Web 上の学術リソースリポジトリの構築

難波英嗣
広島市立大学大学院 情報科学研究科
nanba@hiroshima-cu.ac.jp

1. はじめに

近年、ある研究の成果を他の研究者が再現できるよう、実験に用いたデータやツールを学術リソースとして Web 上で公開するのが一般的になってきている。Web 上に存在するこのようなリソースは、同じ分野の研究者だけでなく、非専門家にとっても最新の研究成果を利用した商用サービスを実現する上で非常に有用である。しかし、ある分野でそのリソースがどの程度一般的に使われているのかを知るのは、常に論文をチェックしているわけではない非専門家にとって容易ではない。そこで、本研究では、学術論文集合から学術リソースリポジトリを自動的に構築する手法を提案する。

計算機科学の分野では、論文の著者自身が開発したシステム、ベースラインシステム、実験に用いたデータ等を Web 上で公開し、その URL を論文中に記載することが少なくない。もし論文中の URL を自動抽出し、分類整理すれば、それらは有益な学術リソースリポジトリになりうる。

本稿では、次節で関連研究について述べ、3 節で学術リソースリポジトリの構築方法を説明する。また、提案手法の有効性を確認するために行った実験について報告する。4 節で本稿をまとめる。

2. 関連研究

2.1 論文と Web ページ間の引用関係の解析の解析

Web ページと論文間の引用(リンク)関係を扱った研究は、大きく、以下の 2 つのグループに分けることができる。

1. オンライン・ジャーナルなどの Web 上でアクセス可能な論文がどのような Web ページから引用されているのかを分析する研究[5, 8]
2. 論文中でどのような Web ページを引用しているのかを分析する研究[6, 9]

以下、それぞれの関連研究について述べる。Kousha ら[5]は、Web 上に存在する論文に引用(リンク)している Web ページについて、様々な側面から分析を行っている。この分析の観点として、例えば、引用元の Web ページを国別に分類したり、サイトのドメイン(org/com/edu など)別に分類したりしている。また、ある論文(誌)を引用する平均 Web ページ数とインパクトファクタの間には一定の相関があることを報告している。インパクトファクタとの相関に関して、Vaughan ら[8]も同様の結果を報告している。近年では、研究の影響力(research impact)を測るための尺度として altmetrics（例えば、https://www.altmetric.com など)が広く知られてきているが、この尺度では、ソーシャルメディアやオンラインニュース等における論文引用を利用している。

Lawrence ら[6]は、論文中で引用されている Web ページの持続性について調査している。Web ページは時間が経過するにつれ、ページそのものが消滅してしまうことがある。そこで、論文中で引用されている Web ページがどのくらい消滅するのかを、論文の著作年ごとに分けて集計している。その結果、論文が発表されて 5 年以上経過すると、論文中で引用されている Web ページの過半数は消滅すると報告している。一方で、こうして消滅してしまった Web ページの大半は、URL が変わっているだけで、検索エンジンなどを利用して同一内容の Web ページをすぐに見つけることができたり、同一でなくても関連性の高いページを見つけることができたりするとも報告している。Yang ら[9]は、3 つのデータセット(The Chinese Social Science Citation Index / Communication of the ACM, IEEE Computer / MEDLINE)について、論文中の URL を分析している。分析の観点として、以下のものが挙げられる。

- Web サイトのドメイン: com/net/org/edu/gov/ac/int など
- Web ページのタイプ: html/pdf/doc/ppt/動的なもの(php/jsp/asp)
- URL の頻度
- URL の深さ(URL に含まれる/の数)
- URL の長さ: URL の文字数

以上の関連研究を概観すると、Web ページと論文間の引用の性質を統計的に分析することに焦点を当てている。これに対し、本稿では、論文中での Web ページの引用を、第三者が再利用するためのシステムを構築することに主眼を置いている点が従来研究と異なる。

2.2 引用論文の分散表現

ある論文 d_t が他の論文 d_s から引用される時、その引用の文脈を読めば、論文 d_t の内容がある程度把握できる。この考え方に基づけば、論文の引用記号を単語とみなし、word2vec[7]を適用することで、引用論文の分散表現を獲得することができる[3]。

Han ら[2]は、paragraph vector[7]の考え方に基づいた引用論文の分散表現を獲得する手法を提案している。paragraph vector とは、ある文書 d と、文書 d 中に出現する単語列を入力とし、単語列の次に出現する単語を予測できるよう文書 d のベクトルを学習するモデルである。今、論文 d_s が論文 d_t を文脈 C で引用する場合を考える。Han らは、上述の文書 d、文書 d 中に出現する単語列、単語列の次に出現する単語をそれぞれ論文 d_s、文脈 C、論文 d_t と考え、論文 d_s と文脈 C を入力とし、論文 d_t が出力されるように論文 d_s のベクトルを学習するモデルを提案している。

本研究では、引用論文ではなく、論文中に記載された URL に word2vec を適用し、各 URL の分散表現を獲得する。次にこれを各 URL の内容を表現するキーワードの付与に利用する。

3. 学術リソースリポジトリの構築

本研究では、各リソースが分類、よく利用される順に並べることにより、学術リソースリポジトリを構築する。構築は 3 つの手順(1)学術論文からの学術リソースの所在(URL)の抽出、(2)各学術リソースへのキーワードの付与、(3)学術リソースの分類から構成される。これらの手順について、3.1 節、3.2 節、3.3 節でそれぞれ説明する。

3.1 学術論文からの学術リソースの所在の抽出

学術論文中の学術リソースの所在(URL)を抽出するためのルールベースの抽出器を構築した。抽出器を評価するため、ACL Anthology コーパス[1]から “http”という文字列を含んだ 2,212 文をランダムに選択し、文中の URL の個所に人手でタグ付けし、抽出器による抽出結果との一致度を再現率と精度により評価した。実験の結果、再現率 0.940、精度 0.896 が得られた。この抽出器を用いて ACL Anthology コーパスの 31,812 論文から、67,834 件の URL(異なり数 34,982 件)を抽出した。

3.2 学術リソースへのキーワードの付与

提案手法

このステップでは、各学術リソース(URL)に対し、その内容を示すキーワードが複数付与される。本研究では、このキーワード付与に“W2V-URL”という手法を提案する。まず、各 URL を ID 番号に置き換える。次に、ACL Anthology コーパス中の 31,812 論文を結合して作成したテキストファイルに word2vec[7]を適用し、各 URL の分散表現を獲得する。最後に、各 URL の分散表現と類似する単語の上位 n 件(=1, 3, 5, 10)を抽出し、これらを URL に付与するキーワードとする。最後のステップにおいて、記号、ストップワード(https://www.textfixer.com/tutorials/common-english-words.txt)、および ACL Anthology 中の頻度が 100 未満のものを事前に除外しておく。

比較手法

以下に示す 4 種類の提案手法および 2 種類のベースライン手法を用いて実験し、結果を比較する。

- W2V-URL@1 (提案手法): W2V-URLでURLと最も類似する単語
- W2V-URL@3 (提案手法): W2V-URLでURLと類似する単語上位3件
- W2V-URL@5 (提案手法): W2V-URLでURLと類似する単語上位5件
- W2V-URL@10 (提案手法): W2V-URLでURLと類似する単語上位10件
- TITLE (ベースライン手法): 各Webページ(URL)のTITLEタグ間の語句
- SNIPPET (ベースライン手法): Googleを用いてそのWebページを検索した時のスニペット

データセット

まず、ACL Anthology コーパス中の論文内に出現する URL のうちの 22 個をランダムに選定した。次に、各 URL を被験者に示し、Web ページの内容をよく読んでもらった上でおおよそ 10 個程度のキーワードを付与してもらった。以下は、Java ベースの統計的自然言語処理パッケージのひとつである MALLET(http://mallet.cs.umass.edu)に付与されたキーワードの例である。

machine learning mallet document classification sequence tagging topic modeling

評価と実験結果

提案手法およびベースライン手法を再現率と精度で評価した。実験結果を表 1 に示す。

表 1 各 URL へのキーワード付与に関する評価

手法	再現率	精度
W2V-URL@1 (提案手法)	0.042	0.647
W2V-URL@3 (提案手法)	0.088	0.479
W2V-URL@5 (提案手法)	0.107	0.350
W2V-URL@10 (提案手法)	0.134	0.222
TITLE (ベースライン手法)	0.199	0.481
SNIPPET (ベースライン手法)	0.236	0.379

TITLE および SNIPEET 手法は、URL ごとにそれぞれ 4.9 語および 19.0 語収集した。表 1 からわかるとおり、精度では提案手法のひとつである W2V-URL@1 が最も優れていたが、再現率ではベースライン手法の方が提案手法よりも優れていた。

考察

以下は、上述の MALLET の Web ページに関して、W2V-URL@10、TITLE、および SNIPPET により収集されたキーワードの例である。正解は下線で示してある。W2V-URL@10 の結果を見てわかるとおり、この手法では、WEKA、NLTK、SVM-Light のような MALLET のライバルツールの名前を誤ってキーワードとして収集しているが、これは分散表現の負の効果であると考えられる。

- W2V-URL@10: toolkit mallet weka python lemur csie nltk timbl package svmlight
- TITLE: mallet homepage
- SNIPPET: mallet java based package statistical natural language processing document classification clustering topic modeling information extraction machine learning applications text mallet includes sophisticated tools

一方で、評価に用いたデータだけでなく、他の URL についても実際にキーワードを付与した結果、提案手法には以下のような効果があることも分かった。

- Web ページが英語以外の言語で記述されている場合でも、提案手法は英語キーワードを出力することができる。
- HTML 以外のファイル(例えば pdf ファイル)でも、提案手法はキーワードを付与することができる。

TITLE 手法は、再現率でも精度でも比較した手法の中で 2 番目に良い値が得られている。しかし、TITLE 手法が適用できない Web ページは数多く存在している。ACL Anthology から抽出した 34,982 件の URL のうち、TITLE 手法で文字列が抽出できたものは 8,960 件(25%)に過ぎなかった。本実験では、この 8,960 件の中から 22 件を選んで実験を行ったため、比較的良好な結果が得られているが、実際には 75%の Web ページ(URL)には TITLE 手法ではキーワードが全く付与されなかった。

3.3 学術リソースの自動分類

このステップでは、以下の手順で学術リソース(URL)の自動分類を行う。

1. 学術論文をカテゴリ分類する分類器を構築する。
2. ACL Anthology コーパス中のすべての論文を分類する。
3. カテゴリごとにキーワードが付与された URL を論文中での引用数順にならべて表示する。

ここで、カテゴリごとにどのような学術リソースが利用されているのかを示すことを目的としている。例えば、機械翻訳システムを構築する際に、ある形態素解析ツールが頻繁に用いられている場合には、「機械翻訳」カテゴリにそのツールが分類される。

機械学習のためのデータセット

カテゴリの決定および機械学習用のデータセット構築のため、ACL Anthology コーパスに含まれる会議の開催プログラムを可能な限り収集した。これらのプログラムでは、各論文はひとつのセッションに割り当てられている。このセッション名を各論文のカテゴリ名とみなし、機械学習の際の教師データとして用いる。ここで、例えばセッション名“parsing”と “syntactic analysis”はほぼ同じ意味で使われていると考えられるため、このようなセッション名は統一する。この他に“summarization and generation”のようなテキスト要約と生成の両カテゴリに属するようなセッション名に含まれる論文は教師データから除外する。この結果、表 2 に示すデータセットが構築された。

表 2 カテゴリおよびカテゴリごとの論文数

カテゴリ名	論文数
Machine translation	335
Semantics	299
Syntax	192
Information extraction	173
Sentiment analysis	119
Discourse and dialogue	114
Machine learning	67
Morphology	50
Language resources	47
Summarization	46
Question answering	44
Information retrieval	29
Generation	29
Vision	27
Text categorization	24

分類器の構築

論文を分類するための学習器として fastText[4] を用いた。word2vec の次元として 100 を用い、5 分割交差検定を行ったところ、再現率、精度ともに 0.682 が得られた。この分類器を用いてすべての論文を分類し、カテゴリごとに URL を引用数順に並べたリストを作成した。

4. おわりに

本稿では、学術リソースリポジトリを自動構築する手法について述べた。まず、ACL Anthology コーパス中の 31,812 論文から 67,834 件の URL を抽出した。次に、各 URL に対してキーワードを付与する手法を提案した。最後に、URL をカテゴリごとに分類した。

謝辞

本研究は、科研費基盤研究(A)(15H01721)の支援を受けて行われた。

参考文献

1. Aizawa, A., Sagara, T., Iwatsuki, K., and Topic, G.: Construction of a New ACL Anthology Corpus for Deeper Analysis of Scientific Papers, Proceedings of the Third International Workshop on SCIentific DOCument Analysis (SCIDOCA2018) (2018).
2. Han, J., Song, Y., Zhao, W.Z., Shi, S., and Zhang, H.: hyperdoc2vec: Distributed Representations of Hypertext Documents, Proceedings of the 56th Annual Meeting of the Association for Computational Linsuistics, pp. 2384-2394 (2018).
3. Huang, W., Wu, Z., Liang, C., Mitra, P., and Giles, C.L.: A Neural Probabilistic Model for Context Based Citation Recommendation, Proceedings of the 29th AAAI Conference on Artificial Intelligence, pp. 2404-2410 (2015).
4. Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T.: Bag of Tricks for Efficient Text Classification, arXiv:1607.01759v3 [cs.CL] (2016).
5. Kousha, K. and Thelwall, M.: How Is Science Cited on the Web? A Classification of Google Unique Web Citations. Journal of the American Society for Information Science and Technology, 58(11), pp.1631-1644 (2007).
6. Lawrence, S., Pennock, D.M., William Flake, G., Krovets, R., Coetzee, F.M., Glover, E., Nielsen, F.A., Kruger, A., and Giles, C.L.: Persistence of Web References in Scientific Research, Computer, 34(2), pp. 26-31 (2001).
7. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, Advances in Neural Information Processing Systems 2013, pp. 3111-3119 (2013).
8. Vaughan, L. and Shaw, D.: Web Citation Data for Impact Assessment: A Comparison of Four Science Disciplines. Journal of the American Society for Information Science and Technology, 56(10), pp. 1075-1087 (2005).
9. Yang, S., Han, R., Ding, J., and Song, Y.: The Distribution of Web Citations. Information Processing & Management, 48, pp. 779-790 (2012).