Wikipediaにおける記事に適したInfoboxテンプレート種類の推定手法

原　圭介
豊橋技術科学大学
知識情報工学課程
hara@la.cs.tut.ac.jp

小林　暁雄
豊橋技術科学大学大学院
情報・知能工学専攻
kobayashi@la.cs.tut.ac.jp

増山　繁
豊橋技術科学大学大学院
情報・知能工学専攻
masuyama@tut.jp

1. はじめに

近年，テキストマイニングを行なう際の情報資源としてWikipediaが注目を集めている.その理由として，語彙網羅性，即時更新性が優れている点と，半構造情報資源である点が挙げられる．これにより，テキストマイニングを行なう際に，フリーテキストと比べて扱いやすい情報資源となっている.そのため，Wikipediaを用いた様々な研究が盛んに行われている.

その中でもInfoboxが有する「記事–属性–属性値」のトリプル(三つ組)は，テキストマイニングを行なう上で非常に有用な情報である.Wu et al.[1]は，InfoboxとWordNet Synset階層を利用し，高精度なオントロジーを構築する手法を提案している．玉川ら[2]は，Infobox から抽出したトリプルを用いて属性間の上位下位関係，属性関係として対称関係，推移関係，関数関係，逆関数関係の推定を行なう手法を提案している.真嘉比ら[3]はInfoboxから抽出したトリプルを用いた質問応答手法を提案している.

しかしながら，Wikipedia には作成したばかりで文章数が少ない記事や，書きかけの記事が多く，Infoboxが付随していない記事は非常に多い．したがって，そのような記事に Infobox を付随させることができれば，Wikipedia をより有用な情報資源にすることが可能であると考えられる.

記事に Infobox を付随させる際には，多数存在するInfobox テンプレートの種類の中から記事に適した種類を推定する必要がある．そこで我々は，記事の概要文とWikipediaが持つ記事の分類体系であるカテゴリに着目し，これを用いて多数存在する Infobox テンプレートの種類の中から，記事に適した種類を推定する手法を提案する.

2. 関連研究

Wu et al.[4]は，すべての一覧記事からInfoboxの種類名がタイトルに含まれている一覧記事を取得し，取得した一覧記事中に掲載されている記事の中から，Infobox の種類名が含まれているカテゴリタグが付されている記事に対して種類を推定するという方法を用いている.この手法は適合率が98.5%，再現率が68.8%と高い性能で種類を推定することが可能となっている.しかしながら，一覧記事に掲載されていない記事やカテゴリ付けが不十分な記事に対しては，この手法を利用することができず，記事の網羅性に欠ける.

我々はこれまで，記事冒頭の定義文に着目し，これを用いて記事にふさわしいと思われる Infobox テンプレートの種類の候補を上位3種類まで推定する手法を提案してきた(原ら[5])．今回はこの手法を，上位1種類のみ用いるという条件のもとで本手法に取り入れ，対象の記事に対し Infobox を付随させるべきか否かを判断し，付随させる場合は，その記事に最も適するInfoboxの種類を1種類推定する.

3. Infoboxについて

Infobox テンプレートは記事の様々な属性についての要約情報を表示することを目的としたテンプレートである(図 1 参照)．これによって閲覧者は記事をすべて読まなくとも，Infobox内の要約情報を見るだけで，記事に関しての重要な情報を容易に得ることができる.

図 1.　Infobox の例

Wikipedia 内に存在する記事には，企業や人物，建造物など，様々な分野の記事が存在するため，分野によって必要な属性は変化する.そのため，これらの様々

な分野の記事に対応できるように，Infobox テンプレートは現在 1,200 種類ほどユーザーによって構築されている．

4. カテゴリについて

Wikipedia では，さまざまな分野の記事に対し，カテゴリを用いて分野ごとに分類し，記事の体系化を行なっている．これにより，利用者は閲覧中の記事に対する関連記事を容易に発見することができる．記事に付随しているカテゴリの例を図 2 に示す．

図 2.　Wikipedia 記事におけるカテゴリ記述の例

5. 提案手法

前処理として，種類候補の推定手法を 2 つ用いて，それぞれ記事に適した Infobox の種類候補を取得する．1 つ目は，我々がこれまで提案してきた，記事冒頭の定義文に含まれる名詞を用いた手法によって推定した Infobox の種類候補(以下，インデックス候補と呼ぶ)である．2 つ目は，Infobox が付随している記事から作成した文書ベクトル(以下，種類ベクトルと呼ぶ)と，Infobox が付随していない記事から作成した文書ベクトルとの類似度を計算することによって推定した Infobox の種類候補(以下，類似度候補と呼ぶ)である．

次に，記事が属するカテゴリのカテゴリ名と，そのカテゴリに属している他の記事を用いて，2 つの種類候補から，記事にもっとも適する Infobox の種類を 1 種類推定する．本手法の概要図を図 3 に示す．

5.1　種類ベクトルを用いた類似度候補の推定手法

種類ベクトルを用いた類似度候補の推定手法の具体的な手順を以下に示す．

Step 1. Infobox が付随している記事を Infobox の種類ごとにまとめた，記事集合を作成する．
Step 2. 記事集合中の記事すべてから，概要文とカテゴリ名を抽出し，形態素解析を行なう．
Step 3. 記事集合を 1 つの文書とみなし，各記事集合の形態素から種類ベクトルを作成する．
Step 4. Infobox が付随していない記事の概要文とカテゴリ名を形態素解析し，文書ベクトルを作成する．
Step 5. Infobox が付随していない記事の文書ベクトルと各種類ベクトル間のコサイン類似度を求め，コサイン類似度の高い上位 32 個に対応する Infobox の種類を，類似度候補とする．

ベクトルを作成するにあたって，形態素の重みに TF-IDF，ベクトルの正規化にコサイン正規化を用いた．

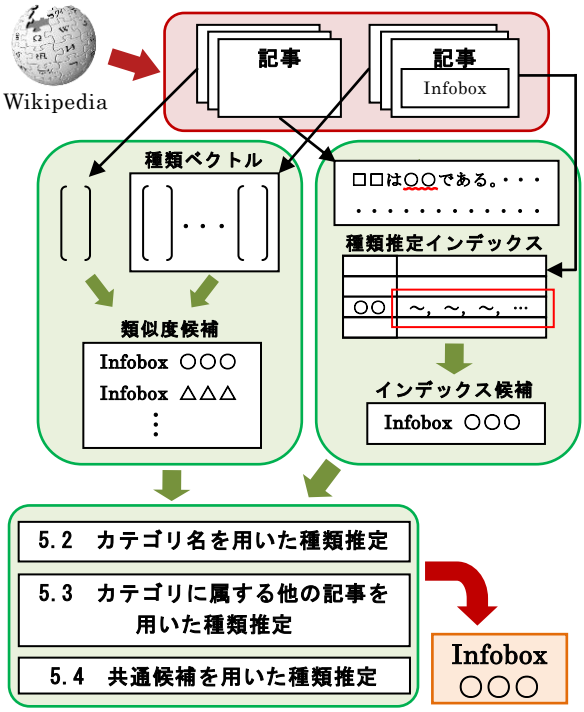

図 3.　提案手法の概要図

5.2　カテゴリ名を用いた種類推定

Infobox の種類名とカテゴリ名は類似している場合が多い．例えば，“大学”という種類名に対して，“日本の大学”，“イングランドの大学”など，類似しているカテゴリ名が多数存在する．この関係を用いて，藤原ら[6]が行なってきたカテゴリ名における記述パターンの分析を参考に，カテゴリ名とのパターンマッチングによる種類推定を行う．具体的な手順を以下に示す．

Step 1.　Infobox が付随していない記事のカテゴリ名をすべて抽出する．
Step 2.　インデックス候補と，抽出したカテゴリ名すべてを用いてパターンマッチングを行なう．
Step 2.a. インデックス候補に対し，カテゴリ名がマッチした場合，インデックス候補を記事に最も適した種類として推定する．
Step 2.b. いずれのカテゴリ名にもマッチしなかった場合，Step 3 に進む．
Step 3.　類似度候補と，抽出したカテゴリ名すべてを用いてパターンマッチングを行なう．

Step 3.a. 類似度候補のいずれかに対し，カテゴリ名がマッチした場合，マッチした Infobox の種類を記事に最も適した種類として推定する．

Step 3.b. いずれのカテゴリ名にもマッチしなかった場合，5.3 章で述べる手法に移る．

5.3 カテゴリに属する他の記事を用いた種類推定

対象の記事が属しているカテゴリにおいて，そのカテゴリに属している他の記事をすべて参照すると，対象の記事に適した種類の Infobox を有している記事が含まれている場合が多い．例えば，“夏目漱石”の記事は“日本の小説家”というカテゴリに属しているが，このカテゴリに属している他の記事を参照すると，“Infobox 作家”という種類の Infobox が用いられている記事が多数存在する．この特徴を用いて，記事に適した Infobox の種類推定を行う．また，カテゴリに属している他の記事のうち，候補である Infobox の種類が用いられている記事が存在する割合が 8%以上であるという条件下で，正解率が最も高いことを予備的検討で確認したため，8%を閾値とする．具体的な手順を以下に示す．

Step 1. Infobox が付随していない記事が属しているカテゴリ全てに対し，そのカテゴリに属している他の記事すべてを取得する．

Step 1.a. 取得したすべて記事のうち，インデックス候補である種類の Infobox が用いられている記事が存在する割合が 8%以上である場合，その Infobox の種類を記事に最も適した種類として推定する．

Step 1.b. 類似度候補の最上位である種類の Infobox が用いられている記事が存在する割合が 8%以上である場合，その Infobox の種類を記事に最も適した種類として推定する．

Step 1.c. Step 1.a と Step 1.b のどちらでもない場合，5.4 章で述べる手法に移る．

5.4 共通候補を用いた種類推定

インデックス候補と類似度候補との間に同じ Infobox の種類が存在している場合，それが記事に対して最も適する Infobox の種類であると考えられる．そこで，その場合は，両方の候補に存在する Infobox の種類を推定する種類とした．

また，共通候補を用いた種類推定手法においても適した種類が推定できなかった場合，Infobox を付随させるべきではない，個別の属性を持たない抽象的な概念の記事(“プライバシー”，“思考”，“認識”など)であると判断した．

6. 評価実験

本手法を評価するために実験を行なった．実験には，日本語版 Wikipedia の記事データとして 2012 年 6 月 3 日のダンプデータ[1]を用いた．また，形態素解析を行なう際の形態素解析エンジンとして MeCab[2]を用いた．

実験に用いる記事としては，Wikipedia の記事からランダムに 1,000 ページを取得し，その中から，一覧記事と曖昧さ回避ページを除いた 945 ページを用いた．

7. 実験結果と考察

6 章で述べた条件の下で評価実験を行なった結果，正解率が 81.7%となった．各種類推定法における推定誤り数を表 1 に示す．

表 1. 各種類推定法における推定誤り数

種類推定法	ページ数
カテゴリ名を用いた種類推定	30
カテゴリに属する他の記事を用いた種類推定	18
共通候補を用いた種類推定	20

どの種類推定法も適用されなかった場合，本手法では，対象の記事が抽象的な概念の記事であるとして，Infobox を付随させるべきではないと判断している．そのような記事は 290 ページ存在した．しかしながら，この判断によって Infobox を付随させるべき記事に Infobox の種類が推定されなかったパターンは 103 ページ存在した．

7.1 カテゴリ名を用いた種類推定の考察

推定誤りの例として“笠島 (丸亀市)”という記事に“建築物”が推定されてしまった例がある．この記事は地区名に関する記事であるが，記事本文に建築物に関する記述があるため，“香川県の建築物”というカテゴリが付随している．これにより推定誤りが起こってしまった．

これは，上位下位関係だけでなく，トピックや関連を表すカテゴリを記事に付随させることを許可している Wikipedia の特徴が原因であると考えられる(白川ら[7])．このため，カテゴリ名を種類推定に用いる際には，どのカテゴリが記事項目と上位下位関係にあるかを判断する必要がある．

また，Infobox を付随させるべきでない記事に対して，種類を推定してしまった例も存在した．その例として“準貴族”という記事に“貴族・コサック”が推定されてしまった例がある．この記事は身分に関する記事であり，“貴族”というカテゴリが付随している．

[1] Wikipedia ダンプデータ(http://dumps.wikimedia.org/jawiki/)

[2] MeCab(http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html)

これにより推定誤りが起こってしまった．

このような推定誤りを解決するためには，カテゴリ名を用いた種類推定を行なう前に，抽象的な概念の記事であるか否かを判断する処理が必要であると考えられる．

7.2 カテゴリに属する他の記事を用いた種類推定の考察

推定誤りの例として“第 88 回全国高等学校サッカー選手権大会”という記事に“国際サッカー大会情報ボックス”が推定されてしまった例がある．この記事は国際サッカー大会に関する記事ではないので,“スポーツ大会シリーズ”という種類が正しいと考えられる．これは，この記事が属している“2010 年のサッカー”というカテゴリの下に国際サッカーの大会に関する記事が多く存在している事が原因であると考えられる．

7.3 共通候補を用いた種類推定の考察

推定誤りの例として“ICC プロファイル”という記事に“Infobox Software”が推定されてしまった例がある．この記事においては，インデックス候補と類似度候補の両方に“Infobox Software”という Infobox の種類が存在したため，この種類が推定されてしまった．しかしながら，この記事はカラーマネジメントにおけるデータの仕様について記載された，抽象的な概念についての記事であり，推定誤りである．

このような誤りの原因として，抽象的な概念の記事の定義文の文末と，具体的な記事の定義文の文末に，同じ名詞が含まれている場合が多いことが挙げられる．この例においては,“ICC プロファイル”の記事の定義文の文末に“データ”という名詞が存在するが,“tz database”という記事の定義文の文末にも同じ名詞が存在する．また,“tz database”という記事には,“Infobox Software”という種類の Infobox が用いられている．これにより，インデックス候補に“Infobox Software”が挙げられてしまい，結果として誤った種類推定を行なってしまっている．

これは，インデックス候補を推定する際に，記事冒頭の定義文に含まれる名詞のみを用いて推定していることに起因すると考えられる．定義文に含まれる名詞のみでは，対象の記事が抽象的な概念の記事であるか否かを判断することは不可能である．このような推定誤りに関しても，共通候補を用いた種類推定を行なう前に，抽象的な概念の記事であるか否かを判断する処理が必要であると考えられる．

8. 終わりに

本稿では，Wikipedia における，Infobox がまだ付随していない記事に対して Infobox の種類を推定する手法の提案，および，その評価を行なった．対象の記事に対し Infobox を付随させるべきか否かを判断しなければならないという条件と，推定する種類を 1 種類まで絞らなくてはならないという条件の 2 つの条件下で，種類推定における正解率が 81.7%と高い正解率を得ることができた．したがって，今回提案した本手法は，Infobox の種類を推定するにあたって有効な手法であると考えられる．

しかしながら，考察から更なる改善が必要であることが判明した．その中でも特に重要な改善すべき点は，どの種類推定法も適用されなかった記事が非常に多く存在した点である．本手法では，カテゴリ名を用いた種類推定，カテゴリに属する他の記事を用いた種類推定，共通候補を用いた種類推定の 3 つを用いて，種類推定を試みているが，どの種類推定法も適用されなかった記事に対して正しい種類推定を行なうために，種類推定法を新たに考案する必要がある．その他の改善すべき点としては，対象の記事に対する Infobox を付随させるべきか否かの判断の精度である．各種類推定法において，Infobox を付随させるべきか否かの判断誤りがいくつか存在した．したがって，Infobox を付随させるべきか否かを判断する処理を新たに考案し，各種類推定法を施す前に，その処理を施す必要がある．

今後はこのような改善すべき点を解決していき，正解率の向上を目指すとともに，Infobox 内の属性の値を補完する手法についても検討していく予定である．

参考文献

[1] F. Wu and D. S. Weld：Automatically Refining the Wikipedia Infobox Ontology, WWW08, pp.635-644 (2008).

[2] 玉川 奨，関本 有佳，森田 武史，山口 高平，日本語 Wikipedia からプロパティを備えたオントロジーの構築，第 25 回 人工知能学会全国大会論文集，2J3-NFC2-5(2011).

[3] 真嘉比 愛，Stijn De Saeger，鳥澤 健太郎，呉 鍾勲，山本和英，Wikipedia Template から抽出した意味的関係インスタンスによる質問応答手法，言語処理学会第 18 回年次大会発表論文集，pp. 703-706(2012).

[4] F. Wu and D. S. Weld ： Autonomously Semantifying Wikipedia, CIKM07, pp.41-50 (2007).

[5] 原 圭介，小林 暁雄，増山 繁，Wikipedia 記事項目定義文の特徴を利用した記事に適した Infobox テンプレート種類の推定法，第 10 回情報学ワークショップ（WiNF2012），ISSN1884-3387(2012).

[6] 藤原 嵩大，吉岡 真治，Wikipedia の階層関係を分析するためのカテゴリパターンの提案，第 26 回人工知能学会全国大会論文集，2C1-NFC-2-4(2012).

[7] 白川 真澄，中山 浩太郎，原 隆浩，西尾 章治郎，ナイーブベイズによる文書分類のための Wikipedia カテゴリグラフ解析，第 26 回人工知能学会全国大会論文集，2C1-NFC-2-2(2012).