eLIFE
elifesciences.org

FEATURE ARTICLE

CUTTING EDGE

Towards PubMed 2.0

Staff from the National Center for Biotechnology Information in the US describe recent improvements to the PubMed search engine and outline plans for the future, including a new experimental site called PubMed Labs.

NICOLAS FIORINI, DAVID J LIPMAN AND ZHIYONG LU

Introduction

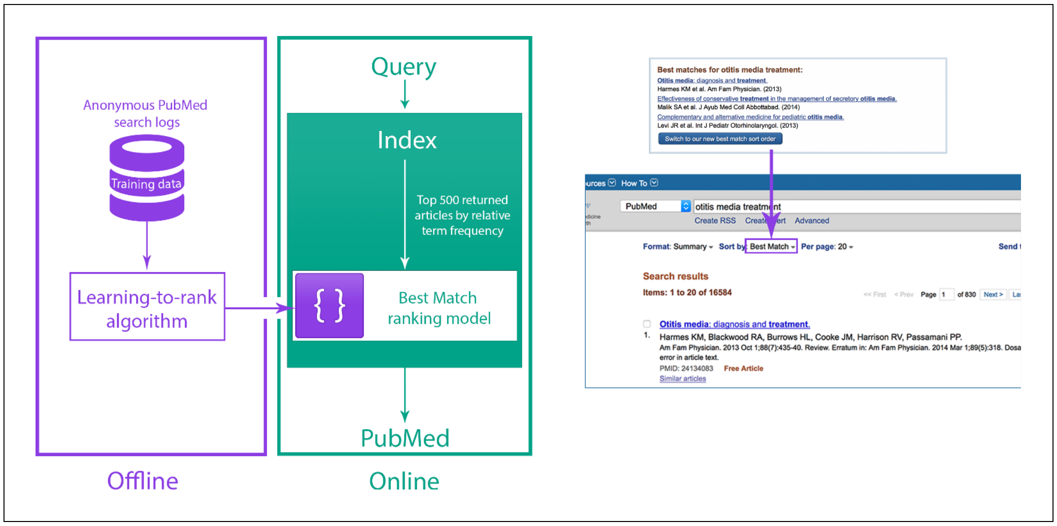
PubMed is a search engine for the biomedical literature that provides access to MEDLINE, a database that contains bibliographic information on more than 27 million articles from more than 7,000 journals, including full text for some 4 million of these articles (***Sayers et al., 2011***). On an average working day approximately 2.5 million users from around the world access PubMed to perform about 3 million searches and 9 million page views. Here, as PubMed celebrates its 20th anniversary, and as the biomedical literature continues to grow, we discuss how PubMed is changing to meet the needs and expectations of its users, and to take advantage of advances in web and information technologies.

Given a query, PubMed (http://www.ncbi.nlm.nih.gov/pubmed) returns the most recent articles first by default. While this reverse chronological sort helps users to find the latest articles on a particular topic or by a specific author, most PubMed users are accustomed to search engines that sort results by relevance. Indeed, over 80% of PubMed users only browse and click results on the first page, which is very similar to the behavior of most web users (***Islamaj Dogan et al., 2009***), even though more useful (and sometimes still recent) papers may be returned on the second page and beyond. Moreover, the need for relevance search becomes more critical in light of the growing use of small-screen devices – today over 20% of total PubMed usage happens on mobile platforms – where users tend to browse results even less.

To accommodate changing user needs, a number of new features have been added to PubMed in recent years, such as faceted search, query auto-suggest (***Lu et al., 2009a***) and author name disambiguation (***Liu et al., 2014***). In this article we describe how we intend to improve PubMed further (in terms of the search quality and also the user interface), and we present our path towards a next-generation system (PubMed 2.0), including a new experimental site called PubMed Labs.

New relevance search algorithm

A new relevance algorithm was recently developed and deployed as one of the sort options in PubMed. It aims to provide the most relevant results and the best quality information within a fraction of a second, in response to a variety of different user information needs. This algorithm is based on a machine learning method called 'learning-to-rank' (***Burges et al., 2011***; ***Liu, 2009***; ***Figure 1***). The initial input to the learning-to-rank algorithm is the list of results returned by a classic term frequency algorithm (which ranks search results based on the relative frequency with which terms from the search query appear in the title and/or abstract of a record; ***Lu et al., 2009b***; ***Robertson et al., 1996***). We chose learning-to-rank because, in addition to its proven performance, it is capable of integrating a wide variety of different signals from the text record (including term frequencies) in order to return improved ranking results.


Anonymous PubMed search logs
Training data
Learning-to-rank algorithm
Offline
Query
Index
Top 500 returned articles by relative term frequency
{ }
Best Match ranking model
PubMed
Online

Best matches for otitis media treatment:
Otitis media: diagnosis and treatment.
Harmes KM et al. Am Fam Physician. (2013)
Effectiveness of conservative treatment in the management of secretory otitis media.
Malik SA et al. J Ayub Med Coll Abbottabad. (2014)
Complementary and alternative medicine for pediatric otitis media.
Levi JR et al. Int J Pediatr Otorhinolaryngol. (2013)
Switch to our new best match sort order

urces How To
PubMed otitis media treatment
Create RSS Create alert Advanced
Format: Summary Sort by: Best Match Per page: 20 Send t
Search results
Items: 1 to 20 of 16584
<< First < Prev Page 1 of 830 Next > La
□ Otitis media: diagnosis and treatment.
1. Harmes KM, Blackwood RA, Burrows HL, Cooke JM, Harrison RV, Passamani PP.
Am Fam Physician. 2013 Oct 1;88(7):435-40. Review. Erratum in: Am Fam Physician. 2014 Mar 1;89(5):318. Dosa error in article text.
PMID: 24134083 Free Article
Similar articles

Figure 1. Relevance search in PubMed. (Left) System overview for the Best Match sort option in PubMed. (Right) Users are encouraged to try the Best Match sort option when PubMed detects a query for which the Best Match sort option could yield better results.

In our ranking model, we combine over 150 distinct signals to find the best results: many of these signals are computed from the number of matches between the search terms and the PubMed record, while others are either specific to a record (for example, publication type, publication date and number of abstract views within the last year) or a search (for example, the length of the query). The new ranking model was trained and tested on benchmarking data obtained from aggregated search logs.

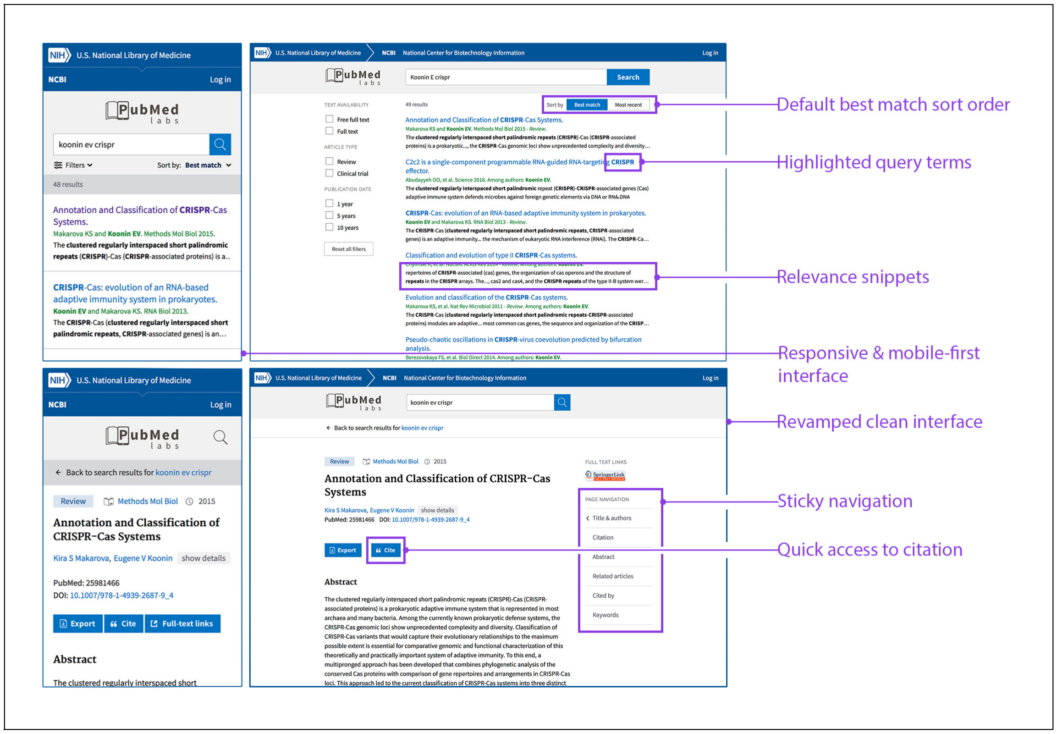
The new ranking algorithm has been available since April 2017 to those who select the 'Best Match' sort option in PubMed, and users who select this option are 17.4% more likely to click on a link than those who select the default date-order algorithm. This figure is derived by comparing the fraction of queries with at least one article click on the first results page for the new algorithm with the same fraction for the default algorithm.

PubMed labs and its new interface

In parallel with these developments we are launching a new experimental site called PubMed Labs (http://www.ncbi.nlm.nih.gov/labs/pubmed) that will allow us to continue our efforts to improve search quality while also working to improve PubMed's user interface: in particular, the interface for PubMed Labs (***Figure 2***) has been designed to provide an optimized user experience on small-screen devices (such as mobiles and tablets). We are taking a 'minimum viable product' approach in building PubMed Labs by launching it with a small number of core features and making iterative improvements and enhancements based on public testing and feedback. For example, PubMed Labs includes only a small subset of PubMed facets (selected on the basis of past usage in PubMed); decisions about the final set of facets will be based on user input and real-world usage. Other new features that we plan to introduce and analyze on PubMed Labs range from simple font color variations to complex algorithmic adjustments. Based on quantitative analysis of usage patterns, as well as direct user input and feedback, we will iteratively improve and enhance the minimum viable product.

Outlook

As in the past, R&D will be key to the success of PubMed 2.0. However, while there has been extensive research into general web searches (e.g. ***Serdyukov et al., 2013***), there has been less published research on usage patterns for online biomedical information resources. However, we do know that there are important differences between the two (***Islamaj Dogan et al., 2009***; ***Wilbur et al., 2006***; ***Herskovic et al., 2007***; ***Hersh and Voorhees, 2009***). For example, PubMed users are more persistent than general web users when searching. Moreover, author names appear in over 35% of user queries in PubMed, and these names are often ambiguous because different authors share the same name and a given author can publish under multiple names (***Névéol et al., 2011***;



NIH U.S. National Library of Medicine | NCBI | Log in | PubMed labs

koonin ev crispr | Filters | Sort by: Best match | 48 results

Annotation and Classification of **CRISPR**-Cas Systems.
Makarova KS and **Koonin EV**. Methods Mol Biol 2015.
The **clustered regularly interspaced short palindromic repeats** (**CRISPR**)-Cas (**CRISPR**-associated proteins) is a..

CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes.
Koonin EV and Makarova KS. RNA Biol 2013.
The **CRISPR**-Cas (**clustered regularly interspaced short palindromic repeats**, **CRISPR**-associated genes) is an…

NIH U.S. National Library of Medicine | NCBI National Center for Biotechnology Information | Log in | PubMed labs

Koonin E crispr | Search

TEXT AVAILABILITY: □ Free full text □ Full text
ARTICLE TYPE: □ Review □ Clinical trial
PUBLICATION DATE: □ 1 year □ 5 years □ 10 years
Reset all filters

49 results | Sort by: Best match | Most recent

Annotation and Classification of **CRISPR**-Cas Systems.
Makarova KS and **Koonin EV**. Methods Mol Biol 2015 - *Review*.
The **clustered regularly interspaced short palindromic repeats** (**CRISPR**)-Cas (**CRISPR**-associated proteins) is a prokaryotic..., the **CRISPR**-Cas genomic loci show unprecedented complexity and diversity…

C2c2 is a single-component programmable RNA-guided RNA-targeting **CRISPR** effector.
Abudayyeh OO, et al. Science 2016. Among authors: **Koonin EV**.
The **clustered regularly interspaced short palindromic** repeat (**CRISPR**)-**CRISPR**-associated genes (Cas) adaptive immune system defends microbes against foreign genetic elements via DNA or RNA-DNA

CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes.
Koonin EV and Makarova KS. RNA Biol 2013 - *Review*.
The **CRISPR**-Cas (**clustered regularly interspaced short palindromic repeats**, **CRISPR**-associated genes) is an adaptive immunity... the mechanism of eukaryotic RNA interference (RNAi). The **CRISPR**-Ca…

Classification and evolution of type II **CRISPR**-Cas systems.
Chylinski K, et al. Nucleic Acids Res 2014 - Review. Among authors: Koonin EV.
repertoires of **CRISPR**-associated (cas) genes, the organization of cas operons and the structure of **repeats** in the **CRISPR** arrays. The..., cas2 and cas4, and the **CRISPR repeats** of the type II-B system wer…

Evolution and classification of the **CRISPR**-Cas systems.
Makarova KS, et al. Nat Rev Microbiol 2011 - *Review*. Among authors: **Koonin EV**.
The **CRISPR**-Cas (**clustered regularly interspaced short palindromic repeats**-**CRISPR**-associated proteins) modules are adaptive... most common cas genes, the sequence and organization of the **CRISP**…

Pseudo-chaotic oscillations in **CRISPR**-virus coevolution predicted by bifurcation analysis.
Berezovskaya FS, et al. Biol Direct 2014. Among authors: **Koonin EV**.

NIH U.S. National Library of Medicine | NCBI | Log in | PubMed labs

Back to search results for koonin ev crispr

Review | Methods Mol Biol | 2015

Annotation and Classification of CRISPR-Cas Systems

Kira S Makarova, Eugene V Koonin | show details

PubMed: 25981466
DOI: 10.1007/978-1-4939-2687-9_4

Export | Cite | Full-text links

Abstract

The clustered regularly interspaced short

NIH U.S. National Library of Medicine | NCBI National Center for Biotechnology Information | Log in | PubMed labs

koonin ev crispr

Back to search results for koonin ev crispr

Review | Methods Mol Biol | 2015

Annotation and Classification of CRISPR-Cas Systems

Kira S Makarova, Eugene V Koonin | show details
PubMed: 25981466 DOI: 10.1007/978-1-4939-2687-9_4

Export | Cite

Abstract

The clustered regularly interspaced short palindromic repeats (CRISPR)-Cas (CRISPR-associated proteins) is a prokaryotic adaptive immune system that is represented in most archaea and many bacteria. Among the currently known prokaryotic defense systems, the CRISPR-Cas genomic loci show unprecedented complexity and diversity. Classification of CRISPR-Cas variants that would capture their evolutionary relationships to the maximum possible extent is essential for comparative genomic and functional characterization of this theoretically and practically important system of adaptive immunity. To this end, a multipronged approach has been developed that combines phylogenetic analysis of the conserved Cas proteins with comparison of gene repertoires and arrangements in CRISPR-Cas loci. This approach led to the current classification of CRISPR-Cas systems into three distinct

FULL TEXT LINKS: SpringerLink

PAGE NAVIGATION: Title & authors | Citation | Abstract | Related articles | Cited by | Keywords

Figure 2. PubMed Labs: interface and features. The PubMed Labs interface on a small-screen device (left) and a desktop device (right). The default sort option in PubMed Labs is the Best Match sort option, but it is also possible to sort by date order. Each search result in PubMed Labs has search terms highlighted in the title and/or newly generated snippets, which are automatically extracted from the sentence(s) in the abstract with matching search terms. Other differences in the new interface are also highlighted.

Islamaj Dogan et al., 2009). This problem is made worse because authors in PubMed are traditionally represented by just their last name and first initial (e.g. Lu Z). To deal with this challenge we recently developed an automated method for author name disambiguation based on machine learning (***Liu et al., 2014***) and are also planning to make greater use of other related resources such as ORCIDs, regardless whether a publication has an explicit ORCID associated with it or not (most PubMed articles do not).

The rapid growth of PubMed Central (https://www.ncbi.nlm.nih.gov/pmc/) – which stores the full text of articles and is growing by more than 1000 articles per day – also presents opportunities for improving PubMed. In addition to making more content available to users, the availability of the full text (including the text in figures and tables) could lead to improved search if we can develop automated methods for coping with the increased complexity associated with full texts (***Cohen et al., 2010***). We are also studying the feasibility of using deep-learning technology (***Mohan et al., 2017***; ***Severyn and Moschitti, 2015***; ***Mikolov et al., 2013***) in large-scale text mining applications.

Looking into the future, we also envision PubMed serving as the central entry point for access to a range of different NCBI literature resources. To this end, we have created a new data structure, known internally as PubOne, to merge documents of different genres (such as PubMed abstracts, full texts from PubMed Central, book chapters and so on) into a single record in order to enable all-in-one access to the biomedical literature. And as we move forward from the minimum viable product, we will continue to assess and incorporate existing features such as MyNCBI into our new system. We also plan to explore if PubMed could use the outputs of various initiatives in open data and alternative metrics to improve search.

The development of PubMed has benefitted greatly from open-source platforms and technologies (such as the learning-to-rank method we used for relevance search) and from user feedback. In return for the former we would like to make our tools and techniques available to the entire scientific community via source code or web APIs. With regards to user feedback, we encourage our users and the biomedical research community to continue to use PubMed, to experiment with PubMed Labs, and to help us jointly create PubMed 2.0.

Acknowledgements

We would like to thank Vadim Miller, Grisha Starchenko, Evgeny Kireev, Maxim Osipov and Rafis Ismagilov for their work on the new relevance search; Rostyslav Bryzgunov, Ievgeniia Radetska, Asta Gindulyte and Martin Latterner for their work on PubMed Labs; Won Kim, Kathi Canese and Sharmin Hussain for their help with online experiments and log analysis; Jim Ostell, John Wilbur, Udi Manber and Sunil Mohan for helpful discussions; and the many NCBI and NLM Library Operations staff working on and contributing to PubMed.

Nicolas Fiorini is at the National Center for Biotechnology Information, US National Library of Medicine, National Institutes of Health, Bethesda, United States
nicolas.fiorini@nih.gov
http://orcid.org/0000-0002-9260-1326

David J Lipman is at the National Center for Biotechnology Information, US National Library of Medicine, National Institutes of Health, Bethesda, United States

Zhiyong Lu is at the National Center for Biotechnology Information, US National Library of Medicine, National Institutes of Health, Bethesda, United States
zhiyong.lu@nih.gov
http://orcid.org/0000-0001-9998-916X

Competing interests: The authors declare that no competing interests exist.

Published 30 October 2017

References

Burges C, Svore K, Bennett P, Pastusiak A, Wu Q, Chapelle O, Chang Y, Liu TY. 2011. Learning to rank using an ensemble of lambda-gradient models. *Proceedings of the Learning to Rank Challenge* **14**:25–35.

Cohen KB, Johnson HL, Verspoor K, Roeder C, Hunter LE. 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics* **11**:492. doi: 10.1186/1471-2105-11-492, PMID: 20920264

Hersh W, Voorhees E. 2009. TREC genomics special issue overview. *Information Retrieval* **12**:1–15. doi: 10.1007/s10791-008-9076-6

Herskovic JR, Tanaka LY, Hersh W, Bernstam EV. 2007. A day in the life of PubMed: analysis of a typical day's query log. *Journal of the American Medical Informatics Association* **14**:212–220. doi: 10.1197/jamia.M2191, PMID: 17213501

Islamaj Dogan R, Murray GC, Névéol A, Lu Z. 2009. Understanding PubMed user search behavior through log analysis. *Database: The Journal of Biological Databases and Curation* **2009**:bap018. doi: 10.1093/database/bap018, PMID: 20157491

Liu TY. 2009. *Learning to Rank for Information Retrieval*. Heidelberg: Springer-Verlag.

Liu W, Islamaj Doğan R, Kim S, Comeau DC, Kim W, Yeganova L, Lu Z, Wilbur WJ. 2014. Author name disambiguation for PubMed. *Journal of the Association for Information Science and Technology* **65**:765–781. doi: 10.1002/asi.23063

Lu Z, Wilbur WJ, McEntyre JR, Iskhakov A, Szilagyi L. 2009a. Finding query suggestions for PubMed. *AMIA Annual Symposium Proceedings* **2009**:396–400. PMID: 20351887

Lu Z, Kim W, Wilbur WJ. 2009b. Evaluating relevance ranking strategies for MEDLINE retrieval. *Journal of the American Medical Informatics Association* **16**:32–36. doi: 10.1197/jamia.M2935, PMID: 18952932

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* **26**:3111–3119.

Mohan S, Fiorini N, Kim S, Lu Z. , 2017. Deep learning for biomedical information retrieval: learning textual relevance from limited click logs. *Proceedings of the 2017 ACL Workshop on Biomedical Natural Language Processing*. doi: 10.18653/v1/W17-2328

Névéol A, Islamaj Doğan R, Lu Z. 2011. Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. *Journal of Biomedical Informatics* **44**:310–318. doi: 10.1016/j.jbi.2010.11.001, PMID: 21094696

Robertson SE, Walker S, Jones S, Hancock-Beaulieu MM, Gatford M. 1996. Okapi at TREC-3. *Proceedings of the Third NIST Text Retrieval Conference*.

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, et al. 2011. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **39**:D38–D51. doi: 10.1093/nar/gkq1172, PMID: 21097890

Serdyukov P, Braslavski P, Kuznetsov SO, Kamps J, Rüger S, Agichtein E, Segalovich I, Yilmaz E. 2013. *Advances in Information Retrieval: Proceedings of the 35th European Conference on IR Research*. Heidelberg: Springer-Verlag.

Severyn A, Moschitti A. 2015. Learning to rank short text pairs with convolutional deep neural networks. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* 373–382. doi: 10.1145/2766462.2767738

Wilbur WJ, Kim W, Xie N. 2006. Spelling correction in the PubMed search engine. *Information Retrieval* **9**:543–564. doi: 10.1007/s10791-006-9002-8, PMID: 18080004