Construction of an Academic Resource Repository

Nanba, Hidetsugu
Hiroshima City University

https://doi.org/10.5109/2230668

出版情報：Proceedings of Toward Effective Support for Academic Information Search Workshop. 1, pp.8-14, 2018-11-22. Faculty of Information Science and Electrical Engineering, Kyushu University
バージョン：
権利関係：

Construction of an Academic Resource Repository

Hidetsugu Nanba

Hiroshima City University, 3-4-1 Ozukahigashi, Asaminamiku, Hiroshima 731-3194 Japan
nanba@hiroshima-cu.ac.jp

Abstract. I define tools, libraries, and data which were created through various research as "academic resources." In this paper, I propose a system that constructs an academic resource repository. I extracted 67,834 URLs from 31,812 research papers in the ACL Anthology corpus. Then, my system annotated keywords to each URL. Finally, these URLs were classified into categories.

Keywords: web citation, word2vec, distributed representation, academic resource, URL.

1 Introduction

I define tools, libraries, and data, which were created through various researches, as "academic resources." Recently, many researchers have rolled out academic resources on the web for the purpose of re-examination by other researchers. These resources are useful not only for researchers in the same research field but also for developers in companies for making products making use of state-of-the-art techniques or data. However, non-academics cannot find these resources easily, or even if they find such resources, they cannot identify whether they are widely used in that field or not, because they do not always check the latest research papers. In this paper, I propose a system that constructs an academic resource repository from research papers automatically.

In the computer science field, authors of research papers mention the locations on the web (URLs) of their systems, baseline systems, and data used in the examinations. If these URLs are extracted from research papers, and classified for each research area, they will become a useful academic resource repository.

The remainder of this paper is organized as follows. In Section 2, I describe related work. In Section 3, I explain the procedure of constructing an academic resource repository, and then conclude in Section 4.

2 Related Work

2.1 Analysis of Citations Between Research Papers and Web Pages

There are several studies focusing on citations between research papers and web pages. These studies can be divided into the following two categories.

- Analyzing properties of web pages that cite online journal papers or online conference papers [5, 7]
- Analyzing properties of web pages that were cited in research papers [6, 9]

In the following, we will describe these research.

Kousha and Thelwall [5] analyzed web pages citing online research papers from various viewpoints. For example, they classified these web pages by the terms of site domains, such as ".org," ".com," or ".edu," and countries, such as ".jp," ".uk," or ".fr." They also identified statistical correlation between ISI[1] citations and web citations. Vaughan and Shaw [7] also reported this correlation. Recently, "altmetrics" [2] has become a well-known scholarly impact metrics, which also uses citations in social media, online news media, and so on, for calculating the research impact of each research paper.

Lawrence et al. [6] investigated the life-span of web pages cited in research papers. Generally, the inaccessibility of web pages increases by the time since the pages appeared on the web. Lawrence et al. counted the number of inaccessible web pages cited from research papers for each year. They reported that more than half of the web pages became inaccessible after five years since the research papers were published. On the other hand, most of these pages can easily be found using web search engines, because these pages do not disappear from the web but are just moved to other web sites. Yang et al. [9] investigated the properties of web pages cited in research papers in three databases, the Chinese Social Science Citation Index; Communication of the ACM, IEEE Computer; and MEDLINE, from the following viewpoints.

- Web site domain: .com, .net, .org, .edu, .gov, .ac, .int, and so on
- Type of web page: html, pdf, doc, ppt, dynamic pages such as php, jsp, asp, and so on
- Frequency of citation
- Length of URL (the number of characters)
- Depth of URL (the number of '/' in URL)

All of these previous researches mainly focused on statistical analysis of the properties of web pages citing or cited from research papers. In contrast, I focus on the construction of a repository, which enables non-academics access various academic resources.

2.2 Distributed Representation of Citations

Han et al. [3] proposed a method to express citations in research papers by distributed representations [7]. They regarded citation marks (symbols) in research papers as words. They applied word2vec [7] to research papers and obtained distributed representations for each citation. They used these representations for citation recommendation and research paper classification. In my work, I apply this idea to URLs in research papers, and obtain distributed representations for each URL. Then

[1] Thomson ISI (the Institute for Scientific Information)

we use these representations for annotating keywords that express the contents of each URL.

3 Construction of an Academic Resource Repository

I construct an academic resource repository, in which each resource is classified into research fields and is ranked according to its popularity. The procedure of the construction consists of three steps: (1) Extraction of academic resource locations (URLs) from research papers, (2) Annotation of keywords to each resource, and (3) Classification of resources into categories. In this section, I mention these steps in Sections 3.1, 3.2, and 3.3, respectively.

3.1 Extraction of Academic Resource Locations from Research Papers

For the extraction of academic resource locations (URLs) in research papers, I created a rule-based URL extraction system. To evaluate my system, I randomly selected 2,212 sentences including a string "http" from the ACL Anthology corpus [1], then manually identified URLs in each sentence. I applied my system to this data and obtained 0.896 of recall and 0.940 of precision, respectively. Finally, I applied the system to 31,812 research papers in the ACL Anthology corpus and then extracted 67,834 URLs in total (34,982 different URLs).

3.2 Annotation of Keywords for Each Resource

Proposed Method
In this step, some keywords are annotated to each resource (URL) for the purpose of showing their explanations to users. For this annotation, I propose a method "W2V-URL" based on the Hangs' method [3]. First, I replace all URLs in research papers by unique numbers. Second, I obtain distributed representations of each URL by applying word2vec [7] to a text file, which was created by concatenating 31,812 research papers in the ACL Anthology corpus. Third, I collect the top n (=1, 3, 5, and 10) words similar to each URL using distributed representations. In the third step, I preliminary removed symbols, stop words[2], and words that appear less than 100 times in the ACL Anthology corpus.

Alternatives
I examined the proposed method and two baseline methods as follows. For TITLE and SNIPPET methods, stop words were removed preliminary.

- W2V-URL@1 (proposed): Most similar one keyword by proposed method
- W2V-URL@3 (proposed): Top three keywords by proposed method
- W2V-URL@5 (proposed): Top five keywords by proposed method
- W2V-URL@10 (proposed): Top ten keywords by proposed method
- TITLE (baseline): Character strings between title tags of each web page (URL)
- SNIPPET (baseline): A snippet obtained by searching web pages using Google

[2] https://www.textfixer.com/tutorials/common-english-words.txt

Datasets

First, I randomly selected 22 URLs that appear in research papers in the ACL Anthology corpus. Second, I asked a human subject to annotate approximately 10 keywords to each URL by reading its web page. The following constitute an example of keywords for the MALLET homepage at http://mallet.cs.umass.edu. MALLET is a Java-based package for statistical natural language processing.

machine learning mallet document classification sequence tagging topic modeling

Evaluation and Results

I evaluated the proposed method and two baseline methods by recall and precision. The experimental results are shown in Table 1.

Table 1. Experimental results for annotating keywords to each URL.

Methods	Recall	Precision
W2V-URL@1 (proposed method)	0.042	0.647
W2V-URL@3 (proposed method)	0.088	0.479
W2V-URL@5 (proposed method)	0.107	0.350
W2V-URL@10 (proposed method)	0.134	0.222
TITLE (baseline method)	0.199	0.481
SNIPPET (baseline method)	0.236	0.379

The TITLE and SNIPPET methods collected 4.9 and 19.0 words for each URL, respectively. As can be seen from Table 1, the proposed method W2V-URL@1 is superior to baseline methods in terms of precision, while the proposed method is inferior to baseline methods in terms of recall.

Discussion

Followings are examples of outputs by W2V-URL@10, TITLE, and SNIPPET methods. I show correctly extracted words with underline. The results showed that W2V-URL@10 mistakenly extracted some words for rival tools, such as WEKA, NLTK, and SVM-Light. This is considered as negative effect of distributed expressions for URLs.

- 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

Although there are some errors due to distributed expressions, there are still the following merits to use.

- The proposed method can output English keywords, even if the web page is written in other languages.
- The proposed method can output keywords, even if the web page is not an HTML file (e.g. pdf file).

The TITLE method obtained the second highest recall and precision scores among all. However, there are many cases that the TITLE method is not applicable. Among 34,982 URLs that I extracted from the ACL Anthology corpus, the number of cases that I could extract character strings between title tags of each web page was only 8,960 (25%). In this experiment, I chose URLs from these 8,960 cases. However, there are various types of web pages such as pdf files [9], and the TITLE method cannot output any keywords for the remaining 75% cases.

3.3 Classification of Resources into Categories

In the final step, I classified URLs into a category using the following procedure.

1. Construct a research paper classifier to a category
2. Classify all research papers in the ACL Anthology corpus
3. Show URLs in research papers for each category by the number of citations together with keywords.

Here, I intend to show what kind of resources are often used for implementing a system for each category. For example, if a part-of-speech tagging tool is often used in implementing machine translation systems, the tool appears in the "machine translation" category.

Datasets for Machine Learning

In order to determine categories and to make training and test datasets, I collected past conference programs about the ACL Anthology corpus. In these programs, each research paper was classified into one of the conference sessions. I regarded each session name as a category of the paper and used the names as training and test datasets for constructing a machine learning-based classifier. Here, session names have a vague aspect. For example, "syntactic analysis" is used in a conference, while "parsing" is used in other conferences, although "syntactic analysis" and "parsing" have the same meaning. I therefore modified such vagueness manually, and finally constructed a dataset for machine learning. Table 2 shows the categories and the number of papers for each category.

Construction of a Classifier

As a machine-learning framework for classifying research papers, we employed fastText [4], which is a neural-based text classifier using word2vec. We used 100 as a

dimension value for word2vec. We conducted five-fold cross-validation. We obtained 0.682 for both recall and precision. Then, I classified all research papers in the ACL Anthology corpus and made a list of URLs ranked by their frequency in each category.

Table 2. Categories and the number of papers.

Category	The number of papers
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

4 Conclusions

In this paper, I propose a system that constructs an academic resource repository. I extracted 67,834 URLs from 31,812 research papers in the ACL Anthology corpus. Then, my system annotated keywords to each URL. Finally, these URLs were classified into categories.

References

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. altmetrics: https://www.altmetric.com (accessed on Oct, 21, 2018)
3. 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).
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).