How Writers Search

Analyzing the Search and Writing Logs of Non-fictional Essays

CHIIR 2016

16 March 2016

Martin Potthast  Matthias Hagen  Michael Völske  Jakob Gomoll  Benno Stein
Bauhaus-Universität Weimar

www.webis.de

<firstname>.<lastname>@uni-weimar.de
How Writers Search

Analyzing the Search and Writing Logs of Non-fictional Essays

CHIIR 2016

16 March 2016

Outline

· Our Dataset: Webis-TRC-12
· Searching and Writing Strategies
· Query Formulation
Webis-TRC-12 (Text Reuse Corpus)

- Text reuse from web sources; lack of public data.

- Source retrieval as well as paraphrasing.

- Beyond text reuse.

- Essay writing task with content reuse; recording search engine interactions as well as text progress.
Dataset Construction
Dataset Construction

TREC Topics
(150)

ClueWeb09
Dataset Construction

TREC Topics (150) → Authors (12)

ClueWeb09
Dataset Construction

TREC Topics (150) → Authors (12) → ClueWeb09 → Search Engine → Interaction Log
Dataset Construction

TREC Topics (150)

Authors (12)

ClueWeb09

Search Engine

Editor

Interaction Log

Revision Log
Dataset Construction

TREC Topics (150)

ClueWeb09

Search Engine

Authors (12)

Interaction Log

70 Queries
90 Clicks

Revision Log

5000 Words
3000 Rev. per essay

© Michael Völske 2016
Writing and Search Behavior

Research Questions

1. Are different writing and searching strategies distinguishable?

2. How consistent are the authors?
Essay Writing Visualization

1 hour and 100% final essay length

Query, result browsing
Result click, reading
Writing session
Short pause
Essay length (relative)

Topic 27

Topic 57

Topic 29

1 hour and 2 hours and 1 hour and
100% final essay length 200% final essay length

100% final essay length

□
□
□
Writing and Search Behavior

- **Essay Growth**: percentage of lengthening revisions

- **Paste Regularity**: variance of number of revisions between adding sources
Writing and Search Behavior

Authors

Essay growth

Paste regularity (Variance)

1 2 5 6 7 14 17 18 20 21 24 25

© Michael Völske 2016
Writing and Search Behavior

Authors

“Boil-down”

“Build-up”

Essay growth

Paste regularity (Variance)
Writing and Search Behavior

### Essay growth vs. Paste regularity (Variance)

Authors:
- 1
- 2
- 5
- 6
- 7
- 14
- 17
- 18
- 20
- 21
- 24
- 25

© Michael Völske 2016
Writing and Search Behavior

Authors

Essay growth

Paste regularity (Variance)

© Michael Völske 2016
Query Formulation
Query Formulation

Exploratory Search

- Complex, open-ended search task
- Increasing topic understanding
- Iterative refinement
Origins of New Query Terms

- 3,718 unique query terms
  - Task description
  - Search Result Page
  - Documents
  - Prior Knowledge
Origins of New Query Terms

- 3,718 unique query terms
  - Task description: 24%
  - Search Result Page: 37%
  - Documents: 31%
  - Prior Knowledge: 8%
Origins of New Query Terms

![Graph showing the origins of new query terms](image-url)
Origins of New Query Terms

First 3 clicks after 6th query
Origins of New Query Terms
New terms in 9th query occur in first 3 clicked results
Origins of New Query Terms

Change of sub-topic
Summary
A new public dataset to learn how writers search: 150 essays written using web sources, fine-grained revision history and search engine interactions.

Two distinct searching and writing strategies

Authors expand subject knowledge while working

Data is available:
www.webis.de → “Corpora” → “Webis-TRC-12”
Summary

- A new public dataset to learn how writers search: 150 essays written using web sources, fine-grained revision history and search engine interactions.

- Two distinct searching and writing strategies

- Authors expand subject knowledge while working

- Data is available: www.webis.de → “Corpora” → “Webis-TRC-12”

- Thank you!

- Questions?