jebari_2021_the_use_of_citation_context_to_detect_the_evolution_of_research_topics_a_large_scale_analysis

Year

2021

Author(s)

Jebari, Chaker and Herrera-Viedma, Enrique and Cobo, Manuel Jesus

Title

The use of citation context to detect the evolution of research topics: a large-scale analysis

Venue

Scientometrics

Topic labeling

Manual

Focus

Secondary

Type of contribution

Established approach

Underlying technique

Manual labeling

Topic labeling parameters

Nr of inspected terms: 7

Label generation

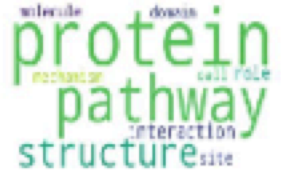
Using the 50 top terms in each topic, we presented these distributions as a wordcloud, where the size of the word is proportional to its probability.

Most publications in MEDLINE/PubMed are manually assigned a set of descriptors from MeSH by biomedical experts at the US NLM.
The descriptors or subject headings are arranged in a hierarchy. When performing a MEDLINE search via PubMed, entry terms are automatically mapped to the corresponding descriptors. In this study, we mapped the seven top terms in each topic to the corresponding descriptor which is used as a topic label.

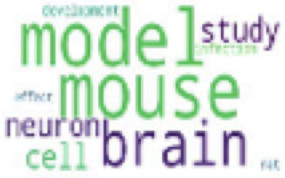
(a) Topic 1: mast cells

(b) Topic 2: risk factors

(c) Topic 3: protein structure, tertiary



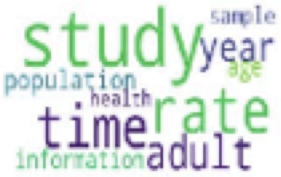
(d) Topic 4: models, theoretical



(e) Topic 5: gene expression profiling

(f) Topic 6: immunity

(g) Topic 7: time and motion studies



(h) Topic 8: behavioral research

(i) Topic 9: clinical trial

(j) Topic 10: receptors, vascular endothelial growth factor

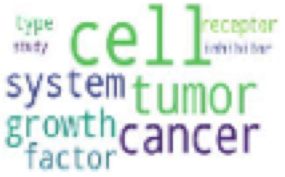


Fig. 3 Discovered topics

Motivation

\

Topic modeling

DTM (with LDA initialization)

Topic modeling parameters

To infer the model parameters variational Kalman filtering and variational wavelet regression are used.

Nr of topics (K): 10

Alpha: 0.01

Beta: 0.005

Nr. of topics

10

Label

Medical Subject Headings (MeSH), a comprehensive controlled vocabulary created and updated by the US NLM to facilitate searching, are used as labels.

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Paper: Biomedical & Life Science

Dataset: Biomedical & Life Science

Problem statement

With the exponential increase in the number of published papers, discovering how topics evolve becomes increasingly important for anybody involved in research, including researchers, institutes, research funding bodies, and decision-makers. This study proposes a large-scale analysis of the evolution of biomedical and life sciences using the citation contexts of the collected papers, or more precisely their citing sentences.

Corpus

Origin: PubMed Central

Nr. of documents: 64,350

Details:

- papers published between 2008 and 2018

Document

Citation context of a paper.

To extract the citation sentences that cite a given paper, we replaced the reference enclosed between <xref ref-type="bibr"*><*/xref> with its PMID. To ensure that all papers are cited by other papers published within the same period (from 2008 to 2018), we decided to remove papers that are not cited at all by papers from 2008 to 2018. After that, for each paper, we extracted the citation sentences where the paper is cited. Table 1 presents the different citations of the paper with PMID=18988837 by four other papers published in different years.

Table 1 Citation context example

PMID	Year	Citation sentence
19683024	2010	It is well known that linkage analysis is powerful in detecting rare and high risk alleles but has limited power in identifying common genetic variants with low-penetrance [<xref rid="R29">15516958</xref>,<xref rid="R30">18988837</xref>].
19818800	2011	Faster and more efficient genotyping methods have propelled studies that seek to identify QTL in many different systems, including humans (<xref rid="R1">18988837</xref>)
20639796	2014	A key assumption in GWAS is what is known as the common disease/common variant hypothesis [<xref rid="R18">18988837</xref>]
20552648	2016	Determining the genetic basis of complex genetic diseases is one of the main challenges in human genetics [<xref rid="R2">18988837</xref>

Besides the citation sentences, we extracted for each paper the publication year, the name of the

journal in which it is published, and the affiliations of the authors.

Pre-processing

- tokenisation
- stopword removal
- POS tagging
- To apply the DTM algorithm, we grouped our papers into 11 time slices (from 2008 to 2018)

```
@article{jebari_2021_the_use_of_citation_context_to_detect_the_evolution_of_rese
arch_topics_a_large_scale_analysis,
    abstract = {With the exponential increase in the number of published papers,
discovering how topics evolve becomes increasingly important for anybody
involved in research, including researchers, institutes, research funding
bodies, and decision-makers. This study proposes a large-scale analysis of the
evolution of biomedical and life sciences using the citation contexts of the
collected papers, or more precisely their citing sentences. Using 64,350 papers
published in PubMed Central between 2008 and 2018, we determined the research
trends for ten research topics. Moreover, we studied how these topics evolve
across countries and across the most common journals in biomedical and life
sciences.},
    author = {Jebari, Chaker and Herrera-Viedma, Enrique and Cobo, Manuel
Jesus},
    date-added = {2023-04-12 22:15:26 +0200},
    date-modified = {2023-04-12 22:15:26 +0200},
    day = {01},
    doi = {10.1007/s11192-020-03858-y},
    issn = {1588-2861},
    journal = {Scientometrics},
    month = {Apr},
    number = {4},
    pages = {2971--2989},
    title = {The use of citation context to detect the evolution of research
topics: a large-scale analysis},
    url = {https://link.springer.com/content/pdf/10.1007/s11192-020-03858-
y.pdf},
    volume = {126},
    year = {2021},
    bdsk-url-1 = {https://link.springer.com/content/pdf/10.1007/
```

```
s11192-020-03858-y.pdf},
    bdsk-url-2 = {https://doi.org/10.1007/s11192-020-03858-y}}
```

#Thesis/Papers/FS