集するための情報源として，質問回答サイトを対象とした．(1)で収集したサジェストを情報要求観点として，各サジェストを含む質問回答事例を 収集した後，トピックモデルを適用して質問回答事例集合を集約した．

(3) サジェストと質問回答事例の間で，ウェ

ブ検索者の情報要求観点における差異の有無について分析した．具体的には，収集された情報要求観点，および，各情報要求観点のもとで収集される情報を定量的に比較した．

(4) (1)のうちの方式 1 の「トピックモデルを用いる方式：各サジェストを情報要求観点として指定して収集したウェブページ集合に対してトピックモデルを適用して，ウェブページ集合の話題集約を行った後，情報要求観点として指定したサジェストそのものを集約する」方式を対象として，各トピックにおけるウェブ検索結果を集約する方式を開発した．この方式では，各トピックに対応するサジェストを利用して，サジェストを網羅する最少数のウェブ検索結果を提示する方式を開発し，従来方式の約 2 倍の集約率を達成した．

(5) 検索エンジン・サジェストを利用することにより，質問回答サイトおよびウェブからノウハウ知識を相補的に収集する方式を開発した．この方式においては，検索エンジン・サジェストを索引として収集される情報に加えて質問回答サイトから得られる情報を相補的に利用し，それらを混合して集約することを実現した．特に，質問回答サイト以外の一般的ウェブページを情報源として，質問回答サイトのノウハウ知識を補充する新ノウハウ知識が収集できることを示した．

(6) ウェブ検索者の関心事項に着目することにより，ウェブ上の情報を多言語(日本語・中国語)間で比較・対照分析し，他国の情報の収集を支援するとともに，言語間の差異発見の過程を支援する方式を実演した．特に，サジェストの集約過程においてトピックモデルを用いる方式を採用し，サジェスト集約を自動で行なった上で，その結果に対して日中間対照分析を行う方式を開発した．

(7) (5)の方式においては，検索エンジン・サジェストを索引として収集される情報に加えて質問回答サイトから得られる情報を相補的に利用し，それらを混合して集約することによって，両方の情報源の特性を活かしてノウハウ知識の収集を実現した．これに対して，日本語および中国語の各言語において収集したノウハウ知識を，日中二言語間で比較・対照分析する方式を実現した．

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 3 件）

① Takehito Utsuro, Chen Zhao, Linghan Xu, Jiaqi Li and Yasuhide Kawada, An Empirical Analysis on Comparing Market Share with Concerns on Companies measured through Search Engine Suggests, Global Journal of Flexible Systems Management, 査読有, 18 巻, 2017, 3-19.

② 聶添, 陳磊, 今田貴和, 宇津呂武仁, 河田容英, 検索エンジン・サジェストを情報源とするウェブ検索者の情報要求観点の日中間対照分析, 知能と情報 査読有, 27 巻, 2015, 527-532.

③ Takehito Utsuro, Yusuke Inoue, Takakazu Imada, Masaharu Yoshioka, and Noriko Kando, Detecting Bursty Topics of Correlated News and Twitter for Government Services, Social Media for Government Services, 査読有 ,2015, 378-389.. 10.1007/978-3-319-27237-5_7

〔学会発表〕（計 29 件）

① Tian Nie, Yi Ding, Chen Zhao, Youchao Lin, Takehito Utsuro, and Yasuhide Kawada. Clustering Search Engine Suggests by Integrating a Topic Model and Word Embeddings., 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2017 年 6 月 26 日, 金沢勤労者プラザ(石川県)

② Chen Zhao, Jiaqi Li, Tian Nie, Yi Ding, Linghan Xu, Takehito Utsuro, Yasuhide Kawada, and Noriko Kando. Identifying Major Contents among Web Pages with Search Engine Suggests by Modeling Topics in a question-answer site, 11th International Conference on Ubiquitous Information Management and Communication, 2017 年 1 月 6 日. B-CON PLAZA (大分県).

③ Jiaqi Li, Chen Zhao, Youchao Lin, Mizuho Baba, Tian Nie, Takehito Utsuro, Yasuhide Kawada, and Noriko Kando A Method of Collecting Know-how Knowledge based on Question-Answer Examples and Search Engine Suggests, 11th International Conference on Ubiquitous Information Management and Communication, 2017 年 1 月 6 日. B-CON PLAZA (大分県).

④ Takakazu Imada, Yusuke Inoue, Lei Chen, Syunya Doi, Tian Nie, Chen Zhao, Takehito Utsuro and Yasuhide Kawada, Analyzing Time Series Changes of Correlation between Market Share and Concerns on Companies measured through Search Engine Suggests,

10th International Conference on Language Resources and Evaluation, 2016年5月26日, Portoroz(スロベニア).

⑤ Takakazu Imada, Yusuke Inoue, Lei Chen, Syunya Doi, Takehito Utsuro and Yasuhide Kawada, An Empirical Study on Optimal Correlation between Market Share and Concerns on Companies measured through Search Engine Suggests, 30th International Conference on Advanced Information Networking and Applications Workshops, 2016年3月24日, クラン・モンタナ(スイス).

⑥ Yusuke Inoue, Takakazu Imada, Syunya Doi, Lei Chen, Takehito Utsuro and Yasuhide Kawada, Selecting Web Search Results of Diverse Contents with Search Engine Suggests and a Topic Model, 30th International Conference on Advanced Information Networking and Applications Workshops, 2016年3月23日, クラン・モンタナ(スイス).

⑦ Takakazu Imada, Yusuke Inoue, Lei Chen, Syunya Doi, Takehito Utsuro and Yasuhide Kawada. Analyzing Concerns on Companies through Statistics of Search Engine Suggests and its Correlation to Market Share, 10th International Conference on Ubiquitous Information Management and Communication, 2016年1月5日,ダナン(ベトナム).

⑧ Yusuke Inoue, Takakazu Imada, Syunya Doi, Lei Chen, Takehito Utsuro and Yasuhide Kawada, Clustering Search Engine Suggests by Modeling Topics of Web Pages collected with Suggests, 10th International Conference on Ubiquitous Information Management and Communication, 2016年1月5日,ダナン(ベトナム)

⑨ Ichiro Moriya, Yusuke Inoue, Takakazu Imada, Takehito Utsuro, Yasuhide Kawada and Noriko Kando, Overviewing the Knowledge of a Query Keyword by Clustering Viewpoints of Web Search Information Needs, 29th International Conference on Advanced Information Networking and Applications Workshops, 2015年3月27日, 光州(韓国)

⑩ Liyi Zheng, Tian Nie, Ichiro Moriya, Yusuke Inoue, Takakazu Imada, Takehito Utsuro, Yasuhide Kawada, and Noriko Kando. Comparative topic analysis of Japanese and Chinese bloggers., 7th International Symposium on Mining and Web, 2014年5月15日, ビクトリア(カナダ)

〔その他〕

・検索エンジンにおける情報集約・俯瞰機能
http://nlp.iit.tsukuba.ac.jp/research/list03-sg.html

６．研究組織

(1)研究代表者

宇津呂　武仁（UTSURO, Takehito）
筑波大学・システム情報系・教授
研究者番号： 90263433

(2)研究分担者

吉岡　真治（YOSHIOKA, Masaharu）
北海道大学・情報科学研究科・准教授
研究者番号： 40290879

乾　孝司（INUI, Takashi）
筑波大学・システム情報系・准教授
研究者番号： 60397031

(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