5L-06

Webサイトを意味的内容の一致度合により分類する手法の検討

吉田奏子†　　寺澤卓也‡
東京工科大学メディア学部メディア学科†‡

1. はじめに

情報検索の方法の 1 つであるキーワード検索では、キーワードをもとに検索エンジンが選んだ Web ページが順番に表示されるが、必ずしも上位に表示されるページに自分が欲しい情報が載っているとは限らない。また、同じような内容の別のページが検索上位に集中すると、その下にある別の内容のページが見つかりづらくなる可能性がある。これを改善するためには、似たような内容のページをグループ化して表示できれば良いと考え、Web サイトを意味ごとに分類する手法を検討することにした。

本研究では、一般的な情報検索のシステムであるキーワード検索の結果から自分の目的の情報に辿りつきやすくするため、Web サイトをページの内容ごとに分類し表示するために必要な手法を調査・検討し、実際に正しくグループ化できるか検証を行う事を目的とする。

2. 関連事例

2.1 Web サイトからの剽窃レポート発見支援システム

「Web サイトからの剽窃レポート発見支援システム」[1]は高橋勇らが発表した研究論文である。この論文では、Web ページからの剽窃の発見を支援するために、以下の 3 つの機能を備えたシステムを Web アプリケーションとして実装した。

- 剽窃元となる Web ページを探す Web 検索機能
- 集めた Web ページのうち、どのページがどの程度レポートの剽窃元の可能性があるか評価する剽窃評価機能
- どの部分が Web ページのコピーかを示す剽窃箇所特定機能

このシステムに対して、特定のアルゴリズムに沿って Web から疑似的に作成した剽窃レポートの検出実験と、実際の授業で出された課題レポートを用いた検出実験が行われた。前者の実験では、すべての剽窃元 Web ページが検出され、後者の実験ではこのシステムで剽窃の可能性が判断されたレポートは、手作業の評価でも剽窃と判断されることが示された。

2.2 doc2vec

doc2vec はニューラルネットワークの理論を利用した自然言語処理ツールである。doc2vec は文章を単語の集合として見てベクトルと捉えることにより、文書間の類似度計算を実現することができる。既存のモデルとの比較実験で、短文・長文感情分析と情報検索を行った結果、いずれも既存のものより精度が良くなっている[2]。

3. 分類手法の検討

先に触れた Web サイトからの剽窃レポート発見支援システムは単語や文章の類似度で剽窃かどうかを判断していたが、同じ単語を用いた文でも違う意味の文章を作ることができるため、それが誤検知につながる可能性もある。また、実際の Web ページには文章だけではなく図や表、プログラム、広告などが埋め込まれており、文章だけでの類似度の判断が難しい場合がある。

本研究では、問題点を解決するために、文章、図、表、プログラムの 4 点で分類の判別をする。本文と直接関係しない広告やメニューバー等は、検証の段階でスクレイピングを行い、検証しやすくする。

文章、図、表、プログラムを個別に分類方法を考え、その後ページ全体の内容で類似度を判断し、分類できるようにする。

4. 分類手法の実装

実装は LinuxOS のコンピュータ上で行った。ディストリビューションには Ubuntu16.04 LTS を選択した。

Study of the method to classify website by similarity of the contents
†Kanako Yoshida ‡Takuya Terasawa
†‡School of Media Science, Tokyo University of Technology

4.1 文章部分の解析

文章解析は、Python とテキスト解析を対象とした機械学習ライブラリである gensim、形態素解析エンジンである MeCab を利用して行う。

スクレイピングを行った複数の文章を Python と MeCab で形態素解析を行った後、gensim に含まれる doc2vec を利用し、文章同士の類似度を測る。doc2vec は文章をベクトルで表す他の方法と違い、語順の考慮ができるため、同じ単語を利用した違う意味の文章の分類もできると考えた。モデル学習時のパラメータは、学習に使う単語の最低出現回数は 1 に、分散表現の次元数は 300 に、コンテキストの文脈幅は 15 とし、学習率は 0.025 とした。

4.2 プログラム部分の解析

プログラムの分類は文章同様 doc2vec で分類することが考えられるが、時間の都合上実装ができなかった。

4.3 図・表部分の解析

図の解析は、Bag of Visual words(Bag-of-Features、Bag-of-Keypoints)と呼ばれる画像の特徴量を 1 次元のベクトルに次元圧縮する手法を用いれば類似度を測ることができる[3]。Python などで利用できるものとして OpenCV が存在し、これを利用すれば図の解析ができると考えられるが、今回は時間の都合上実装ができなかった。

5. 評価

5.1 概要

4.で実装した分類手法の評価を行う。この評価の目的は、グループ化の仕組みが正しく動作しているか確認し、改善、改良の余地があるか調べることである。

5.2 実験内容

分類手法の評価にあたり、Google ウェブ検索における「人工知能」の検索結果から 99 件、「SSL」の検索結果から 83 件、「ルーター」の検索結果から 69 件の web サイトの内容を取得した。それぞれのページで広告や画像などを取り除き、本文部分のみにスクレイピングを行い、個別のテキストファイルに保存して doc2vec で学習させた。スクレイピングした web ページから 3 件を選び、doc2vec の学習モデルをもとにそれぞれ類似するページを取得した。

5.3 実験結果

取得したページの中には比較元と同じジャンルのページが多く、比較元がルーターの解説文であれば取得できたのもルーターの解説文であるなど、概ね内容が似た記事が取得できた。しかし、中には内容が類似しないものも一部含まれていた。

6. おわりに

本研究では、Web ページの中の、文章、図、表、プログラムの 4 点で類似度判別を行い、類似度の高いページ同士をグループ化することを目指した。文章の類似度判別は Python と MeCab、doc2vec を利用することにより、ある 1 ページに対して類似するページを複数取得することができた。類似度は参照したページによるが、ある程度意味的な類似のあるページが取得できると考えられる。今後はパラメータや学習の方法などを検討し文章分類の精度を上げることにより、更に細かいキーワード検索でも分類できるようにする。

本研究において実現できなかったこととしてプログラム、図、表の分類方法を実装すること、文章を含めた全体での分類方法を検討、実装することが挙げられる。また、今回はスクレイピング済みのデータを手元に置いて実験を行ったが、実際の検索ではリアルタイムで処理を行っていくため、なるべく時間のかからない処理方法を考え直す必要がある。

参考文献

[1] 高橋勇, ほか. Web サイトからの剽窃レポート発見支援システム.
電子情報通信学会論文誌 D, Vol.J90-D, No.11, 2007. pp.2989-2999.
[2] Quoc,Le and Tomas, Mikolov.″Distributed Representations of Sentences and Documents.″ Proceedings of The 31st International Conference on Machine Learning (ICML 2014), pp. 1188-1196, 2014.
[3] 永橋知行, ほか. 画像分類における Bag-of-features による識別に有効な特徴量の傾向.
情報処理学会研究報告,コンピュータビジョンとイメージメディア(CVIM)Vol.2009-CVIM-169, 2009, pp1-8.