Using Data Fusion and Web Mining to Support Feature Location in Software

Meghan Revelle, Bogdan Dit, Denys Poshyvanyk

18th IEEE International Conference on Program Comprehension (ICPC'10)

Bug 66914 - [typing] Error Message after undo copy/paste

Status: VERIFIED FIXED

Product: JDT
Component: Text
Version: 3.0
Platform: PC All

Importance: P2 major (vote)
Target Milestone: 3.0 RC3
Assigned To: Tom Hofmann
QA Contact:

Reported: 2004-06-14 08:16 EDT by Ralf Schmauder
Modified: 2004-06-21 04:55 EDT (History)
CC List: 2 users

birsan
dirk_baeumer

See Also:

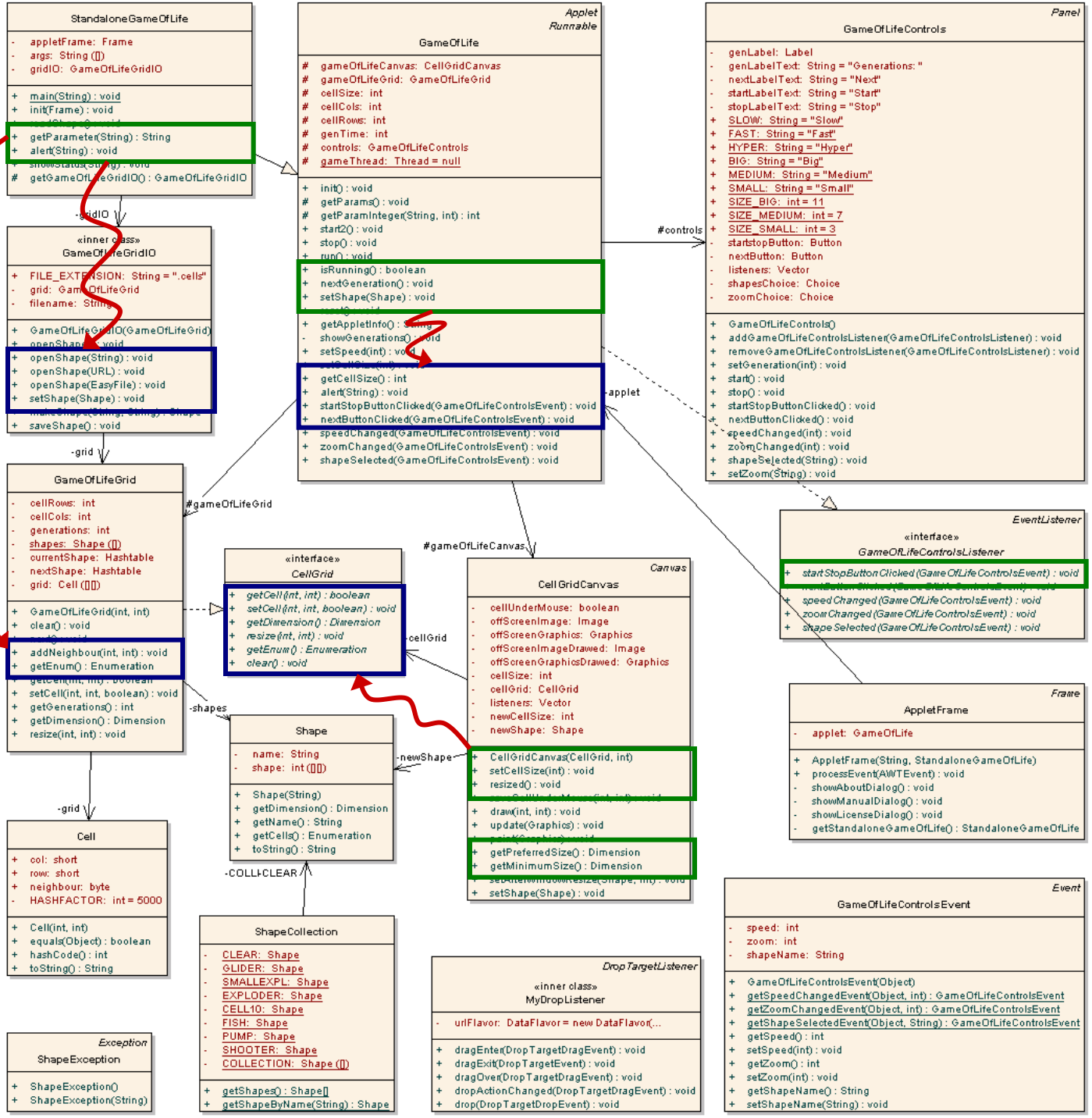
Attachments		
error log (5.43 KB, text/plain) 2004-06-14 08:17 EDT, Ralf Schmauder	*no flags*	Details
LinkedModeUI.diff (4.59 KB, patch) 2004-06-18 09:59 EDT, Tom Hofmann	*no flags*	Details | Diff
Add an attachment (proposed patch, testcase, etc.)		View All

Ralf Schmauder 2004-06-14 08:16:24 EDT Description

```
-create a new Class and generate the main method
-type "sysout" and use the code completion
-type the double quote
-paste Hello World into the double quotes
-try to undo without saving using Ctrl+z

using undo in the menubar does work
```

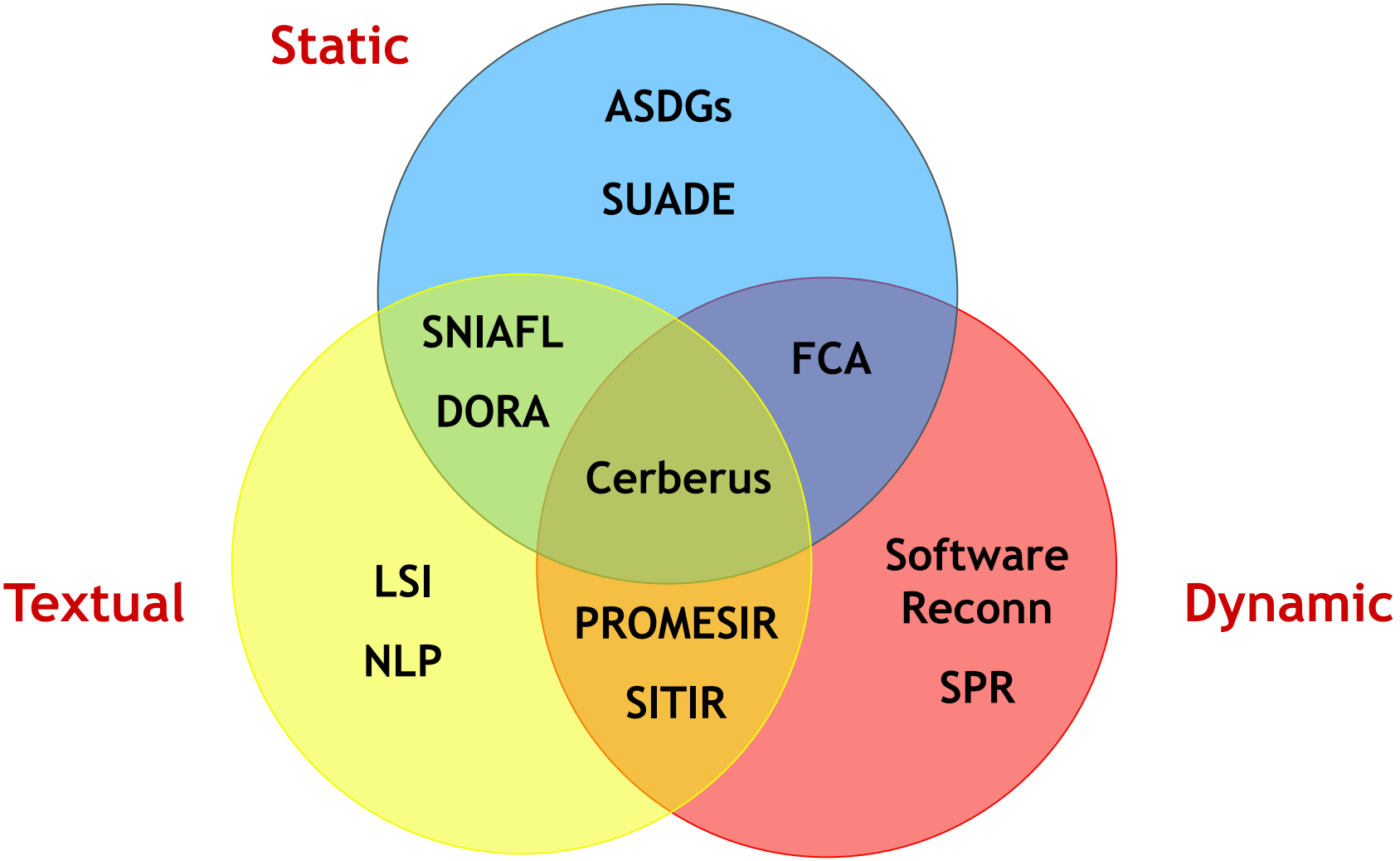
Feature: a requirement that user can invoke and that has an observable behavior.



Feature Location

Impact Analysis

Existing Feature Location Work



Meghan Revelle and Denys Poshyvanyk. "Feature Location in Source Code: A Taxonomy and Survey." Submission to *Journal of Software Maintenance and Evolution: Research and Practice*.

Textual Feature Location

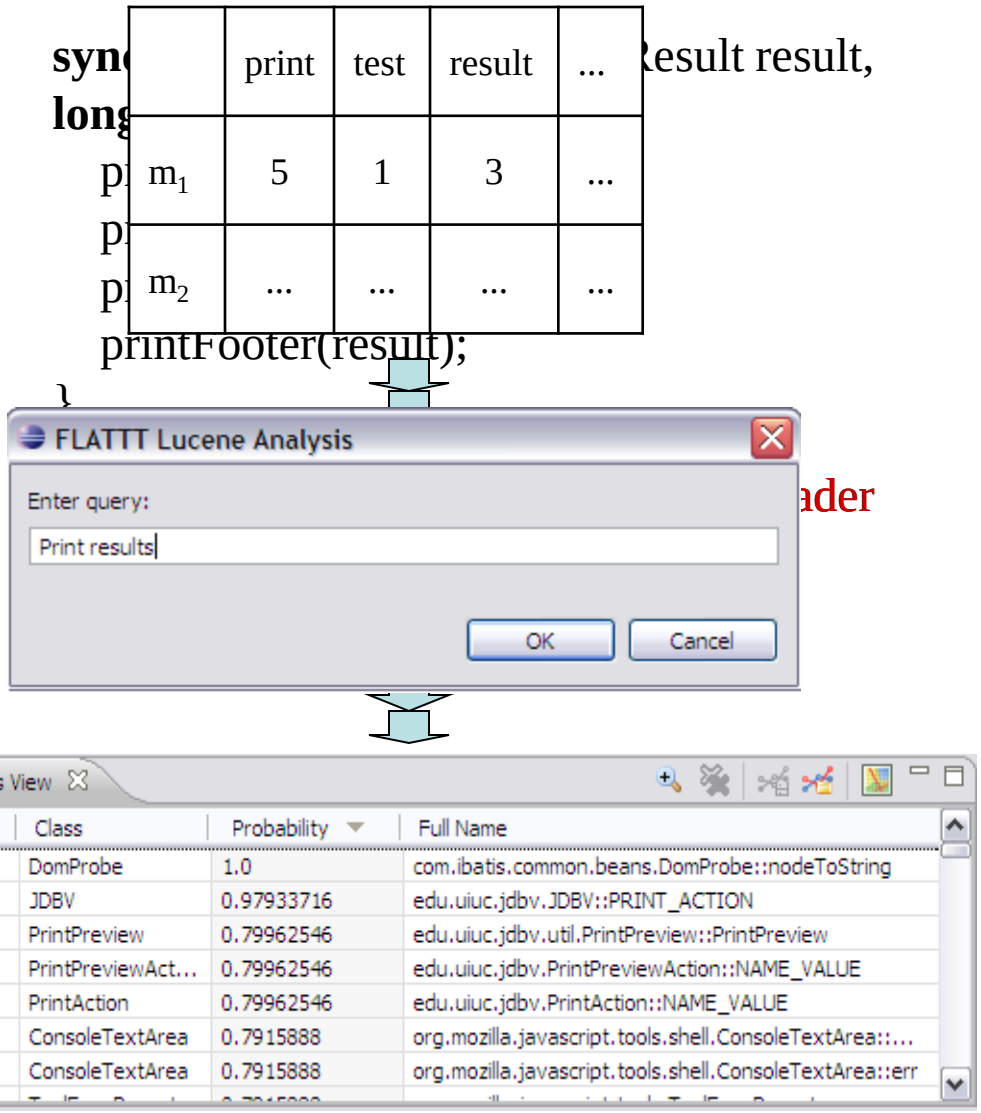
- Information Retrieval (IR)
  - Searching for documents or within docs for relevant information
- First used for feature location by Marcus et al. in 2004*.
  - Latent Semantic Indexing** (LSI)
- Utilized by many existing approaches: PROMESIR, SITIR, HIPIKAT etc.

* Marcus, A., Sergeyev, A., Rajlich, V., and Maletic, J., "An Information Retrieval Approach to Concept Location in Source Code", in Proc. of Working Conference on Reverse Engineering, 2004, pp. 214-223.

** Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R., "Indexing by Latent Semantic Analysis", *Journal of the American Society for Information Science*, vol. 41, no. 6, Jan. 1990, pp. 391-407.

Applying LSI to Source Code

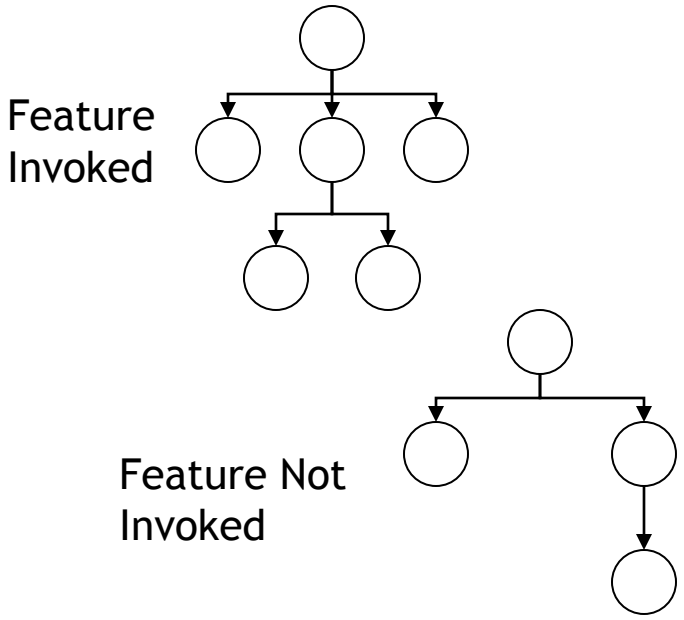
- Corpus creation
  - Choose granularity
- Preprocessing
  - Stop word removal, splitting, stemming
- Indexing
  - Term-by-document matrix
  - Singular Value Decomposition
- Querying
  - User-formulated
- Generate results
  - Ranked list



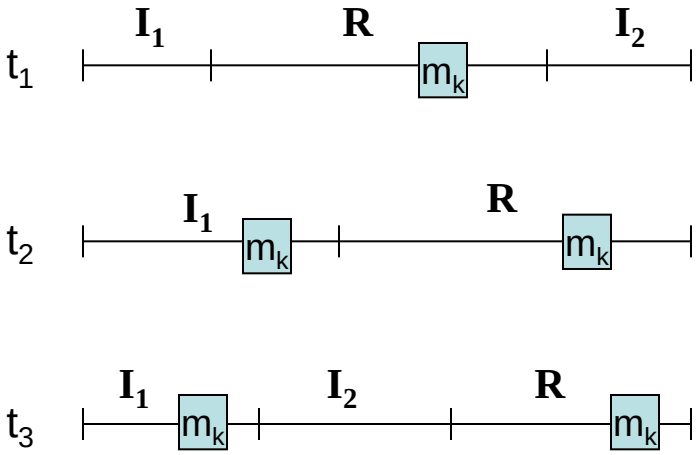
	print	test	result	...
m_1	5	1	3	...
m_2	...	...	...	...

Dynamic Feature Location

Software Reconnaissance*



Scenario-based Probabilistic Ranking (SPR)**



* Wilde, N. and Scully, M., "Software Reconnaissance: Mapping Program Features to Code", *Software Maintenance: Research and Practice*, vol. 7, no. 1, Jan.-Feb. 1995, pp. 49-62.

** Antoniol, G. and Guéhéneuc, Y. G., "Feature Identification: An Epidemiological Metaphor", *IEEE Trans. on Software Engineering*, vol. 32, no. 9, Sept. 2006, pp. 627-641.

Hybrid Feature Location

PROMESIR*

LSI score		SPR score		PROMESIR Score	
m_{15}	0.91	m_{52}	0.80	**m_6**	**0.715**
m_{16}	0.88	m_{47}	0.66	m_{47}	0.70
m_2	0.85	**m_6**	**0.64**	m_{52}	0.70
m_6	**0.79**	m_2	0.53	m_2	0.69
m_{47}	0.74	m_{15}	0.37	m_{15}	0.64
m_{52}	0.60	m_{16}	0.34	m_{16}	0.61
...	...	...	...	...	...

*Probabilistic Ranking of Methods Based on Execution Scenarios and Information Retrieval

Poshyvanyk, D., Guéhéneuc, Y. G., Marcus, A., Antoniol, G., and Rajlich, V., "Feature Location using Probabilistic Ranking of Methods based on Execution Scenarios and Information Retrieval", *IEEE Trans. on Software Engineering*, vol. 33, no. 6, June 2007, pp. 420-432.

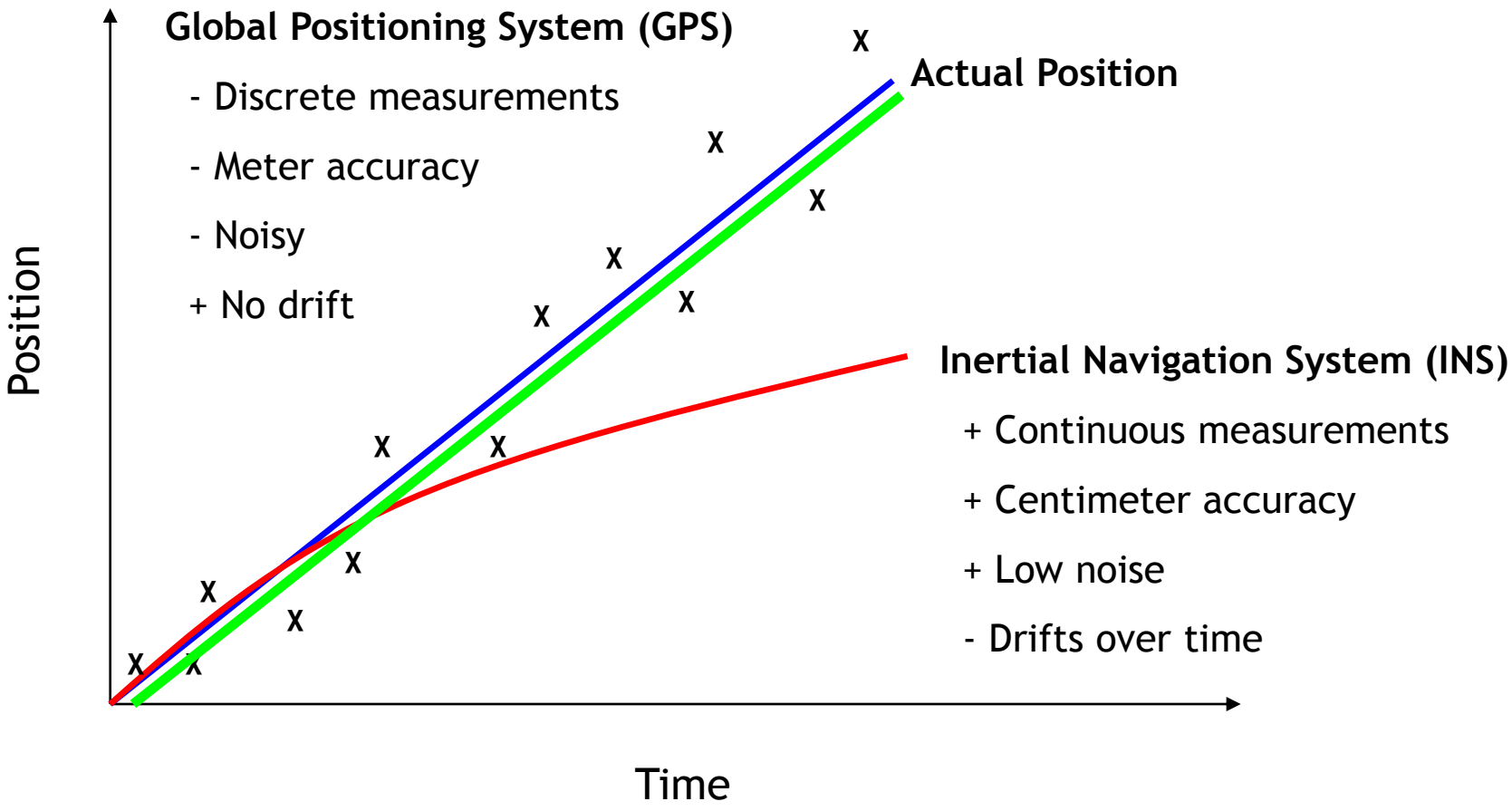
SITIR**

	LSI score	Execution Trace
m_{15}	0.91	main
~~m_{16}~~	~~0.88~~	| m_1
m_2	0.85	| m_2
m_6	**0.79**	| | m_6
m_{47}	0.74	| | m_{15}
~~m_{52}~~	~~0.60~~	| m_3
...	...	| m_{47}
		...

**SIngle Trace and Information Retrieval

Liu, D., Marcus, A., Poshyvanyk, D., and Rajlich, V., "Feature Location via Information Retrieval based Filtering of a Single Scenario Execution Trace", in Proc. of International Conference on Automated Software Engineering, 2007, pp. 234-243.

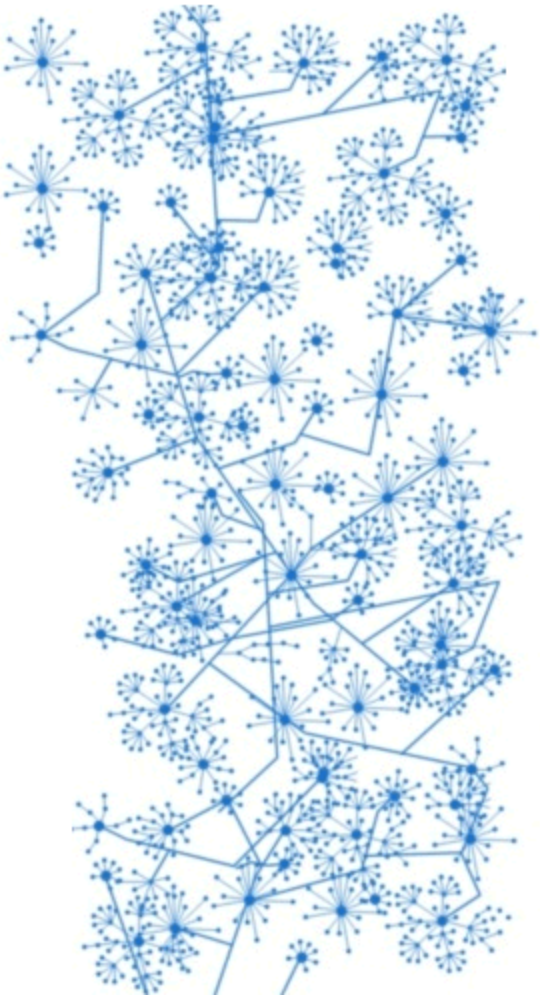
Data Fusion Example

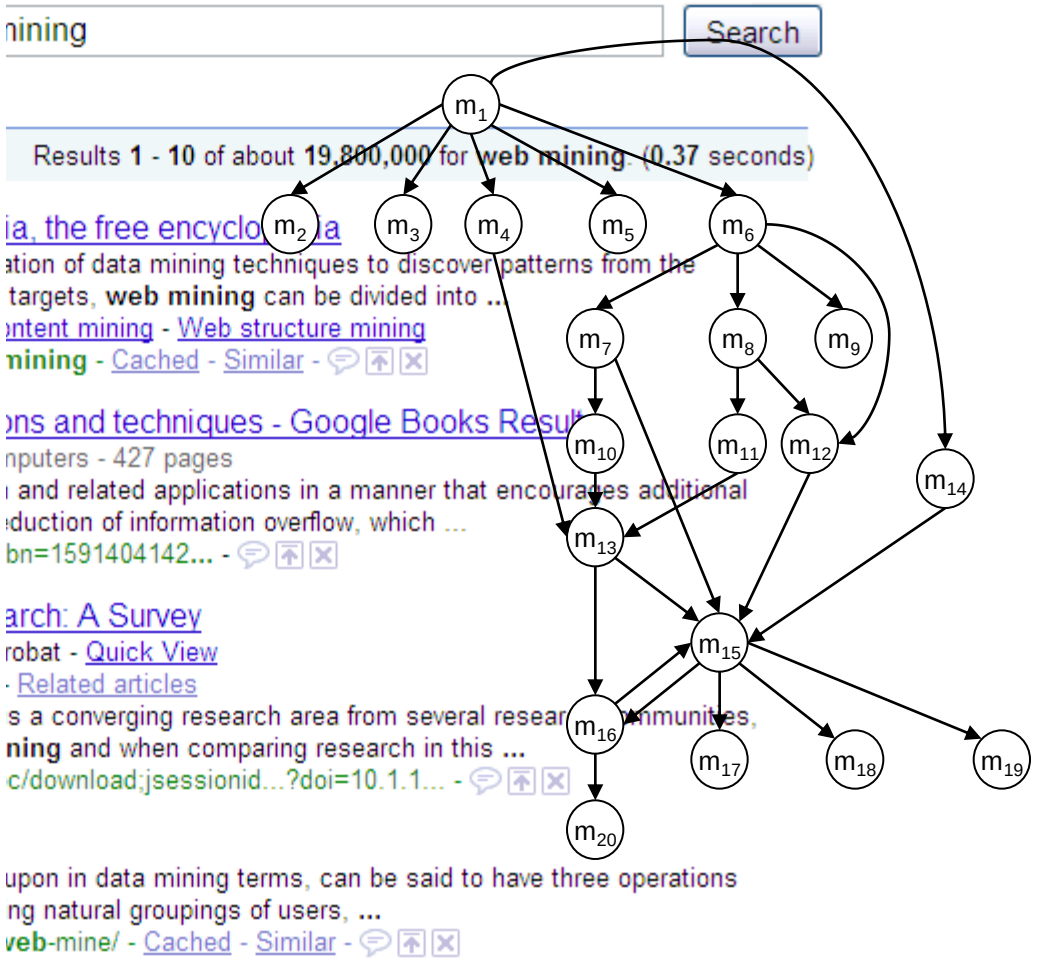


Data Fusion for Feature Location

- Combining information from multiple sources will yield better results than if the data is used separately
  - Previous
    - Textual, Dynamic, and Static
  - Current
    - Textual info from IR
    - Execution info from dynamic tracing
    - **Web mining**

Web Mining




mining
Search
Results 1 - 10 of about 19,800,000 for web mining. (0.37 seconds)
ia, the free encyclo ia
ation of data mining techniques to discover patterns from the
targets, web mining can be divided into ..
ntent mining - Web structure mining
mining - Cached - Similar -
ons and techniques - Google Books Resul
nputers - 427 pages
and related applications in a manner that encourages additional
duction of information overflow, which ...
bn=1591404142... -
arch: A Survey
robat - Quick View
- Related articles
s a converging research area from several resear mmunities,
ning and when comparing research in this ...
c/download;jsessionid...?doi=10.1.1... -
upon in data mining terms, can be said to have three operations
ng natural groupings of users, ...
veb-mine/ - Cached - Similar -
m1
m2
m3
m4
m5
m6
m7
m8
m9
m10
m11
m12
m13
m14
m15
m16
m17
m18
m19
m20

Web Mining Algorithms

PageRank

- Measure the relative importance of a web page
- Used by the Google search engine
- Link from *X* to *Y* means a vote by *X* for *Y*
- A node's PageRank depends on # incoming links and the PageRank of nodes that link to it

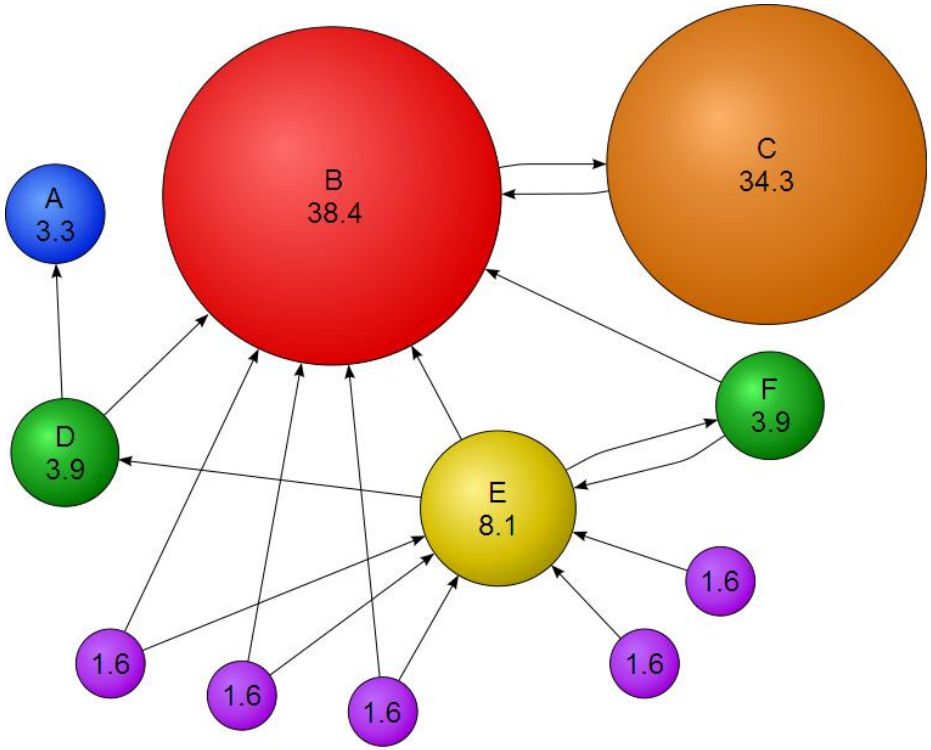


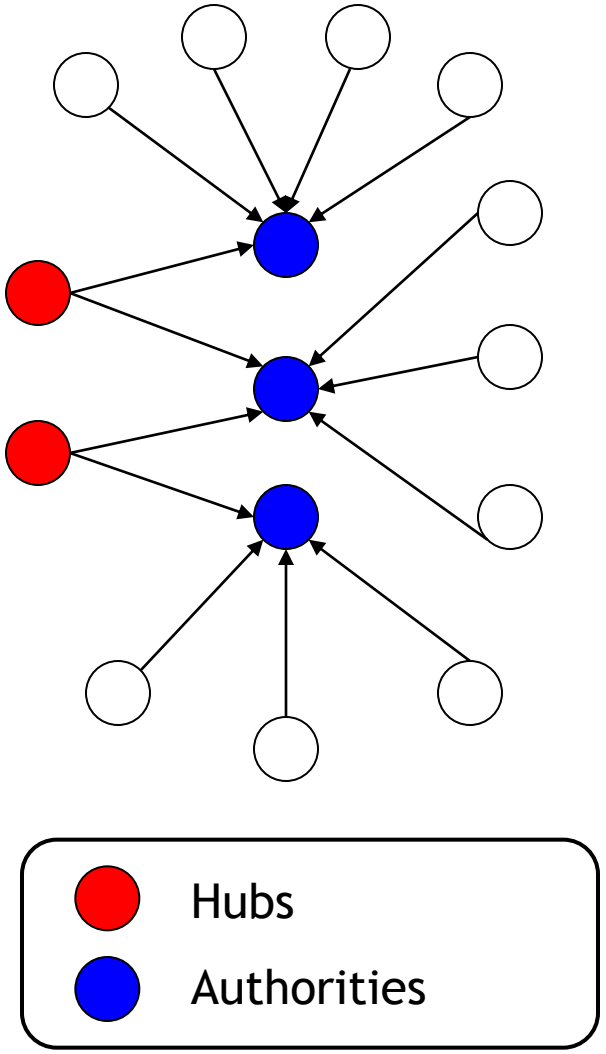
Image source: http://en.wikipedia.org/wiki/Pagerank

Brin, S. and Page, L., "The Anatomy of a Large-Scale Hypertextual Web Search Engine", in Proc. of 7th International Conference on World Wide Web, Brisbane, Australia, 1998, pp. 107-117.

Web Mining Algorithms

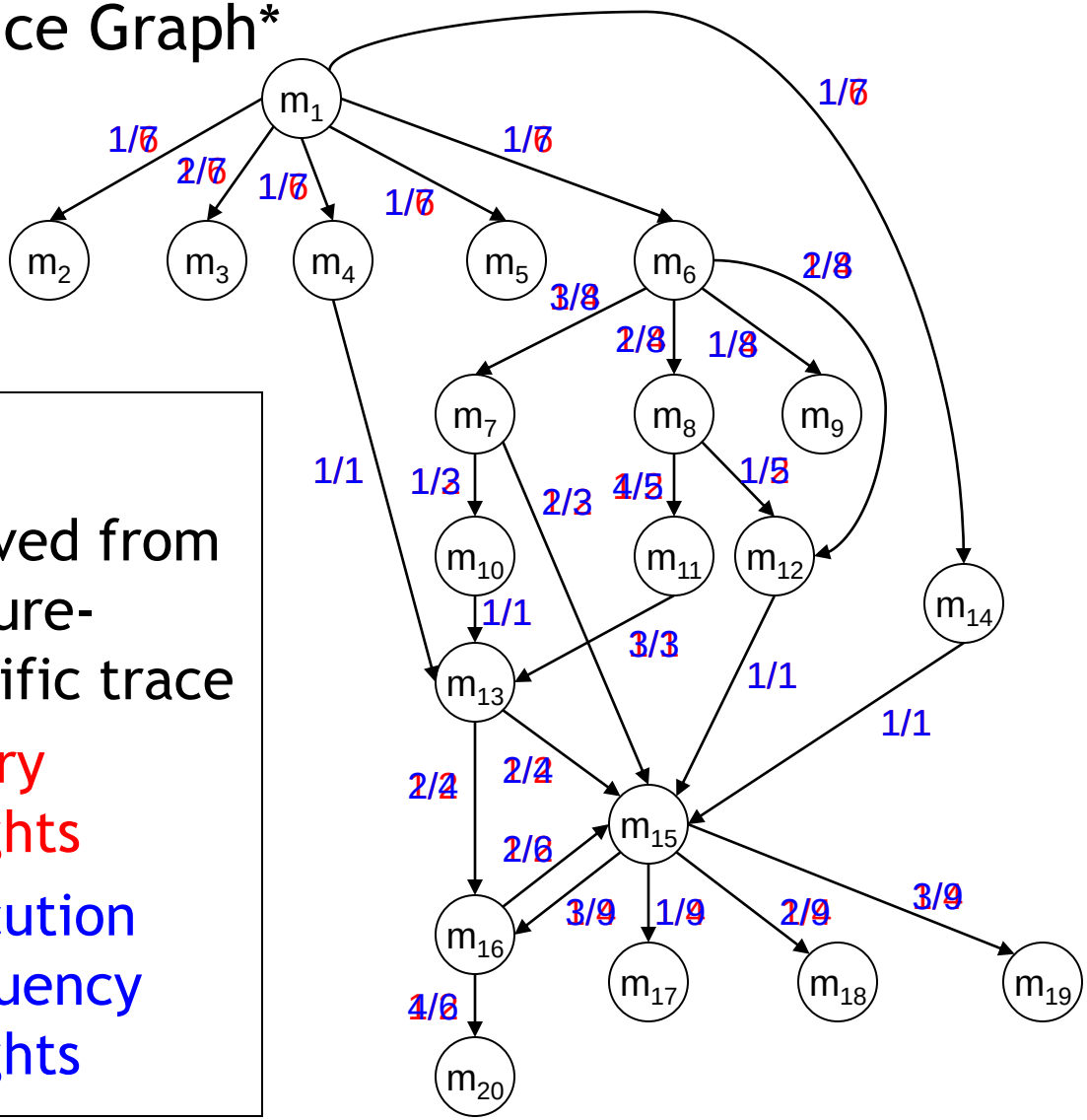
HITS

- Hyperlinked-Induced Topic Search
- Identifies hub and authority pages
- Hubs point to many good authorities
- Authorities are pointed to by many hubs



Kleinberg, J. M., "Authoritative sources in a hyperlinked environment", *Journal of the ACM*, vol. 46, no. 5, 1999, pp. 604-632.

Probabilistic Program Dependence Graph*



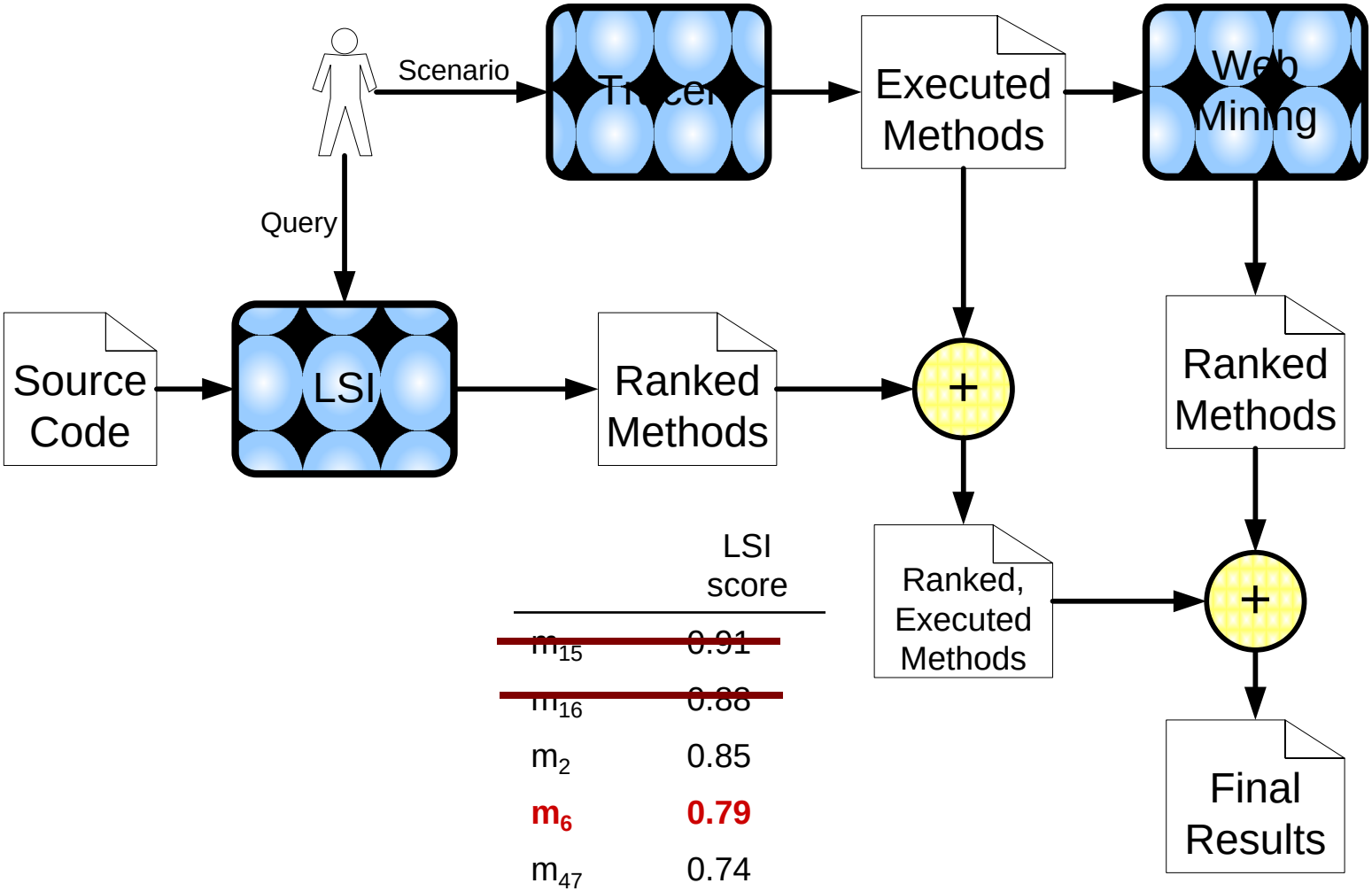
PPDG

- Derived from feature-specific trace
- Binary weights
- Execution frequency weights

15
16
20
13
17
18
19
14
10
12
11
7
8
9
2
3
4
5
6
1

*Baah, G. K., Podgurski, A., and Harrold, M. J. 2008. The probabilistic program dependence graph and its application to fault diagnosis. In *Proceedings of the 2008 International Symposium on Software Testing and Analysis*, 2008.

Incorporating Web Mining with Feature Location



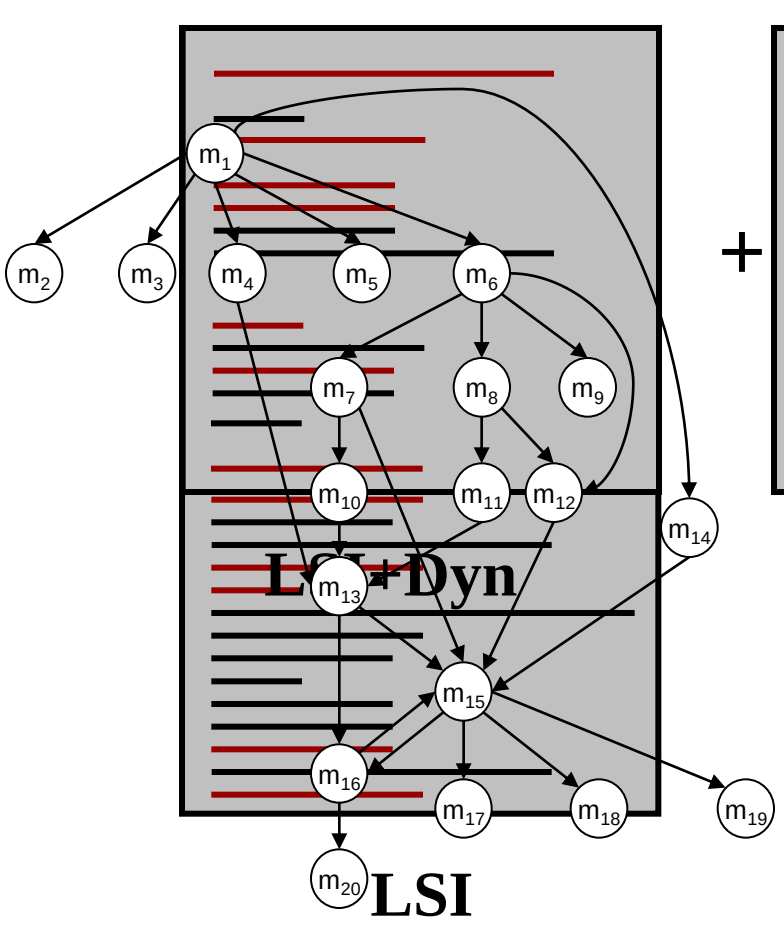
	LSI score
~~m_{15}~~	~~0.91~~
~~m_{16}~~	~~0.88~~
m_2	0.85
m_6	**0.79**
m_{47}	0.74
~~m_{52}~~	~~0.60~~

	PR
m_{15}	0.14
m_{16}	0.09
m_{20}	0.07
m_{13}	0.04
m_{17}	0.001
...	...

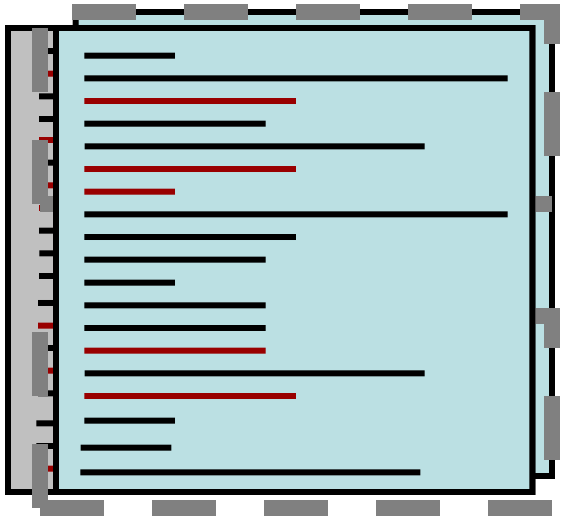
Feature Location Techniques Evaluated

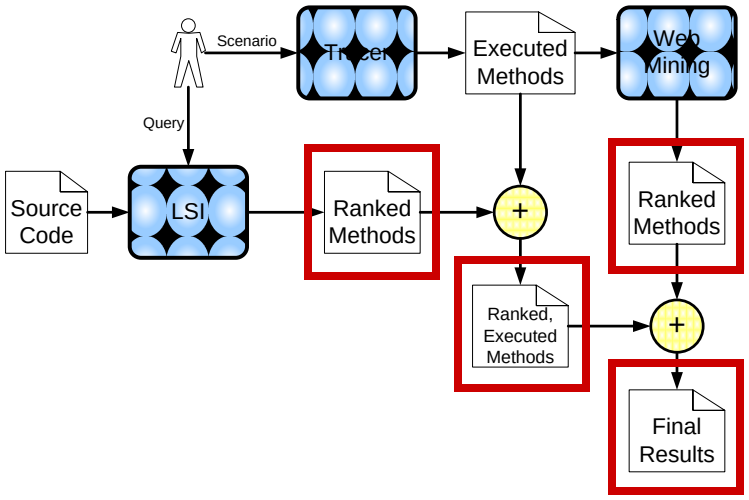
LSI & Dynamic Analysis	Web Mining	LSI, Dyn, & PageRank	LSI, Dyn, & HITS	
LSI LSI+Dyn (baseline)	PR(bin) PR(freq) HITS(h, bin) HITS(h, freq) HITS(a, bin) HITS(a, freq)	LSI+Dyn+PR(bin)^{top} LSI+Dyn+PR(bin)^{bottom} LSI+Dyn+PR(freq)^{top} LSI+Dyn+PR(freq)^{bottom}	LSI+Dyn+HITS(h,bin)^{top} LSI+Dyn+HITS(h,freq)^{top} LSI+Dyn+HITS(a,bin)^{top} LSI+Dyn+HITS(a,freq)^{top}	LSI+Dyn+HITS(h,bin)^{bottom} LSI+Dyn+HITS(h,freq)^{bottom} LSI+Dyn+HITS(a,bin)^{bottom} LSI+Dyn+HITS(a,freq)^{bottom}
Use LSI to rank methods, prune unexecuted	Use web mining algorithm to rank methods.	Use LSI to rank methods. Prune unexecuted. Use web mining algorithm to also rank methods and prune top- or bottom- ranked methods from LSI+Dyn's results.		

Feature Location Techniques Explained



+





Subject Systems

- Eclipse 3.0
  - 10K classes, 120K methods, and 1.6 million LOC
  - 45 features
  - Gold set: methods modified to fix bug
  - Queries: short description from bug report
  - Traces: steps to reproduce bug

Bug 66914 - [typing] Error Message after undo copy/paste

Status: VERIFIED FIXED

Product: JDT
Component: Text
Version: 3.0
Platform: PC All

Importance: P2 major (vote)
Target Milestone: 3.0 RC3
Assigned To: Tom Hofmann
QA Contact:

Reported: 2004-06-14 08:16 EDT by Ralf Schmauder
Modified: 2004-06-21 04:55 EDT (History)
CC List: 2 users

birsan
dirk_baeumer

See Also:

Attachments		
error log (5.43 KB, text/plain) 2004-06-14 08:17 EDT, Ralf Schmauder	*no flags*	Details
LinkedModeUI.diff (4.59 KB, patch) 2004-06-18 09:59 EDT, Tom Hofmann	*no flags*	Details | Diff
Add an attachment (proposed patch, testcase, etc.)		View All

Ralf Schmauder 2004-06-14 08:16:24 EDT Description

```
-create a new Class and generate the main method
-type "sysout" and use the code completion
-type the double quote
-paste Hello World into the double quotes
-try to undo without saving using Ctrl+z

using undo in the menubar does work
```

Subject Systems

- Rhino 1.5
  - 138 classes, 1,870 methods, and 32,134 LOC
  - 241 features
  - Gold set: Eaddy et al.'s dataset*
  - Queries: description in specification
  - Traces: test cases

* http://www.cs.columbia.edu/~eaddy/concerntagger/

Size of Traces

		Min	Max	25%	Med	75%	σ	μ
Eclipse	Methods	88K	1.5MM	312K	525K	1MM	666K	406K
	Unique Methods	1.9K	9.3K	3.9K	5K	6.3K	5.1K	2K
	Size-MB	9.5	290	55	98	202	124	83
	Threads	1	26	7	10	12	10	5
Rhino	Methods	160K	12MM	612K	909K	1.8MM	1.8MM	2.3MM
	Unique Methods	777	1.1K	870	917	943	912	54
	Size-MB	18	1,668	71	104	214	210	273
	Threads	1	1	1	1	1	1	0

Research Questions

- RQ1
  - Does combining web mining algorithms with an existing approach to feature location improve its effectiveness?
- RQ2
  - Which web-mining algorithms, HITS or PageRank, produces better results?

Data Collection & Testing

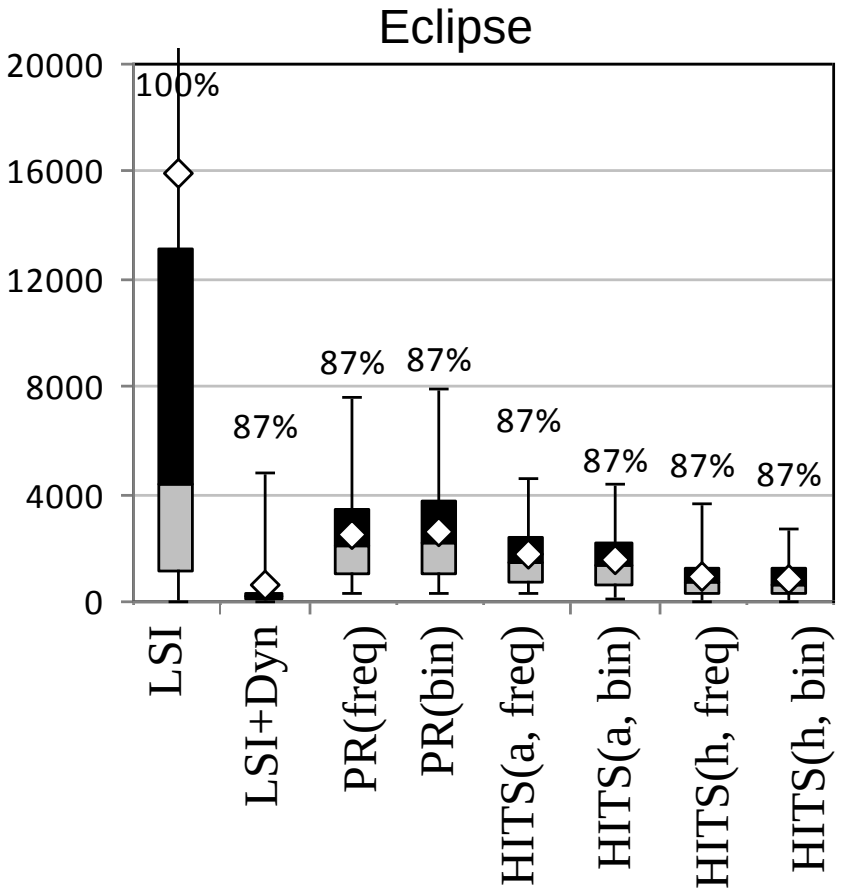
- Effectiveness measure
  - Descriptive statistics
    - 45 Eclipse features
    - 241 Rhino features

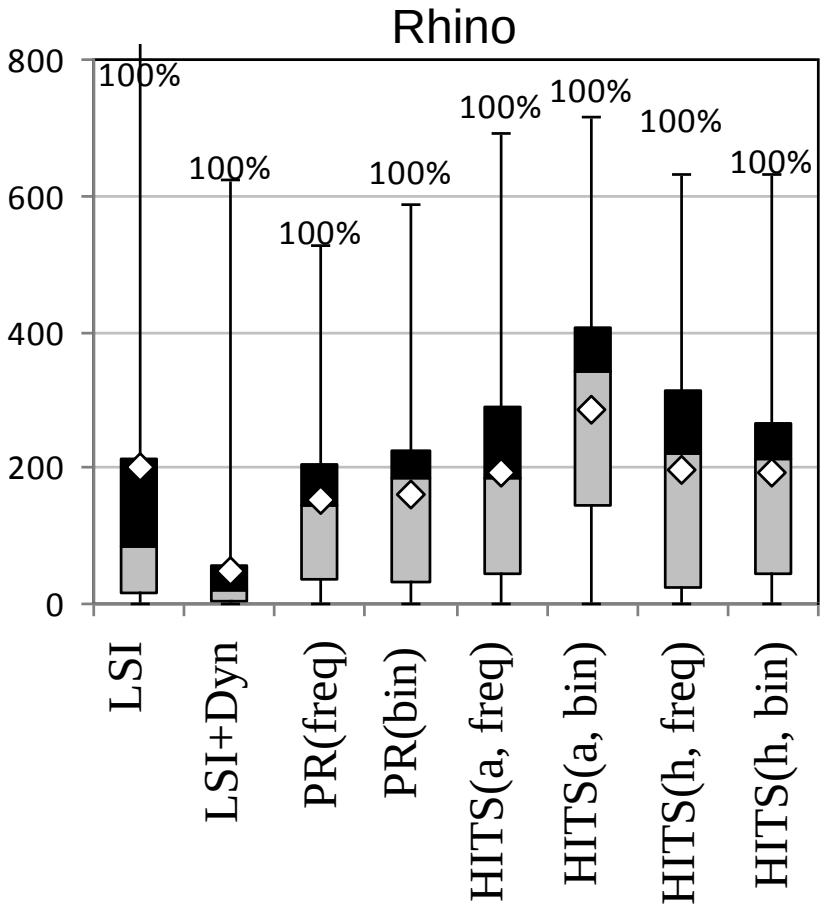
	LSI score
m_{15}	0.91
m_{16}	0.88
m_2	0.85
m_6	**0.79**
m_{47}	0.74
m_{52}	0.60

Effectiveness = 4

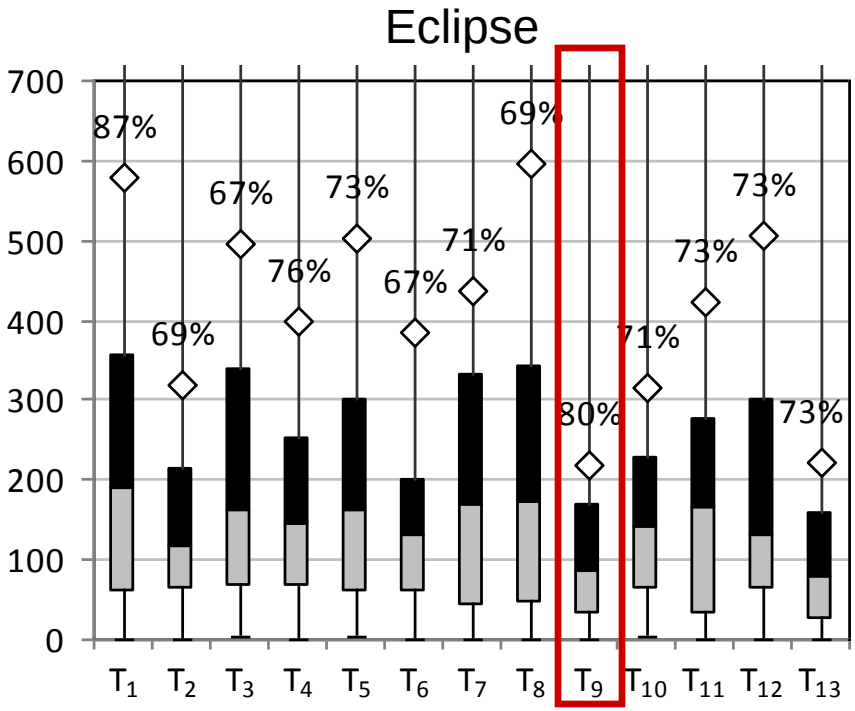
- Statistical Testing
  - Wilcoxon rank sum test
  - Null hypothesis
    - There is no significant difference between the effectiveness of **X** and the baseline (LSI+Dyn).
  - Alternative hypothesis
    - The effectiveness of **X** is significantly better than the baseline (LSI+Dyn).

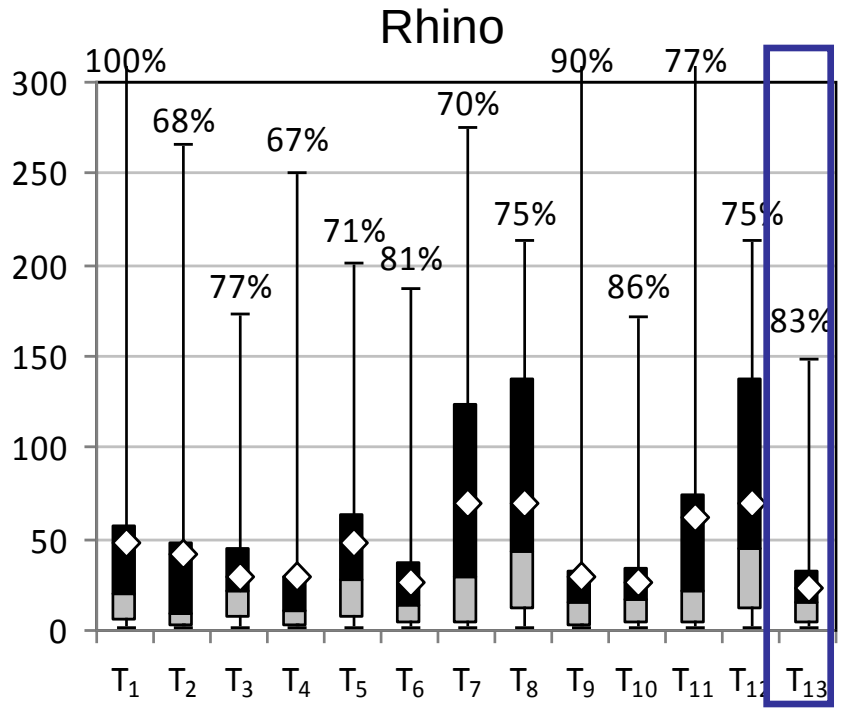
Results: Web Mining Techniques





Results: IR, Dyn, & Web Mining

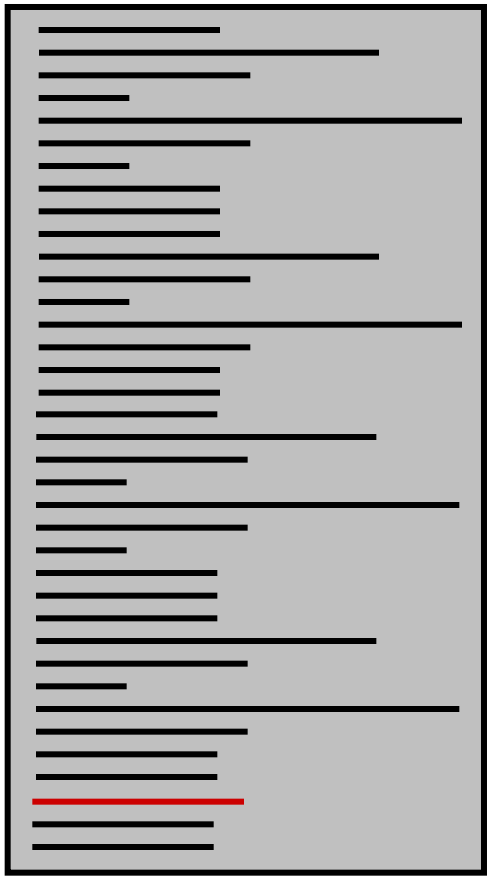




T_1. LSI+Dyn
T_2. LSI+Dyn+PR(freq)$^{\text{top [40, 60]\%}}$
T_3. LSI+Dyn+PR(freq)$^{\text{bot [20, 70]\%}}$
T_4. LSI+Dyn+PR(bin)$^{\text{top [40, 60]\%}}$
T_5. LSI+Dyn+PR(bin)$^{\text{bot [10, 70]\%}}$
T_6. LSI+Dyn+HITS(a, freq)$^{\text{top [30, 70]\%}}$

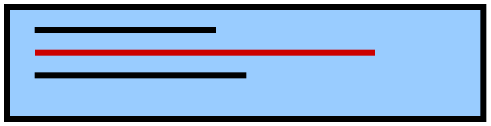
T_7. LSI+Dyn+HITS(a, freq)$^{\text{bot [40, 60]\%}}$
T_8. LSI+Dyn+HITS(h, freq)$^{\text{top [10, 70]\%}}$
T_9. LSI+Dyn+HITS(h, freq)$^{\text{bot [60, 50]\%}}$
T_{10}. LSI+Dyn+HITS(a, bin)$^{\text{top [20, 70]\%}}$
T_{11}. LSI+Dyn+HITS(a, bin)$^{\text{bot [40, 40]\%}}$
T_{12}. LSI+Dyn+HITS(h, bin)$^{\text{top [10, 70]\%}}$
T_{13}. LSI+Dyn+HITS(h, bin) $^{\text{bot [70, 60]\%}}$

A Case in Point: Eclipse exclusion filter



LSI = 1,696

LSI+Dyn = 61



LSI+Dyn+
HITS(h, bin)^{bottom}
= 24

Results of the Wilcoxon Rank Sum test comparing these techniques to the baseline, LSI+Dyn.

α = 0.05.

Null Hypothesis:

There is no significant difference between the effectiveness of ***X*** and the baseline, LSI+Dyn.

	Eclipse	Rhino	Null Hypothesis
PR(bin)	1	1	Not Rejected
PR(freq)	1	1	Not Rejected
HITS(h, bin)	1	1	Not Rejected
HITS(h, freq)	1	1	Not Rejected
HITS(a, bin)	1	1	Not Rejected
HITS(a, freq)	1	1	Not Rejected
LSI+Dyn+PR(bin)^{top}	**< 0.0001**	**< 0.0001**	**Rejected**
LSI+Dyn+PR(bin)^{bottom}	**0.004**	**0**	**Rejected**
LSI+Dyn+PR(freq)^{top}	**< 0.0001**	**< 0.0001**	**Rejected**
LSI+Dyn+PR(freq)^{bottom}	**< 0.0001**	0.74	Not Rejected
LSI+Dyn+HITS(a, freq)^{top}	**0**	**< 0.0001**	**Rejected**
LSI+Dyn+HITS(a, freq)^{bottom}	**< 0.0001**	0.99	Not Rejected
LSI+Dyn+HITS(h, freq)^{top}	**0**	1	Not Rejected
LSI+Dyn+HITS(h, freq)^{bottom}	**< 0.0001**	**< 0.0001**	**Rejected**
LSI+Dyn+HITS(a, bin)^{top}	**< 0.0001**	**< 0.0001**	**Rejected**
LSI+Dyn+HITS(a, bin)^{bottom}	**< 0.0001**	1	Not Rejected
LSI+Dyn+HITS(h, bin)^{top}	**0**	1	Not Rejected
LSI+Dyn+HITS(h, bin)^{bottom}	**< 0.0001**	**< 0.0001**	**Rejected**

Research Questions Revisited

- **RQ1**: Does combining web mining algorithms with an existing approach to feature location improve its effectiveness?
  - Yes

- **RQ2**: Which web-mining algorithms, HITS or PageRank, produces better results?
  - HITS

Best Techniques

- LSI+Dyn+HITS(h, freq)^{bottom}
- LSI+Dyn+HITS(h, bin)^{bottom}

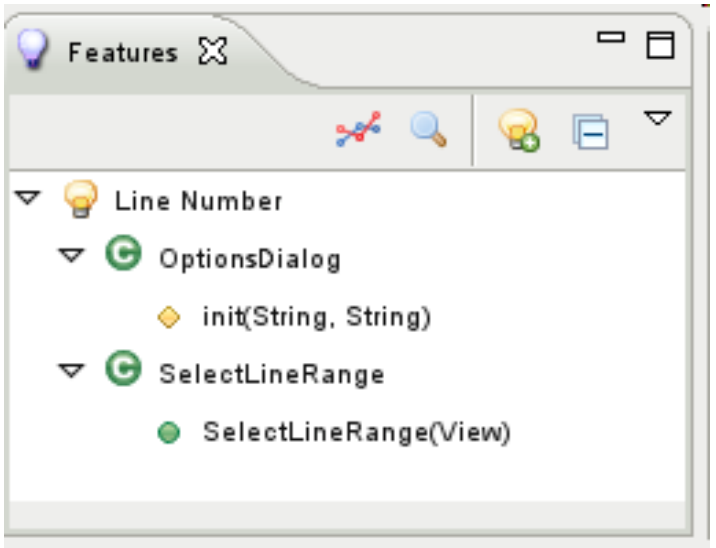
- Methods with low HITS hub values are getters and setters

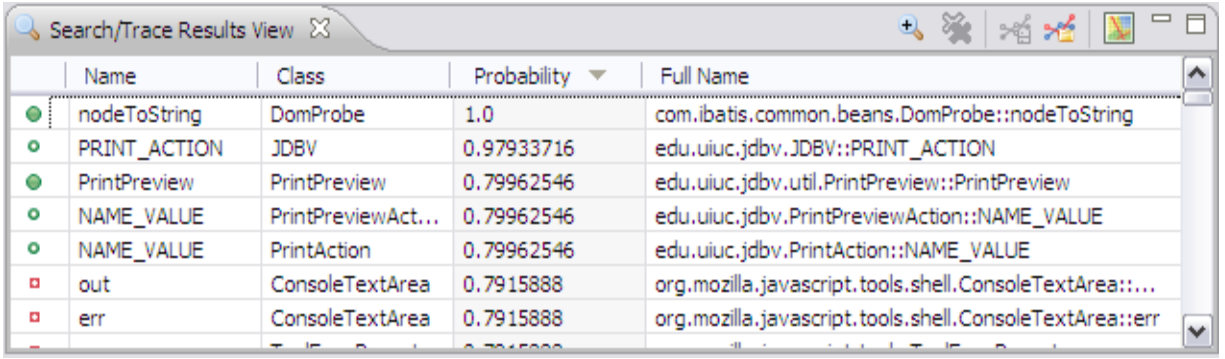
Current Work (not in the paper)

- HITS and PageRank on static vs. dynamic info
- Evaluation first relevant vs. all relevant methods
- Evaluation against fan-in and fan-out and heuristics based on setters and getters
- Impact of thresholds on the filtering power

Tool Support

- FLAT^3
  - Eclipse Plug-in
  - Lucene-based IR
  - Execution tracing
  - Integration
  - Tagging
  - Metrics





http://www.cs.wm.edu/semeru/flat3/

Trevor Savage, Meghan Revelle, and Denys Poshyvanyk. "FLAT3: Feature Location and Textual Tracing Tool." In the Proceedings of the 32nd International Conference on Software Engineering (ICSE'10), Formal Research Tool Demonstration, Cape Town, South Africa, May 2-8, 2010.

Summary

- Proposed and implemented novel methods for feature location based combinations of:
  - Textual analysis, dynamic analysis and web mining
- Evaluated proposed methods on large, open-source systems
- Developed practical tools for the proposed approaches
- Released benchmarks for feature location:
  - http://www.cs.wm.edu/semeru/data/icpc10-data-fusion/

Thank you. Questions?

SEMERU @ William and Mary

http://www.cs.wm.edu/semeru/

denys@cs.wm.edu