What’s Hot and What’s Not: Windowed Developer Topic Analysis

Abram Hindle, Michael W. Godfrey, Richard C. Holt

Software Architecture Group
David R. Cheriton School of Computer Science
University of Waterloo
Canada
http://swag.uwaterloo.ca/
{ahindle,migod,holt}@cs.uwaterloo.ca
Development History
Our blackbox

LDA

LDA
LSI
Example

[Times]

apologies to those with prior LDA/LSI experience
What if we didn't know what section the articles were in?
Documents are represented as word distributions (word counts)
LDA finds independent word distributions that the documents are related to.

Documents can be associated with more than one topic.
Documents are represented as a linear combination of independent topics

Word Distributions

Athlete and Actor

Topics:
Independent
Word Distributions

Sports Entertainment

$C_0 \times$ $+$ $\Rightarrow$ $C_1 \times$
Here are two topics. I don't know what they are about!

These word lists look like: **Sports** and **Entertainment**!

**Topic 1**
- play
- game
- inning
- player
- quarter
- opponent
- ...

**Topic 2**
- gambling
- play
- night life
- comedy
- movie
- theatre
- ...

[Image of robot and pig discussing topics]
Previous authors
* Linstead et al. & Lukins et al.
* analyzed the entire history
* extracted N topics across the entire project's life time.

<table>
<thead>
<tr>
<th>Time</th>
<th>Time1</th>
<th>Time2</th>
<th>Time3</th>
<th>Time4</th>
<th>Time5</th>
<th>Time6</th>
<th>Time7</th>
<th>Time8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Topic analysis (LDA) can group articles around topics. Can we apply this to software?

Can we find topics, that developers were concerned with, from change-logs.
Common topics dominate!
If we did this to newspapers we'd expect to extract the sections, not necessarily local news trends!

<table>
<thead>
<tr>
<th></th>
<th>Time1</th>
<th>Time2</th>
<th>Time3</th>
<th>Time4</th>
<th>Time5</th>
<th>Time6</th>
<th>Time7</th>
<th>Time8</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Opinion</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Sports</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Life</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Auto</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>
Perhaps more local topics are more interesting?

<table>
<thead>
<tr>
<th>Time1</th>
<th>Time2</th>
<th>Time3</th>
<th>Time4</th>
<th>Time5</th>
<th>Time6</th>
<th>Time7</th>
<th>Time8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Border Tension</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Playoffs</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Celebrity Death</td>
<td>⬤</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Copyright Case</td>
<td></td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Local Crime</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
</tbody>
</table>
Local topic analysis is relevant to software because we care about what just occurred and often deal with iterations of development.

One iteration might have totally different development topics than another.
This Rational Unified Process diagram shows different disciplines are used at different times. This is an example of locality.
Local analysis
* a time unit
* an iteration
* short term topics

Global Analysis
* relate local topics globally
* Recurring or dominant topics
Development History

Apply a sliding window
We extract $N$ unique topics per window!
We noticed that some topics are quite similar, they might not appear just locally, they might reoccur globally!
We could group the similar topics together.
Our method of similarity:
* Compare top 10 most frequent words per topic
* Those that match beyond a threshold are immediately similar
* Otherwise 2 topics are similar if they are transitively related
Message

Word Distribution

Topic

Trend

Top 10 Words:
* perforce
* bug #
* POSIX
* Opteron
* ...
MySQL 3.23 Case Study

This plot was created from MySQL changelog topics that could be easily named.
MySQL 3.23 Case Study

Some named topics repeat across multiple periods
MySQL 3.23 Case Study

But many topics are local only to their particular time window!
MaxDB 7.500 Case Study

Auto-Generated Plot

Automated Analysis

Extracted Topics
Repeating trends are colored.
Local unique trends are grey
This gap indicates no development occurred.
We can name topics based on certain kinds of words.

- portability
- bugfixing

'ilities

maintenance
Observations

* some important local topics were not found with global topic analysis
Observations:

Power Law-like distribution of topic recurrence

Topic Frequency ->
Only ~10% of topics reoccur!
Observations:

Count

Topic Frequency ->

implies

with some
Future Work

- OS/2
- CMS
- 32bit
- ...

- Bug
- fix
- correct
- ...

- Clean
- refactor
- improve
- ...

- optimize
- hashes
- inline
- ...

Portability

Maintainability

Correctness

Performance

[Ernst et al.]

Ontological Naming Algorithm
Conclusions