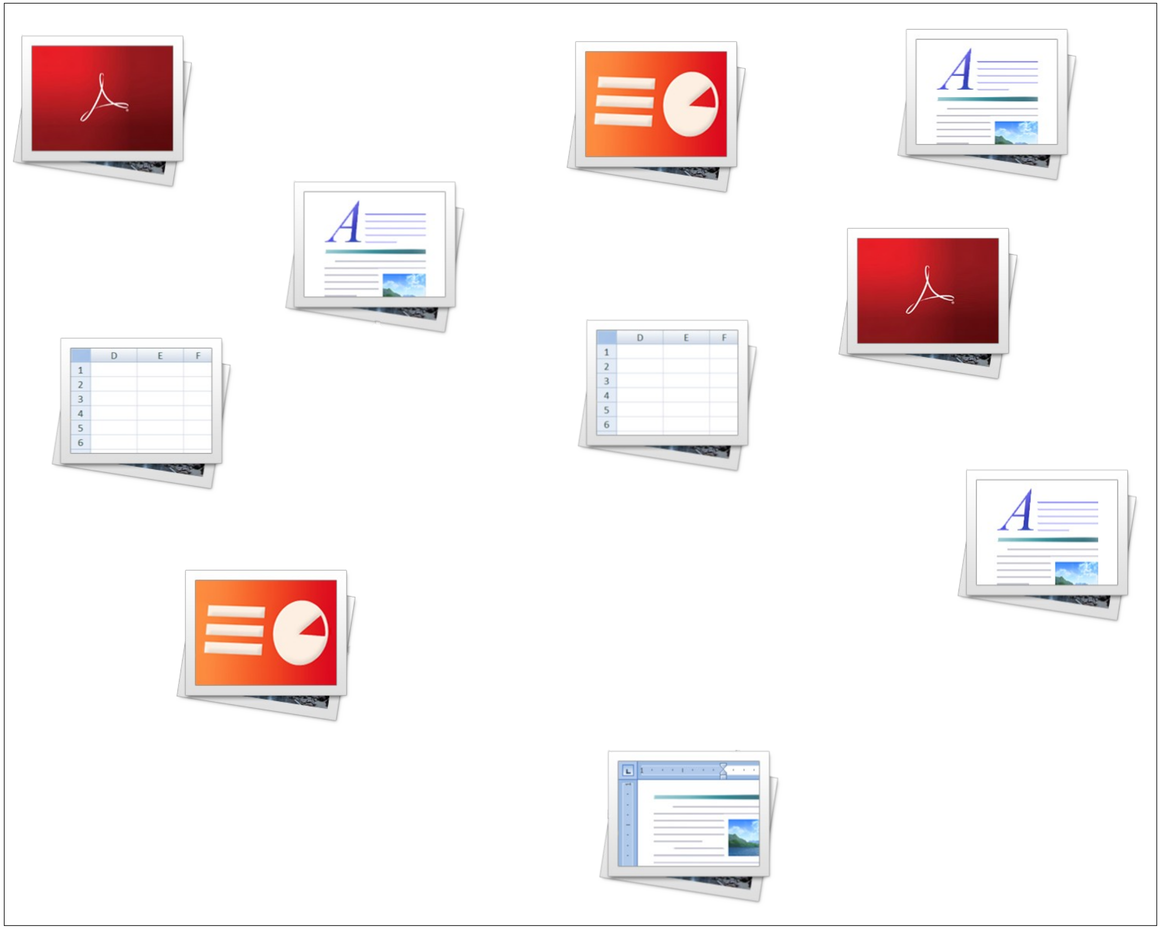
InfoTracker:
Pedigree Tracking in the Face of Ancillary Content

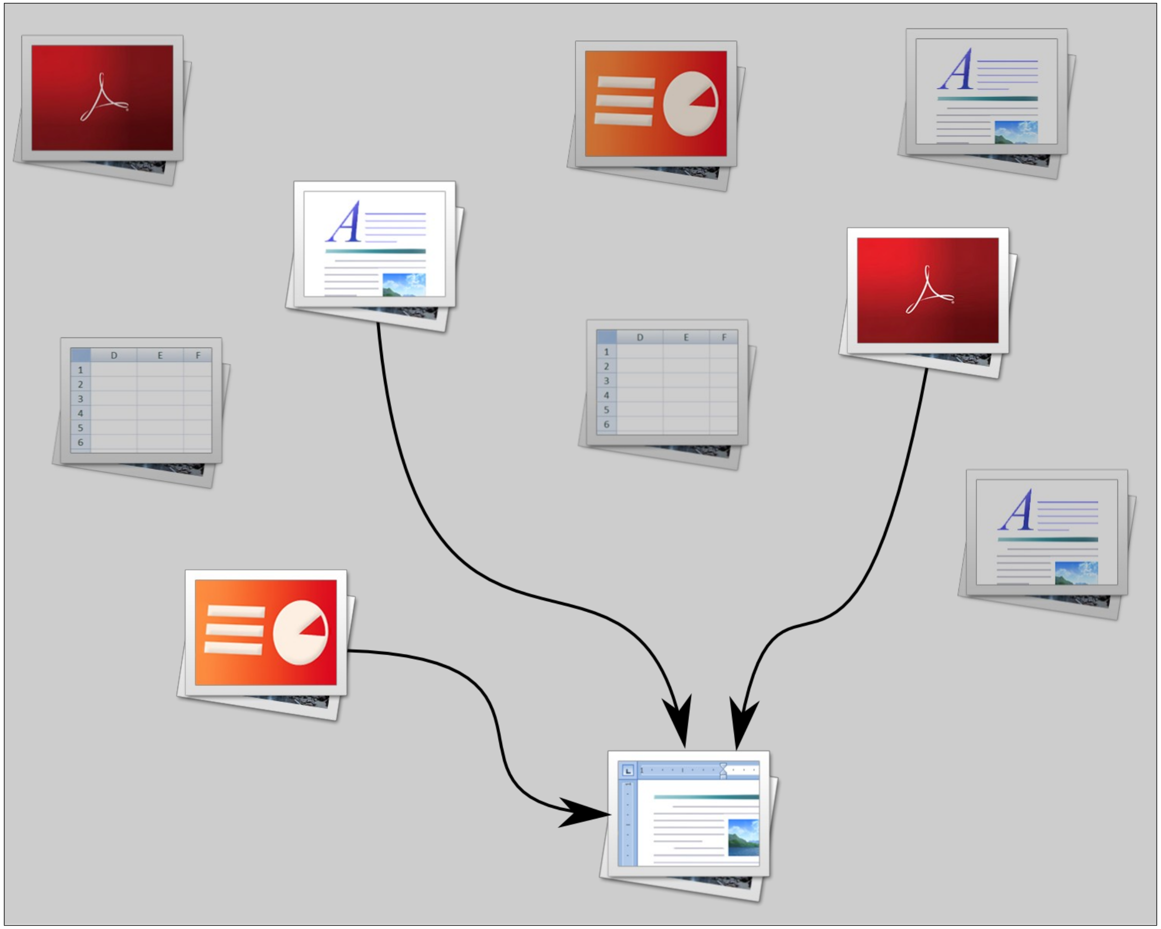
Eugene Creswick, Terrance Goan and Emi Fujioka
Stottler Henke Associates Inc.

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

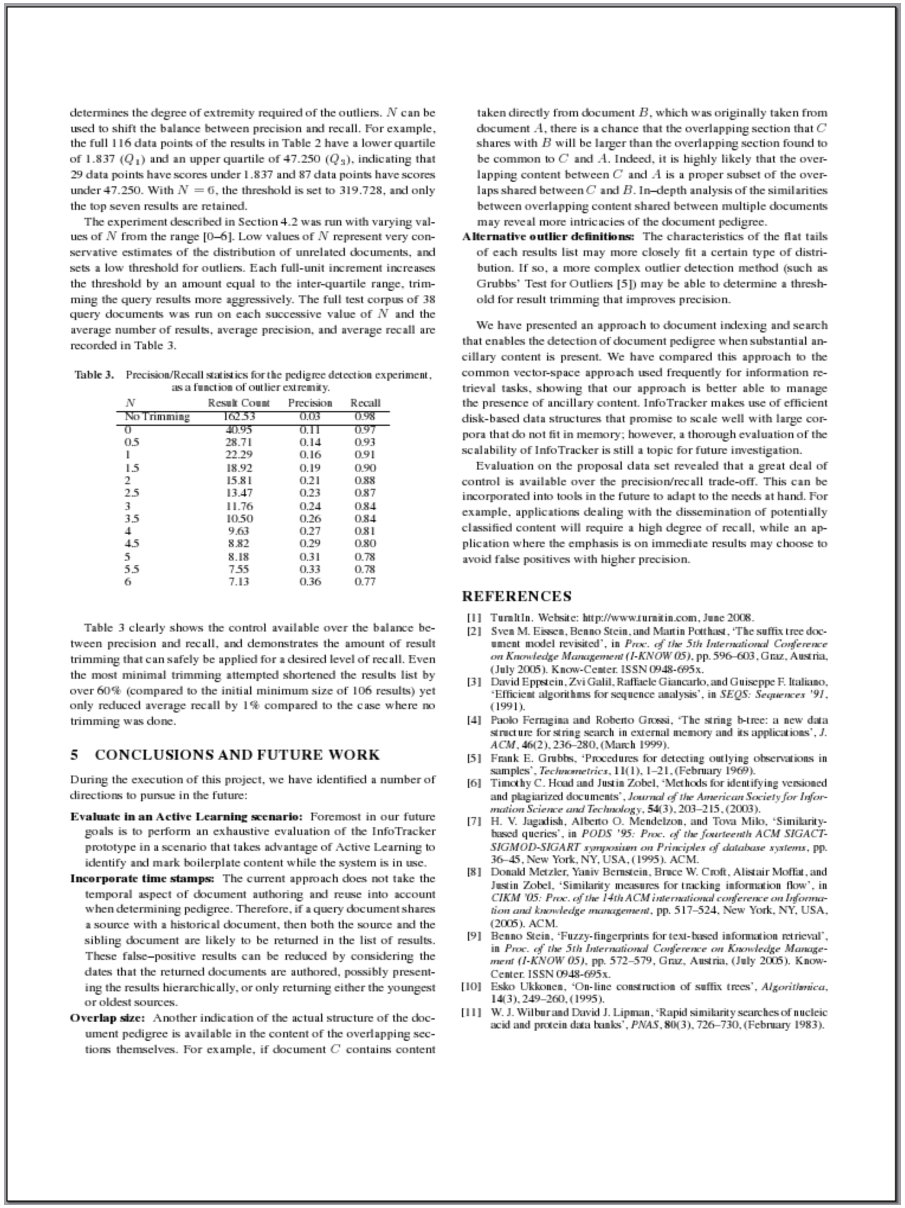
Plagiarism

Information Flow

Security Policies

The Challenge

Common content confuses comparisons



determines the degree of extremity required of the outliers. N 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 1.837 (Q_1) and an upper quartile of 47.250 (Q_3), indicating that 29 data points have scores under 1.837 and 87 data points have scores under 47.250. With $N = 6$, the threshold is set to 319.728, and only the top seven results are retained.

The experiment described in Section 4.2 was run with varying values of N from the range [0–6]. Low values of N represent very conservative estimates of the distribution of unrelated documents, and sets a low threshold for outliers. Each full-unit increment increases the threshold by an amount equal to the inter-quartile range, trimming 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 3. Precision/Recall statistics for the pedigree detection experiment, as a function of outlier extremity.

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

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 shortened the results list by over 60% (compared to the initial minimum size of 106 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:

Evaluate in an Active Learning scenario: Foremost in our future goals is to perform an exhaustive evaluation of the InfoTracker prototype in a scenario that takes advantage of Active Learning to identify and mark boilerplate content while the system is in use.

Incorporate time stamps: The current approach does not take the temporal aspect of document authoring and reuse into account when determining pedigree. Therefore, if a query document shares a source with a historical document, then both the source and the sibling document are likely to be returned in the list of results. These false–positive results can be reduced by considering the dates that the returned documents are authored, possibly presenting the results hierarchically, or only returning either the youngest or oldest sources.

Overlap size: Another indication of the actual structure of the document pedigree is available in the content of the overlapping sections themselves. For example, if document C contains content taken directly from document B, which was originally taken from document A, there is a chance that the overlapping section that C shares with B 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 proper subset of the overlaps shared between C and B. In–depth analysis of the similarities between overlapping content shared between multiple documents may reveal more intricacies of the document pedigree.

Alternative outlier definitions: The characteristics of the flat tails of each results list may more closely fit a certain type of distribution. If so, a more complex outlier detection method (such as Grubbs' Test for Outliers [5]) may be able to determine a threshold for result trimming that improves precision.

We have presented an approach to document indexing and search that enables the detection of document pedigree when substantial ancillary content is present. We have compared this approach to the common vector-space approach used frequently for information retrieval tasks, showing that our approach is better able to manage the presence of ancillary content. InfoTracker makes use of efficient disk-based data structures that promise to 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 future investigation.

Evaluation on the proposal data set revealed that a great deal of control is available over the precision/recall trade-off. This can be incorporated into tools in the future to adapt to the needs at hand. For example, applications dealing with the dissemination 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

[1] TurnItIn. Website: http://www.turnitin.com, June 2008.

[2] Sven M. Eissen, Benno Stein, and Martin Potthast, 'The suffix tree document model revisited', in *Proc. of the 5th International Conference on Knowledge Management (I-KNOW 05)*, pp. 596–603, Graz, Austria, (July 2005). Know-Center. ISSN 0948-695x.

[3] David Eppstein, Zvi Galil, Raffaele Giancarlo, and Guiseppe F. Italiano, 'Efficient algorithms for sequence analysis', in *SEQS: Sequences '91*, (1991).

[4] Paolo Ferragina and Roberto Grossi, 'The string b-tree: a new data structure for string search in external memory and its applications', *J. ACM*, 46(2), 236–280, (March 1999).

[5] Frank E. Grubbs, 'Procedures for detecting outlying observations in samples', *Technometrics*, 11(1), 1–21, (February 1969).

[6] Timothy C. Hoad and Justin Zobel, 'Methods for identifying versioned and plagiarized documents', *Journal of the American Society for Information Science and Technology*, **54**(3), 203–215, (2003).

[7] H. V. Jagadish, Alberto O. Mendelzon, and Tova Milo, 'Similarity-based queries', in *PODS '95: Proc. of the fourteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pp. 36–45, New York, NY, USA, (1995). ACM.

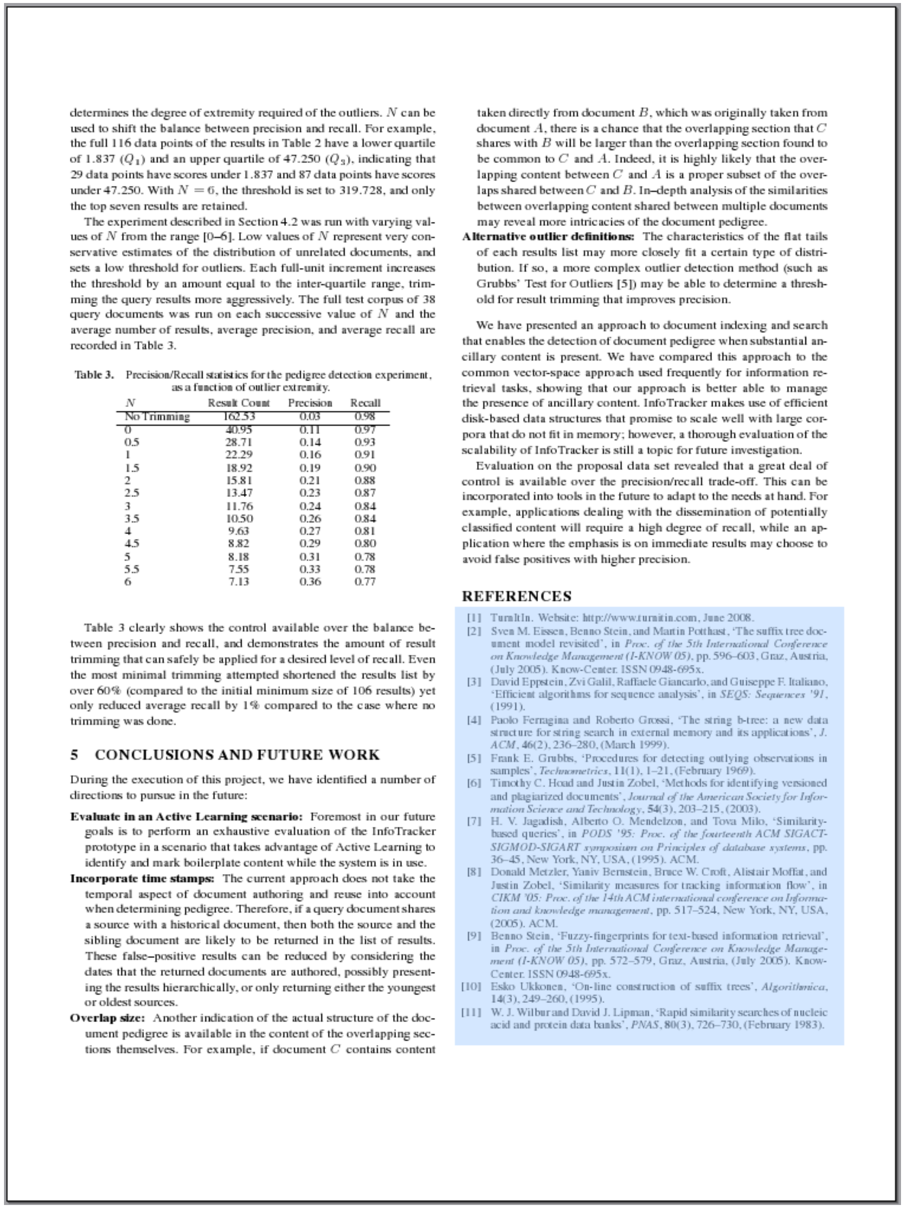
[8] Donald Metzler, Yaniv Bernstein, Bruce W. Croft, Alistair Moffat, and Justin Zobel, 'Similarity measures for tracking information flow', in *CIKM '05: Proc. of the 14th ACM international conference on Information and knowledge management*, pp. 517–524, New York, NY, USA, (2005). ACM.

[9] Benno Stein, 'Fuzzy-fingerprints for text-based information retrieval', in *Proc. of the 5th International Conference on Knowledge Management (I-KNOW 05)*, pp. 572–579, Graz, Austria, (July 2005). Know-Center. ISSN 0948-695x.

[10] Esko Ukkonen, 'On-line construction of suffix trees', *Algorithmica*, 14(3), 249–260, (1995).

[11] W. J. Wilbur and David J. Lipman, 'Rapid similarity searches of nucleic acid and protein data banks', *PNAS*, 80(3), 726–730, (February 1983).

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 116 data points of the results in Table 2 have a lower quartile of 1.837 (Q_1) and an upper quartile of 47.250 (Q_3), indicating that 29 data points have scores under 1.837 and 87 data points have scores under 47.250. With $N = 6$, the threshold is set to 319.728, and only the top seven results are retained.

The experiment described in Section 4.2 was run with varying values of N from the range [0–6]. Low values of N represent very conservative estimates of the distribution of unrelated documents, and sets a low threshold for outliers. Each full-unit increment increases the threshold by an amount equal to the inter-quartile range, trimming 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 3. Precision/Recall statistics for the pedigree detection experiment, as a function of outlier extremity.

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

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 shortened the results list by over 60% (compared to the initial minimum size of 106 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:

Evaluate in an Active Learning scenario: Foremost in our future goals is to perform an exhaustive evaluation of the InfoTracker prototype in a scenario that takes advantage of Active Learning to identify and mark boilerplate content while the system is in use.

Incorporate time stamps: The current approach does not take the temporal aspect of document authoring and reuse into account when determining pedigree. Therefore, if a query document shares a source with a historical document, then both the source and the sibling document are likely to be returned in the list of results. These false–positive results can be reduced by considering the dates that the returned documents are authored, possibly presenting the results hierarchically, or only returning either the youngest or oldest sources.

Overlap size: Another indication of the actual structure of the document pedigree is available in the content of the overlapping sections themselves. For example, if document C contains content taken directly from document B, which was originally taken from document A, there is a chance that the overlapping section that C shares with B 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 proper subset of the overlaps shared between C and B. In–depth analysis of the similarities between overlapping content shared between multiple documents may reveal more intricacies of the document pedigree.

Alternative outlier definitions: The characteristics of the flat tails of each results list may more closely fit a certain type of distribution. If so, a more complex outlier detection method (such as Grubbs' Test for Outliers [5]) may be able to determine a threshold for result trimming that improves precision.

We have presented an approach to document indexing and search that enables the detection of document pedigree when substantial ancillary content is present. We have compared this approach to the common vector-space approach used frequently for information retrieval tasks, showing that our approach is better able to manage the presence of ancillary content. InfoTracker makes use of efficient disk-based data structures that promise to 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 future investigation.

Evaluation on the proposal data set revealed that a great deal of control is available over the precision/recall trade-off. This can be incorporated into tools in the future to adapt to the needs at hand. For example, applications dealing with the dissemination 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

[1] TurnItIn. Website: http://www.turnitin.com, June 2008.

[2] Sven M. Eissen, Benno Stein, and Martin Potthast, 'The suffix tree document model revisited', in *Proc. of the 5th International Conference on Knowledge Management (I-KNOW 05)*, pp. 596–603, Graz, Austria, (July 2005). Know-Center. ISSN 0948-695x.

[3] David Eppstein, Zvi Galil, Raffaele Giancarlo, and Guiseppe F. Italiano, 'Efficient algorithms for sequence analysis', in *SEQS: Sequences '91*, (1991).

[4] Paolo Ferragina and Roberto Grossi, 'The string b-tree: a new data structure for string search in external memory and its applications', *J. ACM*, 46(2), 236–280, (March 1999).

[5] Frank E. Grubbs, 'Procedures for detecting outlying observations in samples', *Technometrics*, 11(1), 1–21, (February 1969).

[6] Timothy C. Hoad and Justin Zobel, 'Methods for identifying versioned and plagiarized documents', *Journal of the American Society for Information Science and Technology*, **54**(3), 203–215, (2003).

[7] H. V. Jagadish, Alberto O. Mendelzon, and Tova Milo, 'Similarity-based queries', in *PODS '95: Proc. of the fourteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pp. 36–45, New York, NY, USA, (1995). ACM.

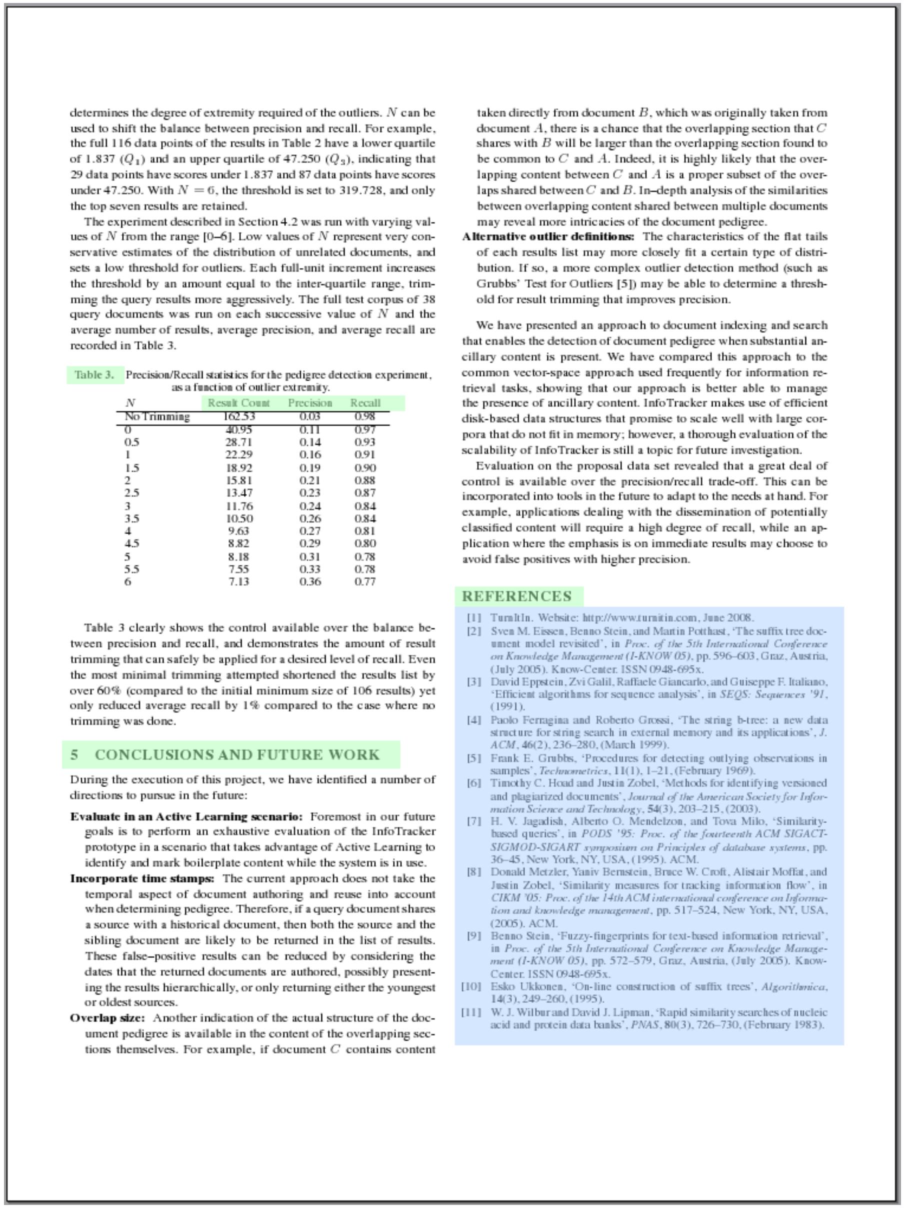
[8] Donald Metzler, Yaniv Bernstein, Bruce W. Croft, Alistair Moffat, and Justin Zobel, 'Similarity measures for tracking information flow', in *CIKM '05: Proc. of the 14th ACM international conference on Information and knowledge management*, pp. 517–524, New York, NY, USA, (2005). ACM.

[9] Benno Stein, 'Fuzzy-fingerprints for text-based information retrieval', in *Proc. of the 5th International Conference on Knowledge Management (I-KNOW 05)*, pp. 572–579, Graz, Austria, (July 2005). Know-Center. ISSN 0948-695x.

[10] Esko Ukkonen, 'On-line construction of suffix trees', *Algorithmica*, 14(3), 249–260, (1995).

[11] W. J. Wilbur and David J. Lipman, 'Rapid similarity searches of nucleic acid and protein data banks', *PNAS*, 80(3), 726–730, (February 1983).

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 116 data points of the results in Table 2 have a lower quartile of 1.837 (Q_1) and an upper quartile of 47.250 (Q_3), indicating that 29 data points have scores under 1.837 and 87 data points have scores under 47.250. With $N = 6$, the threshold is set to 319.728, and only the top seven results are retained.

The experiment described in Section 4.2 was run with varying values of N from the range [0–6]. Low values of N represent very conservative estimates of the distribution of unrelated documents, and sets a low threshold for outliers. Each full-unit increment increases the threshold by an amount equal to the inter-quartile range, trimming 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 3. Precision/Recall statistics for the pedigree detection experiment, as a function of outlier extremity.

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

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 shortened the results list by over 60% (compared to the initial minimum size of 106 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:

Evaluate in an Active Learning scenario: Foremost in our future goals is to perform an exhaustive evaluation of the InfoTracker prototype in a scenario that takes advantage of Active Learning to identify and mark boilerplate content while the system is in use.

Incorporate time stamps: The current approach does not take the temporal aspect of document authoring and reuse into account when determining pedigree. Therefore, if a query document shares a source with a historical document, then both the source and the sibling document are likely to be returned in the list of results. These false–positive results can be reduced by considering the dates that the returned documents are authored, possibly presenting the results hierarchically, or only returning either the youngest or oldest sources.

Overlap size: Another indication of the actual structure of the document pedigree is available in the content of the overlapping sections themselves. For example, if document C contains content taken directly from document B, which was originally taken from document A, there is a chance that the overlapping section that C shares with B 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 proper subset of the overlaps shared between C and B. In–depth analysis of the similarities between overlapping content shared between multiple documents may reveal more intricacies of the document pedigree.

Alternative outlier definitions: The characteristics of the flat tails of each results list may more closely fit a certain type of distribution. If so, a more complex outlier detection method (such as Grubbs' Test for Outliers [5]) may be able to determine a threshold for result trimming that improves precision.

We have presented an approach to document indexing and search that enables the detection of document pedigree when substantial ancillary content is present. We have compared this approach to the common vector-space approach used frequently for information retrieval tasks, showing that our approach is better able to manage the presence of ancillary content. InfoTracker makes use of efficient disk-based data structures that promise to 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 future investigation.

Evaluation on the proposal data set revealed that a great deal of control is available over the precision/recall trade-off. This can be incorporated into tools in the future to adapt to the needs at hand. For example, applications dealing with the dissemination 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

[1] TurnItIn. Website: http://www.turnitin.com, June 2008.

[2] Sven M. Eissen, Benno Stein, and Martin Potthast, 'The suffix tree document model revisited', in *Proc. of the 5th International Conference on Knowledge Management (I-KNOW 05)*, pp. 596–603, Graz, Austria, (July 2005). Know-Center. ISSN 0948-695x.

[3] David Eppstein, Zvi Galil, Raffaele Giancarlo, and Guiseppe F. Italiano, 'Efficient algorithms for sequence analysis', in *SEQS: Sequences '91*, (1991).

[4] Paolo Ferragina and Roberto Grossi, 'The string b-tree: a new data structure for string search in external memory and its applications', *J. ACM*, 46(2), 236–280, (March 1999).

[5] Frank E. Grubbs, 'Procedures for detecting outlying observations in samples', *Technometrics*, 11(1), 1–21, (February 1969).

[6] Timothy C. Hoad and Justin Zobel, 'Methods for identifying versioned and plagiarized documents', *Journal of the American Society for Information Science and Technology*, 54(3), 203–215, (2003).

[7] H. V. Jagadish, Alberto O. Mendelzon, and Tova Milo, 'Similarity-based queries', in *PODS '95: Proc. of the fourteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pp. 36–45, New York, NY, USA, (1995). ACM.

[8] Donald Metzler, Yaniv Bernstein, Bruce W. Croft, Alistair Moffat, and Justin Zobel, 'Similarity measures for tracking information flow', in *CIKM '05: Proc. of the 14th ACM international conference on Information and knowledge management*, pp. 517–524, New York, NY, USA, (2005). ACM.

[9] Benno Stein, 'Fuzzy-fingerprints for text-based information retrieval', in *Proc. of the 5th International Conference on Knowledge Management (I-KNOW 05)*, pp. 572–579, Graz, Austria, (July 2005). Know-Center. ISSN 0948-695x.

[10] Esko Ukkonen, 'On-line construction of suffix trees', *Algorithmica*, 14(3), 249–260, (1995).

[11] W. J. Wilbur and David J. Lipman, 'Rapid similarity searches of nucleic acid and protein data banks', *PNAS*, 80(3), 726–730, (February 1983).

Related Work

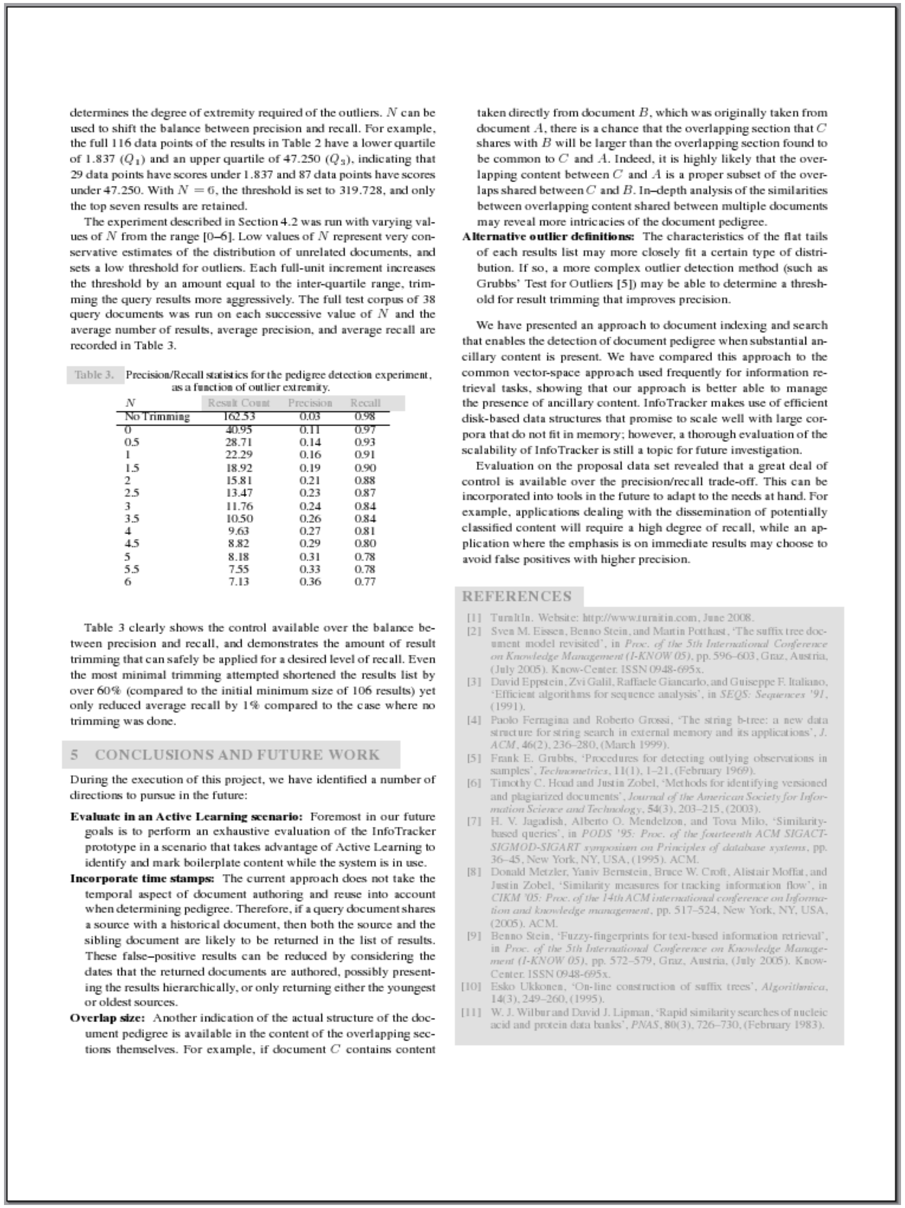
Suffix Tree Document Models

Fuzzy Fingerprints

Hoad & Zobel's Fingerprints

Solution

Ignore the ancillary content



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 116 data points of the results in Table 2 have a lower quartile of 1.837 (Q_1) and an upper quartile of 47.250 (Q_3), indicating that 29 data points have scores under 1.837 and 87 data points have scores under 47.250. With $N = 6$, the threshold is set to 319.728, and only the top seven results are retained.

The experiment described in Section 4.2 was run with varying values of N from the range [0–6]. Low values of N represent very conservative estimates of the distribution of unrelated documents, and sets a low threshold for outliers. Each full-unit increment increases the threshold by an amount equal to the inter-quartile range, trimming 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 3. Precision/Recall statistics for the pedigree detection experiment, as a function of outlier extremity.

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

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 shortened the results list by over 60% (compared to the initial minimum size of 106 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:

Evaluate in an Active Learning scenario: Foremost in our future goals is to perform an exhaustive evaluation of the InfoTracker prototype in a scenario that takes advantage of Active Learning to identify and mark boilerplate content while the system is in use.

Incorporate time stamps: The current approach does not take the temporal aspect of document authoring and reuse into account when determining pedigree. Therefore, if a query document shares a source with a historical document, then both the source and the sibling document are likely to be returned in the list of results. These false–positive results can be reduced by considering the dates that the returned documents are authored, possibly presenting the results hierarchically, or only returning either the youngest or oldest sources.

Overlap size: Another indication of the actual structure of the document pedigree is available in the content of the overlapping sections themselves. For example, if document C contains content taken directly from document B, which was originally taken from document A, there is a chance that the overlapping section that C shares with B 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 proper subset of the overlaps shared between C and B. In–depth analysis of the similarities between overlapping content shared between multiple documents may reveal more intricacies of the document pedigree.

Alternative outlier definitions: The characteristics of the flat tails of each results list may more closely fit a certain type of distribution. If so, a more complex outlier detection method (such as Grubbs' Test for Outliers [5]) may be able to determine a threshold for result trimming that improves precision.

We have presented an approach to document indexing and search that enables the detection of document pedigree when substantial ancillary content is present. We have compared this approach to the common vector-space approach used frequently for information retrieval tasks, showing that our approach is better able to manage the presence of ancillary content. InfoTracker makes use of efficient disk-based data structures that promise to 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 future investigation.

Evaluation on the proposal data set revealed that a great deal of control is available over the precision/recall trade-off. This can be incorporated into tools in the future to adapt to the needs at hand. For example, applications dealing with the dissemination 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

[1] TurnItIn. Website: http://www.turnitin.com, June 2008.

[2] Sven M. Eissen, Benno Stein, and Martin Potthast, 'The suffix tree document model revisited', in *Proc. of the 5th International Conference on Knowledge Management (I-KNOW 05)*, pp. 596–603, Graz, Austria, (July 2005). Know-Center. ISSN 0948-695x.

[3] David Eppstein, Zvi Galil, Raffaele Giancarlo, and Guiseppe F. Italiano, 'Efficient algorithms for sequence analysis', in *SEQS: Sequences '91*, (1991).

[4] Paolo Ferragina and Roberto Grossi, 'The string b-tree: a new data structure for string search in external memory and its applications', *J. ACM*, 46(2), 236–280, (March 1999).

[5] Frank E. Grubbs, 'Procedures for detecting outlying observations in samples', *Technometrics*, 11(1), 1–21, (February 1969).

[6] Timothy C. Hoad and Justin Zobel, 'Methods for identifying versioned and plagiarized documents', *Journal of the American Society for Information Science and Technology*, 54(3), 203–215, (2003).

[7] H. V. Jagadish, Alberto O. Mendelzon, and Tova Milo, 'Similarity-based queries', in *PODS '95: Proc. of the fourteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pp. 36–45, New York, NY, USA, (1995). ACM.

[8] Donald Metzler, Yaniv Bernstein, Bruce W. Croft, Alistair Moffat, and Justin Zobel, 'Similarity measures for tracking information flow', in *CIKM '05: Proc. of the 14th ACM international conference on Information and knowledge management*, pp. 517–524, New York, NY, USA, (2005). ACM.

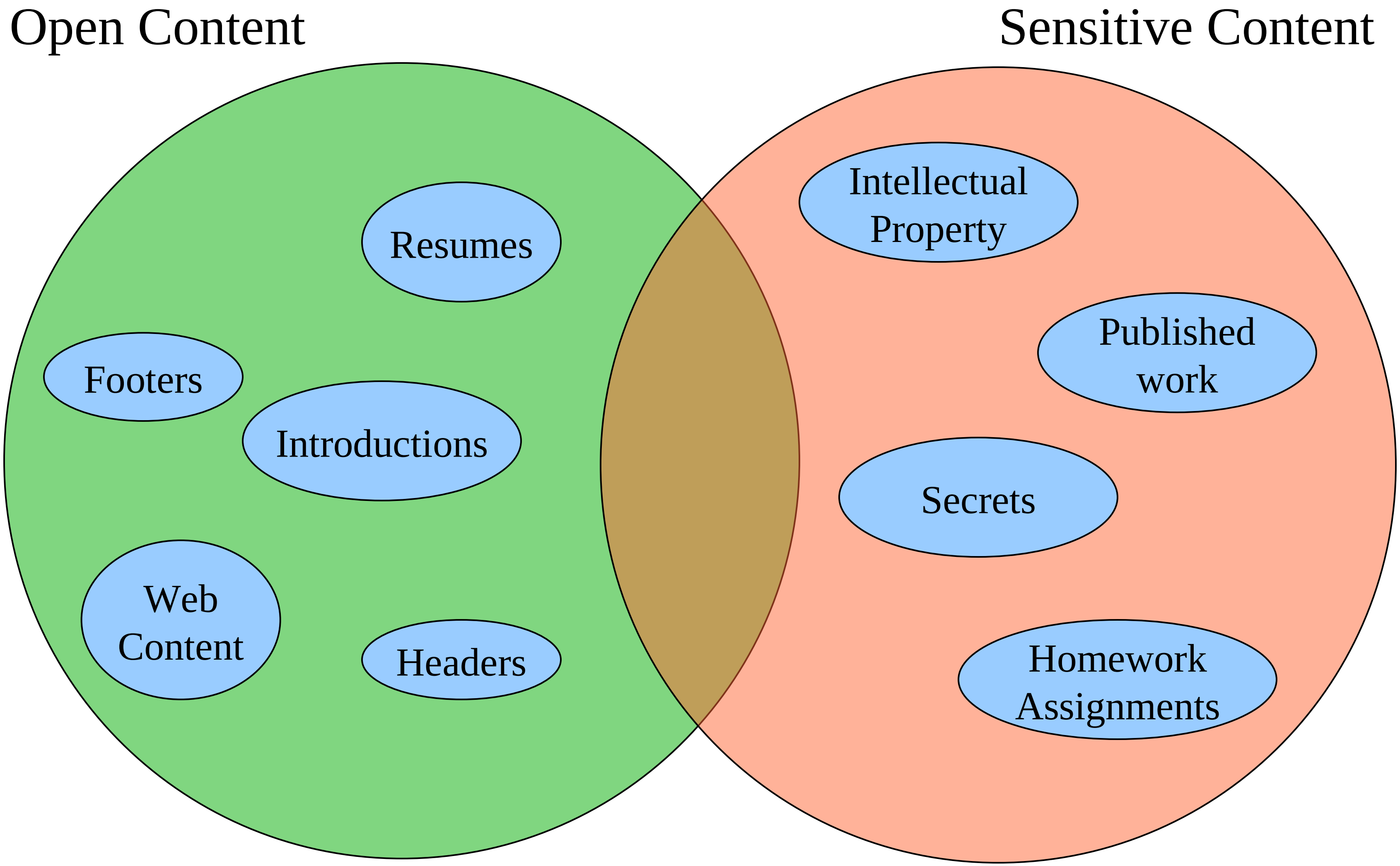
[9] Benno Stein, 'Fuzzy-fingerprints for text-based information retrieval', in *Proc. of the 5th International Conference on Knowledge Management (I-KNOW 05)*, pp. 572–579, Graz, Austria, (July 2005). Know-Center. ISSN 0948-695x.

[10] Esko Ukkonen, 'On-line construction of suffix trees', *Algorithmica*, 14(3), 249–260, (1995).

[11] W. J. Wilbur and David J. Lipman, 'Rapid similarity searches of nucleic acid and protein data banks', *PNAS*, 80(3), 726–730, (February 1983).

How?

How? Use Contrasting Corpora



Open Content

Sensitive Content

Resumes

Footers

Introductions

Web Content

Headers

Intellectual Property

Published work

Secrets

Homework Assignments

Algorithm

Index Both Corpora with one Suffix Tree

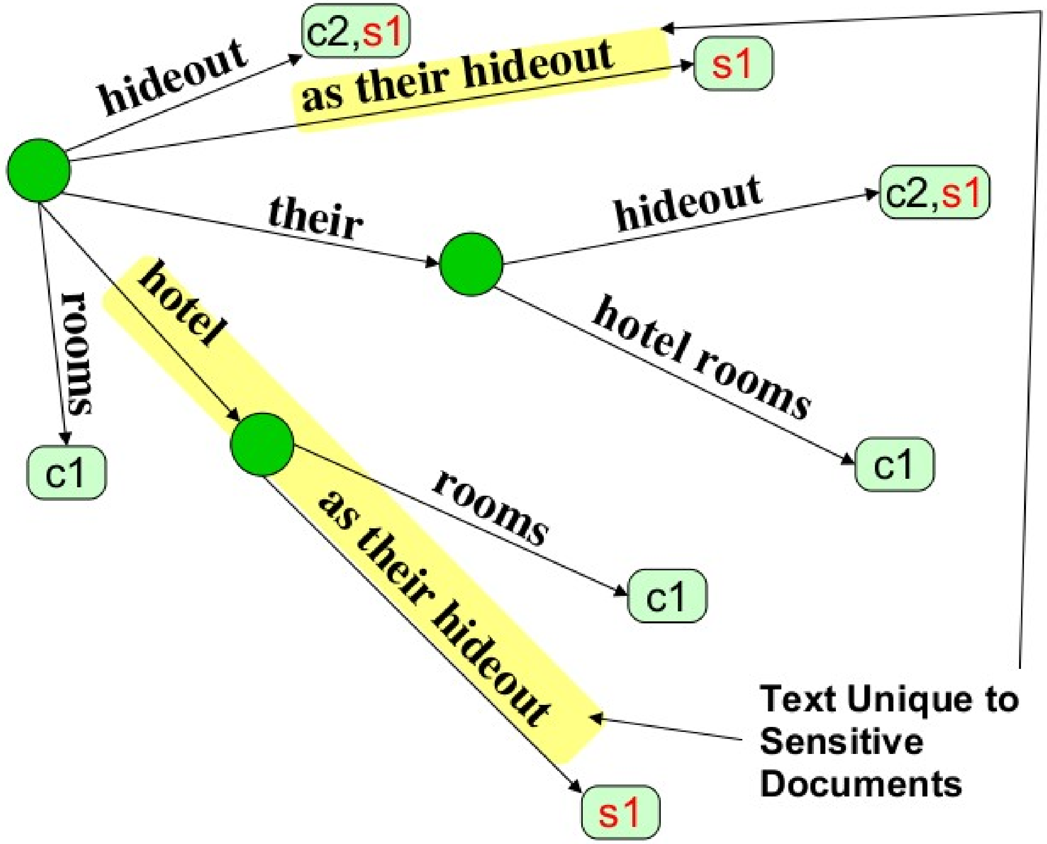
Widely-Used/Common Text
c1="their hotel rooms"
c2="their hideout"

Sensitive Documents
s1="hotel as their hideout"

Suffixes: c1
rooms
hotel rooms
their hotel rooms

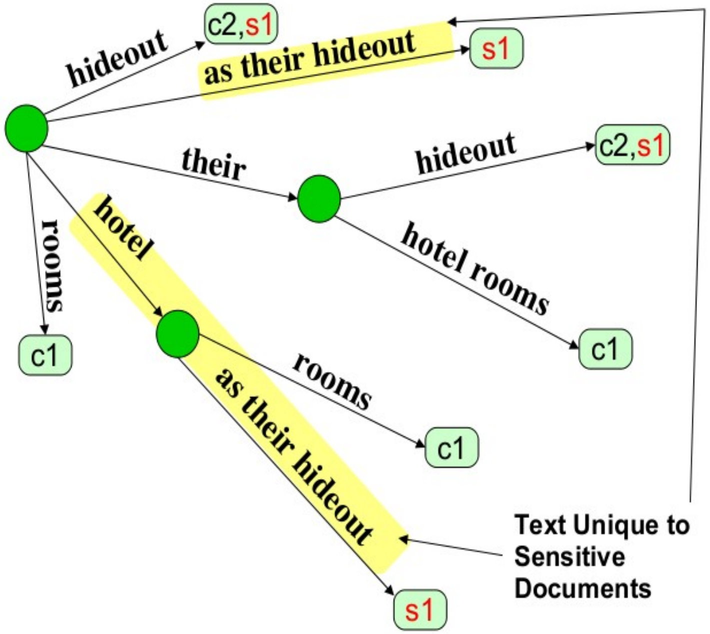
Suffixes: c2
hideout
their hideout

Suffixes: s1
hideout
their hideout
as their hideout
hotel as their hideout



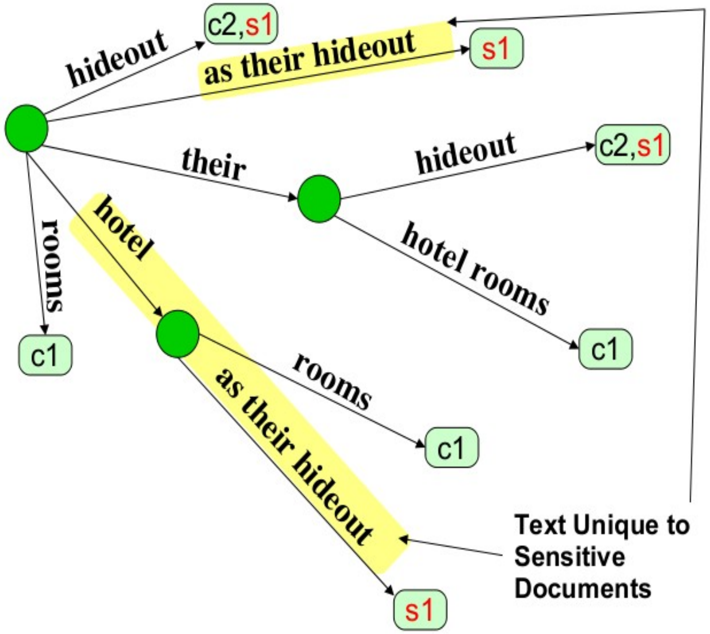
Text Unique to Sensitive Documents

Search for a document



Query: “Hotel rooms as their hideout”

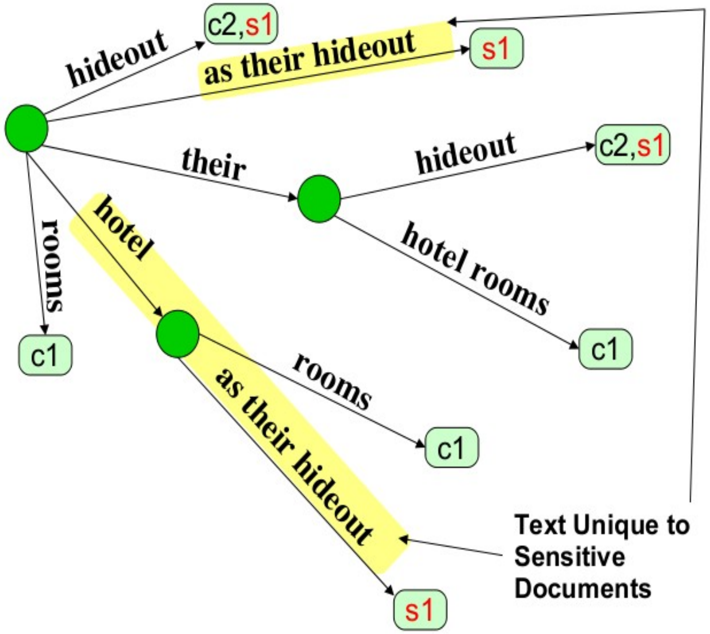
Search for a document



Query: “Hotel rooms as their hideout”

Open: “Hotel rooms”

Search for a document

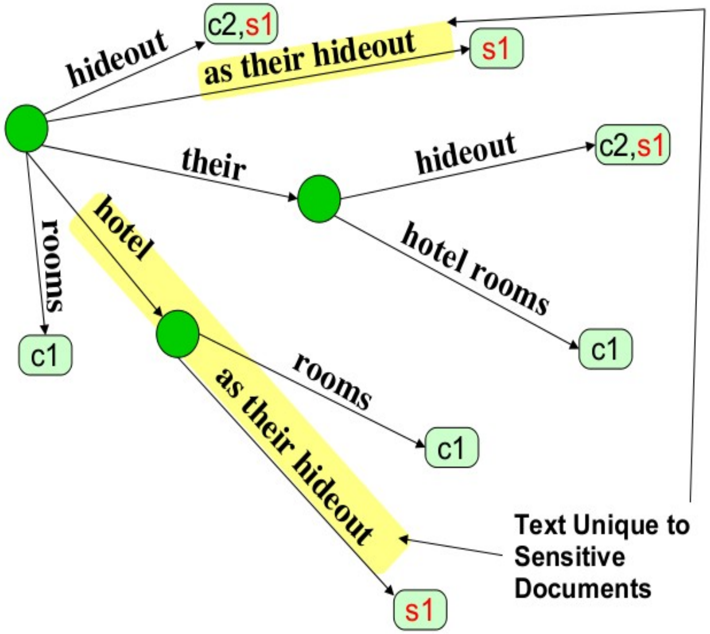


Query: “Hotel rooms as their hideout”

Open: “Hotel rooms”

Open: “rooms”

Search for a document



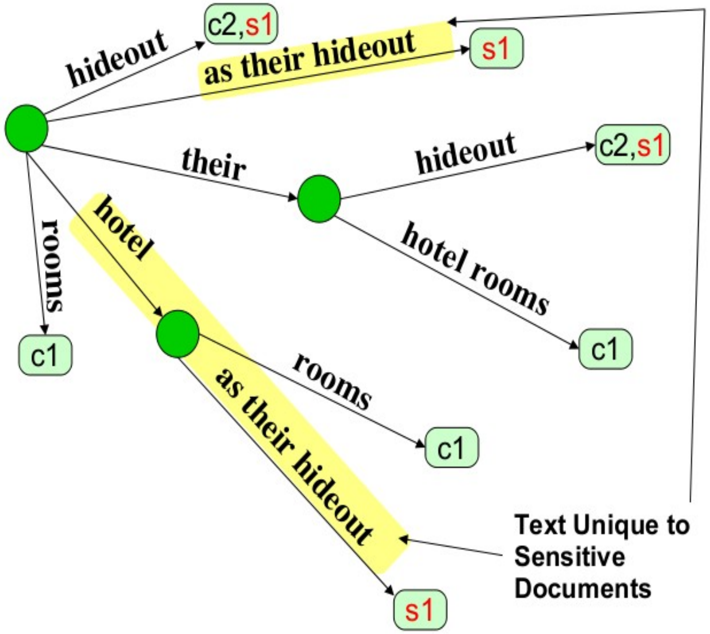
Query: “Hotel rooms as their hideout”

Open: “Hotel rooms”

Open: “rooms”

Sensitive: “as their hideout”

Search for a document



Query: "Hotel rooms as their hideout"

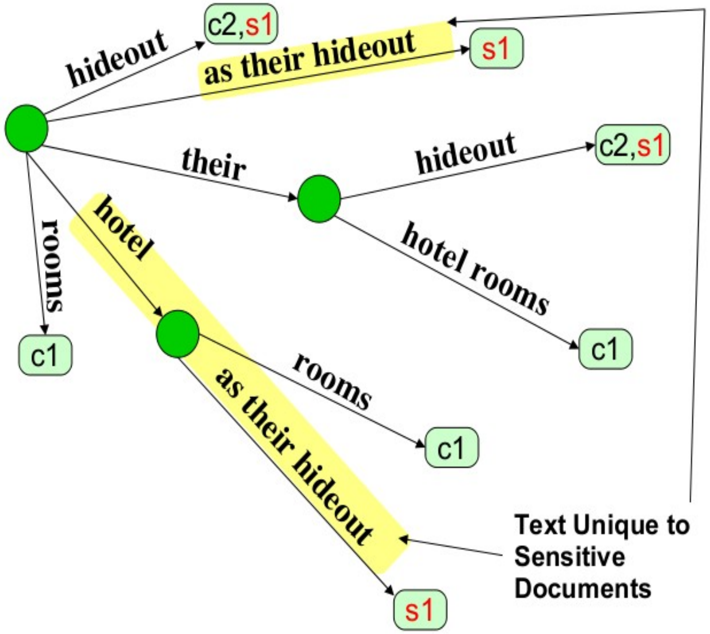
Open: "Hotel rooms"

Open: "rooms"

Sensitive: "as their hideout"

Open: "their hideout"

Search for a document



Text Unique to Sensitive Documents

Query: “Hotel rooms as their hideout”

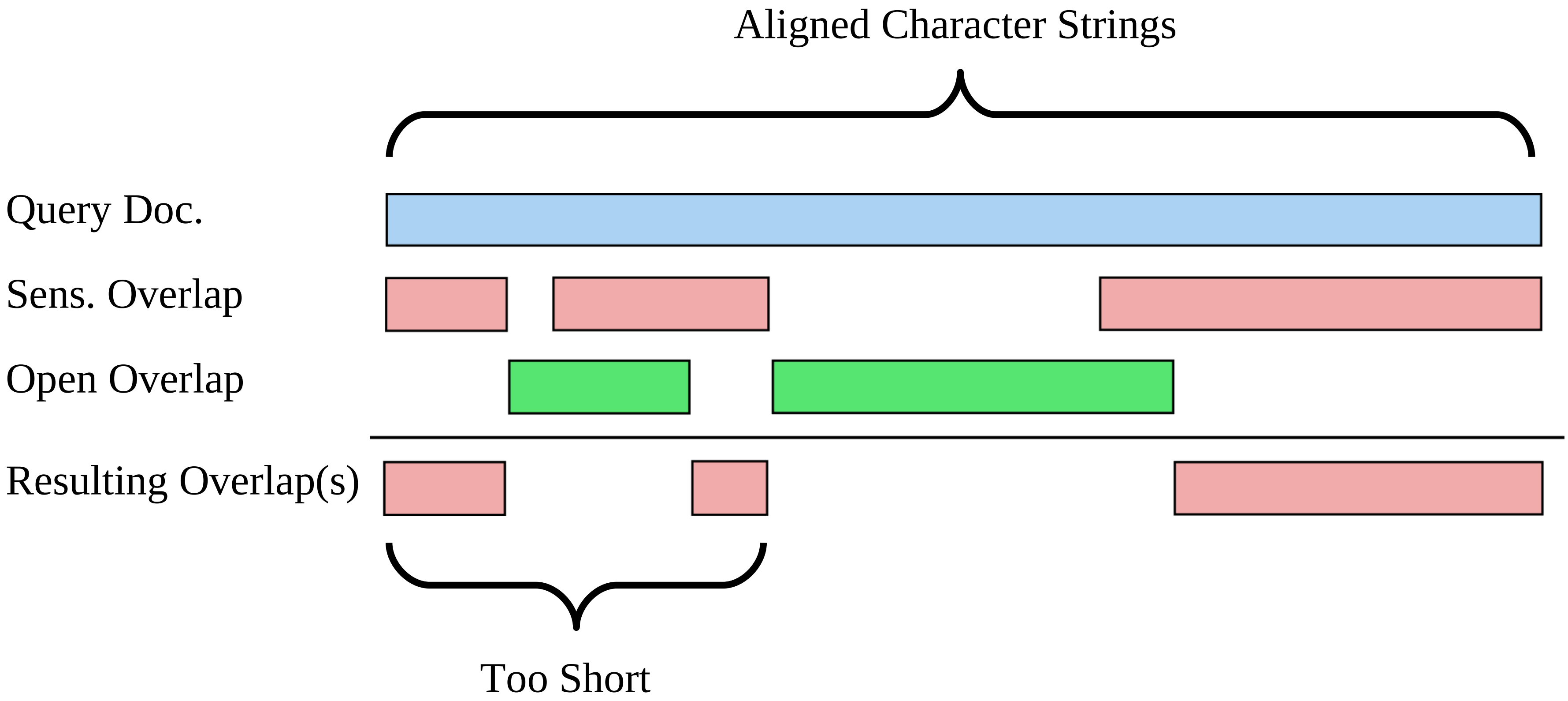
Open: “Hotel rooms”

Open: “rooms”

Sensitive: “as their hideout”

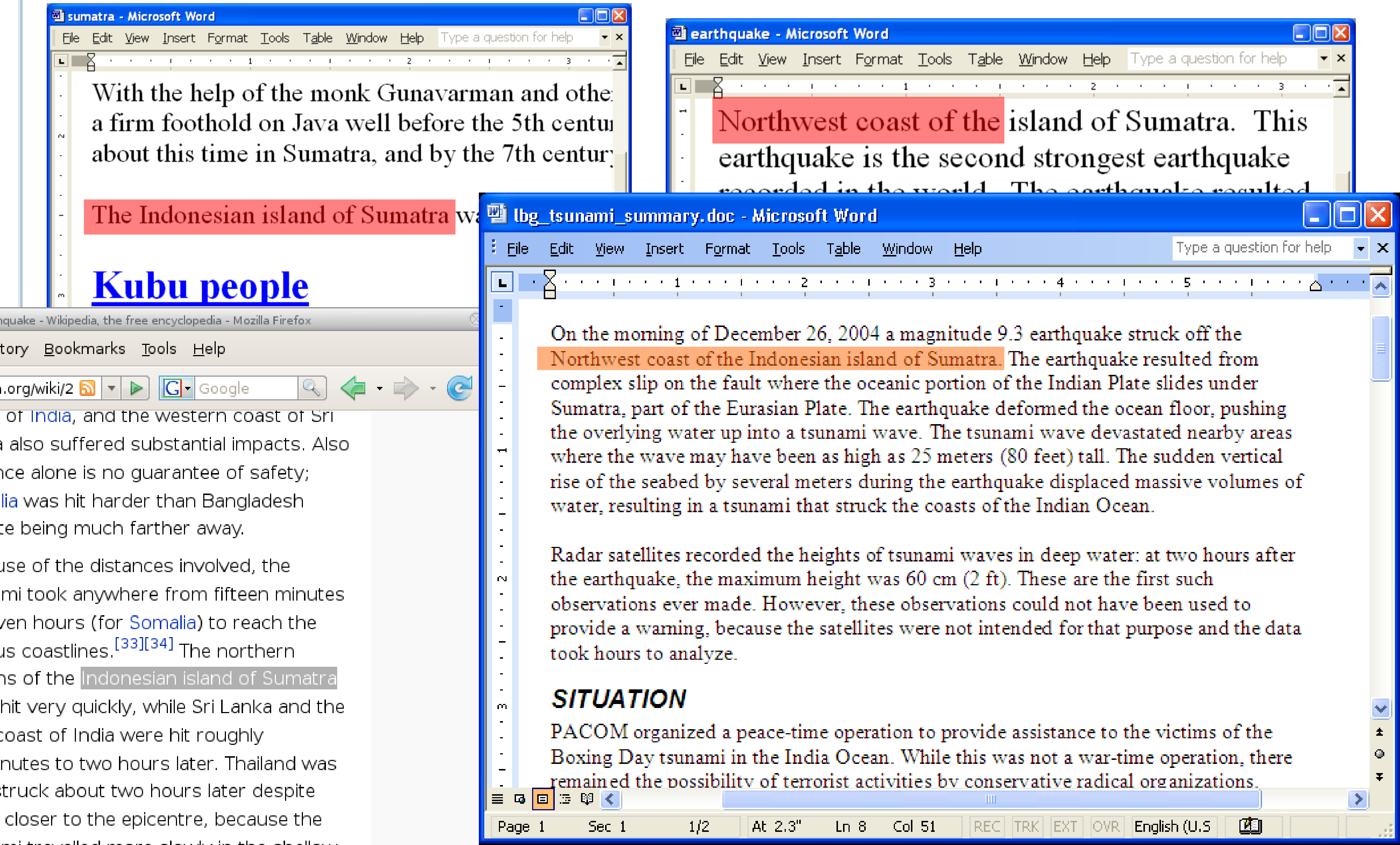
Open: “their hideout”

Filter the resulting string overlaps



Algorithm > Ranking

Overlap-based Ranking


sumatra - Microsoft Word
File Edit View Insert Format Tools Table Window Help
Type a question for help

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

The Indonesian island of Sumatra wa

Kubu people

earthquake - Microsoft Word
File Edit View Insert Format Tools Table Window Help
Type a question for help

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

quake - Wikipedia, the free encyclopedia - Mozilla Firefox
tory Bookmarks Tools Help
.org/wiki/2
Google

of India, and the western coast of Sri
a also suffered substantial impacts. Also
nce alone is no guarantee of safety;
lia was hit harder than Bangladesh
te being much farther away.

use of the distances involved, the
mi took anywhere from fifteen minutes
ven hours (for Somalia) to reach the
us coastlines.[33][34] The northern
ns of the Indonesian island of Sumatra
hit very quickly, while Sri Lanka and the
coast of India were hit roughly
nutes to two hours later. Thailand was
truck about two hours later despite
closer to the epicentre, because the
mi travelled more slowly in the shallow

lbg_tsunami_summary.doc - Microsoft Word
File Edit View Insert Format Tools Table Window Help
Type a question for help

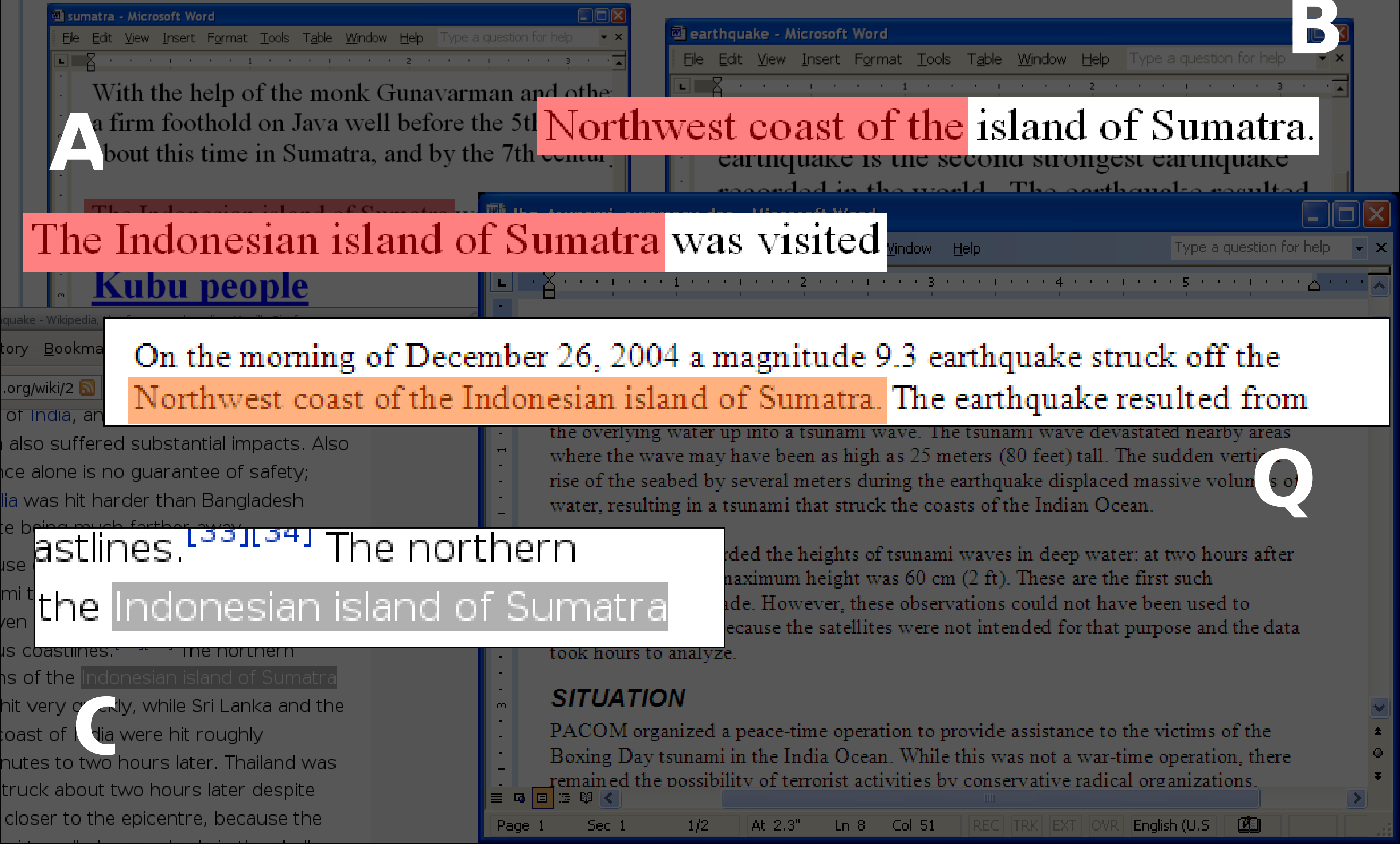
On the morning of December 26, 2004 a magnitude 9.3 earthquake struck off the Northwest coast of the Indonesian island of Sumatra. The earthquake resulted from complex slip on the fault where the oceanic portion of the Indian Plate slides under Sumatra, part of the Eurasian Plate. The earthquake deformed 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


sumatra - Microsoft Word
earthquake - Microsoft Word
File Edit View Insert Format Tools Table Window Help
Type a question for help
A
B
C
Q
Page 1 Sec 1 1/2 At 2.3" Ln 8 Col 51 REC TRK EXT OVR English (U.S

A: With the help of the monk Gunavarman and othe… a firm foothold on Java well before the 5th … bout this time in Sumatra, and by the 7th centu…

The Indonesian island of Sumatra was visited

Kubu people

B: Northwest coast of the island of Sumatra.

earthquake is the second strongest earthquake recorded in the world. The earthquake resulted

Q: 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 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 verti… rise of the seabed by several meters during the earthquake displaced massive volum…s of water, resulting in a tsunami that struck the coasts of the Indian Ocean.

…rded the heights of tsunami waves in deep water: at two hours after … maximum height was 60 cm (2 ft). These are the first such …ade. However, these observations could not have been used to …ecause the satellites were not intended for that purpose and the data took hours to analyze.

SITUATION

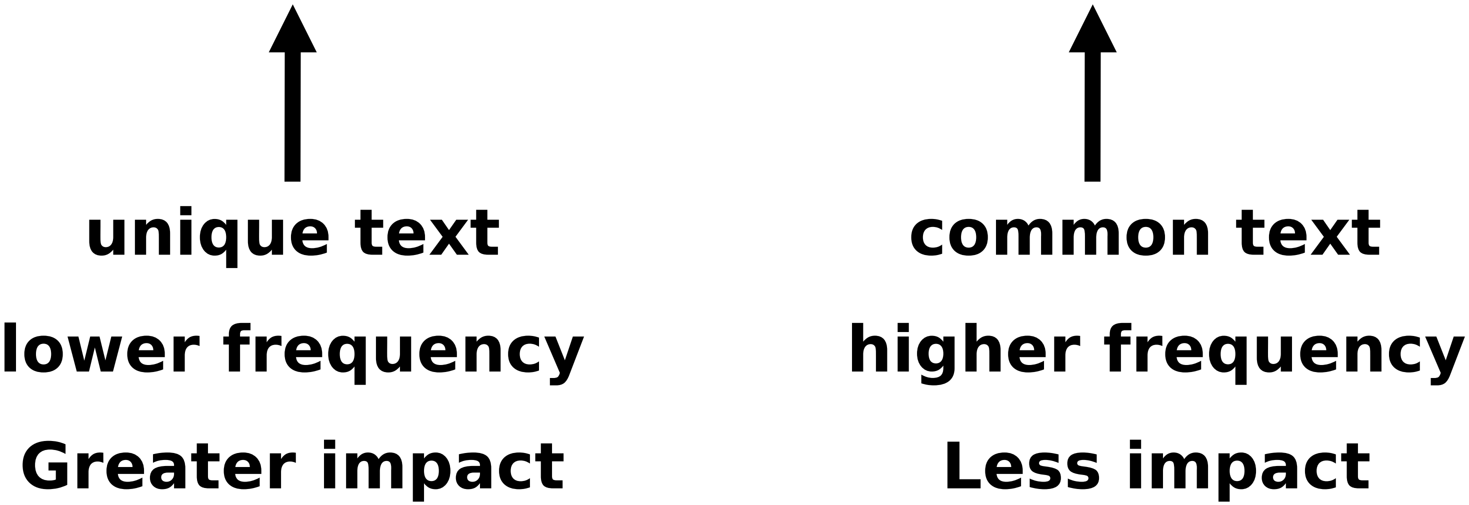
PACOM organized a peace-time operation to provide assistance to the victims of the Boxing Day tsunami in the India Ocean. While this was not a war-time operation, there remained the possibility of terrorist activities by conservative radical organizations.

C: astlines.[33][34] The northern the Indonesian island of Sumatra

of India, an… also suffered substantial impacts. Also …nce alone is no guarantee of safety; …lia was hit harder than Bangladesh …te being much farther away … us coastlines.[33][34] The northern …ns of the Indonesian island of Sumatra hit very quickly, while Sri Lanka and the coast of India were hit roughly …nutes to two hours later. Thailand was …truck about two hours later despite … closer to the epicentre, because the

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

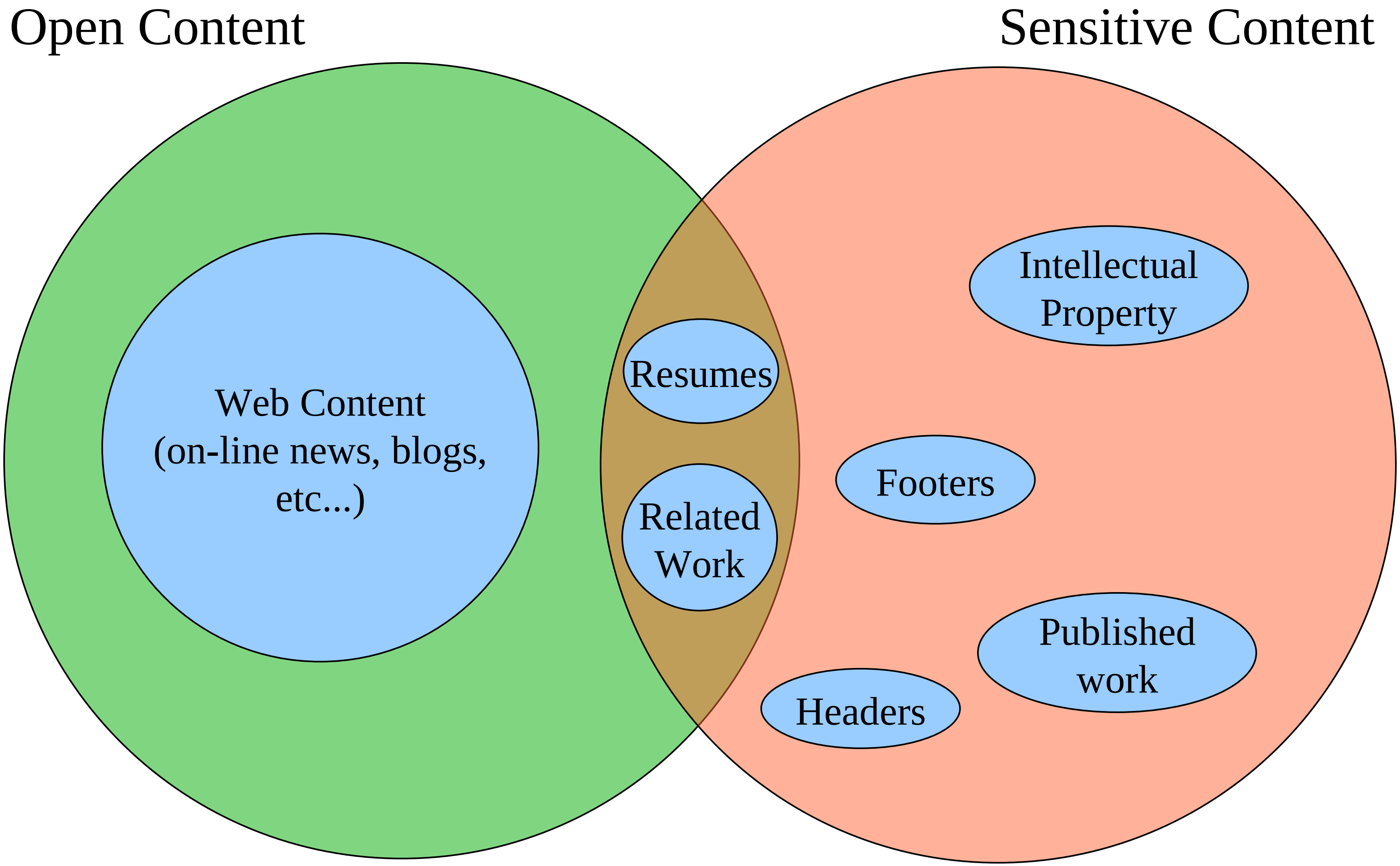
Less impact

Evaluation

InfoTracker was compared to Vector Space

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

Data Set


Open Content
Sensitive Content
Web Content (on-line news, blogs, etc...)
Resumes
Related Work
Intellectual Property
Footers
Published work
Headers

Data Set

272 SBIR proposals

234 historical proposals

38 query proposals

Oracle

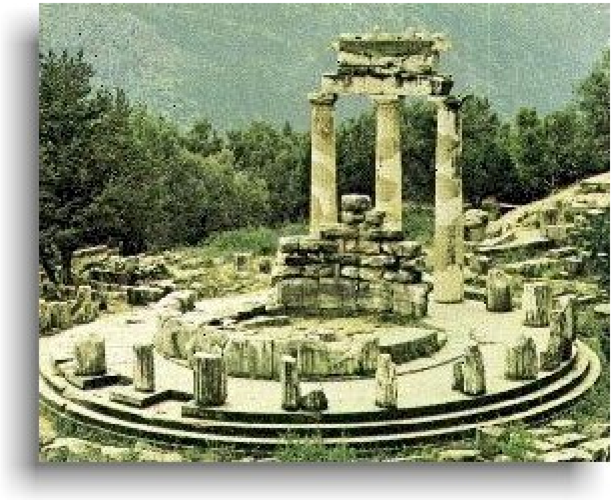


Image from: http://www.marketoracle.co.uk

Evaluation > Results

InfoTracker improved precision / recall

Algorithm	Precision	Recall
Vector Space	0.119	0.764
InfoTracker	0.167	0.913

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

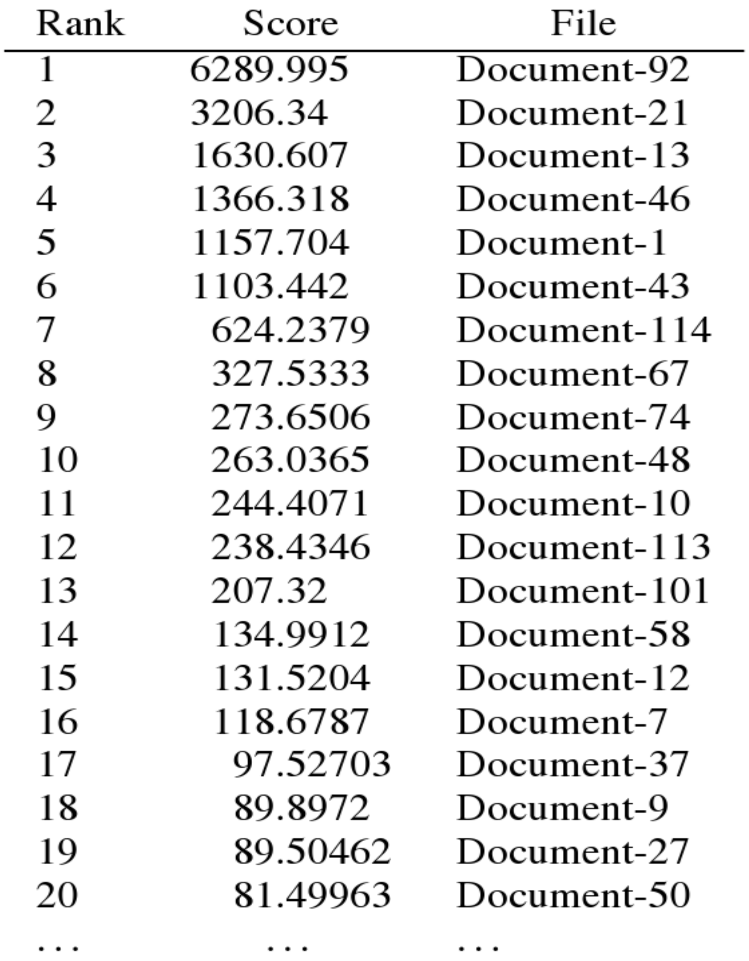
Active Learning

Document time stamps

Overlap size / encapsulation

Questions?

Calculating Precision / Recall

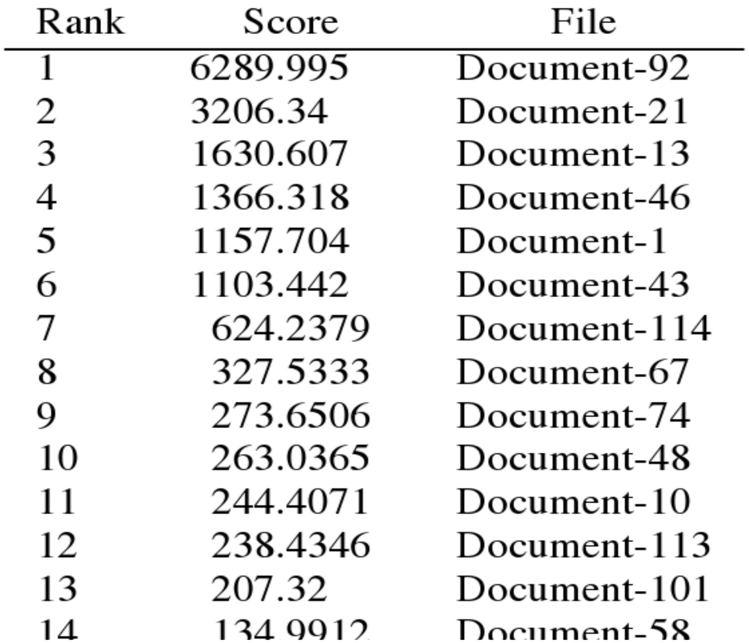


Rank	Score	File
1	6289.995	Document-92
2	3206.34	Document-21
3	1630.607	Document-13
4	1366.318	Document-46
5	1157.704	Document-1
6	1103.442	Document-43
7	624.2379	Document-114
8	327.5333	Document-67
9	273.6506	Document-74
10	263.0365	Document-48
11	244.4071	Document-10
12	238.4346	Document-113
13	207.32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

Calculating Precision / Recall

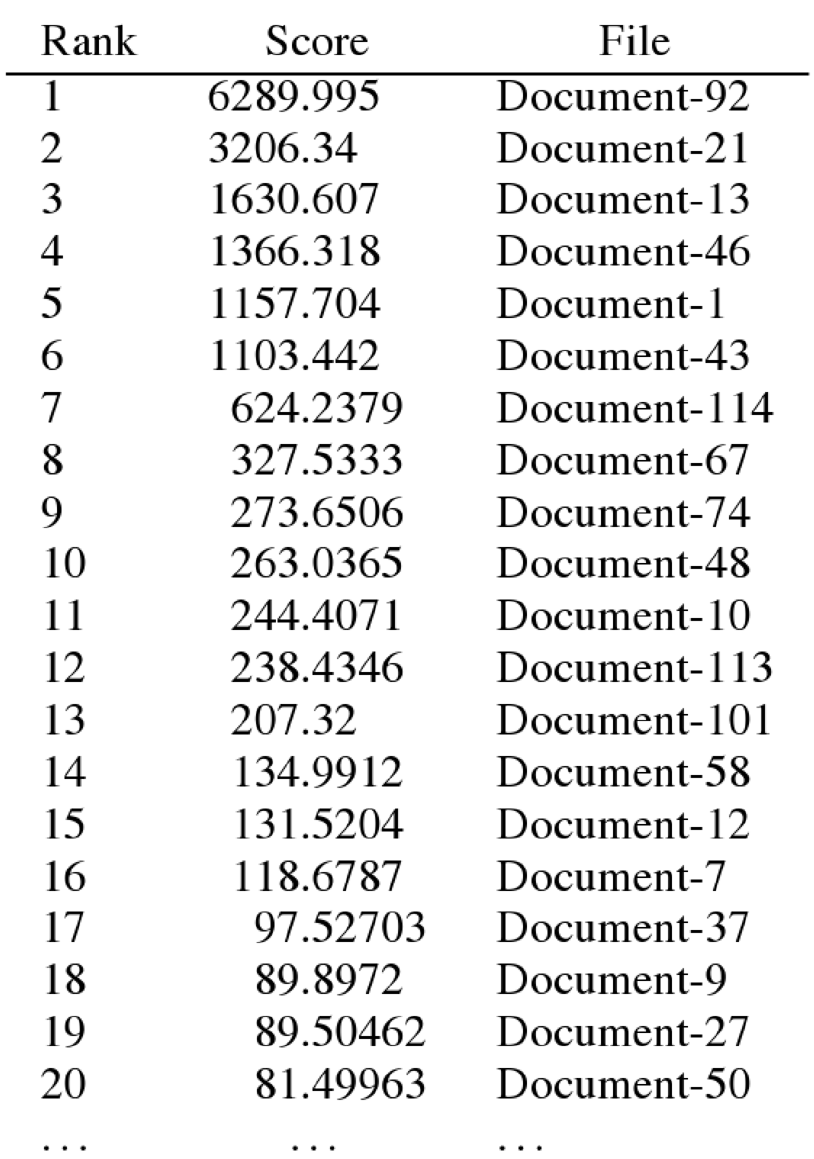
Consider the top 23 results.

(to allow for perfect recall)



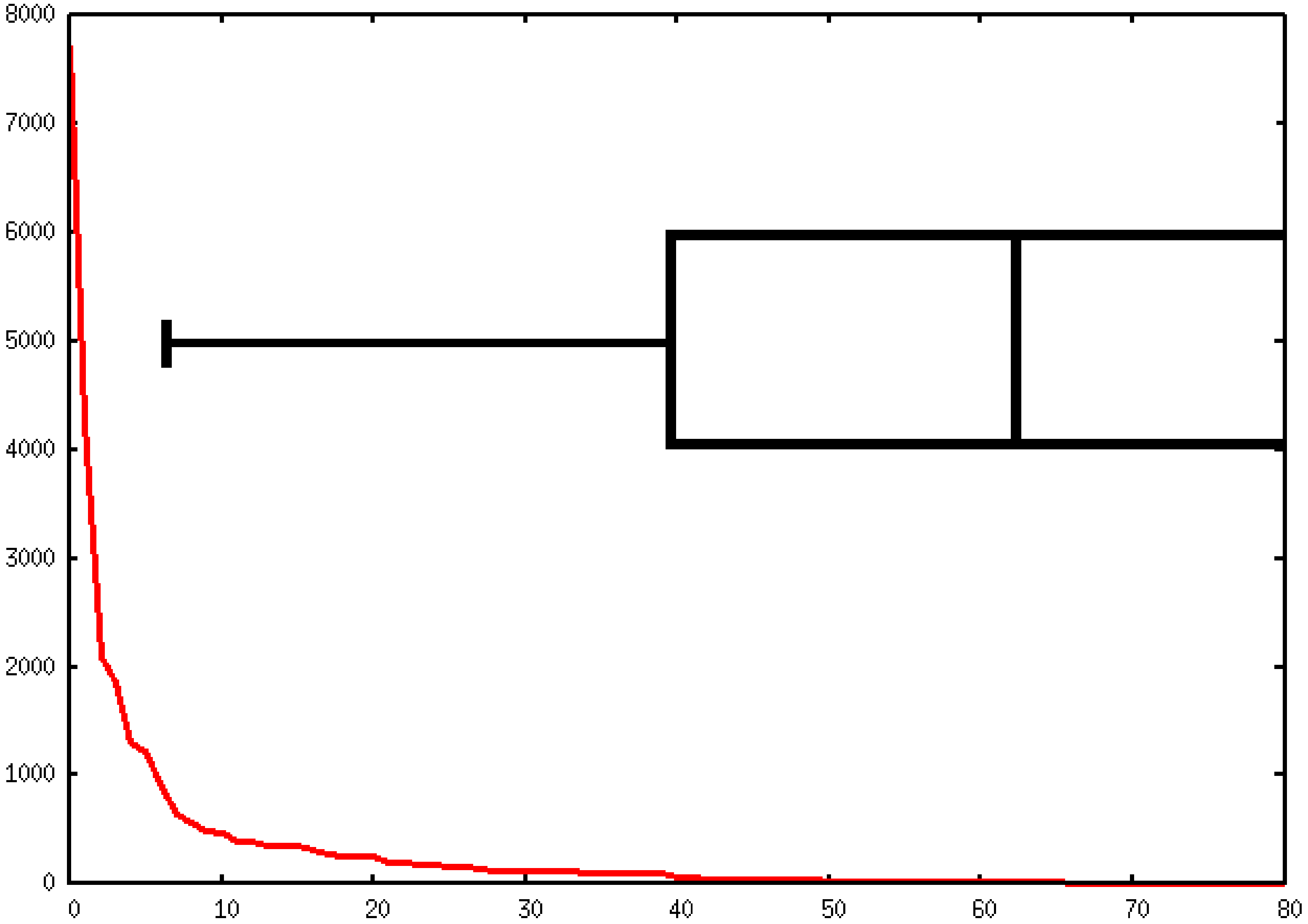
Rank	Score	File
1	6289.995	Document-92
2	3206.34	Document-21
3	1630.607	Document-13
4	1366.318	Document-46
5	1157.704	Document-1
6	1103.442	Document-43
7	624.2379	Document-114
8	327.5333	Document-67
9	273.6506	Document-74
10	263.0365	Document-48
11	244.4071	Document-10
12	238.4346	Document-113
13	207.32	Document-101
14	134.9912	Document-58

Ranking Scores Plummet Quickly

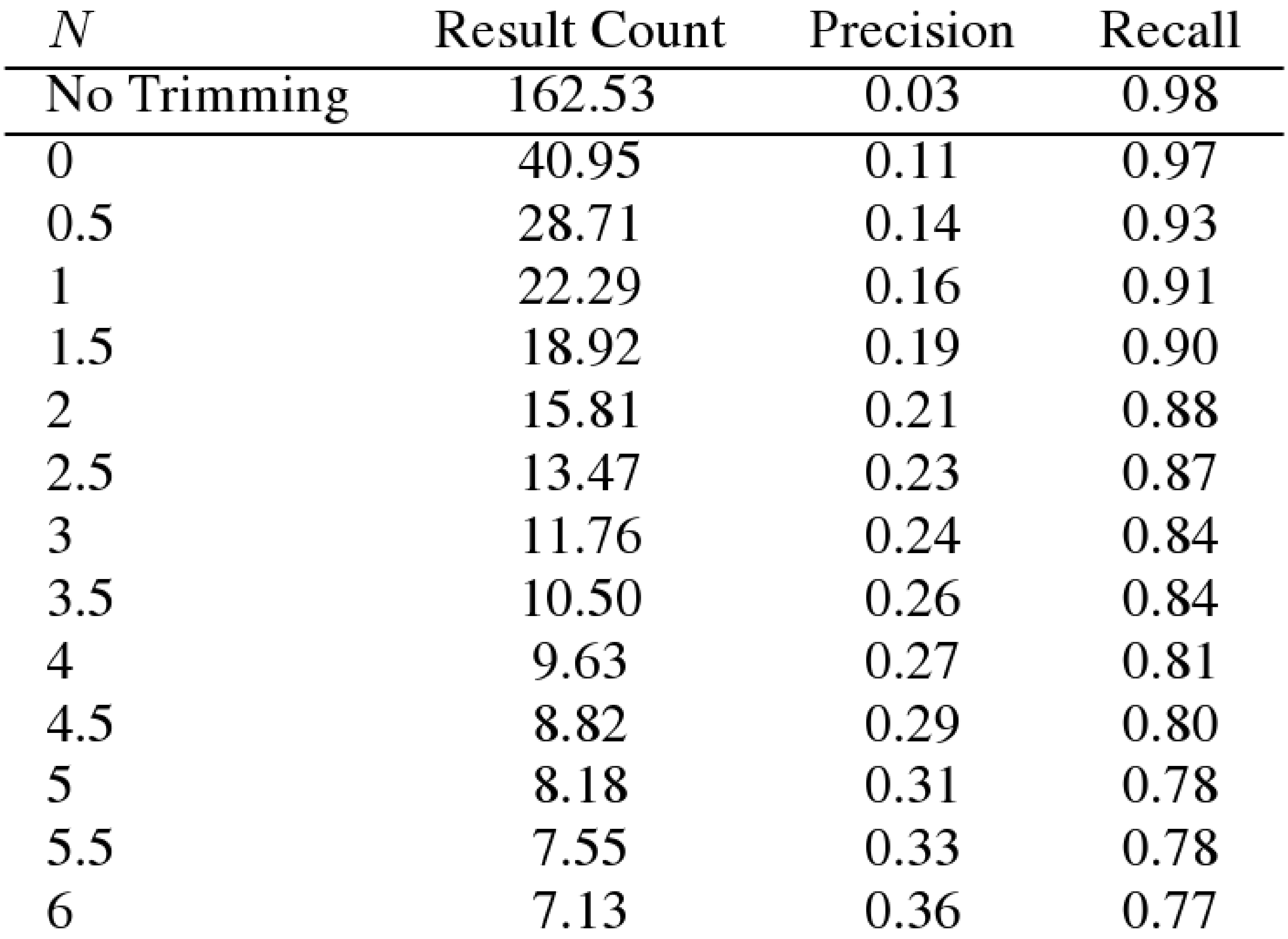


Rank	Score	File
1	6289.995	Document-92
2	3206.34	Document-21
3	1630.607	Document-13
4	1366.318	Document-46
5	1157.704	Document-1
6	1103.442	Document-43
7	624.2379	Document-114
8	327.5333	Document-67
9	273.6506	Document-74
10	263.0365	Document-48
11	244.4071	Document-10
12	238.4346	Document-113
13	207.32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

Ranking Scores Plummet Quickly



Trimming improves precision, retains recall



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