

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)

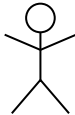
ClueWeb09



Dataset Construction



TREC Topics
(150)

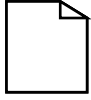


Authors (12)

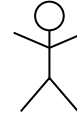
ClueWeb09



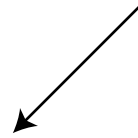
Dataset Construction



TREC Topics
(150)



Authors (12)



ClueWeb09

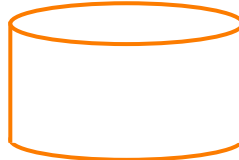


Search Engine

The screenshot shows a search engine interface with the query 'Obama family tree'. The results include links to various websites such as 'iAnMica.com', 'Genealogy Ireland', 'Japan America Society of Greater Philadelphia', and 'McNish Obama's Family Tree has Roots in a Carolina Slave Plantation'. The search engine logo is 'CHAT Noir'.



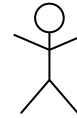
Interaction Log



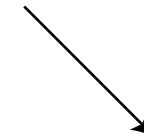
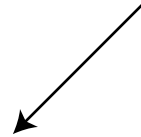
Dataset Construction



TREC Topics
(150)



Authors (12)



ClueWeb09

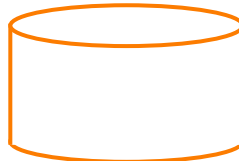


Search Engine

The screenshot shows search results for 'Obama family tree'. The top result is from 'iAfrica.com' with the title 'Country to Launch Obama Family Tree (Page 1 of 1)'. Other results include 'Genealogy Ireland - Obama Irish Family Tree', 'Japan America Society of Greater Philadelphia - Membership', 'Japan American Society of Greater Philadelphia - Obama', 'Michelle Obama's Family Tree has Roots in a Carolina Slave Plantation', and 'Ancestry.com - Reading This Family Tree: Barack Obama was born on 4 August 1961'.



Interaction Log

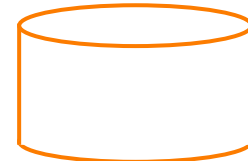


Editor


The screenshot shows a web editor interface with a document titled 'Barack Obama's Family' by John Doe. The text in the editor reads: 'The Family of Barack Obama is an extended clan of African American, Indonesian, and Kenyan heritage. They are best known through the political career of Barack Obama, the current President of the United States. His immediate family is the First Family of the United States the first First Family of African American descent in the United States youngest to enter the White House since the Kennedys. Obama's you family harks back to days of Camelot.' Below this is a URL: 'http://vash115.medien.uni-weimar.de/chatnoir/clueweb?d=100011709993&id='. Further down, it says 'In what follows, we give a detailed overview of Barack Obama's Far light on himself, his immediate and extended family, including maternal relations. Moreover, we give insights into the relations of Michelle O Obama's wife, as well as some distant relations of both.' At the bottom, there is a highlighted sentence: 'Barack Hussein Obama II is the 44th and current President of the United States, the first African American to hold the office. Obama was the junior U.S. Senator from Illinois from 2005 until he resigned following his election'.

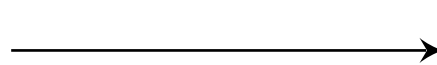


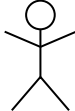
Revision Log



Dataset Construction


TREC Topics
(150)




Authors (12)

ClueWeb09



Search Engine

The screenshot shows a search engine interface with the query 'Obama family tree'. The results list several links, including 'iAfrica.com: Kenya: Country to Launch Obama Family Tree', 'Genealogy Ireland - Obama Irish Family Tree', 'Japan America Society of Greater Philadelphia - Membership', 'Japan American Society of Greater Philadelphia - Membership', 'Michelle Obama's Family Tree has Roots in a Carolina Slave Plantation', and 'Ancestry.com: Barack Obama: The Family'. The interface includes a search bar, filters, and a list of search results with brief descriptions.



Interaction Log

70 Queries
90 Clicks

Editor

The screenshot shows a rich text editor interface. The title is 'Barack Obama's Family' and it is attributed to 'John Doe'. The main text reads: 'The Family of Barack Obama is an extended clan of African American, Indonesian, and Kenyan heritage. They are best known through the political career of Barack Obama, the current President of the United States. His immediate family is the First Family of the United States the first First Family of African American descent in the United States youngest to enter the White House since the Kennedys. Obama's you family harks back to days of Camelot.' There is a link to a ClueWeb09 document. Below the main text, there is a section titled 'Barack Obama' with a snippet of text: 'Barack Hussein Obama II is the 44th and current President of the United States, the first African American to hold the office. Obama was the junior U.S. Senator from Illinois from 2005 until he resigned following his election as President in 2008.'



Revision Log

5000 Words
3000 Rev.

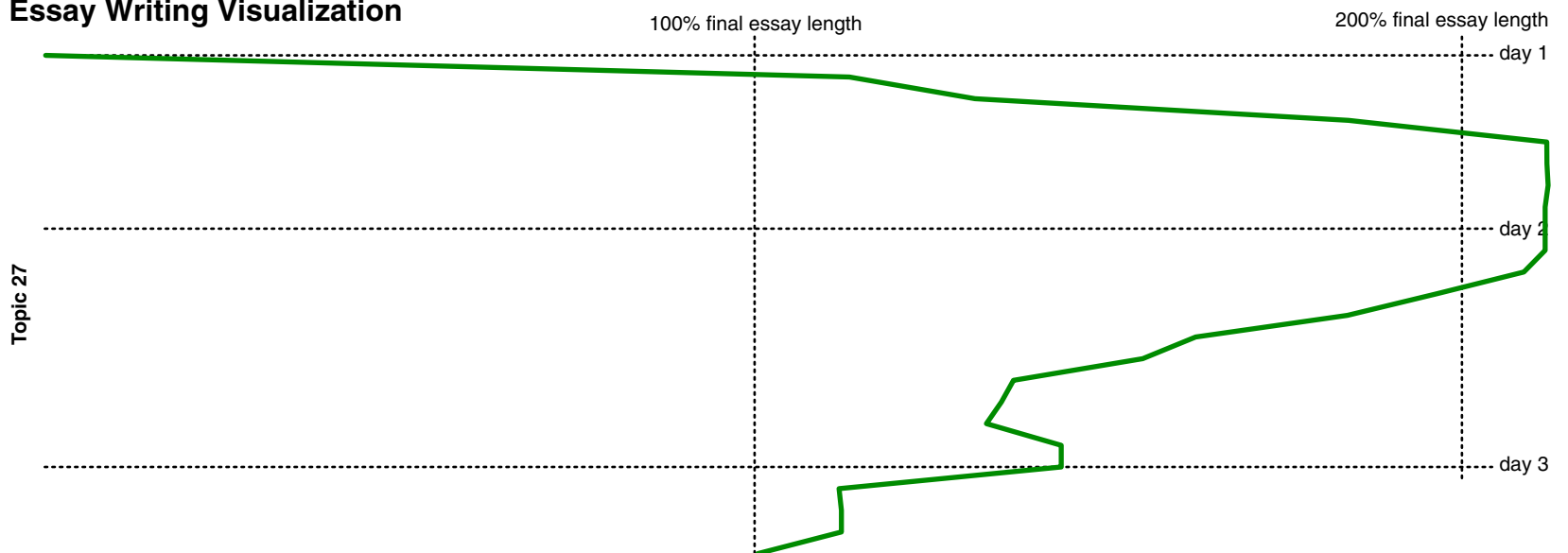
per essay

Writing and Search Behavior

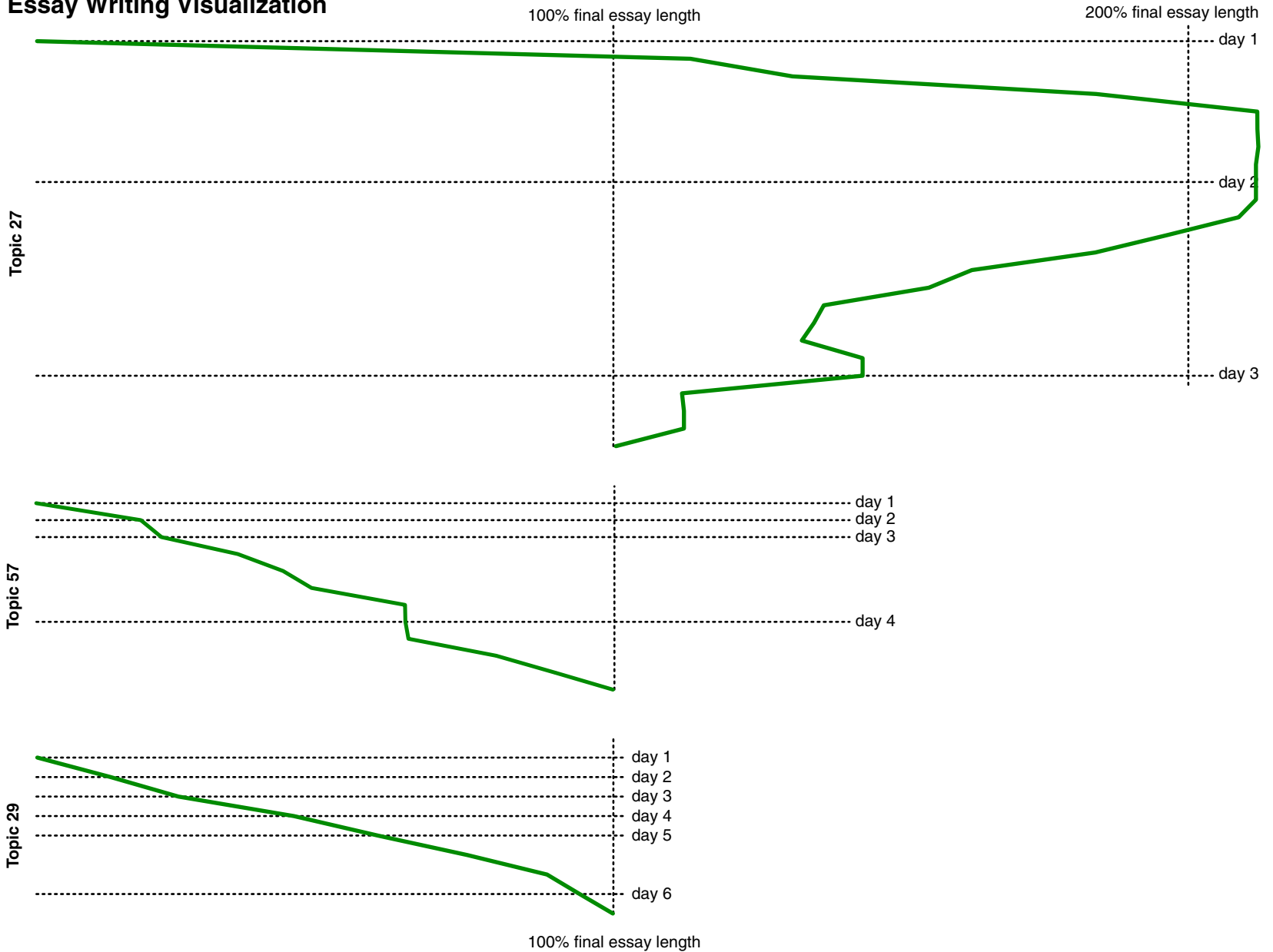
Research Questions

1. Are different writing and searching strategies distinguishable?
2. How consistent are the authors?

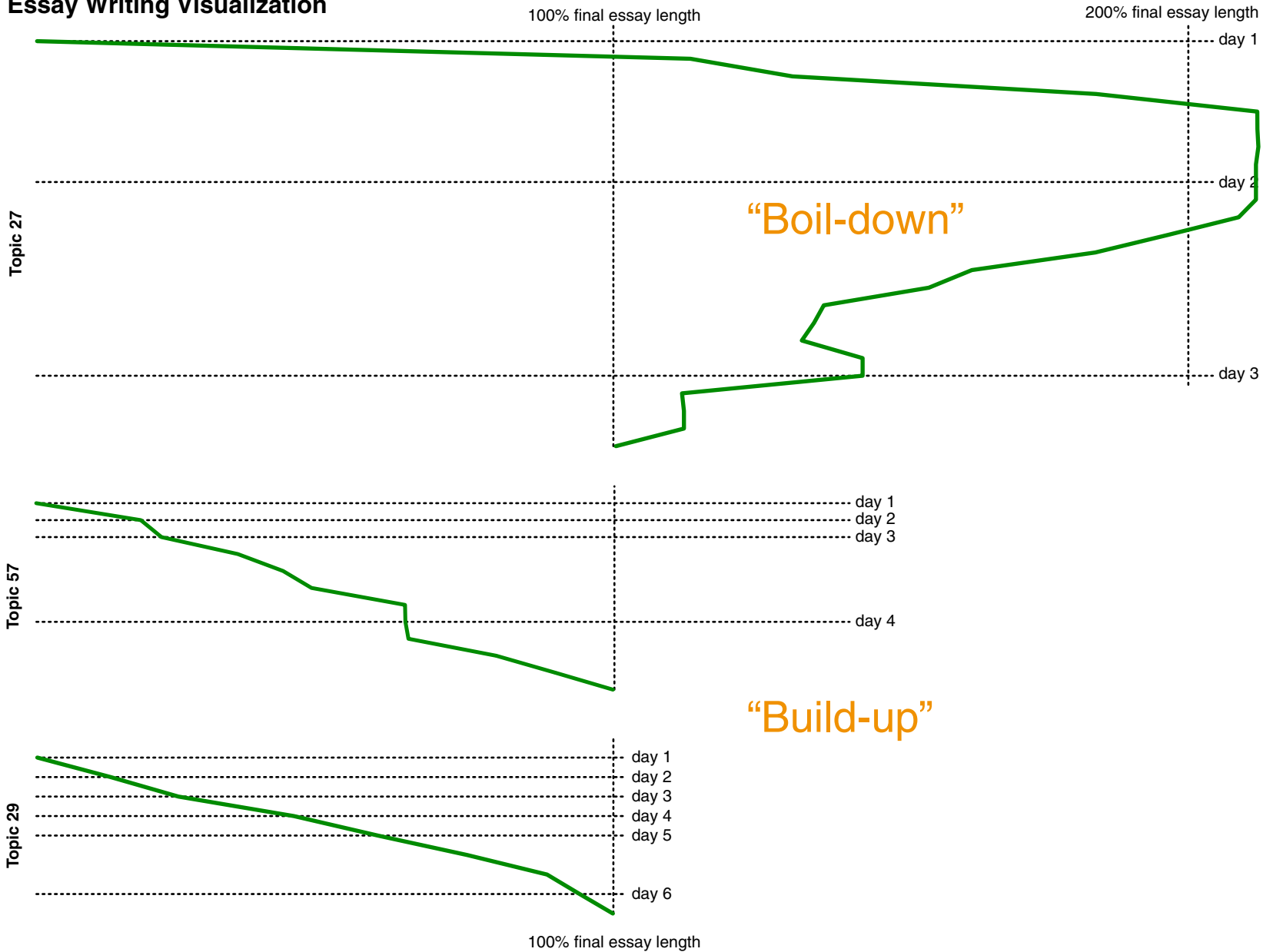
Essay Writing Visualization



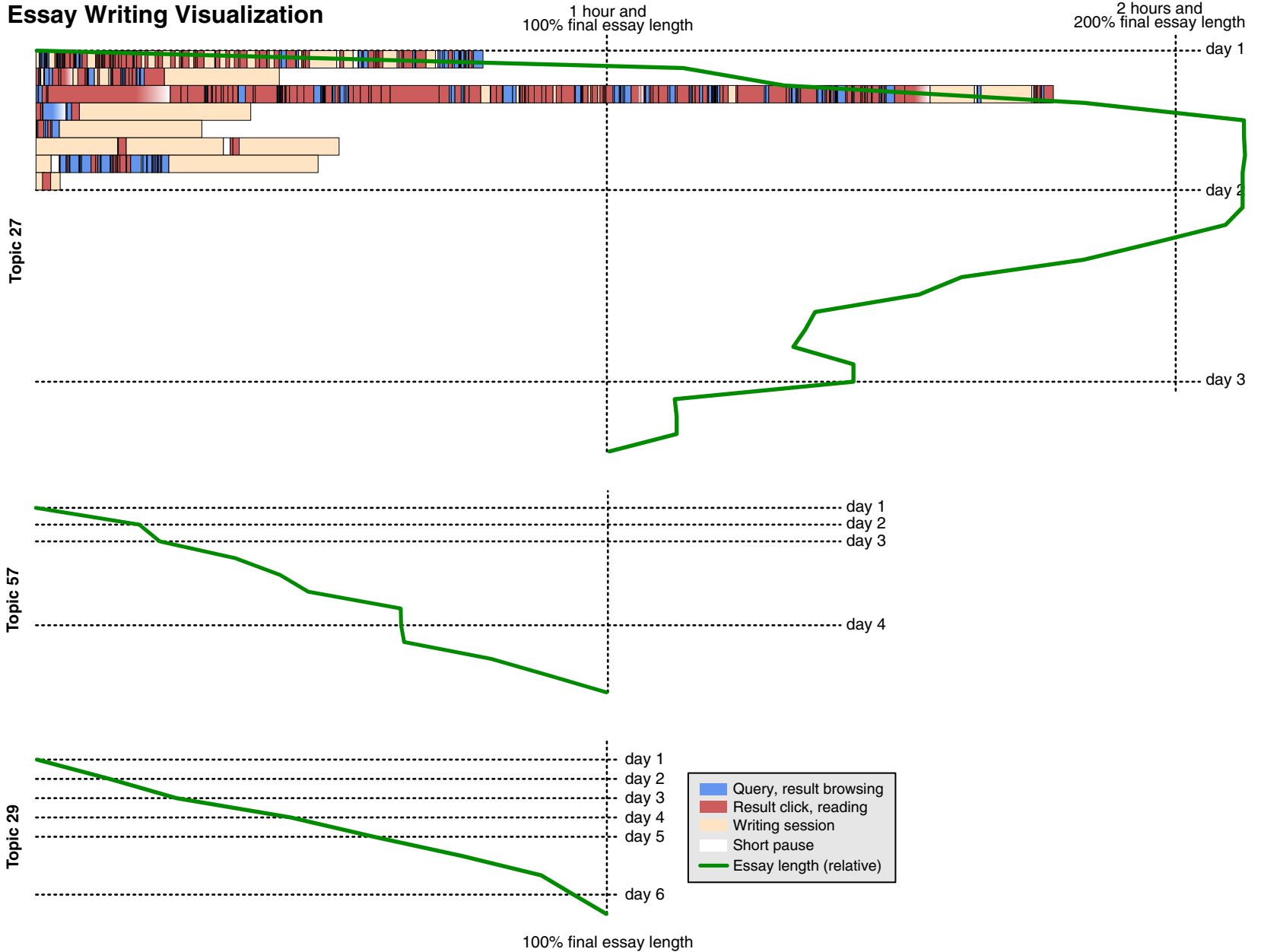
Essay Writing Visualization



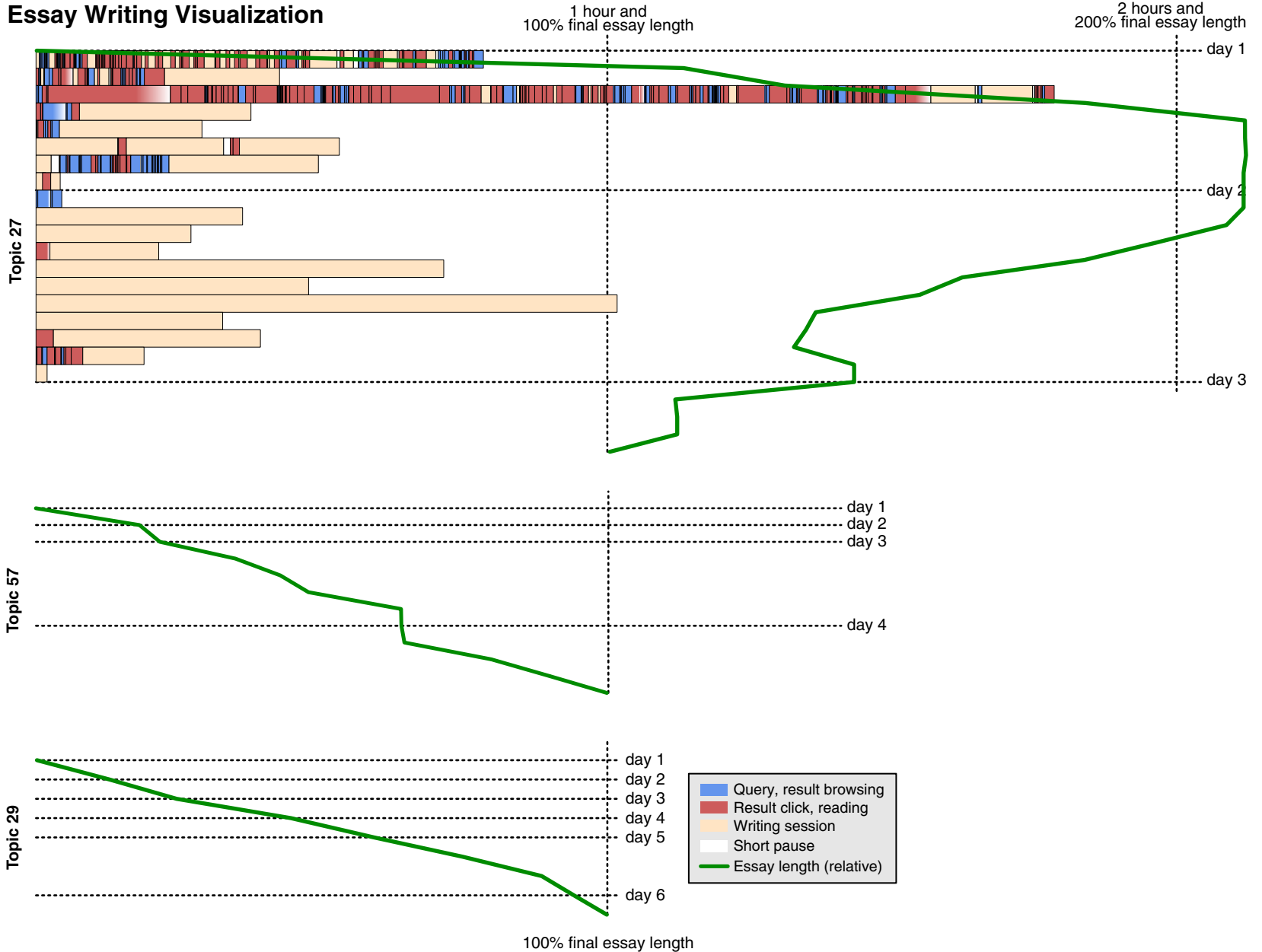
Essay Writing Visualization



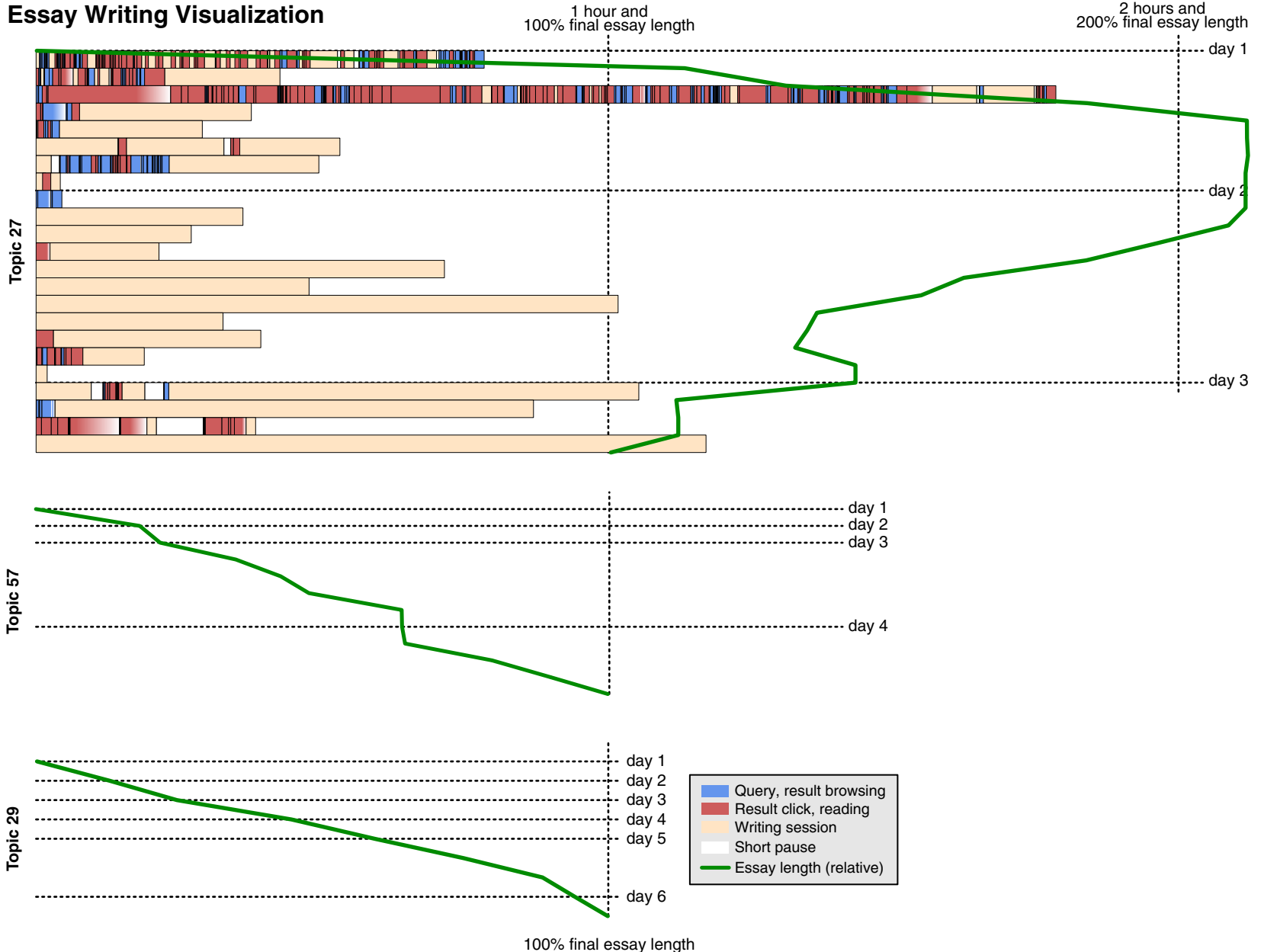
Essay Writing Visualization



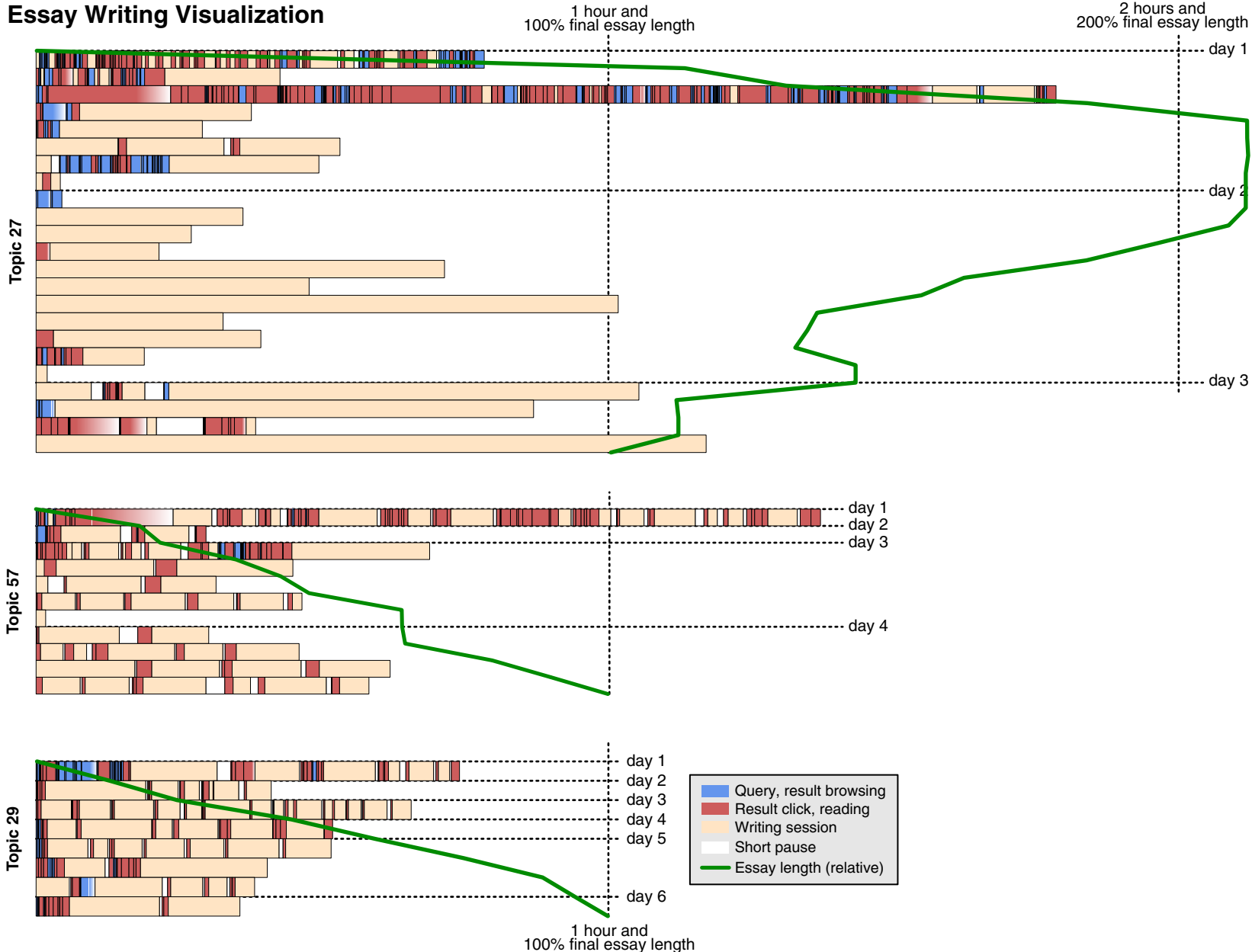
Essay Writing Visualization



Essay Writing Visualization



Essay Writing Visualization



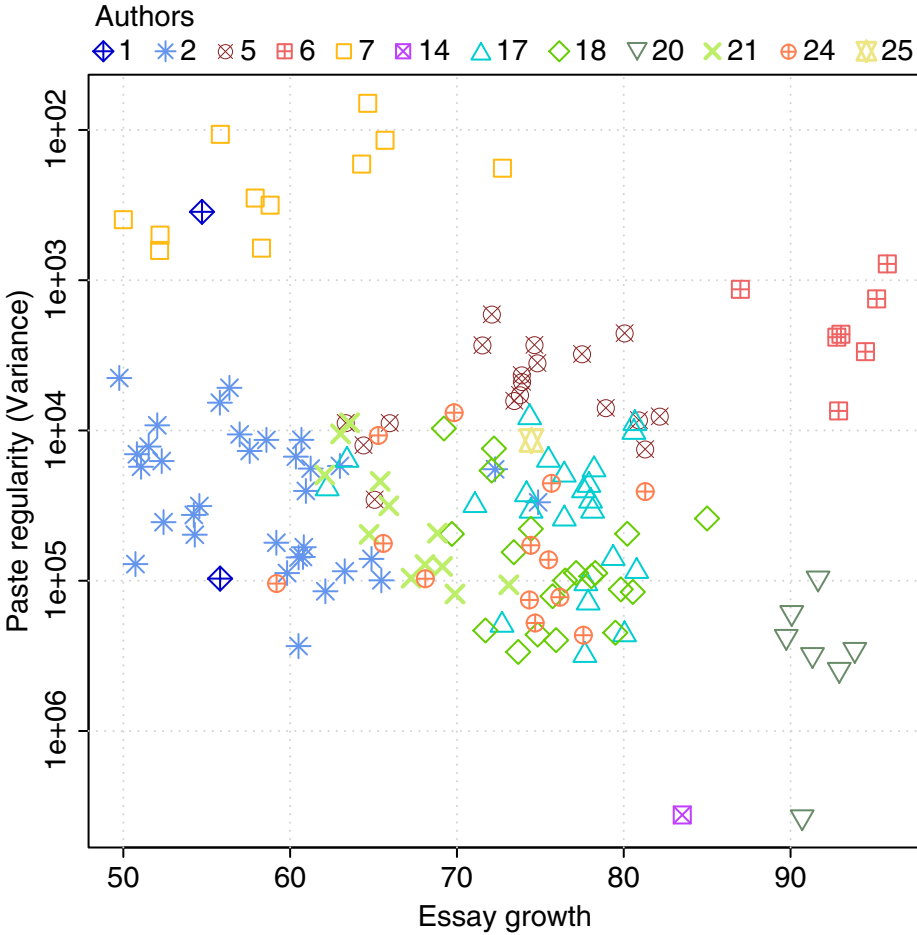
Writing and Search Behavior

Writing and Search Behavior

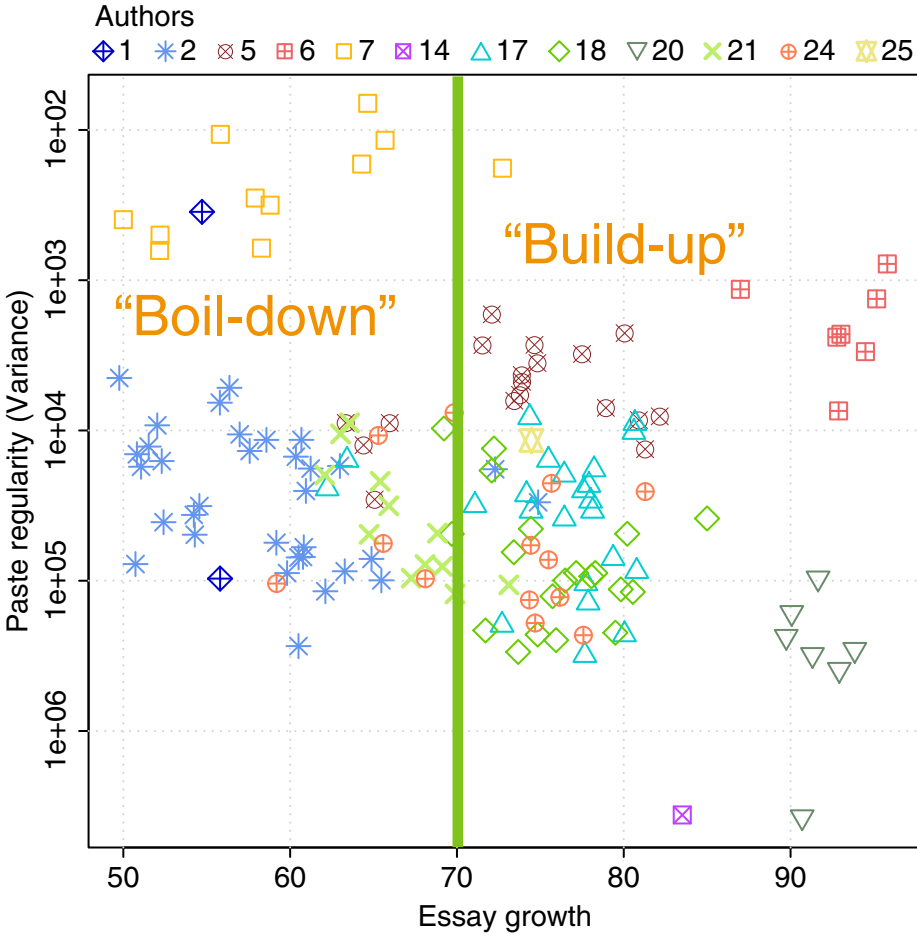
- ❑ **Essay Growth:** percentage of lengthening revisions

- ❑ **Paste Regularity:** variance of number of revisions between adding sources

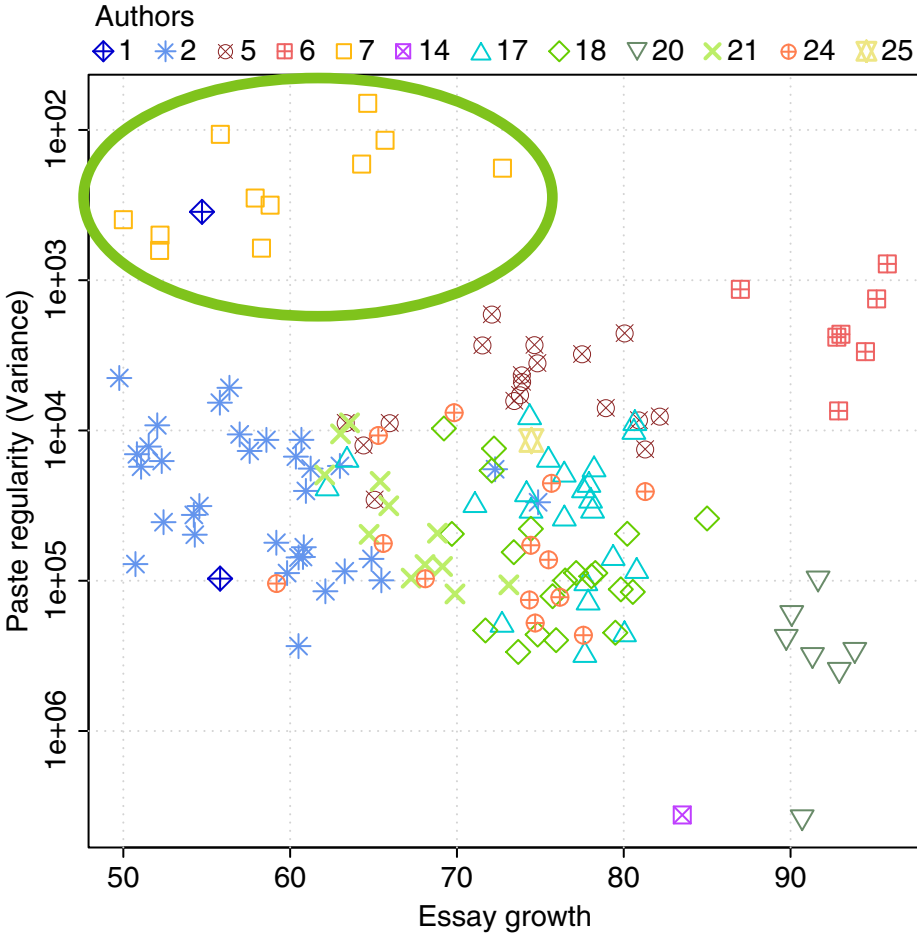
Writing and Search Behavior



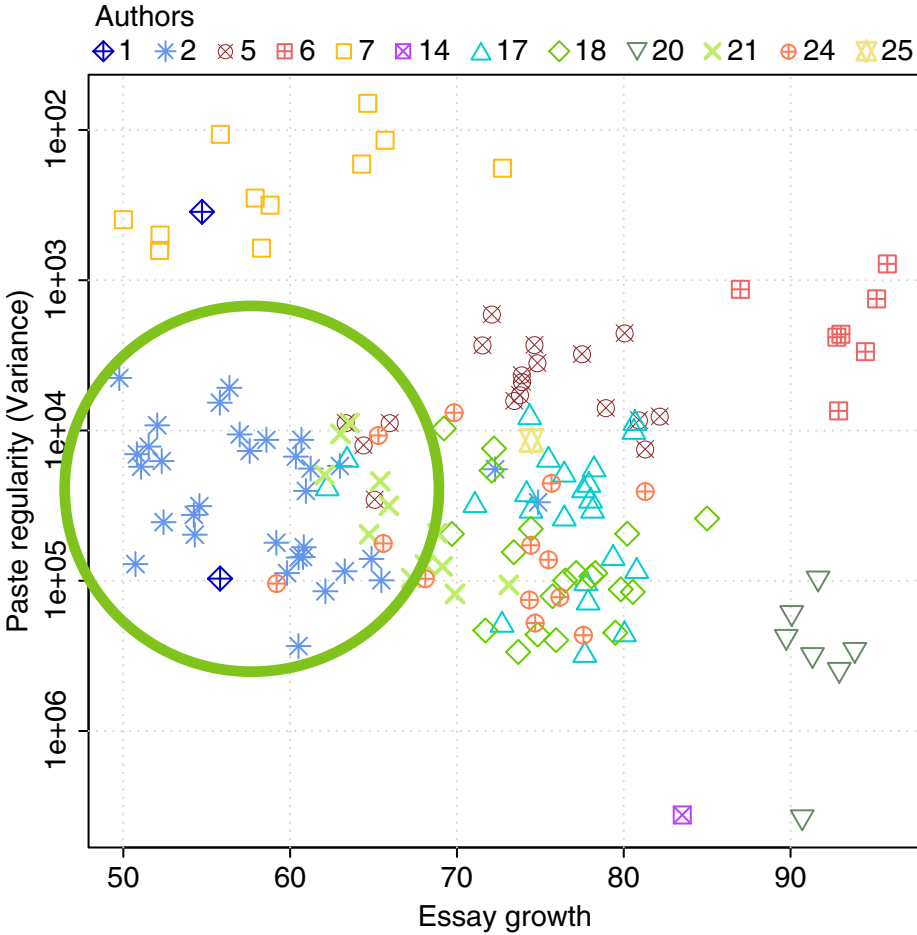
Writing and Search Behavior



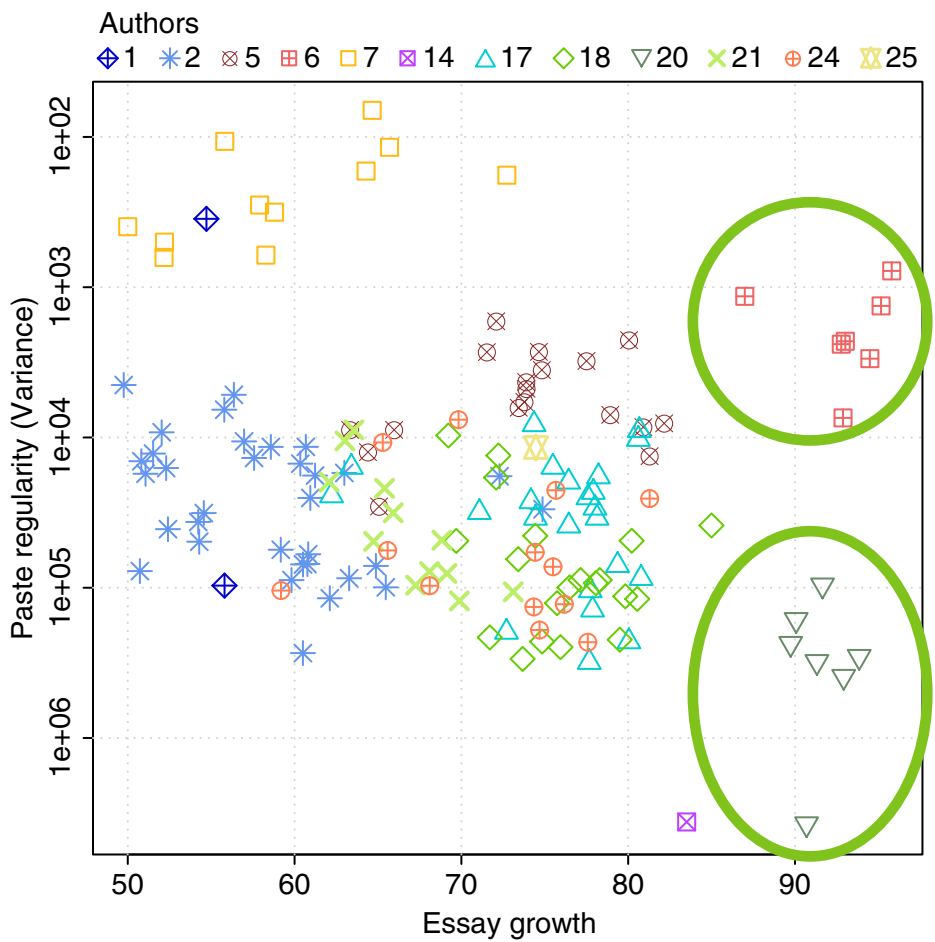
Writing and Search Behavior



Writing and Search Behavior



Writing and Search Behavior



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

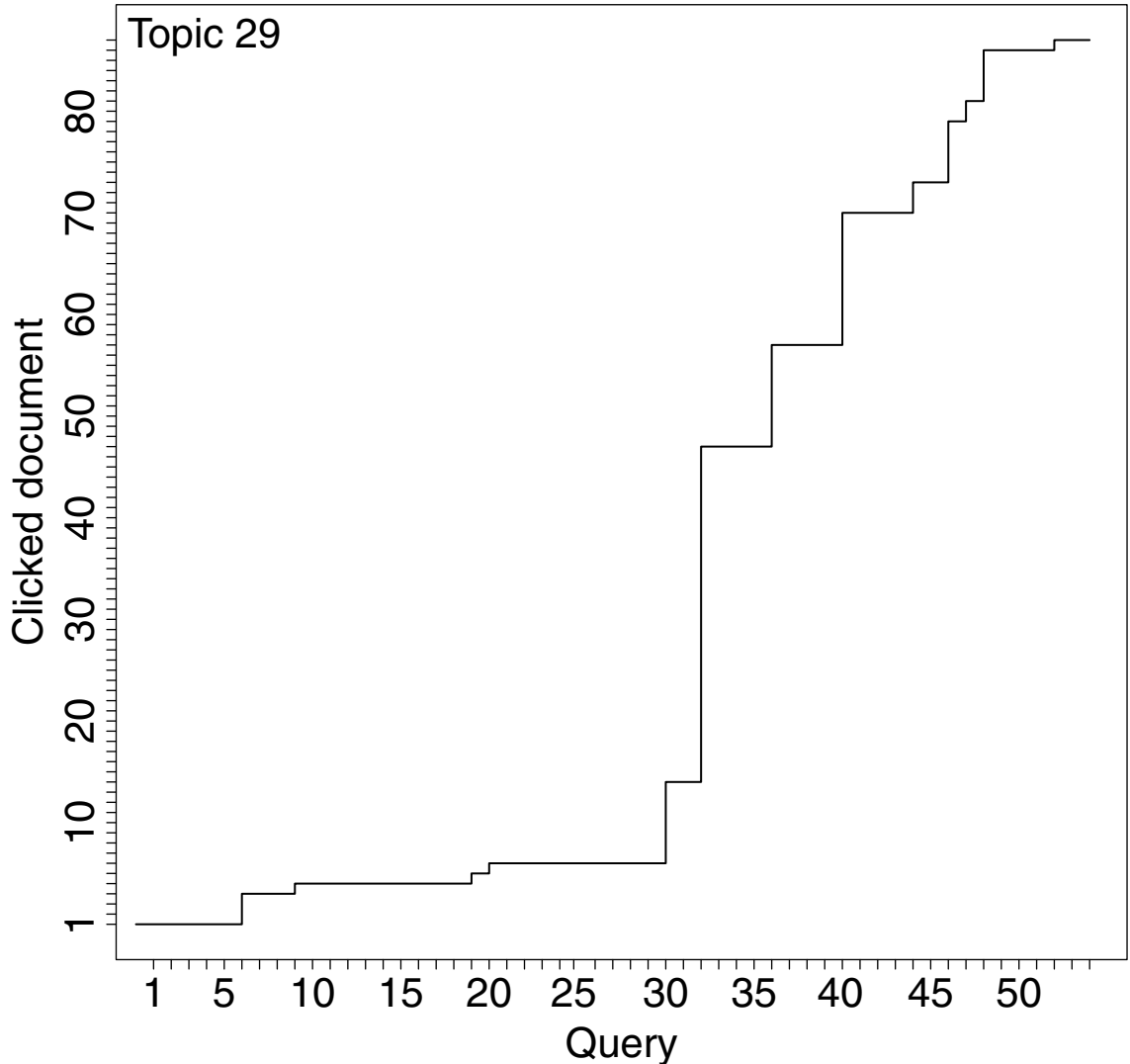
Task description 24%

Search Result Page 37%

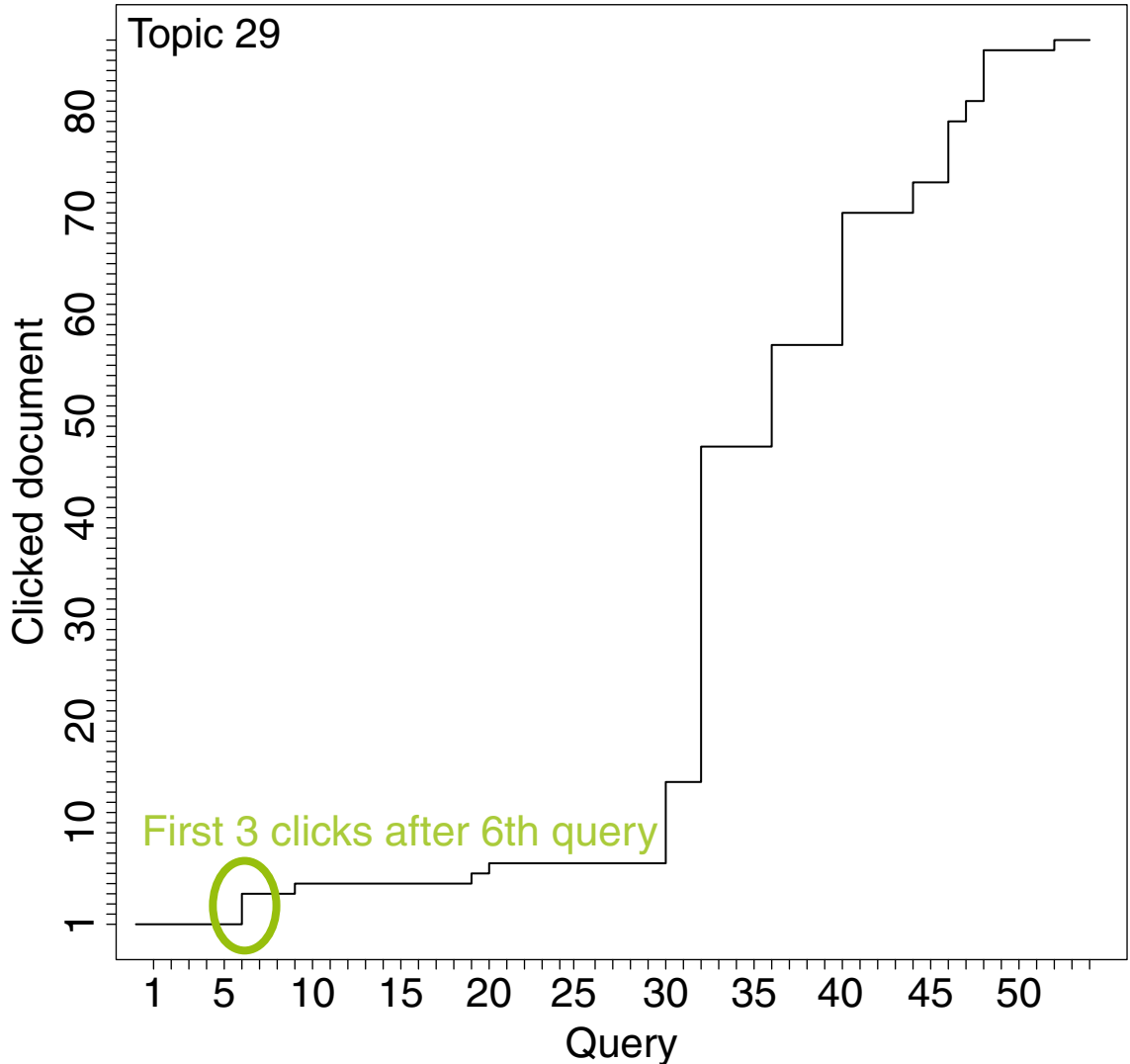
Documents 31%

Prior Knowledge 8%

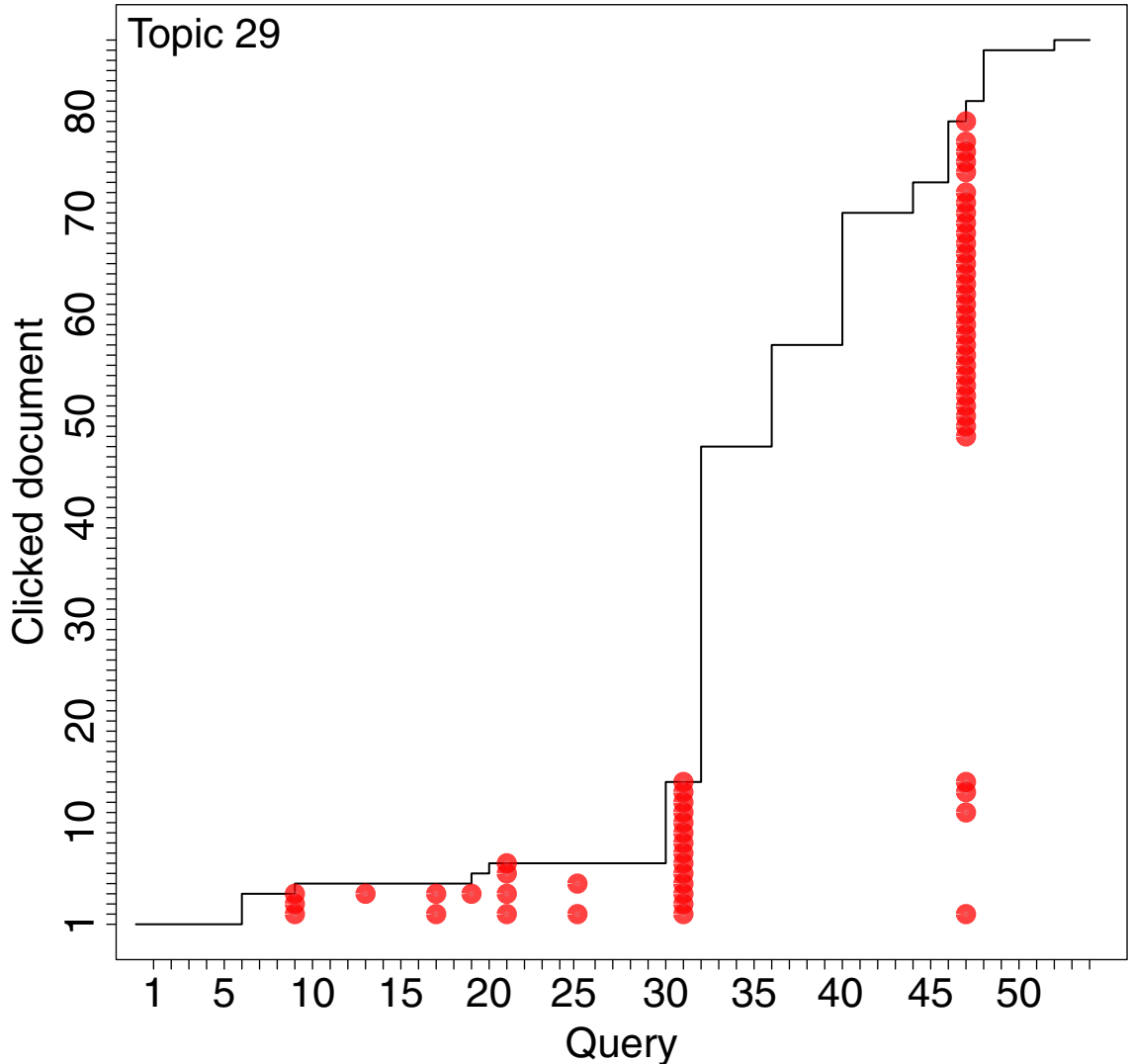
Origins of New Query Terms



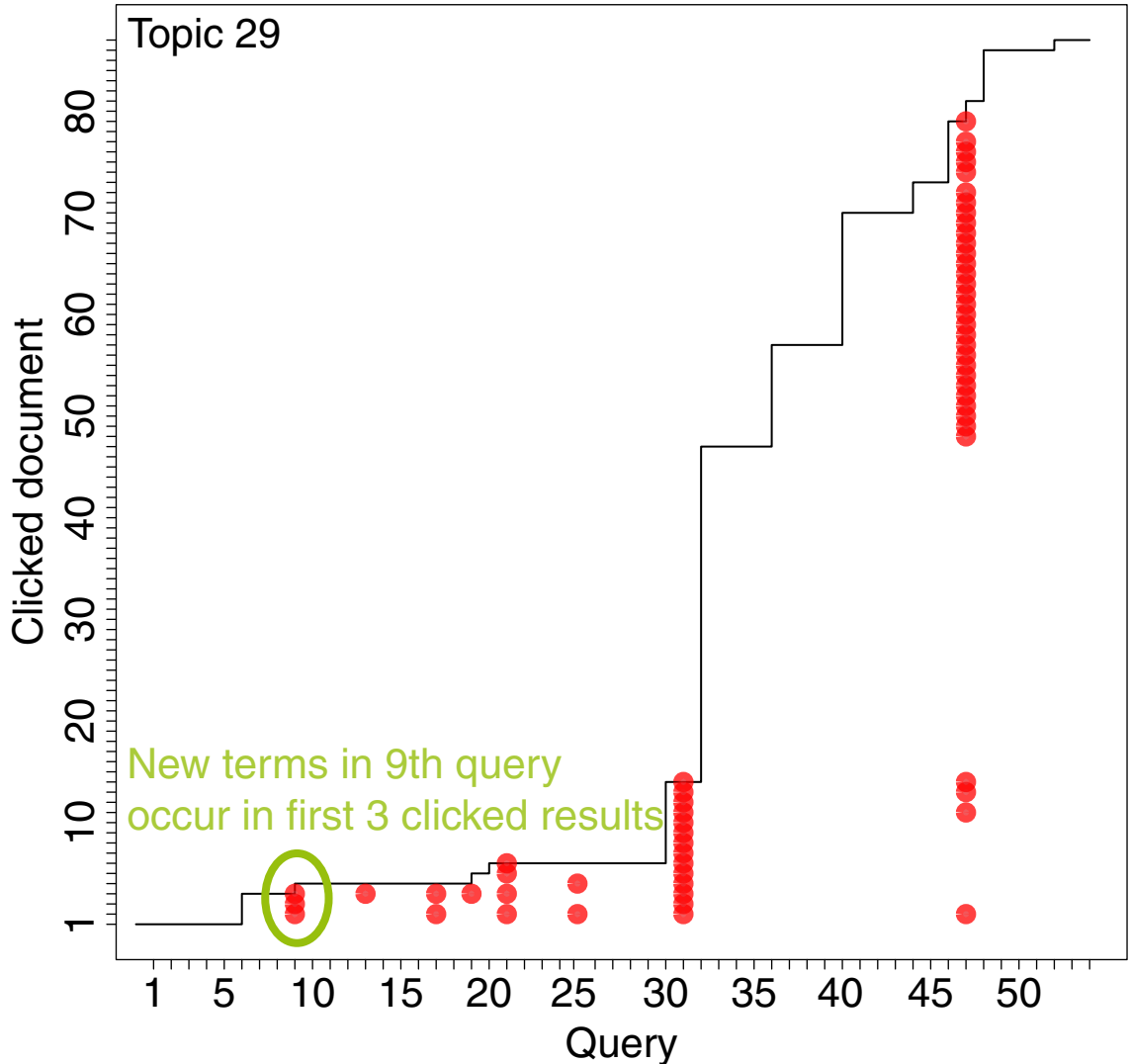
Origins of New Query Terms



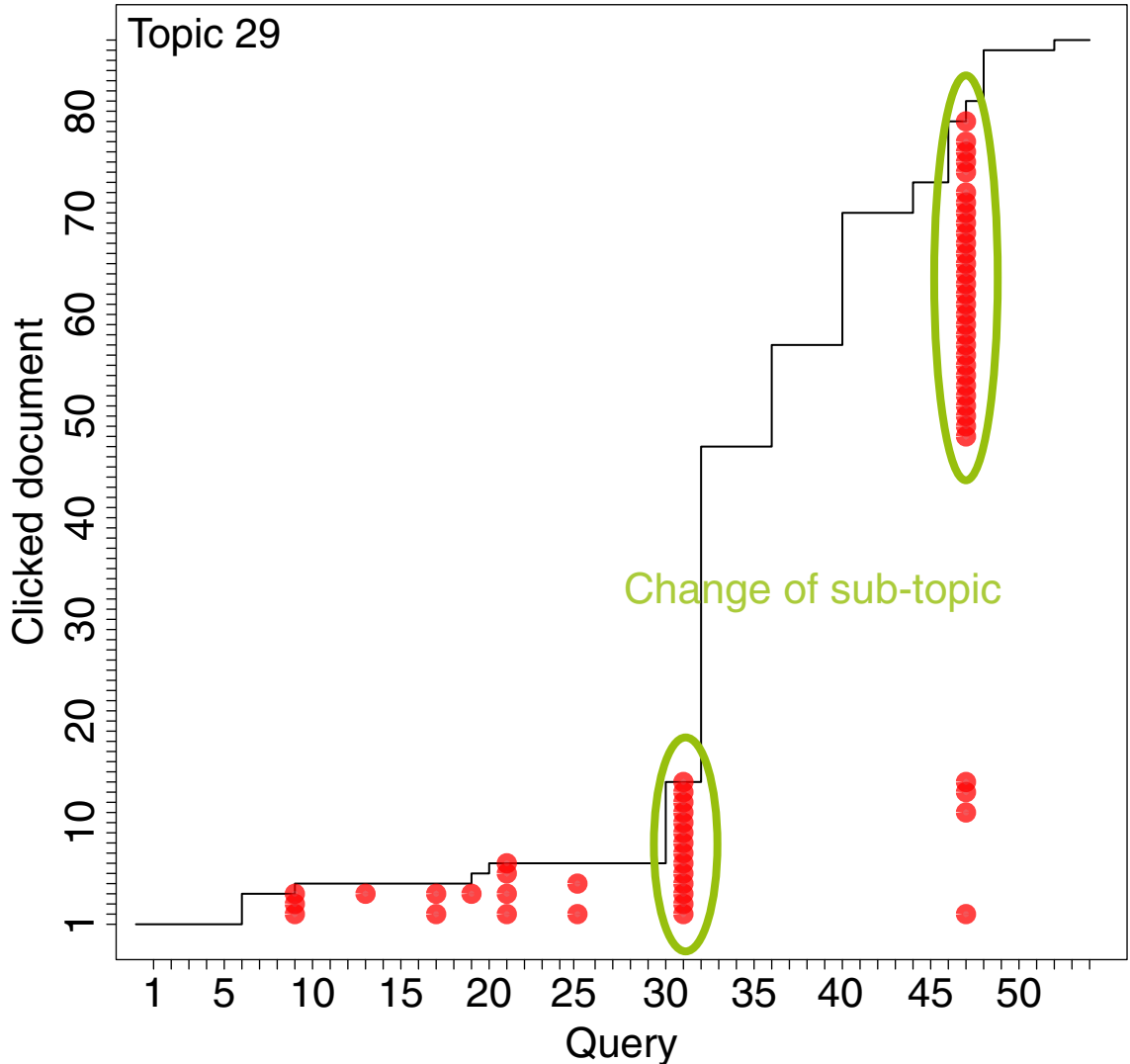
Origins of New Query Terms



Origins of New Query Terms



Origins of New Query Terms



Summary

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?