INDEXING INITIATIVE

Summary of Threshold Studies

Clifford W. Gay
National Library of Medicine
Lister Hill National Center for Biomedical Communications

Table of Contents

Executive Summary

This summary presents just the based conclusions of the various studies reported here. Previous research has shown that the Medical Text Indexer (MTI) performs very differently on citations with and without abstracts. Therefore the studies described here were sometimes conducted on subsets of citations from MEDLINE.

MEDLINE Citations with Abstracts

Attempting to reduce the number of suggestions from the current 25 was shown to reduce MTI performance.

All MEDLINE Citations

Looking for relationships between score, rank, and correctness we examined the characteristics of true terms not included in the recommendations. For terms coming from Related Citations we found that 2/3rds of the true terms were missed. However, the rank of the lowest true term for each citation is virtually random. So there is no rank that will always include the lowest ranking, i.e. last, true term on the candidate list.

Studying term score showed that there is not safe rank to cut off scores, because too many true terms have low scores.

We also looked at all the terms lower than 25 on the candidate list. The precision for terms beyond rank 25 (at least out to 35) is never high enough to match the average precision obtained by the first 25 terms. So including any of those regardless of their score would not improve average performance.

Attempts to filter low scoring terms were not initially successful, but motivated a use of thresholds that includes the filtering of low scores for terms below a certain rank. This approach yields a 0.04 improvement in the F_2 measure for a rank of 13 and a score of 203.

MEDLINE Citations without Abstracts

Looking at low scores terms for this collection we found that terms with scores below 198 had a collective precision of 0.08 and hence might be candidates for exclusion without damaging the F_2 performance. With 16% of the true terms with scores less than 100, and 24% with scores less than 200, filtering all terms below 200 might effect the recall too much to be acceptable, so the rank-score threshold approach was tested. An optimization of those thresholds yielded a rank of 4, and a score of 46.

An analysis of false negatives in that filtering showed an error rate of 0.80 for terms recommended by both Related Citations and MetaMap, so those terms were exempted from the rank-score threshold filtering.

When the effect of title length on performance was measured it was shown that in general longer titles performed better. When partitioning the title only citations by the length of their titles was used a basis for the rank-score thresholding, the combined result was not better than apply the same thresholds to all the citations. Customized thresholds were tried on just three partitions and on a finer grained scheme with a partition for each length. Both showed tiny improvement in the individual partitions (< 1%), but only 0.0005 and 0.0002 increases over the single policy filtering.

A final experiment with partitions looked at the consequence of optimizing the thresholds with the easier to obtain micro-average performance metrics instead of the macro-averaged one that emphasizes the individual citations. The macro-averaged metrics lead to filtering thresholds that provide better performance that is more than the difference in the two averaging methods. Applied to all title-only citations the optimized thresholds become rank 10, score 190. The new thresholds increase the F_2 measure by 0.002 (.307) while increasing the number of terms filtered by 50% and just slightly raising the filtering error. A look after this filtering at the number of recommendations allowed from title only citations confirmed the current limit of 15.

Examination of the individual citations after the threshold filtering suggests that the mild improvement seen in the collective metrics are manifest in the individual citations in generally positive ways that support the filtering as beneficial. Many more citations see increases in F_2 measure than decreases. Most that had their recommendations reduced did not loose any true terms. The increase in precision was seen in many citations; the decrease in recall was limited to many fewer citations.

The studies on title length had shown it to be significant in MTI performance. So we explored the effect of the number of recommendations on the performance of MTI for groups of title only citations separated by the number of words in that title. The average F_2 measure was computed for all initial lists of recommendations and the length producing the maximum F_2 measure was identified. Although not monotonic in our sample the relationship shows that the maximum occurs for longer lists as the length of the title increases. This relationship was used to establish a step function that specifies the length of the recommendation list for any given title length.

Summary of Threshold Studies

Studies of strategies that change the length of the list of recommended MeSH terms from the Medical Text Indexer.

This summary pulls together the explorations and experiments over the last year so that have attempted to identify techniques that will improve the Medical Text Indexer (MTI) performance and reduce the number of spurious recommendations seen by the indexers.

Project Goals. This series of investigations is based on several recommendations from the indexers:

- Provide shorter list for Title only articles. (p.25)
- Make the list shorter. (p17)
- Determine the lowest ranked score for terms selected from the MTI list of suggestions, use as minimum threshold. (p29,31)

These motivated several specific project goals that are addressed by the studies reported here.

1.9.6.2 Determine list size to maximize F_2-measure.

1.9.6.3 Maximize F-measure for classes based on citation length. (by varying the length of the recommendation list.)

1.9.6.4 Determine any correlation between correctness and term score.

The list size is determined in part by the rank at which we cut off the recommended terms list. So we studied the rank and the score of terms recommended by MTI for citations with and without abstracts. Study of overall citation length will be investigated in the future.

MEDLINE Citations with Abstracts

For our first look at recommendation list length we used a large collection with 127,000 MEDLINE citations that had abstracts. Later we look at citations with only titles and mixed collections too. Earlier work found 15 to be the best recommendation list length for citations without abstracts, title only citations. So we need to look at the complement of that set.

Study of Term Rank

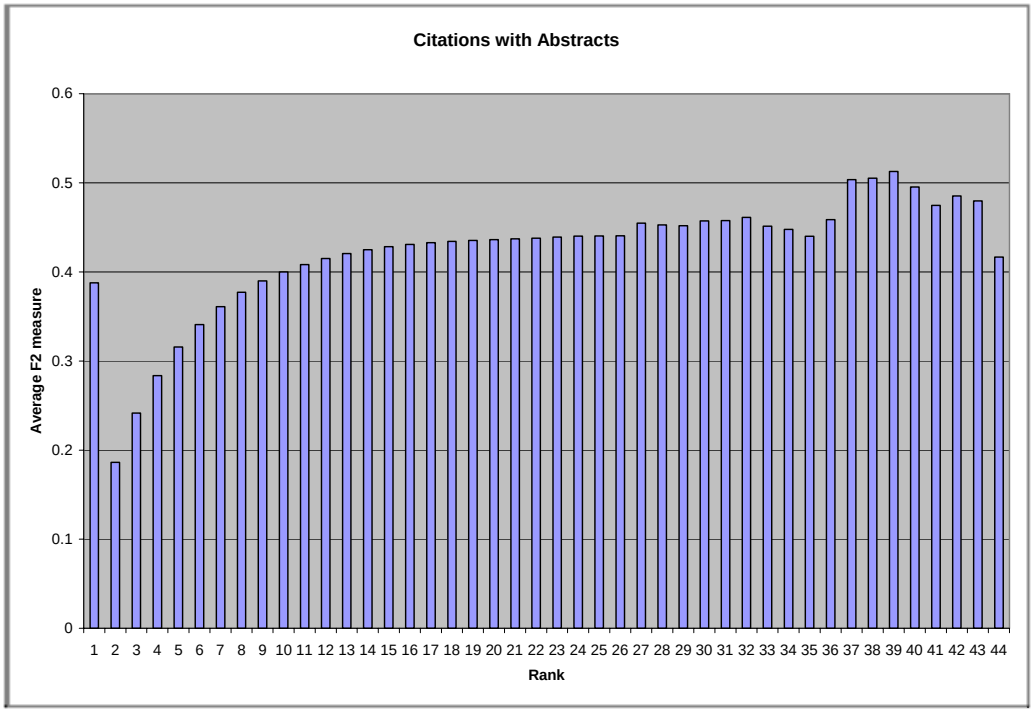

Chart 1. F2 Measure by Rank for Citations with Abstracts

Chart 1 shows the F_2 measure for the citations in the collection as we consider successively longer lists of recommendations. For example, if we cut off all the citations at rank 10, our F_2 measure for performance would be 0.40. Since the metric steadily increases to the current cut off of 25 terms, there is no obvious advantage to a shorter list for citations with abstracts. (The increases that appear after rank 25 reflect the better performance by MTI on check tags and certain look up terms, like geographical terms, that are exempt from the 25 term cut off.)

Since the curve does flatten out at the lower ranks, an experiment was performed by cutting off the terms at 18 instead of 25. This is the highest rank where the F_2-measure is with in 0.005 of the maximum at .4354. This simple filtering using a rank threshold reduced the number of suggestions by 21.8%, but only 84.6% of the excluded terms were actually 'False.' The actual performance change was a decrease of .006, which is too much to adopt as a standard policy.

These results motivate a filtering metric we use for other filtering experiments called error rate. The error rate is the percentage of false positives for the filtering, the filtered terms that actually match the MEDLINE indexing. Above the error rate was 15.4%

All MEDLINE Citations

Our general study of MEDLINE citations includes 1000 citations 602 of which have abstracts. This means that these results do not exclude the citations without abstracts studied in the next section. We looked at term rank and term score following up on the clues from the indexers.

Study of Term Rank

We first look at the terms with the lowest rank in each citation that matches the MEDLINE indexing for that citation, "term is true" for short. We examined the terms in groups determined by the indexing methods that recommended them. We separated all the terms into three classes by source: Related Citations(RC), MetaMap Indexing(MM), both (MM;RC). Since the sources do not always suggest true terms for every citation, the number of lowest true terms is less than the total number of citations. Respectively, the counts for the three classes were 963, 32, 833 lowest true terms. Initially we look at the RC results alone, the largest subset.

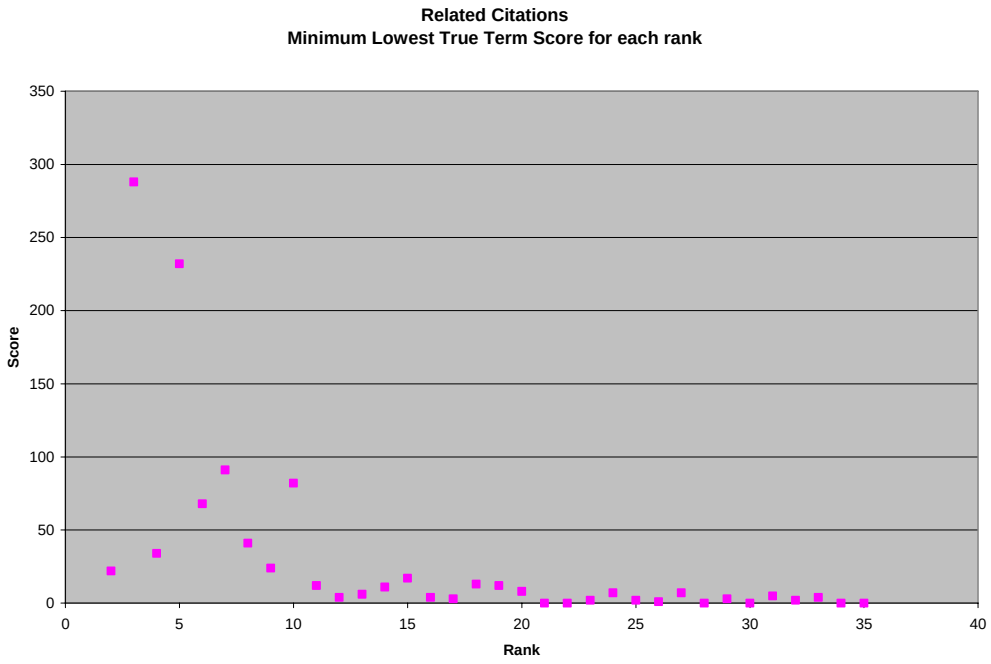

Chart 2. Minimum Lowest True Term Scores

Related Citations – Lowest True

Chart 2 shows the minimum score for the lowest true term when that term appeared at each rank. Although the chart only goes out to 35, the data extends to 113 with very

low minimums. (There are 163 citations where the lowest true term from Related Citations has a score of 0.) When for each rank we show the minimum lowest scoring true term values we see no useful threshold. Only 309 of the 963 are at rank less than or equal to 25. Therefore, over 2/3rds of these lowest true terms are missed.

Chart 3 shows the frequency with which the lowest true score is located at each rank. The charts of frequency with 38 bins (every 3 ranks) do not drop off until over 51. The rank frequency chart shows that the rank of the lowest true RC recommendation is virtually random, being evenly distributed over the full range from 7 to 55. (The highest ranks (1-5) are definitely less frequent.)

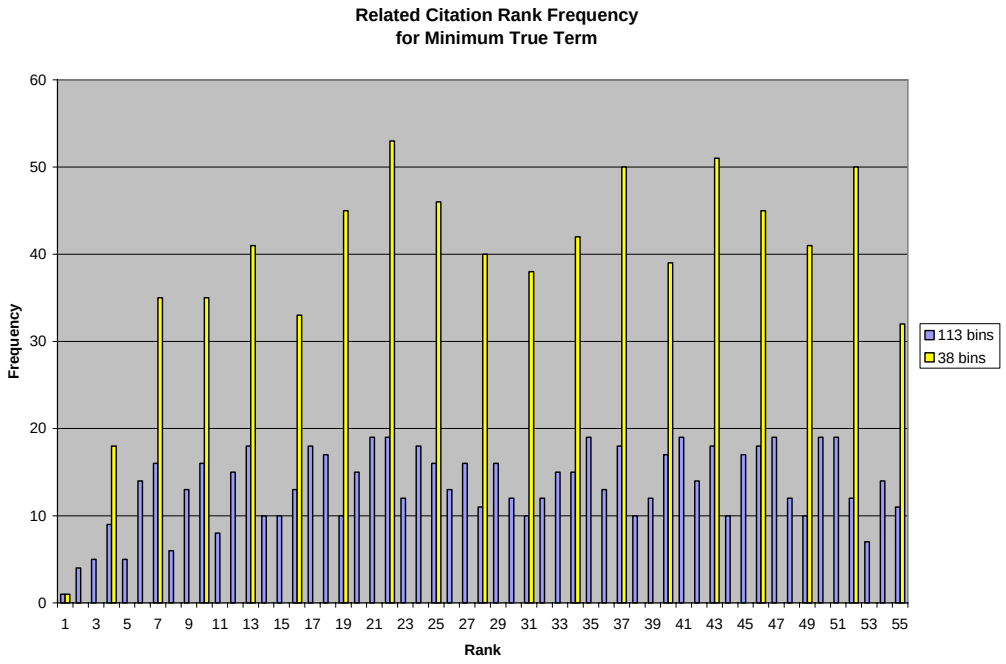

Chart 3. Lowest True Term Rank Frequency.

Thus the conclusion we come to for related citations terms we cannot pick a rank that will always include the lowest ranking, i.e. last, true term on the candidate list.

Study of Term Score

The next feature for analysis is the score of the term. We look again at lowest true terms for each citation, the precision for differing scores, and when scores might indicate true terms at ranks below 25.

Related Citations – Lowest True

Still looking at the lowest true score values, we look at their overall frequency. In Chart 4, those scores are counted for 10 point wide intervals.

For the ranks from 1-35 the most frequent lowest scores for true terms were below 50. For the 341 scores below 500 (out of 448 from rank 35 and lower) we note the following:

- 108 or 32% are below 50.
- 41 (12%) below 10.
- Median 107, Mean 155, Std Dev 148
- Clearly, the distribution is skewed to the lower end and has a large range.

CONCLUSION: There is no easy safe level to cut off scores; too many true terms have low scores.

.

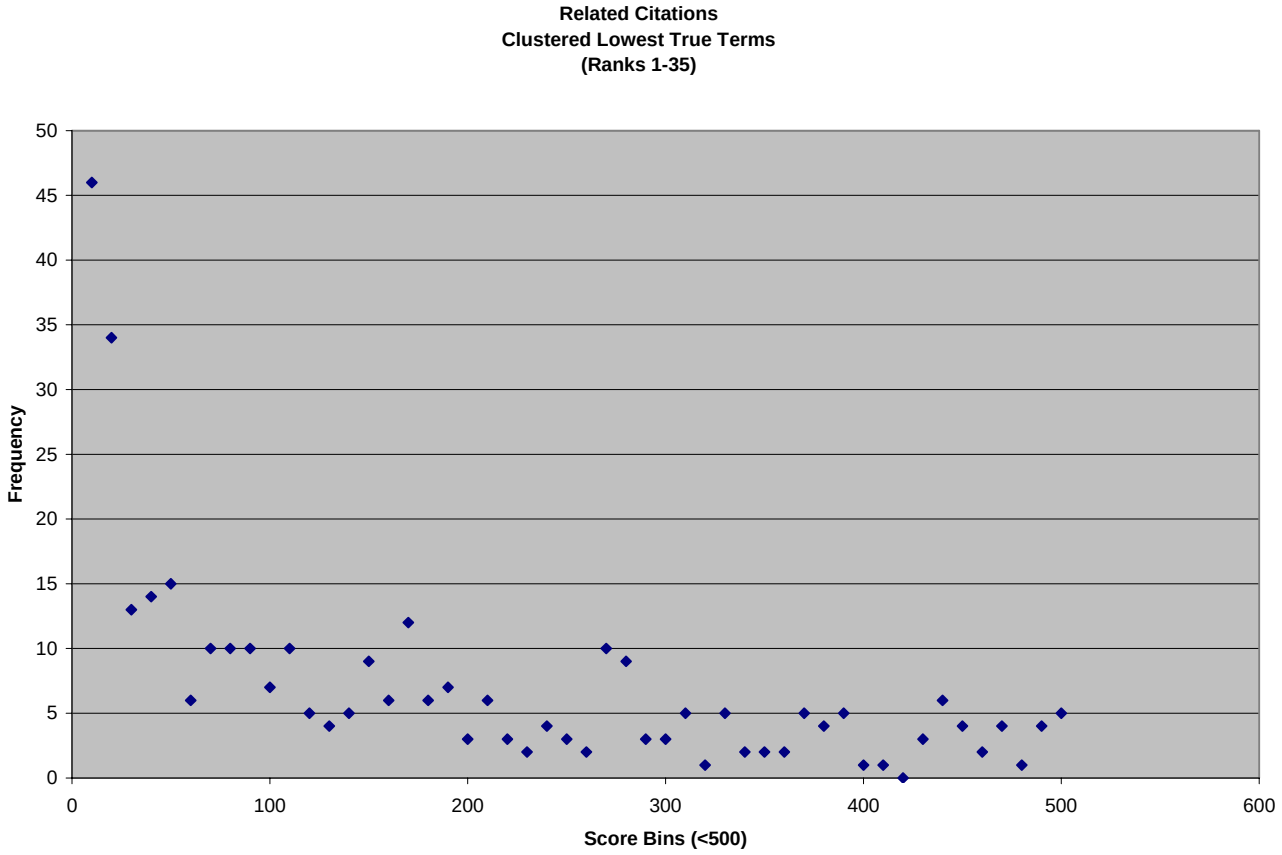

Chart 4. Lowest Score Frequency for Related Citations

Precision by Score

An examination of precision for all terms up to rank 35 showed strong relation to score values.

Hypothesis: If it is true that as with the too general terms, eliminating a class of recommendations that has a precision below 8% will not hurt the F_2 measure. We could remove recommendation with scores below 40 and not hurt the F_2 measure.

Results: An evaluation showed a slight improvement of 0.011 in the F_2 measure.

High Scoring Terms Below Rank 25

Issue: Since removing low scores high on the list improves performance maybe *adding* high scoring terms below the 25 term cut off would help too.

Can we add more terms improving recall without hurting precision?

Question: How many terms between 25 and 35 have scores over 40? If we included them would we raise precision?

Method: Tested this by adding terms with scores above 40 up to rank 35.

Results: This reduced the F-measure by .006.

Discussion: For true terms with scores over 40 in the ranks from 25 to 40 we note:

- The lowest score for each rank is near 40 except for 31 (at 56) and 33 (at 54).
- Frequency distribution for scores has median of 252 and mean of 341.

Checking the precision of higher scoring terms in ranks from 25 to 35 we get the following results:

Scores above	Precision
50	.0914
60	.0863
70	.0881
80	.0893
200	.1308
300	.1466
400	.1414
600	.1611
800	.1673
1000	1631

Table 1. Precision by Score for Low Ranking True Terms

The precision for terms beyond rank 25 (at least out to 35) is never high enough to match the average precision obtained by the first 25 terms. So including any of those regardless of their score would not improve average performance.

Filtering Low Scores

Early tests showed that terms with low scores (<140) have a precision usually less than 10%

Trials of filtering of low scores tended to reduce F_2 measure even for the best performing threshold. The performance of filtering at 100, reduced the F_2 measure by .004. Examination of the filtered true terms showed large numbers of MH-S terms. A check on their performance showed that they had a precision of .31, much better than the current value for the whole collection. With the filtering limited to Main Headings and Entry Terms, the decline in F_2 measure was reduced to .001.

The error rate for this filtering was 7.2% since of the 3747 filtered terms 270 were really true.

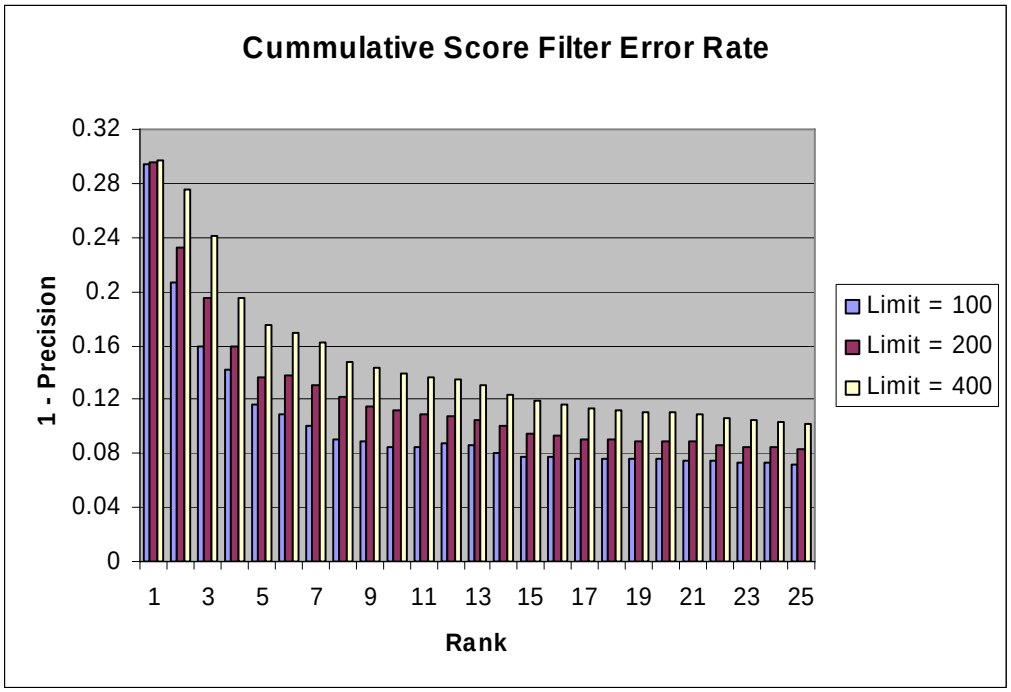

Chart 5. Error Rate by Rank and Limit

Chart 5 shows the error rate by presenting (1 - precision) as it accumulates as we move to higher ranks. As the rank goes up the error rate goes down. So if the filtering is only applied when this error rate goes below the break even point of 8% then we should be able to maintain or improve the f-measure while improving the overall precision. Notice that increasing the score threshold raises the error rate. Also we can maintain the same error rate with a limit of 400 applied at 15 and higher or by a limit of 100 applied at a rank of 5. (The perspective here is non-standard; I have treated the decision to remove a term as the goal. So precision is the ratio(number really false/number filtered) and the error rate reported is (number not really false (i.e. true)/number filtered).)

Rank – Score Filtering

The previous study motivates the strategy we implemented next: Rank-Score thresholds. Since for every score limit the error rate decreased as the rank at which it is applied decreases, we filter terms based on their score and rank.

The Limit and Rank are the controlling parameters for our selection of MTI recommendations based on thresholds. The policy leaves MH-S terms and check tags alone, but for MH and ET terms filters out terms scores below the Limit at any rank Rank or higher (that is the rank number is larger so it appears lower in the list).

When we perform a heuristic search to find the combination of parameters that will maximize the F_2 measure for the test collection we get a score Limit of 203 at a Rank of 13. At these settings there are 245 True terms out of a total of 3762 terms filtered out. (Error rate = 6.5%.) We used micro-averaged statistics during the search, but normally evaluate MTI performance with macro-averaged measures. Instead of treating the whole collection as one classification problem, we compute the statistics for each citation and take the average. Table 2, shows the results.

	Precision	**Recall**	**F_2**
Micro-averaged measures (203, 13)	.30	.49	.436
Macro-averaged measures (203,13)	.30	.50	.426
New optimum: 312,13	.31	.49	.427
No-filtering Baseline	.27	.51	.421

Table 2. Performance with Score Rank Thresholds

We have also learned in experiments with title only citations described below that we can get better performance if the search for optimal filtering thresholds is done with the macro-averaged F_2 measure. So after learning this we returned to the collection to look for new thresholds.

The new optimization gives only a .0006 increase with thresholds of score 312 and rank 13. The final filtering improvement in F_2 measure is 0.0053. Table 3 summarizes the differences between the two sets of thresholds found by different metrics. .The effect of one averaging method over the other is 43% increase in terms removed with a 60% increase in errors.

Thresholds	Decrease in Recommendations	Decrease in True terms	Filtering Error Rate	Change in Precision	Change in Recall	Increase in F_2 measure
Score: 203 Rank: 13	11%	2.7%	6.5%	+0.03	-0.01	.0047
Score 312 Rank: 13	16.3%	4.3%	7.3%	+0.04	-0.02	.0053

Table 3. Comparison of two Score Rank Thresholds

The benefit of this filtering is a reduction in the number of terms suggested because a citation with low scoring terms will have a maximum size of only 13 terms. The other benefit is the improvement in precision (4%). The cost of the trivial overall performance improvement can be seen as the 0.02 drop in recall or as 4.3% drop in true terms.

Choice of metric

If we decide to use the F_4 measure instead of the F_2 used above, then we will get different optimal setting of the thresholds (40 at 16). But these settings filter very few terms (19) and make a trivial change in the metrics over the baseline with no filtering. This lack of significance reinforces the choice of the F_2 measure as an appropriate metric for improving MTI.

MTI Enhancement

This improvement in performance is not very large but sufficient to warrant inclusion in MTI production. It is superior to just a threshold based on score, or rank alone and will be appropriate for most MEDLINE citations.

Analysis of False Negatives

To understand the effect of the filtering based on these thresholds, we examined the errors introduced by this filtering. We identify the terms which match MEDLINE indexing but are removed by this rank-score filtering. The optimal performing thresholds (203, 13) were used. Table 4 shows this analysis.

Source	Recommendations	Recommendations filtered		True	Error rate
MM	5700	6.2%	354	19	.0536
RC	17809	13.2%	2351	197	.0837
MM;RC	7738	2.0%	157	29	.1847
total	31256	12.0%	3762	245	.0651

Table 4. Error Rate by Recommendation Source

The error rate for all the subsets is below the overall precision of the recommendations after filtering (0.3045). This filtering improves the precision for all sources. Thus we can use a uniform policy for all terms.

Threshold Filtering Effects on Individual Citations

Since the margin of improvement is less than 1% and the decrease in recall is 2% and filter error is over 5%, a look at the impact of threshold filtering on individual citations is necessary to justify putting this filtering into production. This investigation seeks to find out how these changes will appear to the indexers using MTI.

F_2 Measure Changes

Here we determine what percentage of citations were actually improved by this level of filtering and how many were actually made worse. First we look at the class of changes and then at the magnitude.

	Number of Citations	Percent	Average Change
Increase	644	44.3%	0.031
No Change	600	40.0%	
Decrease	236	15.7%	-0.056

Table 5. F_2 Changes to Citations after Rank-Score Threshold Filtering.

The decrease in F_2 measure is less than it was for Title Only citations, but the average increase is smaller with more citations unchanged. Here there are almost 3 times as many increases as decreases but magnitude of the increase is a little more than half as large as the decrease in F_2 measure.

When we analyze the magnitude and frequency of the F_2 measure changes we find the following: The maximum change was -0.241, the maximum increase was 0.202. Notice that the average increase (0.031) is substantially more that the collection average of 0.005. If we ignore the unchanged citations we get a mean change of 0.008.

Correct terms Changes

As we see from table 6, 19.2% of the citations lose 1.4 true terms. Note that since we are filtering the process cannot add any true terms. Threshold filtering has a very similar effect on Title Only citations.

	Number of Citations	Percent	Average Change
Increase	0	0.0%	0
No Change	1214	80.1%	
Decrease	288	19.2%	1.4

Table 6. True Term Changes to Citations after Rank-Score Threshold Filtering.

Changes in number of Recommendations.

Table 7 presents the changes in the number of recommendations during filtering.

	Number of Citations	Percent	Average Change
Increase	0	0.0%	0
No Change	558	39.2%	
Decrease	912	60.8%	5.9 terms

Table 7 Recommendation List Size Changes after Rank-Score Threshold Filtering.

An analysis of these values shows that 40.8% of all citations had terms filtered and did not loose any true terms. (80.1% did not loose true terms, 39.2% did not loose any terms.) Note also that the vast majority of those citations have improved F2 measures; only 0.9% were not improved. 114 citations lost 12 terms each, 116 lost 9 and 76 lost 10.

The distribution of number of terms removed is bi-modal with most at either end. Of the 912 that were filtered 362 lost 1 or 2 terms, 306 lost 10, 11, 12 terms (25-13=12 the max that could be filtered.) and 244 were in between. The reason for the two groups is that the Title Only citations in this collection have a max of 15 terms and so (15-13=2) those low scoring citations just loose two terms.

Individual Citation Changes

Although the many of the citations were affected by the filtering (60.8%) usually with 6 terms removed. In only 16% of the whole collection was the recommendation list damaged and then it was 1.4 term (-0.056). 3.5% of the citations that lost good terms, also lost enough incorrect terms to leave their F_2 measure unchanged or improved. Those citations that were affected by the filtering had an average improvement of 0.008. This is probably an outcome we would favor, but is it really a better outcome than the one provided by the 203-13 thresholds?

Evaluating Filtering

When looking at the changes to the individual citations described in the last section left doubts about which of the sets of thresholds would provide the best results for the indexers, we started by improving the metrics describing those changes. When those results did not clarify the choices, other metrics and metric combinations were investigated to quantify and objectify the impression arising from the characterization of the individual citations. So what we report here is the characterization of the changes to the citations by the filtering with several sets of thresholds. Then we define the new combined statistic, explain why we think it captures the essential features of the broader look at the changes to the citations, and justifies the selection of one set of thresholds over the others.

Thresholds	203.13		312 13		350.12		100.20		100.13		305.12	
Partitions	Filtered	All	Filtered	All	Filtered	All	Filtered	All	Filtered	All	Filtered	All
F_2 **Increased**	76%	36%	73%	44%	68%	43%	80%	12%	76%	21%	70%	42%
F_2 **Unchanged**	2%	1%	1%	1%	1%	1%	0%	0%	3%	1%	1%	1%
Not Filtered		53%		39%		36%		86%		73%		40%
F_2 **Decreased**	22%	10%	26%	15%	31%	20%	20%	3%	19%	6%	29%	17%

Table 8. Filtering Partitioned Collections for Candidate Thresholds

Worse – Better Analysis

The new approach to worse-better analysis compares sets of thresholds by describing how each threshold pair partitions the set of citations into those whose F2 measure increases, decreases, or stays the same when terms are filtered, and finally those that are not filtered at all. This analysis is computed with in the context of just the filtered terms too. Other important values are the average increase among those filtered citations which improved and the average decrease among those citations that were damaged by the filtering. These values tell us more about the individual citations since the values are not watered down by all the assorted citations not affected by the filtering. A final distinguishing statistic was the midmean of the F_2 measure. The midmean is the mean of the values between the 25th and the 75th percentiles.

Table 8 presents the partitioning for all the sets of thresholds investigated for this section. Table 9 presents some basic collective metrics as well as the magnitude of the changes to those citations that are in the *improved* or *damaged* partitions. The best value for each metric is indicated by the color red.

	Baseline	203.13	312.13	350.12	100.20	100.13	305.12
Mean F_2 Change	0.42141	+0.00523	+0.00464	+0.00483	+0.00097	+0.00244	+0.00525
F_2 Midmean Change	0.42860	+0.00289	+0.00122	+0.00904	+0.00994	+0.01157	+0.01410
F_2 Median	0.4286	0.4318	0.43690	0.4321	0.4310	0.4310	0.43210
Average F_2 Increase		+0.032	+0.029	+0.038	+0.023	+0.026	+0.037
Average F_2 Decrease		-0.05591	-0.05601	-0.058	-0.057	-0.058	-0.058

Table 9. Threshold Metrics

The Filter Metric

As the tables above show these sets of thresholds although they have many similarities including having F_2 measures above the baseline have a broad range of effects on the individual citations. One filters terms from 64% of the citations and another only 14%. One damages only 3% of the citations, while another damages 31%. These thresholds were all selected based on outstanding performance on one metric on another, which should be used for a production version of MTI? The answer proposed here is a new metric that combines the three most representative metrics. This section defines this filter metric, motivates its use through a description of the path to its discovery, and

shows the results of it application to selecting an optimal set of thresholds for rank-score filtering of MTI indexing.

The Definition of the Filter Metric

The filtering metric combine three indicators of quality for MTI indexing: accuracy of the filtering, the error rate, and the F_2 measure. To properly define these metrics we need first to go back to the confusion matrix from a retrieval task. If we view the filtering as part of the task to select terms that match the MEDLINE indexing terms for an article, then the recommended terms belong to the 'positive' set and those not selected, here those removed by the filtering, belong to the 'negative' set. Those terms in the MEDLINE indexing are 'true' terms; terms not in the MEDLINE indexing are 'false.' This allows us to define a four way partition of all the MTI recommendations before the filtering. See figure 1.

	Recommend	**Filtered**
True	True positive	True negative
False	False positive	False negative

Figure 1. Confusion Matrix for MTI filtering

For accuracy we use the standard definition:

$$Accuracy = \frac{TP + FN}{TP + TN + FP + FN}$$

The Error Rate introduced previously tries to capture a recall like metric for the filtering process: the ratio of the true terms lost and the total terms filtered out.

$$ErrorRate = \frac{TN}{TN + FN}$$

The filtering metric then combines these two statistics with the F_2 measure by turning each into a percentage change from some baseline in order to normalize the values. The normalization allows the statistics to contribute equally to the new Filter Metric. The F_2 measure, F, is compared to the before filtering performance. For Accuracy, A, and Error Rate, E we use mean values for collection of results we are evaluating.($\overline{A}, \overline{E}$.)

With the statistics now normalized to [-1,+1], we apply factors to better equalize the magnitude and range of the values so that all the input statistics matter in the result. These weighting factors were empirically determined and evaluated: $w_F = 20$, $w_E = 2$, and $w_A = 1$.

$$FilterMetric_i = w_F \times \frac{F_i}{F_{baseline}} + w_E \times \frac{E_i}{\overline{E}} + w_A \frac{A_i}{\overline{A}}$$

Problems with the Components Alone

This metric was discovered by successive attempts to use the component metrics alone and then in other combinations. This sections outlines this discovery and thereby the motivation behind this metric.

F2 Measure. When 312.13 thresholds were found with a slightly larger increase in F_2 measure over the baseline MTI, it effects on the individual citations was analyzed and reported above. When compared to the previously considered optimal thresholds, 203.13. The issue of which would truly be better for the indexers was unclear. Based on the data in tables 8 and 9 here is a comparison of those two sets of thresholds:

- 203.13
  - More of the filtered citations are improved (+3%)
  - Fewer of the filtered citations are damaged (-4%)
  - Fewer citations overall were damaged (-5%)
- 312.13
  - More citations over all were improved (+8%)
  - Overall mean F_2 is higher. (+.0006)
  - The midmean F_2 is higher (+.00167)

So with the available metrics no clearly superior set of thresholds is evident. In particular the F_2 measure is not completely convincing on its own when the difference is so small. Table 3 also presents another metric: error rate for the filtering. The error rate for 312.12 is 7.3% and for 203.13 it is 6.5%. The advantages of 203.13 seem to reflect this difference in the accuracy of the filtering. Perhaps that should be the metric we optimize on.

Filtering Accuracy. When we computed the conventional accuracy statistic for both sets of thresholds we found that the two metrics have different ranking for the two sets of thresholds. 312.13 has the better accuracy at 0.417 versus 203.13 at 0.377. So we searched for thresholds from previous trials with high accuracy and an F_2 close to the other candidates since F_2 will always be important. The set of thresholds with an F_2 measure above 0.426 with the highest accuracy was a rank of 12 and a score of 350. Table 10 shows the F_2 and other metrics for the top scoring candidates.

	F2 Measure	Accuracy	Error Rate
Top For Each Metric			
600.13.	0.4240	0.4835	0.0895
100.20	0.4224	0.2983	0.0599
For F2 > .426			
350.12	0.4262	0.4424	0.0776
203.13	0.4261	0.3765	0.0651*
The old optimal sets:			
312.13	0.4267	0.4171	0.0727

Table 10. Threshold Filtering Metrics

A comparison of the 350.12 thresholds to the others using the values in tables 8 and 9 yields these observations:

- Has best midmean for F2: .00618 higher than 312.13
  - Lowest for rate of improved filtered (-5%
  - Lowest overall improved (-1%)
  - Highest damaged filtered citations (+5%)
  - Highest all citations damaged +5%
  - Between others for Overall mean
- Has the largest average increase but also the largest average decrease in individual F2 measures.

Only one of these is a observation showing 350.12 to be a superior performing threshold. The worse/better analysis of 350.12 shows that a threshold chosen by just filtering accuracy does not give a threshold that would be judged better by those other metrics.

Error Rate. When accuracy did not seem particularly useful, we looked more closely at the thresholds with the best error rate score: 100.20. This time a F2 minimum was not applied.

After the worse/better analysis the following observations are possible:

- The best filtering improvement overall 80% (+.04)(point difference)
  - Lowest overall improvement rate (-.25)
- Lowest harm rate for filtered citations (-.02)

- Middling overall harm
- Minusculely larger increase in midmean from 350.12 but over 3 times the increase of 312.13
- Overall F2 mean is equivalent to baseline (+.00097)
- Has lowest average increase and a little higher average decrease in individual F2 measures.

So those results look pretty good so we might be able to find similar filtering accuracy but better F2 performance. An additional 74 new trials were run. The thresholds tested used scores from 90 to 210 usually in increments of 10 for most ranks from 10 to 20. When 100.13 showed promise additional nearby trials were run. At this stage 100.13 has an F2 measure of 0.4239 a wee better than 100.20 which is still in third place for error rate after all the new trials.

A close look at 100.13 using again the data from tables 8 and 9 we observe the following:

- The F2 improvement is half of the others. (0.0024 v 0.0052, 0.0046)
- Midmean improvement is best of all tested (4x 312)
- Filtered improved - close to the best (203)
- Collective improvement - less than contenders. (-.23 -.13):
- Damage very low similar to 100.20,
- Less damage than better performers (203 -- filtered -0.03, collective -0.04)

Now this is a very good showing and we would begin to think this would be the threshold set to put into production. However, it only affects 22.5% of the citations so although its filtering is very good its F2 measure improvement is very small. Some combination of these metrics may be necessary.

Besides picking a set of thresholds for this collection, this investigation was trying to establish a metric to support F2 in selecting or accepting either enhancements with close results or small improvement. To evaluate the metric we used the multiple indicators of the worse/better analysis. However, we seem to be able to measure quality, but still not choose which of two sets of values is the better set.

The competing priorities are the need to filter and the need to filter correctly. Generally the more accurate filtering occurs when there is less filtering. Filter error actually takes this into consideration as the ratio of the errors to the number actually removed. The original confusion arose from the 203.13 being better for just the filtered terms, but 312.13 did better for the overall improvement in individual counts and average F2. May be regular accuracy is helpful. It takes all the errors into account, the negatives and positives.

Metric combinations. During the analysis of the candidates and the individual metrics, several combinations were explored but each seemed to be completely or nearly

dominated by one of the metrics in the combination and did not really add much. So it was clear that to come up with a useful metric that included all three metrics would require some form or normalization instead of the raw difference in scores used initially. Rate of change was natural approach to normalize the F2 measures. For the other metrics to take the same approach required a baseline value and the mean of each metric for all of the trials, the population were evaluating, seemed appropriate. This maps each metric to the range of [-1, 1]. So a simple sum of the results was used.

The Filter Metric Analysis

So when the new filtering metric is applied to all of the trials, one of the previously evaluated candidates appears at the top: 203.13. Table 11 shows all the candidates, their scores for each of the component metrics and the combined filtering metric. It also lists their rank among all 127 trials for the filtering metric. The red color indicates the leader among these candidates for that metric, not the maximum for metric.

Thresholds	F2 change	Normalized Accuracy	Normalized Error Rate	F2 + Error	Accuracy + Error	F2 + Accuracy + Error	Rank
203.13	0.0111	0.0401	0.2028	0.4307	0.2483	0.6790	#1
305.12	0.0125	0.1816	0.1024	0.3533	0.2840	0.6373	#8
312.13	0.0125	0.1521	0.1166	0.3675	0.2687	0.6363	#9
100.13	0.0059	-0.0942	0.2799	0.3979	0.1858	0.5837	#31
350.12	0.0114	0.2221	0.0571	0.2843	0.2792	0.5634	#40
100.200	0.0023	-0.1758	0.2721	0.3189	0.0963	0.4152	#78

Table 11. The various metric combination results for candidate filtering thresholds

This ranking very closely maps to the impressions left after the worse/better analyses, that although 312.13 had the best F_2, that the lower error of 203 made it better. Then when error was emphasized and 100.13 was found, its contribution seemed lacking. So we looked at accuracy. Accuracy led to 350.12, but its error rate made it hard to accept. A simple sum of the three metrics yielded 305.12 as the best candidate. But like its neighbor 312.13 is does too much damage to too many citations. -Since range of values for the favored candidates is much larger for accuracy and error rate than for F_2 a rational combination might weight $F_2$2 more. (Ratio of the maximum values .0125 0.2799 for F_2 and error the value 20 was chosen.) This weighting still chose the same candidate. If we want to keep away from decreases in F2 we have to emphasize Error. The factor on F_2 was kept to balance Accuracy and Error is emphasized with a factor of 2. The resulting combination of all three does not replicate the order for the candidates for any of the statistics or their combinations. This indicates that the new metric reflects merits of the candidates from all of the statistics.

Among these top six candidates the 203.13 set of thresholds is 4th place for F_2 improvement and filtering accuracy. It is 3rd place for low filtering error. It is 1st place for a combination of F_2 and Error and 4th in the combination of accuracy and error. The new metric picks the set of thresholds that filter nearly half of the citations (47%),

so it has an impact on the collection. It improves 76% of those citations thus causing harm to only 10% of the citations in the collection. It still improves the overall performance of the collection and is only 0.00002 less than the best overall performing candidate (305.12).

Production System Testing

With rank-score threshold filtering implemented on the production MTI we tested the selected parameters to confirm their setting. The policy is to filter all terms at or below rank 13 that has a score less than 203 if the term is MH or ET but not MH-S.

MH-S Terms

The recommendations from MTI that are marked MH-S are special terms added to the list by certain post processing rules that select these terms when certain trigger words are found in the text. Previous experience showed that filtering these terms was not helpful. We checked this observation with an analysis of the current data from the baseline run. 63 out of 177 recommended matched Medline indexing terms. (Precision = 0.356) 11 have a score of 0 so the negative effect of any threshold would be severe. (52/166 = 0.313) So the policy to not filter these terms is still a good one.

Baseline Results

Our collection, Medline indexing and baseline MTI recommendations come from the regular production evaluation from April of 2006. The primary performance metrics are shown in the top line of Table 12. The baseline precision is higher with the current production system than it was with the experimental baseline: 0.29 vs. 0.27, so the F_2 is also higher.

Thresholds	Recom.	True	Citation Affected	True IM	Prec	Recall	Used	Prec IM	Recall IM	Used IM	F_2
baseline	32056	9304	0	4153	0.29	0.51	6.20	0.14	0.79	2.77	0.4278
14.190	29207	9105	403	4101	0.30	0.50	6.07	0.15	0.78	2.73	0.4312
13.190	29051	9093	413	4099	0.31	0.50	6.06	0.15	0.78	2.73	0.4314
14.203	29042	9095	416	4099	0.31	0.50	6.06	0.15	0.78	2.73	0.4315
13.203	28873	9082	419	4097	0.31	0.50	6.05	0.15	0.78	2.73	0.4317
14.225	28763	9068	454	4094	0.31	0.50	6.05	0.15	0.78	2.73	0.4314
12.203	28684	9058	459	4088	0.31	0.50	6.04	0.15	0.78	2.73	0.4313
13.250	28217	9011	492	4080	0.31	0.50	6.01	0.15	0.78	2.72	0.4313
12.250	27966	8977	542	4066	0.31	0.49	5.98	0.15	0.77	2.71	0.4306
11.250	27746	8958	542	4061	0.32	0.49	5.97	0.15	0.77	2.71	0.4307
12.312	27157	8895	631	4052	0.32	0.49	5.93	0.15	0.77	2.70	0.4305
11.350	26402	8825	669	4032	0.33	0.49	5.88	0.16	0.77	2.69	0.4309

Table 12. The various metric combination results for candidate filtering thresholds

Experimental Results

All of the experimental trial results are reported in Table 12. The trials are listed in the descending order of the number of citations that had terms removed. (This is almost equivalent to the number of terms filtered out.). Some of the highest values are shown in red. If we note that the difference for these sets of thresholds in F_2 from the lowest to the highest is only 0.0008, we see again why the combination of filter metrics is appropriate.

Table 13 shows the various filter metric results with the combination metric in the rightmost column. The definition of the metric is given on page 12. The best performing set of thresholds, (rank 13, score 203), affected 31% of the citations in the test collection.

Thresholds	F2 change	Normalized Accuracy	Normalized Error Rate	F2 + Error	Accuracy + Error	F2 + Accuracy + Error
14.190	0.0079	0.0212	0.1541	0.3131	0.1754	0.4885
13.190	0.0084	0.0327	0.1497	0.3180	0.1824	0.5004
14.203	0.0086	0.0338	0.1603	0.3332	0.1941	0.5273
13.203	0.0091	0.0463	0.1554	0.3377	0.2016	0.5394
14.225	0.0084	0.0534	0.1321	0.3004	0.1855	0.4859
12.203	0.0082	0.0585	0.1165	0.2802	0.1750	0.4552
13.250	0.0082	0.0909	0.0757	0.2394	0.1667	0.4060
12.250	0.0065	0.1068	0.0318	0.1627	0.1386	0.3013
11.250	0.0068	0.1226	0.0278	0.1634	0.1505	0.3139
12.312	0.0063	0.1629	-0.0110	0.1152	0.1518	0.2670
11.350	0.0072	0.2163	-0.0259	0.1190	0.1903	0.3093

Table 13. The various metric combination results for candidate filtering thresholds

Conclusions

There are some observations about the filtering and the metrics worth noting:

1. The expected pattern of more filtering implies better precision and declining recall.

2. The filtering effects the IM terms less than the NIM terms because the IM terms are usually among the highest scoring terms and are thus less likely to be affected by our algorithm.

3. Although in the experimental system the best performing set of thresholds was not selected by the combination metric, for these trials the 13,203 set has the best F_2 .

So for the selected thresholds we have a +0.0039 (0.91%) improvement in F_2. Essentially, no improvement, but the goal was merely to maintain the F_2. The filtering causes a 9.9% reduction in the number of recommendations. It increases the overall precision by 0.02 (6.9%) and reduces overall recall by 0.01 (2.0%). This filtering makes a helpful improvement by reducing the likelihood of a blooper while only losing about one good term for one in seven citations. Therefore, including rank score filtering in the production MTI system is worthwhile.

Medline Citations without Abstracts

Most of the work that follows uses a test collection of 42,371 MEDLINE citations that have only titles and no abstracts. MTI was unable to provide any suggested MeSH terms for 3112 of these citations. On the other end of the performance scale there are 10 citations with F_2 measure of 0.93 that all have 9 words in the title. (precision: 8/11, recall: 8/8)

This section discusses earlier work with title only citations, new work on true term score properties and rank-score threshold based filtering. It also presents a study of the effect of title length on MTI performance.

Previous Results

F_2 Measure Study

The original work on Title Only citations looked at the F measure for various sized recommendation lists. It established the basic maximum length recommendation list at 15. This yields a macro-averaged F_2-measure of 0.3020. The mode for the F_2 measure distribution is 0.3846 for 953 citations.

Some consequences of this policy are reflected in the test collection and the

C recommendation sets evaluated here:

- max number of recommendations is 27
- mode for num of recommendations is 15 for 21307 citations
- only 430 citations have 1-14 recommendations (~1%)

Term Score

Previous studies of lowest scoring correct terms, or scores of lowest ranked true terms did not yield useful heuristics for selecting better performing sets of terms.

Term Score Revisited

Again the over all goal of the study was to find properties of term scores for the citations with out abstracts that will help predict good recommendations. Those

recommendations from MTI that match the human indexing in MEDLINE we refer to as 'true terms' the true-positives in our performance analysis.

To this end we look into the distribution of term scores, high scoring terms, low scoring terms, and precision.

True terms – False terms

We first collected the distributions for the true and false terms to see if there might be some cross over point.

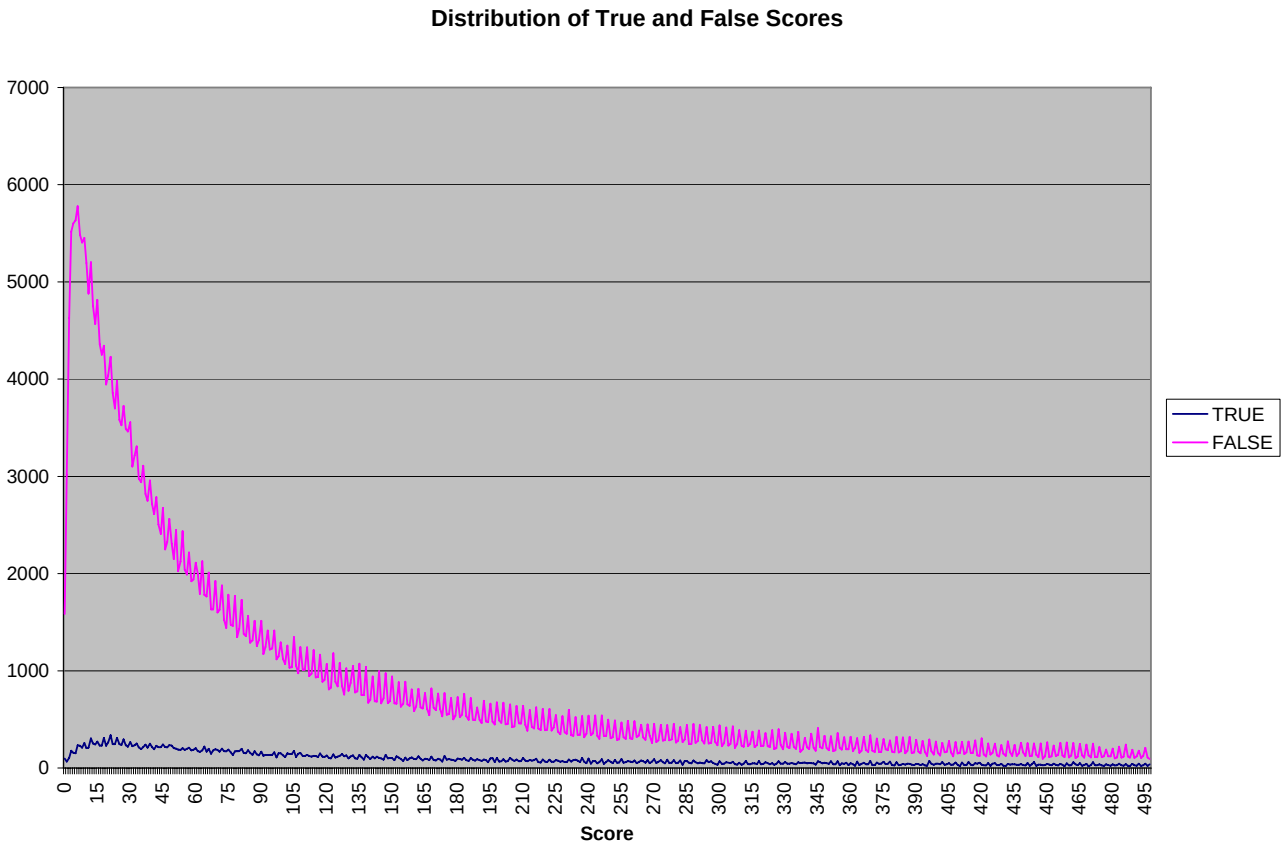

Chart 6. Distribution of True and False scores for Title Only citations.

Chart 6 shows the simple frequency of scores for the class of True terms and the class of False terms.

Half of the trues are 1000 or less and more than half of the false are less than 100. However, 16% of the Trues are also less than 100.

The shape of the curve suggests that for some low scores the number of false terms is many times the number of true terms and filtering below some threshold is likely to improve MTI performance.

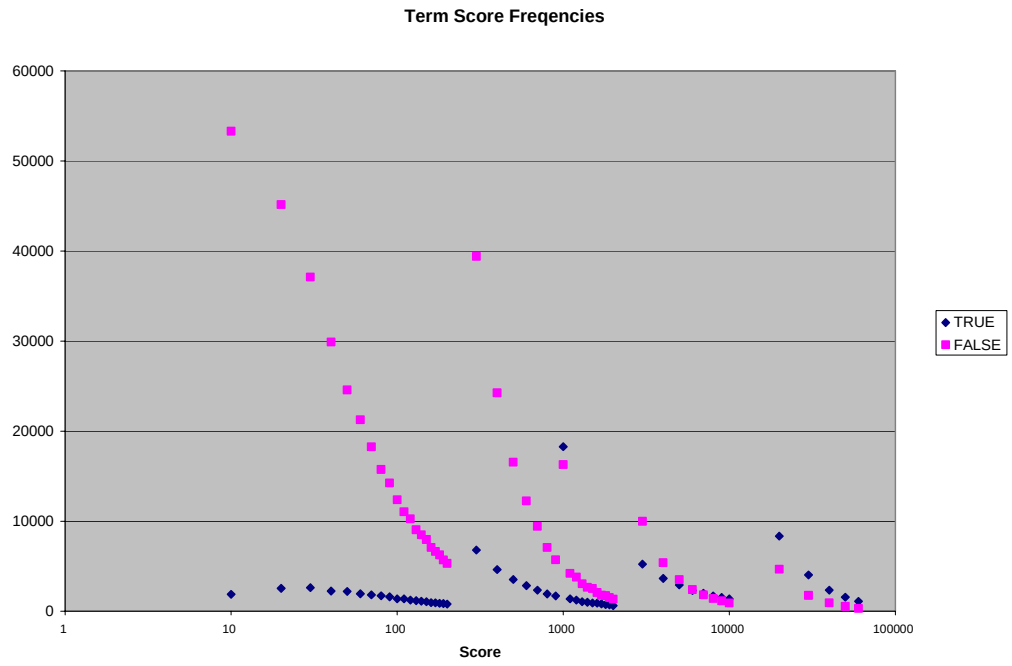

Chart 7. Term Score Frequencies for Title Only Citations.

Chart 7 shows this data accumulated in increasing large buckets (10, 100, 1000, 10,000). (The transitions in the bucket size cause the discontinuities in the curves.) The frequency of both true and false terms decreases as the scores rise, but every where the decrease of false terms is steeper than for the true terms. The cross over point, the score for which the number of true terms with that score does not exceed the number of false terms, is not until the score reaches 6000.

The upper limit of the false terms is close to that for true terms. There are five false terms over a million. So there is no score threshold that guarantees truth. One outcome is that there is a large spike of both true and false terms at 1000 (T-17K, F-11K). This anomaly at 1000 is due to the assignment of 1000 to all checktags that are triggered by other terms.

High Scoring Terms.

To look at the performance of the highest scoring terms, we count the cumulative number of terms with scores from the highest scores down. When the number of True terms with that score or more equals the number of False terms with that score or more is the cross over point we seek. This is the point where filtering out all lower scores would give us a precision of 50%. That score is 1345.

Term Precision v. Term Score

Chart 8 gives a detailed look at the precision and term score. Precision was calculated for range in the partitioning of scores used in chart 7. Some points of interest are marked in yellow on the chart. The special point to the left is at (3000, 0.34) and marks a sudden increase in the slope of the line. The one to the right is the cross over point described above.

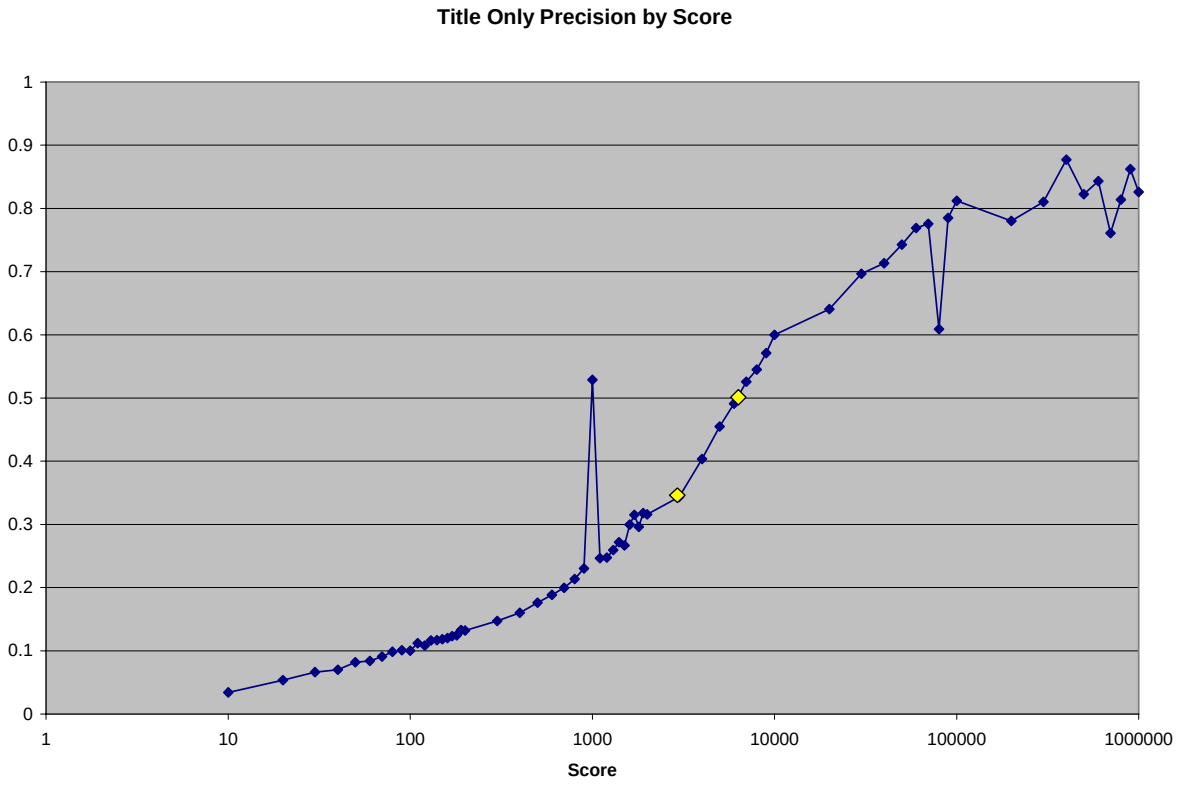

Chart 8. Precision by Score for Title Only Citations.

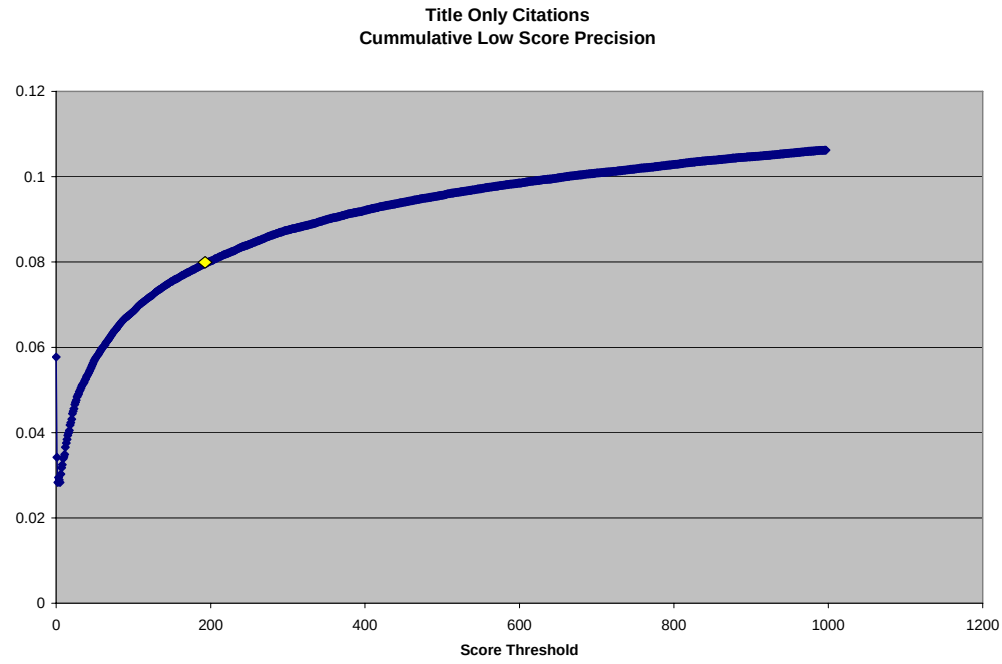

Chart 9. Cumulative Low Score Precision for Title Only Citations.

Low Scoring Terms

Chart 9 looks at the performance of low scoring terms. At each score threshold we calculated the performance for terms with that score or less. In previous studies we have shown that if a class of terms is removed from the suggestion list, the MTI F_2 measure performance will remain constant if the performance of that class on its own is less than 0.08. The chart marks that threshold at a score of 198.

Conclusions

With 16% of the true terms with scores less than 100, and 24% with scores less than 200, filtering all terms below 200 may effect the recall too much to be acceptable. Since there are no other new clues from features of the score data, we will apply the rank-score thresholding to the title-only citations as we have done with citations with abstracts.

Rank – Score Filtering

Initial Results

Table 8 shows the results of applying filtering based on rank-score thresholds. The baseline performance for this collection was an F2 measure of .3300. For this experiment three maxima were found, one each at ranks 4, 5, and 6. If we were going to use these results for production we would probably choose the one with the highest precision, and with the tie below, the fewer filtered true terms: 41, 5. This is an improvement of 0.0078

Score	Rank	Filtered Trues	Precision	F_2
55	**9**	6,829	.2215	.3371
45	**8**	6,701	.2218	.3375
66	**7**	10,235	.2356	.3373
45	**6**	8,291	.2285	.3378
41	**5**	8,137	.2285	.3378
36	**4**	7,688	.2263	.3378

Table 8. Maxima Thresholds.

Finding this ridge of maxima and wondering what the shape of the performance might be, we investigated which thresholds would simply maintain the baseline performance. Table 9 shows the results for a broad range of score values.

Score	Rank	Filtered Trues	Precision	F_2
600000	**11**	10,469	.2251	.3300
3600	**9**	16,806	.2501	.3300
470	**8**	18,985	.2617	.3300
265	**7**	20,037	.2689	.3300
200	**6**	20,879	.2747	.3300
7	**5**	21,252	.2773	.3300

Table 9. Thresholds Maintaining Baseline Performance.

There were insufficient criteria to select threshold for the title only citations, so we moved on to an investigation of title length with the expectation that a more significant improvement or clearer patterns might emerge.

Baseline Filtering

Several changes and error corrections had been applied to MTI, so the title only test collection was reprocessed. One of the changes included outputting a record in the detailed output mode for the citations for which there were no recommendations. These new records allowed the evaluation program that computes the MTI evaluation measures to recognize the missed indexing. This in turn reduced the recall and the F_2 measure. 10 shows the new baseline performance and the performance at the former optimal parameters and the new best thresholds for those ranks. Note that a true maximum emerges with this run and those thresholds were one of the previous optimal parameters.

	Score	Rank	Filtered Trues	Precision	Recall	F_2
No Filtering Baseline				.1881	.3575	.3027
Old Max 6	45	6	7,350	.2310	.3366	.3081
New Max 6	32	6	5,306	.2209	.3424	.3083
Old Max 5	41	5	7,266	.2307	.3368	.3080
New Max 5	28	5	4,973	.2190	.3434	.3083
Old-New Max 4	36	4	6,802	.2283	.3382	.3084

Table 10. New Baseline and Title Only Thresholds

Previous run for these parameters showed improvement of 0.0078 over a. 0.330 baseline. Here the maximum improvement is .0057. This is threshold filtering (4 36) results in a 17.5% decrease (116600/667250) in recommended terms with a filtering precision of 95.9% (111807/116600).

Analysis of False Negatives

When we look at the errors in the rank-score filtering for citations with abstracts we did not find a substantially better performing source. The errors for titleOnly filtering are shown in the table.

Source	Recommendations	Recommendations filtered		True	Error rate
MM	87,854	2.7%	2,380	60	2.52%
RC	602,176	26.7%	160,489	7,453	4.64%
MM;RC	70,864	0.3%	217	175	80.64%
total	760,896	21.4%	163,086	7,688	4.7%

Table 11. Error Rate by Recommendation Source

Since this 0.8 result for the error rate is way over our average precision (.2283) we will make MM;RC terms immune to the filtering.

Title Length

First we look at the distribution of title length, and later at its effect on MTI performance with the goal of optimizing the content of the recommendations list.

Title Length Distribution

Chart 10 shows the distribution of title lengths from the Title Only citations.

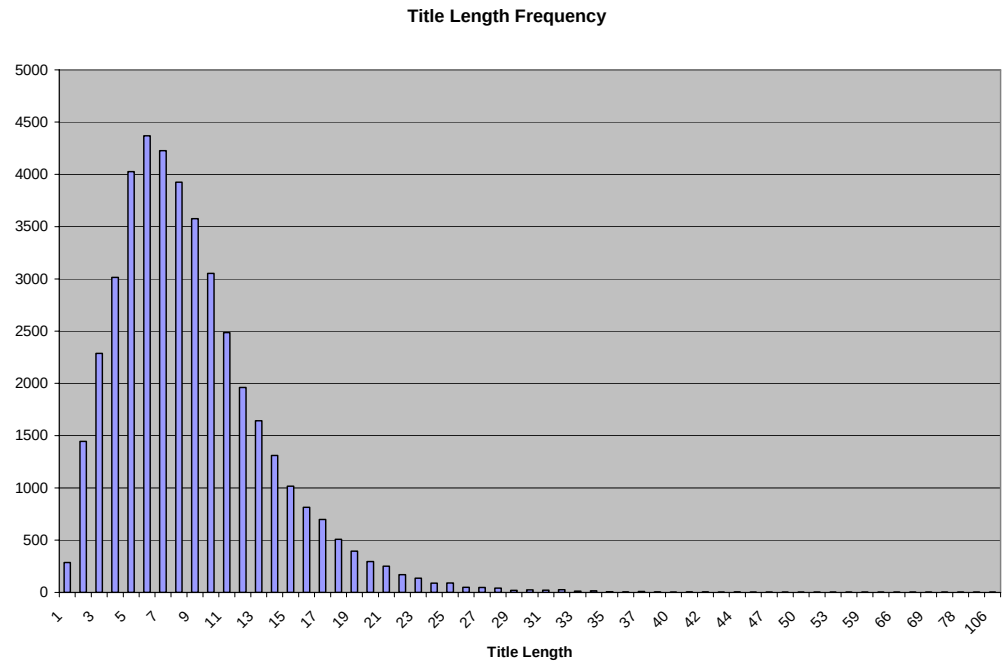

Chart 10. Title Length Frequency for Title Only Citations.

Some observations include:

- max length is 106
- mode length is 6 with 4369 citations
- median length is 8 (56% 1-8)

F-measure - Title Length

Chart 11 shows the F_2 measure for each title length and the F_2 measure accumulated for all of the citations with titles up to each length. Chart 12 focuses on the region containing 99.8% of the citations. It shows the average F_2 measure for citations of a given length (The great variability in the F_2 measure for the longer titles reflects the small frequency at those lengths.)

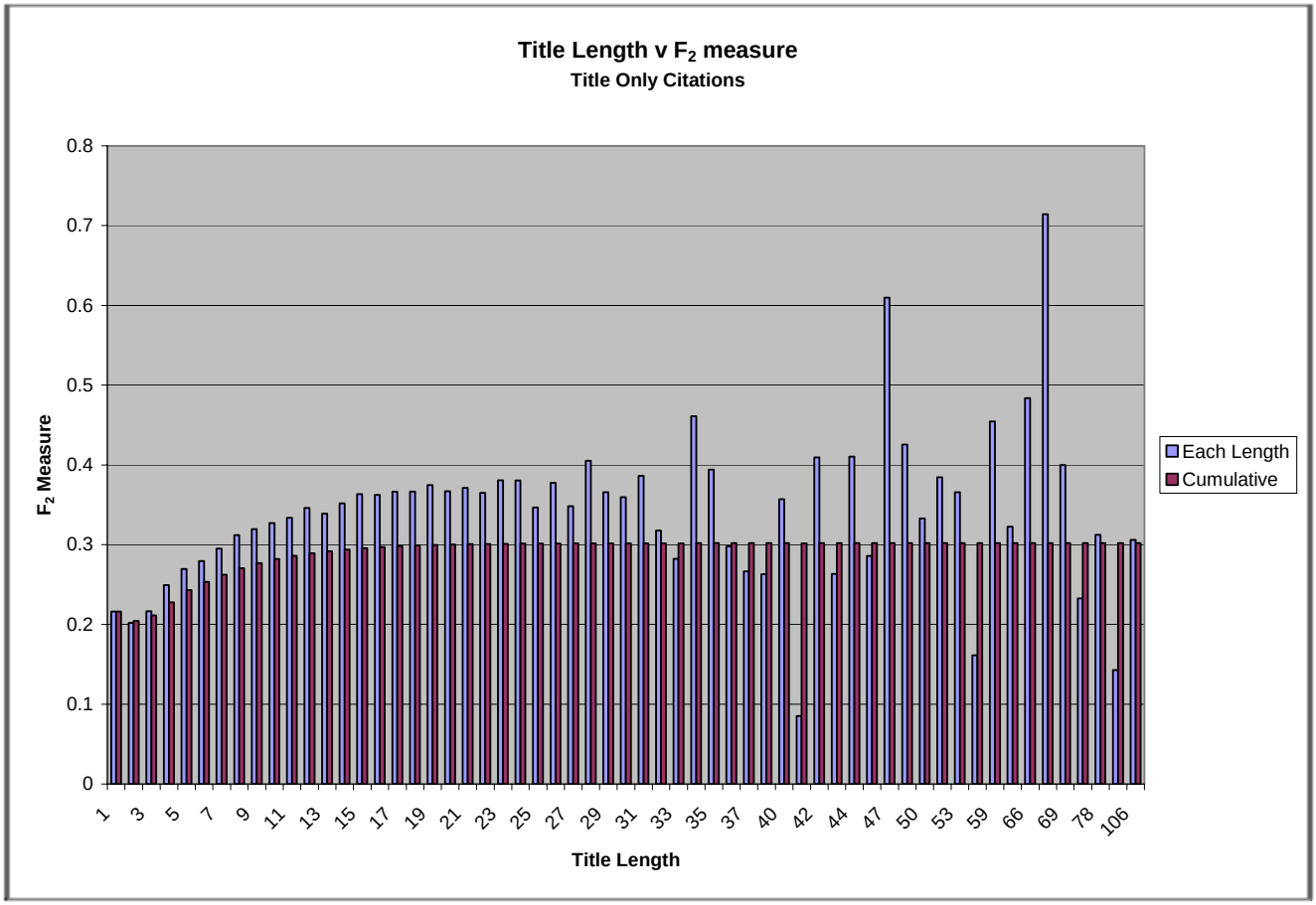

Chart 11. F_2 Measure and Cumulative F_2 Measure by Title Length for Title Only Citations.

Observations:

- Average F_2-measure generally increases as the title gets longer except for a dip at length 2.
- Maximum F_2-measure is 0.93 for 10 instances all with 9 words in title (8/11 prec, 8/8 recall)

- Cumulative F_2-measure reaches its maximum of 0.3019 at 34 words
- Maximum IM F_2-measure is 0.3846 for a couple singletons at 68
- Length 2 is the lowest with .2043

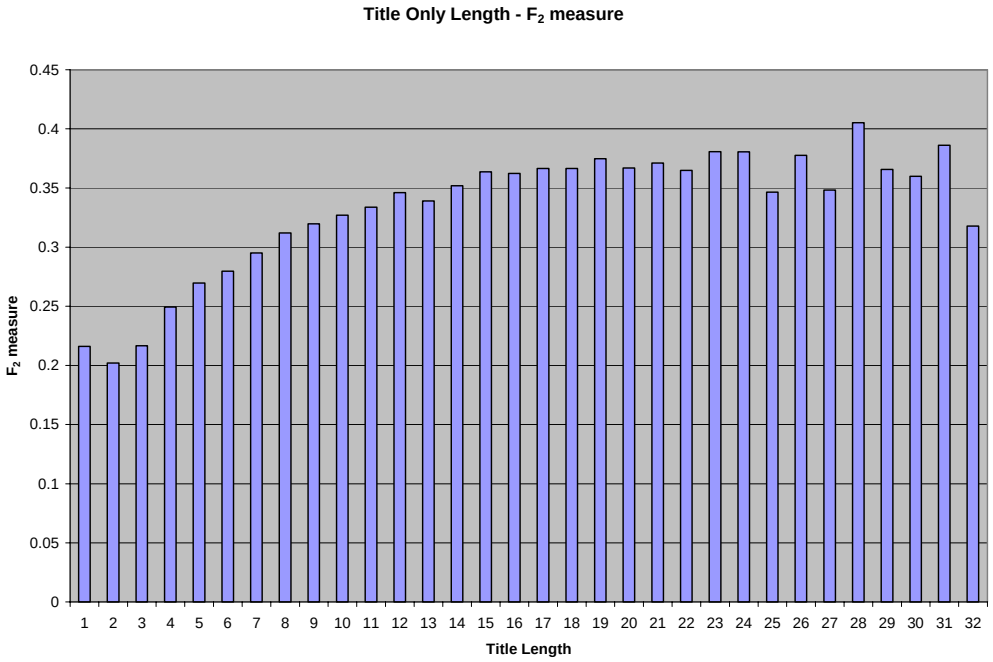

Chart 12. F_2 Measure for most Title Lengths for Title Only Citations

- F2 measure crosses average F2-measure for all titles at length 7.
- Half of the F2 measure maxima occur at length 8 or below.

(Only 75 citations longer than 32 words - so the chart is cut off there)

Discussion: Maximizing rank-score filtering over titles with the core length frequencies might give clearer results than before. Longer titles should do better than average without special thresholds. Shorter ones may need special policies.

On the premise that maximizing score-rank filtering might be more conclusive for the middle sized titles. With the mode for length at 6 and median at 8 we use 6-10 for the middle range. This gives three partitions with the following counts: 11,000 - 19,000 - 12,000. This makes the outer ranges and the middle range about equal in size.

The upper range will be treated normally since the F2 measure for length 11 is 0.3337 already slightly above the average and getting better without filtering.

Rank-Score Filtering

Following the scheme outlined above the Title Only citations were divided into three partitions and the optimal rank-score thresholds were determined for each. Table 12 shows those results and the baselines for each partition before the filtering. The Short

partition has title with 1-5 words (11,060), the Middle partition has titles with 6-10 words (19,148), and the Long partition has titles greater than 10 words (12,162).

	Score	Rank	Filtered True	Filtered	Precision	Recall	F_2
Short Partition		(+.0076)					
baseline					.1419	.2929	.2411
best	21	4	1627	47619	.1829	.2735	.2487
Middle Partition		(+.0041)					
baseline					.1864	.3602	.3034
best	39	6	3115	62628	.2222	.3403	.3075
Long Partition		(+.0028)					
baseline					.2322	.4023	.3508
best	55	8	1271	22949	.2561	.3909	.3536

Table 12. **R**ank-Score Filtering for three Title Length partitions.

The following results from the Short partition show the wide range of thresholds with very similar performance.

Rank	Score	F_2
4	21	2487
5	40	2486
3	18	2485
7	28	2483

When we combine the optimally filtered partitions, the combined result is no better than the baseline filtering for all citations with the same thresholds.

Policy	F_2 Measure	Difference
Baseline (no filtering)	.3027	
Thresholds Only	.3084	+.0057
Partitioned	.3079	+.0052 (-.0005)

Table 14. Overview of Rank-Score Filtering for Title Only Citations.

Partition	Score	Rank	Not Filtered	Filtered	Diff
Short titles	21	4	.2411	.2487	+.0076
Middle titles	39	6	.3034	.3075	+.0041
Long titles	55	8	.3508	.3536	+.0028
All titles	36	4	.3027	.3084	+.0057

Table 15. Summary of Partition based Score Rank Filtering.

Table 15 summarizes the improvements achieved by the partitions separately. Perhaps since the largest improvement was in the smallest partition and the only one with an improvement larger than the improvement for the collection. 16 is similar but shows the performance of the partitions at the thresholds optimized for the full collection too.

Partition	F_2 Measure (36 – 4)	F_2 Measure (Optimized)	Diff
Short titles	.2474	.2487	+.0013
Middle titles	.3071	.3075	+.0004
Long titles	.3529	.3536	+.0007

Table 16. Comparison of Partition Performance for Locally and Globally Optimized Thresholds.

The performance of the partitions individually at the optimal values for the whole collection are always less than their own individual values. Why the collection does better with the common thresholds is unclear.

Although each of the partitions showed better performance individually and the performance for each at the full set parameters were in fact less than their individual optimums, the single policy did better overall.

Fine Grain Partitions

The divergence of both the thresholds themselves and the amount of improvement the filtering brought suggests that perhaps the current partitions are inappropriate clusters. So we looked into finer grained partitions. The short partition was split in to a separate partition for each title length. Optimal thresholds were found for each yielding the following F_2 measures and improvements (Table 17).

Only two of the five partitions have improvements larger than one partition, single policy filtering. When the partition was split in to a separate partition for each title length, the collective performance was just marginally better (.0002) than for the Short partition as a whole. So it was unlikely finer partitions would improve performance for the whole collection.

Partition	F_2 Measure (Optimized)	Improvement
One	.2280	.0091
Two	.2096	.0084
Three	.2220	.0073
Four	.2559	.0097
Five	.2724	.0056
Collective	.2489	.0078
Short Titles (single policy)	.2487	.0076

Table 17. Fine Grain Partition Results

Micro-averaging v. Macro-averaging.

In the preceding experiments on filtering with rank-score thresholds, the optimal thresholds were found using saved results and by computing macro-averaged statistics. (The search for optimal thresholds takes 30-40 trials and this approach made the search tractable.) The collection or partition is treated as a single classification problem (micro-averaging). However, our normal evaluations of MTI performance treat each citation as the classification problem and the results for the collection are the average of the results for the individual citations (macro-averaging). This approach is taken to emphasize the quality of each citation. Because the effects of the filtering are subtle it may be that the averaging technique is hiding significant differences. So we took a look at the micro-averaging results for significant thresholds identified above (Table 10).

	Precision	Recall	F_2
No-filtering Baseline	.19	.38	.302
Rank Score Filtering(4 36)	.23	.36	.304
Rank Score Filtering (6 32)	.22	.36	.305
9 Partitions	.22	.36	.305

Table 18. Micro-averaged Performance for Title Only Citations

Table 18 shows the macro-averaged performance levels achieved with no filtering, two former optimal threshold pairs, and with a new fine partitioning scheme.

To check the premise that the micro-averaging might also effect the value of the optimal thresholds a former set was also evaluated. Notice those thresholds that were not optimal in the micro-averaged trials give a better F_2 measure than the formerly best performing thresholds.

To check to see if a partitioning scheme would show significant improvement with the macro-averaging an evaluation with new set of partitions was performed. It used the finest partitions available: individual lengths for titles from 1-7, and clusters for 8-10 and >10. It used the optimal thresholds determined for each with the micro-averaging technique. The partitioning scheme is within the margin of difference around the multiple set of thresholds with the best performance. At the resolution of the table there is no significant difference. Thus once again the partitioning based on title length did not improve MTI performance.

The rank 6, score 32 threshold has the best F_2 measure (.305 v .304) and a better filtering precision 95.77% (117542/122738) v. 95.46% (140230/146889) It filters 4.1% of the recommendations.

Conclusion: So for now the recommended filtering for Title Only citations changes to a rank score threshold of 6 and 32. However, we have cast doubt on whether this is the optimal value for the threshold filtering.

Rank Score Thresholds Reconsidered

Since the thresholds are set by looking at the micro statistics but the ultimate evaluation is with macro-averaged statistics, reflecting our goals to optimize citation by citation performance, we will try to maximize one partition with a micro evaluation.

We will test this first on a single partition to verify its usefulness before trying the more computationally demanding full collection. We start with partition Six because it is the largest partition. If we do not get more than 1% improvement, then there will be no point in pursuing length anymore.

Partition Test:

Partition Six	**Precision**	**Recall**	**F_2**	**Diff**
No-filtering (Micro-averaged)	.1666	.3352	.2785	
No filtering (Macro-averaged)	.16	.36	.280	
Filtered (6 31) Micro-averaged	.2031	.31671	.2850	+.0065
Filtered(6 31) Macro-averaged	.20	.34	.2841	+.0041
Filtered (8 130) Macro-averaged	.23	.32	.2874	+.0074

Table 19. Macro v. Micro-averaged Performance for Partition Six Citations

Table 19 summarizes the results of trials with macro and micro-averaged performance measures. Notice that the difference in the two baseline F_2 scores, between macro and micro, is just 0.015. After filtering using the macro-averaged score selected thresholds

(6, 31) the difference in the increase is 0.0024, with the micro average showing a bigger improvement over the baseline. But when we optimize the thresholds using the macro-average scores we get a bigger improvement with the macro-averaging, but more significantly we get an 0.0033 improvement in the macro-averaged F_2 measure we would have had without this change in evaluation measure. The increase over the baseline for the threshold filtering goes from 1.5% to 2.6%.

The initial search for optimal thresholds for the Six partition gave interesting but useless results. The search with just 3-digit F_2 values found a broad range of thresholds which all gave the same maximum: 0.287: For a rank threshold of 7, scores from 72 – 248; for a rank of 8, scores from 100 – 340; and for a rank of 10 any score over 110; all yielded the same F_2 measure. Even with 4-digit F_2 measure a tie resulted that was broken by choosing the threshold pair with the lowest filtering error rate.

Full Title Only Collection

Comparing micro and macro-averaging showed the need to optimize with macro-averaging so we did that next for the full collection. All of the performance results reported in this section are based on macro-averaging.

Title Only Citations	Precision	Recall	F_2	Term Reduction	Filtering Error
No-filtering	.19	.38	.3021		
Micro-average based Thresholds (6 32)	.22	.36	.3050	18%	4.23%
Macro-average based Thresholds (10 190)	.23	.35	.3071	27%	5.51%

Table 20. Macro-averaged Performance for Title Only Citations.

This re-optimization was able to increase precision (+0.03 -> +0.04) and with the same size deduction in recall (-0.02 -> 0.03) resulting in a larger (by 0.002) improvement in the F_2 measure (+0.005) over the no-filtering baseline. The filtering error rate is (9923/180053) 5.51% for a 27% decrease in the number of recommendations. So the recommended filtering threshold is rank 10, score 190.

Recommendation List Length

The current limit of 15 non-check tag recommendations for title only citations was based on experiments that calculated the average F_2 measure for each initial list of recommendations. Fifteen was the length that provided the maximum average F_2 measure. Since the threshold filtering has changed the recommendation lists we repeated this study to find the current optimal recommendation list length for title only citations. The new results showed that the F_2 measure reached it maximum (.3071) at 13. But the curve levels out and this turns out to be the same performance as the 15 limit.

Threshold Filtering Effects on Individual Citations

We have shown that for averages based on the individual citations, i.e. micro-averaging, we can filter heavily without hurting MTI performance. This investigation seeks to find out how these changes will appear to the indexers using MTI.

F_2 Measure Changes

Here we determine what percentage of citations were actually improved by this level of filtering and how many were actually made worse. First we look at the class of changes and then at the magnitude.

	Number of Citations	Percent	Average Change
Increase	25691	60.6%	0.037
No Change	8205	19.4%	
Decrease	8477	20.0%	0.087

Table 20. F_2 Changes to Citations after Rank-Score Threshold Filtering.

The size of the average decrease is 2.3 times the average increase. This almost balances there being 3 times as many citations with increases. The extra increase is the slight improvement seen in the collective F_2 measure.

When we analyze the magnitude and frequency of the F_2 measure changes we find the following: The maximum change was -0.4047; the maximum increase was 0.1880. The midrange is 0.108 showing a concentration of smaller changes. The midmean of 0.0150 shows that half of the citations improved three times the amount suggested by the mean alone. The mean change is 0.005066 with standard deviation of 0.055 - consistent with collective statistics.

Correct terms Changes

As we see from table 20, 20% of the citations lose one true term. Note that since we are filtering the process cannot add any true terms. Basically this says that 20% of the citations loose a true term.

	Number of Citations	Percent	Average Change
Increase	0	0.0%	0
No Change	33881	80.0%	
Decrease	8492	20.0%	1.2

Table 20. True Term Changes to Citations after Rank-Score Threshold Filtering.

Changes in number of Recommendations.

Table 21 presents the changes in the number of recommendations before and after filtering.

	Number of Citations	Percent	Average Change
Increase	0	0.0%	0
No Change	5287	12.5%	
Decrease	37086	87.5%	4.86

Table 21 Recommendation List Size Changes after Rank-Score Threshold Filtering.

An analysis of these values shows that 67.5% of citations that had terms filtered and did not loose any true terms. (80% did not loose true terms, 12.5% did not loose any terms.) Note also that the vast majority of those citations have improved F2 measures, only 6.9% were not improved. 2 citations lost 15 terms each, 24 lost 9 and 27 lost 12.

Looking at the frequency of differences in the number of suggestions between the baseline and the 190,10 filtering we get

- Range 0-15
- Mode is 5 for 30218 citations (71.3%)
- Next most frequent change (6) for only 3.5%.

Individual Citation Changes

Although the vast majority of the citations were affected by the filtering (87.5%) usually with 5 terms removed. In only 20% of the whole collection was the recommendation list damaged and then it was only 1 term (-0.087). Those that improved had smaller gains (+0.037) but there were many more of them. This is probably an outcome we would favor. The increase in precision seen in the collective statistics is manifest among many citations and the decrease in recall is isolated in a much smaller number (1/4) of citations.

Summary

Examination of the individual citations suggests that the mild improvement seen in the collective metrics are manifest in the individual citations in generally positive ways that support the filtering as beneficial. Many more see increases in F_2 measure than decreases. Most that had their recommendations reduced did not loose any true terms. The increase in precision was seen in many citations; the decrease in recall was limited to many fewer citations.

Title Length and Number of Recommendations

Here we explore the effect of the number of recommendations on the performance of MTI for groups of title only citations separated by the number of words in that title.

For other groups of citations we have measured the cumulative F_2 measure at every rank to find the list size that yields the maximum F_2 measure for that collection. Looking at which rank each citation reaches it maximum did not supply any patterns

suggesting an optimal length. However, taking the average of the individual citation F_2 measures at each rank will give the overall maximum F_2 measure.

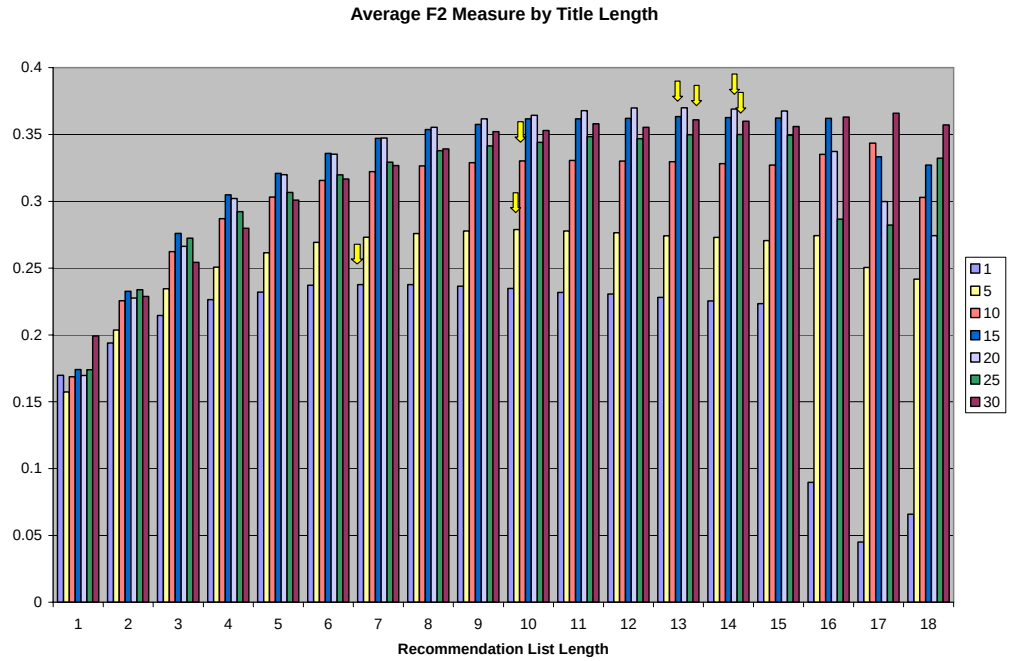

Chart 13. Average F_2 Measure at each List Length and Title Length

The chart 13, Average F_2 Measure at each List Length and Title Length, shows for some title lengths how the F_2 measure increases but reaches a maximum usually before the cut off for main headings at 15 recommended terms. (The arrows show the maxima.)

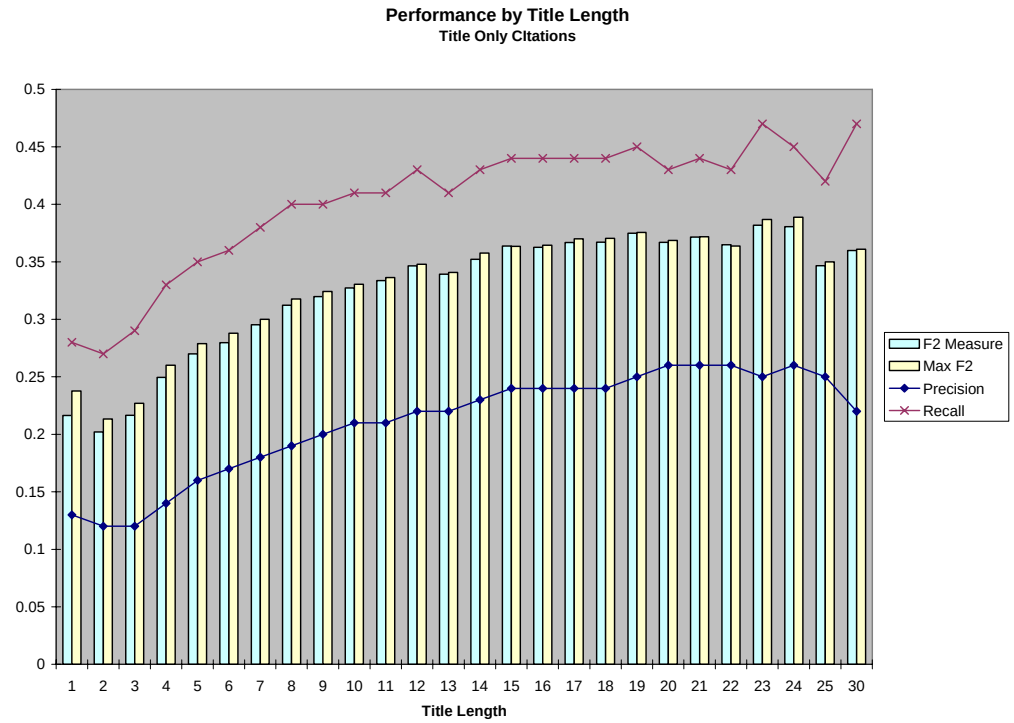

Chart 14. MTI Performance by Title Length for Title Only Citations.

The Performance by Title Length chart (Chart 14) shows that the F_2 measure generally increases as the title length increases. It also shows that if we shorten the list to the length where the maximum appears we get better scores than the current performance. (Yellow bars.) The average improvement up to a title length of 21 is 0.0052. However, we can expect a larger improvement because the technique of this evaluation usually removes the check tags which appear at the end of the list. Now MTI 's performance on check tags is about twice that of its overall performance. So, if the check tags and geographic terms are added on to the list of best performing main headings, the overall performance should be even better. (Note in general how the F_2 measure tracks between the recall and precision but closer to recall.)

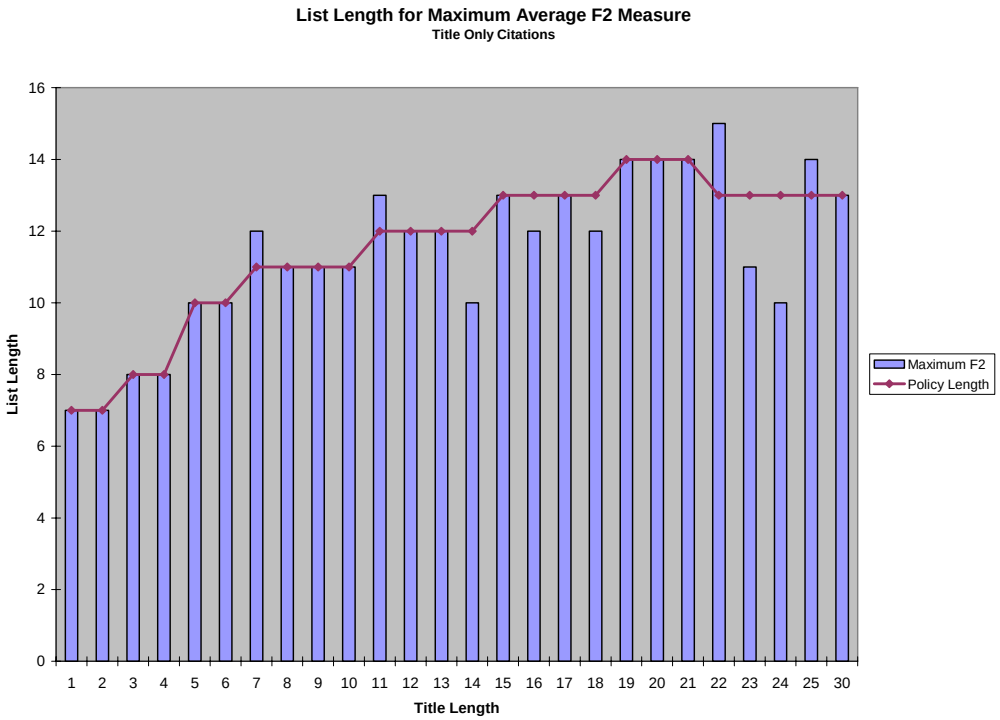

Chart 15. The List Length for Maximum Average F_2 Measure.

Chart 15, The List Length for Maximum Average F_2 Measure, shows a generally steady increase in the list length where the maximum appears as the length of the title increases. The proposed policy setting the length of the main headings recommendations list would be:

Length of Title	Length of MH List
1,2	7
3,4	8
5,6	10
7-10	11
11-14	12
15-18	13
19-21	4
over 21	13

The results for the links from 22 to 30 are erratic 15, 11, 10, 14, 13. This is probably due to the small counts (<170) for those categories. Data were available for all citations over 10, but that partition is dominated by the 15 and under citations. So I ran a trial for the over 21 group finding their max at 13. With a count of 785 this should be a reliable value in the population at large. The red line in the chart shows the

recommended list lengths for each title length. For all title lengths that are not at their maximum values, the performance at the proposed value is still better than the current performance.

Follow up: This policy for list length based on title length needed to be confirmed for the whole collection together. Other partitioned policies investigated before have not resulted in overall improvements. So a version of MTI was build that implements this policy as an option. This option and the rank-score filtering option are evaluated in the next section.

Evaluation of two filtering options

This section presents the results of a fresh evaluation of the two proposed filtering options: rank-score threshold filtering and limiting recommendation list based on title length. A baseline performance was established and the options were tried separately and together.

Environment: The version of MTI was a new one with full implementation of both options. The same title only test collection was used as before, but now the indexing was with 2004 data. The collection was also updated with the indexing from then current 2004 MEDLINE. To provide a stable platform for these evaluations intermediate results from MetaMap Indexing and the Related Citations indexing paths were saved. Then for each trial the MTI processing was completed using the appropriate options.

There are 42,368 citations in the collection. They all have titles but none has an abstract. The human indexers assigned 345,983 MeSH terms to those articles. 124,721 were main headings (IM). That is 8.17 terms per citation.

The Baseline

Although the text in the collection of title only articles did not change and there no other dramatic changes to MTI, there has been an improvement in MTI performance using the default options since the baseline used to evaluate the rank score thresholds.(See Table 20.) The reasons for this change are unclear.

Default Processing	**Precision**	**Recall**	**F_2**	**Diff**
Previous Baseline	.19	.38	.3021	
New Baseline	.22	.37	.3078	+.0057

Table 20. New Baseline v. Previous Baseline

Technical Evaluation

Table 21 presents the basic metrics used for evaluating MTI performance. The values for the baseline or default performance, the filtering thresholds of rank 10 and score 190, the limits on recommendation list size based on title length, and the use of both at the same time. These behaviors will usually be identified in this evaluation by the name of the options that implement these policies: RSfilterTO and limitTitleOnly.

Metric	Baseline	limitTitleOnly	RSfilterTO	Both
Terms recommended	568,659	486,404	444,606	432,811
Terms matching MEDLINE	119,251	114,650	112,025	111,125
Terms matching IM terms	74,095	72,075	70,872	70,484
Precision	0.22	0.24	0.25	0.25
Recall	0.37	0.35	0.35	0.34
Average number of matching terms	2.81	2.71	2.64	2.62
IM Precision	0.14	0.15	0.16	0.17
IM Recall	0.63	0.61	0.60	0.60
Average number of IM terms	1.75	1.709	1.67	1.66
Macro Averaged F_2	0.3078	0.3091	0.3098	0.3095

Table 21. Technical Metrics for Candidate Options and Baseline

The fundamental observation is that these option filter out terms reducing the number of correct recommendations at the same time. However, they manage to filter proportionally the correct and incorrect terms so that the F_2 remains essentially stable. The basic comparison summarized in Table 22 reveals that using both together gives weaker performance than just the rank score thresholds alone and only filters a few more terms. However, this level of evaluation may over look other significant effects of these options especially since the differences are not likely to be statistically significant.

	F_2 delta	terms filtered	Percent Filtered
limitTitleOnly	+0.0013	82,255	14.5%
RSfilterTO	+0.0019	124,053	21.8%
Both	+0.0017	135,848	23.9%

Table 22. Basic Comparison of Candidate Options

Worse-Better Analysis

As we did with the evaluation of the thresholds for citations with abstracts we will look next at the effects of the options on individual citations. We look at the proportion of the different outcome classes for citations. Are they better or worse for the filtering?

We also judge this by some metrics on the distribution of the changes occurring to individual citations.

The partitions shown in Table 23 give the percentage of the filtered citations and the percentage of all citations for the various classes of outcomes. One thing that is clear from these data is that although the performance differences are small the affects of the filtering is pervasive.

Options	**Both**		**limitTO**		**RSfilterTO**	
Outcomes	**Filtered**	**All**	**Filtered**	**All**	**Filtered**	**All**
F_2 **Increased**	68.0%	45.7%	71.3%	33.7%	68.8%	43.9%
F_2 **Unchanged**	8.0%	5.4%	8.6%	4.1%	8.4%	5.3%
Not Filtered		32.9%		52.8%		36.7%
F_2 **Decreased**	24.0%	16.1%	20.1%	9.5%	22.7%	14.4%
Citations	28,436		19,991		26,832	

Table 23. Filtering Partitioned Collections for Candidate Filters

Table 24 shows key statistics for the distribution of the F_2 scores of individual citations and the distribution of the magnitude of the changes in F_2 score for each citation. The change in F_2 is the difference between the F2 for the default and the F_2 for the listed option for a particular citation. The midmean is the mean for the values between the 25th an 75th percentile. The average F_2 increase is the mean of changes in F2 that were positive. The average F_2 decrease is the mean of changes in F_2 that were negative. The F_2 change statistics differ so much from the average increase and decrease because they include many zeros.

Key stats	**both**	**limitTO**	**RSfilterTO**	**default**
F_2 mean	0.30946	0.30910	0.30975	0.30781
F_2 median	0.30000	0.30300	0.30000	0.30300
F_2 midmean	0.30760	0.30696	0.30790	0.30550
F_2 midrange	0.48610	0.48610	0.48610	0.48610
Change Stats				
F_2 change mean	+0.00165	+0.00129	+0.00194	
F_2 change midmean	+0.00669	+0.00155	+0.00572	
Average F_2 increase	0.03459	0.02823	0.03367	
Average F_2 decrease	0.08781	0.08653	0.08842	

Table 24. F_2 Changes Distributions

You may notice that the average increase is much smaller than the average decrease. The reason performance does not suffer for any of the options is that the number of citations with improvements is about three times the number that are damaged by the filtering. (See the increase, decrease numbers in Table 23.)

There are only about 1600 more citations filtered by both than by RSfilterTO alone. LimitTitleOnly is the cleaner filter, but affects only about half of the citations. But unfortunately the portion damage by the filtering is highest with both options. RSfilterTO alone raises the F_2 most. Used together they raise the midmean of the changes in F_2 scores for all citations further than either alone.

Evaluating Filtering

The strictest evaluation of filtering is to look at how many true terms are lost due to the filtering. Table 25 shows how many citations lost terms that matched terms in the MEDLINE indexing.

	both	limitTO	RSfilterTO
True counts down	6837	4025	6111
True - no change	355331	383453	362597

Table 25. Comparison of Citations losing True Terms.

The filtering metric devised originally to evaluate threshold filtering alternatives will be useful here. Table 26 presents the accuracy, the filtering error rate, and the combined filtering metric (See page 14 for the definition of these metrics.). The delta metrics are the percent change from the default if appropriate or the mean for a suitable population. Color is used to highlight the best value for each delta metric.

	Δ F2	Accuracy	Δ Accur	Error	Δ error	Filtering
LimitTO	0.0042	0.3382	-0.1257	0.0559	0.0339	-0.0876
RSfilterTO	0.0065	0.4024	0.0404	0.0582	-0.0060	0.0409
Both	0.0055	0.4200	0.0859	0.0598	-0.0331	0.0583

Table 26. Filter Metrics for Candidate Options

The good news is that with the combined options we loose less than the sum of True terms lost individually. This is reflected in the combination having the best accuracy. Together the options also produced the biggest improvement in accuracy. Since the two options together produce the best overall improvement in the indexing as summarized in the filter metric of 6%, both options should be added to the default MTI processing.