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*

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§
筑波大学大学院 システム情報工学研究科† (株) ログワークス‡ 国立情報学研究所§

1 はじめに

21 世紀の情報社会では，政府機関や会社企業などにとって，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ウェブ検索を行っ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．この問題に対して，文献 [5] 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，日中検索エンジン・サジェストを用いて，国・文化・言語間の差異発見過程を支援する方式を提案した．しかし，文献 [5] の手法においては，サジェストの集約において人手を用いており，相当なコストと時間が要していた．そこで，本論文では，トピックモデルを用いることによりサジェスト集約を自動で行った上で，日中間対照分析を行う (図 1)．

2 検索エンジン・サジェストの収集

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

表 1: 「結婚」に関するサジェストの日中比較対照分析結果 (抜粋)

検索対象+サジェスト			
	日本語側		中国語側
日中共通で観測	結婚 年齢 法律	<=>	结婚 法定年龄(**結婚 法定年齢**)
	結婚 指輪		结婚 戒指(**結婚 指輪**)
	結婚 グリーン		结婚 绿卡(**結婚 グリーンカード**)
	結婚 費用		结婚 费用(**結婚 費用**)
日本語側のみで観測	結婚 住民税		
	結婚 専業主婦		
	結婚 農家		
中国語側のみで観測			结婚 买房(**結婚 分譲住宅を買う**)
			结婚 体检(**結婚 身体検査**)
			结婚 二周年(**結婚 二周年**)

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，本論文では，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，検索対象「結婚」に着目し，Google[1]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，半濁音及び「きゃ」や「びゃ」などの開拗音である．例えば検索窓に「結婚 ゆ」と入力すると,「指輪」や「指輪 相場」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，956 個のサジェストを収集した．

中国語側においては，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) j」と入力すると,「戒指(**指輪**) 」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，277 個のサジェストを収集した．

*Topic Model based Comparative Analysis of Concerns of Web Search between Japanese and Chinese

†Lei Chen, Yusuke Inoue, Ichiro Moriya, Takakazu Imada, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Yashuhide Kawada, Logworks Co., Ltd.
§Noriko Kando, National Institute of Informatics

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

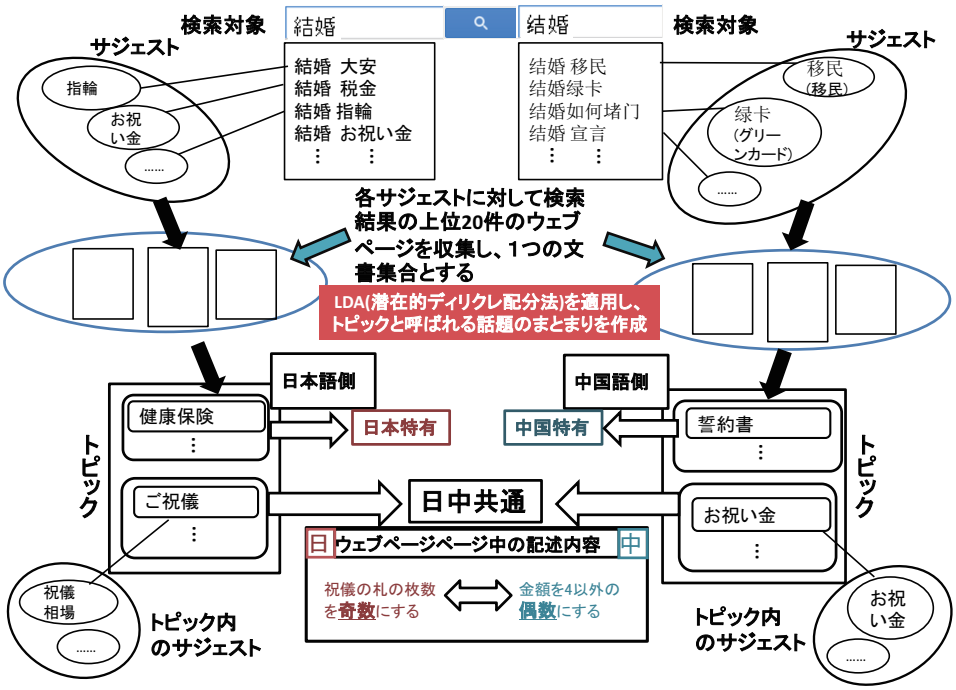

図 1: トピックモデルを用いた検索エンジン・サジェストの日中間対照分析

3 検索エンジン・サジェストの集約

3.1 概要

本節では，図 1 に示すように，トピックモデルを適用することにより，前節において収集したサジェストを自動的に集約し，トピックと呼ばれる話題毎にまとめる [2].

まず，Yahoo! Search BOSS API[2] に対して検索クエリを指定することにより，日本語のサイトと中国語のサイトを対象として，各言語とも上位 20 件のウェブページを別々に収集する．ここでの検索クエリは，検索対象「結婚」および前節において収集した各サジェストの AND 検索の形で作成する．収集された日中各言語のウェブページ集合をそれぞれ D_J および D_C として，D_J，D_C の各々を対象としてトピックモデル (本論文においては，LDA (Latent Dirichlet Allocation [1])) を適用することによって，日中各言語ごとにトピックを推定する．そして，推定されたトピックを用いることによって，各言語ごとにサジェストの集約を行う．

3.2 トピックモデル

本研究では，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 [1]) 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 $P(w|z_n)$ $(w \in V)$，及び，各文書 d におけるトピック z_n の確率分布 $P(z_n|d)$ $(n = 1, \ldots, K)$ を推定する．トピック数 K としては，$K = 50$(日本語側) および，$K = 40$(中国語側) を用いた．

3.3 文書に対するトピックの割り当て

本論文では，各ウェブページに対してトピックを一意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分類する．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．

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

これはつまり，ウェブページ d におけるトピックの分布において，確率が最大のトピックに，ウェブページ d を割り当てていることになる．

3.4 トピックに対する割り当てによる検索エンジン・サジェストの集約

各ウェブページは，検索対象「結婚」および各サジェストの AND 検索によって検索されたものである．したがって，あるウェブページには，一つ以上のサジェストが対応することになる．また，各ウェブページには，トピックが対応付けられている．以上のことから，一つのトピックに対して割り当てられた一つ以上のウェブページに対応するサジェストを収集することにより，一つのトピックに一つ以上のサジェストが割り当てられていることになる．実際に，検索対象「結婚」の場

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

表 2: 「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)

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

合，日本語側の 956 個のサジェストが 50 個のいずれかに割り当てられ，中国語側の 277 個のサジェストが 40 個のトピックのいずれかに割り当てられた．このことから，一般には，各トピックに対して複数のサジェストが対応しており，これによって，複数のサジェストが各トピックに集約されたとみなす．

4 日中間対照分析

4.1 検索エンジン・サジェストの日中間対照分析

本節では，2 節において収集されたサジェストを日中間で比較対照分析する．日本語側でのみ観測されたサジェスト数は 900 個であり，中国語側でのみ観測されたサジェストは 213 個であった．日中で共通に観測されたサジェストは，日本語サジェスト 56 個，中国語サジェスト 64 個であり，日中間では 55 個の対応組となった．一例として，日中共通で観測されたサジェストの例として，表 1 中の日本語側の「検索対象＋サジェスト」

結婚 年齢 法律

と中国語側の「検索対象＋サジェスト」

结婚 法定年龄(**結婚 法定年齢**)

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．一方，日本語側のみで観測されたサジェストの例としては，

結婚 住民税，　結婚 専業主婦，　結婚 農家

があり，中国語側のみで観測されたサジェストの例としては，

结婚 买房(**結婚 分譲住宅を買う**)
结婚 体检(**結婚 身体検査**)
结婚 二周年(**結婚 二周年**)

がある．

4.2 トピックの日中間対照分析

本節では，3 節で集約した日本語および中国語のトピックを対象として，日中間で比較対照分析を行う．図 1 に示すように，日本語特有のトピックの同定，中国語特有のトピックの同定，および，日中共通のトピックの対応付けを行った．その結果，日中共通のトピックとしては，表 2 の例に示すように，

「国際結婚の際に必要となる手続き」(日本語側)
—「国際結婚 (配偶者の国へ移民)」(中国語側)

「ご祝儀について」(日本語側)
—「お祝い金」(中国語側)

等の 11 個の対応組となった．例えば，表 2 中の日中共通トピック「国際結婚」の例では，日中共通に観測されたサジェストとして「グリーンカード」があり，このサジェストによって収集されるウェブページの記述内容のうち日中共通に観測された内容として,「アメリカ人と結婚する際のグリーンカードの申請方法」があった．一方，日中共通トピック「国際結婚」において，日本語側のみで観測されたサジェストとして「ミャ

表 3: 「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)

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	

ンマー人」があり，収集されるウェブページの記述内容としては,「ミャンマー人との結婚」に関するものがある．一方，日中共通トピック「国際結婚」において，中国語側のみで観測されたサジェストとして「移民」,「移民アメリカ」,「移民カナダ」があり，収集されるウェブページの記述内容としては,「アメリカ・カナダへの移民」に関するものがある．

また，表 3 中の日本語特有のトピックの例としては,「健康保険」,「離婚、婚約破棄と慰謝料について」，および,「結婚と六曜」等の 17 個が観測された．このうち，トピック「健康保険」においては,「住民税」,「税金」等のサジェストがあり，収集されるウェブページの記述内容としては,「結婚後の税金」に関するものがある．

一方，中国語特有のトピックとしては，表 4 の例に示すように,「結婚する際に必要な分譲住宅」,「誓約書」，および,「結婚式のゲーム」等の 20 個が観測された．このうち，トピック「結婚する際に必要な分譲住宅」においては,「结婚 买不起房(**結婚 分譲住宅を買えない**)」等のサジェストがあり，収集されるウェブページの記述内容としては,「分譲住宅を持っていない男にとって結婚は困難」というものがある．

以上の結果より，日中検索者の関心の違いおよび日中間の文化間差異を発見する手がかりが容易に得られることが分かった．

5 関連研究

文献 [4] 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た．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 [3] 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た．

表 4: 「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)

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)	

6 おわりに

本論文では，日本語および中国語の検索対象についてのサジェストを収集し，サジェストの自動集約を行うとともに，サジェストおよびトピックを日中二言語間で比較対照分析する手法を提案した．今後は，日中対訳知識を利用することにより，日中間のサジェストの対応付けを自動的に行う手法を確立する．

参考文献

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

[2] 土井俊弥, 井上祐輔, 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検索エンジン・サジェストの集約. 第 29 回人工知能学会全国大会論文集, 2015.

[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, 2013.

[4] 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.

[5] 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, 2014.