Evaluation of Algorithm Performance on Identifying OA

Kristin Antelman, Nisa Bakkalbasi, David Goodman, Chawki Hajjem, Stevan Harnad (in alphabetical order)*

Background

SH and CK shared four data files (Biology 2000 and 2002, Sociology 2000 and 2002) with KA, NB, & DG so they could conduct a manual check of their algorithm's coding of OA and non-OA articles. For a description of these files, see the data posted at http://www.crsc.uqam.ca/lab/chawki/ch.htm

Methodology

- The manual checking was completed during the week of March 14-18, 2005 (Biology 2002) and November 13-16, 2005 (Sociology 2000).
- The first population, ISI biology 2002, contained a total of 54,413 documents, of which 8,113 are coded as OA and 46,300 as non-OA by the algorithm.
    - We drew a random sample of 277 documents from the OA and 277 from the non-OA
    - We selected 28 records using systematic sampling (every 20th record) for our intercoder reliability test. We had a concurrence rate of 93%.
    - For the intercoder test, we used Google, GoogleScholar, Scirus, Yahoo, Alltheweb, Altavista, and eo.st. We found Yahoo, Altavista, eo.st and Scirus to be redundant to Google, GoogleScholar and Alltheweb.
- The second population, ISI Sociology 2000, contained a total of 9371 documents, of which 4405 were articles.
    - We drew a random sample of 354 articles, corresponding to 8% of the population, half being OA and half non-OA,
- Search engines included in the Biology 2002 manual test were: Google, GoogleScholar, and Alltheweb (based on results from the intercoder test, which showed that Yahoo, Altavista and eo.st were redundant). Only Google was used for Sociology 2000 since it was found from the first data set that the additional two search engines together found less than 3% additional OA.
- We coded with the algorithm result column hidden so as not to prejudice our searching.
- All searching was done off-campus.
- Each likely link (see Search Strategy, below) was examined to see if it was or led to an OA copy.

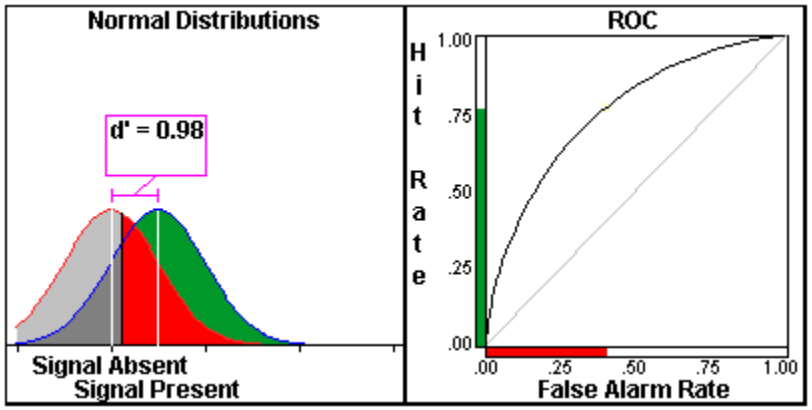
Search Strategy

- Generally, full title was used.
- When it was clear that a complete result set was not obtained (often nothing), it was frequently because of symbols, parenthetical additions, numerals vs. alpha, etc.
- In those cases, title was truncated to remove the special characters.
- Sometimes titles were not sufficiently discriminatory (e.g., no results or too many results not related to the article in question), in which case author last name may be added to the query.
- Links were grouped into types. They were checked according to the following rules ("check first time" means the first time that site is encountered by the coder). Excluded formats (where there is no or virtually no possibility of finding OA) include .ppt, .xls, .doc or .pdf cv's or reports.

Type	Action
institution/organization list of faculty/affiliate publications	check first time
institutional faculty profile	check first time
institutional bibliographic database record	check first time
publisher's journal site	check every time
full text content vendor (e.g., Ingenta, BioOne)	check if no publisher copy in results list
article citation, citation in another document	check if not enough info in results list or excluded by format
faculty cv	check if cv is html
bibliography	check if not excluded by format
non-journal site TOC	check first time
PubMed, PubMed derivative	don't check
other publisher product	don't check
departmental or organizational webspace	check
personal webspace	check
institutional project/research page	check first time
PubMed Central	check
email/listserv/blog	check

Signal Detection Analysis: Biology 2002

Decision Table - Biology 2002				
Manual Detection				
		OA	non-OA	TOTAL
Algorithm Detection	OA	108	164	272
	non-OA	32	240	272
	TOTAL	140	404	544
		Probability	Z-Score	
	Hit rate	0.77143	0.74356	
	False alarm rate	0.40594	-0.23800	
d'=z(H)-z(F)		0.98156		
ß=e -[(z(H)^2-z(F)^2)/2]		0.78027		



Hits
Correct Rejection
False alarm
Miss

ROC: Receiver Operating Characteristic
Hit rate: 0.76812
False alarm rate: 0.40100

The graph is created by using Signal Detection Theory applet developed by WISE project (http://wise.cgu.edu/sdt/sdt.html).

Interpretation of the charts – Biology 2002

The small d' (discriminability index) value (0.98) is an indication that the robot has difficulty in discriminating between the correct response (i.e., *Hit* rate) and the undesirable response (i.e., *False Alarm*). ß (decision bias) value (0.78) indicates the robot is liberal in reporting OA, in other words the robot is accepting a higher false alarm rate, while reporting the highest percentage of hits.

Discrepancy with prior manual check

Hajjem conducted a manual check of 200 records from Biology 2003 (note: not the data set used for our manual check), which found d'=2.45, ß=0.53. One possible explanation for the discrepancy, in addition to sample size, was that the first manual check was drawn from a narrow section of the data (193 records with record numbers beginning 1030, and 8 with record numbers beginning 1032). <http://www.crsc.uqam.ca/lab/chawki/validation.htm>

Findings – Biology 2002

Overall OA rate. The robot found 14% overall OA in Biology 2002. While the hand-checking was testing the accuracy of the robot and not the overall OA rate, the overall OA rate can be estimated using the errors identified in the check of the 1% sample [(missed OA error rate x coded non-OA) + coded OA – (miscoded OA error rate x coded OA)]. The estimated corrected OA rate is 16%. The OA rates between the two methods are similar because there are many more non-OA articles than OA, so the small error on missed OA cancels out the big error on overcoded OA.

Overall OAA

Our observations based on the Biology 2002 sample demonstrate that, due to the algorithm overcoding OA, the OA Advantage is underestimated. The following table shows the average number of citations per OA/non-OA article based on the manual and robot identification.

Avg. # of citations/article		
	OA	non-OA
Algorithm	0.53	0.35
Manual	0.62	0.38

We used the (OA – non-OA) / non-OA ratio to report the OA Advantage for both the algorithm and manual method.

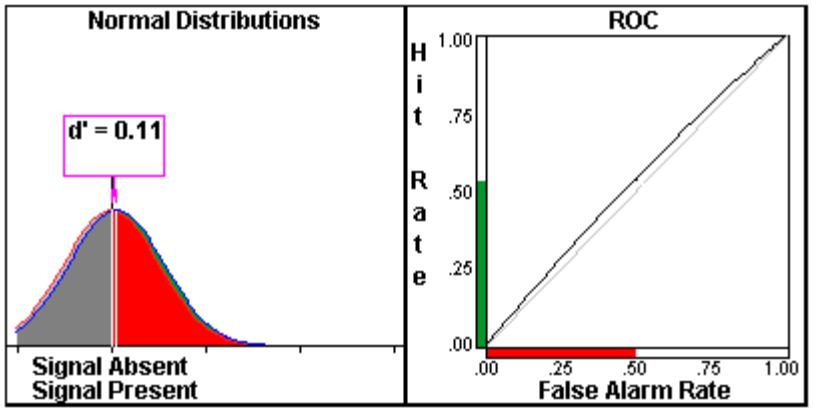
OA Advantage	
Algorithm	50%
Manual	64%

The algorithm underestimated the OA Advantage by 14 percentage points.

Signal Detection Analysis: Sociology 2000

Decision Table - Sociology 2000				
	Manual Detection			
		OA	non-OA	TOTAL
Algorithm	OA	29	148	177
Detection	non-OA	25	152	177
	TOTAL	54	300	354

	Probability	Z-Score
Hit rate	0.53704	0.09297
False alarm rate	0.49333	-0.01671
d'=z(H)-z(F)	0.10968	
ß=e -[(z(H)2-z(F)2)/2]	0.99583	



Hits
Correct Rejection
False alarm
Miss

ROC: Receiver Operating Characteristic
Hit rate: 0.76812
False alarm rate: 0.40100

The graph is created by using Signal Detection Theory applet developed by WISE project (http://wise.cgu.edu/sdt/sdt.html).

Interpretation of the charts – Sociology 2000

Sociology 2000 data indicated poor performance results for the robot. The graph on the left shows the two normal curves overlap almost exactly as the *Hit* rate and the *False Alarm* rates are approximately equal. As can be seen, when the *Hit* rate and the *False Alarm* rate are close to each other, the ROC curve tends towards a straight line indicating poor performance.

Findings – Sociology 2000

Overall OA rate. The robot found 23% overall OA in Sociology 2000. While the hand-checking was concerned with the accuracy of the robot and not the overall OA rate, that rate can be estimated using the errors identified in the check of the 8% sample. Our corrected overall OA rate is 15%, based on articles only.

Overall OAA could not be calculated for Sociology 2000 due to a technical error with the citation counts.

General Conclusions

The robot significantly overcodes for OA. In Biology 2002, 40% of identified OA was in fact OA. In Sociology 2000, only 18% of identified OA was in fact OA. Missed OA was lower: 12% in Biology 2002 and 14% in Sociology 2000.

The sources of the error are impossible to determine since the algorithm did not capture URLs for documents identified as OA.

In conclusion, the robot is not yet performing at a desirable level and future work may be needed to improve the algorithm.

---

*Author affiliations
KA: North Carolina State University Libraries < kristin_antelman@ncsu.edu>
NB: Yale University Library <nisa.bakkalbasi@yale.edu, >
DG: Palmer School of Library and Information Science, Long Island University <dgoodman@liu.edu>
SH: Institut des sciences cognitives, Université du Québec à Montréal <harnad@ecs.soton.ac.uk>
CH: Institut des sciences cognitives, Université du Québec à Montréal <Hajjem@vif.com>