Information Retrieval

Machine Learning Ranking

Index

- Introduction
- Pairwise algorithm
- Training dataset
- Working with the model

Introduction

1,000,000 queries per day

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

"FEEDBACK"

66

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.

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



Pinus pinea - Wikipedia, la enciclopedia libre W https://es.wikipedia.org/wiki/Pinus_pinea Pinus pinea, el pino piñonero, pino manso, pino doncel [2] o pino albar [3] es una especie arbórea de la familia de las pináceas. Su nombre, "pino piñonero", proviene del hecho de Pinus pinaster - Wikipedia, la enciclopedia libre W https://es.wikipedia.org/wiki/Pinus_pinaster Pinus pinaster, el pino rodeno, pino marítimo, pino rubial [1] o pino negral [2] es una especie arbórea de la familia de las pináceas que se extiende por España. Portugal, sur Pinus pinea - Arbolapp www.arbolapp.es/especies/ficha/pinus-pinea/ El pino piñonero aparece en el catálogo de flora protegida de la Región de Murcia con aprovechamiento regulado. Pinus era el nombre latino de los pinos, que se asignaba Pinus pinaster - Arbolapp www.arbolapp.es/especies/ficha/pinus-pinaster/ El pino resinero aparece en los catálogos de flora protegida y amenazada de las comunidades de Baleares, Castilla-La Mancha y Murcia. Pinus era el nombre latino de los Pinus pinea - Wikipedia, la enciclopedia libre

W https://es.wikipedia.org/wiki/Pinus_pinea

Pinus pinaa, el pino piñonero, pino manso, pino doncel [2] o pino albar [3] es una especie
arbórea de la familia de las pináceas Su nombre, "pino piñonero", proviene del hecho de

Pinus pinaster - Wikipedia, la enciclopedia libre

W https://es.wikipedia.org/wiki/Pinus_pinaster

Pinus pinaster, el pino rodeno, pino maritimo, pino rubial [1] o pino negral [2] es una
especie arbórea de la familia de las pináceas que se extiende por España, Portugal, sur

Pinus pinaer, el pino rodeno, pino maritimo, pino rubial [1] o pino negral [2] es una
especie arbórea de la familia de las pináceas que se extiende por España, Portugal, sur

Pinus pinaer - Arbolapp

"a www.arbolapp.es/especies/ficha/pinus-pinae/
El pino piñonero aparece en el catálogo de flora protegida de la Región de Murcia con
aprovechamiento regulado. Pinus era el nombre latino de los pinos, que se asignaba

Pinus pinaster - Arbolapp

"a www.arbolapp.es/especies/ficha/pinus-pinaster/
El pino reinsiero aparece en los catálogos de flora protegida y amenazada de las

comunidades de Baleares, Castilla-La Mancha y Murcia. Pinus era el nombre latino de los

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 biliru	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	(
35192-4	Bilirubin.indirect [Mass or Moles/volume]	Bilirubin.non	Ser/Plas	MSCnc	(
4671-4	Protein C [Mass/volume] in Plasma	Protein C	Plas	MCnc	(
35184-1	Fasting glucose [Mass or Moles/volume] i	Glucose^post	Ser/Plas	MSCnc	C
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	0
14747-0	Glucose [Moles/volume] in Pleural fluid	Glucose	Plr fld	SCnc	(
26474-7	Lymphocytes [#/volume] in Blood	Lymphocytes	Bld	NCnc	1
1975-2	Bilirubin.total [Mass/volume] in Serum or	Bilirubin	Ser/Plas	MCnc	C
14749-6	Glucose [Moles/volume] in Serum or Plas	Glucose	Ser/Plas	SCnc	(
1742-6	Alanine aminotransferase [Enzymatic activ	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	(
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	C
26464-8	Leukocytes [#/volume] in Blood	Leukocytes	Bld	NCnc	1
54439-5	Calcium bilirubinate/Total in Stone	Calcium biliru	Calculus	MFr	C
1920-8	Aspartate aminotransferase [Enzymatic ac	Aspartate am	Ser/Plas	CCnc	

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		Query_ID	Doc
26484-6	Monocytes [#/volume] in Blood	Monocytes	Bld	NCnc	1		#1975-2 #15076-3	
14423-8	Bilirubin.total [Mass/volume] in Synovial	Bilirubin	Synv fld	MCnc	0		#74774-1	
35192-4	Bilirubin.indirect [Mass or Moles/volume]	Bilirubin.non	Ser/Plas	MSCnc	0		#1920-8	
4671-4	Protein C [Mass/volume] in Plasma	Protein C	Plas	MCnc	0		#54439-5	
35184-1	Fasting glucose [Mass or Moles/volume] i	Glucose^pos	Ser/Plas	MSCnc	0		#14747-0 #14423-8	
74774-1	Glucose [Mass/volume] in Serum, Plasma		Ser/Plas/Bld		0		#14749-6	
3082-5	Tyrosine aminotransferase [Mass/volume			MCnc	0		#26464-8	
33870-7	Bilirubin.total [Presence] in Unspecified s		XXX	PrThr	0		#3082-5 #26474-7	
							#14764-5	
14747-0	Glucose [Moles/volume] in Pleural fluid		Plr fld	SCnc	0		#33870-7	
26474-7	Lymphocytes [#/volume] in Blood	Lymphocytes	Bld	NCnc	1	1 qid:1	#4671-4	
1975-2	Bilirubin.total [Mass/volume] in Serum or	Bilirubin	Ser/Plas	MCnc	0		#20442-0	
14749-6	Glucose [Moles/volume] in Serum or Plas	Glucose	Ser/Plas	SCnc	0		#26484-6	
1742-6	Alanine aminotransferase [Enzymatic activ	Alanine amin	Ser/Plas	CCnc	0		#26478-8 #35192-4	
26478-8		Lymphocytes	100007	NFr	1		#35184-1	
15076-3		Glucose	Urine	SCnc	0		#1742-6	
20442-0	Hepatitis B virus DNA [#/volume] (viral loa			NCnc	0		#26464-8 #1920-8	
14764-5	Glucose [Moles/volume] in Serum or Plass	-		SCnc	0		#1920-8 2 #35184-1	
					100		#26478-8	
26464-8	Leukocytes [#/volume] in Blood	Leukocytes	Bld	NCnc	1		#14423-8	
54439-5	Calcium bilirubinate/Total in Stone	Calcium biliru	Calculus	MFr	0		#14747-0	
1920-8	Aspartate aminotransferase [Enzymatic ad	Aspartate am	Ser/Plas	CCnc	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
- The model solves the optimization problem
- 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))

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

subject to: $\forall (d_i, d_j) \in r'_1 : \vec{w} \Phi(q_1, d_i) > \vec{w} \Phi(q_1, d_j) + 1 - \xi_{i,j,1}$... $\forall (d_i, d_j) \in r'_n : \vec{w} \Phi(q_n, d_i) > \vec{w} \Phi(q_n, d_j) + 1 - \xi_{i,j,n}$ $\forall i \forall j \forall k : \xi_{i,j,k} \geq 0$ (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