Conceptually Identifying Academic Papers for Better Query Results

Sameehan Sriharsha [SE23UARI069]

Abstract

Old fashioned and context-blind algorithms used to rank the relevance of documents have glaring flaws; the most obvious being the lack of any semantic reasoning. Research work depends heavily, ever more so in an age of mass data, on the ability of researchers to quickly identify and extract exactly the study or data they require. It is, therefore, crucial that such bottlenecks in the initial phase of research are remedied. This report examines a few novel search and retrieval algorithms that are in use today. Furthermore, in the case of text extraction, theorizes a few methods that may improve accuracy.

Body

There are two major classes of algorithms in use today for searching. The first being regular term comparison, which includes but is not limited to `*BM25*` an improvement over TF-IDF. Another class of algorithms incorporates the use of LLMs in order to facilitate accurate, contextually and semantically appropriate information.

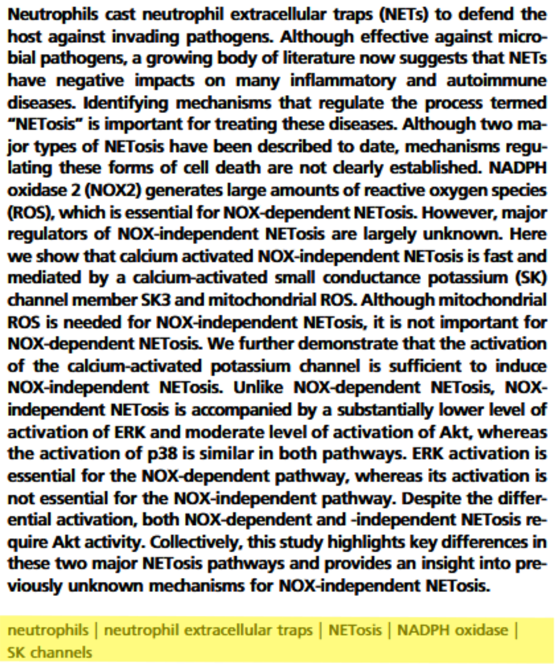
Both classes of algorithm have their benefits, the former being relatively cheap to compute—considering it does not utilize any form of large-scale training. The downside, however, is that the inaccuracies build up the less "involved" a query is with the contents contained within the corpus.

A query is *involved* if it has at least a few technical terms that could be located within a document. Example:

- "what is that trinitrotoluene" is a bad, uninvolved query because 75% of the words will be removed in stop-word processing. The last word however, while technical, might be too broad
- "quality explosive simple" while not technical can provide many more results

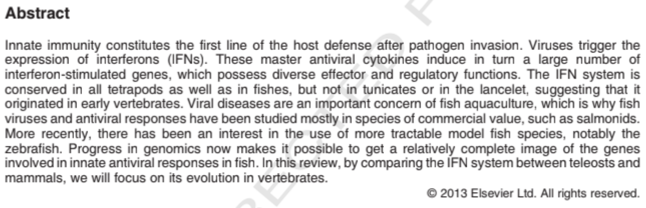
Preprocessing Steps

PDF text preprocessing is difficult when there are no strict standards on the logical flow of information in the document. Some articles provide an abstract and the keywords section in the very beginning of the paper. Others keep it on the final page.



Neutrophils cast neutrophil extracellular traps (NETs) to defend the host against invading pathogens. Although effective against microbial pathogens, a growing body of literature now suggests that NETs have negative impacts on many inflammatory and autoimmune diseases. Identifying mechanisms that regulate the process termed "NETosis" is important for treating these diseases. Although two major types of NETosis have been described to date, mechanisms regulating these forms of cell death are not clearly established. NADPH oxidase 2 (NOX2) generates large amounts of reactive oxygen species (ROS), which is essential for NOX-dependent NETosis. However, major regulators of NOX-independent NETosis are largely unknown. Here we show that calcium activated NOX-independent NETosis is fast and mediated by a calcium-activated small conductance potassium (SK) channel member SK3 and mitochondrial ROS. Although mitochondrial ROS is needed for NOX-independent NETosis, it is not important for NOX-dependent NETosis. We further demonstrate that the activation of the calcium-activated potassium channel is sufficient to induce NOX-independent NETosis. Unlike NOX-dependent NETosis, NOX-independent NETosis is accompanied by a substantially lower level of activation of ERK and moderate level of activation of Akt, whereas the activation of p38 is similar in both pathways. ERK activation is essential for the NOX-dependent pathway, whereas its activation is not essential for the NOX-independent pathway. Despite the differential activation, both NOX-dependent and -independent NETosis require Akt activity. Collectively, this study highlights key differences in these two major NETosis pathways and provides an insight into previously unknown mechanisms for NOX-independent NETosis.

neutrophils | neutrophil extracellular traps | NETosis | NADPH oxidase | SK channels



Abstract

Innate immunity constitutes the first line of the host defense after pathogen invasion. Viruses trigger the expression of interferons (IFNs). These master antiviral cytokines induce in turn a large number of interferon-stimulated genes, which possess diverse effector and regulatory functions. The IFN system is conserved in all tetrapods as well as in fishes, but not in tunicates or in the lancelet, suggesting that it originated in early vertebrates. Viral diseases are an important concern of fish aquaculture, which is why fish viruses and antiviral responses have been studied mostly in species of commercial value, such as salmonids. More recently, there has been an interest in the use of more tractable model fish species, notably the zebrafish. Progress in genomics now makes it possible to get a relatively complete image of the genes involved in innate antiviral responses in fish. In this review, by comparing the IFN system between teleosts and mammals, we will focus on its evolution in vertebrates.

© 2013 Elsevier Ltd. All rights reserved.

In order to minimize the effects of stray, uncaught artifacts there was a brute attempt at simply reducing all non-important words/text into the most basic readable form. This involved *lowercasing*, *symbol-deletion*, *space-shrinking* and *number-removal*.

BM25

BM25 is a popular ranking function that is used to determine which corpus is most relevant. It is a modification of TF-IDF, an older method that did not take into account document length

A) Uncleaned Input The test was conducted by running BM25 on a selection of 5 documents that were not cleaned before being fed into the function.

B) Cleaned Input In contrast to the previous test, the input was cleaned by eliminating stop-words as well as other important features such as: emails, symbols and misspelled words.

Both instances were tested by comparing the results to a training set of query sentences that were known beforehand. Randomization training by generating a new query when required would not have been an effective method to discriminate performance. *It is to be noted that the query strings were cleaned before being passed to the function.*

[RAW]	Query-A	Query-B	Query-C	Query-D	Query-E	Query-F
Document-A	4.76E-03	2.21E-02	2.12E-02	3.06E-01	3.06E-01	1.13E-01
Document-B	4.76E-03	1.51E-02	2.94E-02	2.44E-01	2.44E-01	7.70E-02
Document-C	4.76E-03	2.23E-02	2.68E-02	2.43E-01	2.43E-01	6.93E-01
Document-D	9.86E-01	9.40E-01	9.23E-01	2.07E-01	2.07E-01	1.17E-01

BM25 uncleaned-corpus statistics

[CLEANED]	Query-A	Query-B	Query-C	Query-D	Query-E	Query-F
Document-A	2.96E-03	1.34E-02	3.62E-03	2.45E-01	2.45E-01	7.76E-02
Document-B	2.96E-03	1.89E-02	3.25E-03	2.43E-01	2.43E-01	7.03E-01
Document-C	2.96E-03	1.87E-02	2.68E-03	2.98E-01	2.98E-01	1.08E-01
Document-D	9.91E-01	9.49E-01	9.90E-01	2.14E-01	2.14E-01	1.11E-01

BM25 cleaned-corpus statistics

Embeddings

Word2Vec is an outdated but popular embedding method to convert words to mathematically interactable vectors by way of the assumption: “words are defined by their neighbors”. This idea can be taken further by utilizing embeddings to effectively find a unique representation of a document in the form of a vector.

These vectors can then be utilized alongside a similarly converted version of a search query to determine what the conceptually closest paper is.

Embeddings, however, require extensive training in order to be effective in their distinguishment of words. Additionally tacking on a vector to each document in the database might become expensive depending on how big it is.

Note: At attempt was made to utilize GloVe-vectors as the embedding source. But this proved untenable as the memory use immediately exceeded the maximum capacity assigned by Google’s compute provider

Semantic Sensitive Retrieval using Knowledge Graphs

Newer algorithms that make use of models like BERT in order to capture subtle semantic information are more capable of distinguishing between documents to accurately determine or rank what document is best.

Semantic information is often captured with the use of knowledge graphs. Although, it is to be noted that these models require text remain uncleaned so that they can effectively segregate between important and useless words.

Knowledge graphs improve semantic retrieval in ways that can be generalized to other industry applications.

SciBERT

A fine-tuned variant of BERT for scientific/academic work was used in this test. Similar to BM25, there were two separate pipelines. One for raw data and the other for simple preprocessed pdf text.

The entire document was fed through BERT’s encoder to produce the document embeddings of each file within the corpus. This process was repeated for the query strings as well. Comparison was performed using the cosine similarity metric between the two embedded vectors.

	Document-A	Document-B	Document-C	Document-D	CORRECT
Query-A	2.41E-02	1.29E-02	-5.36E-02	-1.71E-04	O
Query-B	3.08E-01	1.06E-01	1.14E-01	3.13E-03	O
Query-C	5.59E-02	5.51E-02	-5.19E-02	1.72E-02	O
Query-D	-2.19E-02	1.89E-01	-1.81E-02	4.81E-03	O
Query-E	-1.65E-01	8.60E-02	3.52E-02	-1.19E-01	O
Query-F	-7.17E-03	2.93E-01	2.89E-01	1.01E-01	O

Cleaned Dataset Test

	Document-A	Document-B	Document-C	Document-D	CORRECT
Query-A	-2.27E-02	7.78E-02	-1.54E-02	4.33E-02	X
Query-B	2.04E-01	5.59E-02	1.33E-01	4.23E-02	O
Query-C	3.64E-02	1.26E-01	-3.91E-02	1.14E-01	X
Query-D	-5.27E-02	2.05E-01	-3.25E-02	2.16E-02	O
Query-E	-1.47E-01	3.14E-01	6.23E-02	-7.93E-02	O
Query-F	2.35E-02	1.93E-01	2.65E-01	4.17E-02	O

Unclean Dataset Test

The results of the test showed that the unclean dataset polluted the results even though the BERT model should be able to handle such out-of-context tokens. Furthermore, the extremely basic preprocessing phase leaves a lot of ill-formed and stray tokens that could cause more issue—but did not.

Public Health Nurs. 2020;00:1-8.
wileyonlinelibrary.com/journal/phn | 1
© 2020 Wiley Periodicals LLC.

1 | BACKGROUND

Rapid and unprecedented changes worldwide have increased the need for

→

Public health nurs wileyonlinelibrary com journal phn wiley periodicals llc background rapid unprecedented changes worldwide increased need leadership

Conclusions

- ➔ BERT model performance may be improved by running text pre-processing beforehand.
- ➔ Simpler and cheaper functions such as BM25 can still compete against LLMs
- ➔ Embeddings require immense compute power and may demand greater storage

References/Data Utilized

SciBERT: A Pretrained Language Model for Scientific Text (beltagy-etal-2019-scibert)

Iz Beltagy and Kyle Lo, and Arman Cohan

[https://www.aclweb.org/anthology/D19-1371]

NeuralKG: An Open Source Library for Diverse Representation Learning of Knowledge Graphs

[https://arxiv.org/pdf/2202.12571]

ORKG ASK: a Neuro-symbolic Scholarly Search and Exploration System

[https://arxiv.org/pdf/2412.04977]