

Valutazione qualità

Gabriele Pisciotta

12/12/2020

In questa sperimentazione vogliamo valutare la qualità dei match ottenuti attraverso le query per bibref, sia verso l'indice locale sia verso Crossref. Andremo a considerare sia il caso che include i documenti i cui DOI non sono in Crossref, sia il caso che li esclude. Cercheremo in seguito di trovare una soglia di score sotto la quale scartare i risultati ottenuti.

Configurazione della sperimentazione

- Selezionato 10 documenti citanti e relative citazioni (output di BEE)
- Creato un CSV estraendo ogni coppia (DOI, bibref) presente nei documenti citanti
- Per ogni riga del CSV, prendo la bibref ed effettuo la query sia verso l'indice locale, sia verso Crossref, salvando i risultati in altre due colonne del CSV.
- Annoto manualmente il DOI nei casi:
 - DOI esplicito mancante
 - DOI discordanti tra locale e remoto
 - DOI discordanti tra locale ed esplicito
 - DOI discordanti tra remoto ed esplicito.
- Se non è stato possibile trovare un DOI manualmente, inserisco un “???”.
- Escludo dal file CSV le righe appartenenti ai seguenti casi:
 - il doi manuale non è stato possibile trovarlo (valore: ???)
 - il doi trovato manualmente non è incluso nel dump di Crossref (e quindi non sarebbe possibile trovarlo!)

Esempio di tabella risultante

A	B	C	D	E	F
explicit_doi	bibref	retrieved_local_doi	retrieved_remote_doi	manual_doi	
	Mendoza, L., Vilela, R., Voelz, J. 10.1101/cshperspect.a019562	10.1101/cshperspect.a019562	10.1101/cshperspect.a019562	10.1101/cshperspect.a019562	
	Prabhu, RM, and Patel, R. Mu 10.1111/j.1470-9465.2004.00843.x	10.1111/j.1470-9465.2004.00843.x	10.1111/j.1470-9465.2004.00843.x	10.1111/j.1470-9465.2004.00843.x	
	Bittencourt, AL, Marback, R. N 10.4269/ajtmh.2006.75.936	10.4269/ajtmh.2006.75.936	10.4269/ajtmh.2006.75.936	10.4269/ajtmh.2006.75.936	
10.1128/cmr.18.3.556-569.2005	Spellberg, B, Edwards, J, Jr, II 10.1128/cmr.18.3.556-569.2005	10.1128/cmr.18.3.556-569.2005	10.1128/cmr.18.3.556-569.2005	10.1128/cmr.18.3.556-569.2005	
10.1371/journal.pntd.0003984	Blumentrath, CG, Grobusch, A 10.1371/journal.pntd.0003984	10.1371/journal.pntd.0003984	10.1371/journal.pntd.0003984	10.1371/journal.pntd.0003984	
	Gugnani, HC. Entomophthoroi 10.1007/bf00158574	10.1007/bf00158574	10.1007/bf00158574	10.1007/bf00158574	
10.1037/0033-2909.133.3.482	Levin, R, Nielsen, T.A. Disturb 10.1037/0033-2909.133.3.482	10.1037/0033-2909.133.3.482	10.1037/0033-2909.133.3.482	10.1037/0033-2909.133.3.482	
	Schredl, M. Researching Drea 10.1007/978-3-319-95453-0	10.1007/978-3-319-95453-0	10.1007/978-3-319-95453-0	10.1007/978-3-319-95453-0	
10.1037/a0037749	Böckermann, M, Gieselmann, 10.1037/a0037749	10.1037/a0037749	10.1037/a0037749	10.1037/a0037749	

La tabella completa e tutte quelle generate nei passaggi intermedi, è possibile visionarli (qui)

Elementi valutati

Saranno valutati i seguenti valori al fine di caratterizzare la qualità dei match:

- numero di volte in cui il DOI locale è corretto e quello remoto no
- numero di volte in cui il DOI remoto è corretto e quello locale no
- numero di volte in cui sono entrambi errati e identici
- numero di volte in cui sono entrambi errati ma differenti
- numero di volte in cui sono entrambi corretti e identici
- numero di volte in cui troviamo un supplemento del doi corretto come risultato (usando l'indice)
- numero di volte in cui troviamo un supplemento del doi corretto come risultato (usando le API)
- percentuale dei locali sbagliati che sono supplementi sul totale
- percentuale dei remoti sbagliati che sono supplementi sul totale
- percentuale dei locali sbagliati che sono supplementi sugli errori usando l'indice
- percentuale dei remoti sbagliati che sono supplementi sugli errori usando le API
- percentuale di corrispondenza tra DOI locale e DOI manuale (quanta correttezza usando l'indice)
- percentuale di corrispondenza tra DOI remoto e DOI manuale (quanta correttezza usando le API remote)
- percentuale di corrispondenza tra DOI remoto e DOI locale (quanta corrispondenza tra l'indice e le API)
- percentuale entrambi corretti
- percentuale entrambi errati e identici

- percentuale entrambi errati ma differenti
- numero di documenti scartati

Riguardo il caso in cui sono entrambi errati siamo infine interessati a:

- percentuale entrambi errati e identici (su tutti gli errati)
- percentuale entrambi errati ma differenti (su tutti gli errati)

Metriche riassuntive:

- precision
- recall
- F1

Risultati

I seguenti risultati sono stati valutati su un numero di 413 record a seguito di un preprocessing volto ad escludere le righe appartenenti a casistiche indesiderate, come definito nella configurazione della sperimentazione, dal totale 17 sono stati esclusi perchè non posseduti dal dump di Crossref.

- Numero di volte in cui il DOI locale è corretto e quello remoto no: 17
- Numero di volte in cui il DOI remoto è corretto e quello locale no: 5
- Numero di volte in cui sono entrambi errati e identici: 6
- Numero di volte in cui sono entrambi errati ma differenti: 2
- Numero di volte in cui sono entrambi corretti e identici: 383
- Numero di volte in cui vengono restituiti supplementi dall'indice locale: 5
- Numero di volte in cui vengono restituiti supplementi dalle API: 11
- Percentuale locali sbagliati che sono supplementi (sul totale): 1.21
- Percentuale remoti sbagliati che sono supplementi (sul totale): 2.66
- Percentuale locali sbagliati che sono supplementi (sui locali sbagliati): 38.46
- Percentuale remoti sbagliati che sono supplementi (sui remoti sbagliati): 44.00
- Percentuale di corrispondenza tra DOI locale e DOI manuale (quanta - correttezza usando l'indice): 96.85
- Percentuale di corrispondenza tra DOI remoto e DOI manuale (quanta correttezza usando le API remote): 93.95
- Percentuale di corrispondenza tra DOI remoto e DOI locale (quanta corrispondenza tra l'indice e le API): 94.19
- Percentuale entrambi corretti: 92.74
- Percentuale entrambi errati e identici: 1.45
- Percentuale entrambi errati ma differenti: 0.48
- Media score nei corretti 55.75
- Media score negli sbagliati 43.81
- Scartati 0

Se sbagliano entrambi: - Percentuale entrambi errati e identici su tutti gli errati: 75.00 - Percentuale entrambi errati ma differenti su tutti gli errati: 25.00

Metriche riassuntive:

	Local	Crossref
Precision	0.968	0.948
Recall	0.968	0.941
F1	0.968	0.945

Risultati includendo anche i documenti i cui DOI non sono in Crossref

- Numero di volte in cui il DOI locale è corretto e quello remoto no: 17
- Numero di volte in cui il DOI remoto è corretto e quello locale no: 5
- Numero di volte in cui sono entrambi errati e identici: 14
- Numero di volte in cui sono entrambi errati ma differenti: 5
- Numero di volte in cui sono entrambi corretti e identici: 389
- Numero di volte in cui vengono restituiti supplementi dall'indice locale: 6
- Numero di volte in cui vengono restituiti supplementi dalle API: 12
- Percentuale locali sbagliati che sono supplementi (sul totale): 1.40
- Percentuale remoti sbagliati che sono supplementi (sul totale): 2.79
- Percentuale locali sbagliati che sono supplementi (sui locali sbagliati): 25.00
- Percentuale remoti sbagliati che sono supplementi (sui remoti sbagliati): 33.33
- Percentuale di corrispondenza tra DOI locale e DOI manuale (quanta - correttezza usando l'indice): 94.42
- Percentuale di corrispondenza tra DOI remoto e DOI manuale (quanta correttezza usando le API remote): 91.63
- Percentuale di corrispondenza tra DOI remoto e DOI locale (quanta corrispondenza tra l'indice e le API): 93.72
- Percentuale entrambi corretti: 90.47
- Percentuale entrambi errati e identici: 3.26
- Percentuale entrambi errati ma differenti: 1.16
- Media score nei corretti 55.71
- Media score negli sbagliati 38.96
- Scartati 0

Se sbagliano entrambi: - Percentuale entrambi errati e identici su tutti gli errati: 73.68 - Percentuale entrambi errati ma differenti su tutti gli errati: 26.32

Metriche riassuntive:

	Local	Crossref
Precision	0.946	0.927
Recall	0.944	0.918

	Local	Crossref
F1	0.945	0.922

Ricerca della migliore soglia

Procediamo ad effettuare una ricerca della migliore soglia in base al miglior F1 score ottenuto, verificando infine i risultati su un test set. Utilizziamo una grid search per la soglia θ nel range $[0, 50]$. La sperimentazione viene effettuata attraverso la tecnica di validazione hold-out.

Il seguente esperimento è sui dati che includono anche bibref i cui DOI non sono in Crossref.

θ	F1	P	R
0	0.921512	0.921512	0.921512
1	0.921512	0.921512	0.921512
2	0.921512	0.921512	0.921512
3	0.921512	0.921512	0.921512
4	0.921512	0.921512	0.921512
5	0.921512	0.921512	0.921512
6	0.921512	0.921512	0.921512
7	0.921512	0.921512	0.921512
8	0.921512	0.921512	0.921512
9	0.921512	0.921512	0.921512
10	0.921512	0.921512	0.921512
11	0.921512	0.921512	0.921512
12	0.921512	0.921512	0.921512
13	0.921512	0.921512	0.921512
14	0.921512	0.921512	0.921512
15	0.921512	0.921512	0.921512
16	0.921283	0.923977	0.918605
17	0.921283	0.923977	0.918605
18	0.921053	0.926471	0.915698
19	0.923754	0.931953	0.915698
20	0.923754	0.931953	0.915698
21	0.922173	0.931751	0.912791
22	0.922173	0.931751	0.912791
23	0.923529	0.934524	0.912791
24	0.921944	0.934328	0.909884
25	0.921944	0.934328	0.909884
26	0.918759	0.933934	0.90407
27	0.918519	0.936556	0.901163
28	0.918519	0.936556	0.901163
29	0.916914	0.936364	0.898256

θ	F1	P	R
30	0.912072	0.93578	0.889535
31	0.911809	0.938462	0.886628
32	0.906627	0.940625	0.875
33	0.9	0.939873	0.863372
34	0.898331	0.939683	0.860465
35	0.899696	0.942675	0.860465
36	0.894656	0.942122	0.851744
37	0.898773	0.951299	0.851744
38	0.898462	0.954248	0.848837
39	0.879499	0.952542	0.81686
40	0.870662	0.951724	0.802326
41	0.861685	0.950877	0.787791
42	0.843296	0.949091	0.758721
43	0.829508	0.951128	0.735465
44	0.80602	0.948819	0.700581
45	0.805369	0.952381	0.697674
46	0.794567	0.955102	0.680233
47	0.791809	0.958678	0.674419
48	0.770833	0.956897	0.645349
49	0.76007	0.955947	0.630814

- F1 on test: 0.965
- Precision on test: 0.965
- Recall on test: 0.965

Il seguente esperimento è sui dati con esclusione dei record i cui DOI non sono in Crossref

θ	F1	P	R
0	0.966565	0.966565	0.966565
1	0.966565	0.966565	0.966565
2	0.966565	0.966565	0.966565
3	0.966565	0.966565	0.966565
4	0.966565	0.966565	0.966565
5	0.966565	0.966565	0.966565
6	0.966565	0.966565	0.966565
7	0.966565	0.966565	0.966565
8	0.966565	0.966565	0.966565
9	0.966565	0.966565	0.966565
10	0.966565	0.966565	0.966565
11	0.966565	0.966565	0.966565
12	0.966565	0.966565	0.966565

θ	F1	P	R
13	0.966565	0.966565	0.966565
14	0.966565	0.966565	0.966565
15	0.966565	0.966565	0.966565
16	0.966463	0.969419	0.963526
17	0.966463	0.969419	0.963526
18	0.966463	0.969419	0.963526
19	0.967939	0.972393	0.963526
20	0.967939	0.972393	0.963526
21	0.966361	0.972308	0.960486
22	0.966361	0.972308	0.960486
23	0.966361	0.972308	0.960486
24	0.964778	0.972222	0.957447
25	0.964778	0.972222	0.957447
26	0.96319	0.972136	0.954407
27	0.961598	0.97205	0.951368
28	0.961598	0.97205	0.951368
29	0.958398	0.971875	0.945289
30	0.955178	0.971698	0.93921
31	0.95356	0.971609	0.93617
32	0.950156	0.974441	0.927052
33	0.945227	0.974194	0.917933
34	0.943574	0.97411	0.914894
35	0.943574	0.97411	0.914894
36	0.938583	0.973856	0.905775
37	0.938583	0.973856	0.905775
38	0.938583	0.973856	0.905775
39	0.918138	0.972789	0.869301
40	0.907618	0.972222	0.851064
41	0.898693	0.971731	0.835866
42	0.880399	0.970696	0.805471
43	0.870588	0.973684	0.787234
44	0.845361	0.972332	0.74772
45	0.83737	0.971888	0.735562
46	0.822496	0.975	0.711246
47	0.819788	0.978903	0.705167
48	0.807143	0.978355	0.68693
49	0.794224	0.977778	0.668693

- F1 on test: 0.97
- Precision on test: 0.976
- Recall on test: 0.96

Considerazioni

Per entrambi i casi, la migliore soglia che è stata evidenziata è 19. Procediamo quindi ad impostarla all'interno della valutazione complessiva. Il filtraggio consente l'aumento della precision a scapito di una più bassa recall: **preferiremo quindi avere meno risultati ma più corretti.**

Risultati con filtraggio

Caso includendo anche i record i cui DOI non sono in Crossref

- Numero di volte in cui il DOI locale è corretto e quello remoto no: 16
- Numero di volte in cui il DOI remoto è corretto e quello locale no: 6
- Numero di volte in cui sono entrambi errati e identici: 13
- Numero di volte in cui sono entrambi errati ma differenti: 7
- Numero di volte in cui sono entrambi corretti e identici: 388
- Numero di volte in cui vengono restituiti supplementi dall'indice locale: 6
- Numero di volte in cui vengono restituiti supplementi dalle API: 12
- Percentuale locali sbagliati che sono supplementi (sul totale): 1.42
- Percentuale remoti sbagliati che sono supplementi (sul totale): 2.83
- Percentuale locali sbagliati che sono supplementi (sui locali sbagliati): 30.00
- Percentuale remoti sbagliati che sono supplementi (sui remoti sbagliati): 33.33
- Percentuale di corrispondenza tra DOI locale e DOI manuale (quanta - correttezza usando l'indice): 95.28
- Percentuale di corrispondenza tra DOI remoto e DOI manuale (quanta correttezza usando le API remote): 92.92
- Percentuale di corrispondenza tra DOI remoto e DOI locale (quanta corrispondenza tra l'indice e le API): 94.58
- Percentuale entrambi corretti: 91.51
- Percentuale entrambi errati e identici: 3.07
- Percentuale entrambi errati ma differenti: 1.65
- Media score nei corretti 55.90
- Media score negli sbagliati 43.26
- Scartati 6

Se sbagliano entrambi: - Percentuale entrambi errati e identici su tutti gli errati: 65.00 - Percentuale entrambi errati ma differenti su tutti gli errati: 35.00

Metriche riassuntive:

	Local	Crossref
Precision	0.954	0.927
Recall	0.939	0.918
F1	0.947	0.922

Caso escludendo i record i cui DOI non sono in Crossref

- Numero di volte in cui il DOI locale è corretto e quello remoto no: 16
- Numero di volte in cui il DOI remoto è corretto e quello locale no: 6
- Numero di volte in cui sono entrambi errati e identici: 5
- Numero di volte in cui sono entrambi errati ma differenti: 4
- Numero di volte in cui sono entrambi corretti e identici: 382
- Numero di volte in cui vengono restituiti supplementi dall'indice locale: 5
- Numero di volte in cui vengono restituiti supplementi dalle API: 11
- Percentuale locali sbagliati che sono supplementi (sul totale): 1.22
- Percentuale remoti sbagliati che sono supplementi (sul totale): 2.69
- Percentuale locali sbagliati che sono supplementi (sui locali sbagliati): 45.45
- Percentuale remoti sbagliati che sono supplementi (sui remoti sbagliati): 44.00
- Percentuale di corrispondenza tra DOI locale e DOI manuale (quanta - correttezza usando l'indice): 97.31
- Percentuale di corrispondenza tra DOI remoto e DOI manuale (quanta correttezza usando le API remote): 94.87
- Percentuale di corrispondenza tra DOI remoto e DOI locale (quanta corrispondenza tra l'indice e le API): 94.62
- Percentuale entrambi corretti: 93.40
- Percentuale entrambi errati e identici: 1.22
- Percentuale entrambi errati ma differenti: 0.98
- Media score nei corretti 55.95
- Media score negli sbagliati 48.67
- Scartati 4

Se sbagliano entrambi: - Percentuale entrambi errati e identici su tutti gli errati: 55.56 - Percentuale entrambi errati ma differenti su tutti gli errati: 44.44

Metriche riassuntive:

	Local	Crossref
Precision	0.973	0.949
Recall	0.963	0.942
F1	0.968	0.945