Information Retrieval

Machine Learning Ranking

Index

Introduction

1,000,000 queries per day

Thousands of pages contain words from the user's query!

“FEEDBACK”

If we knew the set of pages actually relevant to the user’s query, we could use this as training data for optimizing (and even personalizing) the retrieval function.

”Optimizing Search Engines using Clickthrough Data”

Pairwise algorithm

Optimizing Search Engines using Clickthrough Data

Clickthrough triplets

Clickthrough data in search engines can be thought of as triplets (q, r, c)

Clickthrough triplets

Query

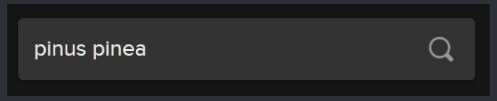
The set of words the user looked for

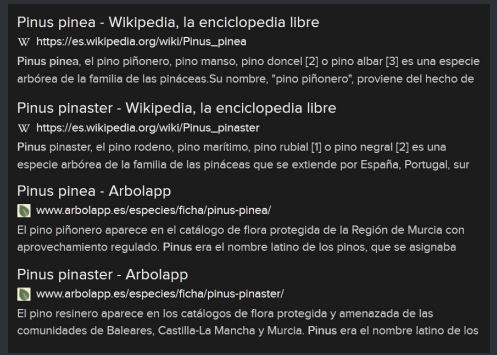
Ranking

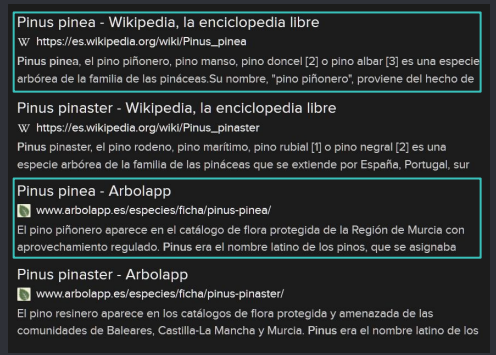
The set of documents presented to the user (search results)

Clicks

Set of documents the user clicked on







Training dataset

Information in clickthrough data

Training set

Query

“White blood cells count”

Ranking

26484-6	Monocytes [#/volume] in Blood	Monocytes	Bld	NCnc
14423-8	Bilirubin.total [Mass/volume] in Synovial	Bilirubin	Synv fld	MCnc
35192-4	Bilirubin.indirect [Mass or Moles/volume	Bilirubin.non	Ser/Plas	MSCn
4671-4	Protein C [Mass/volume] in Plasma	Protein C	Plas	MCnc
35184-1	Fasting glucose [Mass or Moles/volume]	Glucose^post	Ser/Plas	MSCn
74774-1	Glucose [Mass/volume] in Serum, Plasma	Glucose	Ser/Plas/Bld	MCnc
3082-5	Tyrosine aminotransferase [Mass/volume	Tyrosine ami	Plas	MCnc
33870-7	Bilirubin.total [Presence] in Unspecified s	Bilirubin	XXX	PrThr
14747-0	Glucose [Moles/volume] in Pleural fluid	Glucose	Plr fld	SCnc
26474-7	Lymphocytes [#/volume] in Blood	Lymphocytes	Bld	NCnc
1975-2	Bilirubin.total [Mass/volume] in Serum or	Bilirubin	Ser/Plas	MCnc
14749-6	Glucose [Moles/volume] in Serum or Plas	Glucose	Ser/Plas	SCnc
1742-6	Alanine aminotransferase [Enzymatic acti	Alanine amin	Ser/Plas	CCnc
26478-8	Lymphocytes/100 leukocytes in Blood	Lymphocytes	Bld	NFr
15076-3	Glucose [Moles/volume] in Urine	Glucose	Urine	SCnc
20442-0	Hepatitis B virus DNA [#/volume] (viral lo	Hepatitis B vi	Ser	NCnc
14764-5	Glucose [Moles/volume] in Serum or Plas	Glucose^3H p	Ser/Plas	SCnc
26464-8	Leukocytes [#/volume] in Blood	Leukocytes	Bld	NCnc
54439-5	Calcium bilirubinate/Total in Stone	Calcium bilir	Calculus	MFr
1920-8	Aspartate aminotransferase [Enzymatic a	Aspartate am	Ser/Plas	CCnc

Clicks

loinc_num	long_common_name	component	system	property	LABEL
26484-6	Monocytes [#/volume] in Blood	Monocytes	Bld	NCnc	1
14423-8	Bilirubin.total [Mass/volume] in Synovial	Bilirubin	Synv fld	MCnc	0
35192-4	Bilirubin.indirect [Mass or Moles/volume]	Bilirubin.non	Ser/Plas	MSCnc	0
4671-4	Protein C [Mass/volume] in Plasma	Protein C	Plas	MCnc	0
35184-1	Fasting glucose [Mass or Moles/volume] i	Glucose^post	Ser/Plas	MSCnc	0
74774-1	Glucose [Mass/volume] in Serum, Plasma	Glucose	Ser/Plas/Bld	MCnc	0
3082-5	Tyrosine aminotransferase [Mass/volume	Tyrosine ami	Plas	MCnc	0
33870-7	Bilirubin.total [Presence] in Unspecified s	Bilirubin	XXX	PrThr	0
14747-0	Glucose [Moles/volume] in Pleural fluid	Glucose	Plr fld	SCnc	0
26474-7	Lymphocytes [#/volume] in Blood	Lymphocytes	Bld	NCnc	1
1975-2	Bilirubin.total [Mass/volume] in Serum or	Bilirubin	Ser/Plas	MCnc	0
14749-6	Glucose [Moles/volume] in Serum or Plas	Glucose	Ser/Plas	SCnc	0
1742-6	Alanine aminotransferase [Enzymatic acti	Alanine amin	Ser/Plas	CCnc	0
26478-8	Lymphocytes/100 leukocytes in Blood	Lymphocytes	Bld	NFr	1
15076-3	Glucose [Moles/volume] in Urine	Glucose	Urine	SCnc	0
20442-0	Hepatitis B virus DNA [#/volume] (viral loa	Hepatitis B vi	Ser	NCnc	0
14764-5	Glucose [Moles/volume] in Serum or Plas	Glucose^3H p	Ser/Plas	SCnc	0
26464-8	Leukocytes [#/volume] in Blood	Leukocytes	Bld	NCnc	1
54439-5	Calcium bilirubinate/Total in Stone	Calcium bilir	Calculus	MFr	0
1920-8	Aspartate aminotransferase [Enzymatic ac	Aspartate am	Ser/Plas	CCnc	0

Training set

Format the data:

1. Target: Ranking position for each query
2. Query ID: Query Unique Identifier
3. Document ID: Document Unique Identifier(only for information purposes)

loinc_num	long_common_name	component	system	property	LABEL
26484-6	Monocytes [#/volume] in Blood	Monocytes	Bld	NCnc	1
14423-8	Bilirubin.total [Mass/volume] in Synovial	Bilirubin	Synv fld	MCnc	0
35192-4	Bilirubin.indirect [Mass or Moles/volume]	Bilirubin.non	Ser/Plas	MSCnc	0
4671-4	Protein C [Mass/volume] in Plasma	Protein C	Plas	MCnc	0
35184-1	Fasting glucose [Mass or Moles/volume] i	Glucose^post	Ser/Plas	MSCnc	0
74774-1	Glucose [Mass/volume] in Serum, Plasma	Glucose	Ser/Plas/Bld	MCnc	0
3082-5	Tyrosine aminotransferase [Mass/volume	Tyrosine ami	Plas	MCnc	0
33870-7	Bilirubin.total [Presence] in Unspecified s	Bilirubin	XXX	PrThr	0
14747-0	Glucose [Moles/volume] in Pleural fluid	Glucose	Plr fld	SCnc	0
26474-7	Lymphocytes [#/volume] in Blood	Lymphocytes	Bld	NCnc	1
1975-2	Bilirubin.total [Mass/volume] in Serum or	Bilirubin	Ser/Plas	MCnc	0
14749-6	Glucose [Moles/volume] in Serum or Plas	Glucose	Ser/Plas	SCnc	0
1742-6	Alanine aminotransferase [Enzymatic acti	Alanine amin	Ser/Plas	CCnc	0
26478-8	Lymphocytes/100 leukocytes in Blood	Lymphocytes	Bld	NFr	1
15076-3	Glucose [Moles/volume] in Urine	Glucose	Urine	SCnc	0
20442-0	Hepatitis B virus DNA [#/volume] (viral lo	Hepatitis B vi	Ser	NCnc	0
14764-5	Glucose [Moles/volume] in Serum or Plas	Glucose^3H p	Ser/Plas	SCnc	0
26464-8	Leukocytes [#/volume] in Blood	Leukocytes	Bld	NCnc	1
54439-5	Calcium bilirubinate/Total in Stone	Calcium biliru	Calculus	MFr	0
1920-8	Aspartate aminotransferase [Enzymatic ac	Aspartate am	Ser/Plas	CCnc	0


```
#Target Query_ID Document_ID
1 qid:1 #1975-2
2 qid:1 #15076-3
5 qid:1 #74774-1
1 qid:1 #1920-8
1 qid:1 #54439-5
1 qid:1 #14747-0
1 qid:1 #14423-8
6 qid:1 #14749-6
1 qid:1 #26464-8
1 qid:1 #3082-5
1 qid:1 #26474-7
3 qid:1 #14764-5
1 qid:1 #33870-7
1 qid:1 #4671-4
1 qid:1 #20442-0
1 qid:1 #26484-6
1 qid:1 #26478-8
1 qid:1 #35192-4
4 qid:1 #35184-1
1 qid:1 #1742-6
1 qid:2 #26464-8
2 qid:2 #1920-8
1 qid:2 #35184-1
1 qid:2 #26478-8
1 qid:2 #14423-8
1 qid:2 #14747-0
1 qid:2 #14764-5
```

Training set

Each Target value belongs to one Query ID.

Documents with higher Target values are ranked higher.

Order:

- Query 1: B > C > A
- Query 2: A > [B,C]

```
#Target Query_ID Document_ID
1 qid:1 #A
3 qid:1 #B
2 qid:1 #C
2 qid:2 #A
1 qid:2 #B
1 qid:2 #C
```

Working with the model

Training and ranking

Training

1. The model receives the training set in the appropriate format
2. The model solves the optimization problem
3. We obtain a ranking function that has a low number of discordant pairs with respect to the target ranking

OPTIMIZATION PROBLEM 2. (RANKING SVM (PARTIAL))

$$\textit{minimize:} \qquad V(\vec{w}, \vec{\xi}) = \frac{1}{2}\,\vec{w} \cdot \vec{w} + C \sum \xi_{i,j,k} \tag{21}$$

subject to:

$$\forall (d_i, d_j) \in r'_1 : \vec{w}\Phi(q_1, d_i) > \vec{w}\Phi(q_1, d_j) + 1 - \xi_{i,j,1}$$

$$\dots \tag{22}$$

$$\forall (d_i, d_j) \in r'_n : \vec{w}\Phi(q_n, d_i) > \vec{w}\Phi(q_n, d_j) + 1 - \xi_{i,j,n}$$

$$\forall i \forall j \forall k : \xi_{i,j,k} \geq 0 \tag{23}$$

Ranking

- For each query
  a. Order documents with different target values from highest to lowest
  b. Order documents with equal target values according to their features

- Possible features:
  a. Rank in other search engines
  b. Query/content match
  c. Popularity attributes

Thanks!

ANY QUESTIONS?

Sofia Smolianinova - MUCD Erasmus

Álvaro Arranz - EIT Digital

Ignacio Martínez - EIT Health