Wikipedia記事における内容の信憑性を数値化するツールの作成：最適化された高速なオンラインアルゴリズムに向けて

東京学芸大学附属国際中等教育学校5年　2組木下修一　3組岩崎拓斗

研究概要(要旨)

Wikipediaはその知名度と規模に反して、誤情報の存在などを理由に公の場では情報源として活用されることが少ない。信憑性を細分化して評価する先行研究も存在するが、膨大なデータを事前に処理することを求めるオフラインアルゴリズムで構築されている。そのため本研究では、Wikipedia(英語版)において記事を単語レベルに細分化し、逐次処理のオンラインアルゴリズムを用いて**編集履歴からの差分追跡**を行い、高速に単語ごとの情報の信憑性を評価し、それを**視覚化**することを目標とする。

研究の背景・目的

Wikipediaは過度に信頼されることがある一方で、学術的にはほとんど無意味なものとして扱われている。信頼に値する情報とそうでない情報を明確に区別することで、**Wikipediaを有効な情報源として活用できるようにすること**が目的である。

まず、先行研究における評価の仕方の概要を説明する。Wikipediaではそれぞれのページについて編集履歴を見ることができる。これを使うことによって何回の編集を単語・文章が経験し、その編集によって変えられていないかがわかる。単語・文章がページが何回も編集されている間変わらなかったということは、そのページを編集した人がその内容を見て問題ないと判断したということである。Wikipediaのエディターは基本ボランティアで活動しているため、大半は悪い意図を持って編集しているわけではないと考えられるので、長い間残っている単語は信用できると判断可能である。この編集履歴を踏まえ、単語・文章がどのくらいの期間残っているかを考えるのが先行研究における評価の主軸である。

図1. 編集履歴の画面表示
箇条書きの一つ一つが、過去の版を示している。
更新日、著者、容量の順に記載されている。

先行研究の手法ではWikipediaから記事の情報全てを予め取得することが必要である(これを**オフラインアルゴリズム**という)ため時間がかかることがある。本研究に於いては、情報を逐次取得する度に評価を更新するような**処理速度の速いアルゴリズム**（**オンラインアルゴリズム**）**を作成し、それを視覚化する**ことによって研究目的を達成する。

研究の方法

上で示した方法をページの履歴を取得しながらできるように簡略化したものを実装すれば良いと考えた。先行研究ではどのエディターが文章を編集したか、またどこを編集したかなどの情報も考えて最終的な評価を出しているが、研究の第一段階としてはできる限り簡略化した評価の方法、すなわち単語がどのくらいの回数の編集を通して残っているかということのみを計算することを考える。今後はこの方法に少しずつ変化を加えていくことでより先行研究に近いもしくはより精度の高い方法を見つけることを考える。本研究はオンラインアルゴリズムをオフラインにしているため、このオンラインアルゴリズムがオフラインアルゴリズムと同じくらいの精度を出すためにはどのくらい時間がかかるのか、**オンラインアルゴリズムによる高速化を図る**ことによって評価できる。

参考文献

Adler, B. T., Chatterjee, K., De Alfaro, L., Faella, M., Pye, I., & Raman, V. (2008). Assigning trust to Wikipedia content. In Proceedings of the 4th International Symposium on Wikis,1-12

Tichy, W. F. (1984). The string-to-string correction problem with block moves. ACM Transactions on Computer Systems (TOCS), 2(4), 309-321.

引用文献

- Wikipedia. (2023年10月2日). Enteromius teugelsi: Revision history - Wikipedia. https://en.wikipedia.org/w/index.php?title=Enteromius_teugelsi&action=history (閲覧日: 2024年9月4日)
- Wikipedia. (2023年10月2日). Enteromius teugelsi. https://en.wikipedia.org/wiki/Enteromius_teugelsi (閲覧日: 2024年8月23日)

これまでの研究成果

右のフローチャートは現段階の研究で作成したコードを表したものである。まずこのプログラムは取得した古い版と現在の版にある単語をそれぞれマッチさせ、昔の版にあった単語のうちどれが今もあるかを追跡している。これは先行研究でも使われていた、W. Tichyが開発した**マッチングの方法を活用**している。このプログラムではこの手法を用い、できるだけマッチした単語同士を昔の版で繋がっているところは現在の版でも繋がるようにマッチさせている。これを行うことで文章単位でページの内容をマッチさせることができ、頻出する単語、例えば“is”や“the”などが異常に高い評価を得ないようにしている。それによって現在の版にある単語のうち、昔の版とマッチされたものが昔もあったと認識され、ページの計算が終わるごとにユーザーに表示される評価を変更する。

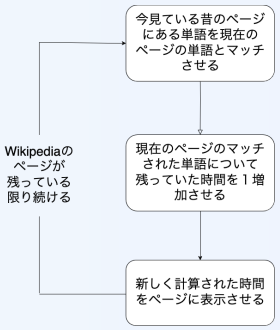


図2. アルゴリズムのフローチャート

このプログラムはこのページのマッチなどを計算する作業をWikipediaから**ページを取得している間に行っている**。履歴の取得はインターネットを通じて行われるため非常に遅いので、評価の計算と同時にやることで表示させるのにかかる時間を削減することに成功している。ユーザーに表示される評価はページに残っている時間から計算されたその情報が正しい確率である。これはある情報が間違っている時に取り除かれる確率が一定だと仮定し、指数関数の形で表される。

このアルゴリズムによって出された評価の例を右に表示する。これは適当なWikipediaのページに対して評価を計算したものである。透明に書かれているところはページが作成された時に一番最初に書かれた文章である。その後、二番目の文章が追加されている。この文章では“Its”が一番濃く現れているが、これはこの単語が一番最近に編集されたものであることが示されている。これは“It’s”という誤植を直したために生まれたものであるが、残されている時間を測るということが機能していることがわかる。

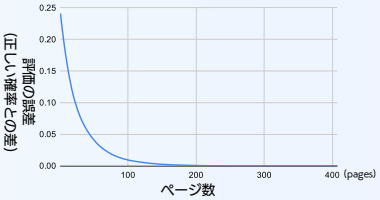
図3. 視覚的表示のイメージ

この視覚的表示は、**ChromeのExtensionを用いて作成**した。WikipediaのHTMLの構造を解析して、本来一つのnodeである文章を単語ごとに分割し、固有のタグを与えている。分割する基準はスペースであり、カンマやピリオドなど不必要なところは分割しない。またanchorも一つのまとまりとして捉え、分割を行っていない。任意のWikipedia(英語版)のページを開くと自動的に解析を開始し、その結果は数十秒程度で視覚的に表示される。

評価

オンラインアルゴリズムの評価としてはどのくらいの速度で真の値に近づくのかを検証した。ランダムなWikipediaのページ100個に対して実験した結果が図4である。およそ100回の編集で誤差が無視できるようになり、これは取得に４秒程度かかるので非常に現実的だと言える。また、アメリカ合衆国などの巨大なページでは履歴が50000以上ある(2024年現在)ため**最大500倍**ほどの高速化になり得る。

図4. 評価に使ったページ数と評価の誤差



今後の展望

まず次にやるべきことは、このアルゴリズムの精度を評価する方法である。これは先行研究ですでに確立されているのですぐできるはずである。今後はそれを用いて精度を高く信憑性を評価する方法を複数考えていきたい。例えば現在は単語がどのくらいの時間残っていたかだけを元にして評価しているが、先行研究でも行われているようにその単語を編集したユーザー、またユーザーがどこを編集しているかなどの複雑な手法を使用していきたい。

また、アルゴリズムと視覚的表現の整合性を向上させたい。現状では表が挿入されているなどの特殊な条件下で分析と視覚的表示にずれが生じていることから、実用性を向上させるためにも整合性を保てるようにしていきたい。