Unsupervised Ranking for Plagiarism Source Retrieval

Kyle Williams, Hung-Hsuan Chen, Sagnik Ray Choudhury and C. Lee Giles

Information Sciences and Technology
Computer Science and Engineering
The Pennsylvania State University

Core Ideas

- The union of the results of multiple queries has a higher probability of containing a true positive that each query individually
  - So submit multiple queries and combine results
- The ranking of the search results does not necessarily reflect the probability of a true positive
  - So re-rank results

PENNSTATE
1855

Approach to Source Retrieval

- Query generation
  - Partition document into 5 sentence paragraphs
  - Queries constructed from non-overlapping sequences of 10 POS tagged words
    - Verbs, nouns, adjectives
  - Multiple queries per paragraph
  - This approach performed better overall than TF-IDF and BM25

- Query Submission
  - Submit the first 3 queries for each paragraph
  - Return 3 results for each query and combine to form a single result set

PENNSTATE
1855

- Result Ranking
  - Re-rank results returned by the queries
  - For each result:
    - Get snippet
    - Calculate similarity between snippet and suspicious document based on 5-word overlaps

For a suspicious document d and result snippet s, the similarity Sim between the snippet and the suspicious documents is given by:

$$Sim(s, d) = S(s) \cap S(d)$$

Where S() is the set of 5-word sequences

- Re-rank results by similarity

PENNSTATE 1855

- Document Downloading
  - Download results in re-ranked order
    - Only consider results that have a similarity above some threshold
    - We required that snippets and the suspicious document must have at least 5 5-word sequences in common
  - Check with Oracle for match
    - Stop if match found
  - Don’t re-download documents that have previously been downloaded for a given suspicious document

PENNSTATE
1855

Results

Retrieval Performance			Workload		Time to 1st Detection		No	Runtime
F_1	Precision	Recall	Queries	Downloads	Queries	Downloads	Detection	
0.47	0.55	0.50	116.40	14.05	17.59	2.45	5	69781436

- Competitive precision and recall with highest F1
- We submitted a relatively large number of queries
  - But queries are cheap! (at least from a bandwidth perspective)
- Relatively few documents downloaded
  - Similarity threshold controlled this

PENNSTATE
1855

Future Ideas

- Better query construction and query selection
- Supervised ranking of search results?

PENNSTATE
1855