InfoTracker: Pedigree Tracking in the Face of Ancillary Content

Eugene Creswick, Terrance Goan and Emi Fujioka
Stottler Henke Associates Inc.

1107 NE 45th St., Suite 310, Seattle, WA 98105
206-675-1169 FAX: 206-545-7227
rcreswick@stottlerhenke.com http://www.stottlerhenke.com
Track Document Pedigree
Track Document Pedigree
Applications

- Plagiarism
- Information Flow
- Security Policies
The Challenge
Common content confuses comparisons

determines the degree of extremality required of the outliers. $N$ can be
used to shift the balance between precision and recall. For example,
the full 114 data points of the results in Table 2 have a lower quartile
of 3.67 ($q_1$) and an upper quartile of 43.280 ($q_3$), indicating that
29 data points have scores under 3.67 and 87 data points have scores
under 43.280. With $N = 10$, the threshold is set to $319.785$, and only
the top seven results are retained.

The experiment described in Section 4.2 was run with varying val-
ues of $N$, from the range [500]. Low values of $N$ represent very con-
servative estimates of the distribution of meaningful documents, and
sets a low threshold for outliers. Each full-and increment increases
the threshold by an amount equal to the intra-quartile range, trim-
ning the query results more aggressively. The full test corpus of 38
query documents was run on each successive value of $N$ and the
average number of results, average precision, and average recall are
recorded in Table 3.

<table>
<thead>
<tr>
<th>$N$</th>
<th>Recall</th>
<th>Precision</th>
<th>Average Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.0</td>
<td>0.95</td>
<td>0.71</td>
<td>0.09</td>
</tr>
<tr>
<td>15.1</td>
<td>0.93</td>
<td>0.74</td>
<td>0.09</td>
</tr>
<tr>
<td>15.9</td>
<td>0.92</td>
<td>0.75</td>
<td>0.09</td>
</tr>
<tr>
<td>10.1</td>
<td>0.91</td>
<td>0.76</td>
<td>0.09</td>
</tr>
<tr>
<td>4.5</td>
<td>0.88</td>
<td>0.77</td>
<td>0.09</td>
</tr>
<tr>
<td>7.5</td>
<td>0.88</td>
<td>0.75</td>
<td>0.09</td>
</tr>
<tr>
<td>6.0</td>
<td>0.86</td>
<td>0.73</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Table 3 clearly shows the control available over the balance be-
 tween precision and recall, and demonstrates the amount of result
trimming that can usually be applied for a desired level of recall. Even
the most minimal trimming shortened the results list by over 90% com-
pared to the initial minimum of 108 results, yet only reduced average recall by 1% compared to the case where no trimming was done.

5 CONCLUSIONS AND FUTURE WORK

During the execution of this project, we have identified a number of
directions to pursue in the future:

Examine an Active Learning scenario: Further our future goals to
be an exhaustive evaluation of the InfoTracker prototype in a scenario
taken from an Active Learning scenario and mark highly relevant con-
cepts. This extends the ability of the system to identify and mark highly
relevant content while the system is in use. Overlap data. Another indication of the activeness of the docu-
ment collection is available in the content of the overlapping docu-
ments themselves. For example, if document $C$ contains content
taken directly from document $E$, which was originally taken from
document $A$, there is a chance that the overlapping section that $C$
shares with $E$ will be larger than the overlapping section found to be
common to $C$ and $A$. Indeed, it is highly likely that the over-
lapping content between $C$ and $A$ is a proper subset of the ove-
lapping content shared between $C$ and $E$. Further analysis of the similarities
to overlapping content shared between multiple documents may
reveal more information about the document sections.

Alternative outlier definitions: The characteristics of the tail of
each results list may more closely fit a certain type of distrib-
ution. For example, a more complex outlier detection method (such as
Gumbel Test for Section 3.0) may be able to determine a thresh-
old for results trimming that improves precision.

We have presented an approach to document indexing and search
that enables the detection of document pedigrees when annotated an-
nuity content is present. We have compared this approach to the
common vector space approach most frequently for information re-
trieval tasks, showing that our approach is better able to manage
the presence of auxiliary content. InfoTracker makes use of efficient
disk-based data structures that promise to scale well with large cor-
pora that do not fit in memory, however, a thorough evaluation of the
scalability of InfoTracker is a task for future investigation.

Evaluation of the proposed data set revealed that a great deal of
content is available over the pre-characterized tail-end. This can be
incorporated into tasks in the future to adapt to the needs at hand. For
example, applications dealing with the classification of potentially
classified content will require a high degree of recall, while an ap-
lication where the emphasis is on immediate results may choose to
avoid false positives with higher precision.

REFERENCES

[1] Turkib, T., "WebPedia: A web-based system for collaborative learning of
Web documents.", in Proc. of the 2nd ACM International Conference on
Web Information and Data Management (WIDM 2001), pp. 1-10, ACM,

effective and efficient method for ranking web pages," in Proc. of the
13th ACM Conference on Information and Knowledge Management (CIK

retrieval in large databases," in Proc. of the 2nd ACM International
Conference on Web Information and Data Management (WIDM 2001),

trieval using a content-based approach," in Proc. of the 2nd ACM
International Conference on Web Information and Data Management

of Query Synonymy on the Performance of Content-based Image Retri-
val Systems," in Proc. of the 13th ACM Conference on Information and

the effect of query synonymy on the performance of content-based image
retrieval systems," in Proc. of the 13th ACM Conference on Information
determines the degree of extremity required of the outlier. \( X \) can be used to shift the balance between precision and recall. For example, the full 
116 data points of the results in Table 2 have a lower quartile of \( X \leq 0.87 \) and an upper quartile of \( X \leq 0.20 \). It indicates that 
29 data points have scores under 1.83 and 27 data points have scores under 1.73. With \( X = 0 \), the threshold is set to 319.78, and only 
the top seven results are retained.

The experiment described in Section 4.2 was run with varying values of \( X \) from the range [0, 1]. Low values of \( X \) represent very 
conservative estimates of the distribution of scattered documents, and 
set a low threshold for outliers. Each fall-off increment increases 
the threshold by an amount equal to the intra-query distance, 
truncating the query results more aggressively. The full test corpus of 38 
query documents was run on each successive value of \( X \) and 
average number of results, average precision, and average recall are 
recorded in Table 2.

Table 2. Precision/Recall statistics for the page detection experiment, 
and function of outlier extremity

<table>
<thead>
<tr>
<th>Value of X</th>
<th>Recall</th>
<th>Precision</th>
<th>Average Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.93</td>
<td>0.78</td>
<td>0.81</td>
</tr>
<tr>
<td>0.5</td>
<td>0.87</td>
<td>0.74</td>
<td>0.80</td>
</tr>
<tr>
<td>1</td>
<td>0.82</td>
<td>0.70</td>
<td>0.78</td>
</tr>
<tr>
<td>1.5</td>
<td>0.78</td>
<td>0.66</td>
<td>0.74</td>
</tr>
<tr>
<td>2</td>
<td>0.74</td>
<td>0.62</td>
<td>0.71</td>
</tr>
<tr>
<td>2.5</td>
<td>0.70</td>
<td>0.58</td>
<td>0.68</td>
</tr>
<tr>
<td>3</td>
<td>0.66</td>
<td>0.54</td>
<td>0.65</td>
</tr>
<tr>
<td>3.5</td>
<td>0.62</td>
<td>0.50</td>
<td>0.62</td>
</tr>
<tr>
<td>4</td>
<td>0.58</td>
<td>0.46</td>
<td>0.58</td>
</tr>
<tr>
<td>4.5</td>
<td>0.54</td>
<td>0.42</td>
<td>0.54</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>0.38</td>
<td>0.50</td>
</tr>
<tr>
<td>5.5</td>
<td>0.46</td>
<td>0.34</td>
<td>0.46</td>
</tr>
<tr>
<td>6</td>
<td>0.42</td>
<td>0.30</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Table 3 clearly shows the control available over the balance between 
precision and recall, and determines the amount of result 
trimming that can be applied for a desired level of recall. Even the 
most minimal trimming approach reduced the results by no more 
than 70%, compared to the minimum size of 106 queries, yet 
only reduced average recall by 1% compared to the case where no 
trimming was done.

5 CONCLUSIONS AND FUTURE WORK
During the execution of this project, we have identified a number of 
approaches to pursue in the future. The evaluation of the hit/track 
algorithm in a realistic scenario with a true set of documents, 
identifies and marks irrelevant content while the system is in use.

Reconcile time stamps: The current approach does not take 
time stamp of document authoring and using into account when determining 
peckage. If the query documents are a source with a historical document, then both the source and 
the result document are likely to be returned in the list of results.

Overlap items: Another indication of the actual structure of the 
document package is available in the content of the overlapping 
items themselves. For example, if document C contains content 
taken directly from document F, which was originally taken from 
document L, there is a chance that the overlapping section that C 
shares with F will be larger than the overlapping section found to be 
common to C and L. Indeed, it is highly likely that the over- 
lapping content between C and L is a proper subset of the overlap 
section shared between C and F. To date, we have not yet analyzed the similarities between overlapping content shared between multiple documents. However, it is expected that the characteristics of the data of each result list may more closely fit a certain type of distri- bution. If so, a more complex outlier detection method such as Gumbel Test for (Section 3) may be able to determine a thresh- old value for result trimming that improves precision.

We have presented an approach to document linking and search 
that enables the detection of document packages when notated as 
irrelevant content is present. We have compared this approach to the 
common vector-space approach and find that it provides better 
retrieval results, showing that our approach is better able to 
manage the presence of irrelevancy content. The tracker makes use of efficient 
database-based data structures that provide good scaling with large 
 corpora that do not fit in memory, respectively. Through evaluation of the 
sensitivity of information retrieval to a set topic for future investigation.

Evaluation of the proposed data set revealed that a great deal of 
control is available over the pre-mentioned hit/track. This can 
be incorporated into text in the future to adapt to the needs of a user. 
For example, applications dealing with the enlargement of potentially 
irrelevant content will require a higher degree of recall, while an 
application where the emphasis is on immediate results may choose 
not to false positives with higher precision.

REFERENCES
Common content confuses comparisons

determines the degree of extremity required of the outliers. $N$ can be used to shift the balance between precision and recall. For example, the full 118 data points of the results in Table 2 have a lower quartile of 3.37 (Q1) and an upper quartile of 24.90 (Q2), indicating that 25 data points have scores under 3.37 and 87 data points have scores under 24.90. With $N = 0$, the threshold is set to 319.72, and only the top seven results are returned.

The experiment described in Sections 4.2 was run for varying values of $N$ from the range [0–14]. Low values of $N$ represent very conservative estimates of the distribution of unrealized documents, and set a low threshold for outliers. Each full-scale increment increases the threshold by an amount equal to the inter-quartile range trimming the query results more aggressively. The full text corpus of 38 query documents was run on each successive value of $N$ and the average number of results, average precision, and average recall are recorded in Table 2.

<table>
<thead>
<tr>
<th>$N$</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>40.53%</td>
<td>0.11</td>
</tr>
<tr>
<td>1</td>
<td>38.11%</td>
<td>0.24</td>
</tr>
<tr>
<td>2</td>
<td>37.54%</td>
<td>0.38</td>
</tr>
<tr>
<td>3</td>
<td>37.36%</td>
<td>0.44</td>
</tr>
<tr>
<td>4</td>
<td>37.51%</td>
<td>0.44</td>
</tr>
<tr>
<td>5</td>
<td>37.86%</td>
<td>0.41</td>
</tr>
<tr>
<td>6</td>
<td>37.93%</td>
<td>0.36</td>
</tr>
</tbody>
</table>

5 CONCLUSIONS AND FUTURE WORK

During the execution of this project, no limitations or directions were pursued.

Evaluate in an Active Learning scenario: Pursuit in our future goals is to perform an exhaustive evaluation of the InfoTracker prototype in scenarios typical of an Active Learning scenario. This includes identifying and marking irrelevant content while the system is in use.

Incorporate time stamping: The current approach does not take into account temporal aspects of document authoring and usage into account when determining pedalies. Therefore, if a query documentation is a source with a historical document, then both the source and the subject document are likely to be returned in the list of results. These false-positive results can be reduced by considering the dates that returned documents are authored, possibly preventing the results hierarchically, or only returning either the youngest or earliest sources.

Overlap data: Another indication of the actual structure of the document peddles is available in the content of the overlapping documents themselves. For example, if document $C$ contains content taken directly from document $D$, which was originally taken from document $E$, there is a chance that the overlapping section that $C$ shares with $D$ will be larger than the overlapping section found to be common to $C$ and $E$. Indeed, it is likely that the overlapping content between $C$ and $D$, $D$ is a proper subset of the overlap shared between $C$ and $E$. In-depth analysis of the similarities between overlapping content shared between multiple documents may reveal more interesting information about the document peddles.

Alternative outlier definitions: The characteristics of the four tables of each class may more closely fit a certain type of distribution. If so, a more complex outlier detection method such as Gumbel Test for Sections 30 may be used to determine a threshold for recall trimming that improves precision.

We have presented an approach to document pooling and search that enables the detection of document peddles when unannotated auxiliary content is present. We have compared this approach to the common vector-space approach most frequently for information retrieval tasks, showing that our approach is better able to manage the presence of auxiliary content. InfoTracker makes use of efficient click-based data structures that can scale well with large corpora that do not fit in memory, however, a thorough evaluation of the scalability of InfoTracker is still a topic for further investigation.

Evaluation of the proposed data set revealed that a great deal of control is available over the pre-constructive recall. This can be incorporated into tools in the future to adapt to the needs of the user. For example, applications dealing with the declassification of potentially classified content will require a high degree of recall, while an application where the emphasis is on immediate results may choose to avoid false positives with higher precision.

REFERENCES

Related Work

- Suffix Tree Document Models
- Fuzzy Fingerprints
- Hoad & Zobel's Fingerprints
Solution
determines the degree of extremeness required of the outliers. \( N \) can be used to shift the balance between precision and recall. For example, the full 114 data points of the results in Table 2 have a lower quartile of 3.87 (Q1) and an upper quartile of 45.78 (Q3), indicating that 37 data points have scores under 3.87 and 87 data points have scores under 45.78. With \( N = 0 \), the threshold is set to 319.72, and only the top seven results are returned.

The experiment described in Section 2.2 was run with varying values of \( N \) from the range [0, 10]. Low values of \( N \) represent very conservative estimates of the distribution of untrusted documents, and sets a low threshold for outliers. Each full-set increment increased the threshold by an amount equal to the inter-quartile range, training the query results more aggressively. The full test corpus of 98 query documents was run on each successive value of \( N \) and the average number of results, average precision, and average recall are recorded in Table 3.

Table 3 clearly shows the control available over the balance between precision and recall, and demonstrates the amount of result trimming that can safely be applied for a desired level of recall. Even the most minimal trimming attempted reduced the results list by over 60% compared to the initial minimum size of 106 results only; reduced average recall by 1% compared to the case where no trimming was done.

## 5. CONCLUSIONS AND FUTURE WORK

During the execution of this project, we have identified a number of directions for future work:

### First Annotate Learning Scenarios

- **Framework for our future goals**: To perform an exhaustive evaluation of the InfoTracker prototype in a scenario that captures the realistic scenarios in which we can identify and mark hostile content while the system is in use.

### Incorporate Time Stamps

- The current approach does not take the temporal aspect of document authoring and mining into account when determining peddledness. Therefore, if a query document contains a source with a historical document, then both the source and the resulting document are likely to be returned in the list of results. These false-positive results can be reduced by considering the date that the retrieved documents are authored, possibly presenting the results hierarchically, or only returning those that are not widely circulated sources.

### Overlap data

- Another indication of the actual structure of the document peddler is available in the context of the overlapping sections in the documents. For example, if document C contains content taken directly from document E, which was originally taken from document A, then there is a chance that the overlapping section that C shares with E will be larger than the overlapping section found to be common to C and A. Indeed, it is highly likely that the overlapping content between C and A is a subset of the overlap shared between C and E. A further analysis of the differences between overlapping content shared between the documents may reveal more information about the document peddler.

### Additional Features

- The characteristics of the false positives list may more closely fit a certain type of distribution. If so, a more complex outlier detection method (such as a Gaussian Test for outliers) may be able to determine a threshold for result trimming that improves precision.

### REFERENCES

1. Title 1, Title 2, Title 3, Title 4, Title 5 (2019).

2. Title 1, Title 2, Title 3, Title 4, Title 5 (2019).

3. Title 1, Title 2, Title 3, Title 4, Title 5 (2019).

4. Title 1, Title 2, Title 3, Title 4, Title 5 (2019).

5. Title 1, Title 2, Title 3, Title 4, Title 5 (2019).

6. Title 1, Title 2, Title 3, Title 4, Title 5 (2019).

7. Title 1, Title 2, Title 3, Title 4, Title 5 (2019).

8. Title 1, Title 2, Title 3, Title 4, Title 5 (2019).

9. Title 1, Title 2, Title 3, Title 4, Title 5 (2019).

10. Title 1, Title 2, Title 3, Title 4, Title 5 (2019).
How?
How? Use Contrasting Corpora

Open Content
- Resumes
- Introductions
- Footers
- Web Content
- Headers

Sensitive Content
- Intellectual Property
- Published work
- Secrets
- Homework Assignments
Algorithm
Index Both Corpora with one Suffix Tree

**Widely-Used/Common Text**
- c1 = “their hotel rooms”
- c2 = “their hideout”

**Sensitive Documents**
- s1 = “hotel as their hideout”

**Suffixes: c1**
- rooms
- hotel rooms
- their hotel rooms

**Suffixes: c2**
- hideout
- their hideout

**Suffixes: s1**
- hideout
- their hideout
- as their hideout
- hotel as their hideout

Text Unique to Sensitive Documents
Search for a document

Query: “Hotel rooms as their hideout”
Search for a document

Query: “Hotel rooms as their hideout”
Open: “Hotel rooms”
Search for a document

Query: “Hotel rooms as their hideout”
Open: “Hotel rooms”
Open: “rooms”
Search for a document

Query: “Hotel rooms as their hideout”
Open: “Hotel rooms”
Open: “rooms”
Sensitive: “as their hideout”
Search for a document

Query:  “Hotel rooms as their hideout”
Open:   “Hotel rooms”
Open:    “rooms”
Sensitive:  “as their hideout”
Open:     “their hideout”
Search for a document

Query: “Hotel rooms as their hideout”
Open: “Hotel rooms”
Open: “rooms”
Sensitive: “as their hideout”
Open: “their hideout”
Algorithm >

Filter the resulting string overlaps

Aligned Character Strings

Query Doc.

Sens. Overlap

Open Overlap

Resulting Overlap(s)

Too Short
Algorithm > Ranking
Overlap-based Ranking

With the help of the monk Gunavarman and others, a firm foothold on Java well before the 5th century, and about this time in Sumatra, and by the 7th century...

The Indonesian island of Sumatra is... Kebu people

Northwest coast of the island of Sumatra. This earthquake is the second strongest earthquake recorded in the world. The earthquake resulted in...

On the morning of December 26, 2004 a magnitude 9.3 earthquake struck off the Northwest coast of the Indonesian island of Sumatra. The earthquake resulted from complex slip on the fault where the oceanic portion of the Indian Plate slides under Sumatra, part of the Eurasian Plate. The earthquake deformed the ocean floor, pushing the overlying water up into a tsunami wave. The tsunami wave devastated nearby areas where the wave may have been as high as 25 meters (80 feet) tall. The sudden vertical rise of the seabed by several meters during the earthquake displaced massive volumes of water, resulting in a tsunami that struck the coasts of the Indian Ocean.

Radar satellites recorded the heights of tsunami waves in deep water: at two hours after the earthquake, the maximum height was 60 cm (2 ft). These are the first such observations ever made. However, these observations could not have been used to provide a warning, because the satellites were not intended for that purpose and the data took hours to analyze.

SITUATION

PACOM organized a peace-time operation to provide assistance to the victims of the Boxing Day tsunami in the India Ocean. While this was not a war-time operation, there remained the possibility of terrorist activities by conservative radical organizations.
Overlap-based Ranking

The Indonesian island of Sumatra was visited.

Northwest coast of the island of Sumatra.

On the morning of December 26, 2004 a magnitude 9.3 earthquake struck off the Northwest coast of the Indonesian island of Sumatra. The earthquake resulted from faultlines. The northern Indonesian island of Sumatra hit very quickly, while Sri Lanka and the coast of India were hit roughly two hours later. Thailand was struck about two hours later despite closer to the epicentre, because the overlying water up into a tsunami wave. The tsunami wave devastated nearby areas where the wave may have been as high as 25 meters (80 feet) tall. The sudden vertical rise of the seabed by several meters during the earthquake displaced massive volumes of water, resulting in a tsunami that struck the coasts of the Indian Ocean.

PACOM organized a peace-time operation to provide assistance to the victims of the Boxing Day tsunami in the India Ocean. While this was not a war-time operation, there remained the possibility of terrorist activities by conservative radical organizations.
Overlap Frequency for Ranking

A: the Indonesian island of Sumatra.
B: Northwest coast of the
C: the Indonesian island of Sumatra.

unique text
lower frequency
Greater impact

common text
higher frequency
Less impact
Evaluation
InfoTracker was compared to Vector Space

- Cosine Similarity
- TF-IDF weighted vectors
- No stop words
Data Set

Open Content

Web Content (on-line news, blogs, etc...)

Related Work

Sensitive Content

Resumes

Footers

Published work

Headers

Intellectual Property
Data Set

- 272 SBIR proposals
- 234 historical proposals
- 38 query proposals
Oracle
Evaluation > Results
### InfoTracker improved precision / recall

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector Space</td>
<td>0.119</td>
<td>0.764</td>
</tr>
<tr>
<td>InfoTracker</td>
<td>0.167</td>
<td>0.913</td>
</tr>
</tbody>
</table>
Contributions / Future Work
Ancillary content can be managed

- Contrasting corpora
- Manual/actively learned tags
- Detecting document sections
(re)Evaluate on Open data

Compare with differing corpora

The Linux Doc. Project
Algorithmic Improvements

Active Learning

Document time stamps

Overlap size / encapsulation
Questions?
## Calculating Precision / Recall

<table>
<thead>
<tr>
<th>Rank</th>
<th>Score</th>
<th>File</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6289.995</td>
<td>Document-92</td>
</tr>
<tr>
<td>2</td>
<td>3206.34</td>
<td>Document-21</td>
</tr>
<tr>
<td>3</td>
<td>1630.607</td>
<td>Document-13</td>
</tr>
<tr>
<td>4</td>
<td>1366.318</td>
<td>Document-46</td>
</tr>
<tr>
<td>5</td>
<td>1157.704</td>
<td>Document-1</td>
</tr>
<tr>
<td>6</td>
<td>1103.442</td>
<td>Document-43</td>
</tr>
<tr>
<td>7</td>
<td>624.2379</td>
<td>Document-114</td>
</tr>
<tr>
<td>8</td>
<td>327.5333</td>
<td>Document-67</td>
</tr>
<tr>
<td>9</td>
<td>273.6506</td>
<td>Document-74</td>
</tr>
<tr>
<td>10</td>
<td>263.0365</td>
<td>Document-48</td>
</tr>
<tr>
<td>11</td>
<td>244.4071</td>
<td>Document-10</td>
</tr>
<tr>
<td>12</td>
<td>238.4346</td>
<td>Document-113</td>
</tr>
<tr>
<td>13</td>
<td>207.32</td>
<td>Document-101</td>
</tr>
<tr>
<td>14</td>
<td>134.9912</td>
<td>Document-58</td>
</tr>
<tr>
<td>15</td>
<td>131.5204</td>
<td>Document-12</td>
</tr>
<tr>
<td>16</td>
<td>118.6787</td>
<td>Document-7</td>
</tr>
<tr>
<td>17</td>
<td>97.52703</td>
<td>Document-37</td>
</tr>
<tr>
<td>18</td>
<td>89.8972</td>
<td>Document-9</td>
</tr>
<tr>
<td>19</td>
<td>89.50462</td>
<td>Document-27</td>
</tr>
<tr>
<td>20</td>
<td>81.49963</td>
<td>Document-50</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>
## Calculating Precision / Recall

Consider the top 23 results.
(to allow for perfect recall)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Score</th>
<th>File</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6289.995</td>
<td>Document-92</td>
</tr>
<tr>
<td>2</td>
<td>3206.34</td>
<td>Document-21</td>
</tr>
<tr>
<td>3</td>
<td>1630.607</td>
<td>Document-13</td>
</tr>
<tr>
<td>4</td>
<td>1366.318</td>
<td>Document-46</td>
</tr>
<tr>
<td>5</td>
<td>1157.704</td>
<td>Document-1</td>
</tr>
<tr>
<td>6</td>
<td>1103.442</td>
<td>Document-43</td>
</tr>
<tr>
<td>7</td>
<td>624.2379</td>
<td>Document-114</td>
</tr>
<tr>
<td>8</td>
<td>327.5333</td>
<td>Document-67</td>
</tr>
<tr>
<td>9</td>
<td>273.6506</td>
<td>Document-74</td>
</tr>
<tr>
<td>10</td>
<td>263.0365</td>
<td>Document-48</td>
</tr>
<tr>
<td>11</td>
<td>244.4071</td>
<td>Document-10</td>
</tr>
<tr>
<td>12</td>
<td>238.4346</td>
<td>Document-113</td>
</tr>
<tr>
<td>13</td>
<td>207.32</td>
<td>Document-101</td>
</tr>
<tr>
<td>14</td>
<td>134.9912</td>
<td>Document-58</td>
</tr>
</tbody>
</table>
### Ranking Scores Plummet Quickly

<table>
<thead>
<tr>
<th>Rank</th>
<th>Score</th>
<th>File</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6289.995</td>
<td>Document-92</td>
</tr>
<tr>
<td>2</td>
<td>3206.34</td>
<td>Document-21</td>
</tr>
<tr>
<td>3</td>
<td>1630.607</td>
<td>Document-13</td>
</tr>
<tr>
<td>4</td>
<td>1366.318</td>
<td>Document-46</td>
</tr>
<tr>
<td>5</td>
<td>1157.704</td>
<td>Document-1</td>
</tr>
<tr>
<td>6</td>
<td>1103.442</td>
<td>Document-43</td>
</tr>
<tr>
<td>7</td>
<td>624.2379</td>
<td>Document-114</td>
</tr>
<tr>
<td>8</td>
<td>327.5333</td>
<td>Document-67</td>
</tr>
<tr>
<td>9</td>
<td>273.6506</td>
<td>Document-74</td>
</tr>
<tr>
<td>10</td>
<td>263.0365</td>
<td>Document-48</td>
</tr>
<tr>
<td>11</td>
<td>244.4071</td>
<td>Document-10</td>
</tr>
<tr>
<td>12</td>
<td>238.4346</td>
<td>Document-113</td>
</tr>
<tr>
<td>13</td>
<td>207.32</td>
<td>Document-101</td>
</tr>
<tr>
<td>14</td>
<td>134.9912</td>
<td>Document-58</td>
</tr>
<tr>
<td>15</td>
<td>131.5204</td>
<td>Document-12</td>
</tr>
<tr>
<td>16</td>
<td>118.6787</td>
<td>Document-7</td>
</tr>
<tr>
<td>17</td>
<td>97.52703</td>
<td>Document-37</td>
</tr>
<tr>
<td>18</td>
<td>89.8972</td>
<td>Document-9</td>
</tr>
<tr>
<td>19</td>
<td>89.50462</td>
<td>Document-27</td>
</tr>
<tr>
<td>20</td>
<td>81.49963</td>
<td>Document-50</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>
Trimming Results >

Ranking Scores Plummet Quickly

[Graph showing a rapid decrease in ranking scores, with a steep drop and a plateau at a lower value.]
Trimming improves precision, retains recall

<table>
<thead>
<tr>
<th>$N$</th>
<th>Result Count</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Trimming</td>
<td>162.53</td>
<td>0.03</td>
<td>0.98</td>
</tr>
<tr>
<td>0</td>
<td>40.95</td>
<td>0.11</td>
<td>0.97</td>
</tr>
<tr>
<td>0.5</td>
<td>28.71</td>
<td>0.14</td>
<td>0.93</td>
</tr>
<tr>
<td>1</td>
<td>22.29</td>
<td>0.16</td>
<td>0.91</td>
</tr>
<tr>
<td>1.5</td>
<td>18.92</td>
<td>0.19</td>
<td>0.90</td>
</tr>
<tr>
<td>2</td>
<td>15.81</td>
<td>0.21</td>
<td>0.88</td>
</tr>
<tr>
<td>2.5</td>
<td>13.47</td>
<td>0.23</td>
<td>0.87</td>
</tr>
<tr>
<td>3</td>
<td>11.76</td>
<td>0.24</td>
<td>0.84</td>
</tr>
<tr>
<td>3.5</td>
<td>10.50</td>
<td>0.26</td>
<td>0.84</td>
</tr>
<tr>
<td>4</td>
<td>9.63</td>
<td>0.27</td>
<td>0.81</td>
</tr>
<tr>
<td>4.5</td>
<td>8.82</td>
<td>0.29</td>
<td>0.80</td>
</tr>
<tr>
<td>5</td>
<td>8.18</td>
<td>0.31</td>
<td>0.78</td>
</tr>
<tr>
<td>5.5</td>
<td>7.55</td>
<td>0.33</td>
<td>0.78</td>
</tr>
<tr>
<td>6</td>
<td>7.13</td>
<td>0.36</td>
<td>0.77</td>
</tr>
</tbody>
</table>