Automated Topic Naming to Support Cross-project Analysis of Software Maintenance Activities

Abram Hindle
Dept. of Computer Science
University of California, Davis
Davis, CA, USA
abram@softwareprocess.es

Neil A. Ernst
Dept. of Computer Science
University of Toronto
Toronto, Ontario, CANADA
nernst@cs.toronto.edu

Michael W. Godfrey
David Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, CANADA
migod@uwaterloo.ca

John Mylopoulos
Dept. Information Eng. and Computer Science
University of Trento
Trento, ITALY
jm@disi.unitn.it

US NSF SHF Medium 0964703
Who Cares About Quality?

Managers

New Developers

Developers

Investors

Customers
What is this commit about?

Added a test for bug #1326 on OSX
What is this commit about?

Added a test for bug #1326 on OSX
What is this commit about?

Added a test for bug #1326 on OSX

Reliability
Maintainability
Portability
But we have many commits..
Developer Topics

Commit

Commit

Developer Topic

Maintainability

purpose?

Reliability

LDA
LSI
Cross Project Relevance

Shared Concepts

- Version Control
- Efficiency
- Usability
- Reliability and functionality (includes correctness)
- Portability
- Maintainability
Quality-related Non Functional Requirements (NFRs)

- Portability
- Usability
- Maintainability
- Reliability and functionality (includes correctness)
- Efficiency

[iso9126]  
[cleland-huang03]  
[ernst10]
Can't we just summarize quality related efforts within this project?

Maintainability
Functionality
Portability
Efficiency
Usability
Reliability

Non-Functional Requirements

time ->
<table>
<thead>
<tr>
<th>Time (months)</th>
<th>Unique Topics</th>
<th>Labelled Developer Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004 Jun</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004 Jul</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004 Aug</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004 Sep</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004 Oct</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004 Nov</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004 Dec</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005 Jan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005 Feb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005 Mar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005 Apr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005 May</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005 Jun</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005 Jul</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005 Aug</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005 Sep</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005 Oct</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005 Nov</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005 Dec</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006 Jan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006 Feb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006 Mar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006 Apr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006 May</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time (months)</td>
<td>Unique Topics</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>---------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Labelled Developer Topics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>efficiency, portability, functionality</td>
<td></td>
</tr>
<tr>
<td></td>
<td>maintainability, portability, functionality</td>
<td></td>
</tr>
<tr>
<td></td>
<td>efficiency, reliability, functionality</td>
<td></td>
</tr>
<tr>
<td></td>
<td>maintainability, portability, functionality</td>
<td></td>
</tr>
<tr>
<td></td>
<td>efficiency</td>
<td></td>
</tr>
<tr>
<td></td>
<td>reliability</td>
<td></td>
</tr>
<tr>
<td></td>
<td>functionality</td>
<td></td>
</tr>
</tbody>
</table>

The table shows a matrix with Time (months) on the left and Unique Topics on the top, with various labels such as efficiency, portability, functionality, maintainability, and reliability.
Example

apologies to those with prior LDA/LSI experience

[Blei]
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

Topics:
Independent Word Distributions

Athlete and Actor

Sports Entertainment

\[ C_0 \times + C_1 \times = \]
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
* ...

These word lists look like: **Sports** and **Entertainment**!
Word bag analysis

Usability

Maintainability

Portability

Reliability

Efficiency
Word Bag Examples

Portability
- portability
- transferability
- interoperability
- documentation
- internationalization
  i18n
  ...

Reliability
- reliability
- failure
- error
- redundancy
- fails
- bug
  ...

Labelled Topics of MaxDB 7.500

Time (months)
Topics of MySQL 3.23

Unique Topics

Tags

Time (months)

functionality

usability

functionality

portability

efficiency

reliability

maintainability

portability

functionality/portability

reliability/usability

functionality/portability

maintainability/reliability/portability

usability

functionality

maintainability

functionality

reliability

usability

maintainability

reliability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability

maintainability

usability
MySQL 3.23 Timeline

*Key Events*

- 3.23.19 GPL’ed
- 3.23.31 prod.
- 3.23.38
- 4.0 alpha
- 3.23.52
- 4.0 production
- 3.23.58
- 5.0 alpha

**Maintainability**

- Maintainability
- Functionality
- Reliability
- Portability

**Usability**

**Functionality**

**Reliability**

**Efficiency**

**Portability**

**None**


Functionality

Portability
ROC Values of Semi-Supervised Word Bags

For NFR:
- portability
- efficiency
- reliability
- functionality
- maintainability
- usability
- total

MaxDB exp2
MySQL exp2
MaxDB exp3
MySQL exp3

ROC Values:
- Portability: 0.45
- Efficiency: 0.5
- Reliability: 0.55
- Functionality: 0.6
- Maintainability: 0.65
- Usability: 0.7
- Total: 0.75

Comparison with other databases:
- MaxDB exp2
- MySQL exp2
- MaxDB exp3
- MySQL exp3
Supervised Multitag Classifiers: MySQL and MaxDB

MySql Classifiers

MaxDB Classifiers
Conclusions

LDA
LSI

[http://softwareprocess.es/name/]
F-1 Measure of Semi-Supervised Word Bags

The graph shows the F-1 Measure of semi-supervised word bags for different NFRs (portability, efficiency, reliability, functionality, maintainability, usability, and total) for MaxDB and MySQL versions MaxDBexp2, MaxDBexp3, MySQLexp2, and MySQLexp3.
Annotation: Stop Words
MaxDB 7.500 Case Study

<table>
<thead>
<tr>
<th>STOP</th>
<th>STOP</th>
<th>topics joined due to similarity</th>
</tr>
</thead>
</table>

2 long trends instead of one
Annotation: Training Sets

Maintainability+

Maintainability-
Annotation: Stop Words

Used in topic analysis or to reduce # of features for learners.
Message

Word Distribution

Top 10 Words:
* perforce
* bug #
* POSIX
* Opteron
* ...

Topic

Trend