

Home | Computational Linguistics | List of Issues | Volume 27 , No. 1 | Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus



Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus

[Mikio Yamamoto](#) and [Kenneth W. Church](#)

Quarterly (March, June, September, December)

160pp. per issue

6 3/4 x 10

Founded: 1974

2018 Impact Factor: 1.319

2018 Google Scholar h5-index: 32

ISSN: 0891-2017

E-ISSN: 1530-9312

Posted Online March 13, 2006

<https://doi.org/10.1162/089120101300346787>

© 2001 Association for Computational Linguistics

Computational Linguistics
Volume 27 | Issue 1 | March 2001
p.1-30



Download Options



Journal Resources

[Editorial Info](#)

[Abstracting and Indexing](#)

[Release Schedule](#)

[Advertising Info](#)

Abstract Authors

Bigrams and trigrams are commonly used in statistical natural language processing; this paper will describe techniques for working with much longer n-grams. Suffix arrays (Manber and Myers 1990) were first introduced to compute the frequency and location of a substring (n-gram) in a sequence (corpus) of length

Author

Resources

Submission

Guidelines

Publication

Agreement

Author Reprints

Reader

Resources

Rights and

Permissions

Most Read

Most Cited

More About

Computational

Linguistics



Metrics



77 Total citations

6 Recent

citations

15 Field Citation

Ratio

n/a Relative

Citation Ratio

Open Access



Computational
Linguistics

Computational
Linguistics is

Open Access.

All content is

freely available

in electronic

format (Full

text HTML,

PDF, and PDF

N. To compute frequencies over all $N(N+1)/2$

substrings in a corpus, the substrings are grouped into a manageable number of equivalence classes. In this

way, a prohibitive computation over substrings is

reduced to a manageable computation over classes.

This paper presents both the algorithms and the code

that were used to compute term frequency (tf) and

document frequency (df) for all n-grams in two large

corpora, an English corpus of 50 million words of Wall

Street Journal and a Japanese corpus of 216 million

characters of Mainichi Shimbun.

The second half of the paper uses these frequencies

to find “interesting” substrings. Lexicographers have

been interested in n-grams with high mutual

information (MI) where the joint term frequency is

higher than what would be expected by chance,

assuming that the parts of the n-gram combine

independently. Residual inverse document frequency

(RIDF) compares document frequency to another

model of chance where terms with a particular term

frequency are distributed randomly throughout the

collection. MI tends to pick out phrases with

noncompositional semantics (which often violate the

independence assumption) whereas RIDF tends to

highlight technical terminology, names, and good

keywords for information retrieval (which tend to

exhibit nonrandom distributions over documents). The

combination of both MI and RIDF is better than either

by itself in a Japanese word extraction task.

Forthcoming

Most Read

See More

Lexicon-Based

Methods for
Sentiment Analysis
(14087 times)

Maite Taboada et al.

Computational Linguistics

Volume: 37, Issue: 2, pp.
267-307

Computational

Linguistics and Deep
Learning (10542
times)

Christopher D.

Manning

Computational Linguistics

Volume: 41, Issue: 4, pp.
701-707

Near-Synonymy

and Lexical Choice
(3675 times)

Philip Edmonds et al.

Computational Linguistics

Volume: 28, Issue: 2, pp.


105-144


(Note that the Most Read numbers are based on the number of full text downloads over the last 12 months.)


Most Cited

See More

Plus) to readers across the globe. All articles are published under a [CC BY-NC-ND 4.0 license](#). For more information on allowed uses, please view the CC license. [Support OA at MITP](#)






 **Lexicon-Based Methods for Sentiment Analysis** (436 times)
Maite Taboada et al.
Computational Linguistics
Volume: 37, Issue: 2, pp. 267-307

 **A Systematic Comparison of Various Statistical Alignment Models** (174 times)
Franz Josef Och et al.
Computational Linguistics
Volume: 29, Issue: 1, pp. 19-51

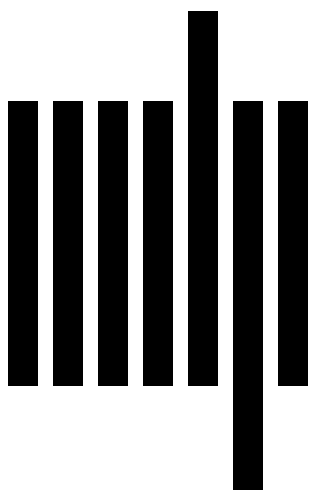
 **Opinion Word Expansion and Target Extraction through Double Propagation** (147 times)
Guang Qiu et al.
Computational Linguistics
Volume: 37, Issue: 1, pp. 9-27

(Note that the Most Cited numbers are based on Crossref's [Cited-by service](#) and reflect citation information for the past 24 months.)

 **Download Options** >

- Favorite 
- Sign up for Alerts 
- Download Citation 
- RSS TOC 
- RSS Citation 
- Submit your article

[Support OA at MITP](#) 



[Journals](#)

[Terms & Conditions](#)

[Privacy Statement](#)

[Contact Us](#)

[Books](#)

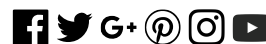
[US](#)

[UK](#)

[Connect](#)

One Rogers Street
Cambridge MA 02142-1209

Suite 2, 1 Duchess Street
London, W1W 6AN, UK



© 2018 The MIT Press
Technology Partner:
[Atypon Systems, Inc.](#)
[CrossRef Member](#)
[COUNTER Member](#)
The MIT Press colophon is registered

in the U.S. Patent and Trademark Office.
[Site Help](#)

