Exploratory Search Missions for TREC Topics

EuroHCIR 2013

Dublin, 1 August 2013

Martin Potthast Matthias Hagen **Michael Völske** Benno Stein

Bauhaus-Universität Weimar

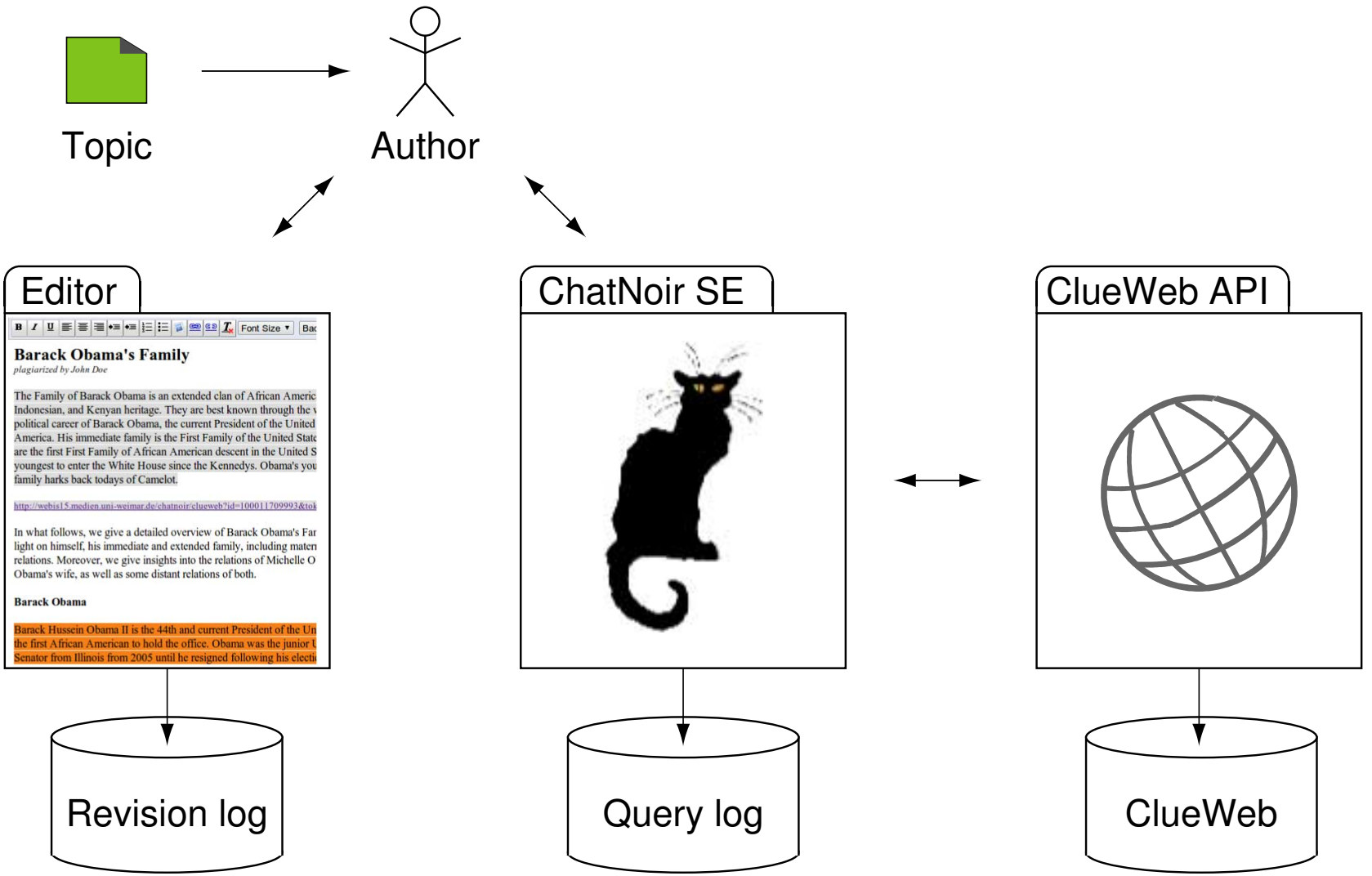
www.webis.de

Background

- Dataset for studying text reuse / plagiarism from web sources
- PAN workshop at CLEF
- Previous datasets: real plagiarism or algorithmically generated
- This time: crowdsourcing to freelance authors

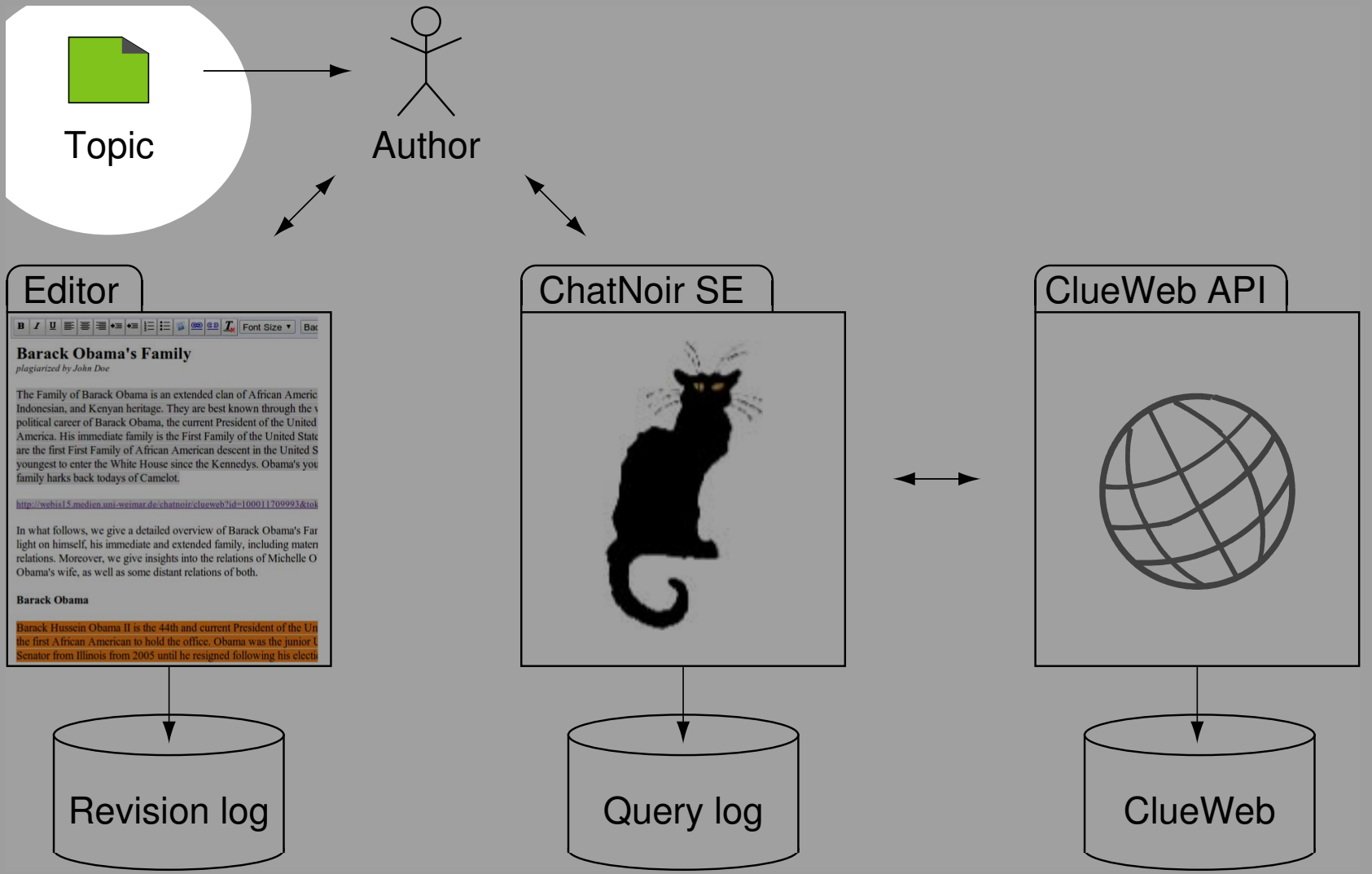
Background

Corpus Construction



Background

Topics



Background

Topics

Example topic:

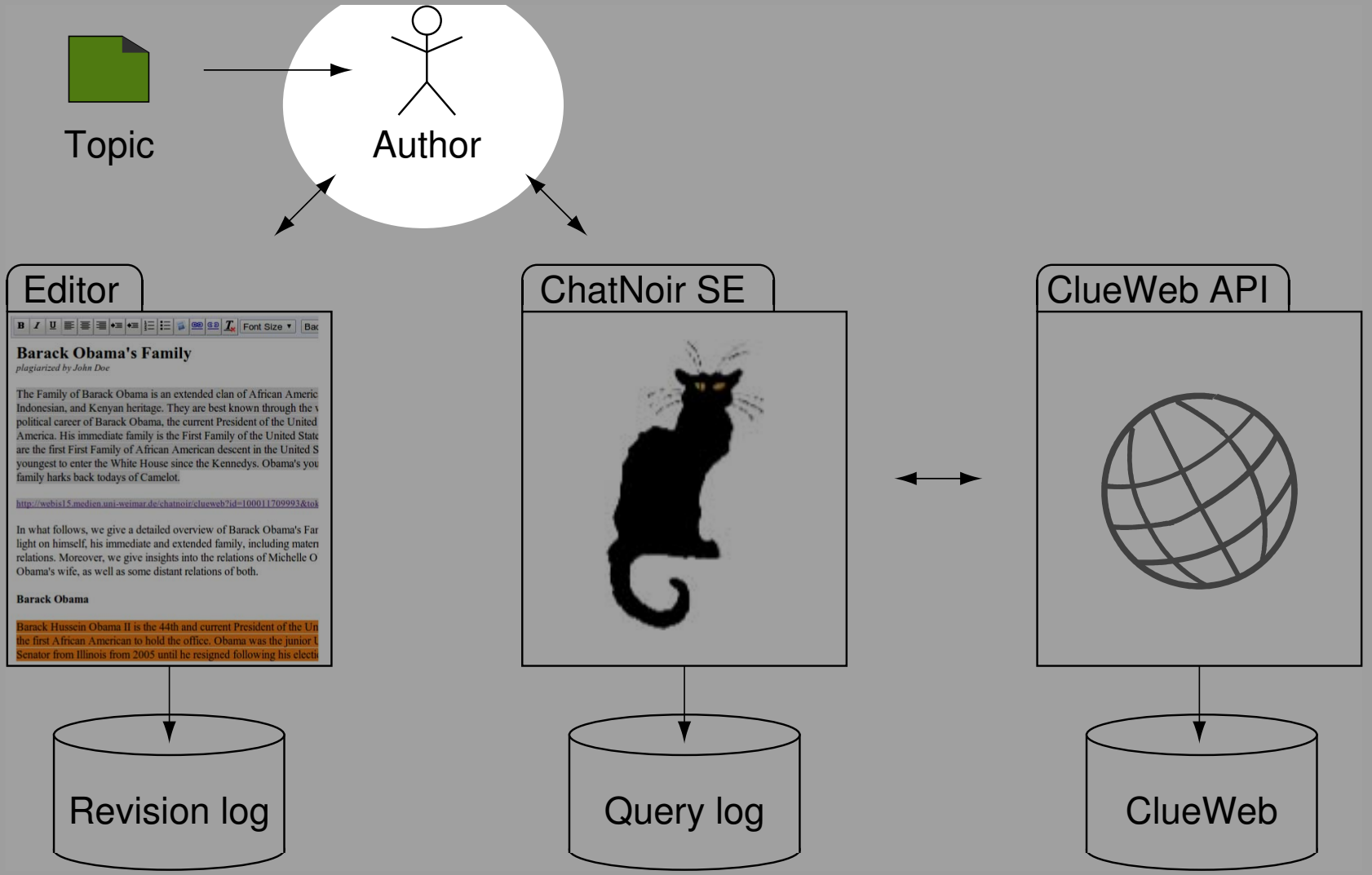
> *Obama's family*.
>
> Write about President Barack Obama's family history, including genealogy, national origins, places and dates of birth, etc. Where did Barack Obama's parents and grandparents come from? Also include a brief biography of Obama's mother.

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.

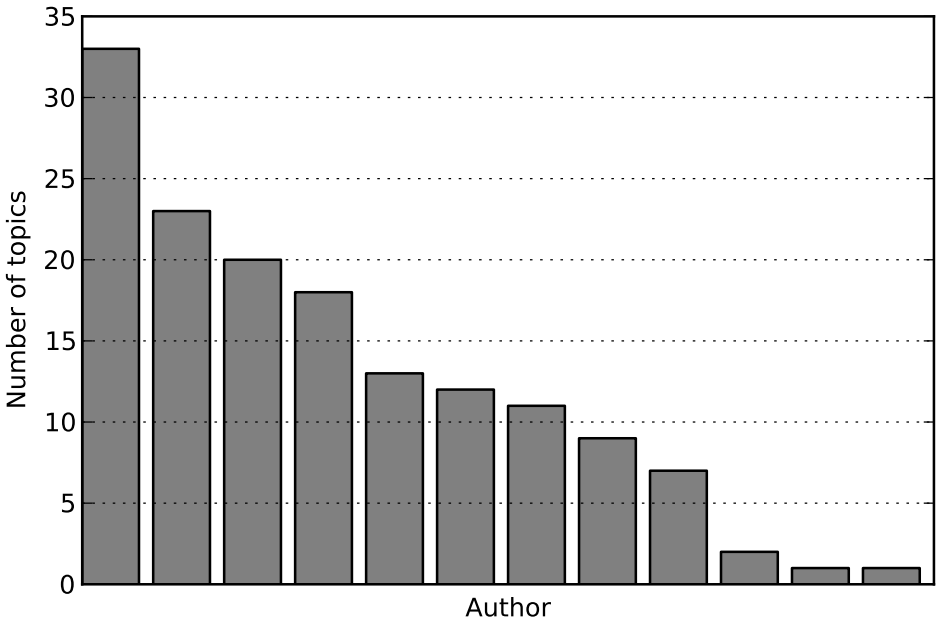
Background

Authors



Background

Authors

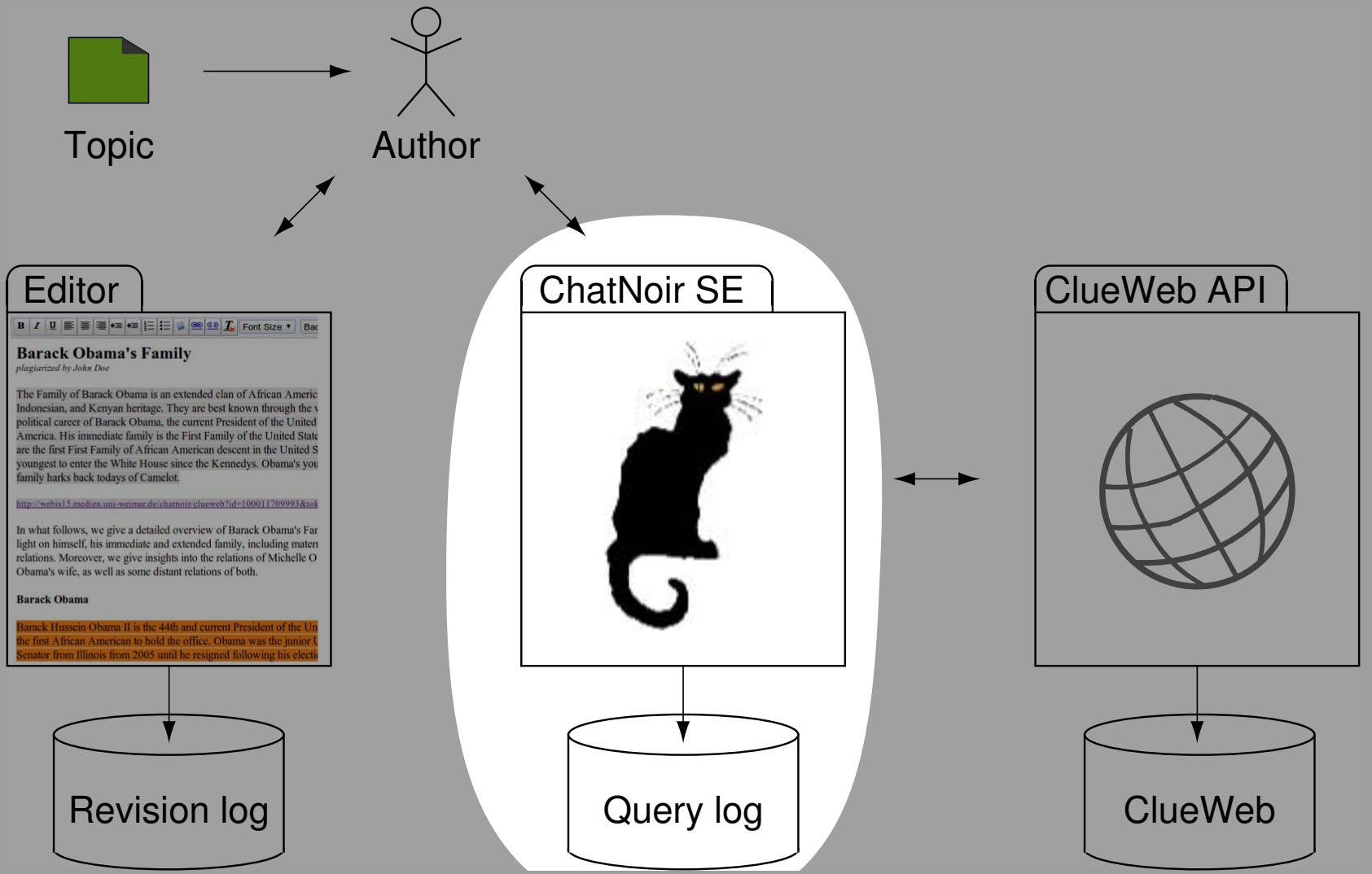


Author Demographics	
Age (Median)	37
Years Writing (Median)	8
Academic degree	
Postgrad	33%
Undergrad	25%
English	
Native	67%
Second Language	33%

- 12 Professional writers hired on oDesk [www.odesk.com]
- Fluent English speakers with several years writing experience

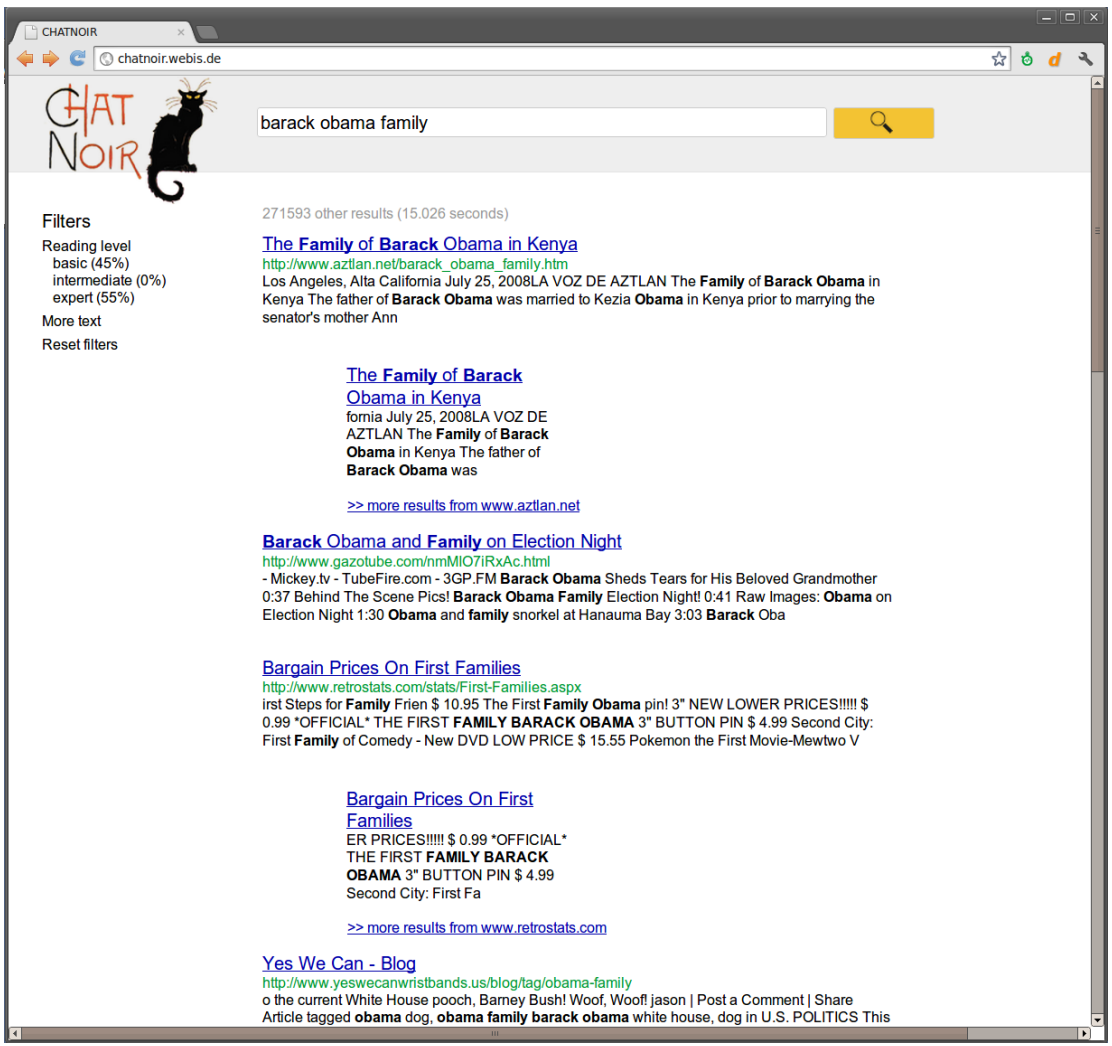
Background

Search engine and query log



Background

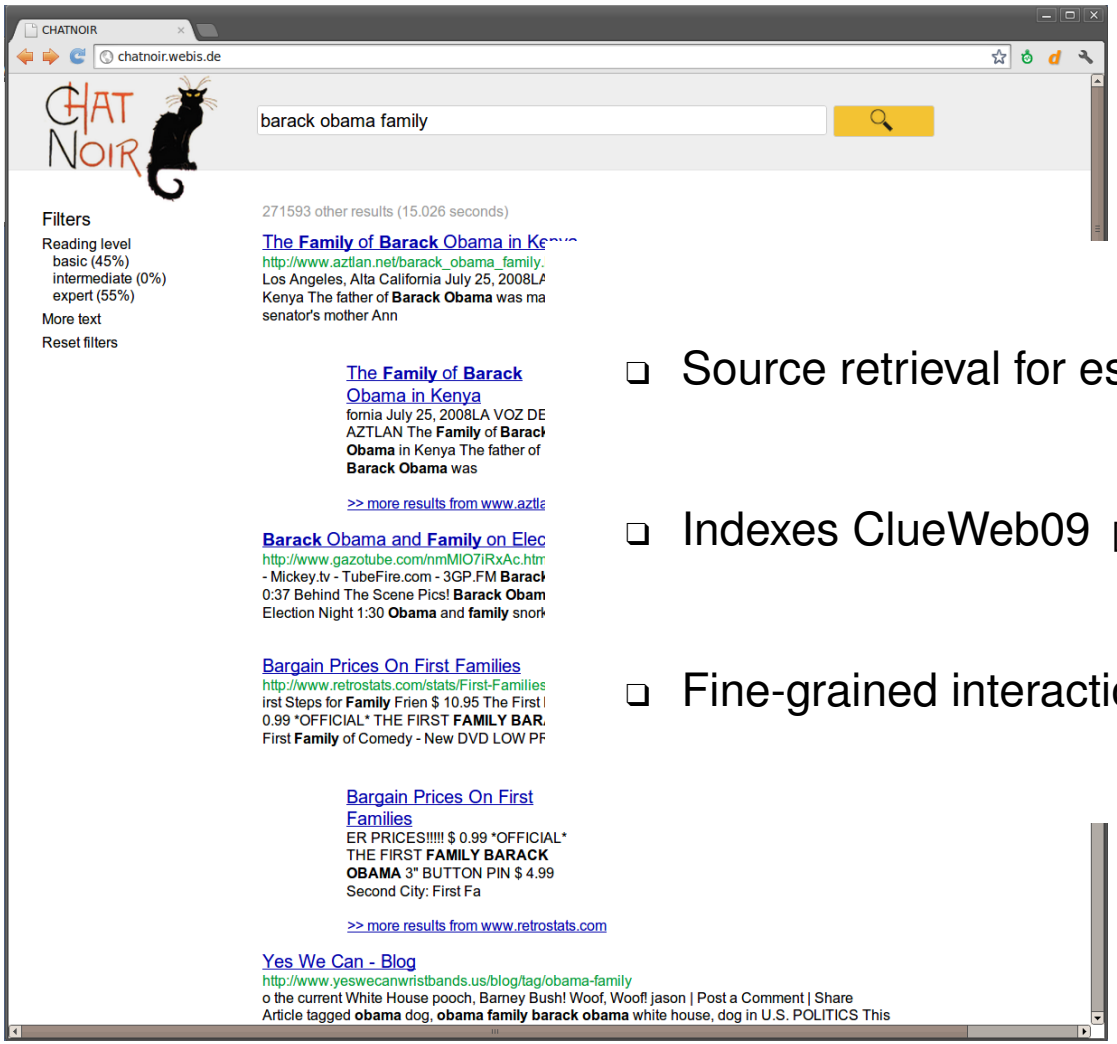
Clueweb search engine



[chatnoir.webis.de]

Background

Clueweb search engine



- Source retrieval for essay writing
- Indexes ClueWeb09 [lemurproject.org/clueweb09]
- Fine-grained interaction logs

[chatnoir.webis.de]

Query logs

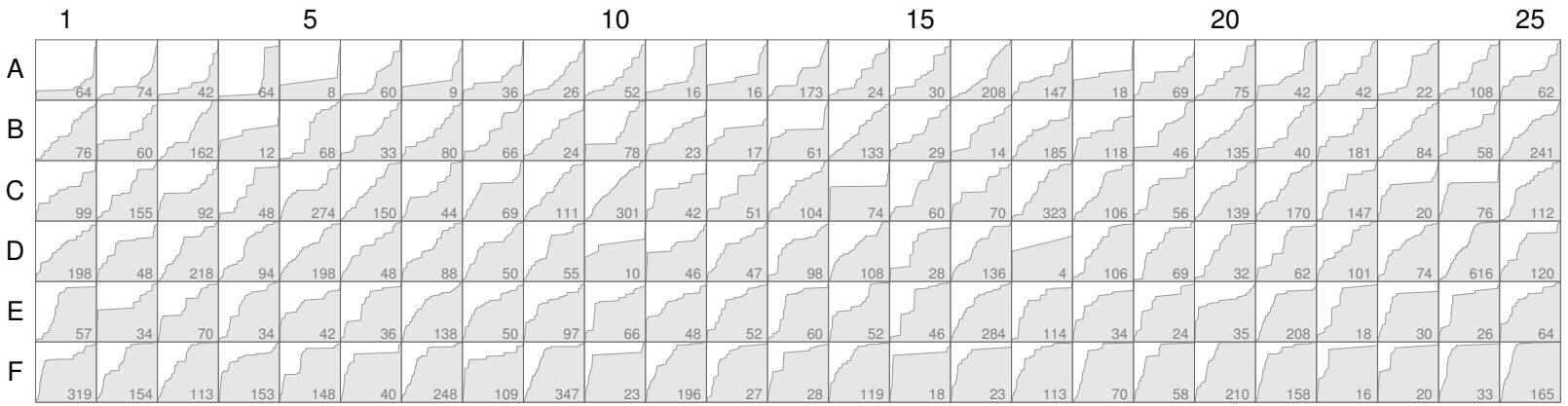
Basic stats

Corpus Characteristic	Distribution min	avg	max	stdev	$\sum$
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	

Findings so far

Findings so far

Distribution of Queries Over Time

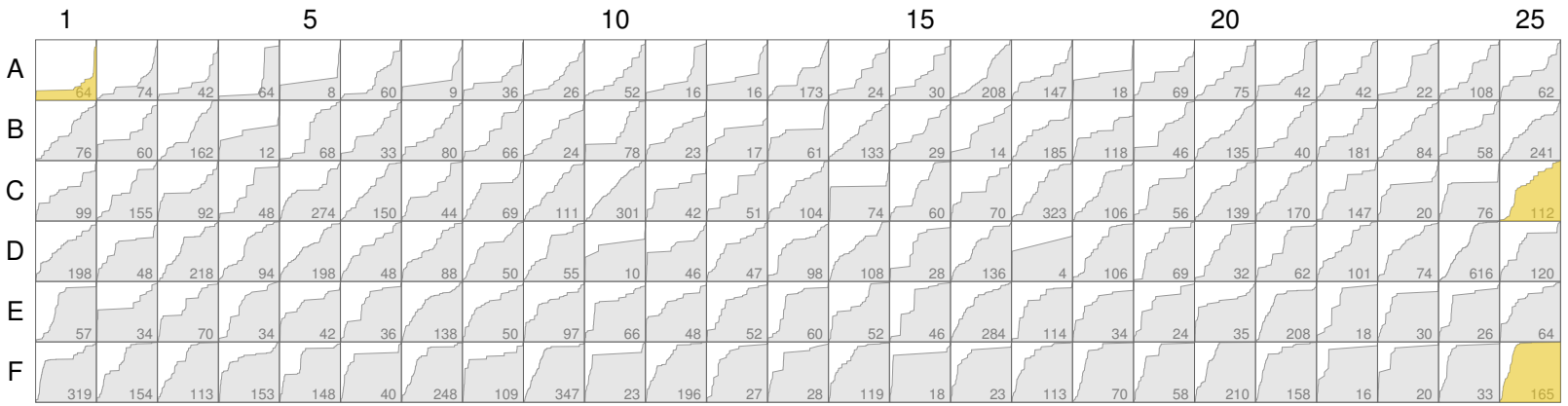


Distribution of queries over time.

- fraction of posed queries (y-axis) over elapsed time (x-axis) between the first query until essay completion
- each cell represents one of 150 essays
- the numbers denote the total amount of posed queries
- the cells are sorted by area under the curve

Findings so far

Distribution of Queries Over Time

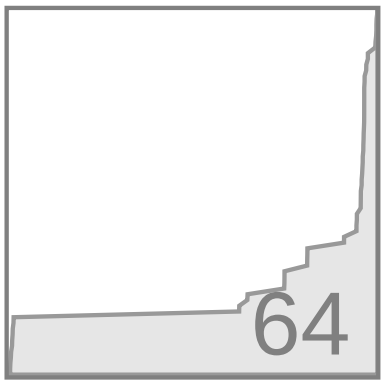


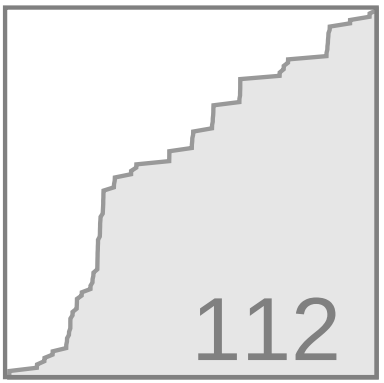
Distribution of queries over time.

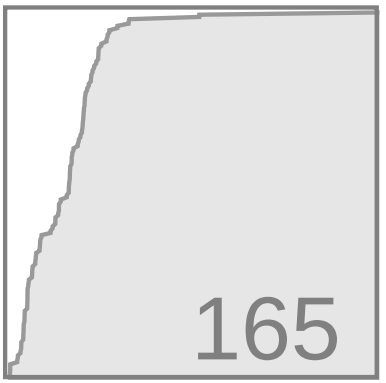
- fraction of posed queries (y-axis) over elapsed time (x-axis) between the first query until essay completion
- each cell represents one of 150 essays
- the numbers denote the total amount of posed queries
- the cells are sorted by area under the curve

Findings so far

Distribution of Queries Over Time







Distribution of queries over time.

- fraction of posed queries (y-axis) over elapsed time (x-axis) between the first query until essay completion
- each cell represents one of 150 essays
- the numbers denote the total amount of posed queries
- the cells are sorted by area under the curve

Summary

1. Corpus of essay writing, focus on text reuse
2. 150 exploratory search missions for TREC web track topics
3. First appraisal indicates wide spectrum of search behavior
4. Data will be publicly available: http://webis.de/research/corpora

Summary

1. Corpus of essay writing, focus on text reuse
2. 150 exploratory search missions for TREC web track topics
3. First appraisal indicates wide spectrum of search behavior
4. Data will be publicly available: http://webis.de/research/corpora

Thank you for your attention!