Anne O'Tate: Value-added PubMed search engine for analysis and text mining

Neil R. Smalheiser*[1], Dean P. Fragnito[2], and Eric E.Tirk[2]

[1] Department of Psychiatry, University of Illinois at Chicago, Chicago, IL 60612, USA

[2] Xornet Inc., 2710 English Road, Rochester, NY 14616, USA

*Corresponding author

Email: neils@uic.edu

Abstract

Over a decade ago, we introduced Anne O'Tate, a free, public web-based tool http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/AnneOTate.cgi to support user-driven summarization, drill-down and mining of search results from PubMed, the leading search engine for biomedical literature. It has been deployed continuously, serving a wide range of biomedical users and needs, and over time has also served as a platform to support the creation of new tools that address additional needs. Here we describe the current, greatly expanded implementation of Anne O'Tate and invite the scientific community to explore how it can assist in analyzing biomedical literature, in a variety of use cases.

Introduction

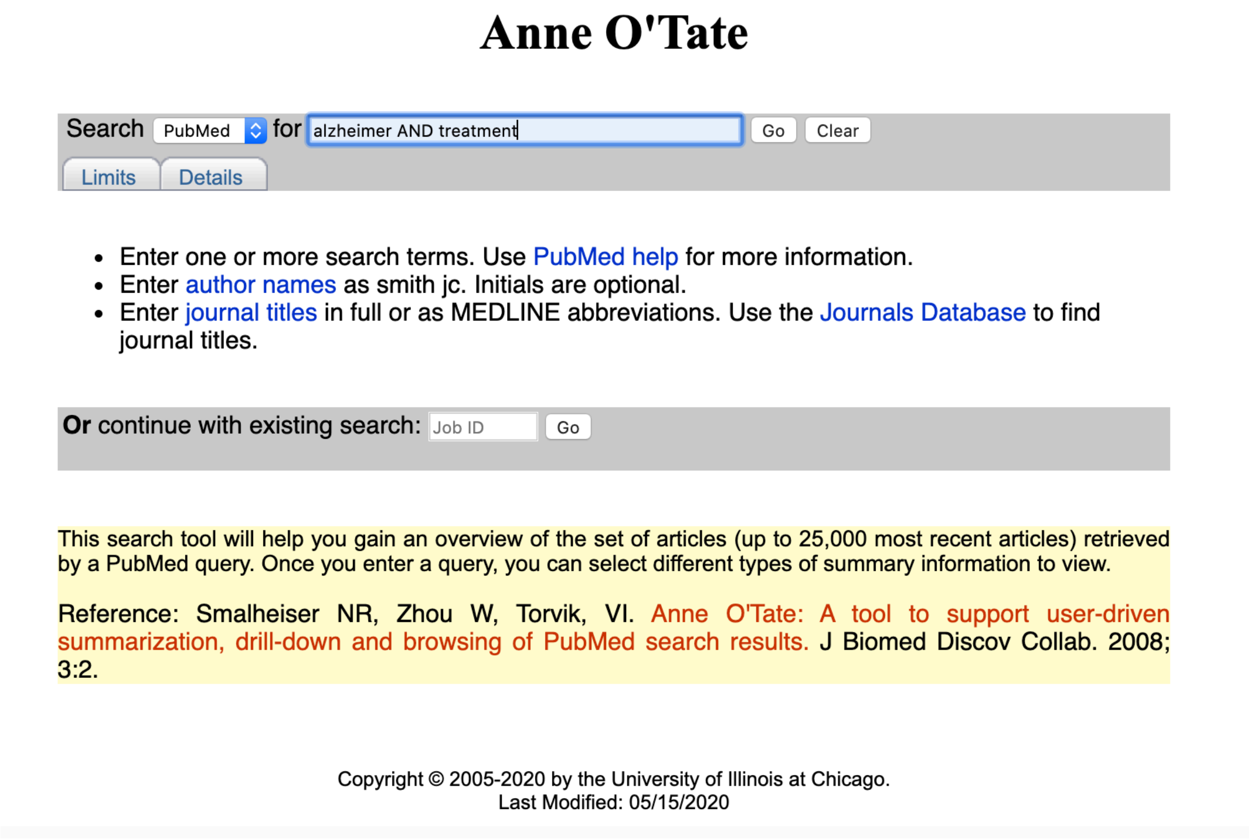
PubMed is the most used, and arguably most advanced search engine, for retrieving biomedical literature [https://www.ncbi.nlm.nih.gov/pubmed/, accessed May 15, 2020]. It has sophisticated automated features such as mapping queries to Medical Subject Heading terms, imputing articles written by the same individual, spelling correction, and other tools that assist a wide variety of users who range from physicians and bench scientists to patients and their relatives. Despite this advanced user-focused engineering, the output of a PubMed query is a simple list of articles, and users have few options for summarizing or mining this output further.

We developed Anne O'Tate to be an integrated, generic tool for summarization, drill-down and browsing of PubMed search results that accommodates a wide range of biomedical users and needs [1]. Briefly, Anne O'Tate allows the user to carry out a PubMed query. After displaying the list of retrieved articles, the user can choose to visualize multiple aspects of the articles to the user, according to pre-defined categories such as the "most important" words found in titles or abstracts; topics; journals; authors; publication years; and affiliations. Clicking on a given item opens a new window that displays all papers that contain that item. One can navigate by drilling down through the categories progressively, e.g., one can first restrict the articles according to author name and then restrict that subset by affiliation. Alternatively, one can expand small sets of articles to display the articles that are most closely related to the set as a whole.

Over the past decade, Anne O'Tate has acquired a steady community of regular visitors and has been reviewed by others [2]. As new needs have become apparent, we have added new facets for summarizing and drilling-down the results of queries. We have also used Anne O'Tate as a platform to support new, more advanced text processing programs. In this report, we present a status update and overview of the tool as it exists currently.

Results

Suppose we enter the query [Alzheimer AND treatment] into the Anne O'Tate query interface (fig. 1). The query box is simplified compared to the PubMed home page but retains hotlinks so that the user can specify query Limits, and see and edit the exact query Details as processed by PubMed. The query is passed to PubMed, which processes the query and returns a list of 42,600 articles in reverse chronological order (fig. 2). Each user session is given a unique job ID and is saved in the webserver for approximately six months; the user can return to the most recent processed query by entering the Job ID into a separate query box on the homepage (fig. 1).



Anne O'Tate

Search PubMed for alzheimer AND treatment

- Enter one or more search terms. Use PubMed help for more information.
- Enter author names as smith jc. Initials are optional.
- Enter journal titles in full or as MEDLINE abbreviations. Use the Journals Database to find journal titles.

Or continue with existing search: Job ID

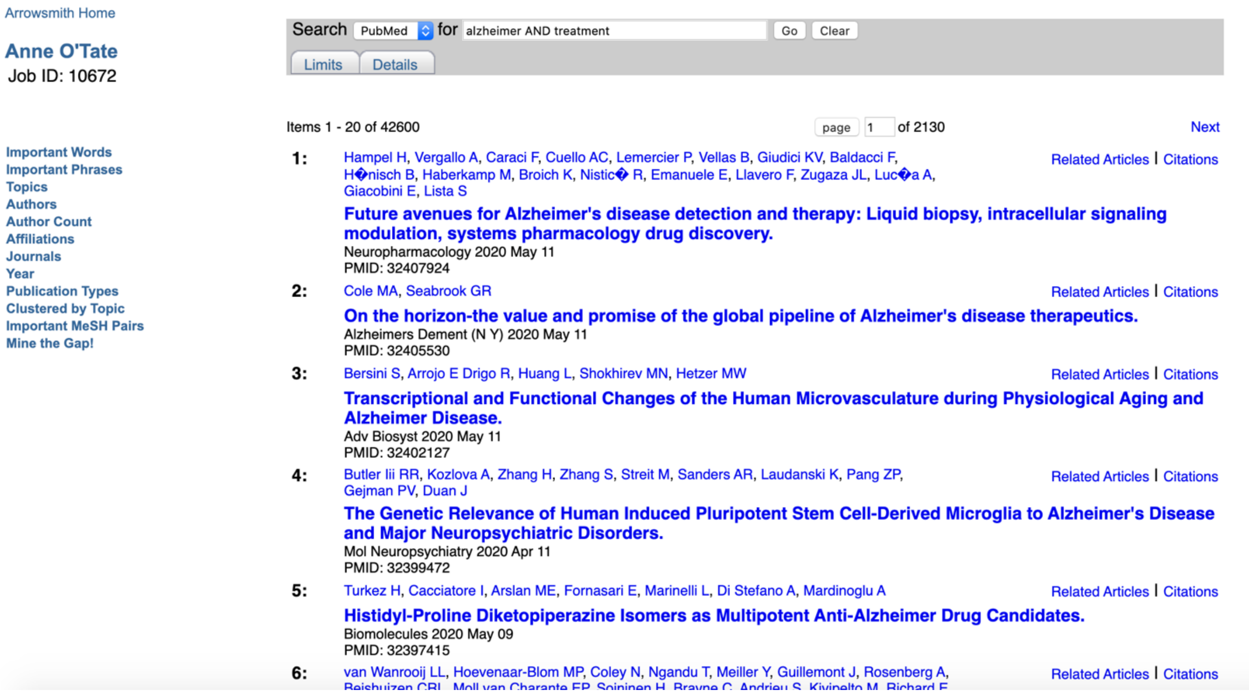
This search tool will help you gain an overview of the set of articles (up to 25,000 most recent articles) retrieved by a PubMed query. Once you enter a query, you can select different types of summary information to view.

Reference: Smalheiser NR, Zhou W, Torvik, VI. Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results. J Biomed Discov Collab. 2008; 3:2.

Copyright © 2005-2020 by the University of Illinois at Chicago.
Last Modified: 05/15/2020

Figure 1. Screenshot of the Anne O’Tate homepage.

Each displayed article (fig. 2) has two hotlinks on its right side: “Related articles” opens a new tab that displays a ranked list of the most related articles as computed by PubMed using the PubMed related article algorithm which is largely based on word usage similarity [3]. The “Citations” hotlink opens a new tab that displays the Citation Cloud surrounding that article (see below).



Items 1 - 20 of 42600 page 1 of 2130

1: Hampel H, Vergallo A, Caraci F, Cuello AC, Lemercier P, Vellas B, Giudici KV, Baldacci F, H�nisch B, Haberkamp M, Broich K, Nistic� R, Emanuele E, Llavero F, Zugaza JL, Luc�a A, Giacobini E, Lista S — Related Articles | Citations
Future avenues for Alzheimer's disease detection and therapy: Liquid biopsy, intracellular signaling modulation, systems pharmacology drug discovery.
Neuropharmacology 2020 May 11
PMID: 32407924

2: Cole MA, Seabrook GR — Related Articles | Citations
On the horizon-the value and promise of the global pipeline of Alzheimer's disease therapeutics.
Alzheimers Dement (N Y) 2020 May 11
PMID: 32405530

3: Bersini S, Arrojo E Drigo R, Huang L, Shokhirev MN, Hetzer MW — Related Articles | Citations
Transcriptional and Functional Changes of the Human Microvasculature during Physiological Aging and Alzheimer Disease.
Adv Biosyst 2020 May 11
PMID: 32402127

4: Butler Iii RR, Kozlova A, Zhang H, Zhang S, Streit M, Sanders AR, Laudanski K, Pang ZP, Gejman PV, Duan J — Related Articles | Citations
The Genetic Relevance of Human Induced Pluripotent Stem Cell-Derived Microglia to Alzheimer's Disease and Major Neuropsychiatric Disorders.
Mol Neuropsychiatry 2020 Apr 11
PMID: 32399472

5: Turkez H, Cacciatore I, Arslan ME, Fornasari E, Marinelli L, Di Stefano A, Mardinoglu A — Related Articles | Citations
Histidyl-Proline Diketopiperazine Isomers as Multipotent Anti-Alzheimer Drug Candidates.
Biomolecules 2020 May 09
PMID: 32397415

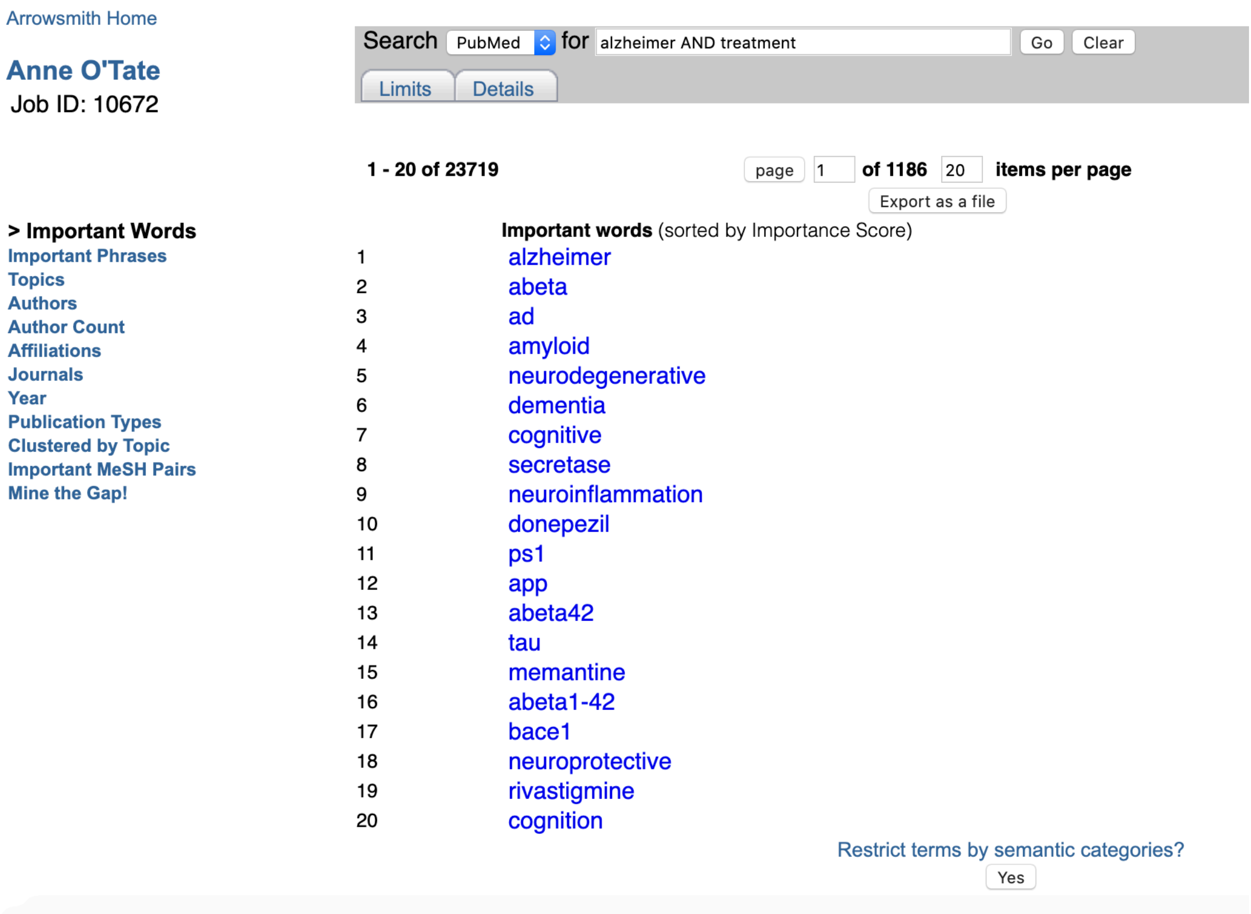
6: van Wanrooij LL, Hoevenaar-Blom MP, Coley N, Ngandu T, Meiller Y, Guillemont J, Rosenberg A, Beishuizen CRL, Moll van Charante EP, Soininen H, Brayne C, Andrieu S, Kivipelto M, Richard E — Related Articles | Citations

Figure 2. Screenshot of the list of articles retrieved from PubMed using the query [Alzheimer AND treatment].

Buttons for focused summarization, drill-down and analysis.

On the left side of the page displaying the list of articles, there are 12 hotlinked facets or "buttons" that the user can choose to mine the set of retrieved articles further (fig. 2). Those buttons which are new or modified since the previous report on Anne O'Tate are marked with asterisks in the following paragraphs. Each button opens up a new tab that displays a processed result. This can be viewed or processed further. We will discuss each button in turn:

Important Words*. This computes the words that are significantly over-represented in the retrieval set (here, 42,500 articles on Alzheimer AND treatment) compared to MEDLINE as a whole. They are ranked in order of their "importance", i.e., the degree to which the word is over-represented [1, 4]. The list can be further filtered using a button to restrict the terms to one or more semantic categories (taken from the UMLS [5] as described [1]). clicking on any one initiates a new query restricted to articles that mention that word in any field of the article's PubMed metadata (title, abstract, or other metadata field). The important words are displayed in stemmed form [6] – for example, the listed word "amyloid" in Fig. 3 comprises mentions of both "amyloid" and "amyloids". Accordingly, clicking on the hotlinked term "amyloid" initiates a new, more restricted PubMed query: [Alzheimer AND treatment AND (amyloid [all fields] OR amyloids [all fields])], and the resulting list of 13,858 articles is displayed in a new tab (Fig. 4). The list of Important Words together with their Importance scores [1, 4] can be filtered according to UMLS semantic category, and can also be exported as a CSV file for use in text mining.



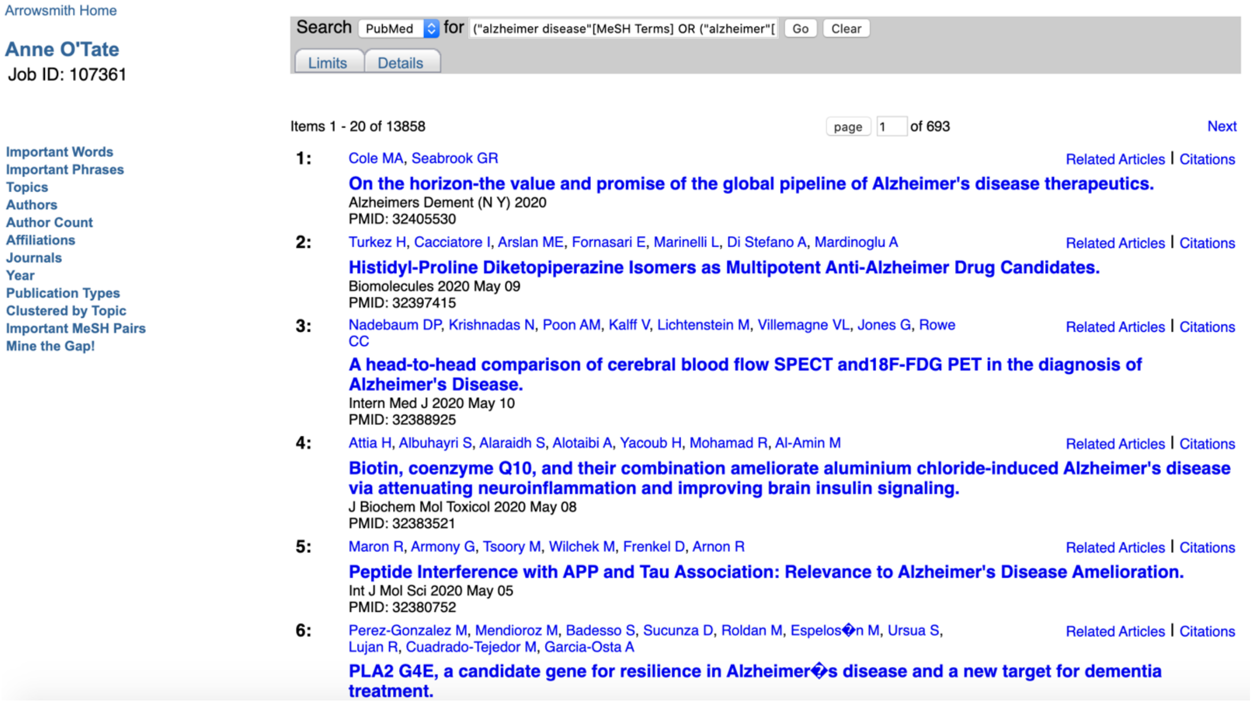
Anne O'Tate
Job ID: 10672

Search PubMed for alzheimer AND treatment

1 - 20 of 23719 page 1 of 1186 20 items per page

	Important words (sorted by Importance Score)
1	alzheimer
2	abeta
3	ad
4	amyloid
5	neurodegenerative
6	dementia
7	cognitive
8	secretase
9	neuroinflammation
10	donepezil
11	ps1
12	app
13	abeta42
14	tau
15	memantine
16	abeta1-42
17	bace1
18	neuroprotective
19	rivastigmine
20	cognition

Figure 3. Screenshot of Important Words calculated for the query [Alzheimer AND treatment].



Items 1 - 20 of 13858

1: Cole MA, Seabrook GR — Related Articles | Citations
On the horizon-the value and promise of the global pipeline of Alzheimer's disease therapeutics.
Alzheimers Dement (N Y) 2020
PMID: 32405530

2: Turkez H, Cacciatore I, Arslan ME, Fornasari E, Marinelli L, Di Stefano A, Mardinoglu A — Related Articles | Citations
Histidyl-Proline Diketopiperazine Isomers as Multipotent Anti-Alzheimer Drug Candidates.
Biomolecules 2020 May 09
PMID: 32397415

3: Nadebaum DP, Krishnadas N, Poon AM, Kalff V, Lichtenstein M, Villemagne VL, Jones G, Rowe CC — Related Articles | Citations
A head-to-head comparison of cerebral blood flow SPECT and18F-FDG PET in the diagnosis of Alzheimer's Disease.
Intern Med J 2020 May 10
PMID: 32388925

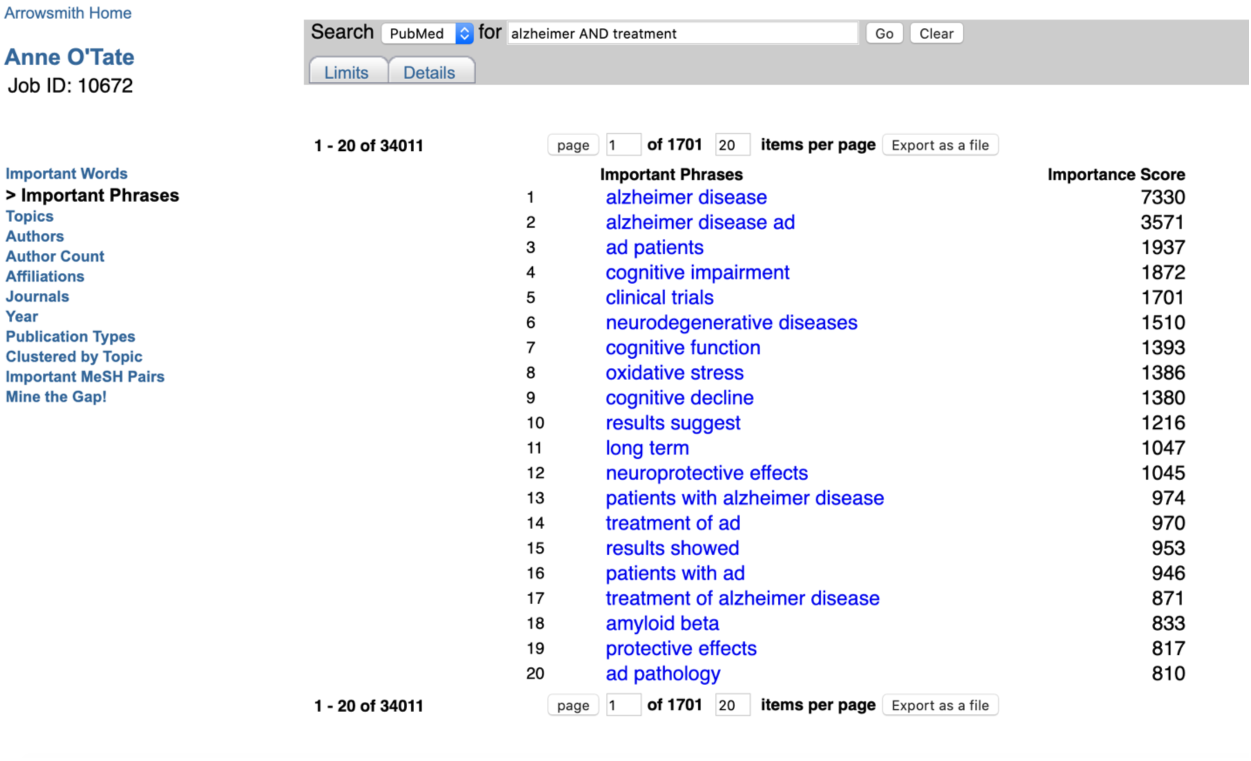
4: Attia H, Albuhayri S, Alaraidh S, Alotaibi A, Yacoub H, Mohamad R, Al-Amin M — Related Articles | Citations
Biotin, coenzyme Q10, and their combination ameliorate aluminium chloride-induced Alzheimer's disease via attenuating neuroinflammation and improving brain insulin signaling.
J Biochem Mol Toxicol 2020 May 08
PMID: 32383521

5: Maron R, Armony G, Tsoory M, Wilchek M, Frenkel D, Arnon R — Related Articles | Citations
Peptide Interference with APP and Tau Association: Relevance to Alzheimer's Disease Amelioration.
Int J Mol Sci 2020 May 05
PMID: 32380752

6: Perez-Gonzalez M, Mendioroz M, Badesso S, Sucunza D, Roldan M, Espelos�n M, Ursua S, Lujan R, Cuadrado-Tejedor M, Garcia-Osta A — Related Articles | Citations
PLA2 G4E, a candidate gene for resilience in Alzheimer�s disease and a new target for dementia treatment.

Figure 4. Screenshot of the list of articles that mention "amyloid" or "amyloids" within the original query.

Important Phrases*. This runs the TopMine algorithm [7] to identify phrases that are important within the titles and abstracts of retrieved articles (without comparison to their frequency in MEDLINE). For the [Alzheimer AND treatment] query, we see phrases such as "Alzheimer disease", "cognitive function", and "oxidative stress" (Fig. 5). Clicking on any phrase initiates a new restricted query [Alzheimer AND treatment AND "exact phrase" [tiab]] whose results are shown in a new tab. Note that, unlike Important Words, Important Phrases are not stemmed and they are mined only from titles and abstracts, not all fields of the PubMed record. The list of Important Phrases can also be exported as a CSV file for use in text mining.



Job ID: 10672

Search PubMed for alzheimer AND treatment

1 - 20 of 34011 page 1 of 1701 20 items per page

	Important Phrases	Importance Score
1	alzheimer disease	7330
2	alzheimer disease ad	3571
3	ad patients	1937
4	cognitive impairment	1872
5	clinical trials	1701
6	neurodegenerative diseases	1510
7	cognitive function	1393
8	oxidative stress	1386
9	cognitive decline	1380
10	results suggest	1216
11	long term	1047
12	neuroprotective effects	1045
13	patients with alzheimer disease	974
14	treatment of ad	970
15	results showed	953
16	patients with ad	946
17	treatment of alzheimer disease	871
18	amyloid beta	833
19	protective effects	817
20	ad pathology	810

1 - 20 of 34011 page 1 of 1701 20 items per page

Figure 5. Screenshot of Important Phrases calculated for the query [Alzheimer AND treatment].

Topics. This displays the Medical Subject Headings (MeSH terms) [8] indexed in the set of retrieved articles, ranked by document frequency.

Authors. This lists author names (defined as lastname, firstinitial) ranked by order of document frequency within the set of retrieved articles. No attempt is made here to disambiguate different individuals sharing the same name. However, each author name on each PubMed article has been disambiguated in a separate project, Author-ity [9, 10]. The Author-ity 2009 release has been linked to Anne O'Tate: Each author name on each article in the displayed list is hotlinked ,and clicking on it opens a new tab which displays the predicted individual if that article is present in the Author-ity 2009 dataset. (The most recent Author-ity beta release contains articles through 2018, but has not yet been disseminated in final form publicly.)

Author count*. This button displays the distribution of the number of author names on each article, over the set of retrieved articles. This allows one to choose, for example, only those articles that are sole-authored, or those having many co-authors. Editorials and reviews are often written by senior individuals as sole-authored papers, whereas clinical trials, genomics projects, or experimental physics papers often are conducted in large teams.

Affiliations. This displays chunks of text that are delimited by commas in the Affiliation field of the PubMed record. For example, "University of Illinois" or "Cold Spring Harbor Laboratory" or "USA" are typical chunks that correspond to institutions, cities, or countries. These are ranked in order of document frequency.

Journals. This displays the names of journals in the set of retrieved articles ranked by document frequency.

Year. This displays the distribution of publication dates by year for the set of retrieved articles. The list can be displayed as a histogram and downloaded if desired.

Publication types*. This displays the distribution of publication types as indexed by MEDLINE or assigned by PubMed for the set of retrieved articles. The publication types are grouped into categories: **Problematic** publication types include Retracted Publication, Published Erratum, Retraction of Publication, and Corrected and Republished Article. **Clinical Trials** comprises 10 different types of articles that include the word "trial" (e.g., Pragmatic Clinical Trial) as well as Multicenter Study. **Article Types** list all other articles (e.g., review, case report, practice guideline, etc.) in descending order of document frequency. Finally, **Research Support** lists sources of research support (e.g., NIH, Extramural or Non-U.S. Gov't) in descending order of frequency.

Clustered by Topic. This runs an algorithm which partitions the set of retrieved articles into 12 different topical categories as evenly as possible in terms of the number of articles in each category [1]. This can be useful in surveying the range of topics that are covered in the set, without being unduly influenced by the most common or the most rare.

Important MeSH pairs*. This considers pairs of MeSH terms that co-occur on individual articles within the set of retrieved articles. For those pairs which co-occur on at least four articles within the set, it displays and ranks those pairs according to the odds ratio (i.e., the frequency within the set divided by frequency within MEDLINE as a whole). MeSH pairs often represent either frank relations (e.g., aspirin TREATS headache) or implicit relationships (e.g., a paper discussing the impact of an earthquake on Puerto Rico may be indexed with both "Earthquakes" and "Puerto Rico"). This may help to characterize the kinds of relations studied within a set of retrieved articles, and to drill down more precisely than if one were choosing individual topics.

Mine the Gap!*. This button runs a tool that identifies "gaps" within the set of retrieved articles, i.e., pairs of MeSH terms that do NOT co-occur on any article within the set, even though the individual MeSH terms occur in more than 10% of the set, and the predicted co-occurrence just by chance would be ≥ 10 articles within the set [11]. Finding gaps in a literature may help to predict which new lines of research may be undertaken in the future. Generally, one only finds gaps satisfying those criteria in sets that contain at least several hundred articles. Clicking on the Mine the Gap! button starts a program that identifies and displays gaps, along with some of their features [11]. An optional button carries out Arrowsmith two-node searches [12, 13] on each gap. This allows the user to view and analyze terms that may bridge the gap [13], and calculates the pR ratio for each gap -- this is a measure of the amount of implicit information shared by the pair of MeSH terms [12]. The tool and its underlying model have been described elsewhere in detail [11].

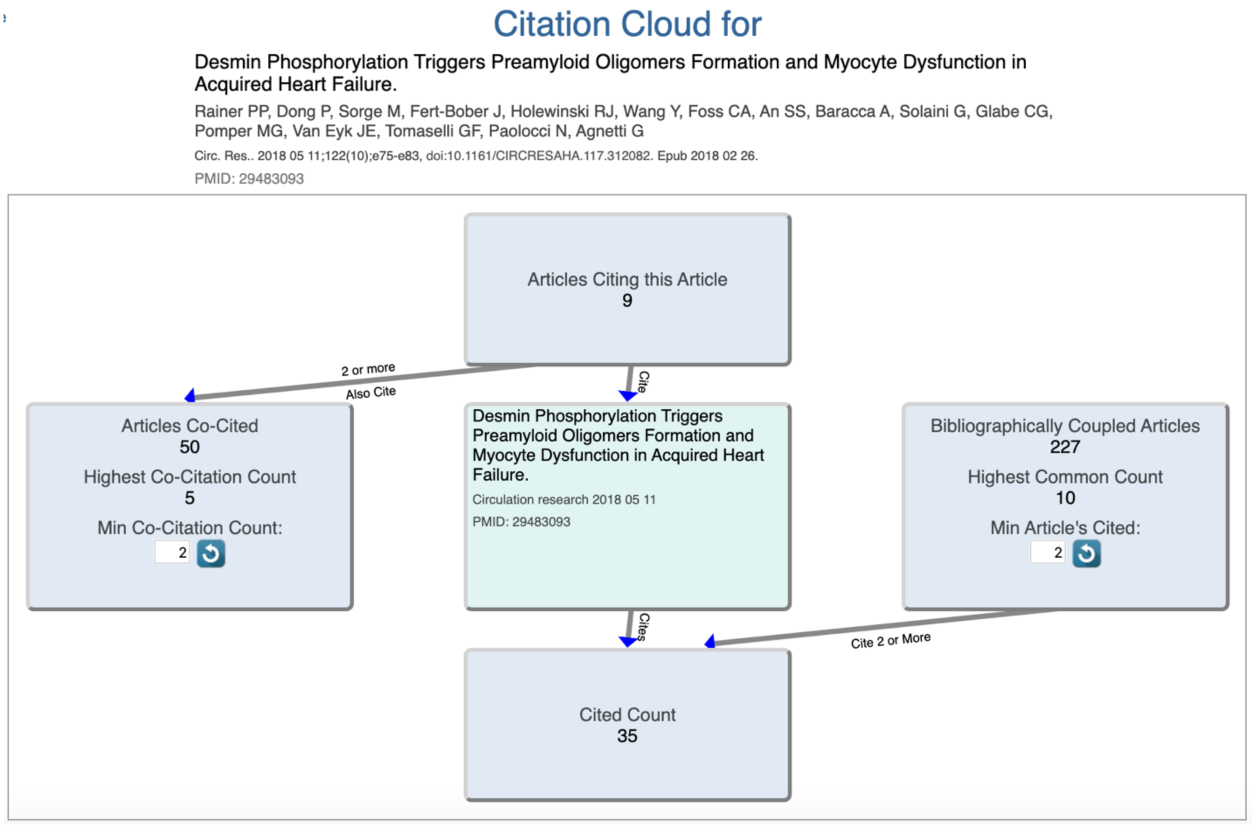
Drill down and expansion of queries

It is an important architectural feature of Anne O'Tate that any new tab displaying a list of articles can be further processed according to any of the 12 buttons on the left side, facilitating easy, progressive, multi-faceted "drill-down" and "slice-and-dice" of the retrieval set as desired. For example, if one would like to find authors who have written recent review articles, one can do the topical query, then click on the Publication Types button, then choose the Review button, then click on the Year button, then choose a given year, and then click on the Authors button to display the list of authors who satisfy all the chosen criteria.

Conversely, any list of displayed articles which is smaller than 50 can be expanded to add new articles that are highly related to multiple articles in the original set. Clicking on the “expand” button at the top of the page runs an algorithm that employs the PubMed related articles algorithm [3] in batch mode[1].

The Citation Cloud*

This tool should greatly enable the study of citations by the scientific community. Clicking on “Citations” next to any displayed article opens a new tab that allows the user to visualize the “citation cloud” around that article: That is, the set of articles cited by it; those which cite it; those which are co-cited with it; and those which are bibliographically coupled to it (Fig. 6). To say that article A and B are co-cited means that A and B are both cited by one or more articles C_i [14]. Co-citation is a measure of similarity not based on textual or topical similarity. Note that the co-citation relationship is not fixed but can vary over time depending on how many newer articles cite both A and B. In contrast, bibliographically coupled (BC) articles cite some of the same articles in their reference lists [15]. This is also a measure of similarity that is not based on textual or topical similarity, and has the distinct advantage that the BC relationship can be calculated for any two articles regardless of when they are published. As well, the BC relationship is stable and will not change over time. Clicking on any box opens a new tab that displays the list of articles in the box. The Citation Cloud tool employs a large dataset of open citations, including iCite [16] as well as other sources. The tool, and some of its typical use cases, are described in detail in a separate publication [17].



Citation Cloud for

Desmin Phosphorylation Triggers Preamyloid Oligomers Formation and Myocyte Dysfunction in Acquired Heart Failure.

Rainer PP, Dong P, Sorge M, Fert-Bober J, Holewinski RJ, Wang Y, Foss CA, An SS, Baracca A, Solaini G, Glabe CG, Pomper MG, Van Eyk JE, Tomaselli GF, Paolocci N, Agnetti G

Circ. Res.. 2018 05 11;122(10);e75-e83, doi:10.1161/CIRCRESAHA.117.312082. Epub 2018 02 26.

PMID: 29483093

Articles Citing this Article
9
2 or more
Also Cite
Cite
Articles Co-Cited
50
Highest Co-Citation Count
5
Min Co-Citation Count:
2
Desmin Phosphorylation Triggers Preamyloid Oligomers Formation and Myocyte Dysfunction in Acquired Heart Failure.
Circulation research 2018 05 11
PMID: 29483093
Bibliographically Coupled Articles
227
Highest Common Count
10
Min Article's Cited:
2
Cites
Cite 2 or More
Cited Count
35

Figure 6. Screenshot of the citation cloud for the target article: Rainer et al., Desmin Phosphorylation Triggers Preamyloid Oligomers Formation and Myocyte Dysfunction in Acquired Heart Failure. Circ Res. 2018 May 11;122(10):e75-e83. See [17] for a detailed description of the Citation Cloud tool.

Finally, note that a message on the starting page says that Anne O'Tate processes the most recent 25,000 articles from any query (fig. 1). This is to avoid slow-downs that may be caused from users entering huge, poorly formed, or malicious queries. However, the situation is a bit more nuanced: The initial query returns an unlimited number of articles from PubMed. Although most of the buttons are limited to processing the most recent 25,000 articles in the list of displayed articles, the Year, Publication Types, and Citation Cloud functions are all unlimited.

Discussion

A plethora of tools have been developed to assist in searching the biomedical literature (reviewed in [18, 19]). Few have been maintained as a free, public service continuously to the present. Although some of these tools do provide faceted summarization (e.g., listing author and journal names for a set of retrieved article) [2], no other tool offers the unique architecture of Anne O'Tate to drill-down progressively according to any of a dozen dimensions, or to expand up again. Nor do any other search web servers serve as open platforms for adding new tools that greatly enhance the ability of users to analyze literature in a sophisticated manner (e.g., Mine the Gap! which identifies research gaps in the set of retrieved articles [11], and the Citation Cloud which displays the entire local network of citations surrounding any given article [17]). We envision Anne O'Tate as a value-added layer on top of the PubMed search engine. We can offer tools that may be too specialized for the government to host, and can add or change items in a more nimble fashion.

Use cases for Anne O'Tate arise on a daily basis. For example, to find suggested reviewers for a given manuscript or proposal, one can enter a topical query (e.g., keywords taken from the title) and examine the list of authors who have published PubMed articles on that topic. Similarly, to help decide which journal is best for submitting your own manuscript, enter a topical query and examine the list of journals which have published similar work. The Important Words and Important Phrases may help to suggest new keywords for refining queries during high recall retrieval projects such as accumulating evidence for a systematic review. Using the Publication one can identify relationships that are specifically studied in the set of retrieved articles. Types button allows a user to track articles in a given set that have been retracted, and the Citation Cloud to trace articles that have subsequently cited them anyway [20, 21]. Studies of scientific innovation may be interested to compare the characteristics of articles that have been single-authored vs. published in collaborative teams [22]. Using the Year button, one can immediately see the trend of growth over time of a given topic. Using the Important MeSH Pairs, Conversely, to identify relationships that have surprisingly NOT been studied at all [11], one can use the Mine the Gap! button.

We expect that Anne O'Tate will continue to evolve over time. For example, the new updated Author-ity author name disambiguation dataset (through 2018) has not yet been officially released; when it has, we expect to link it to the articles retrieved through Anne O'Tate so that any author on any article can be disambiguated. Also, we are in the process of developing a probabilistic automated classifier to classify each article in PubMed according to 46 different predicted publication types – this augments the official indexing which does not include the most recent articles nor articles which are not in MEDLINE journals. We plan to utilize the predictions of that model to improve the current Publication Types button, by displaying additional articles not indexed by MEDLINE but predicted with high confidence to have the characteristics of one or more publication types. Future changes will be driven by our own ongoing research, and by suggestions and feedback from the biomedical and informatics communities.

Materials and Methods

Many of the methods were discussed in detail in the original publications [1, 11, 17]. As one of the reviewers of Anne O'Tate pointed out [2], in the past Anne O'Tate tended to be slow in processing large queries. We therefore devoted considerable effort to optimizing the data retrieval techniques used for passing information between entities Anne O'Tate and PubMed. PubMed only allows a certain number of articles to be fetched in a single request, even when using an API key (to validate us as a known user) and functions optimally on numbers less than that maximum.

PubMed also only allows a certain number of PMIDs (as a list) to be sent within a single query dictated also by maximum HTTP request size. Internally within our server, the SQL database that holds an internal parsed representation of all PubMed records also has a practical limit within queries. To overcome these limits, batch processing was used when necessary, operating within each component's limits within each batch yet still operating on the complete data set in whole, so that queries were processed smoothly by all components of the system. On two of the new buttons –- Year and Publication Types -- the user-facing client-side uses AJAX requests which allow for the long running processes to be more elegantly and efficiently returned to the end user. The end user in these cases sees immediate results to the screen in real time, as the data are pulled from PubMed and/or collected locally. The previous version used timed browser refreshes, which check with the server at each refresh to see if the server-side function has completed. These refreshes sometimes appeared to 'hang' when the server-side processes died inadvertently, leaving the client to refresh infinitely. The errors causing server-side processes to die were fixed, resolving the 'hanging' problem.

Some other changes made to the original implementation of Anne O'Tate were the elimination of the system-level kill commands being issued on the background Perl daemon processes, and more granular batch processing of the data pipeline (as mentioned above). These made user jobs more stable, faster and less likely to hang and resulted in fewer server crashes. The Architecture supporting Anne O'Tate is comprised of LINUX Ubuntu Server 18.04 LTS, Perl 5 version 26, Python 2.7 and 3.6 and MySql 5.7.

Acknowledgments

We thank the users of Anne O'Tate who have given feedback over the years, and who have emailed us whenever the site went down or a bug was detected!

Funding

This research was supported by NIH Grants R01LM010817 and P01AG039347. The study sponsor had no role in study design; in the collection, analysis and interpretation data; in the writing of the report; or in the decision to submit the paper for publication.

Competing Interests

The authors declare no competing interests.

References

1. Smalheiser NR, Zhou W, Torvik VI. Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results. J Biomed Discov Collab. 2008 Feb 15;3:2. doi: 10.1186/1747-5333-3-2.
2. Engwall KD. Anne O'Tate. Journal of the Medical Library Association: JMLA. 2017 Apr;105(2):200.
3. Lin J, Wilbur WJ. PubMed related articles: a probabilistic topic-based model for content similarity. BMC Bioinformatics. 2007 Oct 30;8:423.
4. Smalheiser NR, Zhou W, Torvik VI. Distribution of "Characteristic" Terms in MEDLINE Literatures. Information. 2011 Jun;2(2):266-76.

5. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Stud Health Technol Inform. 2001;84(Pt 1):216-20.
6. Torvik VI, Smalheiser NR, Weeber M. A simple Perl tokenizer and stemmer for biomedical text. Unpublished technical report. 2007 http://arrowsmith.psych.uic.edu/arrowsmith_uic/tutorial/tokenizer_2007.pdf, accessed May 13, 2020.
7. El-Kishky A, Song Y, Wang C, Voss CR, Han J. Scalable Topical Phrase Mining from Text Corpora. Proceedings of the VLDB Endowment. 2014;8(3).
8. Lipscomb CE. Medical subject headings (MeSH). Bulletin of the Medical Library Association. 2000 Jul;88(3):265.
9. Torvik VI, Weeber M, Swanson DR, Smalheiser NR. A probabilistic similarity metric for Medline records: a model for author name disambiguation. JASIST 2005; 56(2): 140-158.
10. Torvik VI, Smalheiser NR. Author name disambiguation in MEDLINE. ACM Transactions on Knowledge Discovery from Data 2009; 3(3):11.
11. Peng Y, Bonifield G, Smalheiser NR. Gaps within the Biomedical Literature: Initial Characterization and Assessment of Strategies for Discovery. Front Res Metr Anal. 2017 May;2. pii: 3. doi: 10.3389/frma.2017.00003.
12. Torvik VI, Smalheiser NR. A quantitative model for linking two disparate sets of articles in MEDLINE. Bioinformatics. 2007 Jul 1;23(13):1658-65.
13. Smalheiser NR, Torvik VI, Zhou W. Arrowsmith two-node search interface: a tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. Comput Methods Programs Biomed. 2009 May;94(2):190-7. doi:10.1016/j.cmpb.2008.12.006.
14. Small H. Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for information Science. 1973 Jul;24(4):265-9.
15. Kessler MM. Bibliographic coupling between scientific papers. American documentation. 1963 Jan;14(1):10-25.
16. Hutchins BI, Baker KL, Davis MT, Diwersy MA, Haque E, Harriman RM, Hoppe TA, Leicht SA, Meyer P, Santangelo GM. The NIH Open Citation Collection: A public access, broad coverage resource. PLoS Biol. 2019 Oct 10;17(10):e3000385. doi:10.1371/journal.pbio.3000385.
17. Smalheiser NR, Schneider J, Torvik VI, Fragnito DP, Tirk EE. The Citation Cloud of a Biomedical Article: Enabling Citation Analysis, submitted to bioRxiv.
18. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. Database (Oxford). 2011 Jan 18;2011:baq036. doi:10.1093/database/baq036.
19. Wildgaard L.E. and Lund H. Advancing PubMed? A comparison of third-party PubMed/Medline tools. *Library Hi Tech* 2016 34(4):669-684.
20. Steen RG. Retractions in the medical literature: how many patients are put at risk by flawed research? J Med Ethics. 2011 Nov;37(11):688-92. doi:10.1136/jme.2011.043133.
21. van der Vet, P.E., Nijveen, H. Propagation of errors in citation networks: a study involving the entire citation network of a widely cited paper published in, and later retracted from, the journal Nature. *Res Integr Peer Rev* **1,** 3 (2016).
22. Leahey E. From sole investigator to team scientist: Trends in the practice and study of research collaboration. Annual review of sociology. 2016 Jul 30;42:81-100.