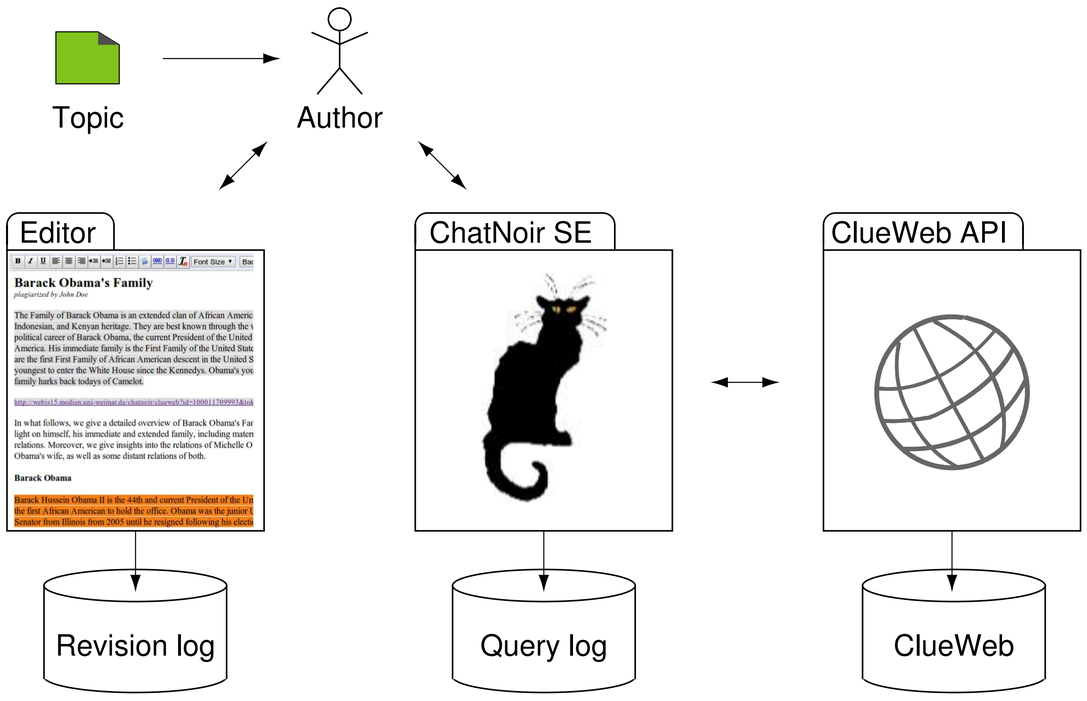
Exploratory Search Missions for TREC Topics

Martin Potthast Matthias Hagen Michael Völske Benno Stein

Bauhaus-Universität Weimar
99421 Weimar, Germany

Corpus Overview



We report on the construction of a new text reuse corpus comprising writing interactions and exploratory search missions.

- 150 essays (based on TREC Web Track topics 2009-2011)
- 12 professional writers hired on a crowdsourcing platform
- Long essay writing task, researching sources using a custom ClueWeb09 search engine
- Writing and search engine interactions recorded in high detail

Data Collection

Authors

Writer Demographics

Age		*Gender*		*Native language(s)*	
Minimum	24	Female	67%	English	67%
Median	37	Male	33%	Filipino	25%
Maximum	65			Hindi	17%
Academic degree		*Country of origin*		*Second language(s)*	
Postgraduate	41%	UK	25%	English	33%
Undergraduate	25%	Philippines	25%	French	17%
None	17%	USA	17%	Afrikaans, Dutch,	
n/a	17%	India	17%	German, Spanish,	
		Australia	8%	Swedish each	8%
		South Africa	8%	None	8%
Years of writing		*Search engines used*		*Search frequency*	
Minimum	2	Google	92%	Daily	83%
Median	8	Bing	33%	Weekly	8%
Standard dev.	6	Yahoo	25%	n/a	8%
Maximum	20	Others	8%		

Topics

Example topic:

Obama's family.
Write about President Barack Obama's family history, including genealogy, national origins, places and dates of birth, etc. Where did Barack Obama's parents and grandparents come from? Also include a brief biography of Obama's mother.

Original topic 001 of the TREC Web Track 2009:

Query. obama family tree

Description. Find information on President Barack Obama's family history, including genealogy, national origins, places and dates of birth, etc.

Sub-topic 1. Find the TIME magazine photo essay "Barack Obama's Family Tree."

Sub-topic 2. Where did Barack Obama's parents and grandparents come from?

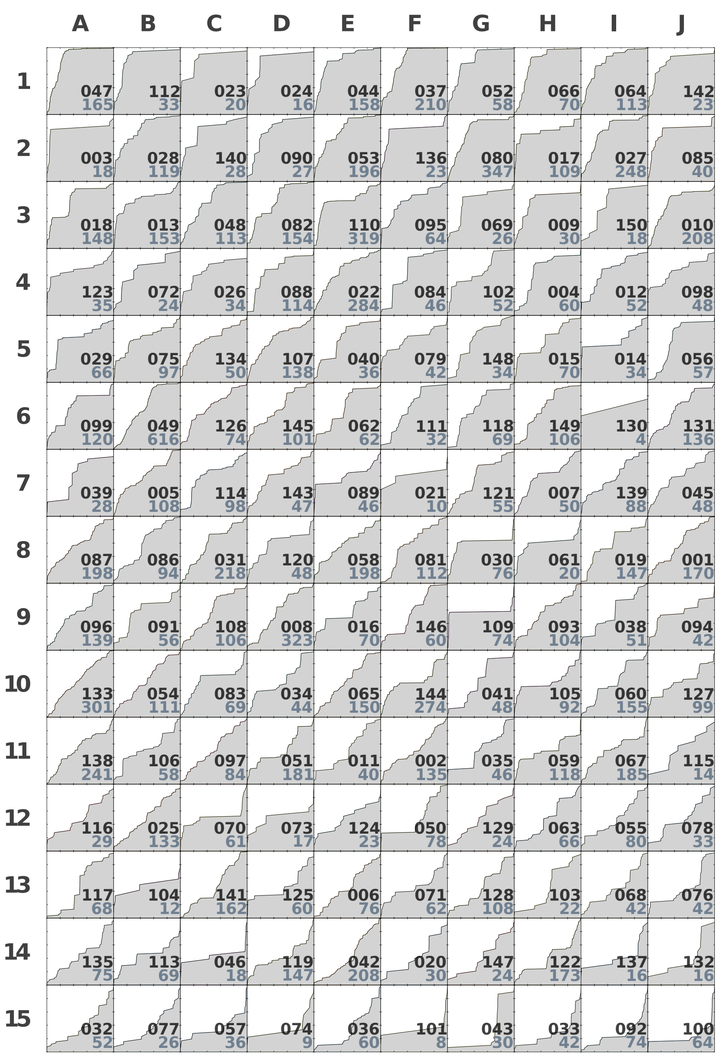
Sub-topic 3. Find biographical information on Barack Obama's mother.

Query log

Corpus Characteristic	**Distribution** min	avg	max	stdev	Σ
Writers					12
Topics					150
Topics / Writer	1	12.5	33	9.3	
Queries					13 651
Queries / Topic	4	91.0	616	83.1	
Clicks					16 739
Clicks / Topic	12	111.6	443	80.3	
Clicks / Query	0	0.8	76	2.2	
Sessions					931
Sessions / Topic	1	12.3	149	18.9	
Days					201
Days / Topic	1	4.9	17	2.7	
Hours					2068
Hours / Writer	3	129.3	679	167.3	
Hours / Topic	3	7.5	10	2.5	

Search mission data will be made available as the Webis-Query-Log-12 (`http://www.webis.de/research/corpora`)

Main Findings

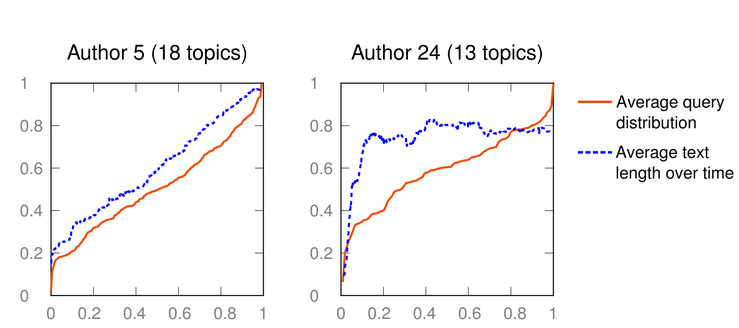


Spectrum of search behavior

- Percentage of queries submitted over time for all 150 search missions
- Ranges from majority of queries issued at the start of the task (A1) to most queries towards the end (J15)
- In between, sets of queries submitted in bursts (e.g F9) or linear increase (A10)

Correlation of searching and writing

- Evidence of distinct text reuse strategies (build-up and boil-down)
- Only the former clearly reflected in the query log



First Conclusions

- Query frequency by itself poor predictor of task completion
- Heavy reliance on search engine indicates need to better support exploratory tasks