# ウェブ検索者の情報要求観点の日中間対照分析*

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

## 1 はじめに

現代の情報社会においては，情報の氾濫，すなわち，いわゆる情報爆発が起こっている．そして，そのように爆発する情報の集約や，俯瞰をするための技術の開発が強く望まれている．中でも，情報爆発が最も顕著に現れているのはウェブである．更に，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．本研究課題においては，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，本研究では，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．そこで，我々は，文献 [7] において，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た．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，本研究では，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，ウェブ検索者の情報要求観点からしか収集できない関心の動向や度合いが多数存在する．そこで，本研究では，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 1 に示す．

## 2 検索エンジン・サジェストからの情報要求観点の収集

### 2.1 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，本論文では，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．

*Comparative Analysis of Viewpoints of Web Search Information Needs between Japanese and Chinese

†Liyi Zheng, Daichi Koike, Tian Nie, Lei Chen, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Takakazu Imada, College of Engineering Systems, School of Science and Engineering, University of Tsukuba
§Yasuhide Kawada, Navix Co., Ltd.
¶Noriko Kando, National Institute of Informatics

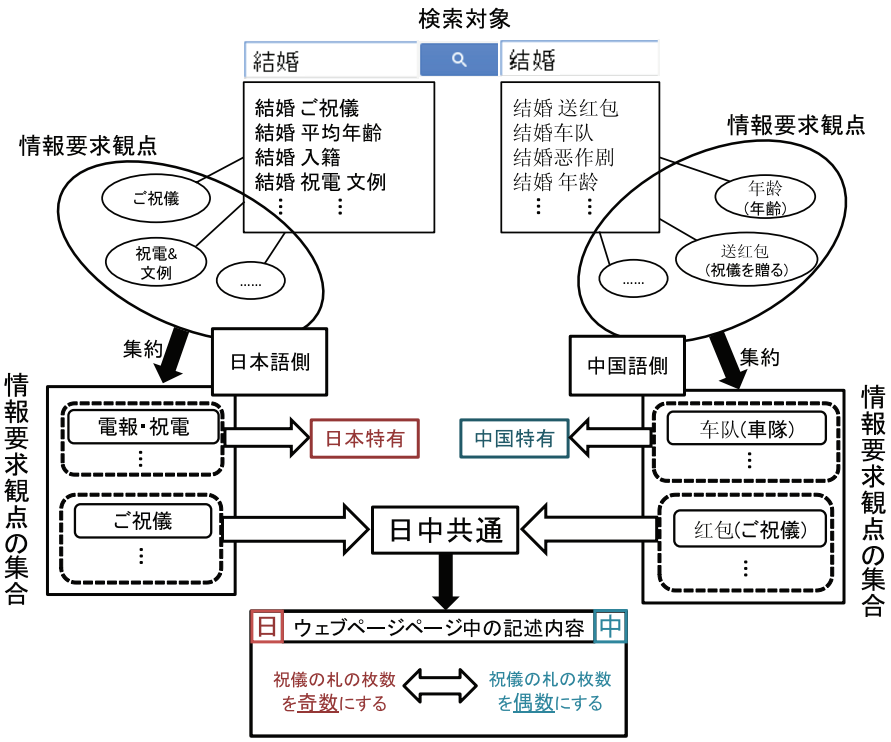


図 1: 検索エンジン・サジェストからの情報要求観点の収集及び日中間比較対照分析

### 2.2 日本語側の収集手順

本論文では，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 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．

### 2.3 中国語側の収集手順

本論文では，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．

## 3 情報要求観点の日中間対照分析

### 3.1 情報要求観点の集約

本節では，2 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．

表 1:「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 |

### 3.2 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 1 に示すように，日本語特有の情報要求観点集合の

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

表 3: 「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 | <=> | 結婚&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<br>男子：18歳以上<br>女子：16歳以上 | 中国の法律においては<br>男子：22歳以上<br>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 |

表 2: 「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 |

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 1 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 2 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．

## 3.3 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 3 の「検索対象＋情報要求観点」の欄に示す．

例えば，表 3 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」

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 3 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．

### 3.4 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 3 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．このことから，日中間の慣習の差異を容易に発見できることが分かった．

## 4 関連研究

本論文の先行研究として，我々は，文献 [7] において，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た．この方式の成果として,「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・言語間の違いを容易に観測することができた．一方，文献 [4] 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 [2] 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，本研究とは大きく異なる．

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 [5, 3, 6, 1] が挙げられる．文献 [5] 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 [3] 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 [6] 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 [1] 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．

## 5 おわりに

本論文では，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，Wikipedia 等を情報源とする日中対訳知識を利用することにより，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確立することが挙げられる．

## 参考文献


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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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, , July 2013.