

Web ページにおける文化差可視化システムの開発

Development of Cultural Differences Visualization System in Web Page

諏訪 智大[†] 宮部 真衣[‡] 吉野 孝[†]
Tomohiro Suwa Mai Miyabe Takashi Yoshino

1. はじめに

多言語間コミュニケーションにおいて、同一の単語を用いて会話をしている場合でも、相手の文化について十分に理解していないために、誤解が生じる可能性がある [1]。文化の違いに基づく誤解を回避するためには、用いている単語に文化差があることを、話者に認識させる必要がある。しかし、相手の文化に関する十分な知識が必要なため、文化差の有無の判断は容易ではない。

我々はこれまでに、文化差検出手法に関する検討を行ってきた。語句に関する文化差の有無に関しては、Wikipedia の項目数に基づく手法、記事内の国名・言語名数に基づく手法、記事における執筆者の意図に基づく手法を統合することにより、比較的高い精度での検出を可能にした [2]。さらに、コミュニケーションにおいて、ユーザが文化差のある語句を「適切に」理解することができるようにするには、各国の間でその語句にどのような差があるのかを理解させる必要がある。

そこで我々は、Wikipedia を用いた文化差可視化サービス SwaChica (スワッチカ) を開発した。本サービスでは、文化の違いを直感的に理解させるために、画像を利用している。しかし画像を提示する際、単純にある語句について画像検索を行い、提示すると、語句の持つ多様性により、意図した画像 (適切な画像) が得られない場合がある [3]。本稿では、文化差検出結果を適切に提示するための、画像検索エンジンと Wikipedia の太字部分を利用した画像検索手法について述べる。

2. 関連研究

画像検索に関するシステム開発や研究は、数多く行われている。画像検索の検索結果として表示する画像は、その画像が見つかったページのタイトルや本文、画像のファイル名などのテキスト情報をもとに判断している。また、画像内容を表す特徴量を考慮して、画像検索を行うシステムも提案されている。

例えば、獅々堀らは、検索語句に適した画像を Web 上から検索するために、画像知識 DB を用いたフィルタリングおよびユーザの画像選択のフィードバックを用いることで、既存の画像検索システム (Google 画像検索) による、画像の検索精度の改善を試みている [4]。出原らは、ユーザ自身がよく知らない物事についての情報を得るために画像検索を行う場合、検索結果として出力された画像が、正解かどうかわからない場合があることに着目した。検索結果の各画像の内容把握を助けるため、文章を元の Web ページから抽出し、補うことを試みている [5]。

本研究においても、画像の選択に、既存の画像検索エンジンの結果を利用する。本研究と既存研究の違いは、Wikipedia のデータを用いることで、概念を一意に特定し、1 つの記事内容と合う「適切な画像」を提示する手法を提案している点である。



図 1: 文化差可視化サービスにおける文化差可視化ページ

3. 文化差可視化サービス SwaChica

文化差可視化サービス SwaChica は、語句を入力すると、Wikipedia で検索語句に関する記事を特定する。そして、その記事内容から文化差の有無を調べ、結果を可視化するサービスである。図 1 のように、それぞれの文化における検索語句についての画像を提示することにより、どのような違いがあるのかを直感的に理解させることを目指している。

現在、SwaChica の提示する画像は、Google 画像検索の検索上位の画像を利用している。文化差検出 [2] では Wikipedia を用いて文化差の有無を調べているため、Wikipedia 記事内の画像を、提示する画像として利用する方法も考えられる。しかし、1 つの画像のみでは固定的なイメージを与える可能性がある。そこで、多様性を持たせるために、本サービスでは画像検索サービスを利用することとした。

4. 画像検索手法の問題点と手法の提案

4.1 現在の画像検索手法の問題点

現在、SwaChica では、Google 画像検索による検索上位の画像を利用している。しかし、単純に検索語句について画像検索を行うと、語句の持つ多義性により、Wikipedia の記事内容とは関係がない画像を含めて提示してしまう場合がある。我々は、意図しない画像を提示することを避けるために、以下の画像検索手法を検討してきた [3]。

- 手法 1 : 画像検索エンジンの多数決方式
- 手法 2 : 検索語の追加による画像検索方式
- 手法 3 : 類似画像検索エンジンによる画像検索方式

[†] 和歌山大学システム工学部, Faculty of Systems Engineering, Wakayama University

[‡] 東京大学知の構造化センター, Center for Knowledge Structuring, The University of Tokyo

手法1の多数決方式では、多数決を行ったとしても、検索上位に存在する、不適切な画像を取り除くことができなかった。検索語句を追加する手法2では、検索語句を追加することで、検索語句の概念を絞ることができ、意図する画像を提示できる可能性を示した。類似画像検索エンジンを用いた手法3では、画像検索結果の上位は、Wikipediaの画像が多く、多様性をもった結果を提示することができなかった。そこで本研究では、手法2を応用した手法を提案する。

4.2 Wikipediaの太字部分を用いた画像検索手法の提案

4.1節で述べたように、検索語句の追加(手法2)により、語句の意図する画像を提示できる可能性がある。しかし、これまでの検討においては、追加する検索語句を手動で選択していた。そこで、追加検索語句を自動的に抽出し、画像を検索する手法を提案する。本提案手法では、追加検索語句として用いる語句として、Wikipediaに登場するBタグ¹で囲まれた太字部分に着目した。Wikipediaにおける太字部分は、その記事内容の語句と同じ概念、または単なる言い換えであると考えた。そこで、1つ目のパラグラフは、記事の見出しについて何を表すのか大まかに説明している文であることを利用する。1つ目のパラグラフでは、検索語句(見出し)が太字で表示されている。1つ目のパラグラフ内に見出し以外の太字部分が存在する場合、その太字部分は、検索語句と特に近い概念を表している可能性がある。提案手法における画像検索手順を以下に示す。

- (1) 1つ目のパラグラフから、記事の見出しを抽出する。
- (2) 追加の検索語句として、1つ目のパラグラフに含まれる見出しとは異なる太字の語句を抽出する。
- (3) 「検索語句 追加の検索語句」でGoogle画像検索をする。

ただし、1つ目のパラグラフに記事の見出し以外の太字部分がない場合、本提案手法を適用することはできず、そのまま検索語句のみで検索することになる。日本人によって文化差があると想定された2209語句[6]について調査を行った結果、曖昧さ回避ページが存在する語句は408語句であった。また、曖昧さ回避ページおよび見出しとは異なる太字部分がともに存在する語句は182語句であった。曖昧さ回避ページが存在する語句は、本提案手法の適用対象である、多義性を持つ語句である。つまり、本提案手法は、多義性を持つ語句のうち、45%(408語句中182語句)について適用可能である。

5. 評価実験

4章で述べた手法と、既存手法による画像検索の比較実験を行った。既存手法とは、単純にWikipediaの見出しでGoogle画像検索する手法である。実験を行うにあたって、実験に用いる語句を以下の基準で選択した。

- (1) Wikipediaに「曖昧さ回避ページ」など(<div class="dablink noprint">)が存在している。
- (2) 記事の1つ目のパラグラフに2つ目の太字部分が存在している。

今回、この条件は、日本語版Wikipediaのみを対象とした。この2つの条件をもとに、過去にSwaChica上で入力された語句と、筆者がWikipediaで無作為に選択した語句のうち、上述した2つの基準を満たした30語句を用いた。

本実験では、それぞれの手法で検索された、上位5枚の画像についてアンケートを作成し、回答してもらった。ただし、直接リンクが禁止されている画像は、予め検索結果から取り除いている。アンケートの質問項目は以下の3つである。

¹HTMLタグの1つで、タグで囲まれたテキストは太字になる。



図2:「ラーメン」の検索結果



図3:「ラーメン 拉麺」の検索結果

- (1) それぞれの画像は適切に語句を表している。
- (2) 5枚の画像は多様な結果を表している。
- (3) 「追加の検索語句」は検索語句の別名として妥当である。

評価者は、中国人5名と日本人5名で、中国人には中国語における結果を、日本人には日本語における結果を提示した。また、(3)については、日本人のみに回答してもらった。

6. 実験結果と考察

日本人による画像の適切性についてのアンケート結果を表1に、多様性についてのアンケート結果を表2に示す。中国人による画像の適切性、多様性についてのアンケート結果を表3に示す。また、追加の検索語句の妥当性についてのアンケート結果を表4に示す。

表1および表2から、日本語のWikipediaに提案手法を適用した結果、一部、適切性の評価が低くなったが、全く異なる概念を表す画像の提示を抑制し、さらに多様性を持った画像の提示を実現した。表3から、中国語のWikipediaに提案手法を適用した結果、適切性、多様性ともに良い評価が得られた。

6.1 画像の適切性について

表1の提案手法について、30語句中14語句の適切性が低くなった。しかし、既存手法における「天津飯」は精度が悪く、漫画に登場する架空の人物の画像が、いくつか提示された。それに対して、提案手法は、全く異なる概念である語句の画像提示を抑制することができた。

表1より、「リンゴ」「花」「オレンジ」については、表4より、追加の検索語句としての妥当性が低かった。「リンゴ」の追加検索語句である「セイヨウリンゴ」は植物学における植物名を表し、いくつか植物の画像が提示されていた。また「花」の追加検索語句である「華」についてYahoo!辞書²で調べた結果、「華やかなこと、はでなこと」という意味であり、概念的な意味を含むため、不適切な画像が提示されたと考えられる。「オレンジ」の追加検索語句である「アマダイダイ」は「オレンジ」の和名であるが、色としての「オレンジ」を表す画像が提示された。この追加の検索語句では、概念を絞ることができず、逆に、広がりを持たせてしまい、オレンジ色の髪の毛の人や、オレンジ色の飲み物の画像をいくつか提示した。

「ラーメン」の追加検索語句である「拉麺」により、店としてのラーメンの画像が提示された。また、表3より、中国語における「ケチャップ」の追加検索語句である「茄汁」(トマト・ソースという意味)により、トマト・ソースを使った料理の画像が多く提示された。

²Yahoo!辞書: <http://dic.yahoo.co.jp>

表 1: 日本人による適切性評価結果

ID	検索語句	既存手法		提案手法		結果
		最頻値	中央値	最頻値	中央値	
1	天津飯	1	4	5	4	○
2	マグロ	4	4	4	4	○
3	リンゴ	5	5	1	2	×
4	タンポポ	4	4	4	4	○
5	マナー	4	4	1	2	×
6	スギ	4	4	4	4	○
7	サンド	4	4	5	4	○
8	イチ	4	4	1	1	×
9	オレンジ	4	4	1	1	×
10	アイス	4	4	4	4	○
11	クリーム	4	4	4	4	○
12	筆	4	4	4	4	○
13	電子メール	4	4	2	2	×
14	ホット	4	4	4	4	○
15	ドッグ	4	4	4	4	○
16	サソリ	5	5	4	4	×
17	コミュニ	3	3	1	2	×
18	ケーション	4	4	4	4	○
19	サクラ	4	4	4	4	○
20	チュー	4	4	5	4	○
21	リップ	4	4	5	4	○
22	タコ	4	4	4	4	○
23	ラーメン	5	4	1	1	×
24	親子丼	4	4	4	4	○
25	ケチャップ	4	4	5	4	○
26	指輪	4	4	4	4	○
27	餃子	4	4	4	4	○
28	チーズ	4	4	4	3	×
29	カエル	5	5	4	3	×
30	ウサギ	5	5	5	3	×
31	花火	5	5	5	4	×
32	バレン	4	4	3	4	×
33	タインデー	4	4	3	4	×
34	クモ	4	4	4	3	×
35	ネギ	4	4	4	4	○

※表中の値は、「1: 強く同意しない」「2: 同意しない」「3: どちらでもない」「4: 同意する」「5: 強く同意する」の5段階の評価基準による。○は、評価が変わらなかった、または上がった語句を表す。×は、評価が下がった語句を表す。

他にも、「マナー」「電子メール」「コミュニケーション」などは、いずれも画像として表すのが難しく、概念的な語句であり、追加の検索語句で AND 検索して概念を絞ることによって、あまり関係のない画像が多数提示された。

6.2 画像の多様性について

表 2 より、Wikipedia の太字部分を追加の検索語句として用いることで、30 語句中 23 語句に対して、多様性を持った画像の提示を実現した。

今回、「(検索語句) (追加の検索語句)」の AND 検索では、Google 画像検索を利用している。したがって、検索結果として提示される画像は絞られ、検索結果の数は少なくなる。既存手法では、他の概念を表す画像が提示される可能性があり、提案手法により、検索語句の概念を限定している。

したがって、提案手法では、検索語句と全く異なる概念の画像を提示することは見られなかった。しかし、同じ概念を表すが、適切性が低い画像を提示するケースが多くなった。提案手法で提示された画像は、多様性を持つことで、適切性の評価が少し下がったと考えられる。

また、既存手法で、適切性の評価が高かったのは、同じような画像が上位に提示されたためであると考えられる。例えば、図 2 より「ラーメン」で検索すると、一杯のラーメンが写った

表 2: 日本人による多様性評価結果

ID	検索語句	既存手法		提案手法		結果
		最頻値	中央値	最頻値	中央値	
1	天津飯	4	4	2	2	×
2	マグロ	4	3	4	4	○
3	リンゴ	1	2	4	4	○
4	タンポポ	2	2	4	4	○
5	マナー	5	4	4	4	×
6	スギ	3	3	4	4	○
7	サンド	4	4	4	4	○
8	イチ	4	4	2	4	×
9	オレンジ	2	3	1	2	×
10	アイス	3	3	4	4	○
11	クリーム	2	2	4	4	○
12	筆	4	4	4	3	×
13	電子メール	4	4	4	3	×
14	ホット	1	2	2	2	○
15	ドッグ	1	1	2	3	○
16	サソリ	4	4	4	4	○
17	コミュニ	2	3	4	4	○
18	ケーション	2	3	4	4	○
19	サクラ	4	3	3	3	×
20	チュー	4	3	3	3	×
21	リップ	2	3	4	4	○
22	タコ	2	3	4	4	○
23	ラーメン	4	4	4	4	○
24	親子丼	2	2	2	3	○
25	ケチャップ	2	2	4	4	○
26	指輪	2	2	4	4	○
27	餃子	2	3	4	2	○
28	チーズ	4	4	4	3	×
29	カエル	2	2	4	4	○
30	ウサギ	2	2	3	3	○
31	花火	2	3	4	4	○
32	バレン	4	4	4	4	○
33	タインデー	4	4	4	4	○
34	クモ	4	3	4	4	○
35	ネギ	2	2	3	3	○

※表中の値は、「1: 強く同意しない」「2: 同意しない」「3: どちらでもない」「4: 同意する」「5: 強く同意する」の5段階の評価基準による。○は、評価が変わらなかった、または上がった語句を表す。×は、評価が下がった語句を表す。

画像が上位に提示される。図 3 より、検索語句を追加することで、ラーメン屋の看板や、ラーメン屋のラーメン以外の商品などが提示される場合が見られた。

今後は、多様性を持った画像を提示しつつ、適切性な画像を提示する検索手法を検討する必要がある。

6.3 追加の検索語句について

表 4 より、検索語句の別名や、英語での言い換え、表記揺れなどは、追加の検索語句としての妥当性が高いという評価が得られた。しかし、一部、妥当性が低い語句が見られた。「茹で餃子」については、別名ではなく、調理の形態による餃子の呼ばれ方の変化を表し、評価が低くなった。「セントバレンタインズデー」については、バレンタインデーの別名であるが、一般的に浸透しているとは言えず、評価が低くなった。他に、追加の検索語句の特徴として見られたものに、「マグロ属」「タンポポ属」「タコ目」などの分類を表す語句の他、学名や和名を表す語句が多く見られた。

このように、追加の検索語句として用いる太字部分は、専門用語を表す場合や、一般的に用いられていない語句を提示するときに用いられる場合も見られる。しかし、表 1 より、これを利用して、適切性の評価が低くなっているわけではない。Wikipedia の太字部分を検索語句として追加することは、概念

表 3: 中国人による評価結果

ID	検索語句	適切性		多様性	
		最頻値	中央値	最頻値	中央値
1	天津飯	4	3	4	4
2	マグロ	4	3	4	4
3	リンゴ	5	5	2	3
4	タンポポ	1	2	2	4
5	マナー	4	4	4	4
6	スギ	4	4	3	3
7	サンドイッチ	3	3	4	4
8	花	5	5	3	3
9	オレンジ	4	4	2	3
10	アイスクリーム	5	4	4	3
11	筆	4	4	4	4
12	電子メール	3	3	2	3
13	ホットドック	5	4	4	4
14	サソリ	5	5	4	4
15	コミュニケーション	4	4	3	3
16	サクラ	4	4	4	4
17	チューリップ	4	4	4	4
18	タコ	4	4	4	3
19	ラーメン	5	4	2	3
20	親子丼	4	4	4	4
21	ケチャップ	2	2	3	3
22	指輪	5	5	4	4
23	餃子	5	4	4	4
24	チーズ	3	4	4	4
25	カエル	2	2	4	4
26	ウサギ	4	4	3	3
27	花火	4	4	4	4
28	バレンタインデー	4	4	5	4
29	クモ	4	4	4	4
30	ネギ	5	4	3	3

※表中の値は、「1: 強く同意しない」「2: 同意しない」「3: どちらでもない」「4: 同意する」「5: 強く同意する」の5段階の評価基準による。

を限定できていると考えられる。今後は、追加する検索語句の妥当性をさらに考慮し、適切な画像を提示する手法を検討する必要がある。

7. おわりに

今回、Wikipedia を用いた文化差可視化サービスにおける文化差検出結果提示のための画像検索手法を提案した。

本研究の貢献は、以下の3点にまとめられる。

- (1) Wikipedia の太字部分を追加検索語句として用いる画像検索手法を提案した。
- (2) 提案手法を用いることによる、適切かつ多様な画像の提示可能性を示した。
- (3) 提案手法を用いることによる、概念が異なる画像の提示を抑制する可能性を示した。

今後は、追加の検索語句としての妥当性を考慮した上での検索語句の生成を検討する必要がある。今回、提案手法は「曖昧さ回避ページ」などが存在し、さらに追加の検索語句としての太字部分が存在する場合のみを想定した。そのため、多義性はあるが、Wikipedia に「曖昧さ回避ページ」などが存在していない語句には適用できない。したがって、「曖昧さ回避ページ」などが存在していない場合における、多義性の有無の判定手法や、太字部分が存在していない場合でも対応できる手法を検討する必要がある。また、多義性の有無に関わらず、太字部分のみが存在する場合でも適用できる可能性があり、検討する必要がある。

今後は、より広い範囲の語句を対象として、画像検索手法の検討を進める。

表 4: 追加の検索語句の妥当性のアンケート結果

ID	検索語句	追加の検索語句	最頻値	中央値
1	天津飯	天津丼	2	3
2	マグロ	マグロ属	2	2
3	リンゴ	セイヨウリンゴ	2	3
4	タンポポ	タンポポ属	2	2
5	マナー	作法	5	5
6	スギ	杉	4	4
7	サンドイッチ	サンドウィッチ	4	4
8	花	華	2	2
9	オレンジ	アマダイダイ	1	4
10	アイスクリーム	ice cream	4	4
11	筆	毛筆	4	3
12	電子メール	e-mail	4	4
13	ホットドッグ	ホットドック	5	5
14	サソリ	蠍	4	4
15	コミュニケーション	Communication	4	4
16	サクラ	桜	4	4
17	チューリップ	鬱金香	2	4
18	タコ	タコ目	2	2
19	ラーメン	拉麺	4	4
20	親子丼	親子どんぶり	4	4
21	ケチャップ	トマト ケチャップ	4	4
22	指輪	指環	5	4
23	餃子	茹で餃子	2	3
24	チーズ	乾酪	3	3
25	カエル	蛙	4	4
26	ウサギ	兎	2	4
27	花火	fir works	2	2
28	バレンタインデー	セントバレン ティンズデー	2	2
29	クモ	蜘蛛	4	4
30	ネギ	葱	4	4

※表中の値は、「1: 強く同意しない」「2: 同意しない」「3: どちらでもない」「4: 同意する」「5: 強く同意する」の5段階の評価基準による。

謝辞

本研究の一部は、独立行政法人科学技術振興機構研究成果最適支援事業 (A-STEP) 探索タイプ「Wikipedia と機械翻訳を用いた多言語用語間における文化差検出サービス」および日本学術振興会科学研究費基礎研究 (B) (22300044) の補助を受けた。

参考文献

- [1] 藤井薫和, 重信智宏, 吉野 孝: 機械翻訳を用いた異文化空間チャットコミュニケーションにおけるアノテーションの評価, 情報処理学会論文誌, Vol.48, No.1, pp.63-71 (2007) .
- [2] 吉野 孝, 宮部真衣, 諏訪智大: Wikipedia を用いた文化差可視化サービスの開発, 情報処理学会, マルチメディア, 分散, 協調とモバイル (DICOMO2011) シンポジウム, pp.2287-2294 (2012) .
- [3] 吉野 孝, 宮部真衣, 諏訪智大: 文化差検出結果提示のための画像選択方法の検討, FIT2012, J-044 (2012) .
- [4] 獅々堀正幹, 小泉大地, 柘植 寛, 北研二: 画像知識データベースを用いた WWW 画像検索システムの開発, 電子情報通信学会論文誌 D-I, J87-D-I(2), pp.154-163 (2004) .
- [5] 出原 博, 藤本 典幸, 竹野 浩, 荻原 兼一: WWW 画像検索における画像周辺の HTML 構文構造を考慮した画像説明文の抽出手法, 電子情報通信学会技術研究報告, データ工学 105(338), pp.19-24, (2005) .
- [6] 吉野 孝, 宮部真衣: 日本語版 Wikipedia における執筆者の意図に着目した日中間の文化差検出手法の検討, 電子情報通信学会, 言語理解とコミュニケーション, NLC2011-58, pp.13-18 (2012) .