日本語と英語 Wikipedia のカテゴリー構造の整合性について

新谷　誠　　綱川隆司　　梶　博行
静岡大学大学院 情報学研究科
{araya, tuna, kaji}@inf.shizuoka.ac.jp

1　はじめに

Wikipedia はウェブ上における有用な情報資源の一つとなりつつあり、多くの言語に対応している。Wikipedia は不特定多数のユーザによって編集されるため、同一の事柄に対しても言語が異なれば内容も異なり、その間の整合性が必ずしもとれていない場合がある。Wikipedia には記事が属するカテゴリーというリンク構造があり、これは一種のタクソノミー（語彙分類体系）として扱える。しかし、タクソノミーを多言語化する際には言語間のカテゴリー構造の整合性をとる必要がある。Melo and Weikum [1] ではマルコフ連鎖を基にした順位付けの方法により多言語のカテゴリーからなるリンク構造から一つのタクソノミーへ統合する方法を提示している。Garcia 達 [2] では、言語に依存しない 20 個の Preprocessing, Syntactic, Structural, Article 素性を利用してヨーロッパ言語 (英語、スペイン語、ドイツ語) と非ラテン文字言語 (アラビア語、ロシア語) のそれぞれに対してカテゴリーの上位下位関係を抽出する実験を行い、適合率と再現率が平均 70% を越える精度となることを示している。

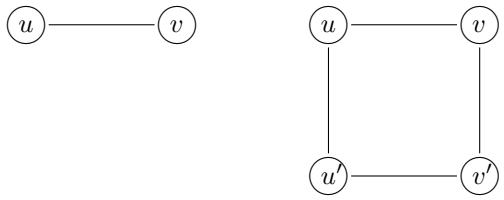
本論文では、言語毎のカテゴリーの適切な粒度を保ちつつ言語間の整合性を得るために必要な提案（カテゴリーの新規作成等）を行う Wikipedia 編集の実用的な支援システム開発のために、日本語 Wikipedia と英語 Wikipedia を対象に基本的なカテゴリーの部分構造の調査を行った。整合性を持った構造は分類とみることができ、編集後には Wikipedia から多言語タクソノミーが得られる。

今回の調査に用いたデータは、2013 年 3 月 28 日の日本語版 Wikipedia(カテゴリー数は 110503 カテゴリー間のリンク数は 225708)、2013 年 4 月 3 日の英語版 Wikipedia(カテゴリー数は 1000736 カテゴリー間のリンク数は 1986710) であり、日本語カテゴリーと英語カテゴリー間のリンク数は 55437 であり略称や別名を除いてある。

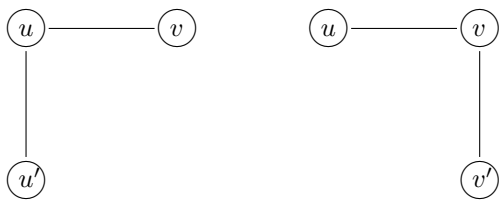
2　カテゴリーのグラフ構造

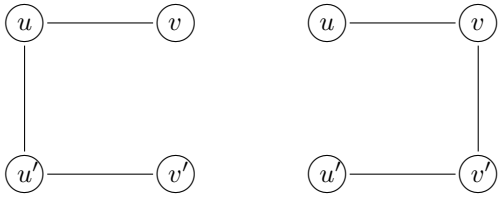
Wikipedia のカテゴリーを点、上位カテゴリーから下位カテゴリーへの関係を弧とした有向グラフを考える。日本語版、英語版の Wikipedia から得られる有向グラフの点の集合をそれぞれ $V(\mathcal{J}), V(\mathcal{E})$、弧の集合をそれぞれ $E(\mathcal{J}), E(\mathcal{E})$ とする。日本語カテゴリーと英語カテゴリー間のリンクの集合を $E(\mathcal{JE})$ とする。また、グラフ $(V(\mathcal{J}), E(\mathcal{J}))$ (または $(V(\mathcal{E}), E(\mathcal{E}))$) において点 $x \in V(\mathcal{J})$ (または $x \in V(\mathcal{E})$) 近傍を $A(x) = \{y \in V(\mathcal{J}) \mid (x, y) \in E(\mathcal{J})\}$ (または $A(x) = \{y \in V(\mathcal{E}) \mid (x, y) \in E(\mathcal{E})\}$) とする。グラフ $(V(\mathcal{J}) \cup V(\mathcal{E}), E(\mathcal{JE}))$ において点 $x \in V(\mathcal{J})$ (または $x \in V(\mathcal{E})$) の近傍を $L(x) = \{y \in V(\mathcal{E}) \mid (x, y) \in E(\mathcal{JE})\}$ (または $A(x) = \{y \in V(\mathcal{J}) \mid (y, x) \in E(\mathcal{JE})\}$) とする。

日本語と英語カテゴリー間リンク $(u, v) \in E(\mathcal{JE})$ を与えたとき、その深さ 1 以下の下位カテゴリーの構造の整合性がある正しい形は次の 2 つの形であり、



整合性がない正しくない形は次の 4 つの形である。





$|\{(u,v) \in E(\mathcal{JE}) \mid A(u) = A(v) = \emptyset\}| = 17035$ である。また、上の 1 つ目以外に次の k_6, k_7 を加え、次のように各個数を定義する。日本語と英語のカテゴリー数の量の差から $k_6 > k_7$ となることが多いと予想できる。

1. $k_1(u,v) = |\{(u',v') \in A(u) \times A(v) \mid (u',v') \in E(\mathcal{JE})\}|$
2. $k_2(u,v) = |\{u' \in A(u) \mid L(u') = \emptyset\}|$
3. $k_3(u,v) = |\{v' \in A(v) \mid L(v') = \emptyset\}|$
4. $k_4(u,v) = |\{u' \in A(u) \mid L(u') \neq \emptyset, A(u) \cap L(u') = \emptyset\}|$
5. $k_5(u,v) = |\{v' \in A(v) \mid L(v') \neq \emptyset, A(v) \cap L(v') = \emptyset\}|$
6. $k_6(u,v) = |\{u' \in A(u) \mid (u',v') \in E(\mathcal{JE}), (v,v''), (v'',v') \in E(\mathcal{E})\ (\exists v'' \in V(\mathcal{E}))\}|$
7. $k_7(u,v) = |\{v' \in A(v) \mid (u',v') \in E(\mathcal{JE}), (u,u''), (u'',u') \in E(\mathcal{J})\ (\exists u'' \in V(\mathcal{J}))\}|$

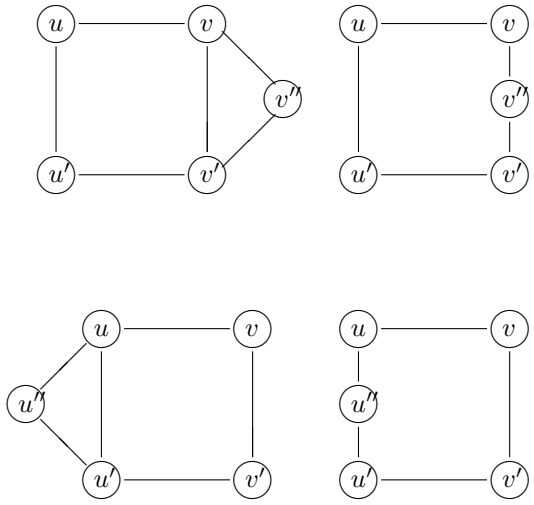
$k_i(u,v)$ を k_i と書くことにする。$\{(u,v) \in E(\mathcal{JE}) \mid A(u) = A(v) = \emptyset\}$ 以外の k_i の基本統計量は表 1 の通りである。

表 1: k_i の基本統計量

	平均	標準偏差	最小値	最大値
k_1	1.30	5.49	0	208
k_2	1.29	7.30	0	841
k_3	8.87	26.00	0	5213
k_4	1.06	4.43	0	354
k_5	0.61	13.52	0	2587
k_6	0.53	2.74	0	216
k_7	0.18	1.85	0	149

表より、「k_2 の最大値 $< k_3$ の最大値」と「k_4 の最大値 $< k_5$ の最大値」であり、大小関係が同じである。また、k_3 の平均値、分散ともに値が大きいことがわかる。

k_6, k_7 が表しているのは、それぞれ次のような構造である。

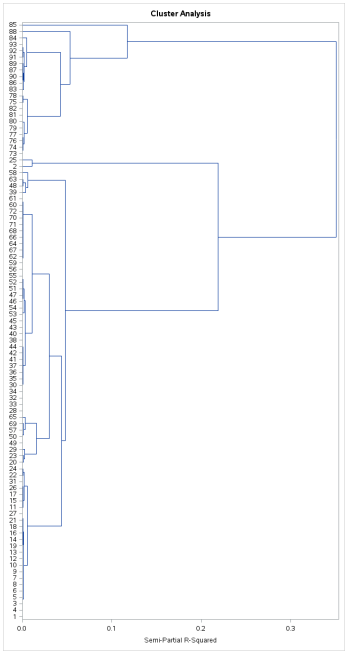


日本語と英語の量の差から $k_6 > k_7$ のとなる傾向の多いことが、k_i の基本統計量からも読み取れる。

3 クラスター分析による考察

日本語と英語の量の差に注目をして、$k_6 > 0$ または $k_7 > 0$ を満たす日本語と英語間リンクについて調査を行うことにする。量の差から $k_6 > k_7$ を満たす日本語と英語間リンクが多いと予想ができ、$k_6 < k_7$ を満たすものは 1346 個あり、$k_6 > k_7$ を満たすものは 5845 個ある。$k_6 < k_7$ を満たすカテゴリーの例としては (国別に分類したカテゴリー, Categories by country) $\in E(\mathcal{JE})$ $((k_i) = (24, 6, 954, 8, 253, 17, 138))$ や (各国の人物 (職業別), People by occupation and nationality) $\in E(\mathcal{JE})$ $((k_i) = (23, 9, 139, 9, 30, 6, 24))$ があげられる。$k_6 > k_7$ を満たすカテゴリーの例としては (自動車の車種, Vehicles by brand) $\in E(\mathcal{JE})$ $((k_i) = (69, 44, 144, 15, 3, 64, 3))$ や (各国の文化, Culture by nationality) $((k_i) = (161, 4, 108, 20, 9, 57, 8))$ があげられる。

データセット $K = \{(k_1, k_2, k_3, k_4, k_5, k_6, k_7) \mid (u,v) \in E(\mathcal{JE}), k_6 + k_7 > 0\}$ に対して統計ソフトウエアの SAS [3] を用いて、クラスター分析 (階層的、ユークリッド距離、Ward 法) を行った。次のデンドログラムが出力される。



使用した命令は表 2 の通りである。

表 2: SAS の命令

```
proc cluster data=K method=ward
outtree=tree;
```

デンドログラムより、クラスター数を 4 としてデータを考察してみる。上から、第 $1, 2, 3, 4$ クラスということにする。データを実際にみると上から順番に k_1 の桁数の大きい順番に並んでいる。k_1 は整合性のある正しい形の個数なので、整合性に関するクラスターが出力されたと評価できる。

第 2 クラスターの上位のカテゴリーには (各国の政治家, Politicians by nationality) $((k_i) = (177, 3, 74, 5, 10, 175, 2))$ と (各国の文化, Culture by nationality) $((k_i) = (161, 4, 108, 20, 9, 57, 8))$ があり k_1 の値が大きいので整合性が高く、対応する英語版のカテゴリーが充実していることがわかる。

第 3 クラスターは、(国別に分類したカテゴリー, Categories by country) $((k_i) = (24, 6, 954, 8, 253, 17, 138))$ と (国関連のテンプレート, Country templates) $((k_i) = (1, 1, 66, 1, 150, 0, 149))$ からなり、それぞれ k_2, k_6 に比べて k_3, k_7 が大きいという特徴を持っている。

第 4 クラスターの中で特徴的なサブクラスターがある。第 2 クラスターの上位と同じ傾向を持っている (各国の歌手, Singers by nationality) $((k_i) = (75, 14, 90, 5, 1, 64, 3))$、(各国の映画, Cinema by country) $((k_i) = (59, 4, 99, 7, 0, 55, 0))$、(アメリカ合衆国の州, States of the United States) $((k_i) = (51, 1, 6, 2, 0, 59, 0))$ といったカテゴリーの組がある。具体的な国名からなるカテゴリーの他に、英語版では地域名からなるカテゴリーも存在しているので対応する英語版のカテゴリーが充実している。

4 整合性向上のための提案

カテゴリー構造の整合性を向上させるために、以下の提案が考えられる。

Wikipedia のカテゴリー名には「各国のサッカー」のように「各国の」ではじまるカテゴリーが多くあり、下位カテゴリーとして「日本のサッカー」のように「各国」を国名で置き換えたカテゴリーとなっている。このような場合には、すべての国名からなるカテゴリーの作成を提案することができる。

$k_6(u, v) > 0$ となるカテゴリーの組 (u, v) に対して、英語カテゴリー v'' $((v, v''), (v'', v') \in E(\mathcal{E}))$ に対応する日本語カテゴリー u'''' $((u, u''), (u'', u') \in E(\mathcal{J})$ の新規作成を提案できる。また、$k > 7(u, v)$ の時にも同様の提案、あるいは対応するカテゴリの削除を提案できる。

5 おわりに

日本語 Wikipedia と英語 Wikipeida のカテゴリ数には約 10 倍の差があり、Wikipedia のカテゴリーに対して、セクション 2 で定義したカテゴリーの部分構造をクラスター分析することで英語版 Wikipedia の充実度を確かめることができた。構造の整合性の考察により、日本語版の量が今後増えるにつれて英語版のカテゴリー構造にならい、日本語版で新カテゴリーの作成を提案することができる。

謝辞 グラフ構造の分析プログラムの実行や SAS の利用のために、本研究の一部は京都大学学術情報メディアセンターのスーパーコンピュータを利用して実施しました。

参考文献

[1] G. de Melo and G. Weikum(2010), "MENTA: inducing multilingual taxonomies from Wikipedia", *Proceedings of the 19th ACM international conference on information and knowledge management*, pp. 1099–1108.

[2] R.D. Garcia, S. Schmidt, C. Rensing and R. Steinmetz(2012), "Automatic taxonomy extraction in different languages using Wikipedia and minimal language-specific information", *Proceedings of the 13th international conference on intelligent text processing and computational linguistics*, LNCS 7181, pp. 42–53.

[3] SAS/STAT(R) 9.3 User's Guide, `https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm`