A Method for Cross-collection Comparison

Sauparna Palchowdhury, Donna Harman, Ian Soboroff
Retrieval Group, ITL, NIST

Introduction

The Cranfield paradigm bounds the experimenter to one test collection (documents, topics and qrels) so that the retrieval systems running on the collections may be compared. In fields like Sociology and Medicine, Meta-Analysis is used to understand the effects of an intervention, say, a new drug or a surgical procedure. When multiple studies across space and time, using different samples of different sizes, report the effect of the same intervention, it is compelling to look at the body of evidence and not only at each study in isolation. It makes sense to synthesize the information because we assume that the studies have enough in common. It is logical to believe that the effect sizes come from a Normal distribution. Meta-analysis uses formulas that are extensions of the formulas used to generate summary statistics, like the mean and variance, *within* each study. Which means that the synthesis procedure summarizes the effects *between* studies by again computing a mean and variance.

Methodology

The goal of the synthesis is to compute a 'summary effect' described by the overall 'effect size' and its precision. Precision is measured by calculating the variance and a confidence interval. The **effect size** is a number that describes the direction and the magnitude of the effect of an intervention. It could be computed in many ways, as long as it gives a plausible indication of the effect. For example, the difference between two scores, or their ratios could be the effect size. For each study one such effect size is computed, along with a **variance**. Then a **weight** is assigned to each effect size to incorporate information about the precision of the experiment. Using the reciprocal of the variance as an indicator of precision is one good way. The **summary effect** computations may not appear intuitive. To put it simply, its a weighted mean of the effect sizes with more weight assigned to more precise studies. Its variance is the the reciprocal of the sum of the weights.

Computations

Within-study variables

Variable	$f()$	Note
$Y = f(m, m1)$	$m - m1$	Difference of means.
	$log(m) - log(m1)$	Ratio of means.
$V = f(sd, sd1)$	$((n + n1)/(n * n1)) * S^2_{pooled} + T^2$	
	Assuming the samples are independent and have the same standard deviation, a pooled standard deviation is included in the computation of variance; $S_{pooled} = (((n-1) * sd^2 + (n1-1) * sd1^2)/(n-1+n1-1))^{0.5}$	
$W = f(V)$	$1/V$	Weights for estimating tau-squared.

Between-study variables

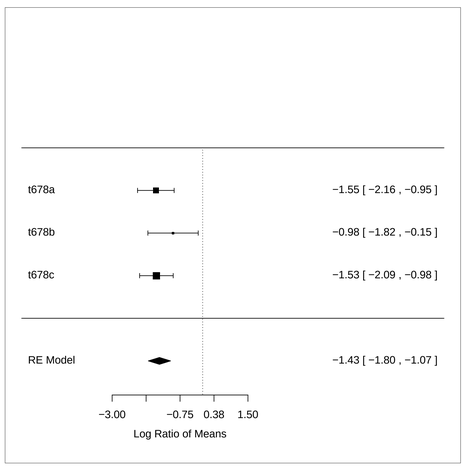
Variable	$f()$	Note
$T = f(W, Y, n)$	$(Q - (n-1))/C$	
	$Q = \Sigma(W * Y^2) - \Sigma(W * Y)^2/\Sigma W$	
	$C = \Sigma W - \Sigma W^2/\Sigma W$	
$V^* = f(V, T^2)$	$V + T^2$	The variance between studies.
$W^* = f(V^*)$	$1/V^*$	Weight assigned to each study.

Summary effect

Variable	$f()$	Note
$M^* = f(W^*, Y)$	$\Sigma(W^* * Y)/\Sigma W^*$	Overall effect size.
$V_{M^*} = f(W^*)$	$1/\Sigma W^*$	Variance of the overall effect size.
	$SE_M = V_M^{0.5}$	Estimated standard error.
	$CI_{95} = M^* \pm 1.96 * SE_M$	95% confidence interval.
	$Z = M^*/SE_M$	Z value.
	$p = 1 - \Phi(\pm\lvert Z\rvert)$	One sided.
	$p = 2 * (1 - \Phi(\lvert Z\rvert))$	Two sided.

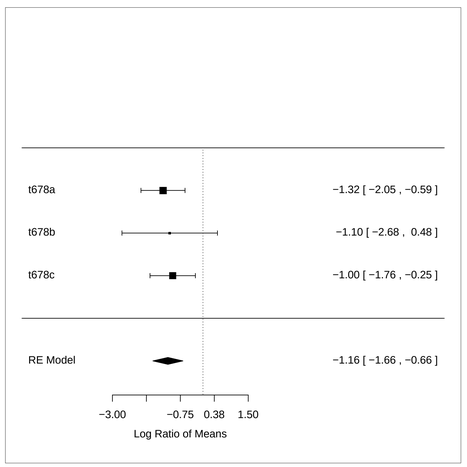
Experiments

Figure 1: **Simple BM25 vs TFIDF**



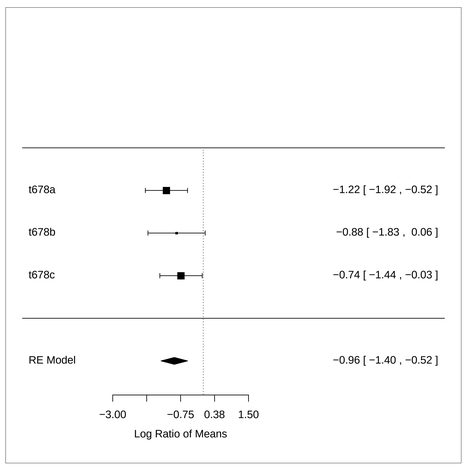
	testcol	MAP	sd	n	MAP1	sd1	n1	Y	V
1	t678a	0.1775	0.1869	30	0.0376	0.0499	30	-1.5520	0.0957
2	t678b	0.1351	0.1662	30	0.0506	0.1001	30	-0.9821	0.1809
3	t678c	0.1529	0.1428	30	0.0330	0.0410	30	-1.5333	0.0805

Figure 2: **TFIDF vs TFIDF without IDF**



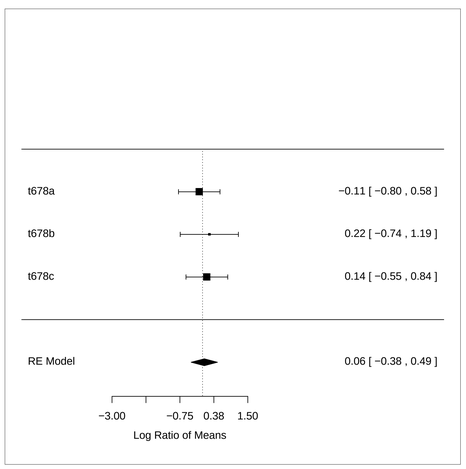
	testcol	MAP	sd	n	MAP1	sd1	n1	Y	V
1	t678a	0.0376	0.0499	30	0.0100	0.0155	30	-1.3244	0.1388
2	t678b	0.0506	0.1001	30	0.0168	0.0664	30	-1.1026	0.6512
3	t678c	0.0330	0.0410	30	0.0121	0.0205	30	-1.0033	0.1471

Figure 3: **TFIDF vs TFIDF with no length normalization**



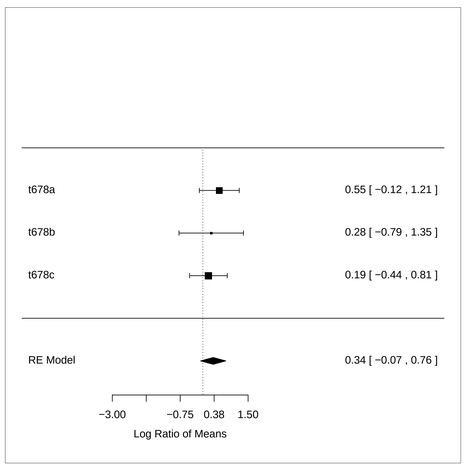
	testcol	MAP	sd	n	MAP1	sd1	n1	Y	V
1	t678a	0.0376	0.0499	30	0.0111	0.0159	30	-1.2201	0.1271
2	t678b	0.0506	0.1001	30	0.0209	0.0369	30	-0.8842	0.2344
3	t678c	0.0330	0.0410	30	0.0158	0.0240	30	-0.7365	0.1284

Figure 4: **TFIDF vs TFIDF with no length normalization and log(TF)**



	testcol	MAP	sd	n	MAP1	sd1	n1	Y	V
1	t678a	0.0376	0.0499	30	0.0337	0.0468	30	-0.1095	0.1230
2	t678b	0.0506	0.1001	30	0.0633	0.1157	30	0.2239	0.2418
3	t678c	0.0330	0.0410	30	0.0381	0.0564	30	0.1437	0.1245

Figure 5: **TFIDF; Porter stemming vs no stemming**



	testcol	MAP	sd	n	MAP1	sd1	n1	Y	V
1	t678a	0.0218	0.0280	30	0.0376	0.0499	30	0.5451	0.1137
2	t678b	0.0383	0.0856	30	0.0506	0.1001	30	0.2785	0.2970
3	t678c	0.0274	0.0334	30	0.0330	0.0410	30	0.1860	0.1010