ウェブ検索者の情報要求観点の集約に関する研究

２０１９年　３月

聶　添

ウェブ検索者の情報要求観点の集約に関する研究

聶 添

システム情報工学研究科
筑波大学

2019年3月

概 要

21 世紀の情報社会では，情報を活用する科学・技術が私たちの豊かで便利な生活を支えている．様々な情報からノウハウが蓄積され，これまでに経験したことのない概念や手法が作り上げられた．情報を活用する事例として，まず，企業は競争相手に勝つために，常に競争相手より早く消費者のニーズを見つけ出し，それを満たした商品の市場導入が求められている．それに対して，企業が様々な手段でウェブから (ウェブアンケート，掲示板など) 大量の情報を集めて分析し，消費者の関心 (消費者ニーズ) を捉えた上で，ビジネス戦略を策定する．また，個人にとって，生涯ではいろんなことが発生する．これらを「ライフイベント」と呼ぶ．典型的なライフイベントとしては 就学， 就職， 結婚， 出産・子育て，教育などのものが挙げられる．それぞれのライフイベントにおいては様々な課題を直面するため，それらを解決する必要がある．例えば，就職については，年収やライフ・ワークバランスなどを考える必要がある．結婚については，相手を探す，資金準備などを考えなければならない．ライフイベントに対して，他人の関心 (自分が知らない・経験したことのない様々な課題) を知ることで，事前にそれらを解決するノウハウを入手し，ライフイベントを円滑に過ごすことができる．しかし，現在の世界においては情報が氾濫している．テレビ番組，ニュースサイト，ツイッター，ブログ，掲示板，EC サイト，ラジオなど，様々な情報源・メディアがたくさん存在しており，政治・経済から恋愛のノウハウまで幅広いジャンルの情報を日々提供している．それらの膨大な情報の中から自分にとって有用な情報を選別することはそれほど容易なことではない．また，年齢層・職種・性別・趣味によって人々が求めている情報も異なる．これらのことから,「自分のニーズに合った答えを情報の山からスピーディーに獲得する」という課題が考えられる．ここで，本論文においては,「自分のニーズに合った答えを情報の山からスピーディーに獲得する」という課題に対して，効率的かつ網羅的に知識を収集するために，人々の関心事項に着目する．

ウェブ検索者が欲しい情報を手に入れる手段の一つとしては，Google 等の検索会社が提供している検索エンジンを利用するのが一般的である．ここで，検索会社は，検索行動支援の一環として，検索エンジン・サジェスト・サービスを提供している．このサービスの特徴として，入力された検索語に対して，関連が強い語を検索エンジン・サジェストとして提示する．ここで，本論文では，ウェブ検索者が詳細な情報を得たい対象を**「クエリ・フォーカス」**と呼ぶ．そして，AND 検索の形でクエリ・フォーカスの次に入力され，より詳細な情報を得たい観点を示す語を**「情報要求観点」**と呼ぶ．検索エンジン・サジェストは，ウェブ検索者の検索履歴の収集結果に基づいて作られている．このことから，検索エンジン・サジェストには，ウェブ検索者の関心事項が反映されていると言える．検索エンジン・サジェ

ストを大量に集めることで，ウェブ検索者の関心事項を網羅的かつ効率的に収集することができる． また，検索エンジン・サジェストは多くの検索者の検索ログに基づいて作られたもので，同じ物事に対して関心を持つ人達の検索ログの話題が重複し，冗長であると考えられる．この問題に対応するためには，収集された検索エンジン・サジェストを適切に集約する必要がある．そこで，本論文では，検索エンジン・サジェストを情報収集の情報源として，ウェブ検索者の情報要求観点を網羅的に収集・集約する手法の確立を目的とする．更に，近年，国間の経済や文化の交流も促進され，お互いの国についての関心が高まりつつある．2015 年に中国人旅行者が日本での「爆買い」が話題になり，流行語年間大賞に選ばれた．2011-2017 年の訪日外国人数は増え続けており，2017 年は前年比 19.3%増の 2,869 万 1,000 人で統計開始以来の最高記録を更新した．観光庁は発表した 2018 年版の観光白書で，2016 年時点の観光国内総生産 (GDP) が約 10.5 兆円に上るとの試算を示した．観光 GDP は宿泊業や訪日外国人客向けの小売業など観光関連の産業がうむ付加価値をまとめたもので，4 年前に比べて 2 兆円も増加している．そして，企業は巨大なビジネスチャンスから外国人のニーズを見つけ出すだめに，文化間差異による嗜好の違いを把握する必要がある．外国人旅行者やビジネスマンは他国で円滑な生活を送るためにも，文化の違いによるその国の特有のトラブルを対応する必要がる．以上のことから，国間の文化間差異を発見することが大切だと考えられる．そこで，本論文では，日本と中国の関心事項の違いを分析することで，日中間文化間差異の発見を支援することをもう一つの目的として取り上げる．

第 1 章では，全体の序論として，研究の背景について述べる．

第 2 章では，ウェブ情報の集約の関連研究と本論文の位置付けについて述べる．

第 3 章では，本論文の研究課題一つである，ウェブ検索者の情報要求観点を網羅的に収集・集約する手法の確立について述べる．これまで，先行研究においては，一つのクエリ・フォーカスに対して，最大約 1,000 個のサジェストを収集する．そして，クエリ・フォーカスにサジェストを加えて，AND 検索によってウェブページの収集を行う．最大約 1,000 個のサジェストに対してこの方法を使うことによってウェブページの収集を行う．ここで，収集されるサジェスト，および，ウェブページ集合の双方において，話題の内容が重複し，冗長である点が問題である．それを解消するために，文書集合に LDA トピックモデルを適用し，話題が類似するサジェストをまとめて俯瞰する手法を提案した．しかし，トピックモデルを用いた話題集約の粒度が粗いという問題点を改善する必要がある．トピックモデルにおいては，トピック数が少ない場合は，集約されたトピックにおける話題の多様性は落ちる．一方，トピック数が多くなる場合には，トピック数が少ない場合には出現しなかった新たな話題を持つトピックが増えるが，たくさんのノイズが含まれるトピックも多数出現する．そして近年，深層学習技術により，単語の意味表現を分散表現によって表す方式が提案され，その有効性が報告されている．この

方式では，大規模なコーパスから語の意味のベクトル表現を学習し，これを用いて各語の周囲の語 (文脈) の予測を行う．そこで，本論文では，分散表現によって表現される話題の粒度が相対的に細かいところに着目し，分散表現とトピックモデルを併用することによって，検索エンジン・サジェストの類似度をより詳細に測定し，トピックごとに細分化することによって，小分類の抽出を行った．評価実験においては，人手によって作成された参照用小分類データに対して，3 つの素性を設定し，二種類の分散表現訓練用コーパスを用いた場合について，小分類集約結果の適合率・再現率の比較を行った．結果として，クエリ・フォーカス「就活」および「結婚」の両方において，分散表現訓練用コーパスとして「Wikipedia の全ページテキストのみ」を用いた場合より,「Wikipedia の全ページテキスト + 検索エンジン・サジェストによって収集されたウェブページテキスト」用いた場合の方が，高い性能が得られることを示した.

第 4 章では，ウェブ情報の集約における日中比較の関連研究と本論文の位置付けについて述べる．

次に，第 5 章において，本論文のもう一つの研究課題である，日中間の文化間差異を発見するための比較対照分析において，トピック単位での日中間比較対照分析とトピック内の小分類単位での日中間比較対照分析について述べる．具体的には，まず，人々が日常生活において最も関心を持つクエリ・フォーカスに対して，一つ目で提案された手法を用いて日中各言語の情報要求観点を網羅的に収集・集約する．次に，集約結果に対して，その内容を分析し，類似する内容のまとめ上げを行い，異なる内容の数を集計する．次に，日中の間でその内容の対応関係をつけ,「日本側でのみ観測した内容」,「中国側でのみ観測した内容」,「日中両言語で観測した内容」に分類して集計する．次に，分析の第 2 ステップにおいては,「日本側でのみ観測した内容」，および,「中国側でのみ観測した内容」を対象として，当該内容が観測されなかったと判定された言語側で追加の検索を行い，本当にその当該内容が検索されないかどうかの検証を行う．さらに，細かい粒度で日中間の文化間差異を比較するために，トピック内の小分類を分析対象とした日中間比較対照分析を行った．結果として，採用活動において重視する点の日中間の違いや両国の雇用形態の違いを発見することができた．これによって，情報要求観点を網羅的に収集することに基づいて，文化間差異を発見することができた.

最後に，第 6 章において，本論文のまとめおよび今後の課題について述べる．

付録 A では，質問解答事例およびウェブからの日中ノウハウ知識の収集および言語間対照分析について述べる．

目 次

図 目 次

表 目 次

第1章　序論

21 世紀の情報社会では，情報を活用する科学・技術が私たちの豊かで便利な生活を支えている．様々な情報からノウハウが蓄積され，これまでに経験したことのない概念や手法が作り上げられた．情報を活用する事例として，まず，企業は競争相手に勝つために，常に競争相手より早く消費者のニーズを見つけ出し，それを満たした商品の市場導入が求められている．それに対して，企業が様々な手段でウェブから (ウェブアンケート，掲示板など) 大量の情報を集めて分析し，消費者の関心 (消費者ニーズ) を捉えた上で，ビジネス戦略を策定する．また，個人にとって，生涯ではいろんなことが発生する．これらを「ライフイベント」と呼ぶ．典型的なライフイベントとしては 就学， 就職， 結婚， 出産・子育て，教育などのものが挙げられる．それぞれのライフイベントにおいては様々な課題を直面するため，それらを解決する必要がある．例えば，就職については，年収やライフ・ワークバランスなどを考える必要がある．結婚については，相手を探す，資金準備などを考えなければならない．ライフイベントに対して，他人の関心 (自分が知らない・経験したことのない様々な課題) を知ることで，事前にそれらを解決するノウハウを入手し，ライフイベントを円滑に過ごすことができる．しかし，現在の世界においては情報が氾濫している．テレビ番組，ニュースサイト，ツイッター，ブログ，掲示板，EC サイト，ラジオなど，様々な情報源・メディアがたくさん存在しており，政治・経済から恋愛のノウハウまで幅広いジャンルの情報を日々提供している．それらの膨大な情報の中から自分にとって有用な情報を選別することはそれほど容易なことではない．また，年齢層・職種・性別・趣味によって人々が求めている情報も異なる．これらのことから,「自分のニーズに合った答えを情報の山からスピーディーに獲得する」という課題が考えられる．ここで，本論文においては,「自分のニーズに合った答えを情報の山からスピーディーに獲得する」という課題に対して，効率的かつ網羅的に知識を収集するために，人々の関心事項に着目する．

ウェブ検索者が欲しい情報を手に入れる手段の一つとしては，Google 等の検索会社が提供している検索エンジンを利用するのが一般的である．ここで，検索会社は，検索行動支援の一環として，検索エンジン・サジェスト・サービスを提供している．このサービスの特徴として，入力された検索語に対して，関連が強い語を検索エンジン・サジェストとして提示する．ここで，本論文では，ウェブ検索者が詳細な情報を得たい対象を**「クエリ・フォーカス」**と呼ぶ．そして，AND 検

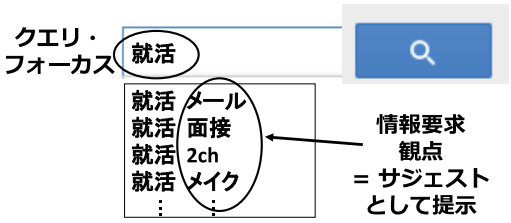


図 1.1: 検索エンジン・サジェストにおける情報要求観点の例

索の形でクエリ・フォーカスの次に入力され，より詳細な情報を得たい観点を示す語を「**情報要求観点**」と呼ぶ (図 1.1)．検索エンジン・サジェストは，ウェブ検索者の検索履歴の収集結果に基づいて作られている．このことから，検索エンジン・サジェストには，ウェブ検索者の関心事項が反映されていると言える．検索エンジン・サジェストを大量に集めることで，ウェブ検索者の関心事項を網羅的かつ効率的に収集することができる．また，検索エンジン・サジェストは多くの検索者の検索ログに基づいて作られたもので，同じ物事に対して関心を持つ人達の検索ログの話題が重複し，冗長であると考えられる．この問題に対応するためには，収集された検索エンジン・サジェストを適切に集約する必要がある．そこで，本論文では，検索エンジン・サジェストを情報収集の情報源として，ウェブ検索者の情報要求観点を網羅的に収集・集約する手法の確立を目的とする．更に，近年，国間の経済や文化の交流も促進され，お互いの国についての関心が高まりつつある．2015 年に中国人旅行者が日本での「爆買い」が話題になり，流行語年間大賞に選ばれた [1]．2011-2017 年の訪日外国人数は増え続けており，2017 年は前年比 19.3%増の 2,869 万 1,000 人で統計開始以来の最高記録を更新した [2]．また，観光庁は，2016 年時点の観光国内総生産 (GDP) が約 10.5 兆円に上るとの試算を示した [3]．観光 GDP は宿泊業や訪日外国人客向けの小売業など観光関連の産業がうむ付加価値をまとめたもので，4 年前に比べて 2 兆円も増加している．そして，企業は巨大なビジネスチャンスから外国人のニーズを見つけ出すだめに，文化間差異による嗜好の違いを把握する必要がある．外国人旅行者やビジネスマンは他国で円滑な生活を送るためにも，文化の違いによるその国の特有のトラブルを対応する必要がる．以上のことから，国間の文化間差異を発見することが大切だと考えられる．そこで，本論文では，日本と中国の関心事項の違いを分析することで，日中間文化間差異の発見を支援することをもう一つの目的として取り上げる．

本論文では，下記の二つの研究課題について取り組む． 一つ目の研究課題として，ウェブ検索者の情報要求観点を網羅的に収集・集約する手法の確立について

[1] https://www.jiyu.co.jp/singo/index.php?eid=00032

[2] https://www.jnto.go.jp/jpn/statistics/data_info_listing/pdf/180116_monthly.pdf

[3] 平成 30 年版観光白書第 II 部 日本経済における存在感が高まりつつある「観光」，http://www.mlit.go.jp/common/001260952.pdf

取り組む．これまで，先行研究においては，一つのクエリ・フォーカスに対して，最大約 1,000 個のサジェストを収集する．そして，クエリ・フォーカスにサジェストを加えて，AND 検索によってウェブページの収集を行う．最大約 1,000 個のサジェストに対してこの方法を使うことによってウェブページの収集を行う．ここで，収集されるサジェスト，および，ウェブページ集合の双方において，話題の内容が重複し，冗長である点が問題である．それを解消するために，文書集合に LDA トピックモデルを適用し，話題が類似するサジェストをまとめて俯瞰する手法を提案した．しかし，トピックモデルを用いた話題集約の粒度が粗いという問題点を改善する必要がある．トピックモデルにおいては，トピック数が少ない場合は，集約されたトピックにおける話題の多様性は落ちる．一方，トピック数が多くなる場合には，トピック数が少ない場合には出現しなかった新たな話題を持つトピックが増えるが，たくさんのノイズが含まれるトピックも多数出現する．そして近年，深層学習技術により，単語の意味表現を分散表現によって表す方式が提案され，その有効性が報告されている．この方式では，大規模なコーパスから語の意味のベクトル表現を学習し，これを用いて各語の周囲の語 (文脈) の予測を行う．そこで，本論文では，分散表現によって表現される話題の粒度が相対的に細かいところに着目し，分散表現とトピックモデルを併用することによって，検索エンジン・サジェストの類似度をより詳細に測定し，トピックごとに細分化することによって，小分類の抽出を行った． 二つ目の研究課題として，一つ目の研究課題において集約された日本・中国の検索者の情報要求観点に対して，日中間の文化間差異を発見するための比較対照分析を行った．具体的には，まず，人々が日常生活において最も関心を持つ「結婚」,「就活」の二つのクエリ・フォーカスに対して，一つ目の研究課題において提案された手法を用いて日中各言語の情報要求観点を網羅的に収集・集約する．次に，集約結果に対して，その内容を分析し，類似する内容のまとめ上げを行い，異なる内容の数を集計する．次に，日中の間でその内容の対応関係をつけ,「日本側でのみ観測した内容」,「中国側でのみ観測した内容」,「日中両言語で観測した内容」に分類して集計する．次に，分析の第 2 ステップにおいては,「日本側でのみ観測した内容」，および,「中国側でのみ観測した内容」を対象として，当該内容が観測されなかったと判定された言語側で追加の検索を行い，本当にその当該内容が検索されないかどうかの検証を行う．これによって，情報要求観点を網羅的に収集すること，および，それに基づき文化間差異を発見することを実現した．

トピックモデルと分散表現の併用によるサジェストの集約

本項目では，評価対象となるクエリ・フォーカスに対して，Google 検索エンジンを用いて，一クエリ・フォーカスにつき約 100 通りの文字列を指定する．具体的には，五十音，濁音，半濁音および「きゃ」や「ぴゃ」などの開拗音である．そして，1 通りの文字列当り約 10 個のサジェストを収集することにより，一クエリ・フォーカスあたり最大約 1,000 個のサジェストを収集し，これを集合 S とする．例えば検索窓に「就活 あ」と入力すると,「あいさつ」や「あなたの強み」等のサジェ

ストが表示されるので，それらの収集を行う．収集したサジェストを用いてウェブページの収集を行う． クエリ・フォーカスにサジェスト s を加えて AND 検索することによって上位 N 件以内に検索されるウェブページ d の集合を $D(s, N)$(ただし，本論文では，$N = 20$ とする) とする． ウェブページの収集では，Google Custom Search API[4] を用いる．また，各ウェブページ d に対して，$d \in D(s, N)$ となるサジェスト s を集めた集合を $S(d)$ として，集めたウェブページの集合を D とする．そして，この D にトピックモデルを適用し，トピックの推定を行い，推定されたトピック分布に従ってサジェストの集約を行う．トピック数を K として，各ウェブページに対して最も確率が高いトピックに割り当てることによって，トピック $z_n (n = 1, \ldots, K)$ のウェブページ集合 $D(z_n)$ を作成する．

各ウェブページは，クエリ・フォーカスに一つのサジェストを加えて AND 検索することによって収集される．そのため，各ウェブページに対して一つ以上のサジェストが対応する．ここで，ウェブページ d にはサジェスト集合 $S(d)$ 中のサジェストが対応する．また，ウェブページ d には，トピック z_n が割り当てられている．すると，トピック z_n に対して割り当てられたウェブページ $d \in D(z_n)$ に対応するサジェスト s を集めることによって，トピック z_n に対してサジェストの集合 $S(z_n)$ が割り当てられる．さらに，トピック z_n におけるサジェスト s の頻度 $f(s, z_n)$ は文書集合 $D(z_n)$ 中におけるサジェスト s の観測回数によって定義される． クエリ・フォーカス「就活」の場合，926 個のサジェストが 50 個のトピックに割り当てられた (図 1.2)．以上のように，検索エンジン・サジェストを用いて収集されたウェブページ集合に対してトピックモデルを適用することによって，検索エンジン・サジェストが集約される．

収集されたサジェストに対して word2vec を適用し，サジェストの分散表現を計算する．そして，得られた分散表現を用いて，同一トピックに集約されているサジェスト同士の類似度を測定することによって，検索エンジン・サジェストの話題をより精密に集約する．この方式によって，語 w_a および w_b の分散表現 v_{w_a} および v_{w_b} を求めた後，分散表現であるベクトル間の余弦類似度を求め，これを分散表現間の類似度と定義する．検索エンジン・サジェストの分散表現を求めるために，分散表現訓練用コーパスを用意する．検索エンジン・サジェストを含む標準的な日本語単語の典型的な用例コーパスとして，日本語 Wikipedia の全ページテキストを用いる．さらに，各クエリ・フォーカスに対して収集された検索エンジン・サジェスト特有の特性を反映した分散表現を得るために，各クエリ・フォーカスに対して収集されたウェブページを加え，これら二種類のテキストデータの混合集合を分散表現訓練用コーパスとして，検索エンジン・サジェストの分散表現を求める．また，ベースラインとして，日本語 Wikipedia の全ページテキストのみを用いて分散表現を訓練した場合との比較を行う．

クエリ・フォーカス「就活」の場合について，トピックモデル推定結果における

[4] `https://cse.google.com/cse/`

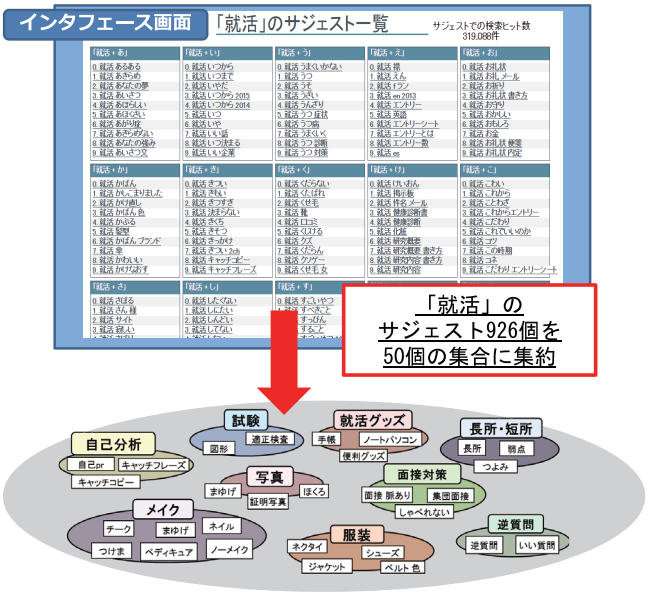


図 1.2: トピックモデルを用いた検索エンジン・サジェストの集約 (クエリ・フォーカス: 「就活」)

トピックのうち,「グループディスカッション，グループワーク，プレゼンテーション，プレゼン」等の検索エンジン・サジェストが割り当てられたトピックを対象として，一部の検索エンジン・サジェスト間で分散表現の類似度を測定した結果を図 1.3 に示す．ここで，分散表現間の類似度の下限値を 0.6 に設定した場合を想定すると,「グループディスカッション，グループワーク，プレゼンテーション，プレゼン」の四つのサジェストのうちの任意の二組のうち，類似度下限値の条件を満たすサジェスト組は,「グループディスカッション，グループワーク」および「プレゼンテーション，プレゼン」の二組のみとなることが分かる．主観評価においても，この二組のみが類似関係にあると言えるため，分散表現の類似度を用いることによって，主観評価の結果と一致する類似度が測定できていることが分かる．

分散表現を用いたサジェスト間類似度の評価を行うために,「就活」および「結婚」の各クエリ・フォーカスを対象として，各トピック中の検索エンジン・サジェストに対して，人手で類似サジェストをまとめることにより,「検索エンジン・サジェストの参照用小分類」を作成した．二つのサジェストの組に対して，(1) サジェスト間の分散表現類似度の下限値，(2) あるサジェストに対して，分散表現類似度の降順に他のサジェストを順位付けした場合の，(a) 全サジェスト中の順位の上限値，(b) トピック内のサジェスト中の順位の上限値，の三つの素性を設定する．そして，三つの素性のあらゆる可能な組み合わせに対して,「検索エンジン・サジェストの参照用小分類」中のサジェスト組の再現率・適合率をプロットする．ここで，分散表現訓練用コーパスとして，(i) Wikipeida のみを用いた場合，(ii) Wikipedia および各クエリ・フォーカスに対して収集されたウェブページの混合集合を用い

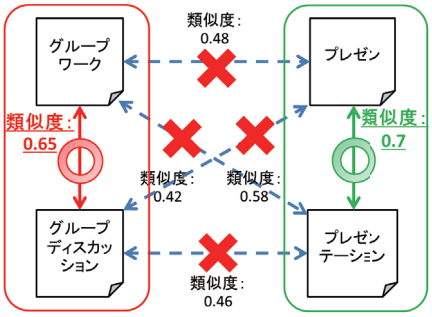


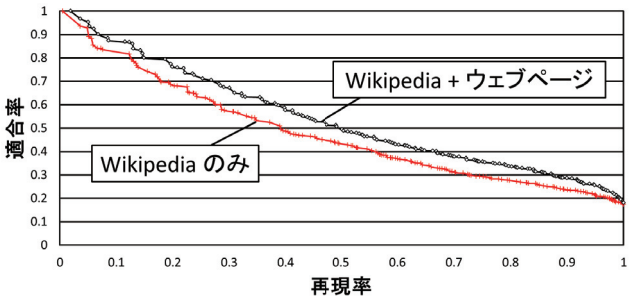
図 1.3: 分散表現を用いたサジェスト間の類似度の測定例 (クエリ・フォーカス:「就活」, 類似度の下限値: 0.6 の場合)

た場合，の間でこのプロットを比較した結果を図 1.4 に示す．この結果から，クエリ・フォーカス「就活」および「結婚」の両方において，分散表現訓練用コーパスとして (ii) を用いた場合の効果を示すことができた．

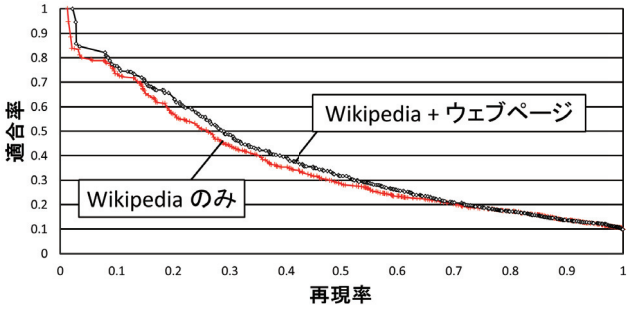
日中間比較対照分析

本項目では，日中文化間差異の発見を支援するために，前章の集約結果に対して，粗い話題の粒度 (LDA の集約結果) と細かい話題の粒度 (分散表現による LDA 集約結果の細分化) の二種類の粒度のもとで，日本語と中国語二言語間で比較対照分析を行う．まず，LDA トピックモデルによる一次集約結果に対して，日中文化間差異の分析を行う．具体的には，クエリ・フォーカス「就活」,「結婚」において，前章で収集された検索エンジン・サジェストとウェブページに LDA トピックモデルを適用した後の集約結果を人手で分析し，類似の内容の質問・回答組をまとめることにより，内容の種類数を集計する．その後，それらの内容を日中間で対応付けることにより，(1) 日中両言語で観測した内容，(2) 日本語側でのみ観測した内容，(3) 中国語側でのみ観測した内容，に分類する.「就活」における分析結果を図 1.5，図 1.6，および，図 1.7 に示す．日本語側のトピック総数と中国語側のトピック総数はそれぞれ 50，60 となっている．まず，日本語側，中国語側のノイズトピック (日本語側: 17 個，中国語側: 19 個) を除外した後，日本語側でのみ観測したトピックの数は 20 個，中国語側でのみ観測したトピックの数は 27 個となった．そして，日中両言語共通のトピックにおいて，日本語側 13 個のトピックと中国語側 14 個のトピックは多対多の対応関係で，人手で分析することによって，9 組の対応組となった．

具体的なトピックの内容について，日中両言語で観測したトピックとして「外国語スキル」,「長所と短所」等を図 1.5 で示す．また，日本語側でのみ観測したトピックとして,「グループ面接などにおいて評価されるところ」,「適性検査と能力検査」,「学校推薦」等があり，それらを図 1.6 で示す．一方，中国語でのみ観測したトピック「エンジニア向けの面接」,「面接者を評価する方法」,「公務員の就活」



(a) クエリ・フォーカス: 「就活」



(b) クエリ・フォーカス: 「結婚」

図 1.4: 検索エンジン・サジェストの類似度の評価結果

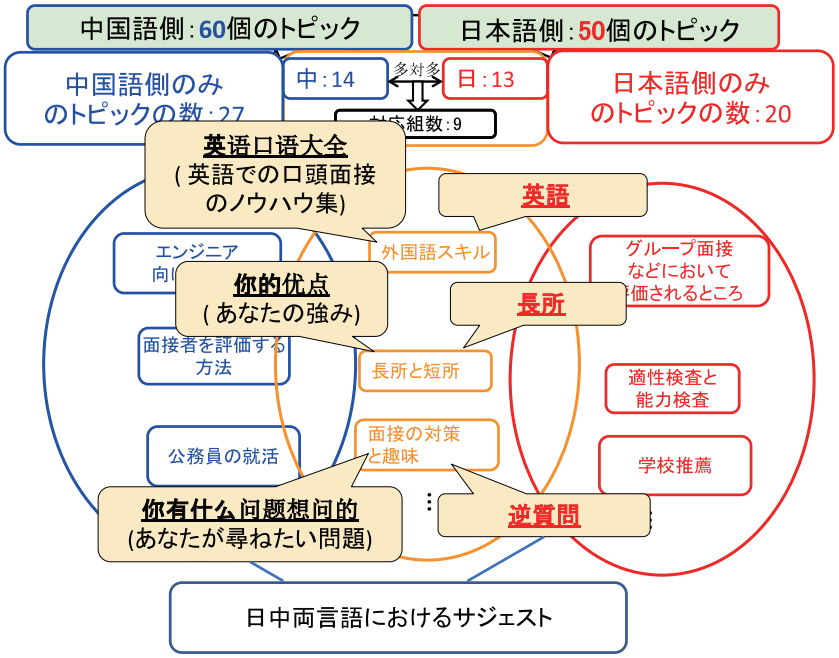


図 1.5: トピックの集約結果の日中間比較対照分析 (1) 日中共通のトピック

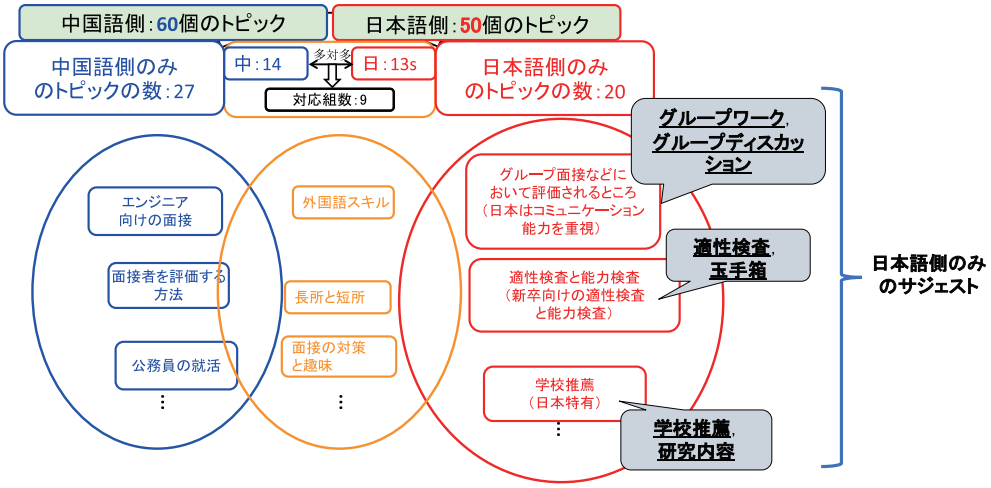


図 1.6: トピックの集約結果の日中間比較対照分析 (2) 日本語側のみのトピック

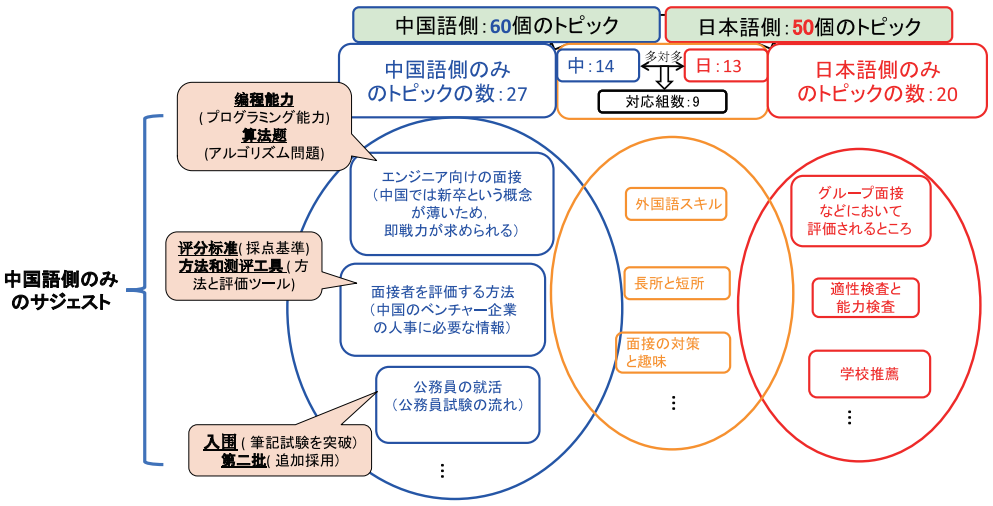


図 1.7: トピックの集約結果の日中間比較対照分析 (3) 中国語側のみのトピック

を図 1.7 で示す．以上の結果から見ると，日本では新卒採用に関心が集まるのに対して，中国の採用活動では即戦力を重視するという文化間差異が明らかになった．

そして，一次集約結果のうち，(3) 日中両側で観測した内容に対して，word2vec から得た分散表現を用いて LDA のトピックの内容を細分化し，人手で類似の内容の質問・回答組をまとめることにより，内容の種類数を集計し，日中間で対応付けることにより，(1) 日中両言語で観測した小分類，(2) 日本語側でのみ観測した小分類，(3) 中国語側でのみ観測した小分類，に分類する．

以上の手順により，LDA トピックモデル，および，word2vec を用いて訓練した分散表現を併用することによって，粗い粒度 (トピックモデル) と細かい粒度 (分散表現) での日中間差異分析を行った．具体的には，日中共通のトピック「面接の対策と趣味」において，日中両側で観測された小分類として,「趣味について」,「逆質問について」等があった．日中共通のトピック「面接の対策と趣味」において，日本語側でのみ観測された小分類として,「集団面接において評価されるところ」,「好印象を与えるための喋り方」等があった．日中共通のトピック「面接の対策と趣味」において，中国語側でのみ観測された小分類として,「転職の理由」,「キャリアパス」等があった．

本論文の構成

本論文は，以下の各章から構成される．第 1 章では，全体の序論として，研究の背景について述べる．第 2 章では，ウェブ情報の集約の関連研究と本論文の位置付けについて述べる．第 3 章では，分散表現とトピックモデルの併用によるサジェストの話題をより精密に集約する手法について述べる．第 4 章では，ウェブ情報の集約における日中比較の関連研究と本論文の位置付けについて述べる．次に，第 5 章において，第 3 章で得た集約結果に対して，日中間比較対照分析について述べる．最後に，第 6 章にて，本論文の結論を述べる．付録 A では，質問解答事例およびウェブからの日中ノウハウ知識の収集および言語間対照分析について述べる．

第2章　ウェブ情報の集約に関する研究(単言語)と本論文の位置付け

2.1　本論文の位置付け

情報集約に関する研究の研究データを集める時に，情報源として執筆者視点でデータを集める，或いは検索者視点でデータを集める，の二種類の方法 (図 2.1) がある．

1. ウェブ執筆者視点でデータを収集する場合，まずウェブ執筆者を選定する．そして，選定されたウェブ執筆者が執筆した文書を収集する．この方法では何百人，何千人の文書を収集できる．

2. ウェブ検索者視点でデータを収集する場合，特定の検索者ではなく，インターネット上にいるウェブ検索者全員を対象として，何千万，何億人の検索ログに基づいて作られた情報要求観点 (＝検索エンジン・サジェスト) を収集する．

ウェブ情報の集約の関連研究 (単言語) と本論文の位置付けに関する詳細は，表 2.1 に記述されている．まず，2.2 節では，ウェブ執筆者に着目した関連研究 [1, 16] について述べる．ウェブ執筆者に着目したウェブ情報を集約する研究の代表例として，文献 [16] が挙げられる．文献 [16] では，ウェブ執筆者であるブロガーに着目し，より効率的な俯瞰を実現するために，ブログ記事単位ではなく，トピックモデルによる同一の興味を持つブロガーコミュニティを生成する手法を提案した．そして，2.3 節では，ウェブ検索者に着目した関連研究 [8, 10] について述べる．ウェブ検索者に着目したウェブ情報を集約する研究の代表例として，文献 [8] が挙げられる．文献 [8] では，ウェブ検索者関心動向を俯瞰するために，ウェブ検索者の関心事項である検索エンジン・サジェストを集約する手法，および，ウェブ検索結果における多様な話題を含むウェブページを選択的に提示手法を提案した．

文献 [8] では，文書集合に LDA トピックモデルを適用し，話題が類似するサジェストをまとめて俯瞰する手法を提案したが，トピックモデルを用いた話題集約の場合には，集約の粒度が相対的に粗いという課題が残る．この課題を解決するた

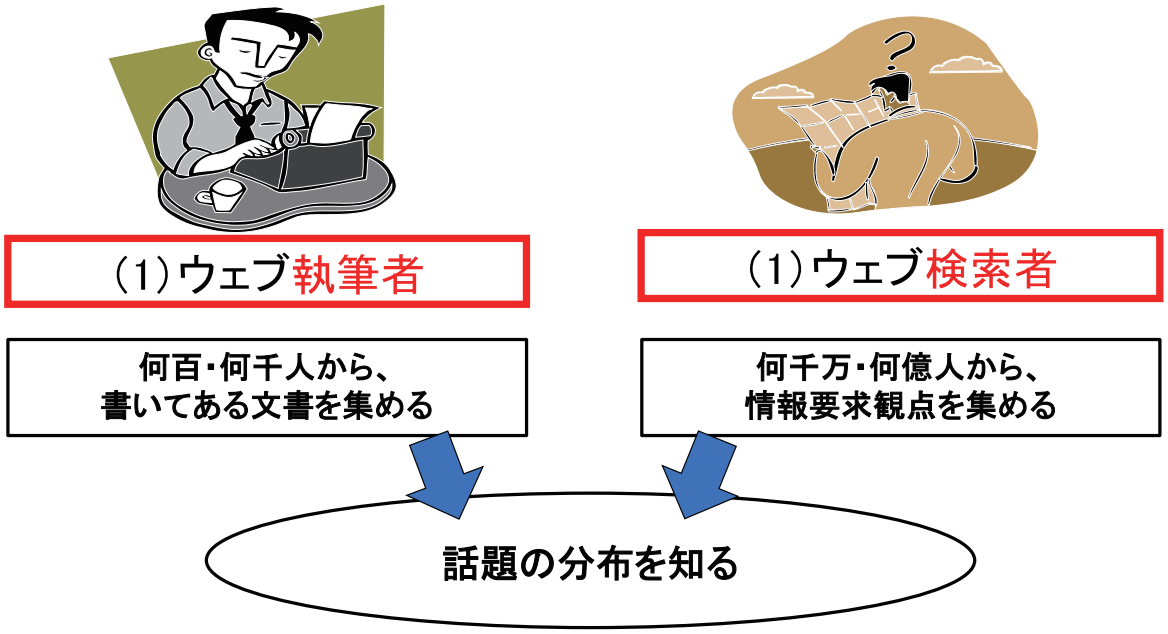


図 2.1: 執筆者視点と検索者視点の比較

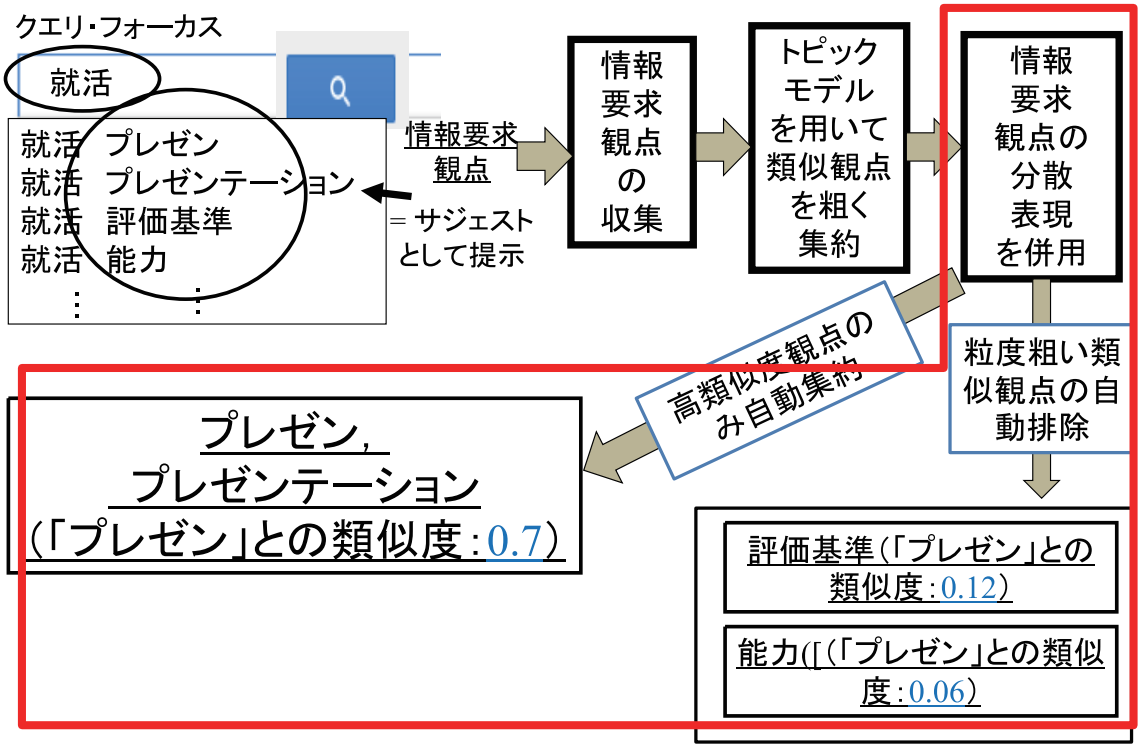
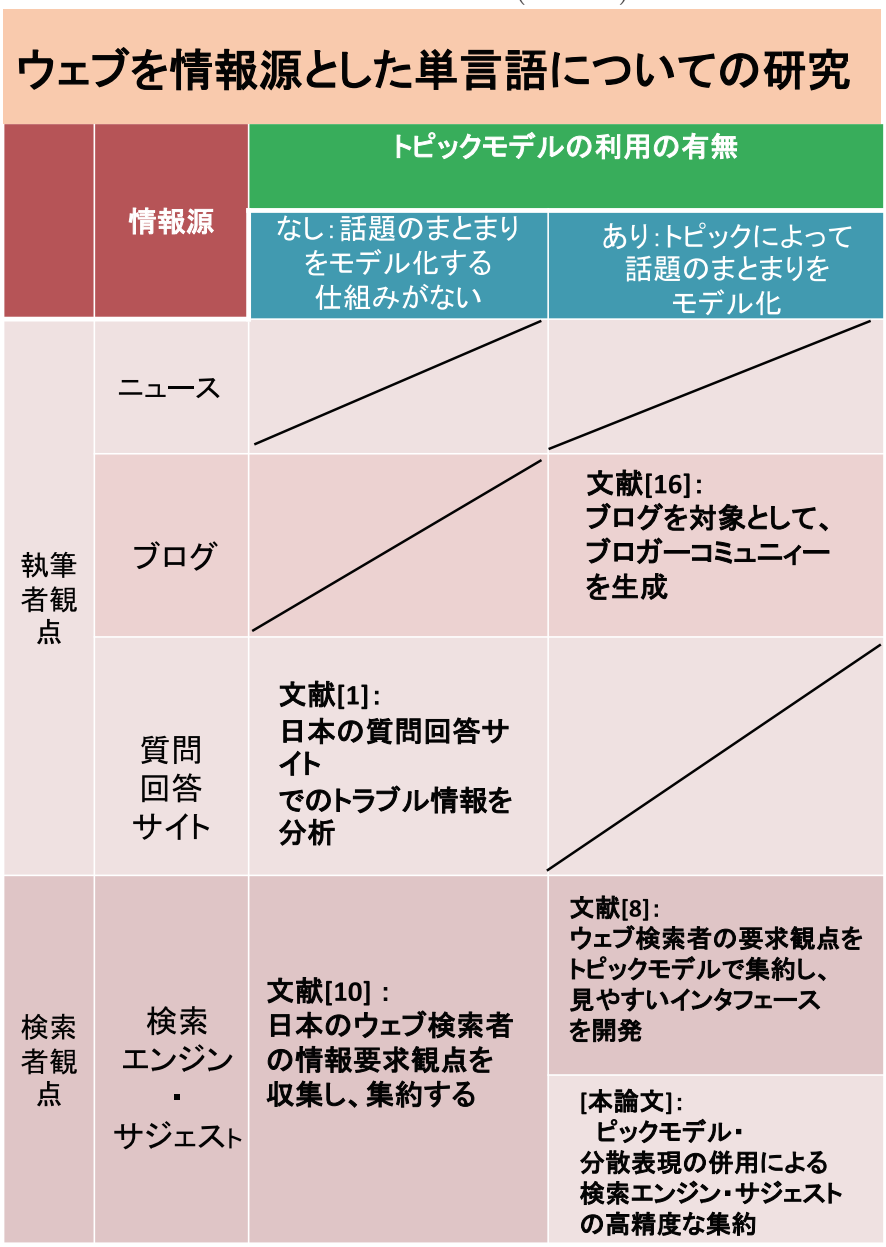


図 2.2: トピック内の小分類単位でのサジェストの集約

表 2.1: ウェブ情報の集約に関する研究 (単言語) と本論文の位置付け



めに，本論文では，分散表現を用いて，トピック内の小分類を同定する手法を提案する．以上の流れを図 2.2 に示す．

図 2.2 の内，本論文と文献 [8] の共通部分である「トピック単位でのサジェストの集約」を以下のステップ 1 に示す．また，本論文の提案手法である「トピック内の小分類単位でのサジェストの集約」を以下のステップ 2 に示す．

ステップ 1 「トピック単位でのサジェストの集約 (文献 [8] と同じ手法)」

1. まず，各クエリ・フォーカスに対して，日本語の検索エンジン・サジェスト，および，中国語の検索エンジン・サジェストを収集する．日本語側においては，五十音などを用いて Google 検索エンジンからサジェストを収集した．中国語側においては，pinyin を用いて Baidu 検索エンジンからサジェストを収集した．
2. 次に，収集された検索エンジン・サジェストを用いて,「クエリ・フォーカス + 検索エンジン・サジェスト」の AND 検索によって検索される上位数十件のウェブページを収集する．
3. 最後に，収集されたウェブページ集合に対して，トピックモデルを適用することで数十個のトピックへ集約する．各ウェブページが検索語であるサジェストの情報を持っているため，ウェブページにトピックを割り当てることで，検索エンジン・サジェストも自動的に数十個のトピックに分類される．

ステップ 2 「トピック内の小分類単位でのサジェストの集約 (図 2.2 の赤枠の部分)」

1. Wikipedia の全件データにステップ 1 で検索エンジン・サジェストによって収集したウェブページを加えて，word2vec を用いてサジェストごとの分散表現を計算する．
2. ステップ 1 の集約結果の中の各トピックに対して，得られた分散表現を用いて検索エンジン・サジェスト同士のコサイン類似度を測定する．そして，低類似度のサジェストを排除し，高類似度のサジェスト同士のみを同じ小分類へ集約する．

2.2 ウェブ執筆者に着目した研究

以下では，ウェブ執筆者に着目したウェブ情報を集約する研究の例として，文献 [16] について述べる．文献 [16] では，日本のブロガーユーザーを対象として，トピックモデルによる同一の興味を持つブロガーコミュニティの生成と拡張の手法を提案した．

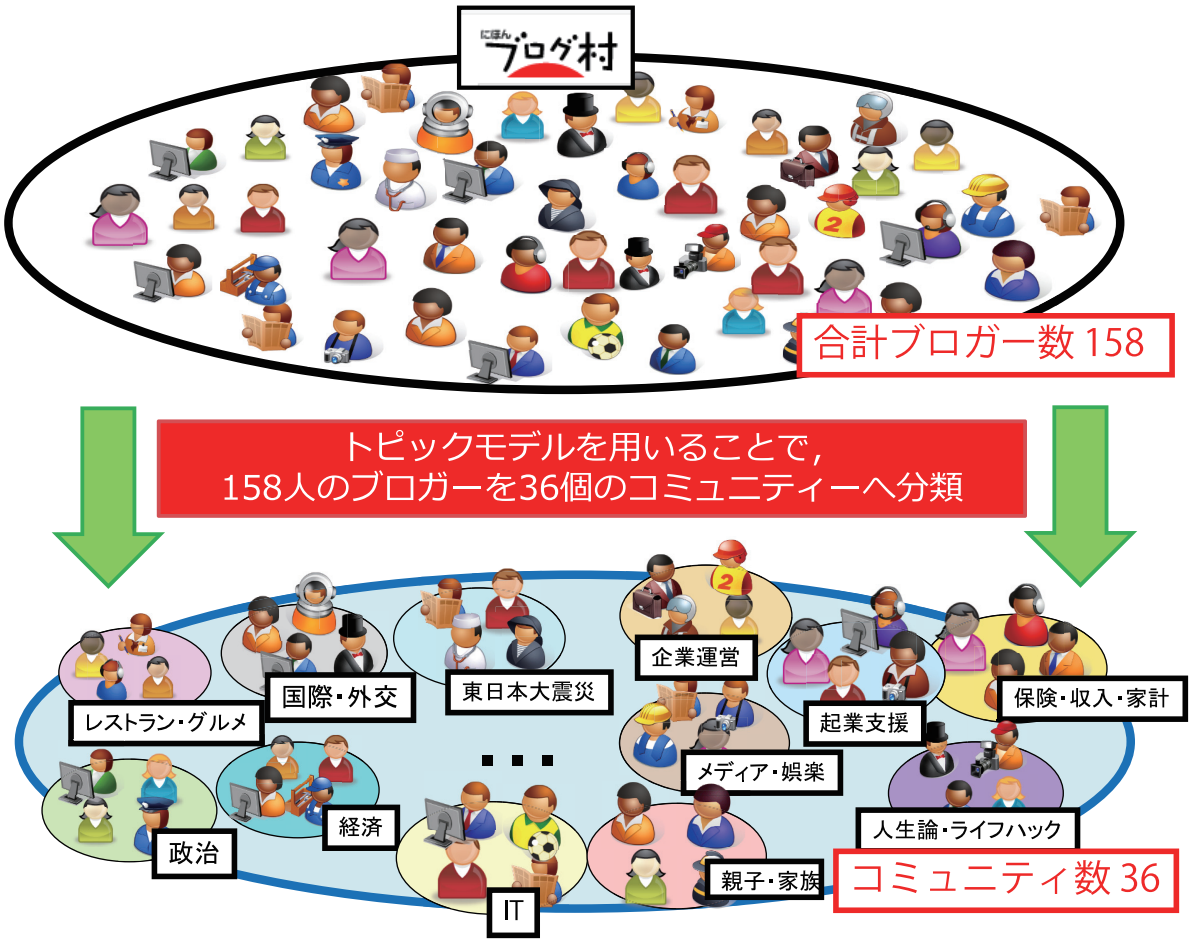


図 2.3: ブロガーからコミュニティを生成する例

2.2.1 トピックモデルを用いたブロガー・コミュニティの収集と俯瞰

文献 [16] におけるブロガー・コミュニティ生成過程について以下に述べる.

1. まず,「にほんブログ村」からブロガーとブログ記事を収集する.「にほんブログ村」は，日本国内最大のブロガーコミュニティであり，各ブロガーは複数の「カテゴリー」および「サブカテゴリー」を持っている．文献 [16] では,「企業」,「ベンチャー」,「経営」,「経済」のカテゴリーから 158 人のブロガーを選定し，各ブロガーのブログ記事を収集した.

2. 次に，収集されたブログ記事集合に対して，LDA トピックモデルを適用し，各ブログ記事におけるトピックの確率分布を推定する．確率最大のトピックをそのブログ記事のトピックとする.

3. そして，人手によって各トピックのまとまり度合を評価し，各トピックに分

類された確率上位 5 記事のうち，3 記事以上が同じ話題であれば，有効なトピックとして判定する．

4. 最終的に，図 2.3 に示すように,「にほんブログ村」から選定された 158 人のブロガーが，36 個のコミュニティへ分類された．

また，文献 [16] では,「にほんブログ村」における課題として,「にほんブログ村」に未登録の外部ブロガーの自動収集・自動登録等の機能が提供されていないという問題を解決するために，コミュニティの自動生成だけでなく,「にほんブログ村」に登録されていないブロガーをコミュニティへ追加することによるコミュニティの自動拡張を行う方式を提案した．その流れを図 2.4 に示す．

1. まず，種コミュニティを生成するために,「にほんブログ村」からブロガーの選定および各ブロガーのブログ記事の収集を行う．収集されたブログ記事集合に LDA トピックモデルを適用することにより，種コミュニティを生成する (図 2.4(a) ブロガー・コミュニティの生成)．

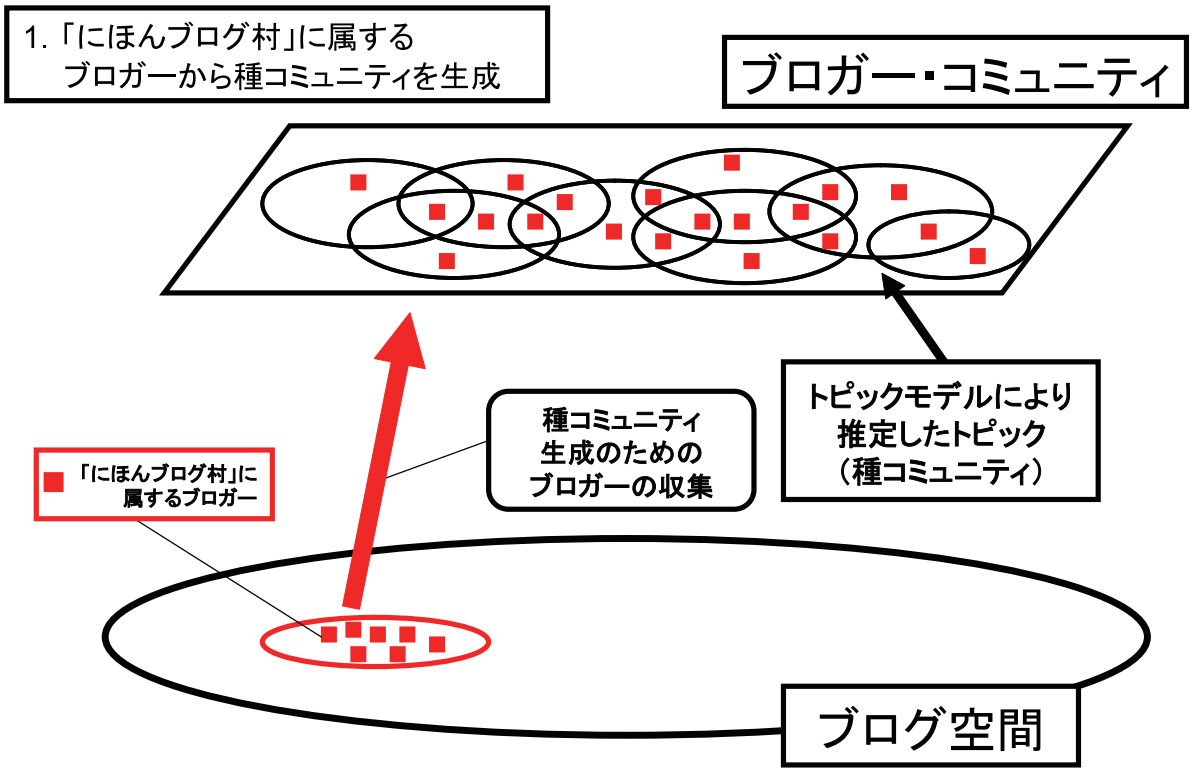
2. 種コミュニティを生成した後，新たに収集したブロガーを種コミュニティへ追加することでコミュニティの拡張を行う．具体的には,「にほんブログ村」以外を情報源として，種コミュニティとの関連の強いブロガー，および，各ブロガーのブログ記事を収集し，種コミュニティを生成時に推定したトピックモデルを用いてブログ記事のトピックの割り当てを行う．以上の手順により,「にほんブログ村」から収集したわずか 158 人のブロガーの記事により生成した種コミュニティを約 3,500 人規模のブロガーコミュニティへ拡張した (図 2.4(b) ブロガー・コミュニティの拡張)．

文献 [16] の手法により,「にほんブログ村」の 4 つのカテゴリー「企業」,「ベンチャー」,「経営」,「経済」から収集した 158 人のブロガーは，表 2.2 の 36 個のコミュニティへ集約された．

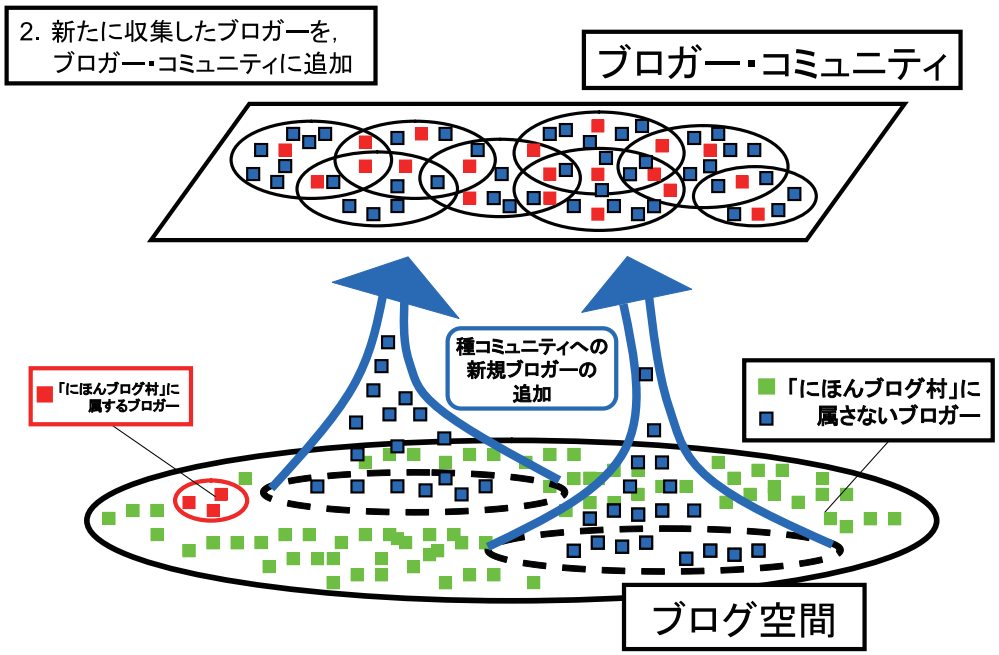
2.3 ウェブ検索者に着目した研究

以下では，ウェブ検索者に着目したウェブ情報を集約する研究の例として，文献 [8] について説明する．文献 [8] では，ウェブ検索者の関心事項 (検索エンジン・サジェスト) を対象として，検索エンジン・サジェストを集約する手法，および，ウェブ検索結果における多様な話題を含むウェブページを選択的に提示する手法を提案した．

図 2.5 に示すように，研究データの情報源として，執筆者視点でデータを集める方と比べて，検索者視点でデータを集める方がいくつかの利点がある．



(a) ブロガー・コミュニティの生成



(b) ブロガー・コミュニティの拡張

図 2.4: ブロガー・コミュニティの生成および拡張 (文献 [16] より引用)

表 2.2: ブロガー・コミュニティにおける「にほんブログ村」のブロガー数の分布 (文献 [16] より引用)

ID	ブロガー・コミュニティの話題	「にほんブログ村」におけるカテゴリー			
		企業	ベンチャー	経営	経済
1	政治	0	0	1	9
2	東日本大震災	1	0	0	6
3	親子・家族	0	4	1	0
4	経済	0	0	1	13
5	保険・収入・家計	4	1	2	5
6	企業運営	1	1	4	2
7	起業支援	0	2	0	0
8	IT	0	3	1	3
9	メディア・娯楽	0	0	2	1
10	国際・外交	1	3	0	4
11	人生論・ライフハック	0	6	11	3
12	レストラン・グルメ	7	1	2	1
13	法律・制度	3	3	2	2
14	事件・時事・社会問題	2	0	0	4
15	スポーツ	1	1	4	1
16	社会学・思想	0	0	0	4
17	建築・住宅	5	1	2	0
18	就職・転職	1	3	1	0
19	株式市場	0	0	0	9
20	ネット通販	0	3	1	1
21	工業	7	0	1	1
22	勉強・スキルアップ	1	1	0	0
23	電気・通信	0	0	0	1
24	起業家向けの勉強会	0	4	3	0
25	農業	14	0	1	0
26	貿易	2	0	0	5
27	インテリア	0	2	1	0
28	仕事論・人生論	0	2	7	1
29	出版	1	2	1	0
30	経営戦略	0	0	3	0
31	財政	0	0	0	2
32	セミナー・勉強会	0	2	1	0
33	融資・金融・経営	1	1	1	0
34	物流業界	3	0	1	0
35	ボディージュエリー	0	0	1	0
36	接客業	0	0	1	0
	合計 (延べ数)	55	46	57	78
	合計 (異なり数)	41	31	39	47

ウェブ検索者

特徴1： 書く人より、検索する人の方が圧倒的に多い

特徴2： 時事的話題など、急に変化する話題の場合、ウェブ検索者による変化への追随が速い

特徴3： 時間的変遷が緩やかな文化・慣習に関する話題の場合も、ウェブ検索者の関心が高く、有用性の高そうな情報の一覧が容易に得られる

図 2.5: ウェブ検索者視点の特徴

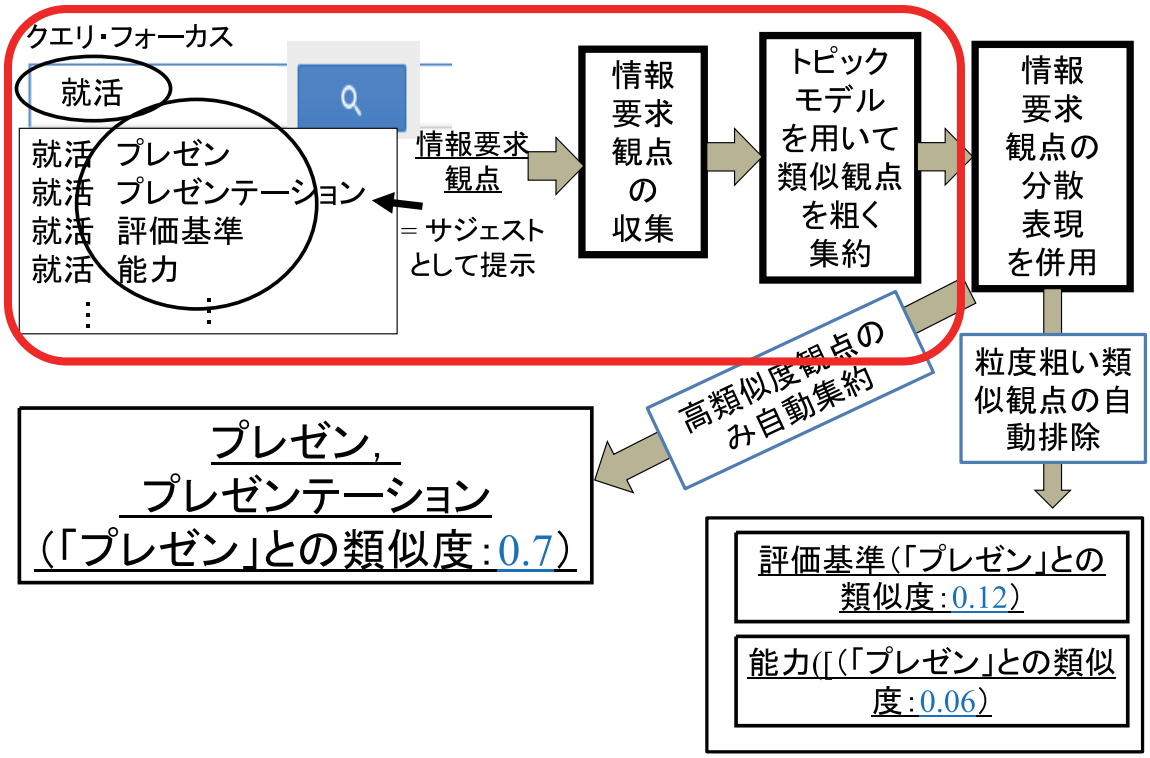


図 2.6: トピック単位でのサジェストの集約

- 一つ目の利点として，書く人より，検索する人の方が圧倒的に多い．つまり，執筆者視点でデータを収集する場合，インターネット上の一部のデータしか収集されていないことに対して，検索者視点でデータを収集する場合，インターネット上のすべてのデータが収集対象となっているため，より多くの情報を収集できる．

- 二つ目の利点は，変遷が急激な時事的話題に関して，ウェブ検索者の関心動向は急に変化する話題への追随が速いため，ウェブ検索者視点でデータを収集する方がより迅速に人々の関心の高い話題を把握できる．

- 三つ目の利点は，時間的変遷が緩やかな文化や慣習の話題でも，人の関心が集まっていれば，ウェブ検索者視点でデータを収集する方が容易に情報を得られる．

2.3.1 検索エンジン・サジェストおよびトピックモデルを用いたウェブ検索結果の集約

以下では，図 2.6 の赤枠部分に沿って，トピックモデルを用いて，トピック単位で検索エンジン・サジェストを集約する流れを示す.

1. まず，各クエリ・フォーカスに対して，日本語の検索エンジン・サジェスト，および，中国語の検索エンジン・サジェストを収集する．日本語側においては，五十音などを用いて Google 検索エンジンからサジェストを収集した．中国語側においては，pinyin を用いて Baidu 検索エンジンからサジェストを収集した.

2. 次に，収集された検索エンジン・サジェストを用いて,「クエリ・フォーカス + 検索エンジン・サジェスト」の AND 検索によって検索される上位数十件のウェブページを収集する.

3. 最後に，収集されたウェブページ集合に対して，トピックモデルを適用することで数十個のトピックへ集約する．各ウェブページが検索語であるサジェストの情報を持っているため，ウェブページにトピックを割り当てることで，検索エンジン・サジェストも自動的に数十個のトピックに分類された.

文献 [8] における日本語サジェストの集約の一例として，図 2.7 に示すように，クエリ・フォーカス「就活」の場合は，まず，五十音を用いて 934 個のサジェストを収集する．次に,「就活」+ サジェストの AND 検索で，検索エンジンからウェブページを収集する．そして，収集されたウェブページ集合に対して，トピックモデルを適用し，数十個のトピックへ集約する．最後に，ウェブページを分類することで，検索語となった「就活」に関連する 934 個のサジェストも 50 個のトピックへ自動的に分類された.

文献 [8] におけるクエリ・フォーカス「就活」の集約結果の具体例を表 2.3 に示す.「就活」に関するサジェストのクラスタリングを行った後，人手で各トピックに分類されたサジェストおよびウェブページの話題を分析し，トピックごとにラベルを付与する．例えば，トピック「髪型」には，“ヘアスタイル 女”，“写真 髪型”，“ロングヘア” などの，好印象を与えるためのヘアスタイルに関するサジェストが含まれていることが分かる.

文献 [8] におけるクエリ・フォーカス「結婚」の集約結果の具体例を表 2.4 に示す．クエリ・フォーカス「就活」と同じように，クエリ・フォーカス「結婚」に関するサジェストのクラスタリングを行った後，人手で各トピックに分類されたサジェストおよびウェブページの話題を分析し，トピックごとにラベルを付与する．例として，トピック「条件，決めて」には，“容姿”，“男性 条件”，“価値観” など

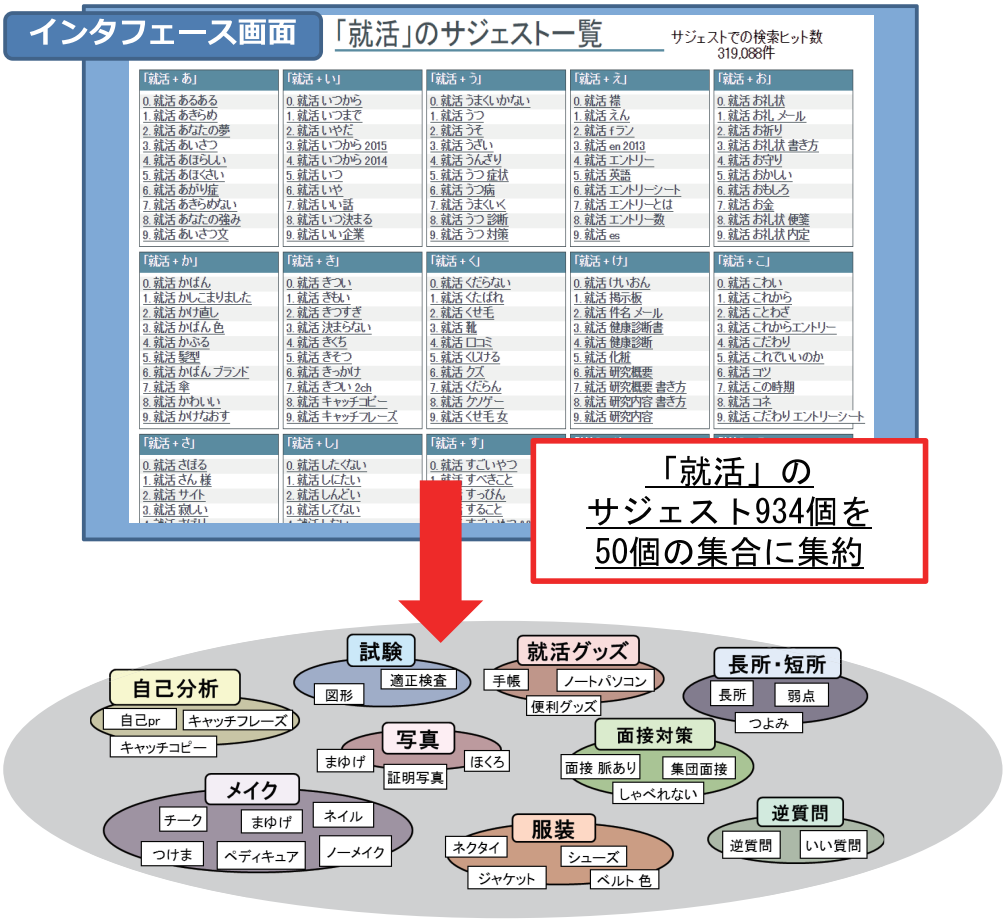


図 2.7: 検索エンジン・サジェストの集約 (クエリ・フォーカス: 「就活」)(文献 [8] より引用)

表 2.3: 提案手法による検索エンジン・サジェストの集約結果の例 (クエリ・フォーカス: 就活)(文献 [8] より引用)

クエリ・フォーカス	人手によりトピックに付与したラベル	トピックに割り当てられたサジェスト (各トピック 10 サジェストを抜粋)
就活	髪型	“ヘアスタイル 女”, “くせ毛 女”, “写真 髪型”, まとめ髪, おだんご, 襟足, ロングヘア, ゆるいパーマ, 美容院, シュシュ
	身に着けるもの	ネクタイ, シューズ, “ベルト 色”, かばん, ピーコート, シャツ , “パンプス おすすめ”, “グレー スーツ”, “ジャケット ボタン”, 防寒
	グループディスカッション	グループワークとは, グループディスカッション, “グループディスカッション テーマ”, 評価, グループワーク対策, 評価基準, プレゼン, “プレゼン 資料”, グループワーク, 能力
	自己分析	“長所 真面目”, 長所, 座右の銘, 軸, どうなりたいか, あなたの夢, こだわり, 将来の夢, どんな人, なりたい自分
	恋愛との両立	“恋愛 両立”, ふられた, 恋愛, 寂しい, 脈あり, 結婚, “うまくいかない 彼氏”, “プレゼント 彼女”, わがまま, プレッシャー
	メイク	ノーメイク, ビューラー, チーク, 化粧, つけま, まつエク, ネイル, まゆげ, “証明写真 メイク”, ペディキュア

表 2.4: 提案手法による検索エンジン・サジェストの集約結果の例 (クエリ・フォーカス: 結婚)(文献 [8] より引用)

クエリ・フォーカス	人手によりトピックに付与したラベル	トピックに割り当てられたサジェスト (各トピック 10 サジェストを抜粋)
結婚	お祝い，メッセージ	“友人　スピーチ”， “お祝い　メッセージ”， “文例　電報　友人”， 祝辞，“電報　バルーン”， 一言メッセージ， “ぬいぐるみ　メッセージ”， ぬいぐるみ電報， ビデオメッセージ， “祝電　文例”
	条件，決めて	“妥協　顔”，ルックス， 見極め，美人，容姿，理想， “男性　条件”， “決め手　女性”， 相手，価値観
	求める収入	高望み，条件， “条件　年収”，収入， 平均年収，高望み， ランキング職業， “条件　ランキング”， 年収，求めるもの
	結婚祝い	プレゼント， “ぬいぐるみ　手作り”， 贈り物， “ぬいぐるみ　うさぎ”， 印鑑，プチギフト， 祝い，ペアウォッチ， サプライズ，寄せ書き
	手続き	“入籍　手続き”， 住所変更，“苗字　変更”， 必要書類，パスポート， 住民票，“会社　手続き”， 外国人，名義変更， グリーンカード
	写真	“写真　東京”， 前撮り，写真，ポーズ， “写真　大阪”， “写真　札幌”，写真だけ， ビデオ，和装， ビデオメッセージ

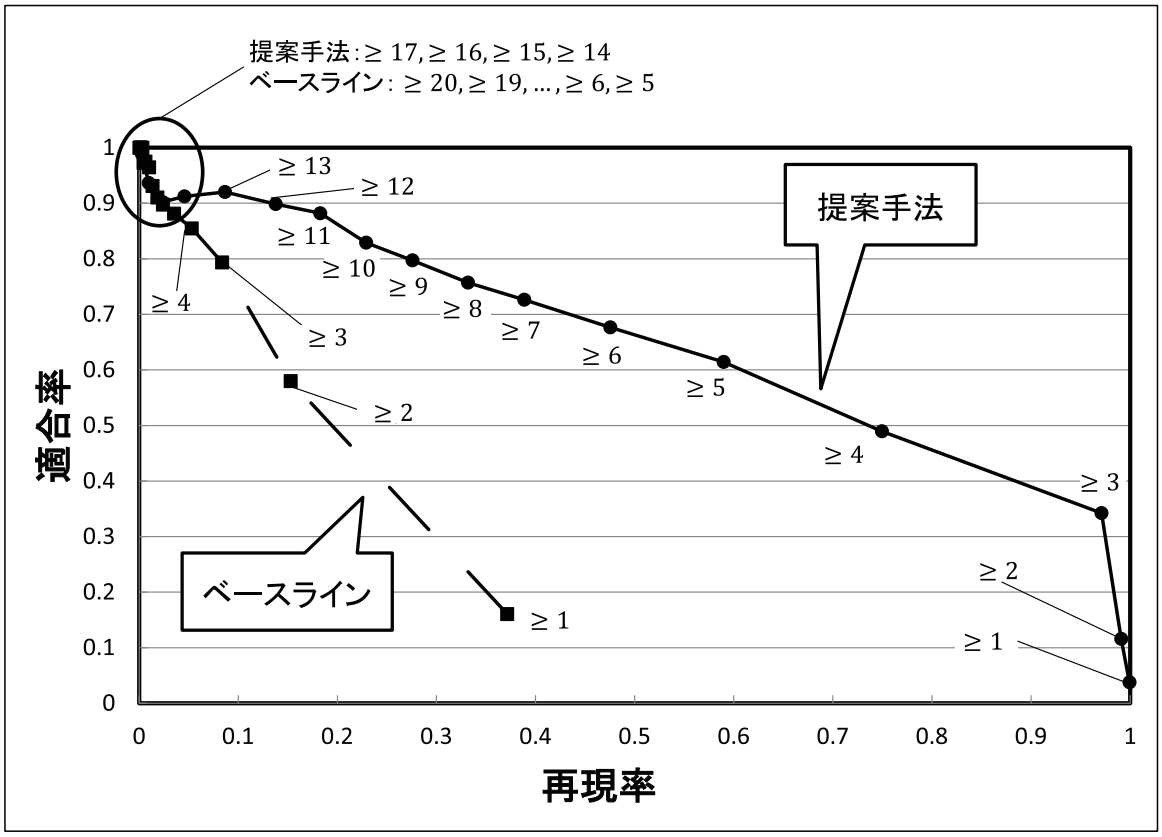


図 2.8: 検索エンジン・サジェストの集約の評価結果 (サジェストの頻度の下限値を変化させた場合)(文献 [8] より引用)

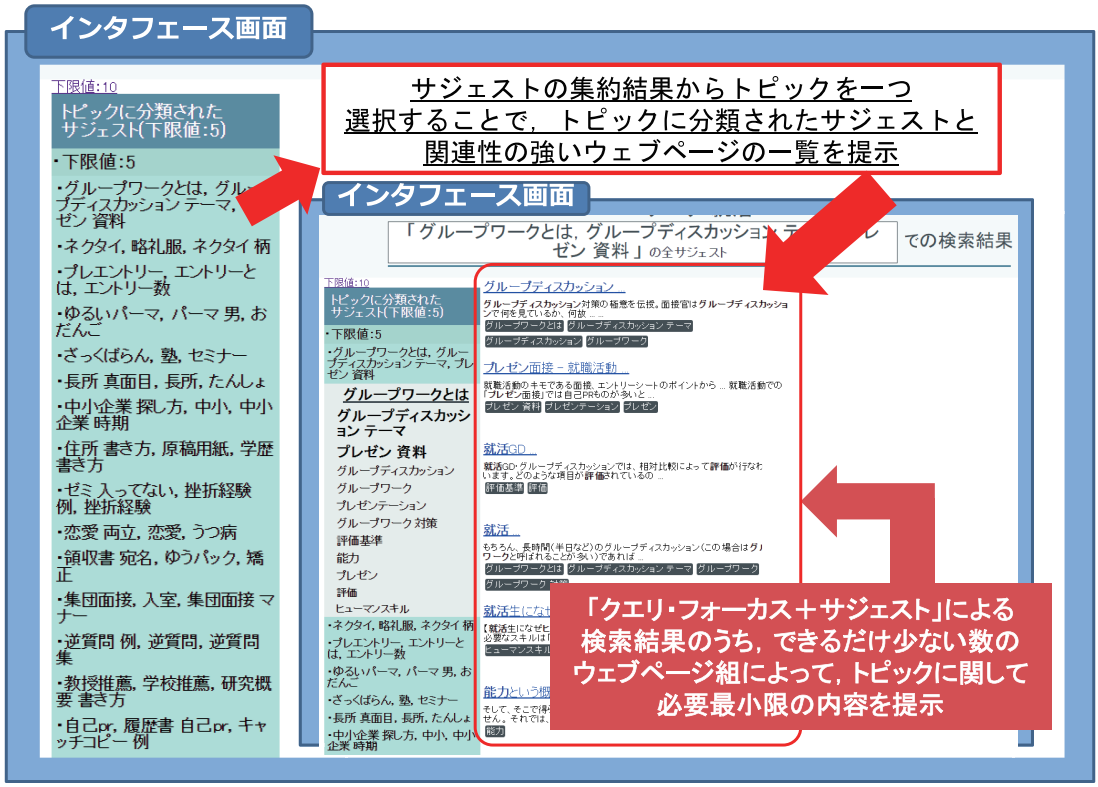


図 2.9: ウェブ検索結果の俯瞰インタフェース画面 (クエリ・フォーカス:「就活」)(文献 [8] より引用)

の，結婚相手を選ぶ時に考える条件に関するサジェストが含まれていることが分かる．

文献 [8] における検索エンジン・サジェストの集約の評価結果を図 2.8 に示す．人手によって作成された参照用サジェスト集合に対して，ベースライン手法の出力における再現率と適合率，および，提案手法の出力における再現率と適合率を算出し，プロットする．この結果において，ベースライン手法より提案手法の方が高い集約精度を達成した．

文献 [8] では，効率良く俯瞰するためのインタフェースを開発した．図 2.9 はインタフェースのキャプチャ画像である．インタフェースの左側は，サジェストの集約結果となるトピックの一覧が表示されている．それぞれのトピックは，その中に分類されたサジェストのうち，関連性の高い上位 3 つで表現されている．また，トピック一つを選ぶと，そのトピックへ分類されたサジェストの一覧も閲覧できる．さらに，各トピックの中に含まれたウェブページの表示に関しては，多様なサジェストを持つウェブページを優先的に提示することによって，できるだけ少ないウェブページ数組で多様な話題を提示している．

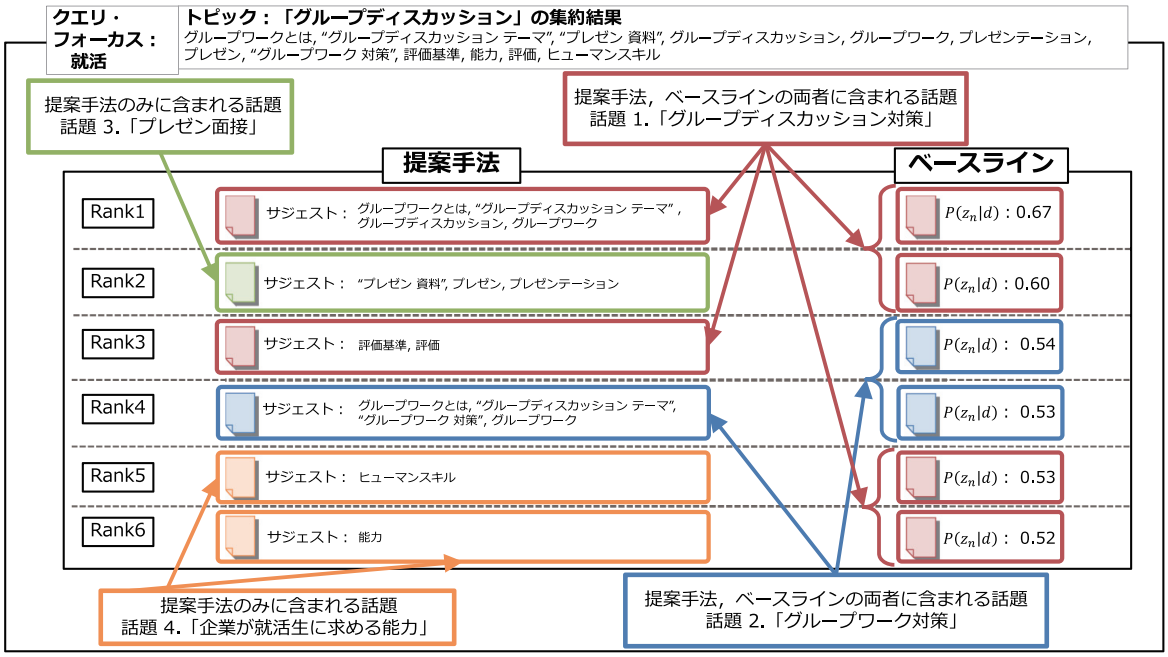


図 2.10: ウェブ検索結果の集約の例 (1) (クエリ・フォーカス: 「就活」, トピック: 「グループディスカッション」)(文献 [8] より引用)

クエリ・フォーカス「就活」のトピック「グループディスカッション」における，文献 [8] の提案手法によるウェブページの集約結果とベースライン手法によるウェブページの集約結果を比較した結果を図 2.10 に示す．ベースライン手法では，単純に各トピックに属するウェブページの確率順でウェブページの選定を行う．これに対して，文献 [8] の提案手法では，ウェブページを選ぶ時に，

1. ウェブページにできるだけ多くのサジェストを含む
2. それぞれのサジェストの順位できるだけ上位である

という考え方でウェブページの選定を行う．両手法によって選定された上位 6 件のウェブページにおいて，ベースライン手法では 2 個の話題が含まれていた．文献 [8] の提案手法では 4 個の話題が含まれていた．具体的に，共通の話題として，話題「グループディスカッション対策」と話題「グループワーク対策」が両方に含まれていた．特に，ベースライン手法における「グループディスカッション対策」のウェブページ数が 4 件となり，総ウェブページ数 (6 件) の半分以上を占めていた．一方，文献 [8] の提案手法における話題「グループディスカッション対策」のウェブページ数と話題「グループディスカッション」のウェブページ数は 2 件，および，1 件となっていた．その他，文献 [8] の提案手法では，話題「プレゼン面接」と話題「企業が就活生に求める能力」が含まれていた．それぞれのウェブページ数は 1 件，および，2 件となっていた．このように，ベースライン手法より，文

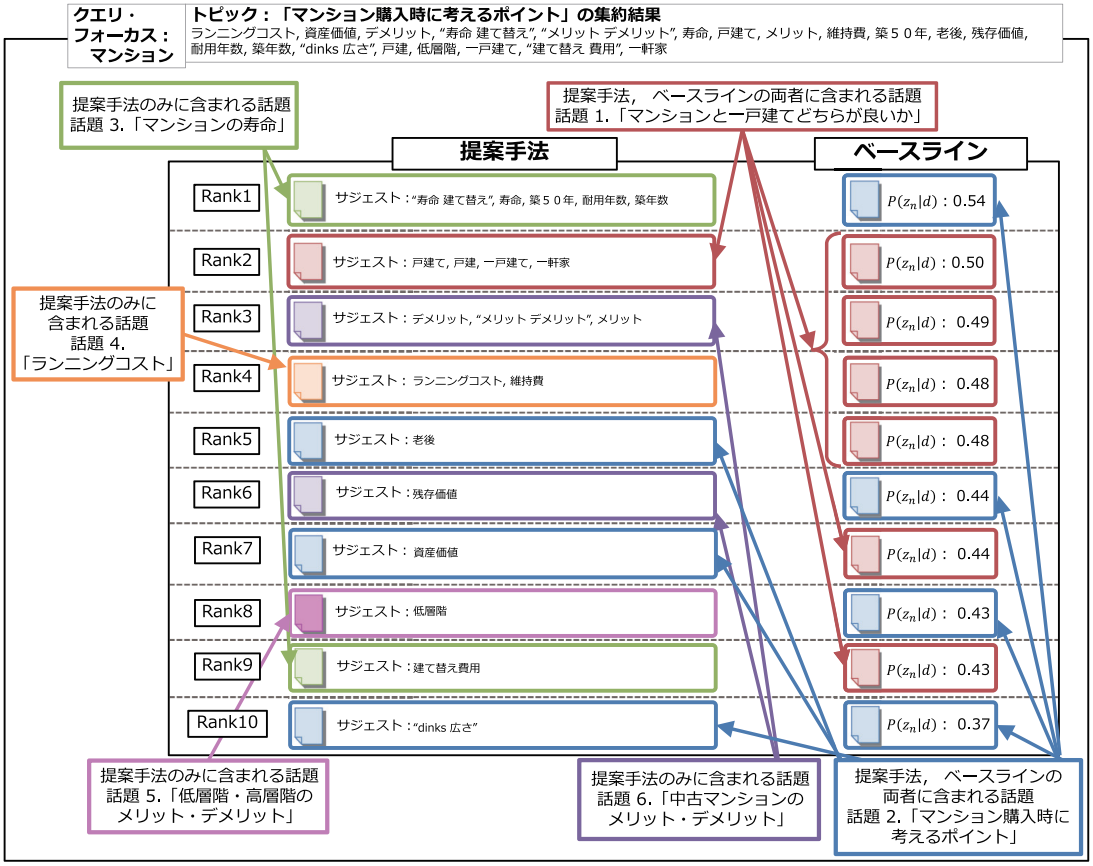


図 2.11: ウェブ検索結果の集約の例 (2) (クエリ・フォーカス: 「マンション」, トピック: 「マンション購入時に考えるポイント」)(文献 [8] より引用)

献 [8] の提案手法の方がウェブページの選定においてより多くの話題を提示できるということが分かる.

図 2.10 と同じように，クエリ・フォーカス「マンション」のトピック「マンションを購入時に考えるポイント」における，文献 [8] の提案手法によるウェブページの集約結果とベースライン手法によるウェブページの集約結果を比較した結果を図 2.11 に示す．両手法によって選定された上位 10 件のウェブページにおいて，ベースライン手法では 2 個の話題しか含まれていなかった．これに対して，文献 [8] の提案手法では 6 個の話題が含まれていた．具体的に，共通の話題として，話題「マンションと一戸建てでどちらが良いか」と話題「マンション購入時に考えるポイント」が両方に含まれていた．特に，ベースライン手法における「マンションと一戸建てでどちらが良いか」のウェブページ数が 6 件となり，総ウェブページの 10 件の半分以上を占めていた．一方，文献 [8] の提案手法における話題「マンションと一戸建てでどちらが良いか」のウェブページ数と話題「マンション購入時に考えるポイント」のウェブページ数は 1 件，および，3 件となっていた．その

他，文献 [8] の提案手法では，話題「マンションの寿命」(2 件)，話題「ランニングコスト」(1 件)，話題「低階層・高階層のメリット・デメリット」(1 件) と話題「中古マンションのメリット・デメリット」(3 件) などが含まれていた．この例においても，ベースライン手法より，文献 [8] の提案手法の方がウェブページの選定においてより多くの話題を提示できるということが分かる．

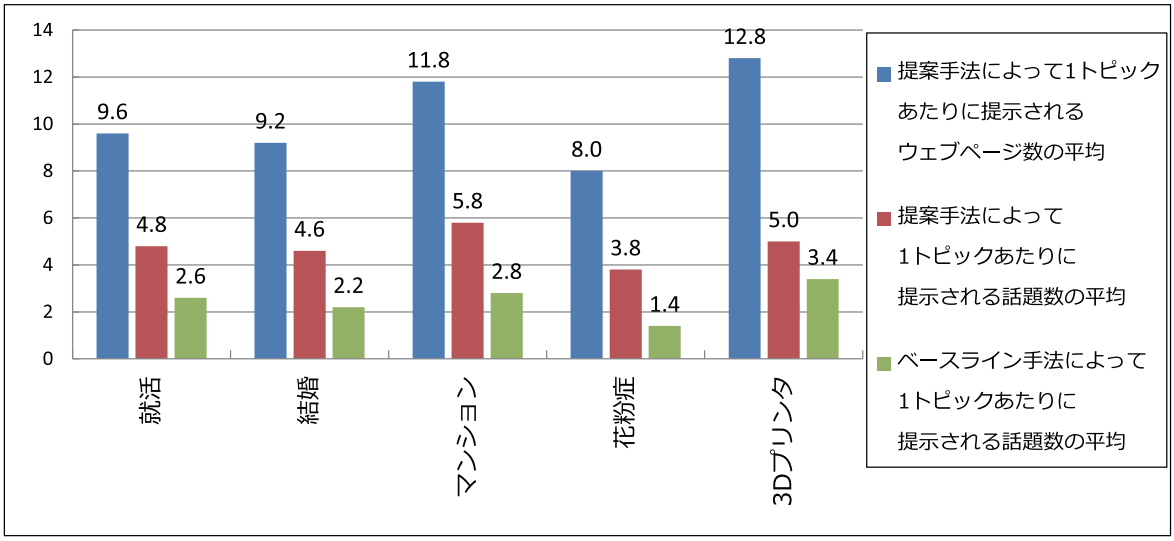
次に，文献 [8] の評価対象となった 5 つのクエリ・フォーカスにおける，提案手法によるウェブ検索結果を集約の評価結果とベースライン手法にによるウェブ検索結果を集約の評価結果を比較した結果を図 2.12 に示す．図 2.12(a) において示すのは，

- 提案手法で提示される 1 トピックあたりのウェブページ数および話題数の平均
- ベースライン手法で提示される 1 トピックあたりのウェブページ数および話題数の平均

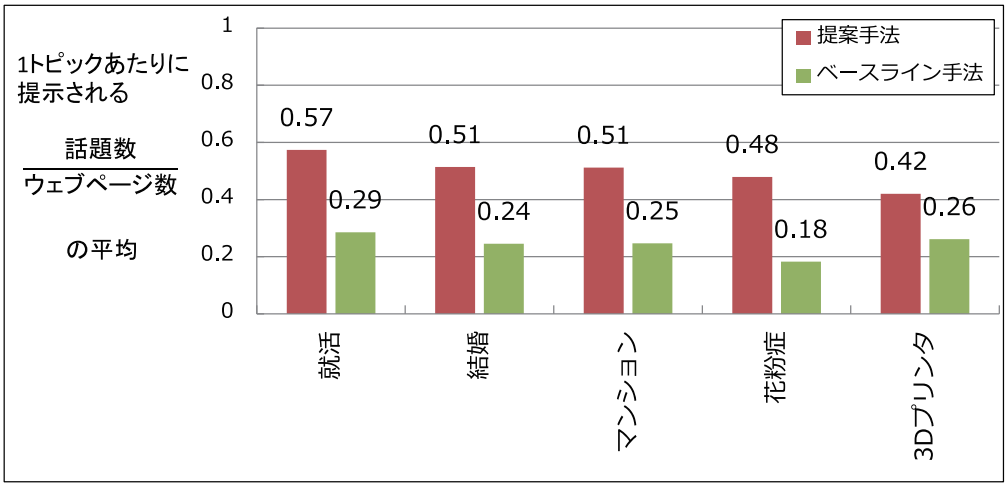
を比較した結果である．図 2.12(b) において示すのは，全トピックに対して，

$$\frac{\text{1 トピックあたりに提示される話題数}}{\text{1 トピックあたりに提示されるウェブページ数}}$$

を計算し，それらの結果の平均を提案手法とベースライン手法の間で比較した結果である．図 2.12(a) と図 2.12(b) の結果から，提案手法の集約結果によって提示される話題の数が，ベースライン手法の集約結果によって提示される話題の数の約 2 倍となり，ベースライン手法と比べて，提案手法の方がより多様な話題を提示できることが分かる．



(a) 1 トピックあたりに提示されるウェブページ数/話題数



(b) (1 トピックあたりに提示される話題数/1 トピックあたりに提示されるウェブページ数) の平均

図 2.12: ウェブ検索結果の集約の評価

第3章　トピックモデルと分散表現の併用による検索エンジン・サジェストの集約

3.1　はじめに

現代社会では，インターネット上に様々な情報が溢れている．ウェブ検索者が欲しい情報を手に入れる手段の一つとしては，Google 等の検索会社が提供している検索エンジンを利用するのが一般的である．ここで，検索会社は，検索行動支援の一環として，検索エンジン・サジェスト・サービスを提供している．このサービスの特徴として，入力された検索語に対して，関連が強い語を検索エンジン・サジェストとして提示する．ここで，本論文では，ウェブ検索者が詳細な情報を得たい対象を**「クエリ・フォーカス」**と呼ぶ．そして，AND 検索の形でクエリ・フォーカスの次に入力され，より詳細な情報を得たい観点を示す語を**「情報要求観点」**と呼ぶ (図 3.1)． 検索エンジン・サジェストは，ウェブ検索者の検索履歴の収集結果に基づいて作られている．このことから，検索エンジン・サジェストには，ウェブ検索者の関心事項が反映されていると言える．そこで，本論文では，検索エンジン・サジェストを情報収集の情報源として，ウェブ検索者の情報要求観点を収集する．

本論文の枠組みにおいて，一つのクエリ・フォーカスに対して，日本語側では最大約 1,000 個のサジェストを収集する．中国語側では最大約 600 個のサジェストを収集する．そして，クエリ・フォーカスにサジェストを加えて，AND 検索によってウェブページの収集を行う．ここで，収集されるサジェスト，および，ウェブページ集合の双方において，話題の内容が重複し，かつ，冗長である点が問題である． そこで，まず，トピックモデル (具体的には，潜在的ディリクレ配分法 (LDA: Latent Dirichlet Allocation) [3]) を適用して話題集約を行う．この手法では，収集されたウェブページ集合に対して LDA を適用することによって，話題ごとにウェブページのクラスタリングを行う．各ウェブページを検索する際には，サジェストが指定されているため，ウェブページごとに少なくとも一つのサジェストが対応付けられる．この対応関係を用いることにより，約 1,000 個のサジェストを数十個程のまとまりに集約する [8].

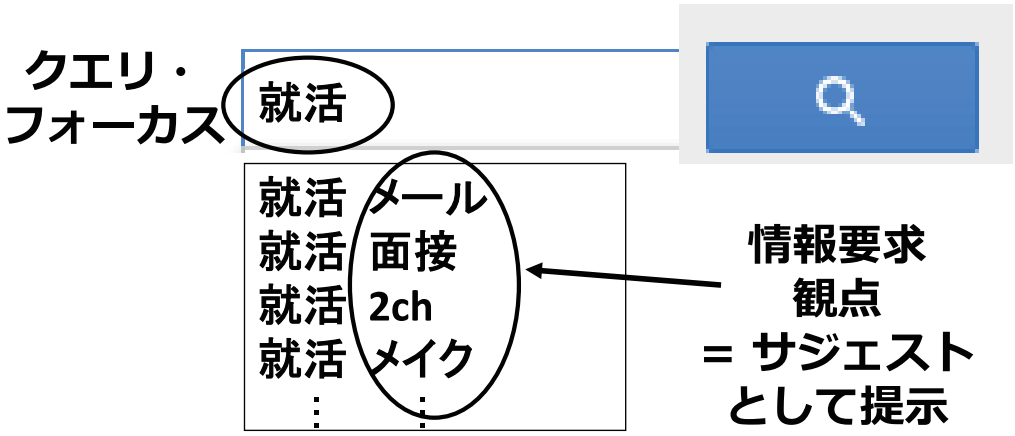


図 3.1: 検索エンジン・サジェストにおける情報要求観点の例

ここで，トピックモデルを用いた集約の問題点として，話題集約の粒度が粗い点が挙げられる．一方，近年，深層学習技術により，単語の意味表現を分散表現によって表す方式が提案され，その有効性が報告されている [18]．この方式では，大規模なコーパスから語の意味のベクトル表現を学習し，これを用いて各語の周囲の語 (文脈) の予測を行う．そこで，本論文では，分散表現によって表現される話題の粒度が相対的に細かいところに着目し，分散表現とトピックモデルを併用することによって，検索エンジン・サジェストの類似度をより詳細に測定し，これを用いることによって，検索エンジン・サジェストの話題をより精密に集約する手法を提案する．

3.2 検索エンジン・サジェストの収集

本論文では，評価対象となるクエリ・フォーカスに対して，日本語側においては Google 検索エンジン [1] を用いて，一クエリ・フォーカスにつき約 100 通りの文字列を指定する．具体的には，五十音，濁音，半濁音および「きゃ」や「ぴゃ」などの開拗音である．そして，一通りの文字列当り最大 10 個のサジェストを収集することにより，一クエリ・フォーカスあたり最大約 1,000 個のサジェストを収集し，これを集合 S とする．例えば検索窓に「就活 あ」と入力すると,「あいさつ」や「あなたの強み」等のサジェストが表示されるので，それらの収集を行う．中国語側においては，日本語側と同じように，Baidu 検索エンジン [2] を用いて，一クエリ・フォーカスにつき約 100 通りの文字列を指定する．具体的には，pinyin を用いて，1 通りの文字列当り約 10 個のサジェストを収集することにより，一クエリ・フォーカスあたり最大約 600 個のサジェストを収集する．クエリ・フォーカス「就

[1] https://www.google.com/
[2] https://www.baidu.com/

表 3.1: クエリ・フォーカス，サジェスト数，ウェブページ数，および，トピック数

クエリ・フォーカス	サジェスト数			ウェブページ数	総トピック数	評価対象トピック数	$\frac{\text{小分類数}}{\text{評価対象トピック数}}$
	ウェブページ収集時	Wikiのみで頻度下限以上	Wiki＋ウェブページにおいて頻度下限以上				
就活(日本)	926	559	671	12,078 (30.8 MB)	50	13	2.85
結婚(日本)	959	709	771	13,256 (31.0 MB)	60	23	4.57
就活(中国)	531	225	409	19,589 (113.0 MB)	60	41	4.41
結婚(中国)	532	282	433	18,950 (110.0 MB)	60	42	4.14

活」および「結婚」に対して収集されたサジェストの数を表 3.1 「サジェスト数，ウェブページ収集時」欄に示す．

3.3 トピックモデルを用いた検索エンジン・サジェストの集約

3.3.1 サジェストを用いたウェブページの収集

収集したサジェストを用いてウェブページの収集を行う． クエリ・フォーカスにサジェスト s を加えて AND 検索することによって上位 N 件以内に検索されるウェブページ d の集合を $D(s, N)$ (ただし，本論文では，日本語側においては，$N = 20$ とし，中国語側においては，$N = 40$ とする)． ウェブページの収集について，日本語側では，Google Custom Search API[3] を用いる．中国語側では，Baidu 検索エンジン [4] を用いる．また，各ウェブページ d に対して，$d \in D(s, N)$ となるサジェスト s を集めた集合を次式 $S(d)$ とする．

$$S(d) \quad = \quad \left\{ s \in S \middle| d \in D(s, N) \right\}$$

日中クエリ・フォーカス「就活」および「結婚」に対して収集したウェブページの数およびテキストサイズを表 3.1 に示す．集めたウェブページの集合を D とす

[3] https://cse.google.com/cse/
[4] https://www.baidu.com/

る．そして，この D にトピックモデルを適用し，トピックの推定を行い，推定されたトピック分布に従ってサジェストの集約を行う．

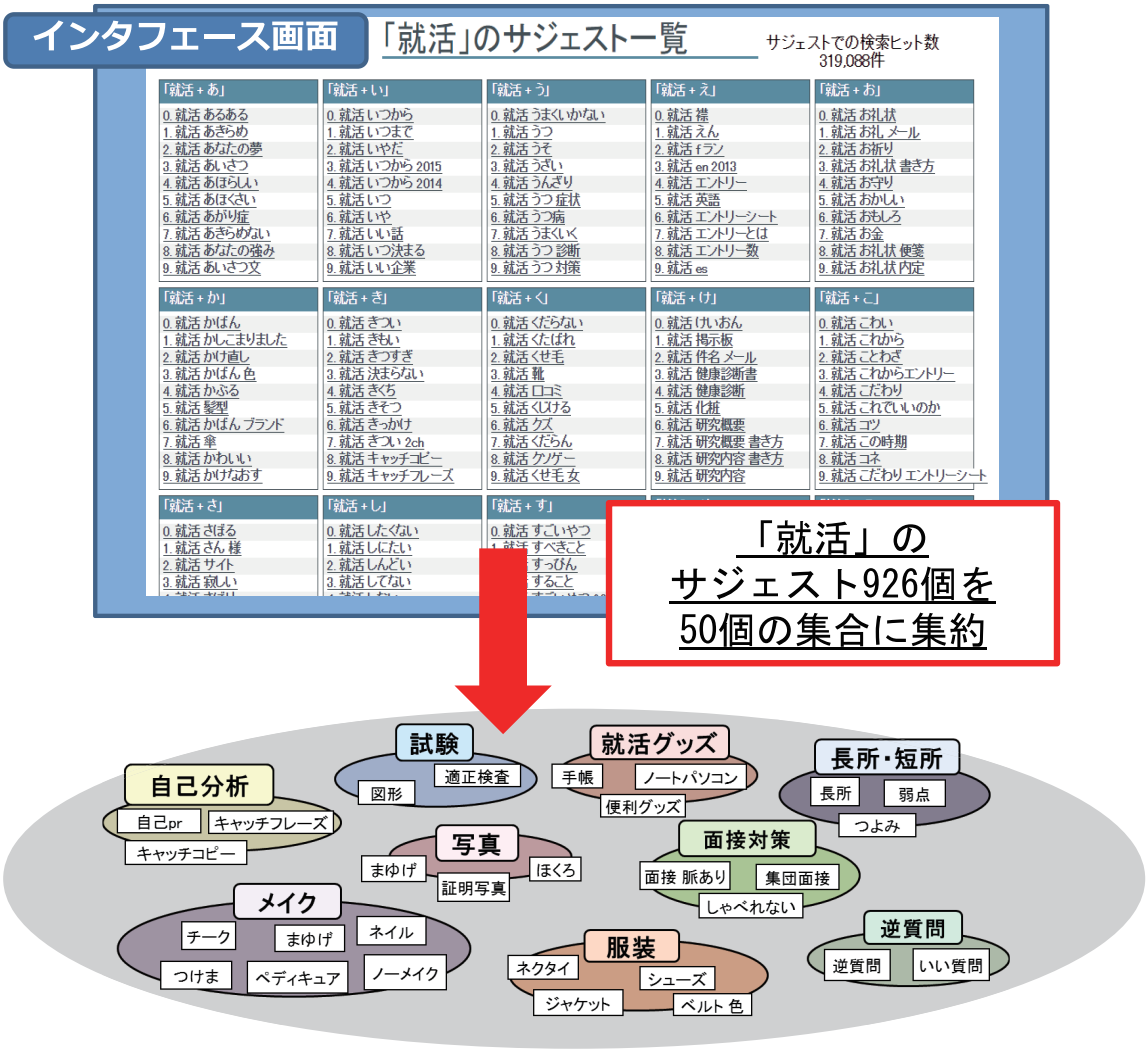


図 3.2: トピックモデルを用いた日本語検索エンジン・サジェストの集約 (クエリ・フォーカス: 「就活」)

3.3.2 トピックモデル

本論文で適用するトピックモデルは，潜在的ディリクレ配分法 (LDA; Latent Dirichlet Allocation) [3] である．トピックモデルの推定においては，入力として，語 w の列によって表す文書集合，および，トピック数 K を指定する．出力として，各トピック z_n $(n = 1, \ldots, K)$ における語 w の確率分布 $P(w|z_n)$ $(w \in V)$ ，および，

各文書 d におけるトピック z_n の確率分布 $P(z_n|d)$ $(n = 1, \ldots, K)$ を得る．LDA のツールとしては GibbsLDA++[5] を用いた．LDA のハイパーパラメータの α，β については，GibbsLDA++のデフォルト値である $\alpha = 50/K$，$\beta = 0.1$ を使い，Gibbs サンプリングの反復回数を 2,000 に設定した． 語 w の集合 V としては，日本語側においては，日本語 Wikipedia 中のタイトルおよびそのリダイレクトの集合[6] を用いた．中国語側においては，ウェブページの文書集合に対して形態素解析[7] を行い，抽出された名詞，形容詞，動詞の集合を用いた． また，トピック数 K としては，K を 10 から 80 まで変化させてトピック推定を行った上で，トピック推定による話題のまとまりが最もよいトピック数 K を人手で選定した．クエリ・フォーカス「就活」においては，日本語側と中国語側それぞれ，$K = 50$，$K = 60$ を採用した．クエリ・フォーカス「結婚」においては，日本語側と中国語側それぞれ，$K = 60$，$K = 60$ を採用した．

3.3.3 文書に対するトピックの割り当て

各ウェブページに対して最も確率が高いトピックに割り当てることによって，ウェブページ集合をトピックに集約する．ウェブページ集合を D，トピック数を K，1 つのウェブページを $d(d \in D)$ とすると，トピック $z_n(n = 1, \ldots, K)$ のウェブページ集合 $D(z_n)$ は次式で表される．

$$D(z_n) \;=\; \left\{ d \in D \;\middle|\; z_n = \underset{z_u\ (u=1,\ldots,K)}{\arg\max}\ P(z_u|d) \right\}$$

3.3.4 トピックに対するサジェスト割り当てによるサジェストの集約

各ウェブページは，クエリ・フォーカスに一つのサジェストを加えて AND 検索することによって収集される．そのため，各ウェブページに対して一つ以上のサジェストが対応する．ここで，ウェブページ d にはサジェスト集合 $S(d)$ 中のサジェストが対応する．また，ウェブページ d には，トピック z_n が割り当てられている．すると，トピック z_n に対して割り当てられたウェブページ $d \in D(z_n)$ に対応するサジェスト s を集めることによって，トピック z_n に対してサジェスト s が割り当てられる．トピック z_n に割り当てられたサジェストの集合 $S(z_n)$ は次の式で表される．

[5] `http://gibbslda.sourceforge.net/`

[6] トピックモデルにおける語の集合として用いる日本語 Wikipedia としては，2014 年 3 月にダウンロードしたエントリ数約 140 万 7,000 のものを用いた．

[7] Jieba (`https://github.com/fxsjy/jieba`) を用いた．

$$S(z_n) \quad = \quad \bigcup_{d \in \ D(z_n)} S(d)$$

さらに，トピック z_n におけるサジェスト s の頻度 $f(s, z_n)$ は下記の式で表される．

$$f(s, z_n) \quad = \quad \left| \left\{ d \in D(z_n) \;\middle|\; s \in S(d) \right\} \right|$$

例えば，日本語のクエリ・フォーカス「就活」の場合，926 個のサジェストが 50 個のトピックに割り当てられた (図 3.2)．以上のように，検索エンジン・サジェストを用いて収集されたウェブページ集合に対してトピックモデルを適用することによって，検索エンジン・サジェストが集約される．

3.4 分散表現を用いた検索エンジン・サジェストの集約

本節では，前節において収集された日本語および中国語のサジェストに対して word2vec [18] を適用し，サジェストの分散表現を計算する．そして，得られた分散表現を用いて，同一トピックに集約されているサジェスト同士の類似度を測定することによって，検索エンジン・サジェストの話題をより精密に集約する [24].

3.4.1 分散表現

分散表現 [18][8] においては，訓練コーパスを単語列 $w_1, w_2, \ldots, w_T$ で表し，位置 t の単語 w_t の前後 δ 語の単語列を文脈窓 $C_{w_t} = (w_{t-\delta}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+\delta})$ とする．そして，文脈窓 C_{w_t} から単語 w_t を予測する条件付き確率分布関数 $p(w_t|C_{w_t})$ を定義し，次式の対数尤度 L を目的関数として，次式を最大化する分散表現を学習する．

$$L \quad = \quad \sum_{t=1}^{T} \log p(w_t|C_{w_t})$$

本論文では，word2vec における分散表現の実装の中でも，特に，skip-gram による実装を用いた場合について述べる．skip-gram では，条件付き確率分布関数 $p(w_t|C_{w_t})$ を，各文脈語 $c \in C_{w_t}$ から単語 w_t を予測する条件付き確率 $p(w_t|c)$ の積に分解する．

$$L^{SG} \quad = \quad \sum_{t=1}^{T} \sum_{c \in C_{w_t}} \log p(w_t|c)$$

[8] 本節における分散表現の定式化は，文献 [25] にしたがう．

そして，コーパス中の全語彙集合を V として，単語 c のベクトル v_c，および，単語 w を予測するベクトル $\tilde{v}_w$ の二種類のベクトルを導入し，さらに，ソフトマックス関数を用いることによって，単語 c から単語 w_t を予測する条件付き確率 $p(w_t|c)$ を次式で定式化する．

$$p(w_t|c) = \frac{\exp(v_c \cdot \tilde{v}_{w_t})}{\sum_{w' \in V} \exp(v_c \cdot \tilde{v}_{w'})}$$

ここで，上式の分母の計算においては，コーパス中の全語彙 V を用いるのではなく，無作為に選んだ K 個の疑似負例単語 $\breve{w}'$ を用いる．さらに，シグモイド関数 σ を用いた近似により，次式を最大化する分散表現を求め，最終的に単語 c のベクトル v_c を得る．

$$L^{SG} = \sum_{t=1}^{T} \sum_{c \in C_{w_t}} (\log \sigma(v_c \cdot \tilde{v}_{w_t}) + \sum_{k=1}^{K} \log \sigma(-v_c \cdot \tilde{v}_{\breve{w}'}))$$

3.4.2 分散表現の類似度

前節の方式によって，語 w_a および w_b の分散表現 v_{w_a} および v_{w_b} を求めた後，分散表現であるベクトル間の余弦類似度を求め，これを分散表現間の類似度と定義する．

$$Sim(w_a, w_b) = \frac{v_{w_a} \cdot v_{w_b}}{| v_{w_a} || v_{w_b} |}$$

3.4.3 検索エンジン・サジェストの分散表現の作成手順

検索エンジン・サジェストの分散表現を求めるために，分散表現訓練用コーパスを用意する．検索エンジン・サジェストを含む典型的な用例コーパスとして，Wikipedia の全ページテキストを用いる[9]．さらに，各クエリ・フォーカスに対して収集された検索エンジン・サジェスト特有の特性を反映した分散表現を得るために，各クエリ・フォーカスに対して収集されたウェブページを加え，これら二種類のテキストデータの混合集合を分散表現訓練用コーパスとして，検索エンジン・サジェストの分散表現を求める．また，ベースラインとして，日本語，および，中国語 Wikipedia の全ページテキストのみを用いて分散表現を訓練した場合との比較を行う．ここで，分散表現訓練時には，各サジェストが頻度下限値 (本論文では，頻度下限値を 5 とする) を満たす必要がある．分散表現訓練用コーパスと

[9]日本語コーパス: 2016 年 2 月時点の日本語 Wikipedia 全 100 万ページ (約 4.8 億語，2.91GB) を用いる．中国語コーパス: 2016 年 2 月時点の中国語 Wikipedia 全 80 万ページ (約 3 億語，1.4GB) を用いる．

して，(i) Wikipeida のみを用いた場合，および，(ii) Wikipedia および各クエリ・フォーカスに対して収集されたウェブページの混合集合を用いた場合，の両方について，この頻度下限値を満たす検索エンジン・サジェストの数を表 3.1 に示す．なお，word2vec を適用する際には，分散表現の次元数を 256 とし，前後の文脈幅 $\delta = 8$ とする．

3.4.4 サジェスト間の類似度測定例

一例として，日本語クエリ・フォーカス「就活」の場合について，トピックモデル推定結果におけるトピックのうち,「グループディスカッション，グループワーク，プレゼンテーション，プレゼン」等の検索エンジン・サジェストが割り当てられたトピックを対象として，一部の検索エンジン・サジェスト間で分散表現の類似度を測定した結果を図 3.3 に示す．ここで，分散表現間の類似度の下限値を 0.6 に設定した場合を想定すると,「グループディスカッション，グループワーク，プレゼンテーション，プレゼン」の四つのサジェストのうちの任意の二組のうち，類似度下限値の条件を満たすサジェスト組は,「グループディスカッション，グループワーク」および「プレゼンテーション，プレゼン」の二組のみとなることが分かる．主観評価においても，この二組のみが類似関係にあると言えるため，分散表現の類似度を用いることによって，主観評価の結果と一致する類似度が測定できていることが分かる．

3.4.5 評価手順

分散表現を用いたサジェスト間類似度の評価を行うために，日中の「就活」および「結婚」の各クエリ・フォーカスを対象として，各トピック中の検索エンジン・サジェストに対して，人手で類似サジェストをまとめることにより，「検索エンジン・サジェストの参照用小分類」を作成した．その際，(i) サジェストの意味的まとまりがないトピック，および，(ii) 逆に全てのサジェストが均一的にまとまっており，より詳細な小分類を作ることが困難なトピック，は評価対象から除外した．その結果，表 3.1 の「評価対象トピック数」欄に示す数のトピックが評価対象となった．また，評価対象のトピック一つ当たりの小分類数の平均は，表 3.1 の「小分類数/評価対象トピック数」欄に示す値となった．

次に，二つのサジェストの組に対して，

1. サジェスト間の分散表現類似度の下限値，
2. あるサジェストに対して，分散表現類似度の降順に他のサジェストを順位付けした場合の

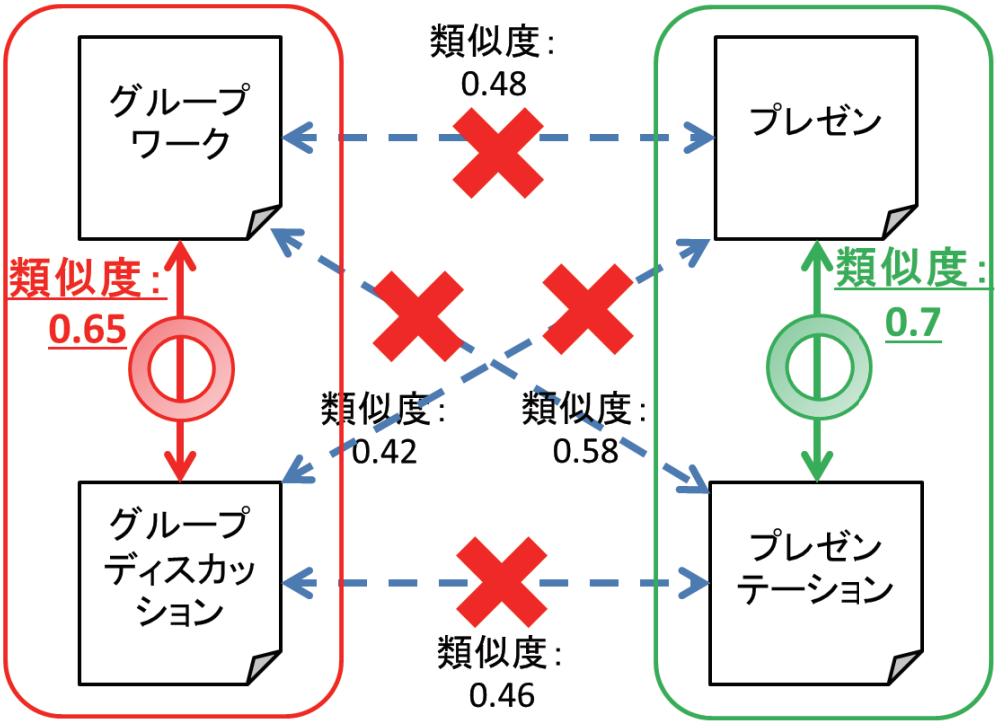


図 3.3: 分散表現を用いた日本語サジェスト間の類似度の測定例 (クエリ・フォーカス: 「就活」, 類似度の下限値: 0.6 の場合)

(a) 全サジェスト中の順位の上限値,
(b) トピック内のサジェスト中の順位の上限値.

の三つの素性を設定する.

そして，三つの素性のあらゆる可能な組み合わせに対して,「検索エンジン・サジェストの参照用小分類」中のサジェスト組の集合を R，三つの素性の組み合わせを満たすサジェスト組の集合を S_f として，次式で算出される再現率・適合率をプロットする.

$$\text{再現率} \ = \ \frac{\left|R \cap S_f\right|}{\left|R\right|}, \quad \text{適合率} \ = \ \frac{\left|R \cap S_f\right|}{\left|S_f\right|}$$

3.4.6 評価結果

ここで，分散表現訓練用コーパスとして,

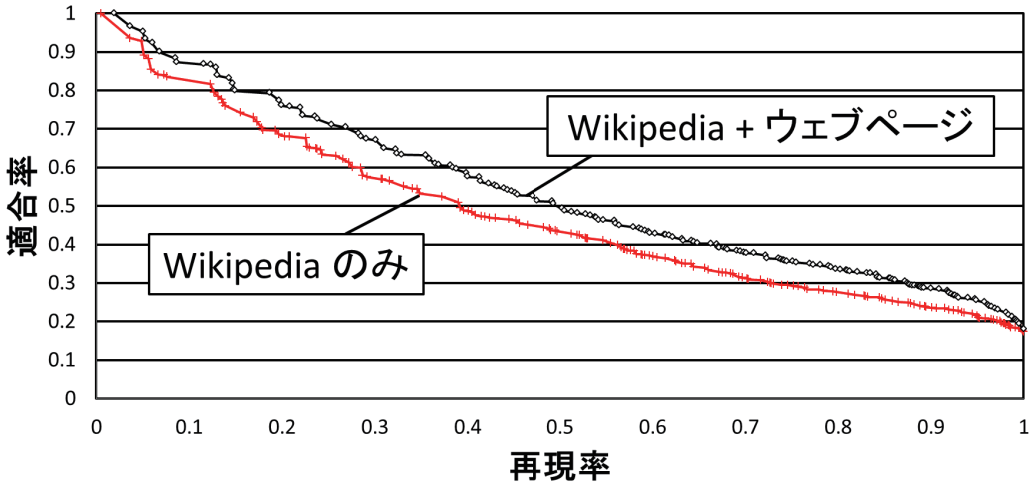
(i) Wikipeida のみを用いた場合,
(ii) Wikipedia および各クエリ・フォーカスに対して収集されたウェブページの混合集合を用いた場合,

の間でこのプロットを比較した結果を図 3.4 と図 3.5 に示す．この結果から，日中クエリ・フォーカス「就活」および「結婚」の両方において，分散表現訓練用コーパスとして (ii) を用いた場合の有効性を示すことができた.

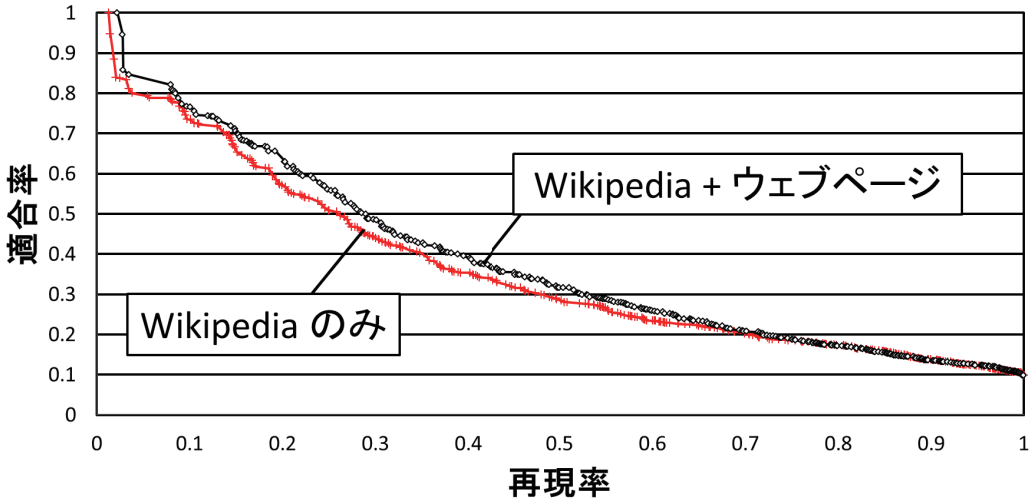
3.5 関連研究

本論文に関連して，従来型のトピックモデルにおいては，語そのものに対する確率分布としてトピックモデルを推定するのに対して，[12] においては，語に対して求めた分散表現に対する確率分布としてトピックモデルを推定する方式を提案している．一方，本論文では，従来型のトピックモデルにおいても，粒度の粗いトピックとしては十分良質なトピックが得られているという立場に立ち，その上で，各トピックに対してより詳細かつ精密な話題集約を実現するために，検索エンジン・サジェストの分散表現を利用する方式を提案した.

その他の関連研究としては，クリックスルーデータを用いて検索クエリのクラスタリングを行う手法 [14, 5, 11] が挙げられる．文献 [14, 5] では，大量のクリックスルーデータを用いて検索クエリのクラスタリングを行った後，出力されたクラスタをベースにして検索クエリを推薦する手法を提案している．文献 [11] では，検索クエリのクラスタリングをユーザ毎に行う手法を提案している．この研究では，ユーザの個人特徴に着目し，各ユーザの趣味を考慮した検索クエリのクラス

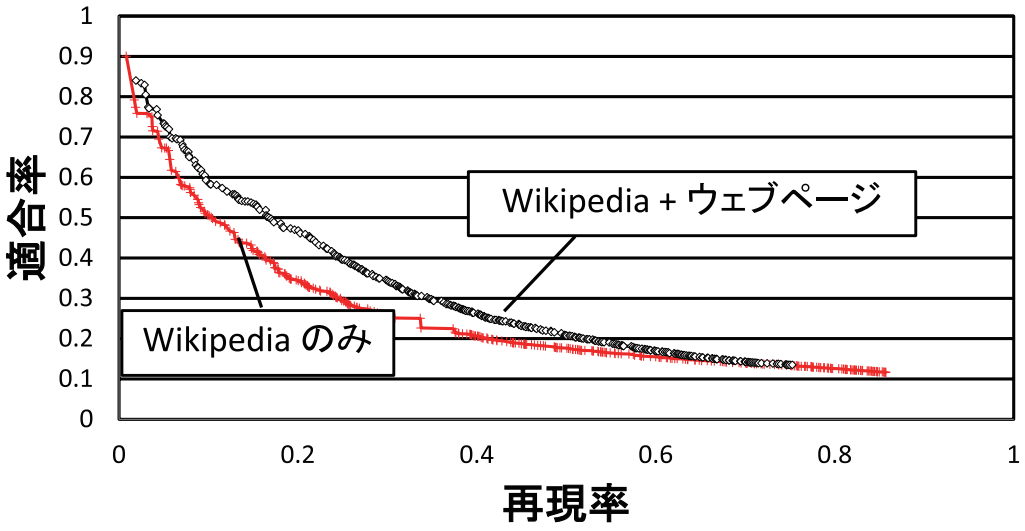


(a) クエリ・フォーカス: 「就活」

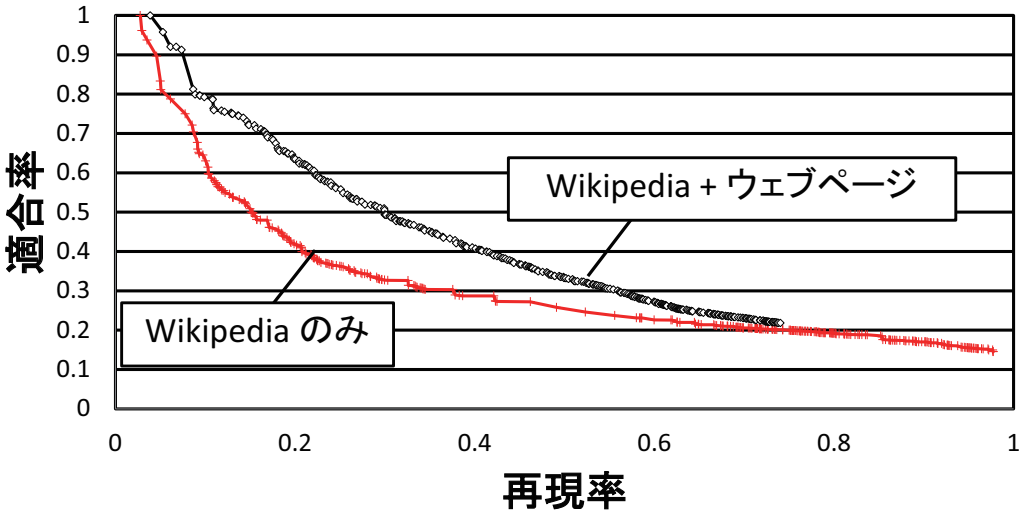


(b) クエリ・フォーカス: 「結婚」

図 3.4: 日本語検索エンジン・サジェストの類似度の評価結果



(a) クエリ・フォーカス: 「就活」



(b) クエリ・フォーカス: 「結婚」

図 3.5: 中国語検索エンジン・サジェストの類似度の評価結果

タリングを行っている．評価実験では，30 人程度の検索ユーザを対象として，約 150 個の検索クエリについて, クラスタリング結果を評価している．

3.6　本章のまとめ

本章では，検索エンジン・サジェストとウェブページの間の関連性に着目し，日本語と中国語の文書集合にトピックモデルを適用することによって，サジェストの自動集約を行なった．また，トピックモデルによる話題集約の粒度は粗い，という問題点に対して，分散表現によって表現された語の意味表現を併用することによって，より詳細かつ精密な話題集約手法を提案した．

第4章　日中間比較対照分析に関する研究と本論文の位置付け

4.1　本論文の位置付け

ウェブを情報源とした文化間差異発見支援の研究 [20, 7, 30, 21, 23, 4] と本論文の位置付けに関する詳細は，表 4.1 に記述されている．このうち，文献 [21] では，執筆者観点に着目し，日中質問回答サイトから収集した質問回答事例を比較対照分析することで，日中間文化間差異を発見する手法を提案した．文献 [30] においては，執筆者観点に着目し，文献 [16] の手法を用いて，日中ブログを対象として，日中間の差異を発見する手法を提案した．文献 [23] では，検索者観点に着目し，日本と中国の検索エンジンサジェストを収集し，比較対照分析することで日中間文化間差異を発見する手法を提案した．ただし，文献 [23] ではトピックモデルを用いていない．一方，本論文では，より細かい粒度で日中間の文化間差異を発見するために，トピックモデルと分散表現を併用することによってい検索エンジン・サジェストの高精度の集約を行い，その上での日中間比較対照分析による日中間文化間差異を発見する手法を提案した．

4.2　ウェブ執筆者に着目した研究

4.2.1　日中質問回答サイトの比較対照分析による文化間差異発見支援

日本と中国の間には，過去の歴史に起因する理由などもあり，様々な問題において相反する意見や見解が多い．しかし，近年では，二国間の経済や文化の交流も促進され，お互いの国についての関心が高まりつつあるので，日中間の文化間差異を発見することが大切だと考えられる．ここで，これまで，その国独特の文化の実態は，その国に直接行ってみなければ知ることが難しいのが実情であった．しかし，多くの人にとっては，時間的制約や経済的制約があるため，インターネットやテレビを通して間接的にその国の文化を体験することしかできなかった．また，従来より，文化間の差異の発見に関する研究を実施するためには，一次資料

表 4.1: 日中間比較対照分析に関する研究と本論文の位置付け

ウェブを情報源とした文化間差異発見支援の研究

	情報源	トピックモデルの利用の有無	
		なし：話題のまとまりをモデル化する仕組みがない	あり：トピックによって話題のまとまりをモデル化
執筆者観点	ニュース		**文献[7]：ニュースを対象として、日中間話題自動対応付け**
	ブログ	**文献[20]：ブログを対象として、日英間の差異を発見**	**文献[30]：ブログを対象として、日中間の差異を発見**
	質問回答サイト	**文献[21]：日中質問回答サイトでのトラブル情報を対象として、文化間差異を発見**	
検索者観点	検索エンジン・サジェスト	**文献[23]：ウェブ検索者の情報要求観点を収集し、日中間の差異を発見する**	**文献[4]：ウェブ検索者の要求観点をトピックモデルで集約し、文化間差異を発見**
			[本論文]：ピックモデル・分散表現の併用による検索エンジン・サジェストの高精度な集約&日中対照分析

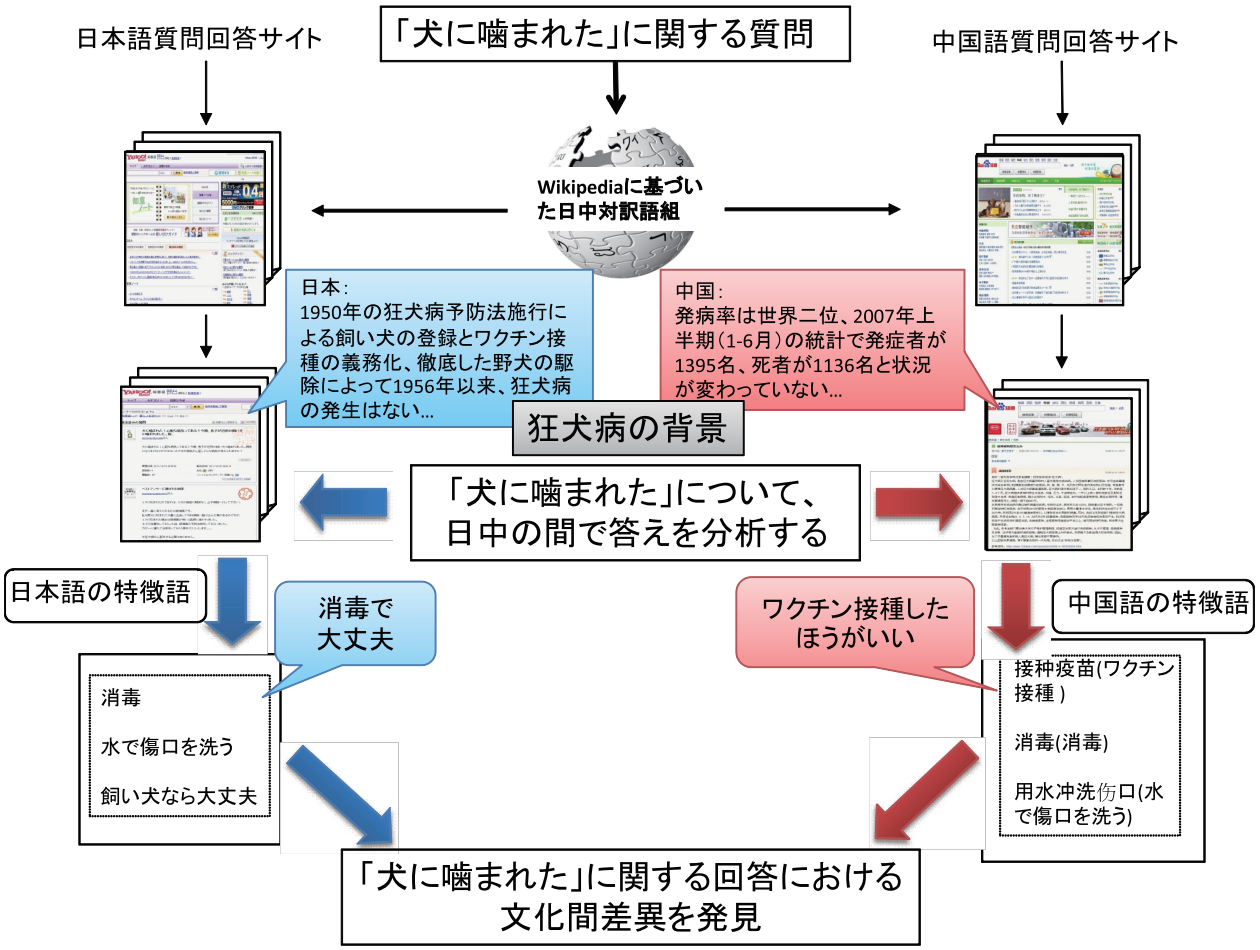


図 4.1: 日中質問回答サイトの比較対照分析による文化間差異発見支援

の分析やヒアリング調査などの手法を適用する必要があったが，このような状況においては，それらの研究を実施するためには莫大なコストが必要となっていた．

ところが，近年，情報技術の進展に伴い，インターネットを利用することによって，各国の文化の実態の一端を把握することが可能になりつつある．そこで，文献 [21] では，インターネット上においても，特に，各国の質問回答サイトを情報源とすることによって，直接その国を訪問することなく各国の文化の実態についての情報を収集し，二国間の文化間差異の発見を支援する方式の確立をした．具体的には，日本語と中国語の質問回答サイトから，ある同一の問題についての日本語と中国語の質問と回答を収集し，回答中の記述を対象として，日本と中国の間の文化的対照性を分析する方法を提案する．ここで，中崎らは，文献 [20] において，日英ブログサイトにおける文化間差異の発見を支援する方式を提案したが，文献 [21] では，日中質問回答サイトにおける文化間差異発見支援のタスクを対象としてこの方式を適用する．文献 [21] において提案する「日中質問回答サイトの比較対照分析による文化間差異発見支援」の枠組みを図 4.1 に示す．

まず，文献 [21] では，日本語質問回答サイトとして Yahoo!知恵袋 [1] を，また，中国語質問回答サイトとして，Baidu(百度) 知道 [2]，および，搜搜問問 [3] を，それぞれ利用する．そして，特定の話題を表すクエリを用いて，日中各言語の質問回答サイトから質問・回答組を収集し分析を行う．分析の第 1 ステップにおいては，日中各言語において 100∼200 組程度の質問・回答組を対象として，質問・回答組の内容を分析し，類似する内容のまとめ上げを行い，異なる内容の数を集計する．次に，日中の間でその内容の対応関係をつけ,「日本側でのみ観測した内容」,「中国側でのみ観測した内容」,「日中両側で観測した内容」に分類して集計する．次に，分析の第 2 ステップにおいては,「日本側でのみ観測した内容」，および,「中国側でのみ観測した内容」を対象として，当該内容が観測されなかったと判定された言語側で追加の検索を行い，本当にその当該内容が検索されないかどうかの検証を行う．

例えば，図 4.1 に示すように,「犬に噛まれた」という話題を対象とした場合，中国語の質問回答サイトにおいて最も多い回答としては,「ワクチンを接種したほうがよい」といった緊張感のある回答の件数が圧倒的になるのに対して，日本語の質問回答サイトにおいて最も多い回答としては,「消毒をすれば大丈夫」といった緊張感の少ない回答の件数が圧倒的となる．この背景には，中国は狂犬病発症率が世界第二位であるのに対して，日本では 1,956 年以来狂犬病の発生が記録されていない，といった社会的文化的差異の影響が大きく関わっているということができる．

[1] http://chiebukuro.yahoo.co.jp/
[2] http://zhidao.baidu.com/
[3] http://wenwen.soso.com/

表 4.2: 各話題において用いたクエリ・クエリの出現数・分析対象質問・回答組数

話題	日本側			中国側 (Baidu / 搜搜)		
	クエリ	クエリの出現数	分析対象質問・回答組数	クエリ	クエリの出現数	分析対象質問・回答組数
犬に噛まれる	犬に噛まれた	136	124	被狗咬了	743/499	124/135
刺身と寄生虫	刺身 and 寄生虫	89	87	刺身 and· 寄生虫	307/調査中	76/—
原発	原発	1,412	131	原子能发电	750/調査中	93/—
喫煙	喫煙	18,534	149	吸烟	741/344	141/149

日中質問回答サイトの質問・回答事例

日本側の質問回答サイトの質問・回答事例としては，Yahoo!知恵袋から提供されている 2004 年 4 月 1 日 ~2009 年 4 月 7 日の 5 年間の質問・回答事例のデータ (質問: 16,257,413 件，回答: 50,053,894 件) を分析対象とした．質問には，カテゴリ情報が付与されており，最下位層の分類として 453 種のカテゴリが存在している．453 種のカテゴリは，それぞれ親カテゴリ，さらにその親カテゴリを持つ三層構造になっており，各カテゴリに数万 ~ 数十万の質問が含まれている．

一方，中国側の質問回答サイトの質問・回答事例は，2012 年 12 月 ~2013 年 2 月の期間に，Baidu(百度) 知道，および，搜搜問問のサイトから収集した．2012 年 12 月の時点で，Baidu(百度) 知道に掲載されていた解決済質問数は 215,755,535 件，搜搜問問に掲載されていた解決済質問数は，207,374,517 件であった．Baidu(百度) 知道，および，搜搜問問のサイトのいずれも，質問には 3~4 階層のカテゴリ情報が付与されており，最上位層のカテゴリ数は 14 種類であった．

分析対象の質問事例の収集

文献 [21] では，特に，日本と中国の間で文化的な差異が大きいことが期待できる話題として，以下の 4 つの話題を選定した．以下では，各話題およびその選定理由を示す．

- 「犬に噛まれる」· · · 中国は狂犬病発症率が世界二位であるため，犬に噛まれば狂犬病が発症する危険性が高いと考えている．

表 4.3: 日中間における質問回答組の内容の比較: 第 1 ステップ / 第 2 ステップ

話題		日本側でのみ観測した内容の数	中国側でのみ観測した内容の数	日中両側で観測した内容の数
第 1 ステップのみ	犬に噛まれる	10 / —	6 / —	9 / —
	刺身と寄生虫	10 / —	9 / —	1 / —
第 1/第 2 ステップ	喫煙	10 / 6	13 / 9	5 / 13
	原発	8 / 7	3 / 2	4 / 6

- 「刺身と寄生虫」··· 中国では生魚の衛生管理面での信頼性が日本よりも低いと考えられている.
- 「原発」··· 日中両国においては，原発への関心の度合いに差異がある (ただし，日本側において分析対象とした質問・回答事例は 2009 年時点までのものである).
- 「喫煙」··· 日本では 20 歳未満は喫煙禁止であり，公共の場所での喫煙は禁止されているが，中国では喫煙できる年齢に関する法律はなく，公共の場所での喫煙が禁止されたのも最近のことである.

そして，各話題について，表 4.2 に示す日中両言語のクエリを用いて，各クエリを文字列として含む質問・相談組を収集した．特に，Baidu(百度) 知道，および，搜搜問問においては，収集可能な質問・回答組の上限が，それぞれ，750 組，および，500 組であったので，この上限の範囲内で収集可能な質問・回答組を収集した．収集された質問・回答組数を，表 4.2 中の「クエリの出現数」欄に示す．次に，各クエリを文字列として含む質問・回答組を人手で分析し，分析対象とする話題に関連する質問・回答組を選別し，最終的に表 4.2 中の「分析対象質問・回答組数」欄に示す個数の質問・回答組を分析した.

日中文化間差異の分析

分析手順

「第 1 ステップ」

まず，前節で収集された日中両言語の質問・回答組の内容を人手で分析し，類似の内容の質問・回答組をまとめることにより，内容の種類数を集計する．その

表 4.4: 話題「犬に噛まれた」において観測された内容および質問・回答組数 (第 1 ステップのみ)

話題		日本側の質問・回答組数	中国側の質問・回答組数 (Baidu/搜搜)
日本側でのみ観測	**飼い主の責任の所在に関する相談 (9 件)** **近所との人間関係についての相談 (8 件)** **飼い犬に噛まれたことがある回答者への相談 (8 件)** **ペットと飼い主との関係についての相談 (7 件)** **破傷風の往診を勧める (7 組) / 犬に噛まれても大丈夫 (6 組)** **犬に噛まれた傷跡についての相談 (6 件)** 犬のワクチン接種を勧める / 狂犬病の保険についての相談 飼い犬が人を噛まない理由についての質問	計 57	—
中国側でのみ観測	**噛んだ犬の様子を観察し狂犬病の犬かどうか判断することを勧める (18 件)** 出血がなければ大丈夫 / 潜伏期間が過ぎれば大丈夫 狂犬病のワクチンの値段に関する質問 犬に噛まれたことについて悩んでいる精神疾患についての相談 ワクチンを接種した犬に噛まれた場合は狂犬病になるとは限らない	—	計 26 / 計 22
日中両側で観測	**狂犬病ワクチンの接種を勧める**	1	**68 / 72**
	狂犬病の基礎知識を説明し，必要があればワクチン接種を勧める	1	**15 / 33**
	狂犬病の基礎知識を説明	2	6 / 1
	噛まれたことを心配している質問に対し対策を回答しているが，回答者に狂犬病の知識はない.	**28**	4 / 3
	犬に噛まれたことに関する質問および回答において法律関係の事項を含む	**20**	1 / 0
	ペットが犬に噛まれたのでアドバイスを求めている (狂犬病への言及なし)	6	3 / 1
	犬恐怖症対策に関する相談	5	1 / 0
	犬に噛まれた時の感触についての相談	3	0 / 1
	狂犬病の往診を勧める	1	0 / 2
合計		124	124 / 135

表 4.5: 話題「刺身と寄生虫」において観測された内容 (第 1 ステップのみ，抜粋)

日本側でのみ観測された内容	中国側でのみ観測された内容	日中両側で観測された内容
特定の魚の刺身が心配という相談に対して，問題ないという回答 (42 件) **特定の魚の刺身が心配という相談に対して，魚種によっては寄生虫が危険という回答 (14 件)** 生の魚は心配という相談に対して，食べない方がよいという回答 (5 件) その他 26 件の大半は「刺身は安全」という回答	**魚の刺身が心配という相談に対して，寄生虫が危険なので食べない方がよいという回答 (50 件)** **魚の刺身が心配という相談に対して，海水魚なら問題ないという回答 (11 件)** その他 15 件の大半は「生魚を不安視」する回答	川魚の刺身についての相談に対し，食べない方がよいという回答 (日本 4 件，中国 2 件)

表 4.6: 話題「喫煙」において観測された内容 (第 1 ステップ・第 2 ステップ)

(a) 日本側または中国側でのみ観測された内容

日本側でのみ観測	中国側でのみ観測
喫煙者が禁煙の風潮に対するクレームを相談 (27 件) **喫煙者に対する印象についての質問と回答 (悪い印象であるという回答 (25 件)**，中立的な印象であるという回答) 喫煙者と非喫煙者が共存する仕組みについて相談と回答 喫煙を始めた時期についての質問と回答 国会議員の喫煙に関する発言についての議論	**喫煙のよい点についての質問と回答 (喫煙者は魅力的，喫煙すると集中できる，国の税収増) (29 件)** **喫煙の仕方についての質問と回答 (13 件)** 中国の喫煙率についての質問と回答 (2009 年当時で約 25%) 喫煙した方が大人びて見えるかという質問と回答 (関係ないと回答) その他，3 種類の内容

(b) 日中両側で観測された内容

喫煙の害についての質問と回答 (日本 18 件，中国 215 件) / 同僚の喫煙による被害の回避策について相談 (日本 37 件) その他，10 種類の内容

後，それらの内容を日中間で対応付けることにより，

(1) 日本側でのみ観測した内容，

(2) 中国側でのみ観測した内容，

(3) 日中両側で観測した内容，

に分類する．分析対象とした 4 つの話題について，上記の (1)∼(3) の内容の数を表 4.3 に示す．

「第 2 ステップ」

次に，話題「喫煙」，および，「原発」について，「(1) 日本側でのみ観測した内容」が中国側の Baidu(百度) 知道，および，捜捜問問において観測できないかどうか，収集した全質問・回答組 (「原発」については，750 組 (Baidu(百度) 知道),「喫煙」については，741 組 (Baidu(百度) 知道)，および，344 組 (捜捜問問)) を対象として調査を行う．同様に，「(2) 中国側でのみ観測した内容」が，日本側の Yahoo! 知恵袋において観測できないかどうか，収集した全質問・回答組 (「原発」については，1,412 組,「喫煙」については，18,534 組) を対象として調査を行う．分析対象とした 2 つの話題について，この第 2 ステップの後の (1)∼(3) の内容の数を同様に表 4.3 に示す．

分析結果

表 4.3 に示すように，4 つの話題のいずれにおいても，「(1) 日本側でのみ観測した内容」，および，「(2) 中国側でのみ観測した内容」が一定数観測されていることが分かる．

次に，話題「犬に噛まれた」について，上述の (1)∼(3) の内容の詳細，および，該当する質問・回答組数を表 4.4 に示す．表中では，同一の内容について 6 件以上の質問・回答組数が観測できた場合に，その内容を**太字**で示す．この結果から分かるように，日中両側で観測された内容「狂犬病ワクチンの接種を勧める」，および，「狂犬病の基礎知識を説明し，必要があればワクチン接種を勧める」に該当する件数が，中国側では圧倒的多数を占めるのに対して，日本側では 1 件のみとなっており，日中間の関心の度合いの違いが，狂犬病の危険度の違いを如実に反映した結果となった．同様に，中国側でのみ観測された内容としては，「噛んだ犬の様子を観察し狂犬病の犬かどうか判断することを勧める」があり，しかもその件数は圧倒的多数となった．一方，日本側では，狂犬病以外の「法律関係の事項」，「飼い主の責任の所在」，「近所との人間関係」，「ペットと飼い主の関係」，といった観点での質問・回答組が多数を占めることが分かる．

同様に，話題「刺身と寄生虫」について，上述の (1)~(3) の内容のうち主なものの抜粋を表 4.5 に示す．この結果から分かるように，日本では，通常家庭や外食店で食べられる刺身については寄生虫の心配はないが，魚種によっては寄生虫が危険な種類もあり，専門的な知識が必要である，といった回答が多数派を占め，社会全体が十分に管理されているという印象を持つが，中国では,「生の魚は寄生虫が危険である」という回答が多数派を占めるという結果であった．

また，話題「喫煙」についても，同様の抜粋を表 4.6 に示す．

関連研究

文献 [21] の先行研究として，文献 [20] においては，ある話題について，日本語と英語のブログ記事を収集し，人々の関心事項や意見などに関する文化間差異発見過程を支援する方式を提案した．この手法においては,「捕鯨」や「臓器移植」など，日本と欧米の社会制度上の違いや食文化の違いが大きい話題について，関心の差異を観測できている．一方，文献 [27] においては，ある話題に関する日英ブログ記事集合において，日本語と英語の観点を比較対照分析する手法を提案している．

また，複数情報源からのニュースの言語間差異分析に関する研究として，文献 [28, 26, 29, 2] が挙げられる．文献 [28] は，32 言語における千件以上の伝染病に関するレポートを分析し，まとめあげる研究を行っている．文献 [26] では，32 言語のニュース記事集合から特定の人物名を収集し，その人物の人間関係やその人物に関連する各国のニュース記事を継続的に解析する研究を行っている．文献 [29] は，複数の国のメディアが発信するニュースに基づいて，同一事象に対する各国のニュースの伝え方の違いを分析する手法を提案している．文献 [2] では，9 言語間における同一事象に対する主観情報の差異分析の研究を行っている．その他，文献 [7, 6, 31] では，日中の時系列ニュースを情報源として，時系列トピックモデルによって得られる日中単言語のトピックの間の言語間対応をとることにより，同じ話題に関するニュース記事の集合を持つ日中各言語のトピックを同定する方式を提案している．これらの研究は主にニュース記事を対象に分析を行っている点で文献 [21] とは異なる．

まとめ

文献 [21] では，日本語と中国語の質問回答サイトから，ある同一の問題についての日本語と中国語の質問と回答を収集し，回答中の記述を対象として，日本と中国の間の文化的対照性を分析する方法を提案した．今後は，特定の話題について収集した日中両言語の質問・回答組の間で，専門的内容を表す用語の対訳関係

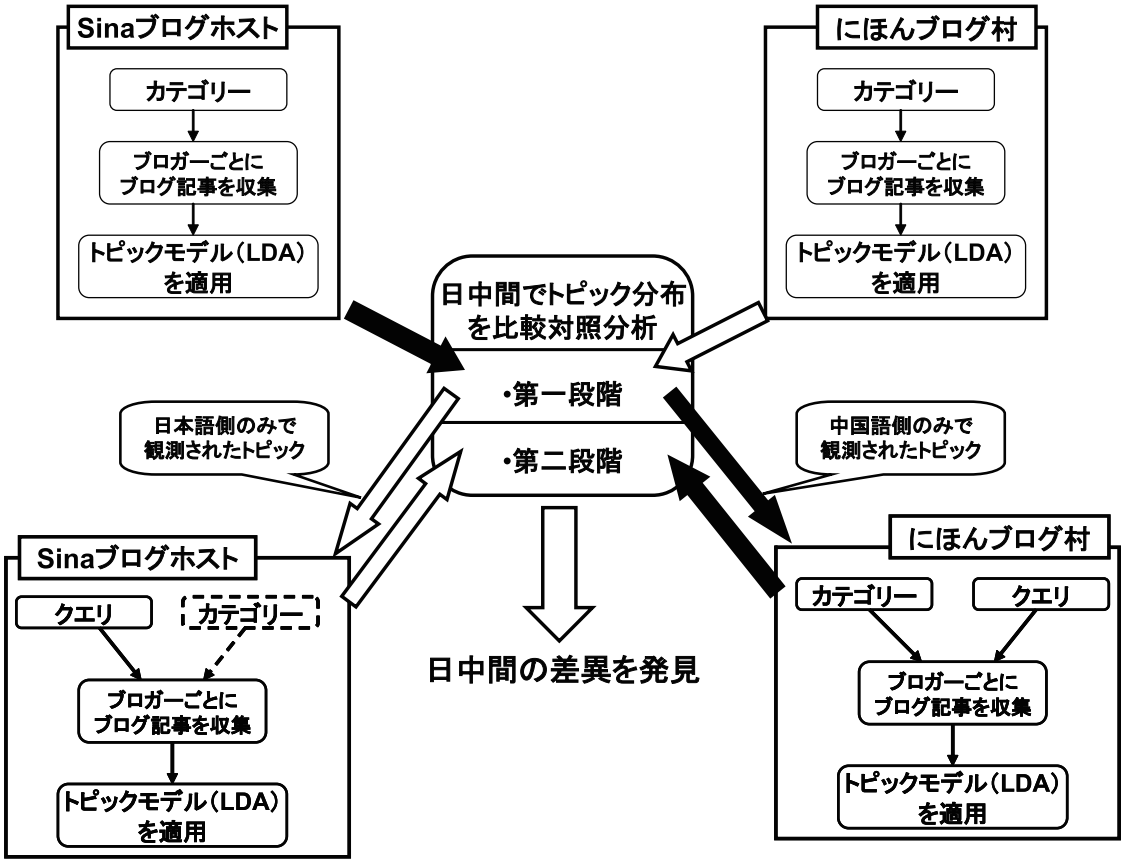


図 4.2: 日中ブロガー・コミュニティの比較対照分析 (文献 [30] より引用)

を利用することにより，日中間の文化間差異の有無を自動判定する方式について研究を行う．

4.2.2 日中ブロガー・コミュニティの収集・俯瞰・対照分析

以下では，ウェブ執筆者に着目した文化間差異発見支援の研究の例として，文献 [30] について説明する．文献 [30] では，日本のブロガー・コミュニティと中国のブロガー・コミュニティを対象として，トピックモデルを用いた日中ブロガー・コミュニティの比較対照分析を行う手法を提案した．

文献 [30] における日中ブロガー・コミュニティの比較対照分析の枠組みを図 4.2 に示す．

第一段階

1. まず，日本語ブログ記事の収集，および，中国語ブログ記事の収集を行う．日本語側においては，日本最大級のブロガー・コミュニティ「にほんブログ村」から，各分析対象となるカテゴリーに属する最大300ブロガーを選定し，

表 4.7: 「Sinaブログホスト」および「にほんブログ村」のカテゴリー数・ブロガー数 (文献 [30] より引用)

	カテゴリー数	サブカテゴリー数	ブロガー数
にほんブログ村	121	約 5,500	681,041
Sina ブログホスト	23	—	12,570

一ブロガーにつき，最新の 50 記事を収集する．中国語側においては，中国最大級の大手メディア運営会社「Sina」が提供しているブログサービス「Sinaブログホスト」から，各分析対象となるカテゴリーに属するブロガーを選定し，一ブロガーにつき，最新の 50 記事を収集する．

2. そして，収集された日本語ブログ記事集合，および，中国語ブログ記事集合に対して，LDA トピックモデルを適用し，日本と中国それぞれのブロガー・コミュニティを生成する．その後，手動で生成された日中ブロガー・コミュニティを比較し，中国語側でのみ観測したブロガー・コミュニティと，日本語側でのみ観測したブロガー・コミュニティを選定する．

第二段階

1. 第二段階では，方言語側でのみ観測されたブロガー・コミュニティが本当に相手言語側に存在しないかどうかの検証を行う．中国語側でのみ観測されたブロガー・コミュニティに対して,「にほんブログ村」のカテゴリー情報を用いた検証と，Yahoo! Search BOSS API[4] を用いた検証の二種類の検証を行う．まず，中国語側でのみ観測されたブロガー・コミュニティに含まれる話題と「にほんブログ村」のカテゴリー情報と照合し，対応するカテゴリーの情報がある場合は，第一段階で行ったブログ記事の収集の手順に沿って,「にほんブログ村」からブログ記事を再収集し，その後，トピックモデルによって新たなブロガー・コミュニティを生成する．また，対応するカテゴリーの情報がない場合は，Yahoo! Search BOSS API を用いてブログ記事を収集し，トピックモデルによって新たなブロガー・コミュニティを生成する．

2. 「Sina ブログ」では，有用なカテゴリー情報を新たに利用する可能性が低いため，日本語側でのみ観測されたブロガー・コミュニティについては，Yahoo! Search BOSS API を用いた検証のみで検証を行う．中国語側と同じように，Yahoo! Search BOSS API でブログ記事を収集した後，トピックモデルを適用し，新たなブロガー・コミュニティを生成する．

[4] http://developer.yahoo.com/search/boss

表 4.8: 分析対象ブロガー数およびブログ記事数 (文献 [30] より引用)

	カテゴリー	ブロガー数	ブログ記事数
日本語	「健康」	300	9,380
中国語		268	7,708
日本語	「軍事」	23	966
中国語		124	1,868
日本語	「介護」	300	7,253
中国語		—	—

「にほんブログ村」および「Sina ブログホスト」の詳細を表 4.7 に示す.「にほんブログ村」の登録ブロガー数は約 68 万人で，各ブロガーは，121 のカテゴリー，および，約 5,500 のサブカテゴリーに登録されている．一方,「Sina ブログホスト」の登録ブロガー数は約 2.5 億人で，そのうち，閲覧可能な 12,570 人の人気ブロガーが 23 個のカテゴリーに登録されている.

また，文献 [30] の分析対象ブロガー数とブログ記事数の詳細を表 4.8 に示す.「健康」,「介護」,「軍事・防衛」を分析対象カテゴリーとする.「健康」カテゴリーにおいては，日本語側の 300 ブロガーに対して 9,380 記事を収集し，中国語側の 268 ブロガーに対して 7,708 記事を収集した.「軍事」カテゴリーにおいても同様に，日本語側の 23 ブロガーに対して 966 記事を収集し，中国語側の 124 ブロガーに対して 1,868 記事を収集した．また,「介護」カテゴリーについては，中国語側ではブログ記事が正しく収集されなかったため，日本語側でのみブログ記事の収集を行った. 300 ブロガーに対して 7,253 件のブログ記事を収集した.

文献 [30] における日中比較対照分析の結果について，図 4.2 の「第一段階」の日中比較対照分析の結果の例を表 4.9，表 4.10，表 4.11 に示す．また，図 4.2 の「第二段階」の日中比較対照分析の結果の例を表 4.12 で示す.

まず，図 4.2 の「第一段階」の日中比較対照分析の結果のうち，カテゴリー「健康」において日中共通に観測された話題のコミュニティについて説明する．表 4.9 に示すように,「飲食と健康」,「歯の健康」および,「健康についての研究」などの健康との関連が強いブロガー・コミュニティが日中両側で観測された．一方，表 4.10 に示すように，図 4.2 の「第一段階」の日中比較対照分析の結果のうち，カテゴリー「健康」において日本語側でのみ観測された話題のコミュニティの例として，治療法に関するブロガー・コミュニティ「育毛」や「レイキ療法」などが観測された．また，表 4.11 に示すように，図 4.2 の「第一段階」の日中比較対照分析の結果のうち，カテゴリー「健康」において中国語側でのみ観測された話題のコミュニティの例として,「二十四節気と健康」が観測された．また，中国伝統文化に関連する「太極拳の健康価値」などのブロガー・コミュニティも観測された.

表 4.9: 日中共通に観測された話題のコミュニティ (「健康」カテゴリー，図 4.2 の「第一段階」の日中比較対照分析後)(文献 [30] より引用)

話題の大分類	「にほんブログ村」		「Sinaブログホスト」	
	ブロガー・コミュニティの話題	ブロガー数	ブロガー・コミュニティの話題（日本語訳）	ブロガー数
病気全般	病気全般、種々の病気について	5	各种疾病（病気全般）	6
美容商品	美容商品、アロマに関する話題	9	美容产品（美容商品）	13
こころの健康	臨床心理士のカウンセラー、レイキヒーリングについて	4	心理健康（心の健康）（2トピック）	18
	ヒーリング能力伝授について	6		
	心理カウンセリングの話題、カウンセリングの講座	2	心理学的理论知识（心理学の知識）	5
	心の健康	22		
飲食と健康	健康な食品に関する紹介、食品と栄養について	12	健康的饮食（飲食健康）（2トピック）	25
	糖砂栄養素の健康効果	2		
	食事療法、食育、レシピ紹介	12	饮食和营养（飲食と栄養）	21
	薬膳、薬膳茶の紹介	2	健康的食谱（健康な飲食レシピ）	15
歯の健康	歯科矯正	2	牙齿的健康(歯の健康)	6
筋力トレーニング	体操、ハピトレなどの筋力トレーニング（2トピック）	20	运动和健康(運動と健康)	3
整体	整体院のお知らせ	6	鼻などの整容(整容)	6
			整形(整形)	4
健康についての研究	免疫力、糖鎖、サプリメント、アンチエイジングについての健康に関する研究	9	关于健康的研究（健康に関する研究）	8
	睡眠、運動、血液などの健康情報	13		

表 4.10: 日本語側でのみ観測された話題のコミュニティ (「健康」カテゴリー, 図 4.2 の「第一段階」の日中比較対照分析後)(文献 [30] より引用)

話題の大分類	「にほんブログ村」	
	ブロガー・コミュニティの話題	ブロガー数
色彩心理テスト	カラーセラピストによる色に関する話題	12
	オーラソーマに関する知識と体験	6
	フラワーセッションセラピー	4
育毛	育毛のためのシャンプー、育毛剤、植毛	12
眼科	レーシックなどの角膜手術	6
寒さ対策	寒さ対策:厚い衣類と靴下	8
足の健康	インソールと足の健康	4
健康講座	健康に関する各種講座	4
禁煙	禁煙に関する話題、体験	5
旅行体験	旅行の体験記、観光日記	7
ヨガ	ヨガインストラクター、ヨガレッスンについて	9
鍼灸、指圧	鍼灸に関しての話題	10
	指圧講座・指圧治療の紹介	3
神秘な治療法	レイキに関する交流会、セミナー	5
	病腺霊感法などの触手療法	3
	真氣光というスピリチュアル系の内容	4
	気功に関する内容、神秘系、スピリチュアルな内容の気功	2
日常記録	日常の出来事 (2トピック)	20

表 4.11: 中国語側でのみ観測された話題のコミュニティ (「健康」カテゴリー, 図 4.2 の「第一段階」の日中比較対照分析後)(文献 [30] より引用)

話題の大分類	「Sinaブログホスト」	
	ブロガー・コミュニティの話題 (日本語訳)	ブロガー数
性教育	性教育(性教育)	14
各種病気	男性病的预防与治疗 (男性病の予防と治療法)	9
	女性病的预防与治疗 (女性病の予防と治療法)	20
	关于糖尿病 (糖尿病に関して)	8
	高血压，心脏病的注意事项 (高血圧、心臓病の注意事項)	2
	癌症的预防与早期癌症的发现 (ガンの予防と早期発見)	5
	关于肝炎 (肝炎について)	3
二十四節気と健康	二十四节气相对应的健康注意事项 (二十四節気の各節気での健康に関する注意事項) (2トピック)	4
製薬企業	社会和制药业 (社会と製薬企業)	9
医者と患者	对于医疗制度的批判和探讨 (医療制度に対する批判と検討)	9
	关于医疗政策的探讨 (医療政策についての検討)	2
	医患关系 (医患関係)	2
経済	房地产,股票等 (不動産、株など)	6
健康(中国固有)	道德健康学 (道徳健康学)	3
	中医与民间土方 (漢方医学と中国民間療法)	9
	经脉和太极拳的健康价值 (ツボ、太極拳の健康価値)	5
子供の健康	儿童健康 (子供の健康)	5

表 4.12: 「健康」カテゴリーのブロガーから生成されたコミュニティにおける日中間差異の例 (図 4.2 の「第二段階」の日中比較対照分析後)(文献 [30] より引用)

	話題	日本語コミュニティ		中国語コミュニティ	
		コミュニティ収集手順	話題	コミュニティ収集手順	話題
中国語側のみで観測	**「二十四节气与健康」（二十四節気と健康）**	「二十四節気 and 健康」をクエリとして、ブログ記事を検索したが、収集されなかった		「健康」カテゴリーのブロガーから生成されたコミュニティの一つとして存在	1年間における各節気での健康に関する注意事項
日本語側のみで観測	**「育毛」**	「健康」カテゴリーのブロガーから生成されたコミュニティの一つとして存在	・日中共通の治療方法（育毛剤） ・日本でのみ観測された治療方法（植毛）	「生发」（育毛）、「生发剂」（育毛剤）、「秃头」（ハゲ）をクエリとして、ブログ記事を収集して、トピックモデルを適用し、「育毛」に関するコミュニティを生成	・中国でのみ観測された治療方法：自然療法（例：生姜を頭皮に塗る）
	「レイキ療法」	「健康」カテゴリーのブロガーから生成されたコミュニティの一つとして存在	代替医療であり、一種の手当て療法である（日本の民間療法）	「灵气疗法」（レイキ療法）、「灵气」（レイキ）、「灵能力」（ヒーリング）、「苏摩」（ソーマ）をクエリとして、ブログ記事を収集して、トピックモデルを適用し、「レーキ療法」に関するコミュニティが生成されなかった	

最後に，図 4.2 の「第二段階」の日中比較対照分析の結果の例を表 4.12 に示す.

図 4.2 の「第一段階」において中国語側でのみ観測されたブロガー・コミュニティ「二十四節気と健康」については，このブロガー・コミュニティに対して Yahoo! Search BOSS API を用いて検証を行ったが，日本語のブログ記事を収集できなかった．Google 検索エンジンで同様な検証を行った結果,「二十四節気と健康」に関連するウェブページは収集できたが,「二十四節気と健康」について書かれたブログ記事は収集できなかった．これは,「二十四節気と健康の間に関連性がある」という認識が日本ではまだ広まっていないからであると考えられる.

図 4.2 の「第一段階」において日本語側でのみ観測されたブロガー・コミュニティ「育毛」については,「第二段階」の検証において,「育毛」に関連するクエリの中国語訳を用いてブログ記事を収集し，新たなブロガー・コミュニティを生成した．そこで，日中間比較対照分析を行った結果，日中両側において観測されたブロガー・コミュニティとして,「育毛剤を用いた治療」が観測された．一方，日中のうちの方言語側でのみ観測されたブロガー・コミュニティとして，日本語側では「植毛」があり，中国語側では「生姜を頭皮に塗る等の自然療法」が観測された．このように，日中間の差異を発見できた．またブロガー・コミュニティ「レイキ療法」に関しては，中国語ブログ記事の収集を試みたが，ブロガー・コミュニティを生成することができなかった．この結果から,「レイキ療法」は日本に特有の民間療法であることが分かった.

4.3 ウェブ検索者に着目した研究

4.3.1 検索エンジン・サジェストの日中間比較対照分析 (トピックモデルを使用しない場合)

近年，世界のグローバル化が進展するのに伴って，国内外の様々な局面において，異文化間での交流機会が増大している．そして，異文化間の交流機会が増大するのと同時に，様々な問題も併せて引き起こしている．このグローバル社会における出来事の根底には，各国特有の歴史的背景や文化的特異性が根強く横たわっている．文献 [23] においては，各国特有の歴史的背景や文化的特異性を同定するために，ウェブ上の情報を多言語 (日本語・中国語) 間で比較・対照分析することにより，言語間の差異を発見するというアプローチをとる．特に，文献 [23] では，ウェブ検索者の関心動向に着目し，研究を行った．

ウェブ上の情報の一例として，近年，一般個人が自由に情報を発信するツールであるブログが世界中で普及し，各地域の人々がそれぞれインターネット上で個人の意見や評判を発信することが可能になった．ここで，文献 [30] においては，ウェブ執筆者の関心動向を収集するための情報源として，日中ブログを用いて，国・文化・言語間の差異発見過程を支援する方式を提案している．しかし,「尖閣諸島」等の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，文献 [23] では，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する．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，ブログにおける言及数を情報源とする場合の遅延を克服でき，関心動向を迅速に把握する．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，ブロガー等のウェブ執筆者の関心動向を直接的に収集するアプローチ (例えば，文献 [30]) では，執筆内容を収集し関心動向を集約・同定するまでの間に膨大な計算を必要とするため，収集可能なウェブ執筆者の関心動向の範囲が制限される点に問題があった．これに対して，ウェブ検索者の情報要求観点を情報源とする場合には，情報要求観点そのものを直接的に収集することができるため，相対的に網羅性の高い関心動向収集が可能となる．そこで，文献 [23] では，日中検索エンジン・サジェストを情報源として，ウェブ検索者の情報要求観点を収集し，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方式を提案する．検索エンジン・サジェストから情報要求観点を収集する手順，および，情報要求観点と検索結果のウェブページ中の記述内容を日中間で比較し，対照分析する手順の概要を図 4.3 に示す．

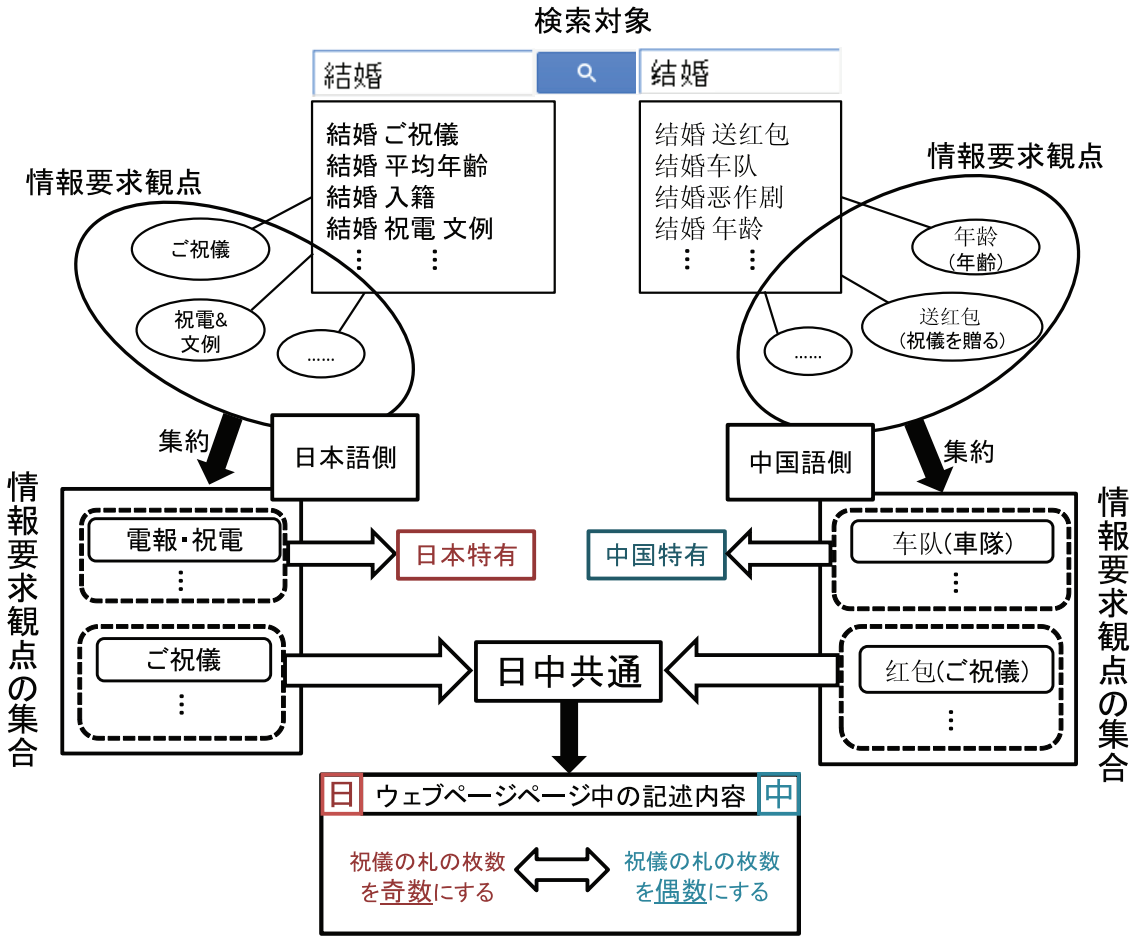


図 4.3: 検索エンジン・サジェストからの情報要求観点の収集および日中間比較対照分析

検索エンジン・サジェストからの情報要求観点の収集

検索エンジン・サジェスト

各検索エンジン会社においては，ウェブ検索者の検索ログが蓄積されており，多数のウェブ検索者が検索したキーワードに対して，検索者が強い関心を持つ語を抽出し，検索エンジン・サジェストとして提示するサービスを提供している．ここで，文献 [23] では，詳細な情報を検索したい対象を**「検索対象」**と呼ぶ．また，検索対象に対して，検索者が AND 検索の形で二つ目以降のキーワードとして指定し，検索対象に対して詳細な情報を得るために用いる観点を**「情報要求観点」**と呼ぶ．

すると，検索エンジン・サジェストとして提示される言葉は,「検索対象」に対して，多数のウェブ検索者が「情報要求観点」として指定した語に相当しており，ウェブ検索者の関心事項そのものを反映していることが分かる．そこで，本節では，検索エンジン・サジェストに着目することによって，ウェブ検索者に焦点を当て，情報要求観点の収集を行う．

日本語側の収集手順

文献 [23] では，検索対象「結婚」に着目し，Google 検索エンジン [5] に対して，一検索対象当り約 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．100 通りの文字列とは具体的には，五十音，濁音，半濁音および「きゃ」や「びゃ」などの開拗音である．例えば検索窓に「結婚 みょ」と入力すると,「苗字」や「苗字 変更」などがサジェストとして掲示されるので，それらを収集することにより，919 個の情報要求観点を収集した．

中国語側の収集手順

文献 [23] では，検索対象「结婚（結婚）」に着目し，Google 検索エンジンに対して，一検索対象当り 28 通りの文字列を指定し，最大 280 語のサジェストを収集する．28 通りの文字列とは具体的には，中国語のピン音の部首である．例えば検索窓に「结婚（結婚）h」と入力すると,「送红包(**祝儀を贈る**)」などがサジェストとして掲示されるので，それらを収集することにより，248 個の情報要求観点を収集した．

[5] https://www.google.com/

情報要求観点の日中間対照分析

情報要求観点の集約

本節では，4.3.1 節において収集した情報要求観点を人手で集約し，話題ごとにまとめる．その結果,「結婚」においては，日本語側の 919 個の情報要求観点は 148 個の集合に集約された．一方，中国語側の 248 個の情報要求観点は 42 個の集合に集約された．

集約後の情報要求観点集合の対照分析

本節では，前節で集約した日本語と中国語の情報要求観点集合を対象として，日中間で比較対照分析を行う．図 4.3 に示すように，日本語特有の情報要求観点集合の同定，中国語特有の情報要求観点集合の同定，および，日中共通の情報要求観点集合の対応付けを行う．

その結果，日中共通の情報要求観点集合は，

ご祝儀関連　　結婚式のマナー関連
結婚の手続き関連　　結婚年齢関連

などの 23 個の対応組であった．日本語特有の情報要求観点集合としては，表 4.13 に示すように,「電報・祝電」(電報の書き方),「入籍について」(入籍の仕方)，および,「苗字について」(苗字変更関連) などがあった．一方，中国語特有の情報要求観点集合としては，表 4.14 に示すように,「車隊」(中国語では，立派な結婚式には車隊が必要),「婚前の健康診断」(中国では一般的である婚前健康診断の手順と項目)，および,「婚礼でのゲーム」(中国のみでの慣習) などがあった．

情報要求観点の比較対照分析

本節では，前節で得られた日中共通の情報要求観点集合を対象として，集合中の情報要求観点を日中間で比較対照分析する．検索対象「結婚」の分析結果のうち，情報要求観点集合として「ご祝儀関連」および「結婚年齢関連」についての分析結果を表 4.15 の「検索対象＋情報要求観点」の欄に示す．

例えば，表 4.15 の「結婚年齢関連」において，日中共通で観測された情報要求観点として，日本語側の「検索対象＋情報要求観点」

結婚 何歳　　結婚 何歳から

と中国語側の「検索対象＋情報要求観点」

表 4.13: 「結婚」において日本語側のみで観測された情報要求観点集合およびウェブページ中の記述内容の抜粋

電報・祝電	
検索対象＋情報要求観点	ウェブページ中の記述内容
結婚&電報	電報の書き方
結婚&電報&メッセージ	
結婚&電報&文例	
結婚&祝電&文例	
結婚&祝電	
入籍について	
検索対象＋情報要求観点	ウェブページ中の記述内容
結婚&入籍日	入籍と結婚式はどちらが先
結婚&入籍	
結婚&入籍&順番	
結婚&入籍&流れ	入籍の仕方
結婚&入籍&手続き	
苗字について	
検索対象＋情報要求観点	ウェブページ中の記述内容
結婚&苗字&仕事	結婚後、会社で名乗るのは新姓かどうか
結婚&苗字&会社	
結婚&苗字	苗字変更に伴う手続き
結婚&苗字&変更	
結婚&苗字&同じ	同じ苗字同士が結婚した場合について

表 4.14: 「結婚」において中国語側のみで観測された情報要求観点集合およびウェブページ中の記述内容の抜粋

車隊	
検索対象＋情報要求観点	ウェブページ中の記述内容
结婚车队 （結婚車隊）	中国では、立派な結婚式には車隊が必要
婚前の健康診断	
検索対象＋情報要求観点	ウェブページ中の記述内容
结婚体检 （婚前の健康診断）	（中国では一般的である） 婚前健康診断の手順と項目
结婚体检项目 （婚前健康診断の項目）	
婚礼でのゲーム	
検索対象＋情報要求観点	ウェブページ中の記述内容
结婚游戏 （結婚式でのゲーム）	婚礼で、新郎新婦が しなければならないゲーム
结婚开门游戏 （結婚式でのゲーム）	
结婚恶作剧 （結婚式でのゲーム）	
结婚闹新房节目 （新居でのゲーム）	新居で、新郎新婦が しなければならないゲーム

结婚年龄规定(**法律上の婚姻適齢**)
结婚法定年龄(**法律上の婚姻適齢**)
结婚年龄限制(**法律上の婚姻適齢**)

はほぼ同一の内容に対応するので,「日中共通で観測」として，日中間の対応を付ける．表 4.15 の「結婚年齢関連」において,「ウェブページ中の記述内容」の欄に示すように，実際に，これらの情報要求観点に対して検索されるウェブページからは，日中間の法律上の差異を容易に発見できることが分かる．

一方，中国語側のみで観測された情報要求観点としては，

结婚年龄测试(**結婚年齢予測の心理テスト**)

があり，これらの情報要求観点に対して検索されるウェブページからは,「心理テストで，結婚年齢を予測する」という中国特有の情報が得られる．

日中共通の情報要求観点によって収集されたウェブページ中の記述内容の比較対照分析

本節では，前節で得られた日中共通の情報要求観点を対象として，それらの情報要求観点に対して検索されるウェブページ中の記述内容を日中間比較対照分析し，日中間差異の有無についての検証を行う．

例えば，表 4.15 の「ご祝儀関連」の「日中共通で観測」の欄に示すように，日中共通の情報要求観点に対して検索されるウェブページ中の記述内容からは，日中両言語において「祝儀袋の書き方，ご祝儀の金額の目安」という情報が得られることが分かる．一方，日本語側特有の記述内容として,「ご祝儀の札の枚数を奇数にする」があり，中国語側特有の記述内容として,「祝儀の札の枚数を偶数にする」がある．このように，提案方式によって，日中間の慣習の差異の発見の過程を効果的に支援できることが分かった．

関連研究

文献 [30] においては，特定の話題について，日本語ブログ記事，および，中国語ブログ記事を収集し，国・文化・言語間の差異発見過程を支援する方式を提案している．この方式では,「健康」や「軍事」など，日本と中国との間で習慣の違いや主張の差異が大きい話題について，ブログ空間における国・文化・言語間の違いを観測している．一方，文献 [27] においては，特定の話題に関するブログ記事集合において，日本語・英語二言語での観点を分類・比較・対照分析する手法が提案されている．また，文献 [21] においては，日中質問回答サイトを対象として，トラブル情報の比較対照分析を行い，文化間差異発見支援を行う方式を提案

表 4.15: 「結婚」の情報要求観点およびウェブページ中の記述内容の日中間比較対照分析の抜粋

ご祝儀関連						
検索対象＋情報要求観点				ウェブページ中の記述内容		
	日本語側		中国語側	日中共通の内容	日本語側独特の内容	中国語側独特の内容
日中共通で観測	結婚&ご祝儀	<=>	结婚&送红包（結婚 AND 祝儀）	祝儀袋の書き方、ご祝儀の金額の目安	ご祝儀の札の枚数：偶数は割り切れてしまうため、縁起が悪いとされるので、枚数を奇数にする	ご祝儀の札の枚数：夫婦二人が対になるため、枚数を偶数にする
	結婚&ご祝儀&親族		结婚红包上怎么写（祝儀袋の書き方）			
	結婚&ご祝儀&相場		结婚红包怎么写（祝儀袋の書き方）			
	結婚&ご祝儀&兄弟		结婚红包送多少（ご祝儀の金額）			
日本語側のみで観測	結婚&ご祝儀&お返し				ご祝儀のお返しのマナー	
	結婚&二次会&祝儀				二次会のご祝儀のマナー	
中国語側のみで観測						

結婚年齢関連						
検索対象＋情報要求観点				ウェブページ中の記述内容		
	日本語側		中国語側	日中共通の内容	日本語側独特の内容	中国語側独特の内容
日中共通で観測	結婚&何歳	<=>	结婚年龄规定（法律上の婚姻適齢）		日本の法律においては 男子：18歳以上 女子：16歳以上	中国の法律においては 男子：22歳以上 女子：20歳以上
	結婚&何歳から		结婚法定年龄（法律上の婚姻適齢）			
			结婚年龄限制（法律上の婚姻適齢）			
日本語側のみで観測						
中国語側のみで観測			结婚年龄测试（結婚年齢予測の心理テスト）			心理テストで、結婚年齢を予測する

している．ただし，これらのブログおよび質問回答サイトを対象とした研究においては，トピックモデルによって話題のまとまりを同定する過程が欠如しており，比較的小規模な文書集合を対象とした人手による分析に重点が置かれている点が，文献 [23] とは大きく異なる．

一方，複数情報源からのニュースの多言語間差異分析を行っている研究として，文献 [28, 26, 29, 2] が挙げられる．文献 [28] は，32 言語における 1,000 以上の情報源を分析し伝染病に関するレポートをまとめあげる研究を行っている．文献 [26] では，32 言語におけるニュース記事群から特定の人物名を収集し，その人物の人間関係やその人物について言及している各国のニュース記事を継続的に分析する研究を行っている．文献 [29] は，複数の国の代表的なメディアが発信するニュースを情報源として，同一事象に対する各国のニュースの伝え方の差異分析方式を提案している．文献 [2] では，9 言語間における同一事象に対する主観情報の差異分析の研究を行っている．これらの研究は主にニュース記事を対象に分析を行っている点で本論文とは異なる．

まとめ

文献 [23] では，Google 検索エンジンに対して，日本語および中国語の検索対象についての情報要求観点を収集し，日中二言語間で，情報要求観点の比較対照分析を行う方式を提案し，その適用事例について報告した．

今後の課題として，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際に，分析者のコストを削減するため，まず，検索エンジン・サジェストを自動分類する方式 [10] を導入することが必要である．また，Wikipedia 等を情報源とする日中対訳知識を利用して日中間の情報要求観点の対応付けを自動的に行う手法を開発しこれを導入することにより，人手で日中間の情報要求観点の対応付けを行う際のコストを削減することが必要である．

第5章　検索エンジン・サジェストの日中間比較対照分析

本章では，より細かい話題の粒度で日中文化間差異の発見を支援するために，第 3 章の集約結果に対して，粗い粒度 (LDA トピックモデルの集約結果)，および，細かい粒度 (分散表現による LDA 集約結果の細分化) の二種類の話題の粒度のもとで，日本語・中国語二言語間の比較対照分析を行う．本章における分析対象のクエリ・フォーカスとしては,「結婚」および「就活」を用いる．

a) まず，粗い話題の粒度で日中文化間差異を分析するために，LDA トピックモデルによる一次集約結果に対して，日中文化間差異の分析を行う．具体的には，LDA トピックモデルの集約結果に対して，各トピックに属するサジェスト，および，ウェブページの内容を分析することにより，日中トピックを「日中共通で観測したトピック」,「日本語のみで観測したトピック」,「中国語のみで観測したトピック」の三種類に分類する．この手順により，日中間の話題の違いをトピック単位で比較することによって，粗い話題の粒度で日中間の文化間差異を発見できる．

b) 次に，より細かい話題の粒度で日中間差異を分析するために，分散表現を用いる．具体的には，日中それぞれの言語の Wikipedia の本文全ページ，および，クエリ・フォーカス「就活」を用いて収集したウェブページ集合の混合文書集合を訓練用コーパスとして word2vec を適用し，分散表現を得る．得られた分散表現を用いて，LDA トピックモデルによる一次集約結果を細分化し，各トピックにおける小分類を抽出する．a) で得た「日中共通で観測したトピック」に対して，その中に含まれる小分類の比較対照分析を行い,「日中共通で観測した小分類」,「日本のみで観測した小分類」,「中国のみで観測した小分類」の三種類に分類する．この手順により，日中間の話題の違いをトピック内の小分類単位で比較することによって，細かい話題の粒度で日中間の文化間差異を発見できる．

5.1 トピック単位でのサジェストの日中間比較対照分析

ウェブ執筆者に着目した日中文化間差異の発見を支援する研究として，文献 [23] が挙げられる．この文献においては，検索者観点に着目し，日本と中国の検索エンジン・サジェストを収集し，人手で比較対照分析することによって日中間文化間差異を発見する手法を提案した．しかし，文献 [23] では，人手で集約作業を行っていたため，相当なコストが必要であった．この問題を解決するために，文献 [4] では，ウェブ検索者の関心事項に着目し，トピックモデルを用いてサジェスト集合をクラスタリングを行った上で，二言語間で比較対照分析する手法を提案した．

以下では，トピックモデルを用いることにより，トピック単位での日中間比較対照分析の流れを示す．本論文と文献 [4] の共通部分である，トピック単位での日中比較対照分析の枠組は，図 5.1 に示すように，大きく以下の 3 つに分けられる．

1. データの収集: まず，日本および中国のサジェストの収集を行う．日本語側は，平仮名などを用いて，Google 検索エンジン [1] を情報源として，1 クエリ・フォーカスにつき約 1,000 個のサジェストを収集できる．中国語側は，pinyin を用いて，Baidu 検索エンジン [2] を情報源として，1 クエリ・フォーカスにつき約 600 個のサジェストを収集できる．そして，収集されたサジェストを用いて，検索エンジンでクエリ・フォーカス＋サジェストの AND 形式で検索を行い，Google Custom Search API[3] を用いて上位 20 件の日本語のウェブページを，また，Baidu 検索エンジンを用いて上位 40 件の中国語のウェブページを，それぞれ収集する．

2. トピックモデルによる集約: 収集された日本語のウェブページ集合と中国語のウェブページ集合に対して，LDA トピックモデルを適用し，数十個のトピックへ集約する．各ウェブページには一つ以上の検索語エンジン・サジェストが付与されるため，ウェブページを集約することにより，サジェストも自動的に数十個のトピックへ集約される．

3. 日中間比較対照分析: 集約結果に対して，各トピックに分類されたウェブページおよびサジェストを分析し，日中間比較対照分析を行う．人手で日中の間で比較対照分析を行い，日中トピックを「日本のみで観測したトピック」，「中国のみで観測したトピック」,「日中共通で観測したトピック」の三種類に分類する．

[1] https://www.google.com/
[2] https://www.baidu.com/
[3] https://cse.google.com/cse/

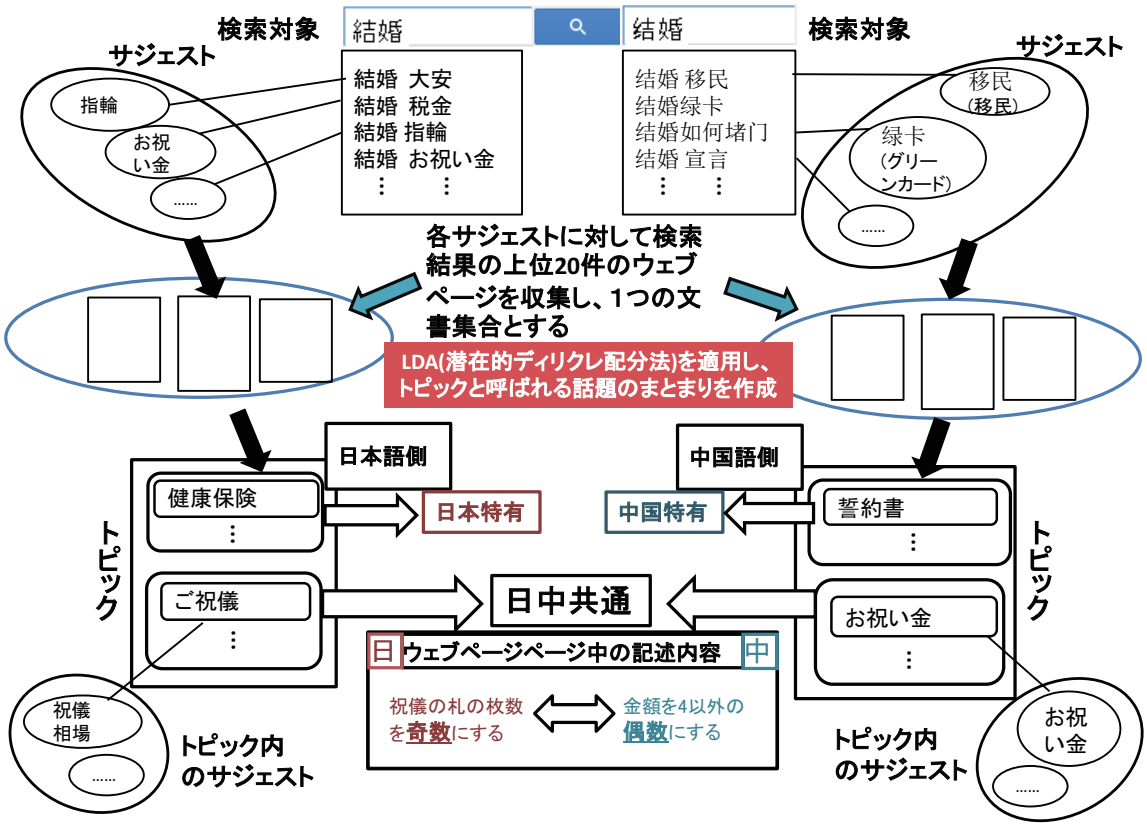


図 5.1: トピックモデルを用いた検索エンジン・サジェストの日中間対照分析 (文献 [4] より引用)

表 5.1: 「結婚」において日中両側で観測されたトピックおよびウェブページ中の記述内容 (抜粋)(文献 [4] より引用)

共通のトピック ・・・ 日本語側：国際結婚の際に必要となる手続き　中国語側：国際結婚(配偶者の国へ移民)						
トピックに対応付けられた検索対象「結婚」+サジェスト				トピック内のサジェストに対応したウェブページの記述内容		
	日本語側		中国語側	日中共通の内容	日本語側独自の内容	中国側独自の内容
日中共通で観測	結婚 グリーンカード	<=>	结婚 绿卡(結婚 グリーンカード)	アメリカ人と結婚する際のグリーンカードの申請方法について	ハワイでのグリーンカード申請の体験談	アメリカ人との偽装結婚でグリーンカードの申請について
日本語側のみで観測	結婚 ミャンマー人 結婚 ミャンマー　手続き 結婚 ミャンマー女性				ミャンマー人との結婚について	
中国語側のみで観測			结婚 移民(結婚 移民) 结婚移民 美国(結婚 移民アメリカ) 结婚移民加拿大(結婚 移民カナダ)			アメリカへ、カナダへの移民について

共通のトピック ・・・ 日本語側：ご祝儀について　中国語側：お祝い金						
トピックに対応付けられた検索対象「結婚」+サジェスト				トピック内のサジェストに対応したウェブページの記述内容		
	日本語側		中国語側	日中共通の内容	日本語側独自の内容	中国側独自の内容
日中共通で観測	結婚 祝儀 相場 結婚 お祝い金	<=>	结婚 礼金(結婚 祝い金) 结婚 份子钱(結婚 祝い金)	新郎・新婦との関係によって金額の目安が異なる	昔は、偶数は割り切れてしまい、縁起が悪いとされたため、札の枚数は奇数がよい。	金額を4以外の偶数にするのが一般的
日本語側のみで観測	結婚 ご祝儀 包み方 結婚 祝儀袋 書き方 結婚 ご祝儀袋				結婚 ご祝儀袋について	
中国語側のみで観測			结婚 人情(結婚 義理のつきあい)			結婚に伴うつきあいが面倒くさい

クエリ・フォーカス「結婚」の例

クエリ・フォーカス「結婚」の場合の例について以下に述べる.

文献 [4] における日中間比較対照分析の結果の内，日中両側で観測されたトピックおよびウェブページの記述内容の例を表 5.1 に示す．日中共通のトピック「国際結婚」においては，サジェスト「グリーンカード」が両側で観測されており，実際にこのサジェストで検索されたウェブページの日中共通の記述内容として,「アメリカ人と結婚する時のグリーンカードの申請手続き」があった．一方，このトピックにおいて，日本語側のみで観測されたサジェストは「ミャンマー人」,「ミャンマー　手続き」等があり，それらに関連するウェブページの記述内容は「ミャンマー人と結婚する時のポイント」であった．また，中国語側のみで観測されたサジェストは「移民」,「移民アメリカ」,「移民カナダ」等があり，それらに関連するウェブページの記述内容は「アメリカへの移民とカナダへの移民」であった.

文献 [4] における日中間比較対照分析の結果の内，日本語側のみで観測されたトピックおよびウェブページの記述内容の例を表 5.2 に示す．日本語側独自のトピックとして,「健康保険」,「離婚，婚約破棄と慰謝料について」,「結婚と六曜」等があった．また，各トピックに分類されたサジェストおよびウェブページの記述内容として，トピック「結婚と六曜」では，サジェスト「赤口」,「大安」,「六曜」,「仏滅」等があり，それらに関連するウェブページの記述内容は,「結婚式は大安に行うのが良いといった情報」等の，結婚式の日取りに関する内容であった.

文献 [4] における日中間比較対照分析の結果の内，中国語側のみで観測された

表 5.2: 「結婚」において日本語側のみで観測されたトピックおよびウェブページ中の記述内容 (抜粋)(文献 [4] より引用)

日本語独自のトピック：健康保険	
トピックに紐付けられた検索対象+サジェスト	トピック内のサジェストに対応したウェブページの記述内容
結婚 住民税	結婚後の税金について
結婚 税金	
結婚 年金手帳	結婚後の社会保険について
結婚 社会保険	

日本語独自のトピック：離婚、婚約破棄と慰謝料について	
トピックに紐付けられた検索対象+サジェスト	トピック内のサジェストに対応したウェブページの記述内容
結婚 破談慰謝料	婚約破毀の慰謝料請求について
結婚 口約束	
結婚 別居生活	別居婚についての相談
結婚 ずっと別居	

日本語独自のトピック：結婚と六曜	
トピックに紐付けられた検索対象+サジェスト	トピック内のサジェストに対応したウェブページの記述内容
結婚 赤口	結婚式は大安に行うのが良いといった情報
結婚 大安	
結婚 六曜	
結婚 仏滅	

トピックおよびウェブページの記述内容の例を表 5.3 で示す．中国語側独自のトピックとして,「結婚する際に必要な分譲住宅」,「誓約書」,「結婚式のゲーム」等があった．また，各トピックに分類されたサジェストおよびウェブページの記述内容として，トピック「結婚する際に必要な分譲住宅」では，サジェスト「结婚 买不起房(**結婚 分譲住宅を買えない**) 」等があり，それらに関連するウェブページの記述内容は,「分譲住宅を持っていない男にとって結婚は困難」等の，中国の男性が結婚する際に必要な条件に関する内容であった．

クエリ・フォーカス「就活」の例

クエリ・フォーカス「就活」の場合の例について以下に述べる．

図 5.1 で示した流れに沿って，日中文化間差異の発見を支援するために，第 3 章の集約結果に対して，トピック単位で人手によって行った日中間比較対照分析の分析結果について説明する．

まず，中国語側の 60 個のトピックと日本語側の 50 個のトピックにおいて，中国語側でのみ観測されたトピックは 27 個であり，日本語側でのみ観測されたトピックは 20 個であった．日中で共通に観測されたトピックは，中国語トピック 14 個,

表 5.3: 「結婚」において中国語側のみで観測されたトピックおよびウェブページ中の記述内容 (抜粋)(文献 [4] より引用)

中国語独自のトピック：結婚する際に必要な分譲住宅	
トピックに紐付けられた検索対象+サジェスト	トピック内のサジェストに対応したウェブページの記述内容
结婚 买不起房(結婚 分譲住宅を買えない)	中国では分譲住宅を持っていない男にとって結婚は困難
结婚 买房(結婚 分譲住宅を買う)	
结婚 没钱怎么办(結婚 金がないどうする)	結婚したいけど、金がない、どうすればいい

中国語独自のトピック：誓約書	
トピックに紐付けられた検索対象+サジェスト	トピック内のサジェストに対応したウェブページの記述内容
结婚 保证书(結婚 誓約書)	結婚式で新郎が新婦に誓う言葉
结婚 宣言(結婚 宣言)	

中国語独自のトピック：結婚式のゲーム	
トピックに紐付けられた検索対象+サジェスト	トピック内のサジェストに対応したウェブページの記述内容
结婚 如何整新郎(結婚 どうやって新郎をからかう)	結婚式の日に新婦の部屋を閉めて新郎が入れないようにして新郎をからかうゲーム
结婚 如何堵门(結婚 扉閉め)	

日本語トピック 13 個であり，日中間では 9 個の対応組となった．

次に，トピック単位で行った日中間比較対照分析の分析結果において，日中両側で観測されたトピックおよびウェブページ中の記述内容の例を図 5.2 に示す．日中共通のトピックとして,「外国語スキル」,「長所と短所」,「面接の対策と趣味」等があった．また，日中共通のトピックに分類されたサジェストおよびウェブページの記述内容として，トピック「外国語スキル」に分類されたサジェストは，中国側は「英语口语大全(英語での口頭面接のノウハウ集)」などがあり，日本語側は「英語」などがあった．それらに関連するウェブページの記述内容は「就活の役に立つ外国語スキル」であった．トピック「長所と短所」に分類されたサジェストとして，中国語側では「你的优点(あなたの強み)」があり，日本語側では「長所」があった．また，それらのサジェストに関連するウェブページの記述内容は「面接での自分の長所と短所の伝え方」であった．一方，トピック「面接の対策と趣味」に分類されたサジェストは，中国語側では「你有什么问题想问的(あなたが尋ねたい問題)」等があり，日本語側では「逆質問」等があった．また，それらに関連するウェブページの記述内容は「逆質問の仕方と好印象を持たれる趣味」であった．

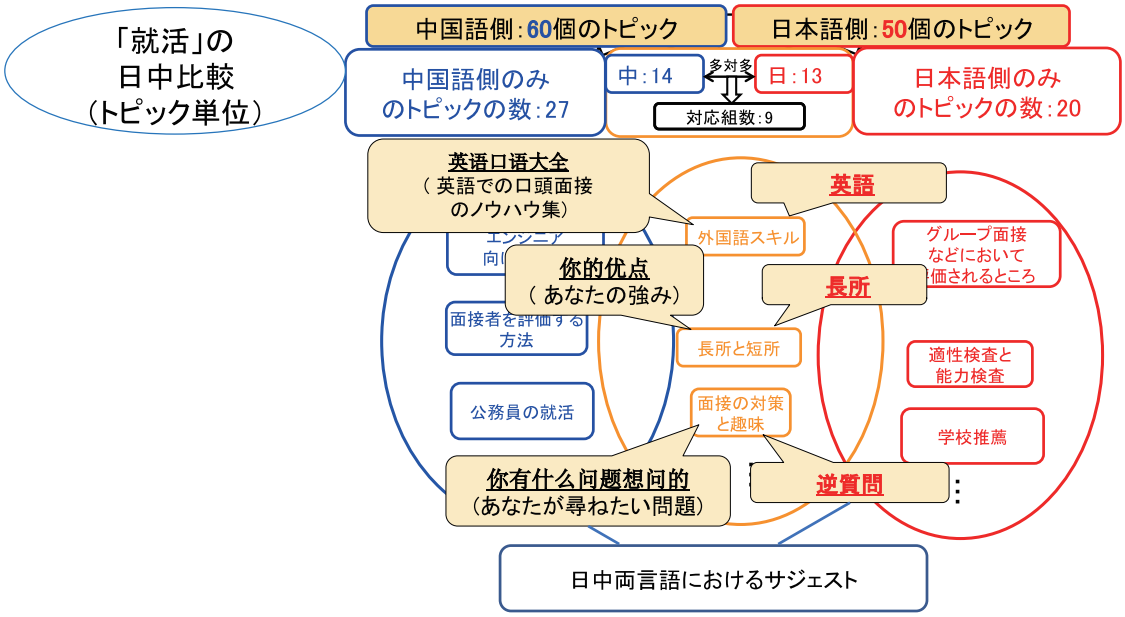


図 5.2: クエリ・フォーカス「就活」におけるトピック単位でのサジェストの日中間比較対照分析 (日中共通の話題)

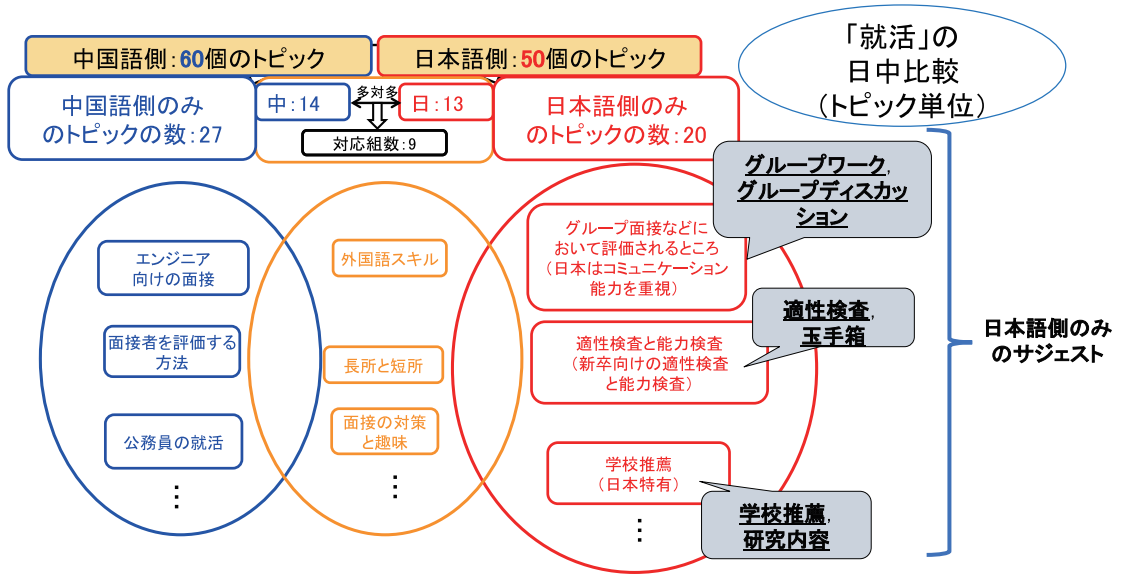


図 5.3: クエリ・フォーカス「就活」におけるトピック単位でのサジェストの日中間比較対照分析 (日本語側のみの話題)

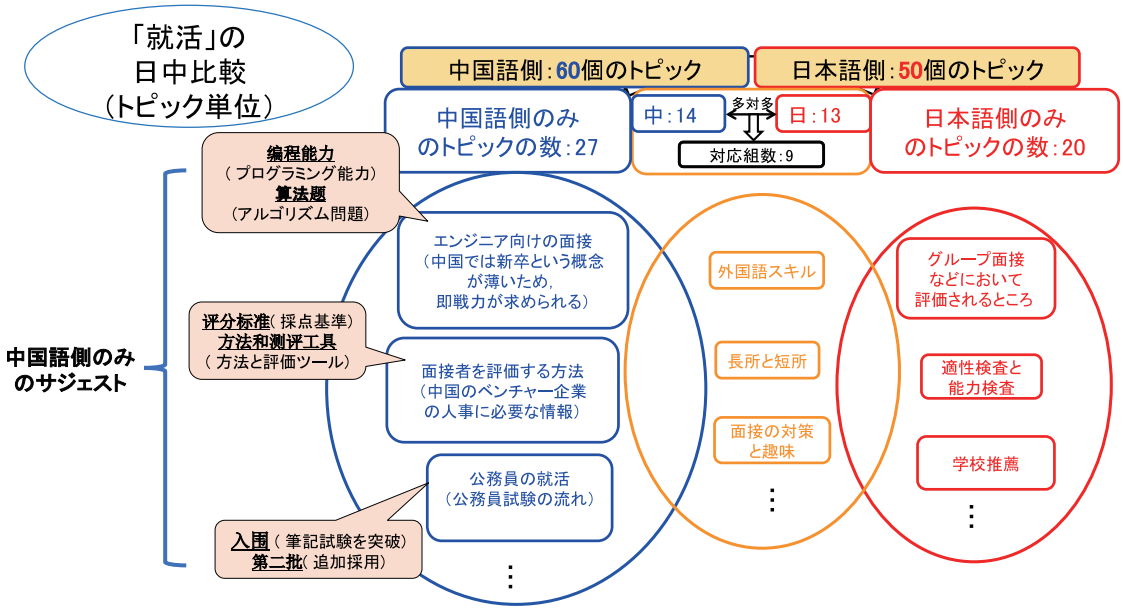


図 5.4: クエリ・フォーカス「就活」におけるトピック単位でのサジェストの日中間比較対照分析 (中国語側のみの話題)

また，トピック単位で行った日中間比較対照分析の分析結果において，日本語側のみで観測されたトピックおよびウェブページ中の記述内容の例を図 5.3 に示す．日本語側独自のトピックとして，「グループ面接などにおいて評価されるところ」，「適性検査と能力検査」，「学校推薦」等があった．また，各トピックに分類されたサジェストおよびウェブページの記述内容として，トピック「グループ面接などにおいて評価されるところ」に分類されたサジェストとしては，「グループワーク」，「グループディスカッション」等があり，それらに関連するウェブページの記述内容は「コミュニケーション能力をアピールするために，グループ面接において注意すべきポイント」であった．一方，トピック「適性検査と能力検査」に分類されたサジェストとしては，「適性検査」，「玉手箱」があり，それらに関連するウェブページの記述内容は「新卒向けの適性検査と能力検査」であった． また，トピック「学校推薦」に分類されたサジェストとしては，「学校推薦」，「研究内容」等があり，それらに関連するウェブページの記述内容は「学校推薦のメリット・デメリット」であった．

そして，トピック単位で行った日中間比較対照分析の分析結果において，中国語側のみで観測されたトピックおよびウェブページ中の記述内容の例について説明する．図 5.4 に示すように，中国語側独自のトピックとして，「エンジニア向けの面接」，「面接者を評価する方法」等があった．また，各トピックに分類されたサジェストおよびウェブページの記述内容として，トピック「エンジニア向けの面接」に分類されたサジェストとしては，「编程能力(**プログラミング能力**)」，「算法题(**アルゴリズム問題**)」

等があり，それらに関連するウェブページの記述内容は「エンジニアの実力を測るための面接問題」であった．一方，トピック「面接者を評価する方法」に分類されたサジェストとしては,「评分标准(採点基準)」,「方法和测评工具(方法と評価ツール)」があり，それらに関連するウェブページの記述内容は「面接者の実力を正しく評価するための採点基準やツール」であった．

5.2 トピック内の小分類単位での日中間比較対照分析

前節では，トピックモデルによる集約結果である各トピックを対象として，粗い話題の粒度のもとで日中間比較対照分析を行った．一方，本節では，第 3 章の集約結果に対して，トピック内の小分類を分析対象として，より細かい話題の粒度のもとでの日中間比較対照分析について述べる．本節における分析対象のクエリ・フォーカスとしては,「就活」を用いる．

本論文の手法と文献 [4] の手法の共通部分である「トピック単位での日中間比較対照分析」の各手順を以下のステップ 1 に示す．また，本論文の提案手法である「トピック内の小分類単位での日中間比較対照分析」の各手順を以下のステップ 2 に示す．

ステップ 1 トピック内の小分類単位での日中間比較対照分析 (文献 [4] と同じ手法)

1. まず，各クエリ・フォーカスに対して，日本語の検索エンジン・サジェスト，および，中国語の検索エンジン・サジェストを収集する．日本語側においては，五十音などを用いて Google 検索エンジン [4] からサジェストを収集した．中国語側においては，pinyin を用いて Baidu 検索エンジン [5] からサジェストを収集した．

2. 次に，収集された検索エンジン・サジェストを用いて,「クエリ・フォーカス + 検索エンジン・サジェスト」の AND 検索によって検索される上位 N 件 (日本語の場合は $N = 20$，中国語の場合は $N = 40$) のウェブページを収集する．

3. そして，収集されたウェブページ集合に対してトピックモデルを適用し，数十個のトピックへ集約する．各ウェブページには一つ以上の検索語エンジン・サジェストが付与されるため，ウェブページを集約することにより，サジェストも自動的に数十個のトピックへ集約される．各ウェブページには一つ以上の検索語エンジン・サジェストが付与されるため，ウェブページを集約することにより，サジェストも自動的に数十個のトピックへ集約される．

[4] https://www.google.com/
[5] https://www.baidu.com/

4. 最後に，人手で日中間比較対照分析を行い，日中二言語の各トピックを,「日中共通で観測したトピック」,「日本語のみで観測したトピック」,「中国語のみで観測したトピック」の三種類に分類する．

ステップ 2 トピック内の小分類単位での日中間比較対照分析

1. Wikipedia の本文全データに加えて，ステップ 1 で検索エンジン・サジェストによって収集したウェブページ集合を混合した混合文書集合を訓練データとして，word2vec を用いてサジェストごとの分散表現を訓練する．

2. ステップ 1 の集約結果における各トピックに対して，分散表現の類似度に基づき検索エンジン・サジェスト同士の類似度を測定する．そして，低類似度のサジェスト組を除外し，高類似度のサジェスト組のみを一つの小分類へ集約する．

3. ステップ 1 において「日中共通に観測されたトピック」に対して，人手によって，トピック内の小分類の日中間比較対照分析を行い,「日中共通で観測した小分類」,「日本のみで観測した小分類」,「中国のみで観測した小分類」の三種類に分類する．

以下では，クエリ・フォーカス「就活」において，日中共通のトピック「面接の対策と趣味」を例として，図 5.5，図 5.6，および，図 5.7 に沿って，トピック内の小分類単位での日中間比較対照分析の結果について述べる．このトピックにおいては，日中で共通に観測された小分類数は 3 個であり，日本語側でのみ観測された小分類数は 3 個であり，中国語側でのみ観測された小分類数は 4 個であった．

まず，日中共通のトピック「面接の対策と趣味」において，日中両側で観測された小分類を図 5.5 に示す．日中両側で観測された小分類として,「趣味について」,「逆質問について」等があった．また，各小分類に分類されたサジェストおよびウェブページの記述内容として，小分類「趣味について」においては，中国語側はサジェスト「兴趣爱好(**趣味と好きなこと**)」があり，日本語側は「趣味」等があった．それらに関連するウェブページの記述内容は「趣味についての答え方」であった．小分類「逆質問」においては，中国語側はサジェスト「你有什么问题想问的(**あなたが尋ねたい問題**)」があり，日本語側は「逆質問」等があった．それらに関連するウェブページの記述内容は「プラスになる逆質問」であった．

次に，日中共通のトピック「面接の対策と趣味」において，日本語側でのみ観測された小分類を図 5.6 に示す．日本語側のみで観測された小分類として,「集団面接において評価されるところ」,「好印象を与えるための喋り方」等があった．また，各小分類に分類されたサジェストおよびウェブページの記述内容として，小分類「集団面接において評価されるところ」においては，サジェスト「集団面接」,「最終面接」などがあり，それらに関連するウェブページの記述内容は「集団面接

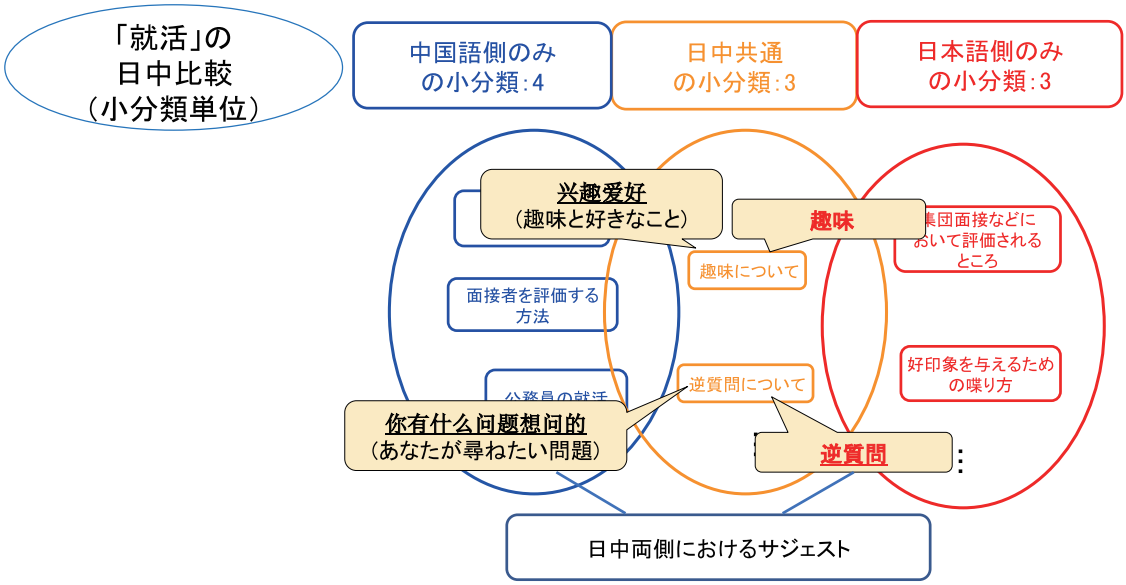


図 5.5: クエリ・フォーカス「就活」におけるトピック内の小分類単位でのサジェストの日中間比較対照分析 (日中共通の話題)

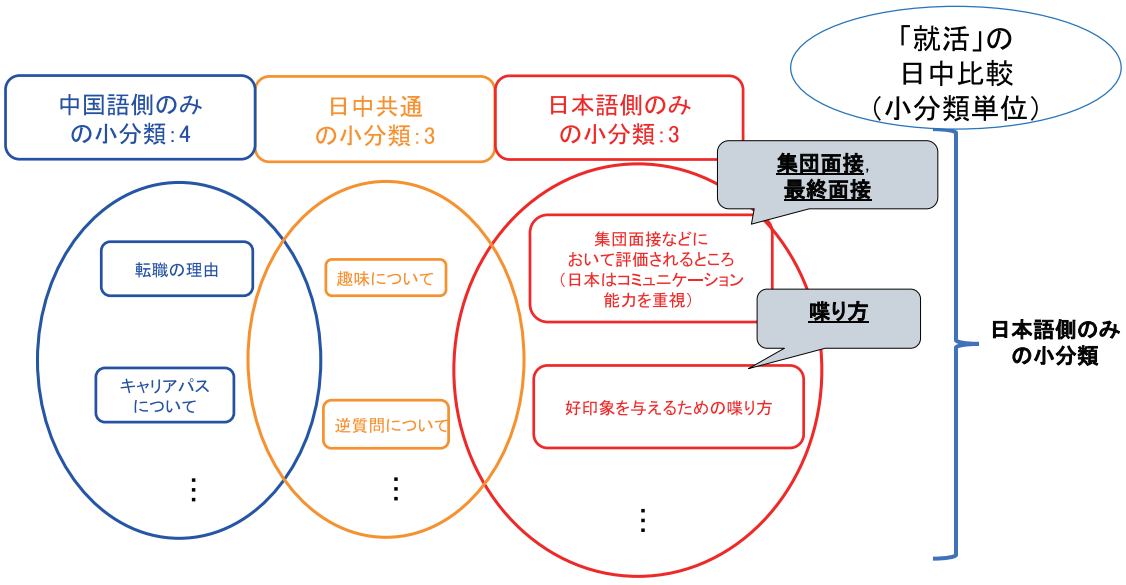


図 5.6: クエリ・フォーカス「就活」におけるトピック内の小分類単位でのサジェストの日中間比較対照分析 (日本語側のみの話題)

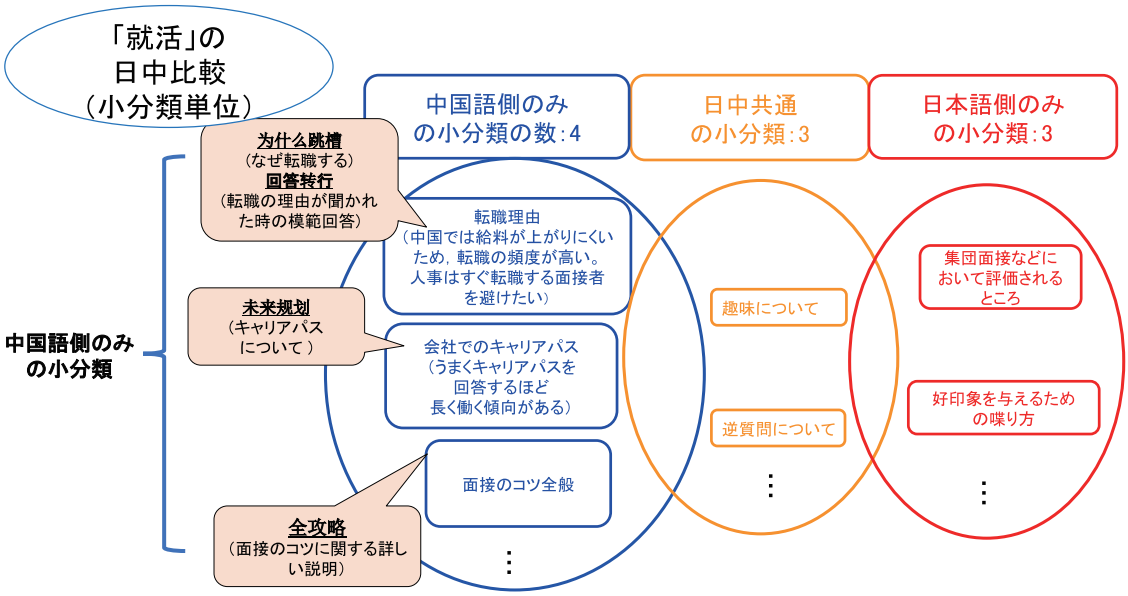


図 5.7: クエリ・フォーカス「就活」におけるトピック内の小分類単位でのサジェストの日中間比較対照分析 (中国語側のみの話題)

におけるテクニック」であった．小分類「好印象を与えるための喋り方」においては，サジェスト「喋り方」等があり，それらに関連するウェブページの記述内容は「面接官に好印象を与えるための喋り方」であった．

最後に，日中共通のトピック「面接の対策と趣味」において，中国語側でのみ観測された小分類を図 5.7 に示す．中国語側のみで観測された小分類として，「転職の理由」，「キャリアパス」等があった．また，各小分類に分類されたサジェストおよびウェブページの記述内容として，小分類「転職理由」においては，サジェスト

- 「回答转行(**異業種への転職の理由が聞かれた時の模範回答**)」，および，
- 「为什么跳槽(**なぜ転職する**)」

等があり，それらに関連するウェブページの記述内容は「転職の理由を答えるための模範解答集」であった．一方，小分類「キャリアパス」においては，サジェスト「未来规划(**キャリアパスについて**)」等があり，それらに関連するウェブページの記述内容は「会社でのキャリアパスに関する内容」であった．

5.3 本章のまとめ

5.1 節，および，5.2 節の手順により，

- トピックモデルによる集約結果である各トピックを対象とした，粗い話題の粒度での日中間比較対照分析
- トピック内の小分類を分析対象とした，細かい話題の粒度での日中間比較対照分析

の二種類の話題の粒度で日中間比較対照分析を行った．

トピック単位での日中間比較対照分析の結果において，日本語側でのみ観測したトピックは,「就活における適性検査」や「学校推薦での就活」等の新卒採用における日本独自の仕組みに関するトピックであった．一方，中国語側でのみ観測したトピックは,「エンジニア向けの面接」等の試験内容に関するトピックであった．これは，中国では新卒採用の概念が薄いため，企業にとって即戦力となる人材が求められていることを反映していると考えられる．

トピック内の小分類単位での日中間比較対照分析の結果においても，日本語側でのみ観測した小分類は,「好印象を与えるための喋り方」等，就活生が関心を持つ内容であった．一方，中国語側でのみ観測した小分類は,「転職理由」等，転職活動に関する内容であった．背景として，中国と日本は雇用の形態が異なり，中国の企業では，日本のような無期限の雇用契約を結ぶことがないため，転職する人が非常に多いという状況がある．

結果から見ると，日本では新卒採用に関心が集まっているのに対して，中国の採用活動では即戦力となる人材を重視するという，就職活動・採用活動における文化が一因となっていることが分かった．また，日本企業の終身雇用に対して，中国では有期雇用の契約社員という形態しかないという両国間の雇用形態の違いも一因となっていることが分かった．以上から，本論文の手法は日中文化間差異発見支援において有効であることが確認できた．

第6章　結論

本論文では，**ウェブ検索者の情報要求観点を網羅的に収集・集約する**手法の確立という課題について論じた．さらに，**日中間の文化間差異を発見するための比較対照分析**の課題について論じた．

人々の関心を知ることは，企業のビジネス戦略の策定や個人のライフイベントの課題解決など，様々な場面において有用である．しかし，情報が爆発している現在のインターネット世界においては，多種多様な情報源・メディアが存在しており，それらの膨大な情報の中から自分にとって有用な情報を選別することはそれほど容易なことではない．そこで，本論文では，効率的かつ網羅的に知識を収集するために，ウェブ検索者の情報要求観点 (検索エンジン・サジェスト) に着目する．**ウェブ検索者の情報要求観点を網羅的に収集・集約する**手法においては，まず，ウェブ検索の情報要求観点を網羅的に収集するアプローチとして，分析対象となるクエリ・フォーカスについて，日本語側では，五十音，濁音，半濁音および「きゃ」や「ぴゃ」などの開拗音を用いて，1 クエリ・フォーカスにつき約 1,000 個の検索エンジン・サジェストを収集した．中国語側では，Pinyin を用いて，1 クエリ・フォーカスにつき約 600 個の検索エンジン・サジェストを収集した．その後，収集された検索エンジン・サジェストを用いて,「クエリ・フォーカス + 検索エンジン・サジェスト」の AND 検索の形でウェブページの収集を行った．具体的に，日本語側においては検索される上位の 20 件のウェブページを収集し，中国語側においては検索される上位の 40 件のウェブページを収集した．最終的に，日中両側において，1 クエリ・フォーカスにつき，それぞれ約 20,000 件のウェブページを収集した．ここで，集めた検索エンジン・サジェストとウェブページ集合には，その中に含む話題の内容が冗長であるという問題があった．本論文では，この問題を解消するために，分散表現を用いることでより細かい粒度で話題を表現できるという点に着目し，検索エンジン・サジェストにおける精密な集約に取り組んだ．本論文の方式では，ウェブページ集合にトピックモデルを適用し，ウェブページにトピックを割り当てることで，検索語となった検索エンジン・サジェストの自動集約を行う．その後，word2vec によって検索エンジン・サジェストの分散表現を求めることで，検索エンジン・サジェストを数百次元のベクトルで表現する．それに基づいて検索エンジン・サジェスト同士のベクトル間の余弦類似度を求め，トピック内の小分類の抽出を行った．分散表現訓練用コーパスとして,「Wikipedia の全ページテキストのみ」と「Wikipedia の全ページテキスト + 検索エンジン・サ

ジェストによって収集されたウェブページテキスト」の二種類を用いた．評価実験においては，人手によって作成された参照用小分類データに対して,「サジェスト間の分散表現類似度の下限値」,「あるサジェストに対して，全サジェスト中の順位の上限値」，および「あるサジェストに対して，トピック内のサジェスト中の順位の上限値」の三種類の素性を設定し，前述の二種類の分散表現訓練用コーパスを用いた小分類集約結果の適合率・再現率を計算し，比較を行った．結果として，クエリ・フォーカス「就活」および「結婚」の両方において，分散表現訓練用コーパスとして「Wikipedia の全ページテキストのみ」を用いた場合より,「Wikipedia の全ページテキスト + 検索エンジン・サジェストによって収集されたウェブページテキスト」用いた場合の方が高性能であることを示した．

また，近年，日本と中国の経済や文化の交流が促進され，お互いの国についての関心が高まりつつある．2017 年の中国からの訪日旅行者数は約 736 万人に達しており，過去の最高記録を更新した．外国人旅行者のニーズに応えるために，文化間差異による嗜好の違いを把握する必要がある．一方，外国人旅行者には，訪問先における生活や文化の違いをよく理解したいという欲求もある．以上のことから，国間の文化間差異を知ることが大切だと考えられる．本論文では，**日中間の文化間差異を発見するための比較対照分析**については，粗い話題の粒度と細かい話題の粒度の二種類の粒度で，クエリ・フォーカス「就活」における日中間比較対照分析を行った．粗い話題の粒度での日中間比較対照分析においては，LDA トピックモデルによる一次集約結果に対して，各トピックに分類された検索エンジン・サジェストとウェブページの内容を人手で分析し,「日本語側のみで観測したトピック」,「中国語側のみで観測したトピック」,「日中共通で観測したトピック」の三種類のトピックに分類した．細かい粒度での日中間比較対照分析においては，word2vec によって求めた検索エンジン・サジェストの分散表現を用いて，LDA トピックモデルによる一次集約結果の各トピック内の小分類を抽出した後，前述の人手で分類した「日中共通で観測したトピック」に対して，その中に含まれる小分類を比較対照分析することで,「日本語側のみで観測した小分類」,「中国語側のみで観測した小分類」,「日中共通で観測した小分類」の三種類に分類した．分析結果については，粗い話題の粒度での日中間比較対照分析において，中国語側でのみ観測したトピックとしては，求職者の実力を測るための試験内容に関するトピックが多く観測された．一方，日本語側でのみ観測したトピックとしては，新卒採用における日本独自の仕組に関するトピックが多く観測された．細かい話題の粒度での日中間比較対照分析において，中国語側でのみ観測された小分類としては，転職活動に関する内容があった．日本語側でのみ観測した小分類は，就活生が関心を持つ内容に関する小分類であった．このように，日中文化間差異発見支援のアプローチにより，採用活動で重視するポイントの違いや両国の雇用形態の違いを発見することができ，本論文の提案手法の有効性が確認できた．

今後の課題の一つとして，本論文の第 3 章で提案された検索エンジン・サジェストに対するクラスタリングの手法を改善することが挙げられる．本論文の第 3 章

の枠組みは，第一段階であるトピックモデルによる検索エンジン・サジェストの一次集約と，第二段階である分散表現を用いて小分類を抽出，という二段階の構成になっている．第一段階と第二段階はそれぞれ独立しており，第一段階においては，文字列そのものに対してクラスタリングを行っており，第二段階で用いた検索エンジン・サジェストの分散表現を考慮していなかった．そのため，第 3 章の手法においては，分散表現の性能を活かしきれていないという弱点がある．近年，深層学習に基づく自然言語処理の技術が画期的に発展しており，機械学習技術を用いて，単語や文書に対して意味表現ベクトルを求め，クラスタリングを行う手法が多数提案された．文献 [17] においては，単語の分散表現に対して，混合ガウスモデル (Gaussian mixture model) によってクラスタリングを行い，各単語がクラスタに属する確率分布を求めた後，確率分布と単語の idf 値を用いて，単語の分散表現を拡張し，それに基づいた高精度の文書ベクトルを生成する手法を提案している．また，ベクトルを集約対象とするクラスタリングの手法として，K-means 法 [15] や，自己組織化マップ (Self-Organizing Map: SOM) [9] 等の手法が挙げられる．これらの手法を用いて，検索エンジン・エンジン・サジェストの文字列ではなく，分散表現に対してクラスタリングを行う方式が分類精度の向上に有効であると期待される．

また，もう一つの課題として，本論文で述べた手法においては，人手によって日中トピックの対応付け作業を行うため，コストがかかるという弱点がある．近年，word2vec [18] をはじめ，深層学習により分散表現を求める手法が多く提案された．例えば，文献 [13] の手法では，二言語のパラレルコーパスを用いて訓練することによって，二言語の単語の分散表現を同一ベクトル空間中で表現でき，これによって，同じ言語の単語間の類似度と，異なる言語の単語間の類似度を測定できる．ここで，本論文の第 3 章で述べた，トピックモデルを用いた検索エンジン・サジェストの集約において，各トピックを，そのトピックに属する確率が上位である数個のサジェストの集合とみなした上で，文献 [13] の言語横断類似単語推定方式を組み込むことが有望である．これによって，第 5 章の日中間比較対照分析においては，以下の手順によってトピックの日中間自動対応付けを実現できると期待される．具体的には，既存の日中パラレルコーパスから日中検索エンジン・サジェストの分散表現を求めた後，日中トピックの間の類似度を自動的に計算し，一定の下限値を満たす日中トピックを対応付ける方式の導入が有望であると考えられる．

付 録A　質問回答事例およびウェブから収集されたノウハウ知識の日中間対照分析

A.1　はじめに

21 世紀の情報社会では，政府機関や企業などにとって，グローバル化により，自国の情報だけではなく他国の情報も重要となっている．近年のインターネットの普及により，非常に多くの人がウェブサイトを閲覧して情報を収集している．そうしたウェブ閲覧者の多くは，自らの関心事項について，Google，Yahoo!，Baidu といった検索エンジンを用いてウェブ検索を行っている．ここで，ウェブ検索者・ウェブ閲覧者が，検索エンジンを用いて他国の情報を得ることはそれほど容易なことではない．そこで，本研究においては，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アプローチをとる．文献 [22] では，特に，文献 [19] において質問回答事例，および，ウェブから収集したノウハウ知識に対して，日中間で比較対照分析を行う手法を提案する．

文献 [22] の全体の流れを図 A.1 に示す．文献 [19] の手法においては，まず，質問回答サイトから収集した質問回答事例，および，検索エンジン・サジェストを索引として収集されたウェブページの混合文書集合に対してトピックモデルを適用することにより，話題のまとまりを生成する．次に各話題を「ノウハウ知識」,「ノウハウ以外の知識」,「意見」,「その他」の 4 つに分類することで，ノウハウ知識を選定する．文献 [22] では,「就活」および「結婚」を検索対象として収集されたノウハウ知識に対して日中間で比較対照分析を行った．その結果，検索対象「就活」において，面接に関して，日本特有のノウハウ知識として「敬語の使い方」,「就活メイク」，中国特有のノウハウ知識として「就活面接ショー番組」，日中共通のノウハウ知識として「就活生の服装」等が収集された．

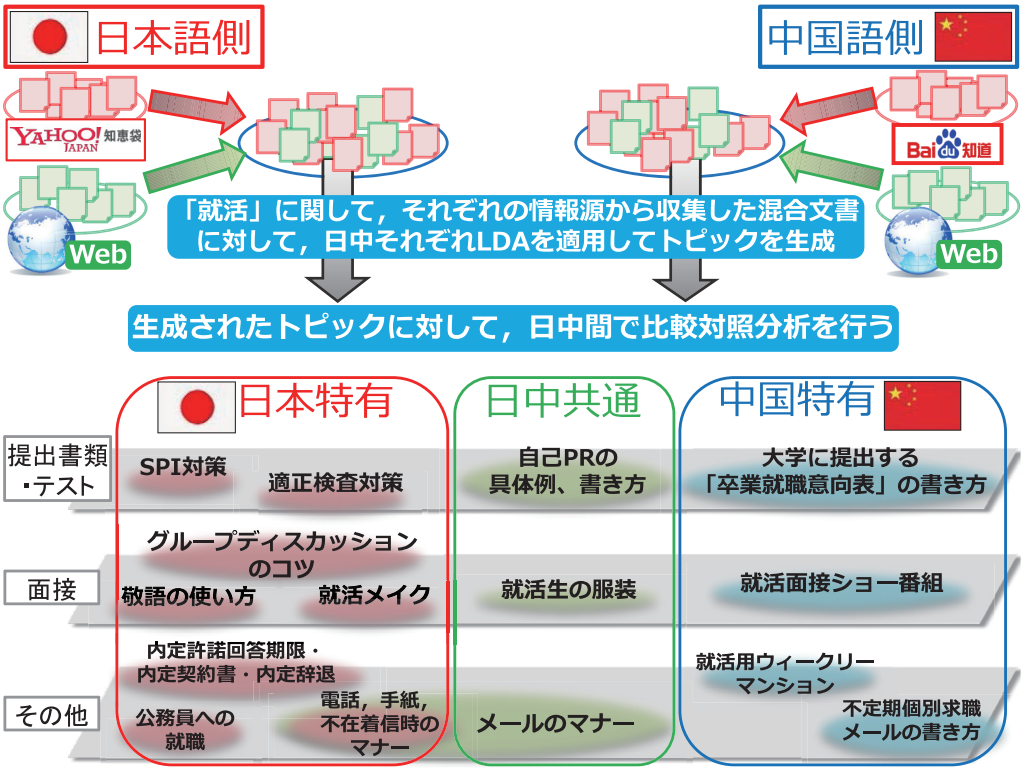


図 A.1: 質問回答事例およびウェブから収集されたノウハウ知識の日中間対照分析の流れ (検索対象:「就活」の抜粋)

A.2　質問回答事例の収集

日本側の質問回答事例のデータとして，Yahoo!知恵袋[1] から提供されている 2004 年 4 月 1 日 ~2009 年 4 月 7 日の 5 年間の質問回答事例のデータ (質問: 16,257,413 件，回答: 50,053,894 件) を用いた．文献 [22] では，カテゴリ名，質問タイトル，質問本文のいずれかに検索対象 q が含まれている質問を抽出し，その質問に対する回答本文全てを結合し，一つの質問回答事例 d_q を作成した．各検索対象 q あたりの質問回答事例の文書集合を $D_q = \{d_q^1, \ldots, d_q^k\}$ と定義する．

一方，中国側の質問回答事例は，2014 年 11 月 ~2015 年 1 月の期間に，Baidu(百度) 知道[2] から収集した．2015 年 1 月の時点で，Baidu 知道に掲載されていた解決済質問数は 354,412,701 件であった．

各検索対象において，日中それぞれの言語において収集した質問回答事例の数を表 A.1 に示す．

[1] http://chiebukuro.yahoo.co.jp/
[2] http://zhidao.baidu.com/

表 A.1: 各検索対象におけるサジェスト数，および,混合文書集合の記事数

検索対象	言語	質問回答事例数	ウェブページ		質問回答事例数＋ウェブページ数
			サジェスト数	ページ数	
就活	日本語	11,366	934	13,221	24,587
	中国語	754	209	3,054	3,808
結婚	日本語	35,426	956	14,409	49,835
	中国語	753	248	4,085	4,838

A.3 検索エンジン・サジェストを用いたウェブページの収集

各検索エンジン会社においては，ウェブ検索者の検索ログが蓄積されており，多数のウェブ検索者が検索したキーワードに対して，検索者が強い関心を持つ語を抽出し，検索エンジン・サジェストとして提示するサービスを提供している．ここで，検索エンジン・サジェストとして提示される語は，検索対象に対して，多数のウェブ検索者が AND 検索の形で二つ目以降に入力した語を情報源として抽出されたものである．そこで，文献 [22] では，検索エンジン・サジェストには，ウェブ検索者の関心事項そのものが反映されていると考え，ウェブ検索者の関心事項を収集する目的で，検索エンジン・サジェストを収集する．

日本語側においては，Google[3] 検索エンジンに対して，一検索対象当り 100 通りの文字列を指定し，最大 1,000 語のサジェストを収集する．100 通りの文字列とは具体的には，五十音，濁音，半濁音および「きゃ」や「びゃ」などの開拗音である．例えば検索窓に「就活 ね」と入力すると,「ネクタイ」や「ネクタイ 結び方」などがサジェストとして掲示されるので，それらを収集することにより，934 個のサジェストを収集した．

中国語側においては，Google 検索エンジンに対して，一検索対象当り 28 通りの文字列を指定し，最大 280 語のサジェストを収集する．28 通りの文字列とは具体的には，中国語のピン音の部首である．例えば検索窓に「求职(就活) j」と入力すると,「简历(エントリーシート)」などがサジェストとして掲示されるので，それらを収集することにより，209 個のサジェストを収集した．

各検索対象において，日中それぞれの言語において収集したサジェストの数を

[3] https://www.google.com/

表 A.2: ノウハウ知識の話題数

検索対象	大分類の数	トピック数 (ノウハウ知識/LDA 適用時)		話題数				
		日本語側	中国語側	日本特有	中国特有	日中共通	合計 日本	合計 中国
就活	6	33/50	24/30	40	15	日:10, 中:16	50	31
結婚	4	26/50	25/30	24	19	日:11, 中:12	35	31

表 A.1 に示す.

ここで，ある検索対象に対して収集されたサジェストの集合を$\mathbb{S}$とすると，$s \in \mathbb{S}$となるサジェストsに対して，検索対象との AND 検索により上位N件以内に検索されるウェブページpの集合を$\mathbb{P}(s, N)$ (ただし，文献 [22] においては，$N = 20$とする) とし，各検索対象あたりのウェブページの文書集合D_wを$D_w = \bigcup_{s \in \mathbb{S}} \mathbb{P}(s, N)$と定義する．なお，ウェブページの収集には Yahoo! Search BOSS API [4] を用いた.

A.4 トピックモデルの適用

A.2 節および A.3 節で収集した質問回答事例の文書集合D_qとウェブページの文書集合D_wの混合文書集合$D_{qw} = D_q \bigcup D_w$を作成する．各検索対象における混合文書集合の記事数を表 A.1 に示している．文献 [22] では，トピックモデルとして潜在的ディリクレ配分法 (LDA; Latent Dirichlet Allocation) [3] を用いる．LDA を用いたトピックモデルの推定においては，語wの集合をVとして，語$w (w \in V)$の列によって表現された文書の集合と，トピック数Kを入力として，各トピックz_n $(n = 1, \ldots, K)$における語wの確率分布$P(w|z_n)$ $(w \in V)$，および，各文書bにおけるトピックz_nの確率分布$P(z_n|b)$ $(n = 1, \ldots, K)$を推定する [5]．本研究では，各文書に対して確率が最大のトピックを一意に割り当てることにより，各文書を分類することとした．記事集合をD，トピック数をK，1 つの文書を$d (d \in D)$とすると，トピック$z_n (n = 1, \ldots, K)$の記事集合$D(z_n)$は以下の式で表される.

$$D(z_n) = \left\{ d \in D \,\middle|\, z_n = \underset{z_u \ (u=1,\ldots,K)}{\arg\max} P(z_u|d) \right\}$$

[4] http://developer.yahoo.com/search/boss

[5] 推定のためのツールとしては，GibbsLDA++を用いた．LDA のハイパーパラメータであるα, β としては，$\alpha = 50/K$，$\beta = 0.1$，Gibbs サンプリングの反復回数は 2,000 を用いた.

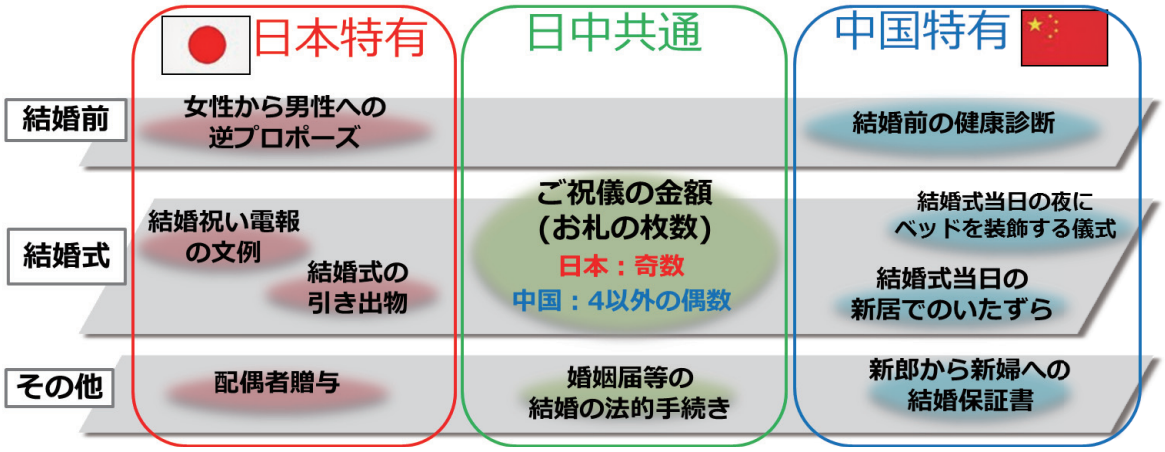


図 A.2: 質問回答事例およびウェブから収集されたノウハウ知識の日中間対照分析の例 (検索対象: 「結婚」の抜粋)

A.5　ノウハウ知識の収集

A.4 節の手順に従い，各トピックに割り当てられた確率上位 20 件の記事を分析したところ，トピックによっては，いずれかの情報源に偏るものがあった．そこで，今回の分析では，情報源ごとに確率上位 10 件の記事を分析し，そのうち 3 件以上同一とされる話題があった場合に，そのトピックの話題として抽出した[6]．これにより各トピックの情報源毎に最大 3 つの話題を抽出した．なお，話題分析の際には，各トピックにおける確率 $P(w|z_n)$ の高い語 w とトピックおよびウェブページに割り当てられたサジェストを参照して分析を行う．

次に，各トピックから得られた各話題を，1) ノウハウ知識，2) ノウハウ以外の知識，3) 意見，4) その他，の 4 種類に分類する．このうち,「ノウハウ知識」は，やり方についての情報など閲覧した人の行動につながるものである．具体的にはレシピサイト，方法や手順が書かれているもの，対策やマナー，コツなどがノウハウ知識にあたる．文献 [22] では，ユーザの行動につながる知識は全てノウハウ知識であるとみなした．収集されたノウハウ知識の話題数を表 A.2 に示す.「ノウハウ以外の知識」は，それを見てもユーザの行動に影響を与えない情報である．例えば,「芸能人の結婚」がこれにあたる.「意見」は，多くの人の意見を求める相談や，自分の意見を主張しているものである．例えば,「就活中の友人，恋人とのつきあ

[6]異なるトピックから同一の話題が収集される場合においても，文献 [22] の分析の範囲においては，別の話題として数えた．

表 A.3: 中国特有のノウハウ知識の詳細説明

検索対象	話題	説明
就活	大学に提出する「卒業就職意向表」の書き方	就活生が自分の技術や就職希望企業について「卒業就職意向表」を記述して大学に提出する．大学側は企業による訪問面接の参加の斡旋をしてくれる．
	就活面接ショー番組	就活生と企業がスタジオで実際に面接をしてその場で選考を行う番組
	就活用ウィークリーマンション	主に就活生が利用する廉価型のウィークリーマンション
	不定期個別求職メールの書き方	希望する企業に対して就活生が不定期の時期に自己アピールおよび選考依頼のメールを送る際のメールの書き方
結婚	結婚前の健康診断	中国では，結婚前に健康診断を受ける慣習がある．
	結婚式当日の新居でのいたずら	結婚式当日，親しい友人が新居に集まり新郎新婦にいたずらをする．
	結婚式当日の夜にベッドを装飾する儀式	結婚式当日の夜に，赤い寝具を用いてベッドを装飾するという中国独特の儀式を行う．
	新郎から新婦への結婚保証書	結婚の際，新郎が守るべき約束を結婚保証書に明記し渡す．

いかたについて」や「結婚後の嫁姑の問題」がこれにあたる.「その他」は，上記3つのいずれにも分類できないものである．例えば,「結婚占い」がこれにあたる.

A.6　ノウハウ知識の日中間対照分析

前節で収集されたノウハウ知識に対して，日中間で比較対照分析を行ったところ，各検索対象における，日本特有のノウハウ知識の数，中国特有のノウハウ知識の数，日中共通のノウハウ知識の数はそれぞれ表 A.2 に示す結果となった[7].

検索対象「就活」においては，図 A.1 に示すように，日本特有のノウハウ知識として「SPI 対策」,「敬語の使い方」,「内定許諾回答期限・内定契約書・内定辞退」等 40 個のノウハウ知識が収集された．中国特有のノウハウ知識としては,「大学に提出する卒業就職意向表の書き方」,「就活面接ショー番組」,「就活用ウィークリーマンション」等 15 個のノウハウ知識が収集された (図 A.1 における中国特有のノウハウ知識の詳細な説明を表 A.3 に示す)．日中共通のノウハウ知識としては,「自己 PR の具体例，書き方」,「就活生の服装」,「メールのマナー」等，日本語側 10 個，中国語側 16 個のノウハウ知識が収集された[8].

また，検索対象「結婚」においては，図 A.2 に示すように，日本特有のノウハウ知識として「女性から男性への逆プロポーズ」,「結婚祝い電報の文例」,「配偶者贈与について」等 24 個のノウハウ知識が収集された．中国特有のノウハウ知識としては,「結婚前の健康診断」,「結婚式当日の新居でのいたずら」,「新郎から新婦への結婚保証書」等 19 個のノウハウ知識が収集された (図 A.2 における中国特有のノウハウ知識の詳細な説明を表 A.3 に示す)．日中共通のノウハウ知識としては,「ご祝儀の金額 (お札の枚数)」,「婚姻届等の結婚の法的手続き」等，日本語側 11 個，中国語側 12 個のノウハウ知識が収集された[9].

A.7　関連研究

文献 [30] においては，特定の話題について，日本語ブログ記事，および，中国語ブログ記事を収集し，日中両国の文化間差異の発見を支援する方式を提案している．しかし，ブログを情報源とする場合，日中両国の文化間の差異をウェブ検

[7]日本語側の「メール，電話，手紙，不在着信時のマナー」のノウハウ知識のうち「メールのマナー」の部分は，日中共通のノウハウ知識として，日中共通の話題のうちの日本語側 10 個のうちの 1 つとして数えた．一方，残りの「電話，手紙，不在着信時のマナー」の部分は日本特有のノウハウ知識 40 個のうちの 1 つとして数えた．このため，日本特有の 40 個と日中共通の日本語側の 10 個を加えた計 50 個が，合計欄の日本語側 49 個よりも 1 つ多くなっている.

[8]中国語側のノウハウ知識「エントリーシート」,「自己分析」，および,「面接の対策」は複数のトピックから収集されたため，日本語側に比べて中国語側の話題数が多くなっている.

[9]中国語側のノウハウ知識「ご祝儀の金額 (お札の枚数)」は複数のトピックから収集されたため，日本語側に比べて中国語側の話題数が多くなっている.

索者の視点から効率よく収集することが容易でないという問題があった．その他，文献 [21] においては，日中質問回答サイトを対象として，トラブル情報の比較対照分析を行い，日中両国の文化間の差異発見過程を支援する方式を提案している．また，文献 [4] においては，日中検索エンジン・サジェストを用いて，ウェブ検索者の関心事項に着目することにより，ウェブ上の情報から国・文化・言語間の差異発見過程を支援する方式を提案している．

A.8　本章のまとめ

文献 [22] では，質問回答事例，および，検索エンジン・サジェストを索引として収集されたウェブページ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提案した．

謝辞

本論文は筆者が筑波大学大学院システム情報工学研究科において行った研究成果をまとめさせて頂いたものです.

本論文を作成するにあたり，指導教員である筑波大学大学院システム情報工学研究科知能機能システム専攻の宇津呂武仁教授には，修士課程在学時に，研究というものの面白さと厳しさ，研究に向かう姿勢や基礎を教えて頂きました．また，社会人博士課程入学のご相談から研究の指導および生活のアドバイスに至るまで終始一貫して暖かいご指導を賜り，心から御礼申し上げます.

本論文の審査過程において，システム情報系古賀弘樹教授，矢野博明教授，星野准一准教授，乾孝司准教授には，お忙しい中，論文をまとめるに当たって副査としてご指導賜りました．心から御礼申し上げます.

自然言語処理研究室の皆様には，多くのご支援を下さり，様々な形でお世話になりました．心から御礼申し上げます.

そして，心が折れそうになった時に励みの言葉をくれた人々に感謝いたします.最後に，いつでも温かく見守り，励ましてくれた家族に心から感謝いたします.

参考文献

[1] 新井翔太, 聶添, 宇津呂武仁, 河田容英. 「契約・解約」に関する消費者トラブル相談・回答事例の分析. 第 27 回人工知能学会全国大会論文集, 2013.

[2] M. Bautin, L. Vijayarenu, and S. Skiena. International Sentiment Analysis for News and Blogs. In *Proc. ICWSM*, pp. 19–26, 2008.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[4] 陳磊, 井上祐輔, 守谷一朗, 今田貴和, 宇津呂武仁, 河田容英, 神門典子. トピックモデルを用いたウェブ検索者の関心の日中間対照分析. 言語処理学会第 21 回年次大会論文集, pp. 696–699, 2015.

[5] J. Guo, X. Cheng, G. Xu, and H.-W. Shen. A structured approach to query recommendation with social annotation data. In *Proc. 19th CIKM*, pp. 619–628, 2010.

[6] S. Hu, Y. Takahashi, L. Zheng, T. Utsuro, et al. Cross-lingual topic alignment in time series Japanese / Chinese news. In *Proc. 26th PACLIC*, pp. 532–541, 2012.

[7] 胡碩, 高橋佑介, 鄭立儀, 宇津呂武仁, 吉岡真治, 神門典子. 日中時系列ニュースにおけるバースト・トピックの推定と二言語間対応付け. 言語処理学会第 19 回年次大会論文集, pp. 204–207, 2013.

[8] 井上祐輔, 今田貴和, 陳磊, 徐凌寒, 宇津呂武仁, 河田容英. 検索エンジン・サジェストおよびトピックモデルを用いたウェブ検索結果の集約. 第 8 回 DEIM フォーラム論文集, 2016.

[9] T. Kohonen. *Self-Organizing Maps*. Springer, 3rd. edition, 2000.

[10] 小池大地, 鄭立儀, 今田貴和, 守谷一朗, 井上祐輔, 宇津呂武仁, 河田容英, 神門典子. ウェブ検索者の情報要求観点の集約. 言語処理学会第 20 回年次大会論文集, pp. 328–331, 2014.

[11] K. W.-T. Leung, W. Ng, and D. L. Lee. Personalized concept-based clustering of search engine queries. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20, No. 11, pp. 1505–1518, 2008.

[12] X. Li, J. Chi, C. Li, J. Ouyang, and B. Fu. Integrating topic modeling with word embeddings by mixtures of vMFs. In *Proc. the 26th COLING*, pp. 151–160, 2016.

[13] T. Luong, H. Pham, and C. D. Manning. Bilingual word representations with monolingual quality in mind. In *Proc. HLT-NAACL*, pp. 151–159, 2015.

[14] H. Ma, H. Yang, I. King, and M. R. Lyu. Learning latent semantic relations from clickthrough data for query suggestion. In *Proc. 18th CIKM*, pp. 709–718, 2008.

[15] J. Macqueen. Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.

[16] 牧田健作, 鈴木浩子, 小池大地, 鄭立儀, 宇津呂武仁, 河田容英, 神門典子. トピックモデルを用いたブロガー・コミュニティの収集と俯瞰. 第 5 回 DEIM フォーラム論文集, 2013.

[17] D. Mekala, V. Gupta, B. Paranjape, and H. Karnick. SCDV: Sparse composite document vectors using soft clustering over distributional representations. In *Proc. EMNLP*, pp. 659–669, 2017.

[18] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proc. 26th NIPS*, pp. 3111–3119, 2013.

[19] 守谷一朗, 井上祐輔, 今田貴和, 聶添, 宇津呂武仁, 河田容英, 神門典子. 質問回答事例および検索エンジン・サジェストを用いたノウハウ知識の相補的収集. 第 7 回 DEIM フォーラム論文集, 2015.

[20] 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, Vol. 25, No. 5, pp. 613–622, 2010.

[21] 聶添, 新井翔太, 宇津呂武仁, 河田容英. 日中質問回答サイトの比較対照分析および文化間差異発見支援. 第 27 回人工知能学会全国大会論文集, 2013.

[22] 聶聶添, 守谷一朗, 井上祐輔, 今田貴和, 李雪山, 宇津呂武仁, 河田容英, 神門典子. 質問回答事例およびウェブから収集されたノウハウ知識の日中間対照分析. 言語処理学会第 21 回年次大会論文集, pp. 948–951, 2015.

[23] 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, Vol. 27, No. 1, pp. 527–532, 2015.

[24] Tian NNie, Yi Ding, Chen Zhao, Youchao Lin, and Takehito Utsuro. A method of subtopic classification of search engine suggests by integrating a topic model and word embeddings. *International Journal of Software Innovation*, Vol. 6, No. 3, pp. 67–78, 2018.

[25] 岡崎直観. 言語処理における分散表現学習のフロンティア. 人工知能学会誌, Vol. 31, No. 2, pp. 189–201, 2016.

[26] B. Pouliquen, R. Steinberger, and J. Belyaeva. Multilingual Multi-document Continuously-updated Social Networks. In *Proc. Workshop: Multi-source, Multilingual Information Extraction and Summarization*, pp. 25–32, 2007.

[27] 鈴木浩子, 横本大輔, 牧田健作, 宇津呂武仁, 河田容英, 福原知宏. Wikipedia を知識源とする日英ブログ記事集合の観点分類と言語間対照分析. 情報処理学会研究報告, Vol. 2011–DBS–153, , 2011.

[28] R. Yangarber, C. Best, P. von Etter, F. Fuart, D. Horby, and R. Steinberger. Combining Information about Epidemic Threats from Multiple Sources. In *Proc. Workshop: Multi-source, Multilingual Information Extraction and Summarization*, pp. 41–48, 2007.

[29] M. Yoshioka. IR Interface for Contrasting Multiple News Sites. In *Proc. 4th AIRS*, pp. 516–521, 2008.

[30] 鄭立儀, 小池大地, 宇津呂武仁, 河田容英, 神門典子. 日中ブロガー・コミュニティの収集・俯瞰・対照分析. 情報処理学会研究報告, Vol. 2013–DBS–157/2013–IFAT–111, , 2013.

[31] L. Zheng, T. Utsuro, and M. Yoshioka. Bursty topics in time series Japanese / Chinese news streams and their cross-lingual alignment. In *Proc. 13th PACLING*, 2013.

業績リスト

査読付き論文雑誌

1. 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 (日本知能情報ファジィ学会誌), Vol. 27, No. 1, pp. 527–532, March 2015.

2. Tian Nie, Yi Ding, Chen Zhao, Youchao Lin, and Takehito Utsuro. A method of subtopic classification of search engine suggests by integrating a topic model and word embeddings. *International Journal of Software Innovation*, Vol. 6, No. 3, pp. 67–78, July-September 2018.

査読付き国際会議論文

1. Shota Arai, Tian Nie, Takehito Utsuro, Yasuhide Kawada, and Noriko Kando. Collecting and classifying examples of consumer troubles on "contract and cancellation" in a question-answer site. In *Proceedings of the 10th International Symposium on Natural Language Processing*, pp. 171–178, October 2013.

2. 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. In *Proceedings of the 2014 28th International Conference on Advanced Information Networking and Applications Workshops*, pp. 664–669, May 2014.

3. 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. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pp. 1917–1923, May 2016.

4. 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 *Proceedings of the 11th International*

Conference on Ubiquitous Information Management and Communication, January 2017.

5. 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. In *Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication*, January 2017.

6. 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. In *Proceedings of the 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, pp. 581–586, June 2017.

全国大会，研究会報告，その他の研究発表

1. 新井翔太, 聶添, 宇津呂武仁, 河田容英, 神門典子. 「契約・解約」に関する消費者トラブル相談事例の分類と分析. 言語処理学会第 19 回年次大会論文集, pp. 94–97, March 2013.

2. 新井翔太, 聶添, 宇津呂武仁, 河田容英. 「契約・解約」に関する消費者トラブル相談・回答事例の分析. 第 27 回人工知能学会全国大会論文集, June 2013.

3. 聶添, 新井翔太, 宇津呂武仁, 河田容英. 日中質問回答サイトの比較対照分析および文化間差異発見支援. 第 27 回人工知能学会全国大会論文集, June 2013.

4. 鄭立儀, 小池大地, 聶添, 今田貴和, 陳磊, 宇津呂武仁, 河田容英, 神門典子. ウェブ検索者の情報要求観点の日中間対照分析. 言語処理学会第 20 回年次大会論文集, pp. 332–335, March 2014.

5. 守谷一朗, 井上祐輔, 今田貴和, 聶添, 宇津呂武仁, 河田容英, 神門典子. 質問回答事例および検索エンジン・サジェストを用いたノウハウ知識の相補的収集. 第 7 回データ工学と情報マネジメントに関するフォーラム—DEIM フォーラム— 論文集, March 2015.

6. 井上祐輔, 守谷一朗, 今田貴和, 聶添, 宇津呂武仁, 河田容英, 神門典子. 質問回答事例および検索エンジン・サジェストを情報源とするノウハウ知識の収集インタフェース. 言語処理学会第 21 回年次大会論文集, pp. 700–703, March 2015.

7. 聶添, 守谷一朗, 井上祐輔, 今田貴和, 李雪山, 宇津呂武仁, 河田容英, 神門典子. 質問回答事例およびウェブから収集されたノウハウ知識の日中間対照分析. 言語処理学会第 21 回年次大会論文集, pp. 948–951, March 2015.

8. 聶添, 徐凌寒, 趙辰, 宇津呂武仁, 河田容英. サジェストおよびトピックモデルを用いた多様な話題のウェブページの選択的収集. 第 30 回人工知能学会全国大会論文集, June 2016.

9. 宇津呂武仁, 徐凌寒, 聶添, 趙辰, 李佳奇, 河田容英. 企業名に関する関心動向のトピックモデリングを用いた日中市場シェアの分析. 第 30 回人工知能学会全国大会論文集, June 2016.

10. Chen Zhao, Yi Ding, Tian Nie, Jiaqi Li, Takehito Utsuro, Yasuhide Kawada, and Noriko Kando. Identifying Web pages with major contents based on search engine suggests and topic modeling. 第 9 回データ工学と情報マネジメントに関するフォーラム—DEIM フォーラム— 論文集, March 2017.

11. 聶添, 丁易, 趙辰, 李佳奇, 宇津呂武仁, 河田容英. トピックモデルおよび分散表現の併用による検索エンジン・サジェストの集約. 第 31 回人工知能学会全国大会論文集, May 2017.

12. Chen Zhao, Tian Nie, Yi Ding, Jiaqi Li, Takehito Utsuro, and Yasuhide Kawada. Identifying major documents with search engine suggests by unsupervised subtopic labeling. 第 31 回人工知能学会全国大会論文集, May 2017.