Topical Sequence Profiling

Tim Gollub* Nedim Lipka† Eunyee Koh† Erdan Genc* Benno Stein*

TIR @ DEXA 5. Sept. 2016

* Webis Group
Bauhaus-Universität Weimar
www.webis.de

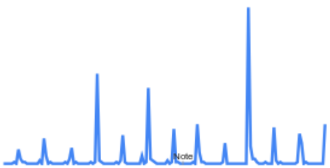
† Big Data Experience Lab
Adobe Systems
www.research.adobe.com

Google Trends

Explore topics

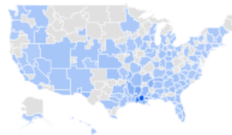
Featured insights

MTV Video Music Awards 2016



Search interest in the VMAs

Louisiana floods

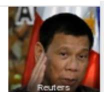


Search interest by metro area, past week

Stories trending now

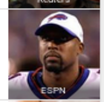
1 Davao, Philippines, Rodrigo Duterte

Past 24h



2 Buffalo Bills, Manny Lawson, Jerome Felton

Past 24h



3 Food and Drug Administration, Antibacterial soap, Antibi...

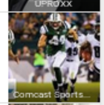
Past 24h

4 Jon Polito, Coen brothers, The Big Lebowski

Past 24h

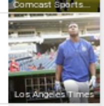
5 Tennessee Titans, Justin Hunter, Bishop Sankey, NFL

Past 24h



6 Los Angeles Dodgers, Vin Scully, KTLA

Past 24h

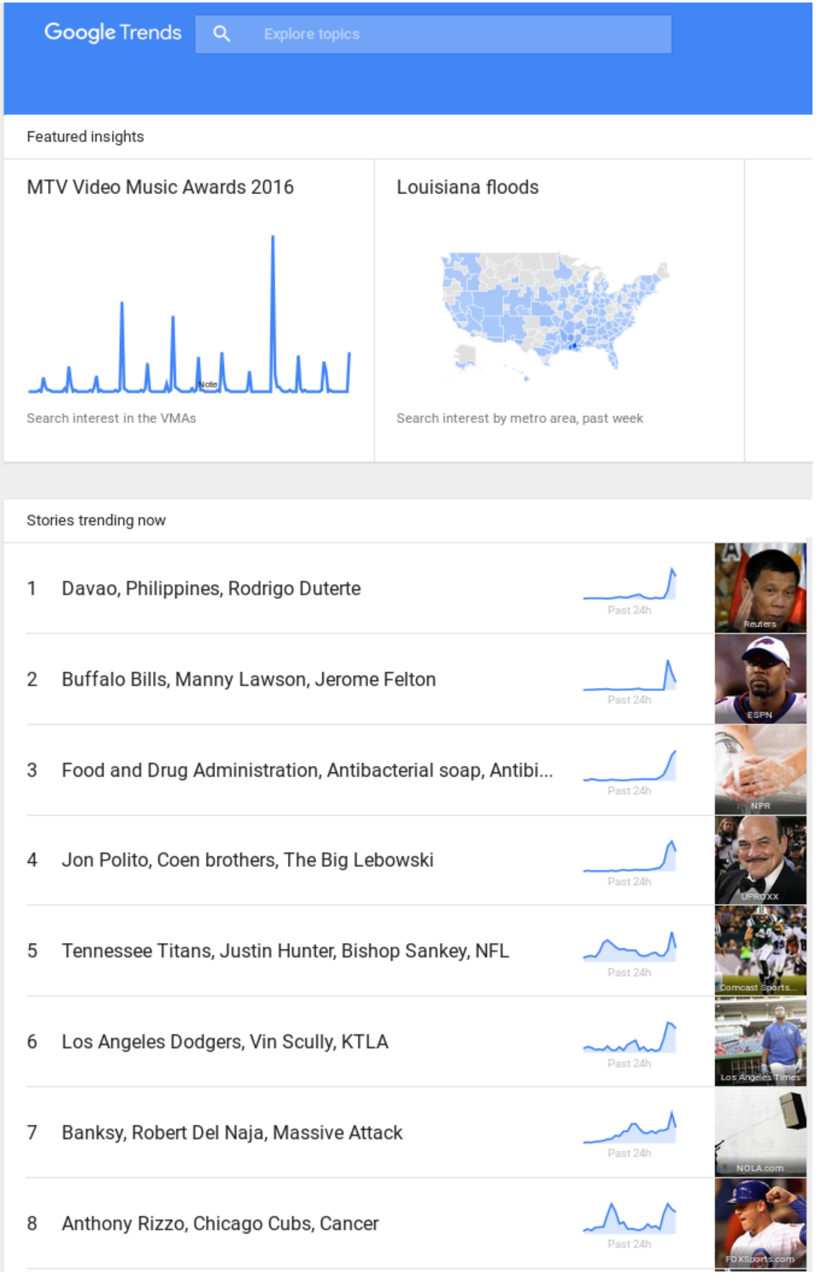


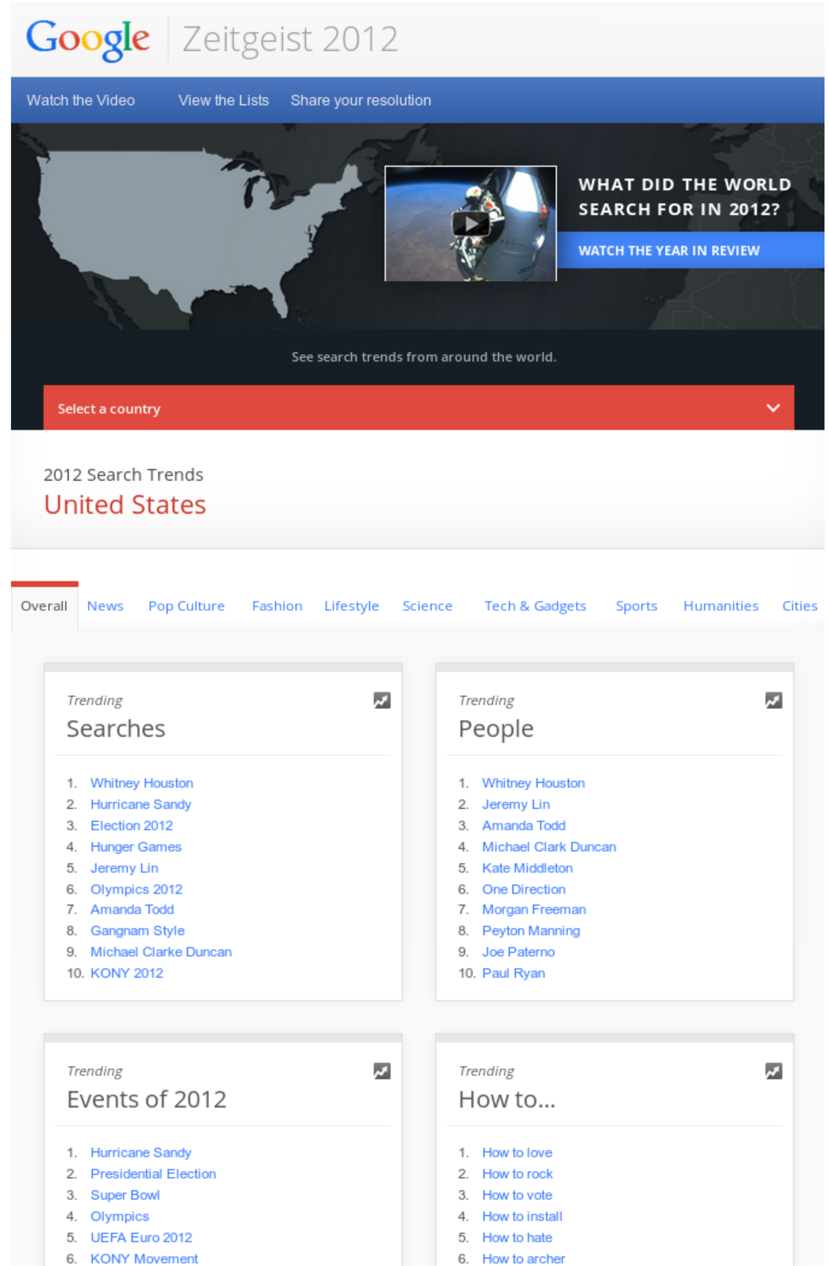
7 Banksy, Robert Del Naja, Massive Attack

Past 24h

8 Anthony Rizzo, Chicago Cubs, Cancer

Past 24h


Google Trends
Explore topics
Featured insights
MTV Video Music Awards 2016
Note
Search interest in the VMAs
Louisiana floods
Search interest by metro area, past week
Stories trending now
1 Davao, Philippines, Rodrigo Duterte
Past 24h
Reuters
2 Buffalo Bills, Manny Lawson, Jerome Felton
Past 24h
ESPN
3 Food and Drug Administration, Antibacterial soap, Antibi...
Past 24h
NPR
4 Jon Polito, Coen brothers, The Big Lebowski
Past 24h
UPROXX
5 Tennessee Titans, Justin Hunter, Bishop Sankey, NFL
Past 24h
Comcast Sports...
6 Los Angeles Dodgers, Vin Scully, KTLA
Past 24h
Los Angeles Times
7 Banksy, Robert Del Naja, Massive Attack
Past 24h
NOLA.com
8 Anthony Rizzo, Chicago Cubs, Cancer
Past 24h
FOXSports.com


Google Zeitgeist 2012
Watch the Video
View the Lists
Share your resolution
WHAT DID THE WORLD SEARCH FOR IN 2012?
WATCH THE YEAR IN REVIEW
See search trends from around the world.
Select a country
2012 Search Trends
United States
Overall
News
Pop Culture
Fashion
Lifestyle
Science
Tech & Gadgets
Sports
Humanities
Cities
Trending
Searches
1. Whitney Houston
2. Hurricane Sandy
3. Election 2012
4. Hunger Games
5. Jeremy Lin
6. Olympics 2012
7. Amanda Todd
8. Gangnam Style
9. Michael Clarke Duncan
10. KONY 2012
Trending
People
1. Whitney Houston
2. Jeremy Lin
3. Amanda Todd
4. Michael Clark Duncan
5. Kate Middleton
6. One Direction
7. Morgan Freeman
8. Peyton Manning
9. Joe Paterno
10. Paul Ryan
Trending
Events of 2012
1. Hurricane Sandy
2. Presidential Election
3. Super Bowl
4. Olympics
5. UEFA Euro 2012
6. KONY Movement
Trending
How to...
1. How to love
2. How to rock
3. How to vote
4. How to install
5. How to hate
6. How to archer

Topical Sequence Profiling

		Resource	
		Stream	Collection
Register	Query		
	Social Media		
	Articles, Papers		
	Books		

Topical Sequence Profiling

Register	Resource: Stream	Resource: Collection
Query	Google Trends	Google Zeitgeist
Social Media		
Articles, Papers		
Books		
Time range:	incomplete	complete
Analysis:	differential	covering

Topical Sequence Profiling

		Resource	
		Stream	Collection
Register	Query	Google Trends	Google Zeitgeist
	Social Media	Retweet statistics in Twitter	Wikipedia reverts analysis
	Articles, Papers	Provenance analytics	Topical Sequence Profiling
	Books	–	Discourse analysis
	Time range:	incomplete	complete
	Analysis:	differential	covering

Topical Sequence Profiling

Problem Statement

Given a sequence $\mathcal{D} = \{D_1, \ldots, D_n\}$ of text collections,

such as

- a series of the annual proceedings of a conference,
- a news feed of a certain period, or
- the social media mentions of an entity over time frames,

provide (statistical) insights about its content.

Topical Sequence Profiling

Problem Statement (continued)

A *topic embedding* $\mathbf{T}$ of $\mathcal{D} = \{D_1, \ldots, D_n\}$:

$$\mathbf{T} \;=\; \begin{array}{c|c} & \begin{array}{ccc} D_1 & \cdots & D_n \end{array} \\ \begin{array}{c} t_1 \\ \vdots \\ t_k \end{array} & \begin{bmatrix} \mathbf{T}_{11} & & \mathbf{T}_{1n} \\ & \ddots & \\ \mathbf{T}_{k1} & & \mathbf{T}_{kn} \end{bmatrix} \\ \hline \text{Coverage} & \begin{array}{ccc} c_1 & \cdots & c_n \end{array} \end{array}$$

Topical Sequence Profiling

Problem Statement (continued)

A *topic embedding* $\mathbf{T}$ of $\mathcal{D} = \{D_1, \ldots, D_n\}$:

$$
\begin{array}{rl}
 & \begin{array}{ccc} D_1 & \cdots & D_n \end{array} \\
\mathbf{T} = \begin{array}{r} \text{Cluster analysis} \\ \vdots \\ \text{Retrieval model} \end{array} & \begin{bmatrix} 4\% & & 40\% \\ & \ddots & \\ 7\% & & \mathbf{T}_{kn} \end{bmatrix} \\
\hline
\text{Coverage} & \begin{array}{ccc} 72\% & \cdots & c_n \end{array}
\end{array}
$$

Topical Sequence Profiling

Problem Statement (continued)

A *topic embedding* $\mathbf{T}$ of $\mathcal{D} = \{D_1, \ldots, D_n\}$:

$$\mathbf{T} \;=\; \begin{array}{r c} & \begin{array}{ccc} D_1 & \cdots & D_n \end{array} \\ \begin{array}{r} \text{Cluster analysis} \\ \vdots \\ \text{Retrieval model} \end{array} & \left[\begin{array}{ccc} 4\% & & 40\% \\ & \ddots & \\ 7\% & & \mathbf{T}_{kn} \end{array} \right] \\ \hline \text{Coverage} & \begin{array}{ccc} 72\% & \cdots & c_n \end{array} \end{array}$$

A topic embedding $\mathbf{T}$ characterizes $\mathcal{D} = \{D_1, \ldots, D_n\}$ if:

1. $\mathbf{T}$ is *representative* for each of the text collections $D \in \mathcal{D}$.
2. $\mathbf{T}$ highlights *informative topic developments* within $\mathcal{D}$.
3. $\mathbf{T}$ is *small* enough to be surveyed quickly.

Topical Sequence Profiling

Problem Statement (continued)

A *topic embedding* $\mathbf{T}$ of $\mathcal{D} = \{D_1, \ldots, D_n\}$:

$$\mathbf{T} = \begin{array}{r|ccc} & D_1 & \cdots & D_n \\ \text{Cluster analysis} & \multicolumn{3}{c}{\left[\begin{array}{ccc} 4\% & & 40\% \\ & \ddots & \\ 7\% & & \mathbf{T}_{kn} \end{array}\right]} \\ \vdots & & & \\ \text{Retrieval model} & & & \\ \hline \text{Coverage} & 72\% & \cdots & c_n \end{array}$$

1. *Representativeness.*
   A topic embedding $\mathbf{T} = \{t_1, \ldots, t_k\}$ is representative if for every collection $D \in \mathcal{D}$ the percentage of documents that address at least one of the topics is above a predefined threshold $c \in [0; 1]$.

Topical Sequence Profiling

Problem Statement (continued)

A *topic embedding* $\mathbf{T}$ of $\mathcal{D} = \{D_1, \ldots, D_n\}$:

$$\mathbf{T} = \begin{array}{r c} & \begin{array}{ccc} D_1 & \cdots & D_n \end{array} \\ \begin{array}{r} \text{Cluster analysis} \\ \vdots \\ \text{Retrieval model} \end{array} & \left[\begin{array}{ccc} 4\% & & 40\% \\ & \ddots & \\ 7\% & & \mathbf{T}_{kn} \end{array} \right] \\ \hline \text{Coverage} & \begin{array}{ccc} 72\% & \cdots & c_n \end{array} \end{array}$$

2. *Informativeness.*
   The informativeness of a topic's development is assessed by the variance of the topic distribution. The higher the variance, the more informative is its development. The mean topic variance is used to assess the informativeness of the whole embedding $\mathbf{T}$.

$$\frac{1}{k} \sum_{i=1}^{k} \text{Var}(\mathbf{T}_{i:})$$

Topical Sequence Profiling

Problem Statement (continued)

A *topic embedding* $\mathbf{T}$ of $\mathcal{D} = \{D_1, \ldots, D_n\}$:

$$\begin{array}{rc}
 & \begin{array}{ccc} D_1 & \cdots & D_n \end{array} \\
\mathbf{T} = \begin{array}{r} \cancel{\text{Cluster analysis}} \\ \vdots \\ \text{Retrieval model} \end{array} & \left[\begin{array}{ccc} \cancel{4\%} & & \cancel{40\%} \\ & \ddots & \\ 7\% & & \mathbf{T}_{kn} \end{array} \right] \\
\hline
\text{Coverage} & \begin{array}{ccc} 72\% & \cdots & c_n \end{array}
\end{array}$$

3. *Minimality.*
   A topic embedding $\mathbf{T}$ is minimal if no topic can be removed without losing representativeness.

Topical Sequence Profiling

Problem Statement (continued)

A *topic embedding* $\mathbf{T}$ of $\mathcal{D} = \{D_1, \ldots, D_n\}$:

$$\mathbf{T} = \begin{array}{r c} & \begin{array}{ccc} D_1 & \cdots & D_n \end{array} \\ \begin{array}{r} \cancel{\text{Cluster analysis}} \\ \vdots \\ \text{Retrieval model} \end{array} & \left[\begin{array}{ccc} \cancel{4\%} & & \cancel{40\%} \\ & \ddots & \\ 7\% & & \mathbf{T}_{kn} \end{array}\right] \\ \hline \text{Coverage} & \begin{array}{ccc} 72\% & \cdots & c_n \end{array} \end{array}$$

3. *Minimality.*
   A topic embedding $\mathbf{T}$ is minimal if no topic can be removed without losing representativeness.

Sought is the minimum representative topic embedding $\mathbf{T}^*$ with the highest mean topic variance. $\mathbf{T}^*$ is called the topical sequence profile of $\mathcal{D}$.

Topical Sequence Profiling

Approach

Our approach comprises two steps:

(1) *Topic Acquisition.*

- Acquisition and assignment of topics to documents.
- Result: a representative but not minimal topic embedding $\mathbf{T}$.
- We consider Wikipedia articles as topics with their titles as explicit labels.

Operationalization: ESA-inspired model with heuristic candidate search.

(2) *Topic Selection.*

- Find $\mathbf{T}^*$ given $\mathbf{T}$.

Operationalization: Greedy topic selection.

Topical Sequence Profiling

Approach: (1) Topic Acquisition

1. Represent each Wikipedia article (title $\sim$ topic) as BM25 vector t.

   Represent each document in $\mathcal{D}$ as BM25 vector d.

   Represent the topic domain of $\mathcal{D}$ as centroid $\bar{d}$ of all vectors, $\bar{d} = \sum_{i=1}^{|\mathcal{D}|} d_i$.

2. For each Wikipedia article:

   Compute the similarity $\rho(t, \bar{d})$ of the <Wikipedia article, topic domain> pair.

   Remember the top similar Wikipedia articles as seed documents.

3. Explore (BFS) the Wikipedia link graph starting with the seed documents.

   For each unassigned document in $\mathcal{D}$:

   Compute the similarity $\rho(t, d)$ of the <Wikipedia article, document> pair.

   If $\rho(t, d) > \rho(t, \bar{d})$ then assign topic t to document d.

Topical Sequence Profiling

Approach: (1) Topic Acquisition

1. Represent each Wikipedia article (title $\sim$ topic) as BM25 vector t.

   Represent each document in $\mathcal{D}$ as BM25 vector d.

   Represent the topic domain of $\mathcal{D}$ as centroid $\bar{d}$ of all vectors, $\bar{d} = \sum_{i=1}^{|\mathcal{D}|} d_i$.

2. For each Wikipedia article:

   Compute the similarity $\rho(t, \bar{d})$ of the <Wikipedia article, topic domain> pair.

   Remember the top similar Wikipedia articles as seed documents.

3. Explore (BFS) the Wikipedia link graph starting with the seed documents.

   For each unassigned document in $\mathcal{D}$:

   Compute the similarity $\rho(t, d)$ of the <Wikipedia article, document> pair.

   If $\rho(t, d) > \rho(t, \bar{d})$ then assign topic t to document d.

Topical Sequence Profiling

Approach: (1) Topic Acquisition

1. Represent each Wikipedia article (title $\sim$ topic) as BM25 vector t.

   Represent each document in $\mathcal{D}$ as BM25 vector d.

   Represent the topic domain of $\mathcal{D}$ as centroid $\bar{d}$ of all vectors, $\bar{d} = \sum_{i=1}^{|\mathcal{D}|} d_i$.

2. For each Wikipedia article:

   Compute the similarity $\rho(t, \bar{d})$ of the <Wikipedia article, topic domain> pair.

   Remember the top similar Wikipedia articles as seed documents.

3. Explore (BFS) the Wikipedia link graph starting with the seed documents.

   For each unassigned document in $\mathcal{D}$:

   Compute the similarity $\rho(t, d)$ of the <Wikipedia article, document> pair.

   If $\rho(t, d) > \rho(t, \bar{d})$ then assign topic t to document d.

Topical Sequence Profiling

Approach: (2) Topic Selection

Result of the topic acquisition step:

$$\begin{array}{cc} & \begin{array}{ccc} D_1 & \cdots & D_n \end{array} \\ \mathbf{T} = \begin{array}{c} t_1 \\ \vdots \\ t_n \end{array} & \begin{bmatrix} \mathbf{T}_{11} & & \mathbf{T}_{1n} \\ & \ddots & \\ \mathbf{T}_{1m} & & \mathbf{T}_{mn} \end{bmatrix} \\ \hline & \begin{array}{ccc} 100\% & \cdots & 100\% \end{array} \end{array}$$

Greedy topic selection strategy:

1. Sort the topics of the embedding by ascending diversity Var$(\mathbf{T}_{i:})$.

2. For each topic t:
   If the embedding $\mathbf{T}$ without t is representative, remove t from $\mathbf{T}$.

Notes:

- The above greedy strategy returns a minimal topic embedding $\widehat{\mathbf{T}}^*$—precisely, a topic embedding that can not be made smaller.
- However, the topic embedding $\widehat{\mathbf{T}}^*$ may not be minimum, i.e., $|\widehat{\mathbf{T}}^*| < |\mathbf{T}^*|$.
- Computing $\mathbf{T}^*$ is an NP-hard problem. For its decision variant (existence of a $\mathbf{T}$ for a k-bounded mean topic variance) the NP-completeness may be shown.
- Since the above strategy removes topics in ascending order of their diversity, the algorithm strives for maximizing the mean topic diversity.
- The heuristic is effective, if the topics are orthogonal (since the coverage percentages won't "add up" on few text collections).

Topical Sequence Profiling

Related Work

Existing approaches focus on (single) trend detection—not on coverage.*

(a) Apply LDA topic detection to $\mathcal{D}$, irrespective the topic distributions in the individual text collections. [Griffiths:2004, Hall:2008, Yeung:2011]

(b) Apply LDA topic detection to each text collection $D \in \mathcal{D}$ individually, then align the topics across the individual text collections. [Swan:2000, Wang:2005]

(c) Model time explicitly as a parameter of the LDA topic model. [Blei:2006, Wang:2006]

*) To showcase results, topics are cherry-picked or the hottest/coldest topics are presented. The extend to which the presented topics cover (= characterize) the sequence as a whole is not examined.

Topical Sequence Profiling

Case Study

$\mathcal{D}$ = SIGIR conference proceedings from 2007 to 2015.

	Year	# Papers
D_1	2007	198
D_2	2008	193
D_3	2009	193
D_4	2010	214
D_5	2011	232
D_6	2012	216
D_7	2013	205
D_8	2014	226
D_9	2015	193

Topical Sequence Profiling

Case Study

$\mathcal{D}$ = SIGIR conference proceedings from 2007 to 2015.

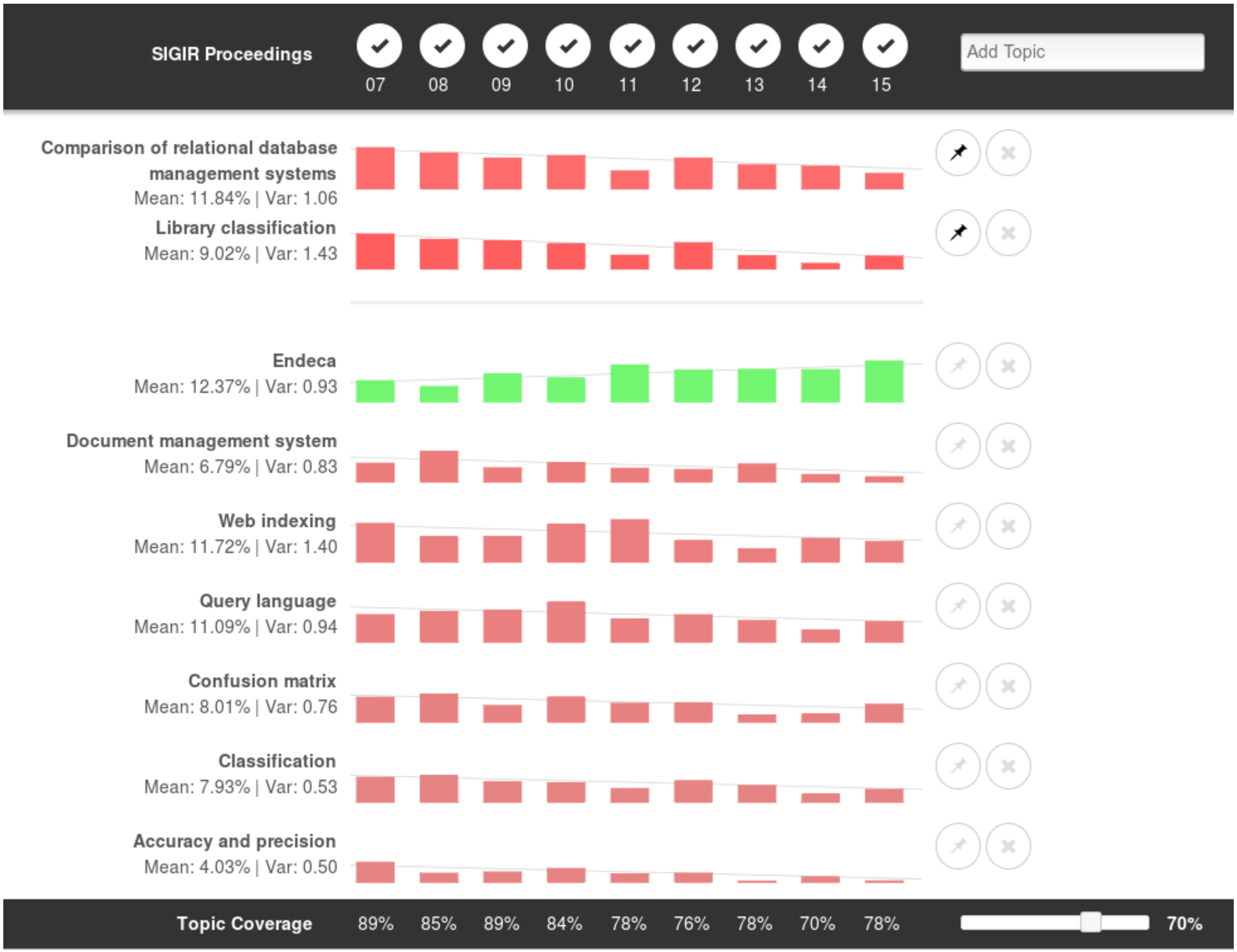
	Year	# Papers
D_1	2007	198
D_2	2008	193
D_3	2009	193
D_4	2010	214
D_5	2011	232
D_6	2012	216
D_7	2013	205
D_8	2014	226
D_9	2015	193

Wikipedia articles (seed documents) determined in the acquisition phase, along with the averaged similarity $\rho(t, \bar{d})$:

1. Concept Search (0.678)
2. Information Retrieval (0.593)
3. Human-Computer Information Retrieval (0.588)
4. Web Query Classification (0.582)
5. Enterprise Search (0.549)
6. Search engine technology (0.540)
7. Document retrieval (0.539)
8. Cognitive models of information retrieval (0.524)
9. Federated search (0.524)
10. Web search query (0.518)

Topical Sequence Profiling

Case Study [Demo: local, web]



Topical Sequence Profiling

Summary and Outlook

- Topical sequence profiling means to *cover* a set $\mathcal{D}$ of text collections.
- Such a cover should be representative, informative, and minimal; a cover is called topic embedding $\mathbf{T}$.
- We determine a minimal $\widehat{\mathbf{T}}^*$ within two steps:
  (1) topic acquisition with Wikipedia, and
  (2) heuristic topic selection by variance maximization.
- Determining the optimum cover $\mathbf{T}^*$ is NP-hard.

What is missing + further steps:

- *Thorough* evaluation of our approach:
  - effectiveness of Wikipedia-based topic acquisition
  - comparison to LDA-based approaches
  - approximation quality of the greedy selection heuristic
- Tap the potential of user interaction

Topical Sequence Profiling

Summary and Outlook

- Topical sequence profiling means to *cover* a set $\mathcal{D}$ of text collections.
- Such a cover should be representative, informative, and minimal; a cover is called topic embedding $\mathbf{T}$.
- We determine a minimal $\widehat{\mathbf{T}}^*$ within two steps:
  (1) topic acquisition with Wikipedia, and
  (2) heuristic topic selection by variance maximization.
- Determining the optimum cover $\mathbf{T}^*$ is NP-hard.

What is missing + further steps:

- *Thorough* evaluation of our approach:
  - effectiveness of Wikipedia-based topic acquisition
  - comparison to LDA-based approaches
  - approximation quality of the greedy selection heuristic
- Tap the potential of user interaction

Thank you for listening!

Topical Sequence Profiling

Related (1): Cluster Labeling

Topical sequence profiling is closely related to cluster labeling:

(a) $\mathcal{D}$ can be considered as a clustering of documents.
Assign to each $D \in \mathcal{D}$ some topic t as label.

(b) The categorization of a set $D \in \mathcal{D}$ can be considered as non-exclusive clustering task. Assign the topics $t \in \mathbf{T}$ as labels to the clusters in D.

Property	View (a)	View (b)
Unique	✖	✓
Summarizing	✓	✓
Expressive	✓	✓
Discriminating	✖	✓
Contiguous	✖	✓
Irredundant	✓	✓
Representative	✓	✓
Diverse	✓	✖
Minimal	✓	✓

In topical sequence profiling the desired properties of cluster labels [Stein:2004b] depend on whether we consider view (a) or view (b).

Topical Sequence Profiling

Related (2): Cluster Analysis

Topic acquisition and assignment in topical sequence profiling is closely related to the principle of descriptive cluster analysis. [Hoppe:2010]