Combining Formal Concept Analysis with Information Retrieval for Concept Location in Source Code

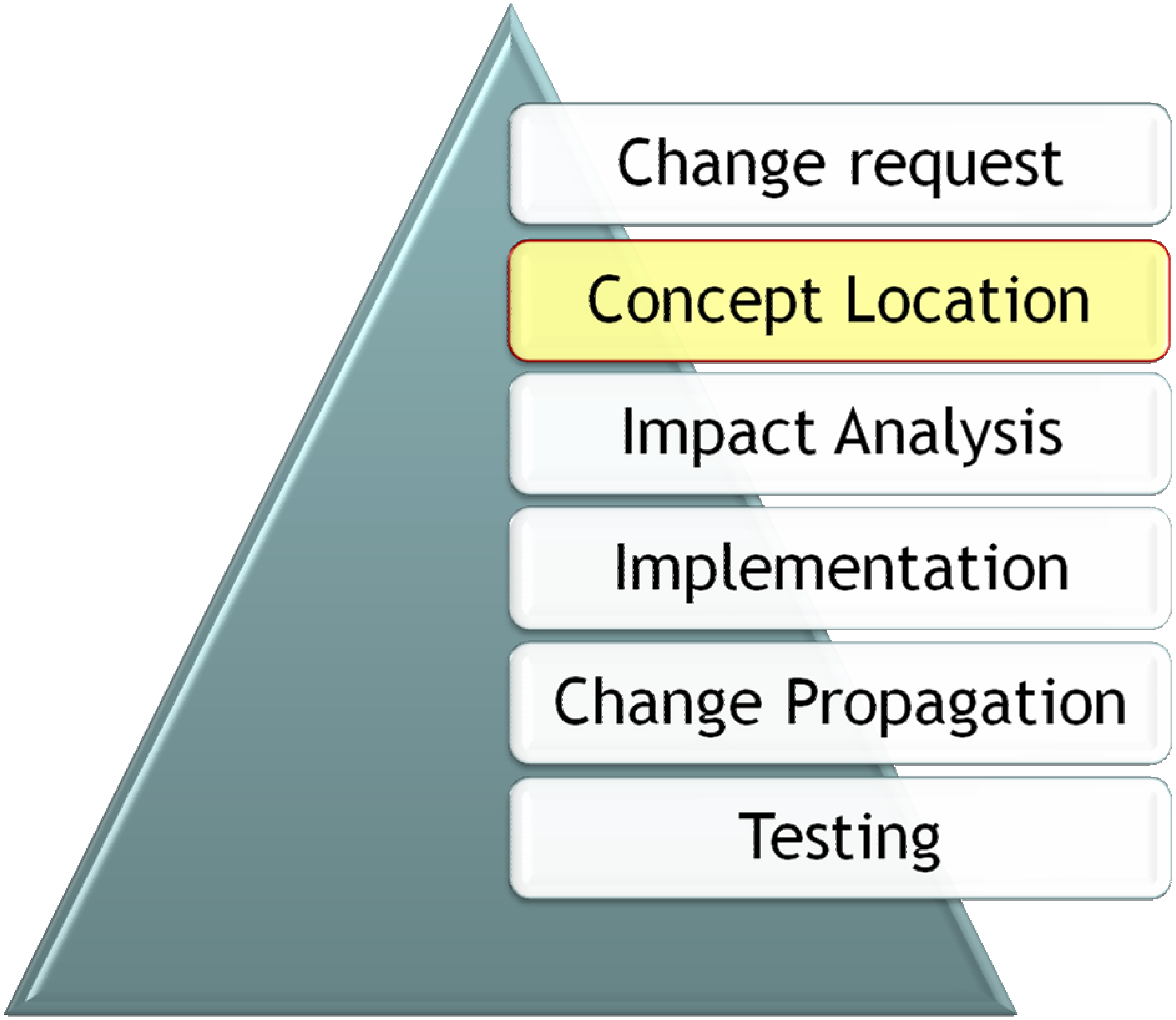
Denys Poshyvanyk and Andrian Marcus

SEVERE Group @

WAYNE STATE UNIVERSITY

ICPC 2007, Banff, Alberta

Incremental Change of Software



Change request

Concept Location

Impact Analysis

Implementation

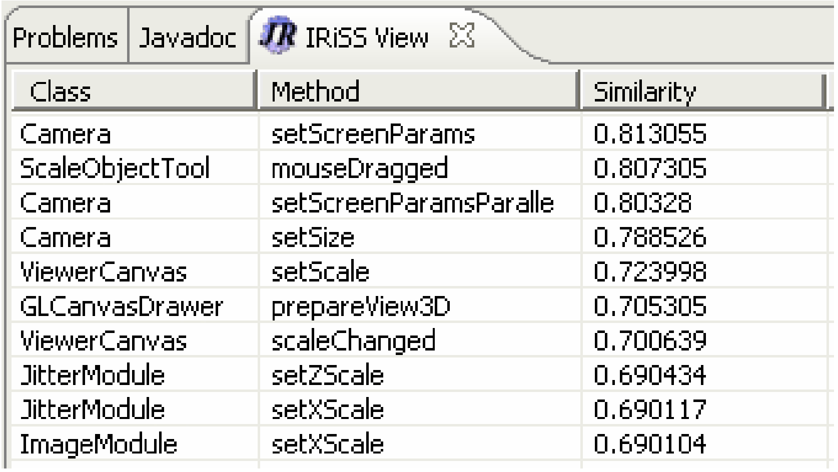
Change Propagation

Testing

State of the Art in Concept Location

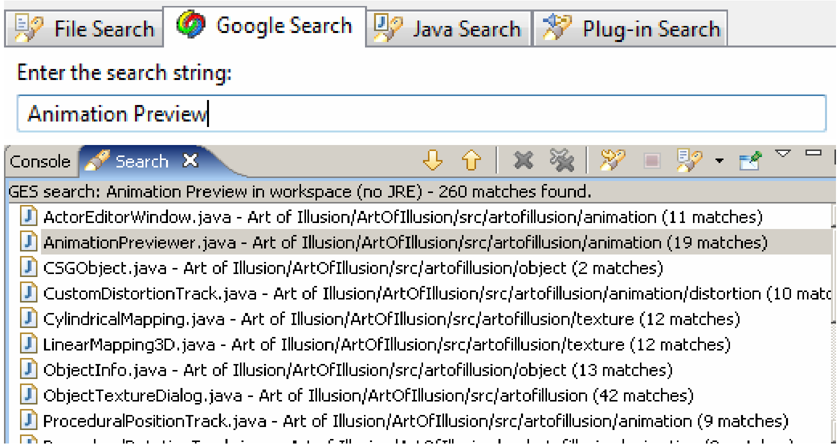
- Static
  - Dependency based search [Rajlich’00]
  - Information Retrieval based methods [Marcus’04]
- Dynamic
  - Execution traces - Reconnaissance [Wilde’92]
  - Scenario based probabilistic ranking [Antoniol’06]
- Combined
  - FCA + Execution Traces [Eisenbarth’03][Tonella’04]
  - IR + Execution Traces [Poshyvanyk’07]
  - …

JIRiSS



Class	Method	Similarity
Camera	setScreenParams	0.813055
ScaleObjectTool	mouseDragged	0.807305
Camera	setScreenParamsParalle	0.80328
Camera	setSize	0.788526
ViewerCanvas	setScale	0.723998
GLCanvasDrawer	prepareView3D	0.705305
ViewerCanvas	scaleChanged	0.700639
JitterModule	setZScale	0.690434
JitterModule	setXScale	0.690117
ImageModule	setXScale	0.690104

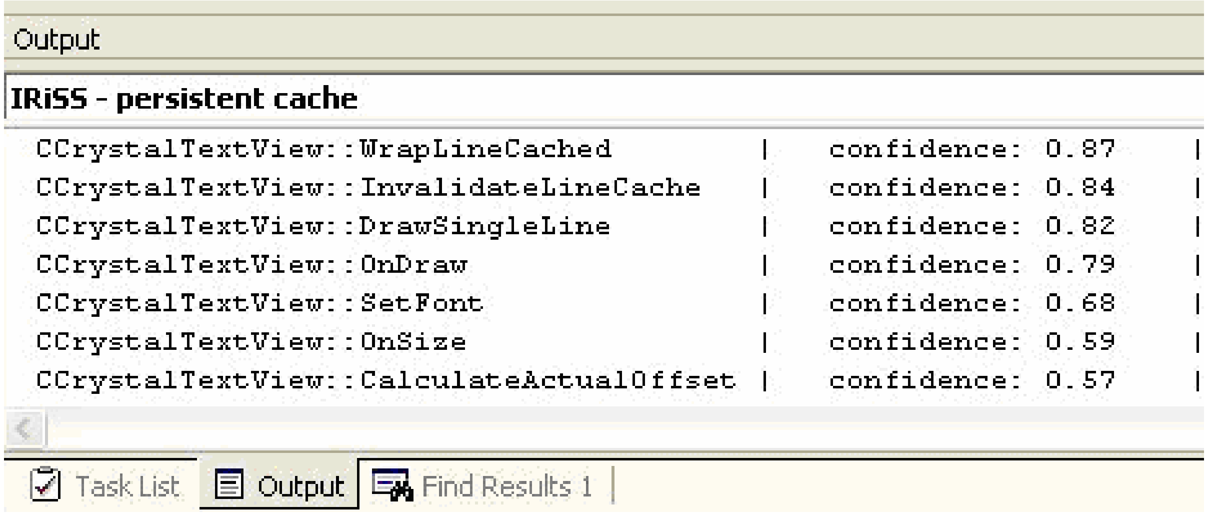
GES



Enter the search string: Animation Preview

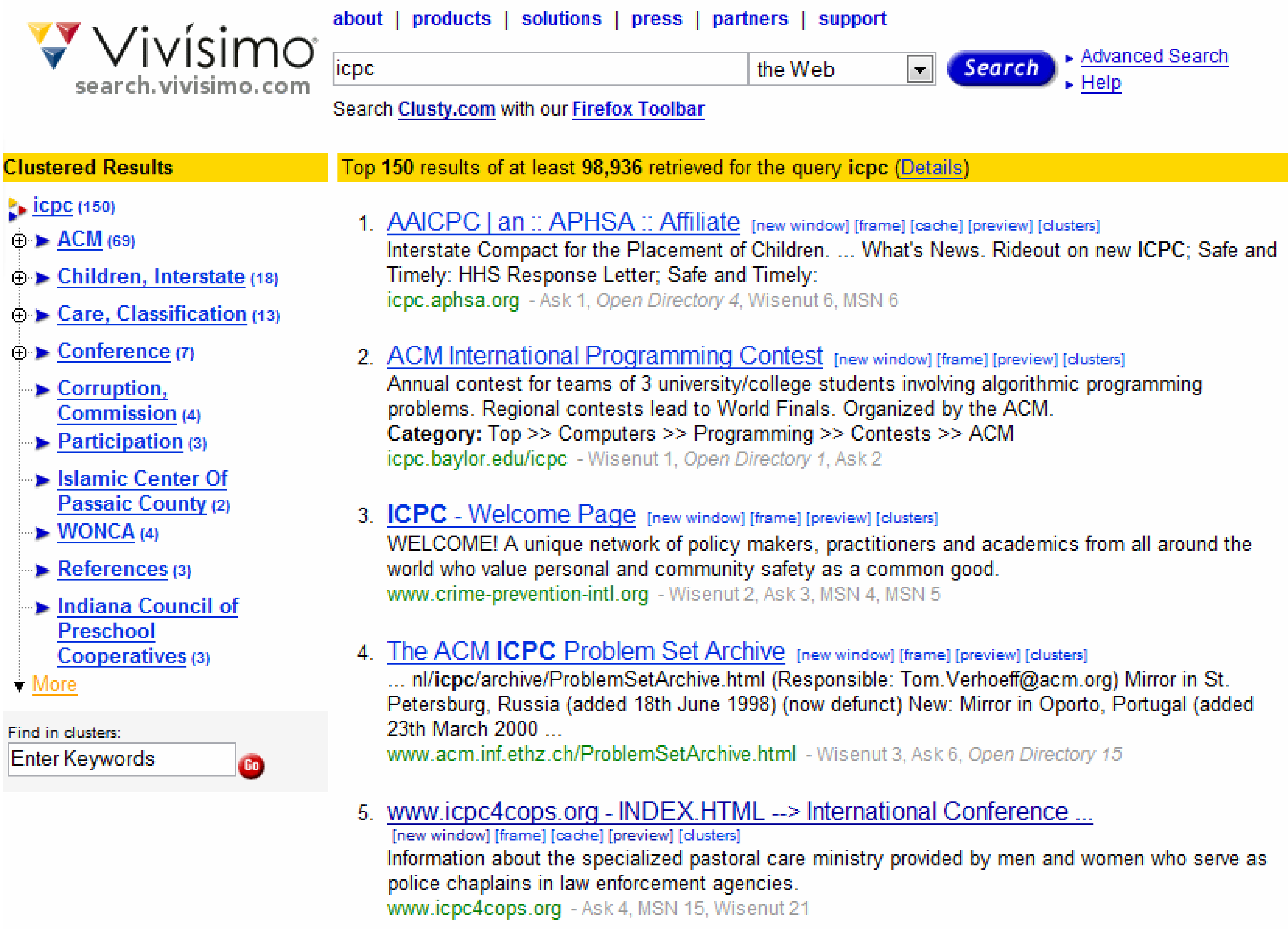
```
GES search: Animation Preview in workspace (no JRE) - 260 matches found.
ActorEditorWindow.java - Art of Illusion/ArtOfIllusion/src/artofillusion/animation (11 matches)
AnimationPreviewer.java - Art of Illusion/ArtOfIllusion/src/artofillusion/animation (19 matches)
CSGObject.java - Art of Illusion/ArtOfIllusion/src/artofillusion/object (2 matches)
CustomDistortionTrack.java - Art of Illusion/ArtOfIllusion/src/artofillusion/animation/distortion (10 matc
CylindricalMapping.java - Art of Illusion/ArtOfIllusion/src/artofillusion/texture (12 matches)
LinearMapping3D.java - Art of Illusion/ArtOfIllusion/src/artofillusion/texture (12 matches)
ObjectInfo.java - Art of Illusion/ArtOfIllusion/src/artofillusion/object (13 matches)
ObjectTextureDialog.java - Art of Illusion/ArtOfIllusion/src/artofillusion (42 matches)
ProceduralPositionTrack.java - Art of Illusion/ArtOfIllusion/src/artofillusion/animation (9 matches)
```

IRiSS



IRiSS - persistent cache

```
CCrystalTextView::WrapLineCached          |   confidence: 0.87     |
CCrystalTextView::InvalidateLineCache     |   confidence: 0.84     |
CCrystalTextView::DrawSingleLine          |   confidence: 0.82     |
CCrystalTextView::OnDraw                  |   confidence: 0.79     |
CCrystalTextView::SetFont                 |   confidence: 0.68     |
CCrystalTextView::OnSize                  |   confidence: 0.59     |
CCrystalTextView::CalculateActualOffset   |   confidence: 0.57     |
```



Search Clusty.com with our Firefox Toolbar

Clustered Results

- icpc (150)
- ACM (69)
- Children, Interstate (18)
- Care, Classification (13)
- Conference (7)
- Corruption, Commission (4)
- Participation (3)
- Islamic Center Of Passaic County (2)
- WONCA (4)
- References (3)
- Indiana Council of Preschool Cooperatives (3)
- More

Find in clusters:

Enter Keywords Go

Top **150** results of at least **98,936** retrieved for the query **icpc** (Details)

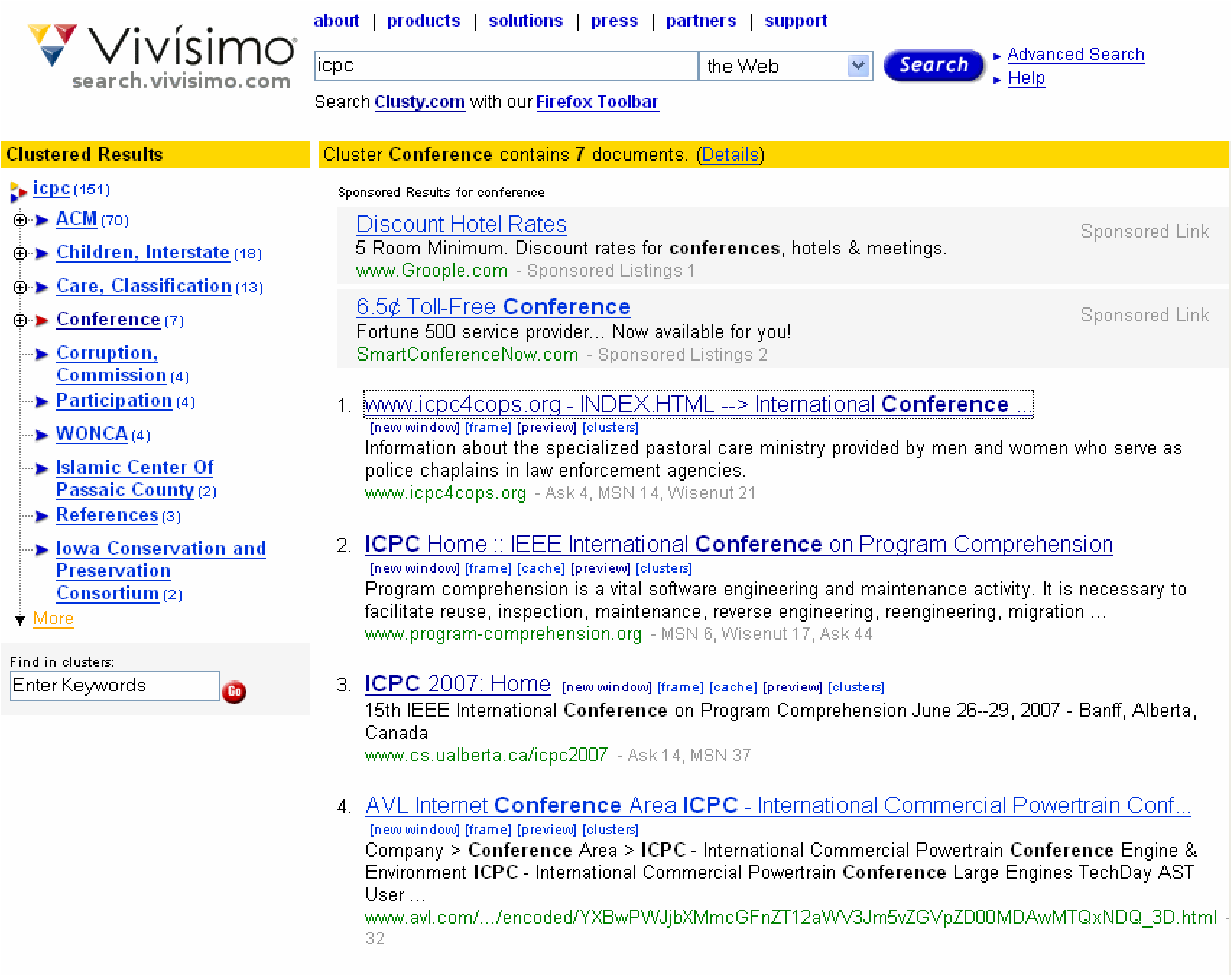
1. AAICPC | an :: APHSA :: Affiliate [new window] [frame] [cache] [preview] [clusters]
Interstate Compact for the Placement of Children. ... What's News. Rideout on new **ICPC**; Safe and Timely: HHS Response Letter; Safe and Timely:
icpc.aphsa.org - Ask 1, *Open Directory* 4, Wisenut 6, MSN 6

2. ACM International Programming Contest [new window] [frame] [preview] [clusters]
Annual contest for teams of 3 university/college students involving algorithmic programming problems. Regional contests lead to World Finals. Organized by the ACM.
Category: Top >> Computers >> Programming >> Contests >> ACM
icpc.baylor.edu/icpc - Wisenut 1, *Open Directory 1*, Ask 2

3. **ICPC** - Welcome Page [new window] [frame] [preview] [clusters]
WELCOME! A unique network of policy makers, practitioners and academics from all around the world who value personal and community safety as a common good.
www.crime-prevention-intl.org - Wisenut 2, Ask 3, MSN 4, MSN 5

4. The ACM **ICPC** Problem Set Archive [new window] [frame] [preview] [clusters]
... nl/**icpc**/archive/ProblemSetArchive.html (Responsible: Tom.Verhoeff@acm.org) Mirror in St. Petersburg, Russia (added 18th June 1998) (now defunct) New: Mirror in Oporto, Portugal (added 23th March 2000 ...
www.acm.inf.ethz.ch/ProblemSetArchive.html - Wisenut 3, Ask 6, *Open Directory 15*

5. www.icpc4cops.org - INDEX.HTML --> International Conference ...
[new window] [frame] [cache] [preview] [clusters]
Information about the specialized pastoral care ministry provided by men and women who serve as police chaplains in law enforcement agencies.
www.icpc4cops.org - Ask 4, MSN 15, Wisenut 21



about | products | solutions | press | partners | support

icpc | the Web | Search

▸ Advanced Search
▸ Help

Search Clusty.com with our Firefox Toolbar

Clustered Results

- icpc (151)
  - ACM (70)
  - Children, Interstate (18)
  - Care, Classification (13)
  - Conference (7)
    - Corruption, Commission (4)
    - Participation (4)
    - WONCA (4)
    - Islamic Center Of Passaic County (2)
    - References (3)
    - Iowa Conservation and Preservation Consortium (2)
- More

Find in clusters:
Enter Keywords | Go

Cluster **Conference** contains **7** documents. (Details)

Sponsored Results for conference

Discount Hotel Rates — Sponsored Link
5 Room Minimum. Discount rates for **conferences**, hotels & meetings.
www.Groople.com - Sponsored Listings 1

6.5¢ Toll-Free **Conference** — Sponsored Link
Fortune 500 service provider... Now available for you!
SmartConferenceNow.com - Sponsored Listings 2

1. www.icpc4cops.org - INDEX.HTML --> International **Conference** ...
[new window] [frame] [preview] [clusters]
Information about the specialized pastoral care ministry provided by men and women who serve as police chaplains in law enforcement agencies.
www.icpc4cops.org - Ask 4, MSN 14, Wisenut 21

2. **ICPC** Home :: IEEE International **Conference** on Program Comprehension
[new window] [frame] [cache] [preview] [clusters]
Program comprehension is a vital software engineering and maintenance activity. It is necessary to facilitate reuse, inspection, maintenance, reverse engineering, reengineering, migration ...
www.program-comprehension.org - MSN 6, Wisenut 17, Ask 44

3. **ICPC** 2007: Home [new window] [frame] [cache] [preview] [clusters]
15th IEEE International **Conference** on Program Comprehension June 26--29, 2007 - Banff, Alberta, Canada
www.cs.ualberta.ca/icpc2007 - Ask 14, MSN 37

4. AVL Internet **Conference** Area **ICPC** - International Commercial Powertrain Conf...
[new window] [frame] [preview] [clusters]
Company > **Conference** Area > **ICPC** - International Commercial Powertrain **Conference** Engine & Environment **ICPC** - International Commercial Powertrain **Conference** Large Engines TechDay AST User ...
www.avl.com/.../encoded/YXBwPWJjbXMmcGFnZT12aWV3Jm5vZGVpZD00MDAwMTQxNDQ_3D.html - 32

Concept Location with Concept Lattices

1. Creating a corpus of a software system
2. Indexing with Latent Semantic Indexing
3. Formulating a query
4. Ranking methods
5. Selecting words
6. Clustering with Formal Concept Analysis
7. Examining results

Creating a corpus of a software system

- Parsing source code and extracting documents
  - corpus – collection of documents (e.g., methods)

- Removing non-literals and stop words
  - common words in English, standard function library names, programming language keywords

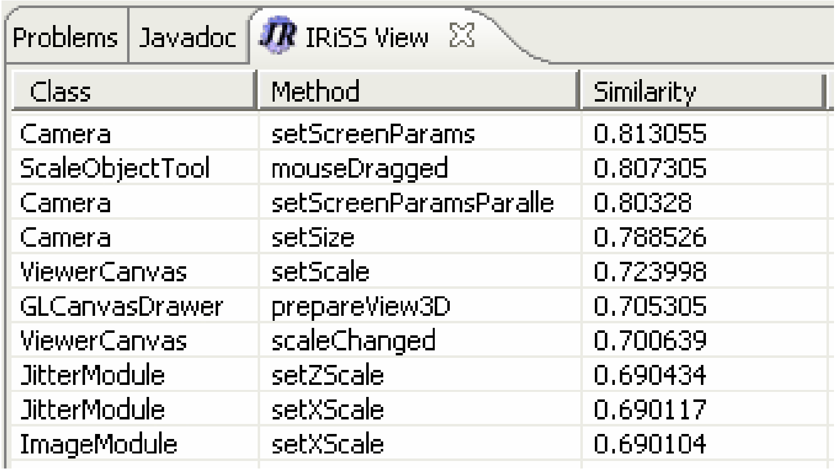
- Preprocessing: split_identifiers and SplitIdentifiers

Concept Location with Concept Lattices

1. Creating a corpus of a software system
2. Indexing with Latent Semantic Indexing
3. Formulating a query
4. Ranking methods
5. Selecting words
6. Clustering with Formal Concept Analysis
7. Examining results

Concept Location with Concept Lattices

1. Creating a corpus of a software system
2. Indexing with Latent Semantic Indexing
3. Formulating a query
4. Ranking methods
5. Selecting words
6. Clustering with Formal Concept Analysis
7. Examining results



Class	Method	Similarity
Camera	setScreenParams	0.813055
ScaleObjectTool	mouseDragged	0.807305
Camera	setScreenParamsParalle	0.80328
Camera	setSize	0.788526
ViewerCanvas	setScale	0.723998
GLCanvasDrawer	prepareView3D	0.705305
ViewerCanvas	scaleChanged	0.700639
JitterModule	setZScale	0.690434
JitterModule	setXScale	0.690117
ImageModule	setXScale	0.690104

Selecting Descriptive Words

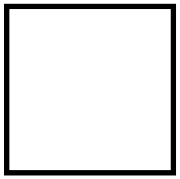
- Ranking criteria – words which are mostly similar to search results than to the rest of the system
- Words in a subset of search results are ranked with LSI

	Page	Paper	Rendering	Device	Printer	
startPage	2	1	1	0	0	…
endPage	2	1	1	0	0	…
getBounds	0	1	2	1	1	…
…						
otherMethod1	3	0	0	2	3	
otherMethod2	2	0	0	3	4	

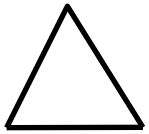
Concept Location with Concept Lattices

1. Creating a corpus of a software system
2. Indexing with Latent Semantic Indexing
3. Formulating a query
4. Ranking methods
5. Selecting words
6. Clustering with Formal Concept Analysis
7. Examining results

FCA Example - classification of geometrical shapes

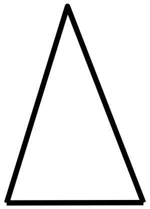


square

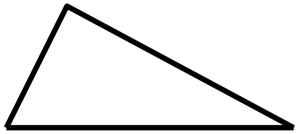


equilateral-triangle

rectangle



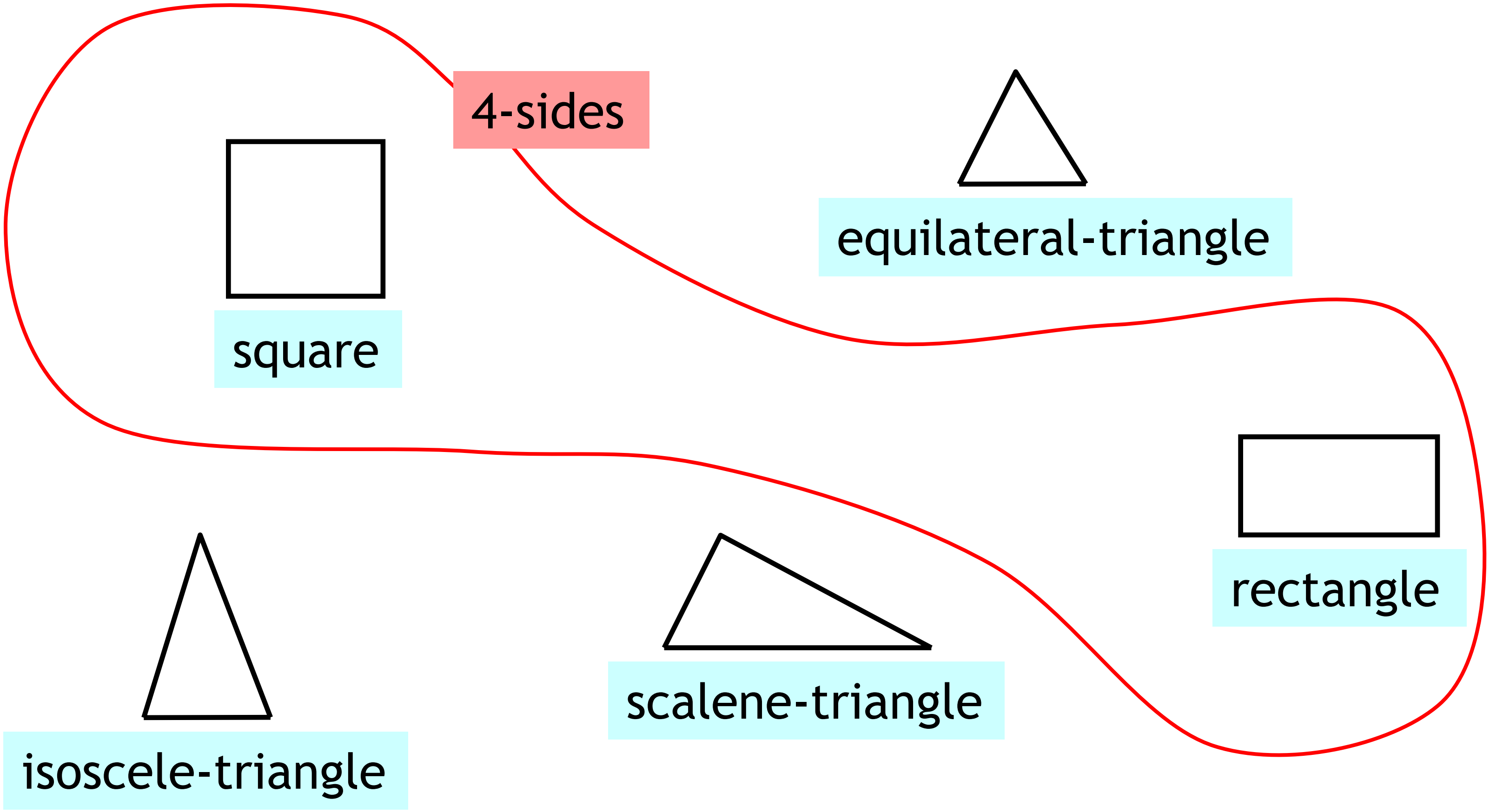
isoscele-triangle



scalene-triangle

Example from Paolo Tonella’s tutorial on Formal Concept Analysis in Software Engineering

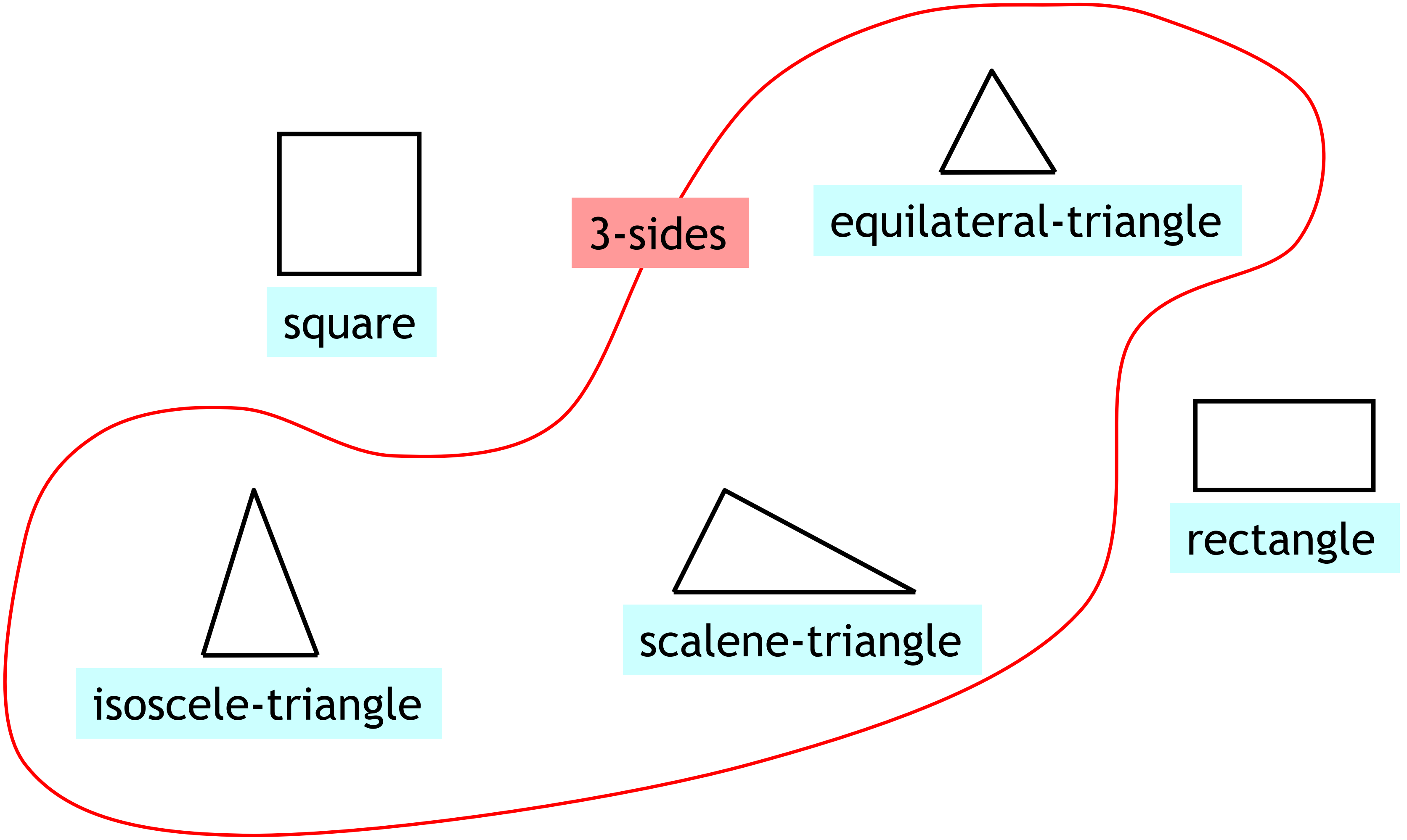
FCA Example - classification of geometrical shapes



Example from Paolo Tonella’s tutorial on Formal Concept Analysis in Software Engineering

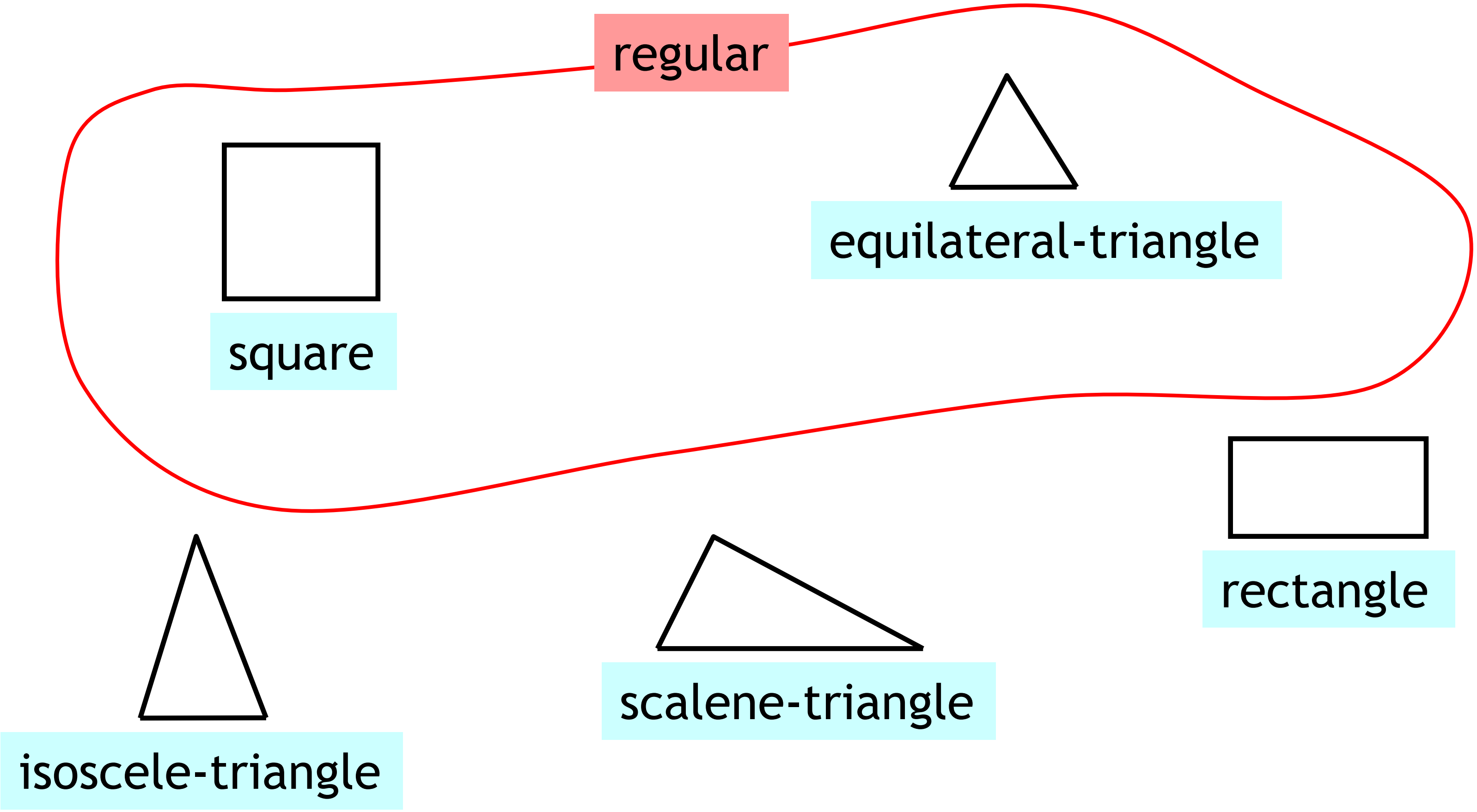
WAYNE STATE UNIVERSITY
SEVERE GROUP
SoftwarE Visualization and Evolution REsearch GROUP

FCA Example - classification of geometrical shapes



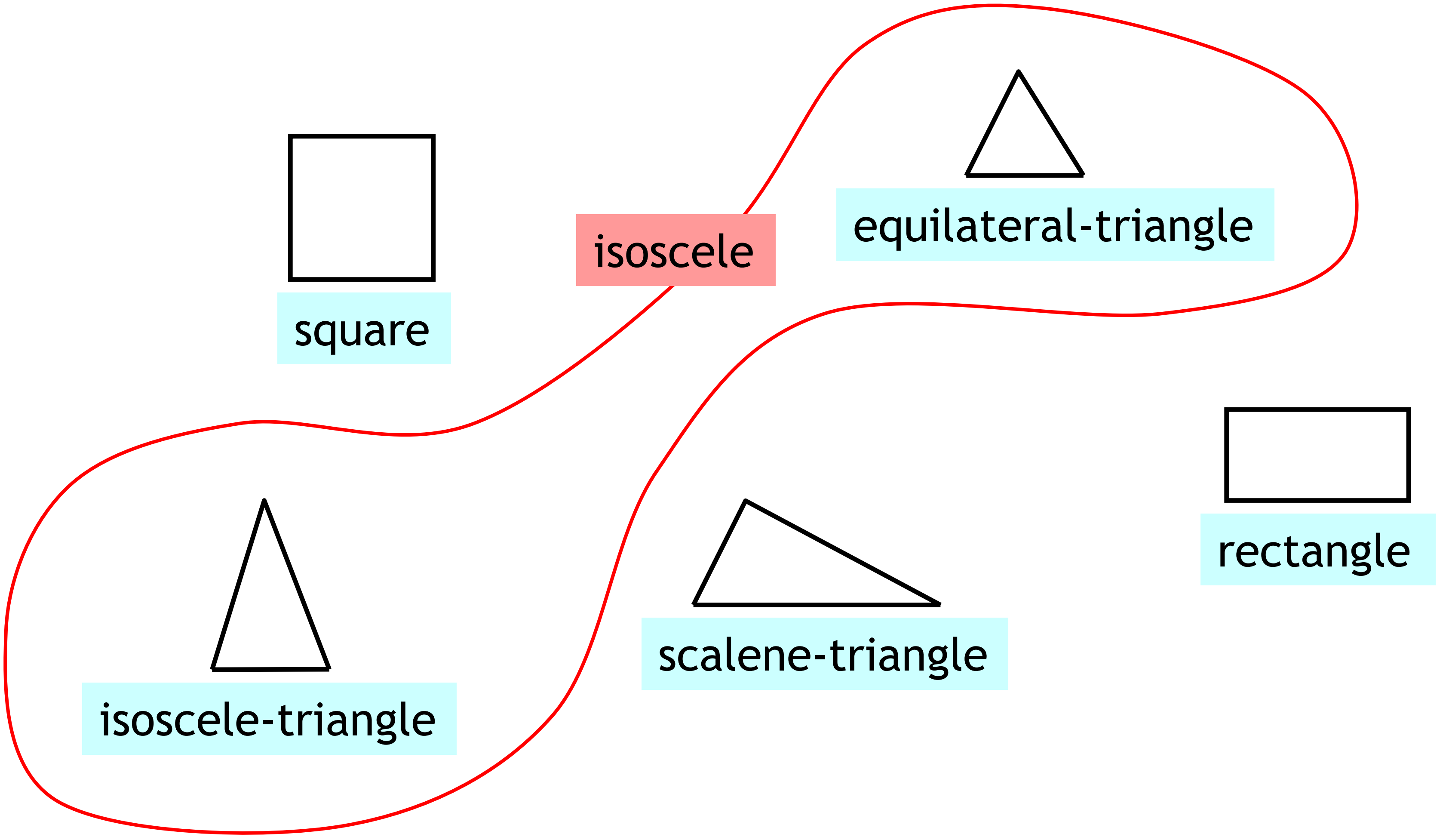
Example from Paolo Tonella’s tutorial on Formal Concept Analysis in Software Engineering

FCA Example - classification of geometrical shapes



Example from Paolo Tonella’s tutorial on Formal Concept Analysis in Software Engineering

FCA Example - classification of geometrical shapes



Example from Paolo Tonella’s tutorial on Formal Concept Analysis in Software Engineering

WAYNE STATE UNIVERSITY
SEVERE GROUP
SoftwarE Visualization and Evolution REsearch GROUP

FCA Example - classification of geometrical shapes

Attributes

Objects

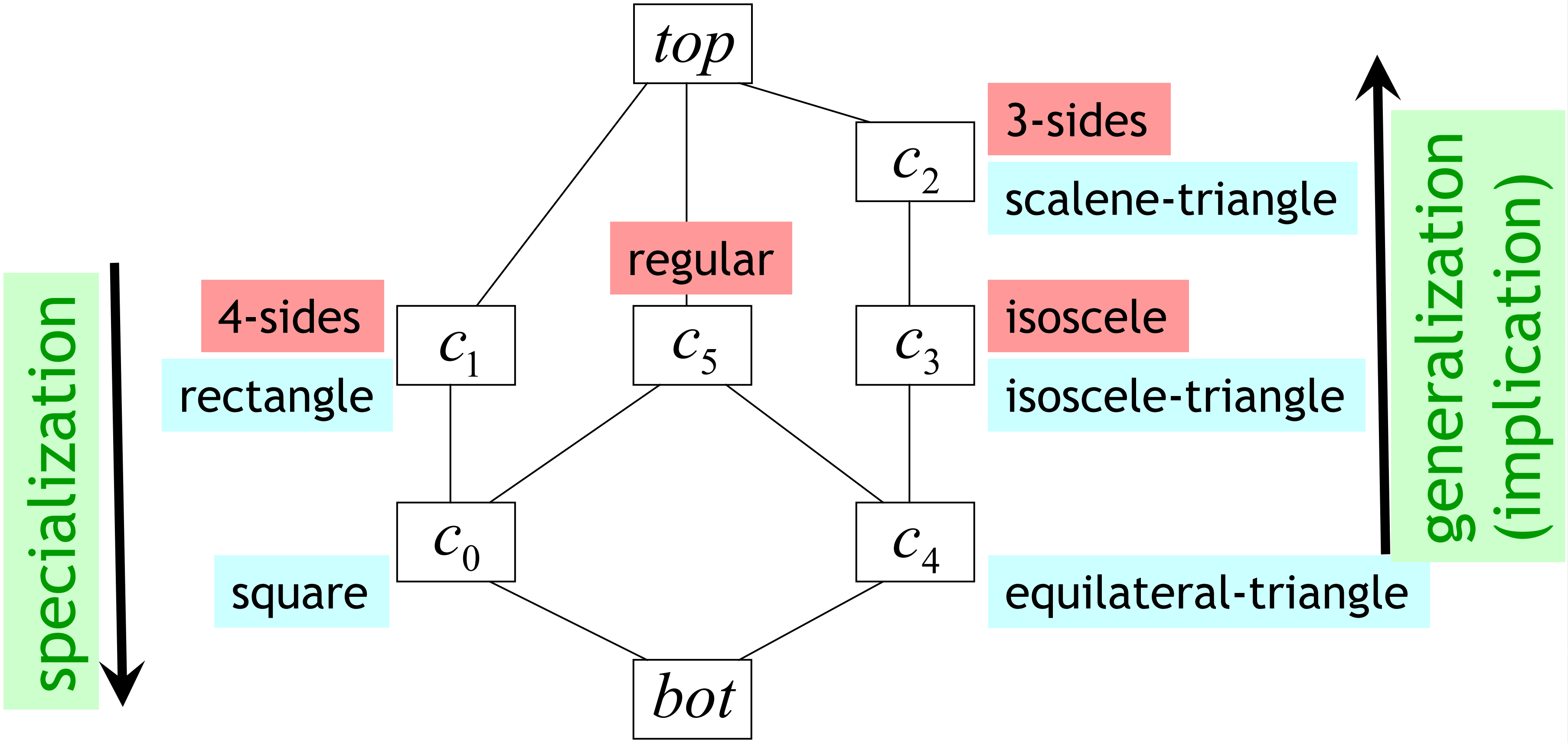
	4-sides	3-sides	regular	isoscele
square	×		×	
rectangle	×			
scalene-triangle		×		
isoscele-triangle		×		×
equilateral-triangle		×	×	×

CONTEXT

Example from Paolo Tonella's tutorial on Formal Concept Analysis in Software Engineering

WAYNE STATE UNIVERSITY

Concept lattice



Example from Paolo Tonella’s tutorial on Formal Concept Analysis in Software Engineering

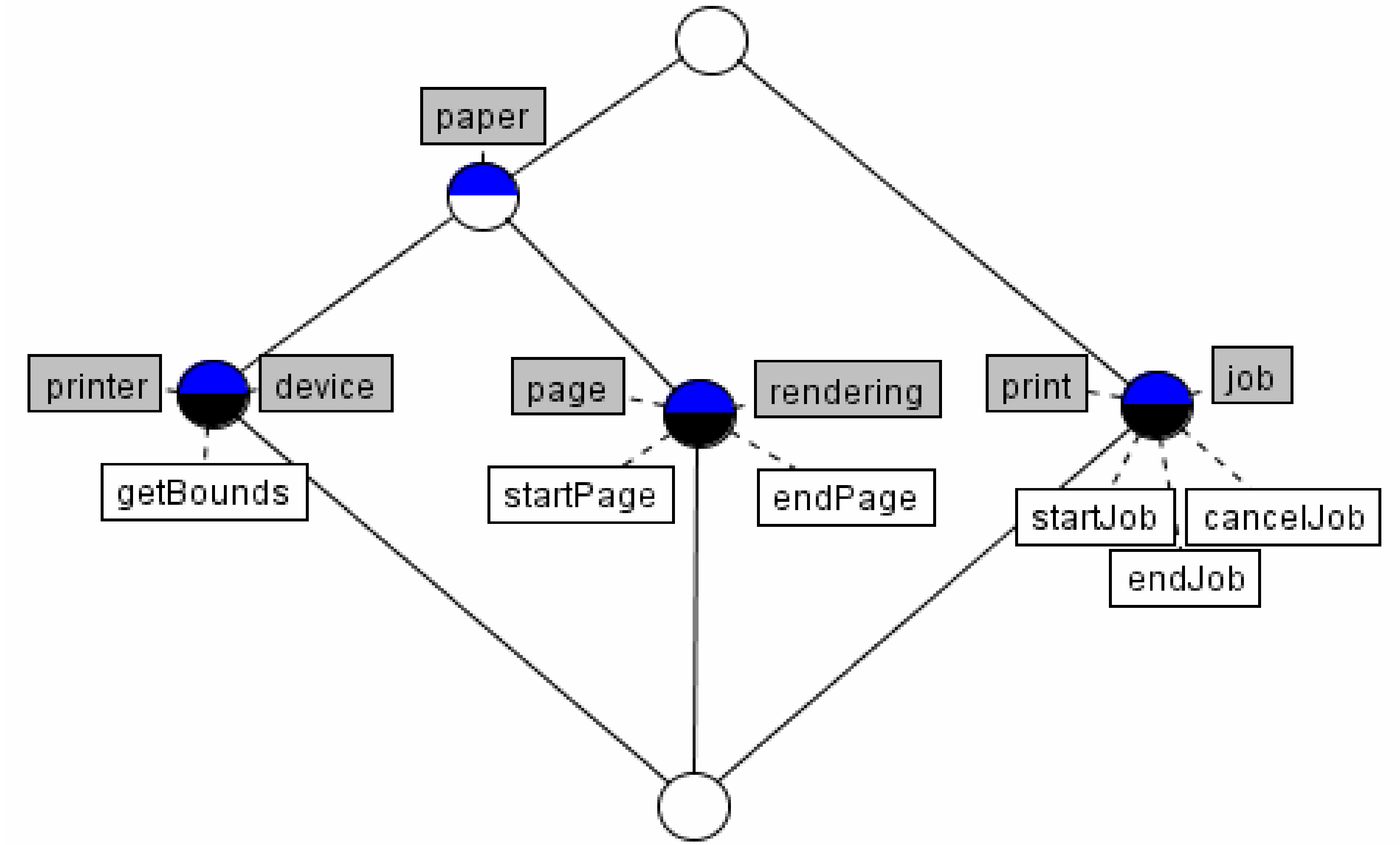
SEVERE GROUP
SoftwarE Visualization and Evolution REsearch GROUP

Clustering Search Results with Formal Concept Analysis

- Subset of methods in search results
- Subset of words selected from search results
- Example: searching for *print page* feature

	printer	print	page	job	device	paper	rendering
startJob		×		×			
endJob		×		×			
cancelJob		×		×			
startPage			×			×	×
endPage			×			×	×
getBounds	×				×	×	

Concept Lattice of Search Results for *Print Page*



Attributes (words) – grey boxes

Objects (methods) – white boxes

Case Study

- Locating concepts associated with bug descriptions
  - Bug fixes contain a set of changed methods
- Source code of Eclipse 3.1
  - vocabulary of unique terms – 56,863[1]
  - number of extracted methods – 86,208

[1] Oxford English Dictionary has 171,476 words in current use

Evaluation of Novel Approach

- Lattices compared with ranked lists:
  - Number of methods: 80, 90, 100
  - Number of words: 10, 15, 20, 25
- Studied properties of lattices:
  - grouping of relevant information
  - browsing overhead of lattice structure
- Redefined measures from [Cigarran'04]:
  - Lattice distillation factor
  - Lattice browsing complexity

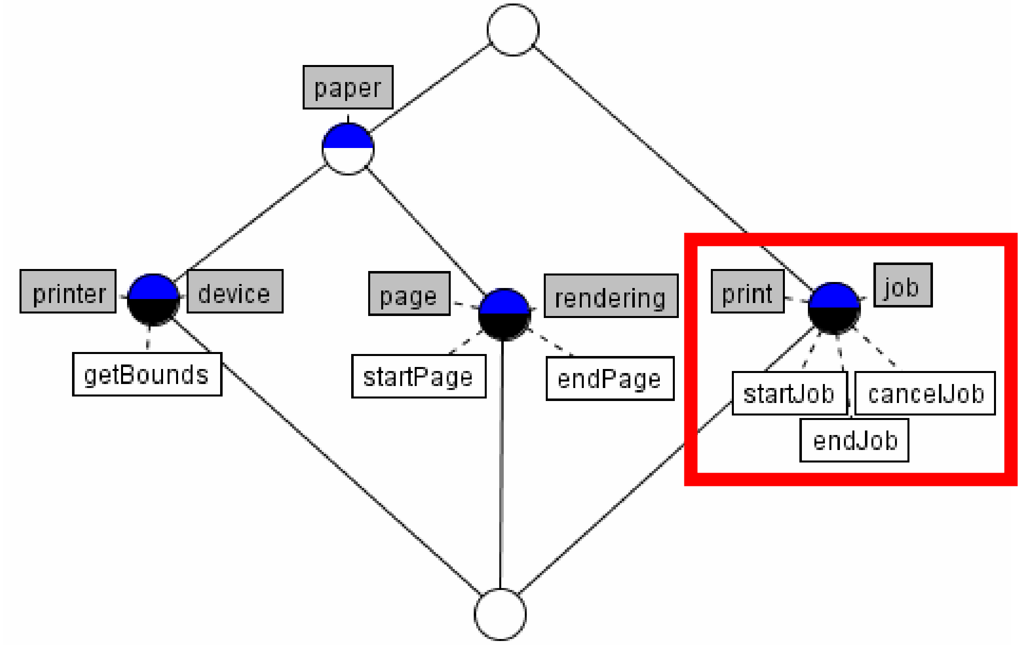
Measures

Lattice Distillation Factor

- how many methods are visited

Lattice Browsing Complexity

- how many concept nodes (categories) are visited while browsing



Ranked list

1. startPage
2. endPage
3. getBounds
4. startJob
5. endJob
6. cancelJob

…

Locating Features in Eclipse - Example

- Table Headers Feature
- "The task list, which uses the native table widget, cannot be ***sorted*** by clicking on the ***table headers***"
- Query: "table headers sorted"
- Associated with bug report# 34160

Locating *Table Headers* Feature

Clustered results into labeled categories

- Table
  - createTable
    - Widget.setData
    - FilteredList.TableUpdater
    - …
    - **Table.createWidget**
  - tableViewer
  - getTable
  - tableValue, keyTable
- Header
  - setHeaderVisible
  - setLineVisible
  - …

Ranked List

1. WidgetTable.put
2. TableTree.getTable
3. EditorsView.getTable
4. SimpleLookupTable.rehash
5. WidgetTable.shells

- …

39.TableTreeEditor.resize

- …

71. Widgets.Table.createWidget

Results for *Table Headers*

Docs	Terms	L_{MBA}	C	C_{VIEW}
100	10	51	17	8
100	15	40	25	11
100	20	38	33	11
100	25	36	39	10
90	10	43	17	8
90	15	36	24	11
90	20	34	31	11
90	25	26	36	11
80	10	38	15	7
80	15	30	22	10
80	20	29	28	10
80	25	23	33	10

- L_{MBA}-rank of the first relevant method in the lattice
- C – number of concepts in the lattice
- C_{VIEW} – number of expanded concepts in the lattice

Conclusions

- Novel representation for the search results in the source code
  - results are ranked and ***structured***
  - labels for topics are ***automatically*** extracted

- Scalable to software of ***any size***
  - fixed number of top search results
  - fixed number of attributes

Current and Future Work

- Different strategies for ranking and selecting words
- Lattice size: best ratio of docs and terms
- Adding IR ranks and similarities to the lattices
- Use FCA and LSI during impact analysis
- More user studies
- Search problem inside the search problem