MACHINE LEARNED RANKING

Rodrigo Pueblas
David Burrell
Angel Igareta
Miguel Pérez

INTRODUCTION TO THE PROBLEM

We were given a dataset of medical results from the LOINC database.

We were asked to build a model capable of ranking the results for any set of given queries:

- “Glucose in blood”
- “Bilirubin in plasma”
- “White blood cells count”

SELECTED APPROACH: PAIRWISE RANKING

We based our approach on the paper “Optimizing Search Engines using Clickthrough Data” by Thorsten Joachims.

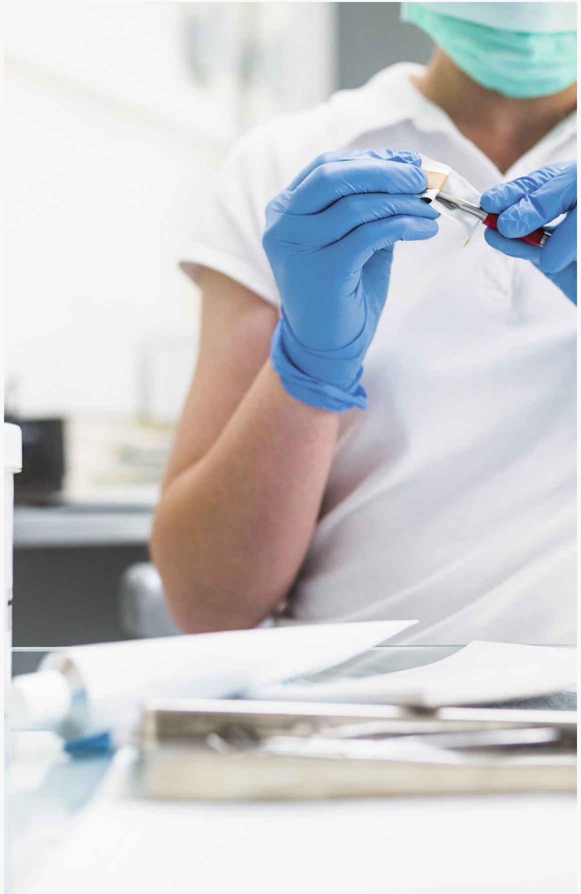
Data can be represented as a triplet (query, ranking, clicks), given a unique ID.

The original method used query data from the web browser’s logs to train the model.

This ranking will compute pairwise comparisons, ending up with a total ranking of documents for each query.

PROPOSED SOLUTION

To compare the textual components of the results (columns) and the queries, we used the cosine similarity metric to compare documents in a pairwise way, after that, ranking.



DATASET GENERATION

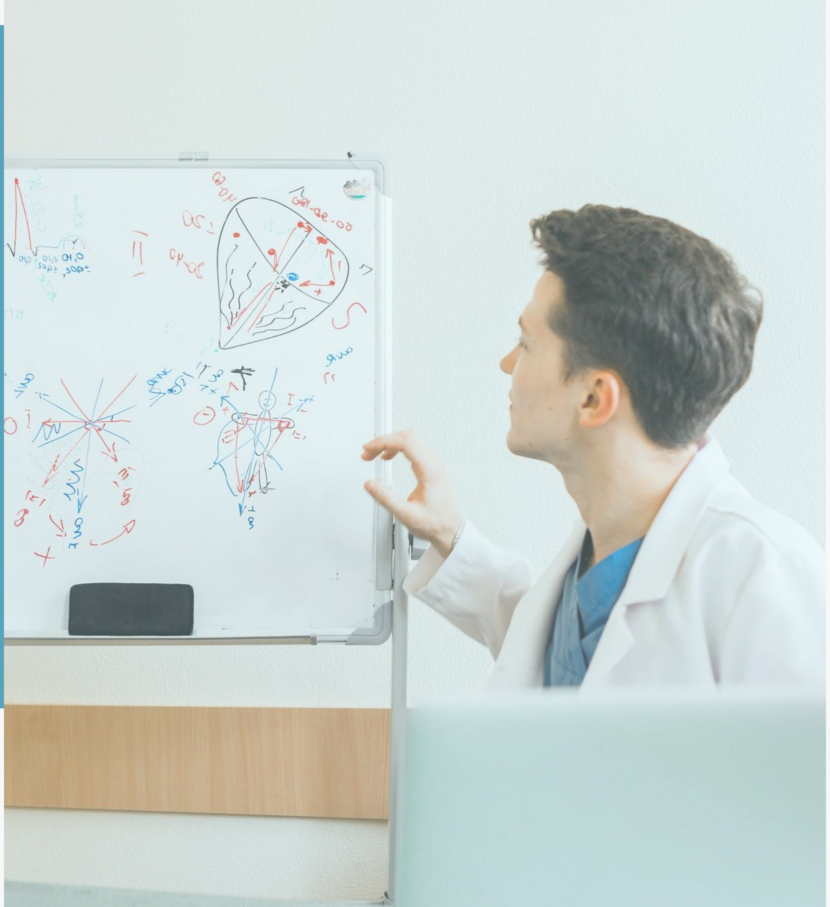
To simulate past user “click data” a set of keywords for each query were compared to each document to creating a matching score

Example for query 1:

"white", "blood", "cells", "count", "hemoglobin", "plasma", "leucocyte"

To add a random element number between 2 - 6 added

Highest score first document / lowest last



IMPLEMENTATION

We converted our textual features using the method mentioned before, and generated our training data set.

We trained a SVM model with our features with promising results, but not perfect. Good start!

The full implementation of the generated dataset and our attempt to build the corresponding model can be found in the following Github repository:

https://github.com/angeligareta/MLRanking/

RESULTS

Glucose in blood

Ranking
Glucose [Moles/volume] in Urine
Glucose [Moles/volume] in Pleural fluid
Glucose [Moles/volume] in Serum or Plasma
Glucose [Mass/volume] in Serum Plasma or Blood
Cholesterol in I-iDL [Mass/volume] in Serum or Plasma

Bilirubin in plasma

Ranking
Bilirubin total [Mass/volume] in Synovial fluid
Bilirubin indirect [Mass/volume] in Serum or Plasma
Bilirubin direct [Mass/volume] in Serum or Plasma
Bilirubin total [Mass/volume] in Serum or Plasma
Cholesterol in I-iDL [Mass/volume] in Serum or Plasma

White blood cells count

Ranking
Billrubin total [Presence] in Unspecfied specimen
Nitrofurantoin [Susceptibility]
Cholesterol [Mass/volume] in Serum or Plasma
Trimethoprim+Sulfamethoxazole [Susceptibility]
Blood group antibody screen [Presence] in Serum or Plasma

CONCLUSIONS

The domain of ranking search result entries is very broad and complex.

In some cases, we can rely on the criterion of the user to make more relevant the results that are most clicked on.

The huge advantage of this approach in this era is the amount of easily accessible, cheap data to support these algorithms.

THANKS FOR WATCHING

David Burrell
Angel Igareta
Rodrigo Pueblas
Miguel Pérez