2I3-NFC4-4

Web上の情報を用いたWikipedia記事の信頼性評価に関する検討

An investigation into evaluating the credibility of Wikipedia contents by using information on the Web

伊藤 雅弘[*1] 中山 浩太郎[*2] 原 隆浩[*3] 西尾 章治郎[*4]
Masahiro Ito Kotaro Nakayama Takahiro Hara Shojiro Nishio

[*1*3*4]大阪大学大学院情報科学研究科
Dept. of Multimedia Eng., Graduate School of Information Science and Technology, Osaka University

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

Recently, Wikipedia, a huge scale Web-based encyclopedia, attracts great attention as an invaluable corpus for knowledge extraction because it has various impressive characteristics such as a huge number of articles, live updates, a dense link structure, brief anchor texts and URL identification for concepts. However, Wikipedia sometimes has an imbalance ofinformation volume for specific categories or incorrect/truthless information. Wikipdia makes it possible to construct vast scale of contents whereas it needs to improve information quality. In this paper, we investigate into a method for evaluating or improving the information quality of Wikipedia contents for knowledge extraction by using information on the Web.

1. はじめに

Wikipedia は,「群集の叡智」と呼ばれる形式の新しいソーシャルメディアであり，知識抽出のためのコーパスとして，その有用性が研究者の間でも非常に注目を集めている．一方で，Wikipedia の情報には内容に偏りがあったり，また虚偽の情報が記載されているケースも存在する．そのため,「誰でも編集可能である」という特性が大規模なコンテンツの構築を可能にしている一方で，情報の信頼性を如何に確保するかが大きな課題となっている．そこで本研究では，Wikipedia において信頼性を評価し向上するための手法を検討することによって，Wikipedia の質や精度の高い意味情報の抽出を目指す．従来研究において，編集履歴に基づいて記事の信頼度を算出する手法が提案されている．Adler ら [Adler 08] や金本ら [金本 08] は，Wikipedia の編集履歴から著者の信頼度を算出し，その著者の信頼度を基に記事の信頼度を求める手法を提案している．著者の信頼度は，その著者が書いた記述の追記，削除，置換を考慮した残存率に基づいて算出される．

本稿では，Wikipedia の信頼性に影響する因子の中で，特に概念構造の分野による偏りや欠落に着目した．例えば，日本語版 Wikipedia 場合，ポップカルチャーに関する項目は非常に充実しており，そのアクセス率は 80%にも上る一方，手薄な分野も存在する．これは，利用者の文化的背景の影響という意味では興味深い現象ではあるが，知識抽出や百科事典的情報源としては網羅性の確保が重要である．そこで，ある程度構造化された情報源である Wikipedia と，ノイズは多いが大量の情報を持つ Web 上の情報を融合することにより，Wikipedia の内容を補完し，より質の高い知識抽出を可能とする手法を検討する．具体的には，Web 検索エンジンを用いることによって，Wikipedia 中の欠落する概念を同定し，提示及び自動的に概念構造に追加する．

連絡先: 伊藤雅弘，大阪大学大学院情報科学研究科，〒 565-0871 大阪府吹田市山田丘 1-5 大阪大学大学院情報科学研究科 マルチメディア工学専攻 マルチメディアデータ工学講座（西尾研究室），06-6879-4513，06-6879-4514，ito.masahiro@ist.osaka-u.ac.jp

2. 提案手法

提案手法では，Web 上の情報を用いることによって，Wikipedia 中の欠落した概念（記事）の発見を試みる．前章で述べたように，Wikipedia では項目が非常に充実している分野がある一方，手薄な分野も存在する．本手法によって，Wikipedia 中の欠落した概念を発見することによって，Wikipedia コミュニティーによる網羅性の向上を促進することができる他，既存の Wikipedia から構築された連想シソーラス [Nakayama 07, Ito 08] の網羅性を向上させることが可能となる．本手法の主な処理の流れを以下に示す．

1. 任意の概念の新規関連概念を抽出するための情報源となる Web ページ集合を収集.
2. Web ページ集合から語集合を収集.
3. 語集合から新規概念集合を推定.

本章の以降では，任意の概念 α に対する新規関連概念の収集に関して，アルゴリズムのそれぞれの処理について詳述する．

2.1 Web ページ集合 P の収集

まず任意の概念 α に対して，連想可能な新規概念を発見するための情報源となる Web ページ集合 P を収集する．ここで，概念 α をクエリとした Web 検索を行った場合，膨大な数の Web ページがヒットするため，ランキング上位の Web ページを収集した場合でも，内容が発散し新規概念発見の精度が下がると考えられる．そこで，解析対象を絞り込むため，また既存の Wikipedia の概念構造に新規概念を追加するという目的から，概念 α の連想概念を取得し，その連想概念を用いて Web 検索のためのクエリを作成する．連想概念は，Wikipedia から構築された連想シソーラスを用い，連想概念の上位 k 個を用いる．概念 α に対する k 個の連想概念集合を A とすると，Web 検索クエリ集合 Q は以下のように定義される．

$$Q = \{\text{"}\alpha_1\text{"} \quad \text{"}A_{11}\text{"}, \cdots, \text{"}\alpha_i\text{"} \quad \text{"}A_{kj}\text{"}\}.$$

ここで，概念 α の同義語を $\alpha_1 \cdots \alpha_i$，概念 A_k の同義語を $A_{k1} \cdots A_{kj}$ とする．

このクエリ集合 Q を用いて，各クエリで Web 検索によって得られたページの上位 n 件を，Web ページ集合 P として収集する．

2.2 語集合 W の収集

収集した Web ページ集合 P から，新規概念の候補となる語集合 W を収集する．語の抽出の範囲としては，ページ中のすべての語を対象とするか，ページ中に出現するクエリに含まれる語の周辺語を対象とするかの 2 つの方法がある．今回は，ページ中のすべての語を対象とした．語の抽出は，有名な固有表現抽出のためのツールである Stanford Named Entity Recognizer[*1] を用いた．

2.3 新規概念集合 B の推定

収集した語集合 W から，新規概念集合 B を推定する．まず，語集合 W を Wikipedia に存在する概念で特定できるもの Wa と，それでないもの Wb に分類する．語が Wikipedia の概念として存在するかの判別処理は，各語と Wikipedia の記事タイトル，リダイレクトページのタイトル，アンカーテキストでのマッチング処理によって行う．語が 3 つのいずれかにマッチした場合，その語が表す概念が Wikipedia に存在するとみなす．

次に，Wikipedia の概念で特定できない語集合 Wb から，ノイズ語の除去を行う．抽出した語の中には，言語解析のミスによって明らかに不適切な語が含まれている場合があり，また連想概念の候補となり得ない関係の薄い，もしくは一般的な語も含まれている．それらの語を除去するために，まず $tf \cdot idf$ を用いて各語のページ集合 P 中での重要度を求め，また概念 α と各語の Web 上での共起性を求め，双方の値が高い語を新規概念とする．ここで，tf はページ集合 P 中の各語の出現頻度，df は各語の Web 検索でのヒット件数する．概念 α と各語の共起性は，それぞれの Web 検索でのヒット件数を f_x, f_y とし，2 語の AND 検索のヒット件数を f_{xy} として，以下の式によって求めた．

$$cooccurrence = \frac{f_{xy}}{min(f_x, f_y)}$$

3. 実験

本章では，提案手法の有効性を検証するために行った実験について述べる．本実験では，Wikipedia に存在する概念 “Nagasaki Prefecture”, “Japanese Garden” を対象に本手法を適用し，それぞれに関連する新規概念の収集を行う．手法中に用いる Web 検索エンジンとして，Yahoo! Search BOSS API[*2] を用いた．

3.1 実験結果と考察

実験によって推定された新規概念候補リストを表 1 に示す．表中の，“Nagasaki Prefectural Art Museum”，“Nagasaki Port”，“Dejima Museum of History” など，また “Sakuteiki”，“Adachi Museum of Art”，“Western Pureland” などは Wikipedia の概念として存在しないが，“Nagasaki Prefecture” や “Japanese Garden” と関係があると考えられるものであった．一方で，“Fukue Shima”，“Saikai City”，“Ginkakuji Temple”, “Nikka Yuko Garden” などは Wikipedia に存在する概念であると確認できた．これは，表記の違いでうまくマッチングできなかったことが原因である．また “Kyushu-Okinawa”, “Sasebo Hotels”, “Art.com” などは，概念として不適切であると考えられる．この中で特に “Art.com” は非常に高い共起性を示していた．

表 1: 新規概念候補リスト (トップ 10 件)

	Nagasaki Prefecture	Japanese Garden
1	Fukue Shima	Ginkakuji Temple
2	Nagasaki Prefectural Art Museum	Art.com
3	Kyushu-Okinawa	Sakuteiki
4	Sasebo Hotels	Adachi Museum of Art
5	Saikai City	Ryoanji Temple
6	Sasebo Japan	Soseki Muso
7	Nagasaki Port	Kyu Hamarikyu Gardens
8	Nagasaki City Office	Daitokuji Temple
9	West of Nagasaki	Nikka Yuko Garden
10	Dejima Museum of History	Western Pureland
11	City of Isahaya	Keiunkan Garden
12	Hirado City	Morikami
13	Hirado Island	Green Dragon Bonsai
14	Isahaya City	Karesansui Type
15	Nagasaki Lantern Festival	Nijo Castle Ninomaru Garden

4. まとめ

本稿では，Wikipedia の分野による偏りに対応した知識抽出を行うため，Web 上の情報を用い Wikipedia 中の欠落している概念を収集する手法の提案を行った．実験の結果より，収集された新規概念候補の中にはいくつかの新規概念が存在したが，言語解析の問題や語と概念とのマッチング処理の問題によって，不適切な語が収集されていることが分かった．今後の課題として，不要語の除去や同義語の判別による収集精度の向上と，収集した新規概念候補と既存概念との関連度を推定することによって，すでに Wikipedia から構築された連想シソーラスに新たな関連概念を追加する手法の検討を行う．

謝辞

本研究は，特別研究員奨励費 (09J04675)，科学研究費補助金基盤研究 (B)(21300032)，特定領域研究 (18049050) およびマイクロソフト産学連携研究機構 CORE 連携研究プロジェクトの助成によるものである．ここに記して謝意を表す．

参考文献

[Adler 08] Adler, B. T., Chatterjee, K., Alfaro, de L., Faella, M., Pye, I., and Raman, V.: Assigning Trust to Wikipedia Content, Technical report (2008)

[Ito 08] Ito, M., Nakayama, K., Hara, T., and Nishio, S.: Association Thesaurus Construction Methods based on Link Co-occurrence Analysis For Wikipedia, in *Proceedings of ACM International Conference on Information and Knowledge Management*, pp. 817–826 (2008)

[Nakayama 07] Nakayama, K., Hara, T., and Nishio, S.: Wikipedia Mining for An Association Web Thesaurus Construction., in *Proceedings of IEEE International Conference on Web Information Systems Engineering*, pp. 322–334 (2007)

[金本 08] 金本 径卓, 鈴木 優, 川越恭二：編集履歴に基づく Wikipedia における記事の信頼度算出手法, 情報処理学会研究報告 第 144 回データベースシステム研究会報告, 第 2008 巻, pp. 31–38 (2008)

*1 http://nlp.stanford.edu/software/CRF-NER.shtml
*2 http://developer.yahoo.com/search/boss/