Wikipediaにおける言語間の差異マイニング

森竜也[1*]　増田英孝[1†]　中川裕志[2‡]　清田陽司[2§]

[1] 東京電機大学大学院未来科学研究科¶
[2] 東京大学情報基盤センター‖

1　はじめに

WikipediaはWeb上で自由に編集が行えるフリー百科事典である．通常の百科事典と大きく異なるのは，一般のユーザ自身が記事を作成してコンテンツを育てていく点である．またWikipediaには多数の言語版が存在し，2010年1月現在では250を超える言語版が運営されている[1]．各言語版は他とは独立しており，ユーザは自分が使用する言語の版に対して記事を作成する．一方で記事の主題はユーザの自由に任されているので，ある事柄についての記事が特定の言語版にのみ存在していたり，ある話題に関する記事が他に比べて特に多く作成されている言語版があるなど，各言語版はそれぞれが異なる成長をしている．本研究ではそのような言語間の差異を言語およびその言語を使う国や文化による興味・関心の違いとしてとして捉え，比較を行うシステムを作成している．言語による興味や関心の違いという評価しにくい性質に対して指標を与えることが目的である．

本稿では作成する差異の比較システムのアルゴリズムを説明する．また実際のWikipediaのデータを用いて比較システムを実行し，システムの効果と問題点の分析を行う．

2　Wikipedia

2.1　Wikipediaのページ

Wikipediaのページは役割によっていくつかの種類がある．ページは種類ごとの名前空間とタイトルを持つ．同一の名前空間内ではタイトルの重複は許されない．特に重要なページの種類が記事とカテゴリである．

記事とは通常の百科事典における項目に相当するページである．1件の記事で1つの事柄について記述される．

カテゴリとは他のページの分類をするためのページである．ページにはユーザが自由にカテゴリを複数設定できる．また新しいカテゴリを作成することも可能である．カテゴリにカテゴリを設定することで階層的なカテゴリ構造が構築されている．

2.2　言語間の差異

Wikipediaの全言語版の総記事数は1000万件を超える．最も記事数が多い版は英語版で約315万件，日本語版は6番目に多く約64万件となっている．しかし日本語版は英語版の抜粋を翻訳したものではなく，日本語を使用するユーザによって独自に作られてきた．そのため相互に共通して持っているページと，一方しか持っていないページがある．これは他の言語版との間にも言えることである．

あるページと他の言語版に存在する同一概念について記述されたページは，言語間リンクという特別なハイパーリンクで接続されている．図1はWikipediaのページ構造の例である．直線の矢印がカテゴリ関係，破線の矢印が言語間リンクを示している．“行政”と“Government”の例のように，言語間でカテゴリ関係が常に一致するとは限らない．

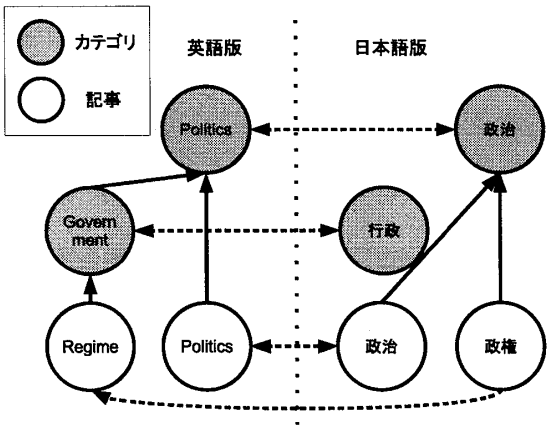

図1: Wikipediaの構造例

3　言語間の比較システム

3.1　比較対象

本研究において言語版による興味・関心の違いとは，ある話題に属する記事の数の差異としている．具体的にはあるカテゴリに属する記事の件数の差異のことで

Mining Differences in Multilinguality of Wikipedia
*Tatsuya Mori
†Hidetaka Masuda
‡Hiroshi Nakagawa
§Yoji Kiyota
¶Graduate School of Science and Technology for Future Life, Tokyo Denki University
‖Information Technology Center, University of Tokyo

ある．カテゴリはページを分類するための Wikipedia の公式のシステムであり，カテゴリの設定は人手で行われるので，信頼性の高い話題の分類が獲得できることを期待している．図 1 に示したようなページ構造を言語版を跨いで比較するには，ページ名，ページの種類（記事あるいはカテゴリ），属するカテゴリ，言語間リンクのデータが必要である．

Wikipedia では全ページの内容を含んだ XML ファイルを言語版ごとに公開している．XML ファイルを解析して上記のデータを全てのページから抽出し記録しておく．今回対象とする言語版は日本語，中国語，ドイツ語，フランス語とした．これらの言語版は記事数が多く，地理的・文化的に共通点と相違点があり，比較対象として効果的であると考えたためである．

3.2 アルゴリズム

本システムでは指定された言語版のカテゴリを起点としてページ構造を再帰的に探索し，カテゴリに属する記事とそのタイトル，およびカテゴリの下位カテゴリとそのタイトルを全て収集する．探索アルゴリズムは以下のとおりである．

1. 指定されたカテゴリを識別する．
2. カテゴリに属する記事を探索結果に追加する．
3. カテゴリの下位カテゴリを取得する．
4. 3 で取得したカテゴリと 1 のカテゴリの名前の類似度が一定の値以上だった場合，3 のカテゴリを探索結果に含め，1 へ戻り再帰的に探索を続ける．類似度が一定値に満たない場合，探索を打ち切る．
5. 下位カテゴリがなくなれば探索を終了する．

アルゴリズム 3 において，カテゴリ同士のタイトル文字列の類似度を算出し，探索の打ち切りを行っている．カテゴリはユーザによって自由に複数設定できるため，関連性の希薄なカテゴリが設定されることがある．その場合，無条件に探索を続けていると始めのカテゴリの話題とはほぼ無関係な記事が結果に含まれてしまう．そこで意味的な関連性の指標としてタイトル文字列を取り，以下の式 1 で類似度を算出し，設定した閾値との比較で探索を続けるか判定を行う．

$$\text{類似度} = \frac{|s(b(T_1), b(T_2))|}{\sqrt{|b(T_1)| \cdot |b(T_2)|}} \quad (1)$$

式中の T_1, T_2 は類似度を算出する文字列である．b は与えた文字列の bi-gram を得る関数である．また s は与えた 2 つの bi-gram のうち共通するものを得る関数である．今回は閾値を実験的に 0.25 とした．

さらに比較システムに問い合わせをし，結果を表示するための Web インタフェースを作成した．

4 評価

表 1 は作成した比較システムを使用して，日本語版の “欧州連合” カテゴリを探索した結果である．図 2 はその時の Web インタフェースである．収集したページの数とそれらのタイトルが閲覧できる．ドイツ語版とフランス語版にページが多く，文化的・政治的な背景が Wikipedia に反映されている例と言える．

表 1: “欧州連合” の比較結果

言語	記事数	カテゴリ数
ドイツ語	1102	46
フランス語	2618	278
日本語	220	22
中国語	32	4

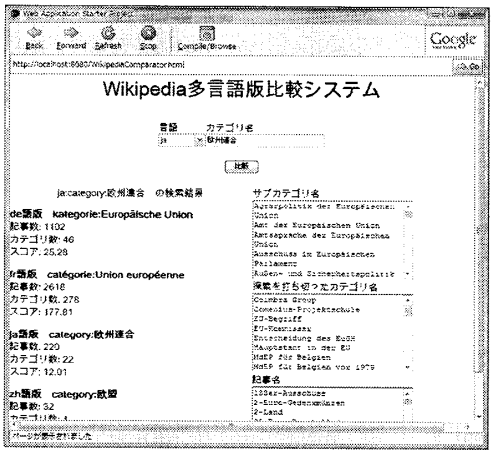

図 2: 比較システムの Web インターフェース

探索結果が不適切な例として，探索アルゴリズム 4 で探索を打ち切ったカテゴリの中に，関連性が高く本来打ち切るべきでないものが含まれている場合があった．タイトル文字列の比較だけではページ間の意味的な関連性を測るには不十分であることが分かった．

5 おわりに

本稿では多言語に展開する Wikipedia の言語間の差異を分析し，興味・関心の指標として提示するシステムを作成した．今後の課題としては探索に含めるページの判定の精度を高めることがある．そのために記事に関するデータをまとめる表であるテンプレートや，同種のページのリストである一覧ページなどを利用することを考えている．

参考文献

[1] Wikimedia Foundation. Wikipedia:全言語版の統計. http://ja.wikipedia.org/wiki/Wikipedia:全言語版の統計.