Government of India Hackathon: Medical Expert Search Engine technical report and documentation

Wei Wang, Yash Pratyush Sinha

June 12, 2017

1 Aim

A tool for Mining MEDLINE/Pubmed for identifying subject Experts. Mediline/Pubmed is an xml based repository of medical papers and their authors and abstracts. Our tool must mine it, and should give a ranked list of experts based on a search query.

2 Requirements

We have two major requirements:

- PubmedMedline data containing at least the paper, it's references it', it's Medical Subject headings (MeSH) & keywords and it's author's names.
- A big medical knowledge graph, similar to wordnet (or babelnet) so as to search and understand the medical search terms much more effectively. Since our extensive search did not let us find any freely available medical wornet, we built our own solution to find similar words using tf-idf.

2.1 Technology

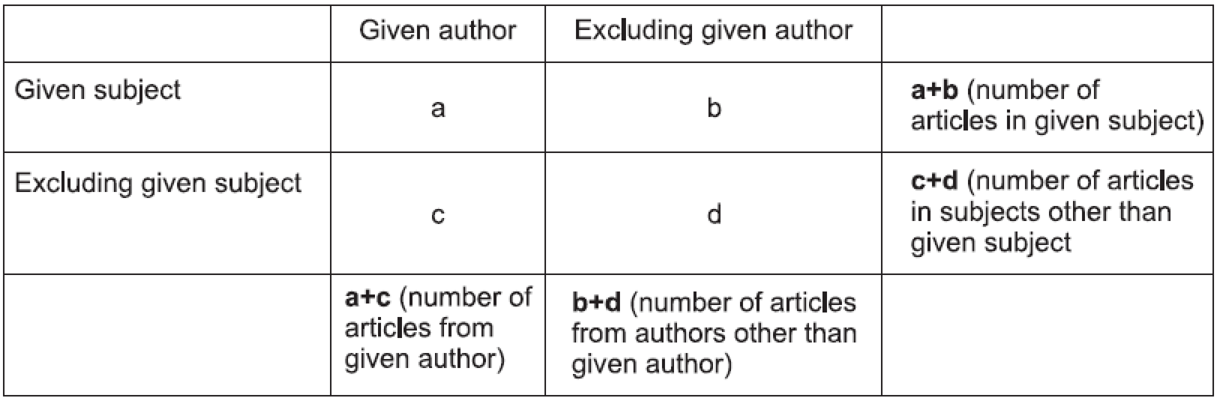
Python, with scikit-learn and pandas for most of the searching, and heavy lifting.
R for some of the parsing.
MySQL for database.
Php for the web application

3 ALGORTIHMS

We have used several algorithms in our project. They are:

3.1 Z-SCORE

Z-score is a method of scoring a particular author according to how many papers has he/she written about that given subject. For this, we first define a few variables:



	Given author	Excluding given author	
Given subject	a	b	**a+b** (number of articles in given subject)
Excluding given subject	c	d	**c+d** (number of articles in subjects other than given subject
	a+c (number of articles from given author)	**b+d** (number of articles from authors other than given author)	

Using these variables, we define the algorithm as follows:

$$z = \frac{|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

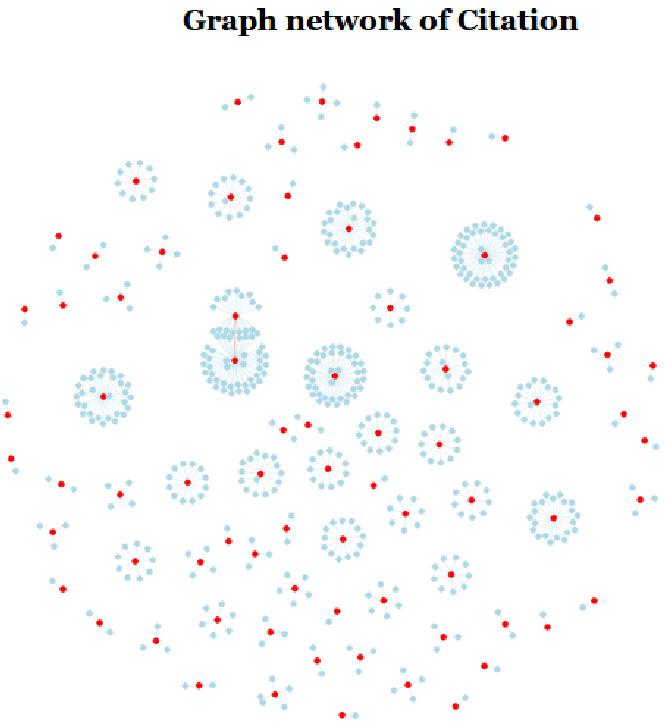
$$n_1 = (a+b) \quad and \quad n_2 = (c+d)$$

$$\hat{p} = \frac{(a+c)}{(a+b+c+d)} \quad and \quad \hat{q} = 1 - \hat{p}$$

$$\hat{p}_1 = \frac{a}{(a+b)} \quad and \quad \hat{p}_2 = \frac{c}{(c+d)}.$$

Now, this z-score increases with higher number of papers in the given field, decreases with diversity and increases with specialization. For more informations, refer to [1]

3.2 Pagerank



We have used Google's pagerank algorithm to devlope a ranking system of all papers where each paper cites another paper. This way we can set papers which have been cited more as more important. For more information about pagerank, refer to [2]

3.3 FLAE

FLAE or First-Last Author Emphasis is a simple algorthim using which one can score individual author for a given paper. In this, the first author should get credit of the whole impact, the last author half, and the credit of the other authors is the impact divided by the number of all authors. For more information, refer to [3].

3.4 Medical Dictionary to Term Similarity builder

This is an in-house developed algorithm which given a list of medical terms and their definitions, can use these definitions to find terms simliar to the one given in query. This is done by assuming each term-definition as a separate document and then calculating the tf-idf for each word in each document/definition. It then multiplies the tf-idf score of common words of each definition to find the similarity between them. By thresholding this similarity score at 0.25, we can say whether two terms are similar or not

4 Overall Method

1. On getting a search term, we find similar term to the given query. We call this list as T.
2. We search all papers with any word of T in their MeSH or Keywords. Each paper is given a 'relevance score' according to how many terms of T are present in it. Using these papers, we built a set of authors A which has to be ranked
3. We calculate weighted a,b,c and d (as in z-score) for each author, where every number of paper is multiplied by the relevance score and it's page rank score.
4. We calculate this weighted z-score for each author and rank them according to it.

References

[1] Developing a Biomedical Expert Finding System Using Medical Subject Headings, Harpreet Singh, PhD, Reema Singh, PhD, Arjun Malhotra, MSc, Manjit Kaur, MCA http://dx.doi.org/10.4258/hir.2013.19.4.243

[2] A Document Clustering and Ranking System for Exploring MEDLINE Citations, YONGJING LIN , Ms, WENYUAN LI , PHD, KEKE CHEN , PHD, YING LIU , PHD Journal of the American Medical Informatics Association Volume 14 Number 5 Sept / Oct 2007

[3] Author Sequence and Credit for Contributions in Multiauthored Publication, Teja Tscharntke, Michael E Hochberg, Tatyana A Rand, Vincent H Resh, and Jochen Krauss doi:10.1371/journal.pbio.0050018