

多言語に展開する Wikipedia の特徴の比較調査

Characteristic Investigation and Comparison among Multilingual Wikipedias

森竜也^{1*} 増田英孝¹ 清田陽司² 中川裕志²
Tatsuya Mori¹ Hidetaka Masuda¹ Yoji Kiyota² Hiroshi Nakagawa²

¹ 東京電機大学大学院未来科学研究科

¹ Graduate School of Science and Technology for Future Life, Tokyo Denki University

² 東京大学情報基盤センター

² Information Technology Center, University of Tokyo

Abstract: 本研究では Web 上のフリー百科事典 Wikipedia の多言語展開に着目し、各言語版を特徴づけるデータを抽出し、比較を行うシステムを作成した。作成システムによって、各言語版におけるカテゴリの発展の度合いを比較できる。またシステムに問い合わせを行う Web インターフェースを作成した。

1 はじめに

Web 上のフリー百科事典 Wikipedia には通常の百科事典にはない様々な性質が存在する。そのため単なる読み物としてだけでなく、機械処理可能な巨大な文書群としても利用されている。Wikipedia から獲得できる情報には、広範なトピックスに渡る膨大な記事とそれらの分類、ハイパーリンクによる記事同士の参照などがある。さらに重要な性質として Wikipedia は 250 を超える世界中の言語で展開されているという点がある。従来、この性質は主に異なる言語版における同一概念の記事を特定するために利用されてきた。本研究は Wikipedia の多言語展開のうち、他の言語版と同一ではない部分に着目し、各言語版 Wikipedia の特徴を浮かび上がらせることを目的としている。記事数の多いいくつかの言語版の Wikipedia を対象として、各版を特徴付けるデータを抽出し、他の版との比較を行うシステムを作成した。作成システムはカテゴリを対象として各言語版における発展の度合いを提示する。また作成した比較システムを Web 上から利用するためのインターフェースを作成した。

2 Wikipedia の多言語展開

2.1 多言語展開の現状

Wikipedia プロジェクトには 2009 年 10 月の時点で 271 の言語版が存在していて、記事の総計は 1000 万件

以上である [1]。ただしそれらのうち継続的な活動がされているのは 170 言語版程度であり、記事数が 10 万件を超えているのは 30 言語版に満たない。最も記事数が多い言語版は英語版で、300 万件を超えている。日本語版は 5 番目に記事数が多い版で、記事数は約 62 万件となっている。表 1 は記事数の多い言語版トップ 15 をまとめたものである。

表 1: 各言語版 Wikipedia の統計 (2009 年 11 月 12 日現在)

記事数の順位	言語	記事数
1	英語	3,081,378
2	ドイツ語	974,651
3	フランス語	868,503
4	ポーランド語	646,911
5	日本語	627,803
6	イタリア語	620,223
7	オランダ語	567,655
8	スペイン語	524,560
9	ポルトガル語	516,608
10	ロシア語	448,285
11	スウェーデン語	334,754
12	中国語	281,902
13	ノルウェー語	233,915
14	フィンランド語	220,258
15	カタルーニャ語	207,344

*E-mail: mori@cdl.im.dendai.ac.jp

2.2 多言語展開の性質を利用した先行研究

Wikipedia では記事に対して言語間リンクという特別なリンクを設定することができる。言語間リンクは他の言語版の Wikipedia に存在する同一あるいは対応する概念の記事へのナビゲーションとして動作するリンクである。例えば日本語版の記事“政権”には英語版“Regime”, フランス語版“Régime,” 中国語版“政权”といった言語間リンクが設定されている。言語間リンクは人手で設定されるだけでなく、ロボットによる自動設定も行われている。ロボットはリンク先を再帰的に解析し、未設定の言語間リンクを発見した場合に元の記事に追加する機能を持っている。例えば日本語版のある記事が英語版への言語間リンクを持っていて、英語版において新しい言語間リンクが追加された場合、日本語版にも追加されることになる。

Wikipedia の多言語展開を利用した先行研究では、主に言語間リンクから他の版における記事名を得ている。Auer らによって開発されている DBpedia[2] はセマンティック Web のための知識ベースを Wikipedia から生成する試みである。DBpedia では英語版 Wikipedia の記事名が基本となっているが、言語間リンクによって他の言語版の情報を獲得している。

言語間リンクから語の翻訳関係を獲得する試みもある。基本的な考え方は、言語間リンクの先にある記事のタイトルを翻訳語とみなすことである。上記の例であれば“政権”の英訳が“Regime”である、という翻訳関係が見出される。言語間リンクは人の判断で設定されることに加えて、収集する際に自然言語処理の手法を用いる必要がなく、高い精度の翻訳語が期待できる。Erdmann らの研究 [3] では言語間リンクを起点として対応する語を取得し、さらにリダイレクトやリンクテキストを収集することで適切な翻訳語を獲得している。

3 各言語版の比較

3.1 各言語版の差異

今まで述べたように Wikipedia は多言語で展開されている。しかし複数の言語版が存在するという意味が Wikipedia の場合、通常の事典や書籍とは異なる。通常の事典の場合、ある言語で書かれた底本が翻訳されることで他の言語版が作られるため、内容に意味的な違いはない。しかし Wikipedia の場合はそれぞれの版が独立・並行してユーザに編集されているため、版ごとに内容に自然と差異が生じる。

表 1 に示した通り、言語版によって存在する記事の数が異なる。日本語版は英語版の約 5 分の 1 の記事数であるが、日本語版に存在するすべての記事を英語版が持っているわけではない。日英両方に存在する記事

と、日本語版にしか存在しない記事、英語版にしか存在しない記事がある。

また Wikipedia では記事を階層的にカテゴリ分類できるが、この分類も版によって違いがある。記事の場合と同様に存在するカテゴリそのものに違いがあるほか、属する記事や下位のカテゴリが異なる場合がある。

言語間リンクから同一概念を獲得する場合には、言語間リンクは版を跨いで等価な情報を示すものと期待されている。しかし版によって異なる言語、異なる人物によって編集されているため、多言語展開には意味的に異なる情報も含まれている。

3.2 差異の比較と調査

意味的に等価な情報を獲得する研究においては、版ごとの差異は好ましいものではない。語の対応関係が正しく取れない、ないし全く獲得できないことがあるためである。しかし本研究では版ごとの差異を、その言語を使う文化や歴史を特徴づけるものとして捉え、積極的に活用する。例えば日本語版にだけ存在し言語間リンクを持たない記事は、日本に特有の話題を扱ったものであるか、世界に先駆け日本で注目されているものであると予想する。またあるカテゴリに着目したとき、そこに属する記事数が他に比べて日本版で特に多い場合も、日本において重要な意味を持っているものと考えられる。本研究の目的は、記事数の多いいくつかの言語版の Wikipedia を対象として、版ごとにその版を特徴付けるデータを抽出し、他の版のデータとの比較を行うシステムを作成することである。

4 各言語版の比較システム

4.1 利用するデータ

Wikipedia では言語毎に全内容を含んだデータファイルが無償で公開しており、自由にダウンロードして利用できる。本研究では日本語のほかに、ドイツ語、フランス語、中国語の Wikipedia を対象とする。日本と中国は東アジア、ドイツとフランスは西ヨーロッパという地理的・文化的な要因が Wikipedia にどのように反映されているか調べるためである。またこれらの言語版は記事数が多いことも理由である。使用するデータファイルは 2009 年 10 月時点で公開されている最新版である。英語版は表 1 に示したように他の言語版に比べて圧倒的に記事数が多いが、本研究では使用していない。英語版は記事数が広範囲に渡ってとても多く、Wikipedia の標準的な言語版のような存在になっている。そのため却って言語特有の情報が得にくくなっているため、本研究では使用しない。

表 2 は使用する 4 つの言語版のエントリ数の統計である。記事数の違いだけでなく、カテゴリの数と割合にも違いがあることが読み取れる。ドイツ語は記事数に対してカテゴリが少ないが、フランス語や中国語ではカテゴリが多い。

4.2 Wikipedia データの解析

公開されているデータファイルを解析し、Wikipedia 内の構造を表現する次のデータをエントリごとに取得しておく。解析には筆者が開発した Wikipedia データ解析アプリケーションである Wik-IE[4] を使用した。Wik-IE は Hadoop[5] を利用することで巨大な Wikipedia データファイルを分散処理することができる。また解析結果のデータは全文検索システム Lucene[6] のインデックスとして保存しておく。

- エントリ名
- エントリの種別（記事あるいはカテゴリ）
- 上位カテゴリ
- 言語間リンク

エントリとは記事やカテゴリなど Wikipedia におけるページの総称である。Wikipedia は Web 上のプロジェクトであり、エントリ名が URL で記述されるため、ユニークであることがあらかじめ保証されている。そこでエントリ名をエントリを識別するためのキーとして使用する。今回は 4 つの言語を対象としているが、Wikipedia で使用されている Wiki システムはどの版でも同じであり、データファイルの形式も共通なため、必要に合わせて他の言語版を追加することが可能である。

図 1 は抽出したエントリ構造の一部の例である。

4.3 Wikipedia データの探索

作成したシステムでは、カテゴリを起点として言語間の比較を行う。システムはインデックスとして記録された Wikipedia のエントリ構造を探索し、次のデータを収集する。

- カテゴリに属する記事
- カテゴリの下位カテゴリ
- 探索を打ち切ったカテゴリ
- 下位カテゴリの名前の類似度によって算出したスコア

探索アルゴリズム

システムは次のアルゴリズムで、データを収集する。

1. 指定されたカテゴリのデータを取得する。
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、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. 追加したカテゴリのスコアを算出し総スコアに足す。
6. 下位カテゴリがなくなれば探索を終了する。

文字列の類似度による探索の打ち切り

探索アルゴリズム 4 においてカテゴリ名の類似度によって取得したカテゴリを検索結果に含めるかどうか判定している。Wikipedia ではユーザが自由な文字列によるカテゴリを 1 つのエントリに対して複数付与できる。ユーザによってカテゴリ付けの切り口や方針に違いがあり、下位カテゴリが必ずしも上位カテゴリの狭義的な意味を持っているとは限らない。そのため無条件に下位カテゴリを再帰的に探索すると、探索を開始したカテゴリとは関連性が希薄なエントリが大量に取得される場合がある。例えば“イスラム教”の下位カテゴリを追っていくと、イスラム教-イスラム世界史-スペインの歴史-バルセロナオリンピック-バルセロナオリンピックにおける各競技-バルセロナオリンピックにおけるレスリング競技、という経路が存在する。あるトピックに関する言語版ごとの統計を取るために探索を行っているので、このようなエントリは探索結果に含めるべきではない。また探索のための時間も余計にかかっている。そこで関連性の低いエントリを排除するため、関連性の高いエントリは文字列の類似度も高いとの想定のもと、探索の打ち切りを行っている。

カテゴリ名の類似度は次式で算出する。

$$\frac{|s(b(T_1), b(T_2))|}{\sqrt{|b(T_1)| \cdot |b(T_2)|}}$$

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。閾値を設定し、上式によって得た値が閾値以下の場合に探索を打ち切る。今回は閾値を 0.25 とした。

図 1 において、カテゴリ c1 を起点に探索を開始した場合、下位エントリはカテゴリ c2, c3, c4、記事 a1, a2 となる。a1, a2 を探索結果に含める。c2, c3, c4 に対して c1 との類似度を算出し、閾値以上であれば探索結果に含めさらに下位エントリを探索する。このようにエントリ構造を再帰的に探索し、エントリを収集する。図 1

表 2: 使用する言語版の Wikipedia の統計

言語	記事数	カテゴリ数	カテゴリ数/記事数
ドイツ語	1024188	71629	0.07
フランス語	854991	122832	0.14
日本語	628266	64628	0.10
中国語	277244	56102	0.20

において, c_4 と c_6 の探索が打ち切られた場合の探索結果を図示したものが図 2 である. 探索結果はカテゴリ c_1, c_2, c_3, c_5 , 記事 $a_1, a_2, a_3, a_4, a_8, a_9$ となる. c_4 と c_6 は探索結果には含まれないが, ユーザに探索を打ち切ったカテゴリを提示するために記録しておく.

スコアの算出

カテゴリの大きさの指標としてスコアを算出する. 探索によって得たカテゴリに対して次式で算出した値の総和をスコアとする.

$$Score_c \cdot p \cdot s(b(T_r), b(T_s)) + q$$

式中の $Score_c$ はこの式による上位カテゴリのスコア, T_r は探索の起点のカテゴリ, T_s は探索によって得た下位カテゴリである. 探索の起点となるカテゴリのスコアは 1.0 で開始する. p は重みづけのための係数で今回は 0.6 に設定した. q はスコアに足し合わせる定数で今回は 0.2 に設定した. この式によって上位カテゴリのスコアを下位カテゴリに伝播させながら全体のスコアを算出する.

4.4 Web インターフェース

作成した探索システムに対して問い合わせと結果の表示を行う Web インターフェースを作成した. 図 3 は Web インターフェースのスクリーンショットである. 探索の起点となる言語とカテゴリ名を与えて探索を開始する. 探索システムは指定された言語版に記述されている言語間リンクから他の言語版のカテゴリ名を取得し, 同様に探索を行う. 画面左の太字のカテゴリ名をクリックすると画面右に探索結果の詳細が表示される.

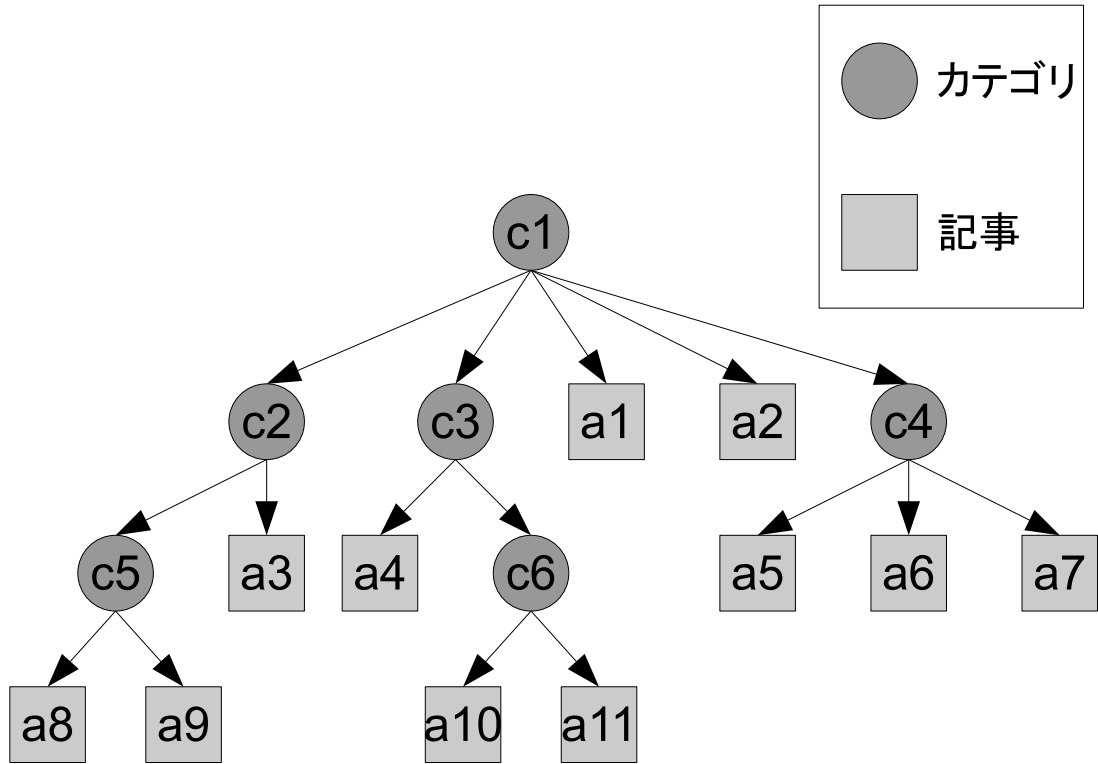


図 1: Wikipedia から抽出したエントリ構造の例

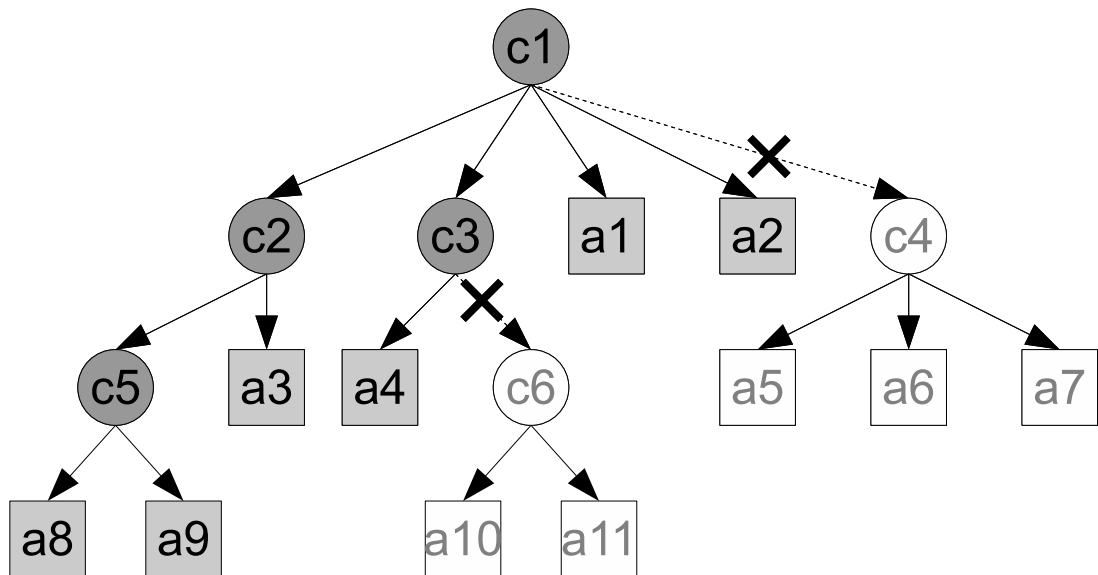


図 2: 図 1 のエントリ構造の探索結果の例



図 3: 比較システムの Web インターフェース

5 評価

作成したシステムとインターフェースを使用して取得した探索結果の実例を提示する．いずれも探索の起点となる言語は日本語版である．

言語版によって差異が小さいカテゴリ

表 3 は“データ構造”の比較結果である．言語版によって記事数にほとんど差がない．言語や文化によって Wikipedia ユーザの興味が偏らないトピックスであるといえる．ただし，フランス語版と中国語版にはサブカテゴリが存在し，スコアに反映されている．日本語版とドイツ語版にはサブカテゴリが存在しないためスコアが 0 となっている．表 2 に示した通り，言語版によって存在するカテゴリ数に違いがある．カテゴリ数が多い言語版ほど探索結果にカテゴリが含まれやすくなるので，スコアが高くなる傾向がある．またサブカテゴリが多い版ほどそのトピックスについて整備されているという想定のもとスコアを算出しているが，カテゴリが精査されている結果サブカテゴリが少ないということも考えられる．

言語版によって差異が大きいカテゴリ

表 4 は“第一次世界大戦”の比較結果である．ドイツ語版とフランス語版はエントリ数が多いが，日本語版と中国語版は比較的エントリ数が少ない．歴史的な経緯が Wikipedia ユーザの興味に反映されている例といえる．

表 5 は“第二次世界大戦”の比較結果である．“第一次世界大戦”の場合と似た傾向の結果が出ているが，日本語版では“太平洋戦争”の探索が打ち切れている．タイトルの文字列の類似度が低いためである．このように探索に含めるべきサブカテゴリであっても文字列の類似度が低いために適切な結果が得られない場合がある．

6 おわりに

本研究では Wikipedia の多言語展開に存在する言語版ごとの内容の差異に着目し，各版を特徴付けるデータの抽出と比較をするシステムを作成した．またその Web インターフェースを作成し，いくつかの比較結果の実例を提示した．

今後の展開としては 5 章で述べた探索システムの問題点を解決する．特に不適切な探索の打ち切りに対しては，ユーザが探索を続けるカテゴリを指定できるようにする，あるいは他の言語版で探索が続いているカテゴリは結果に含めるといった処理が考えられる．また比較に使用できる他のデータを獲得することも考えている．例えば探索したサブカテゴリや記事がどの程度の言語間リンクを持っているかを提示することで，そのトピックスがある言語や文化に特有のものなのか，あ

るいは国際的な興味を持たれているものなのかの指標にする．

参考文献

- [1] Wikimedia Foundation. List_of_Wikipedias. http://meta.wikimedia.org/wiki/List_of_Wikipedias.
- [2] Soren Auer, Chris Bizer, Jens Lehmann, Georgi Kobilarov, Richard Cyganiak, and Zachary Ives. DBpedia: A nucleus for a web of open data. *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference*, pp. 715–728, 2007.
- [3] Maike Erdmann, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. Extraction of bilingual terminology from a multilingual web-based encyclopedia. *Journal of Information Processing*, Vol. 16, July, pp. 68–79, July 2008.
- [4] 森 竜 也. Wik-IE. <http://sourceforge.jp/projects/wik-ie>.
- [5] Apache Software Foundation. Apache Hadoop. <http://hadoop.apache.org>.
- [6] Apache Software Foundation. Apache Hadoop. <http://lucene.apache.org>.

表 3: “データ構造” の比較結果

言語	記事数	カテゴリ数	スコア
ドイツ語	63	1	0.0
フランス語	61	3	1.29
日本語	64	1	0.0
中国語	58	2	0.44

表 4: “第一次世界大戦” の比較結果

言語	記事数	カテゴリ数	スコア
ドイツ語	2877	52	33.29
フランス語	1402	28	17.86
日本語	343	12	6.88
中国語	67	3	1.27

表 5: “第二次世界大戦” の比較結果

言語	記事数	カテゴリ数	スコア
ドイツ語	2927	40	23.79
フランス語	4237	97	61.30
日本語	878	11	6.23
中国語	376	23	14.96