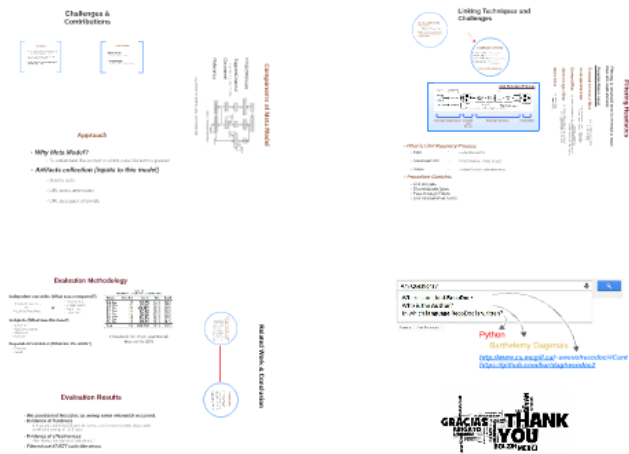


The Process Overview



Recovering Traceability Links between an API and its Learning Resources



By,
Abhishek Kharche

Recovering Traceability Links between an API and its Learning Resources

What is this?
Why we need it?



By,
Abhishek Kharche

What is this? Why we need it?

The screenshot shows a code editor with several lines of Java code. Red arrows originate from specific code elements and point to other files in the project, such as 'ASPX Page', 'Code Refact Page / Code File Page', and 'PageRefactPage'. A red text label 'Method Calls' is overlaid on the arrows.

• Study of how internal components are mapped within API and resources

• Extensive efforts needed in documenting them

• When they should be updated?

The screenshot shows the 'Call Hierarchy' view in Eclipse IDE. It lists various methods that call a specific method, such as 'appendFileExtension(String)', 'appendFragment(String)', 'createURWithCache(String)', 'createDeviceUR(String)', 'createUR(String, boolean, int)', 'createUR(String, boolean, int)', 'createUR(String, boolean, int)', 'createUR(String, boolean, int)', 'createUR(String, boolean, int)', and 'createUR(String, boolean, int)'. A red text label 'Call Hierarchy' is overlaid on the list.



So, what technique is proposed?

- Link code-like-terms (year()) to code elements (DateTime.year())
- What really is a code like term?
e.g. patterns like () for methods
CamelCase for types & <> for XML
- Linking is not spurious because of this process
- Implemented through tool called RecoDoc.

The screenshot shows two windows in Visual Studio. The top window, titled 'Default.aspx', contains JavaScript code for an ASPX page. The bottom window, titled 'AjaxViaNonStaticMethod.aspx.cs', contains C# code for the code-behind page. Red arrows point from the text 'Method Calls' to specific lines in both files: line 27 in the ASPX page and line 28 in the code-behind page.

```

25 //This "CallMyServerMethod" function is created from code behind and will exist in the final rendering of the page
26 CallMyServerMethod(context.flag, context);
27
28
29 }
30
31 //Function that is called on Successful AJAX method call. These are referenced in the "CallMyServerMethod" function that is created from code behind
32 function ServerCallSucceeded(result, context) {
33     document.getElementById("DisplayDate").innerHTML = result;
34 }
35
36 //Function that is called on failure or error in AJAX method call. These are referenced in the "CallMyServerMethod" function that is created from code behind
37 function ServerCallFailed(result, context) {
38     document.getElementById("divDisplay").innerHTML = result;
39 }
40 </script>
41
42 </head>
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

```

Method Calls

- Study of how internal components are mapped within API and resources
- Extensive efforts needed in documenting them
- When they should be updated?

The screenshot shows the 'Call Hierarchy' view in Eclipse IDE. It lists members calling the 'URI' class methods. The methods listed include 'appendFileExtension', 'appendFragment', 'createURIWithCache', 'createDeviceURI', 'createURI', 'createURI(String, boolean, int)', 'createURI(String, boolean)', 'createURI(String)', 'createFileURI', and 'createInputStream'.

```

Members calling 'URI(boolean, String, String, String, boolean, String[], String, String)'
- URI(boolean, String, String, String, boolean, String[], String, String)
  - appendFileExtension(String) : URI - org.eclipse.emf.common.util.URI
  - appendFragment(String) : URI - org.eclipse.emf.common.util.URI
  - createURIWithCache(String) : URI - org.eclipse.emf.common.util.URI
  - createDeviceURI(String) : URI - org.eclipse.emf.common.util.URI
  - createURI(String, boolean, int) : URI - org.eclipse.emf.common.util.URI
  - createURI(String, boolean) : URI - org.eclipse.emf.common.util.URI
  - createURI(String) : URI - org.eclipse.emf.common.util.URI
  - createFileURI(String) : URI - org.eclipse.emf.common.util.URI
  - createInputStream(String) : InputStream - org.eclipse.emf.common.util.URI

```

Call Hierarchy



So, what technique is proposed?

- Link code-like-terms (`year()`) to code elements (`DateTime.year()`)
- What really is a code like term?
 - e.g. patterns like `()` for methods
 - CamelCase for types & `<>` for XML
- Linking is not spurious because of this process
- Implemented through tool called `RecoDoc`.

The Process Overview

Challenges & Contributions



Approach

- Why Meta Model?**
 - To understand the context in which code-like terms in present
- Artifacts collection (inputs to this model)**
 - Source code
 - URLs in documentation
 - URLs in support channels

Evaluation Methodology

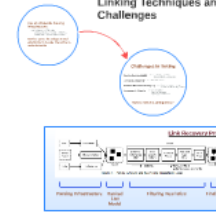
Variable	Value	Results	F1	Recall
Source repository	GitHub	89	0.93	0.74
Source's language	Python	161	0.93	0.74
Subjects (What was the data?)	Code files	180	0.93	0.74
	URLs	124	0.93	0.74
	Documentation	105	0.93	0.74
	Support channels	14	0.93	0.74
	Discussions	12	0.93	0.74
Dependent variables (What are the results?)	Threshold for F1, and Recall was set to 90%			

Evaluation Results

- We considered RecoDoc as wrong when mismatch occurred.
- Existence of heuristics**
 - 14 heuristics out of 20000 code-like terms, 18979 were heuristics, those could be filtered in majority of 100000.
- Existence of effectiveness**
 - After filtering heuristics with heuristics 10.0
- Filtered out 87127 code-like terms



Linking Techniques and Challenges



- What is Link Recovery Process?**
 - Input
 - code-like terms
 - Associated terms
 - code snippets, files, code
 - Output
 - matched list of code elements
- Procedure Contains**
 - Link to terms
 - Documentation types
 - Stack through filters
 - Link to associated terms

Filtering Heuristics



Any Questions?

Where can I find RecoDoc?
Who is the Author?
In which language RecoDoc is written?

Please Enter to search

Python
Barthelemy Dagenais
<http://www.cs.mcgill.ca/~swevo/recoDoc/#Cont>
<https://github.com/bartdagitecodoc2>



Challenges & Contributions

Challenges

- Ambiguity of unstructured natural language
e.g. joda time library (year) in 11 classes
- Reference to External Libraries, so simple mechanical match would fail
- No previously developed technique at such fine level of granularity

Contributions

- **Meta Model**
to represent relationships
- **A Technique**
to link the terms to documentation

Challenges

- Ambiguity of unstructured natural language
e.g. joda time library (year() in 11 classes)
- Reference to External Libraries, so simple mechanical match would fail
- No previously developed technique at such fine level of granularity

Contributions

- ***Meta Model***

to represent relationships

- ***A Technique***

to link the terms to documentation

Approach

- ***Why Meta Model?***
 - To understand the context in which code-like term is present
- ***Artifacts collection (inputs to this model)***
 - Source code
 - URL to documentation
 - URL to support channels

Components of Meta Model

ProjectRelease

SupportChannel

e.g. Mailing lists, Forums

Document

Reference

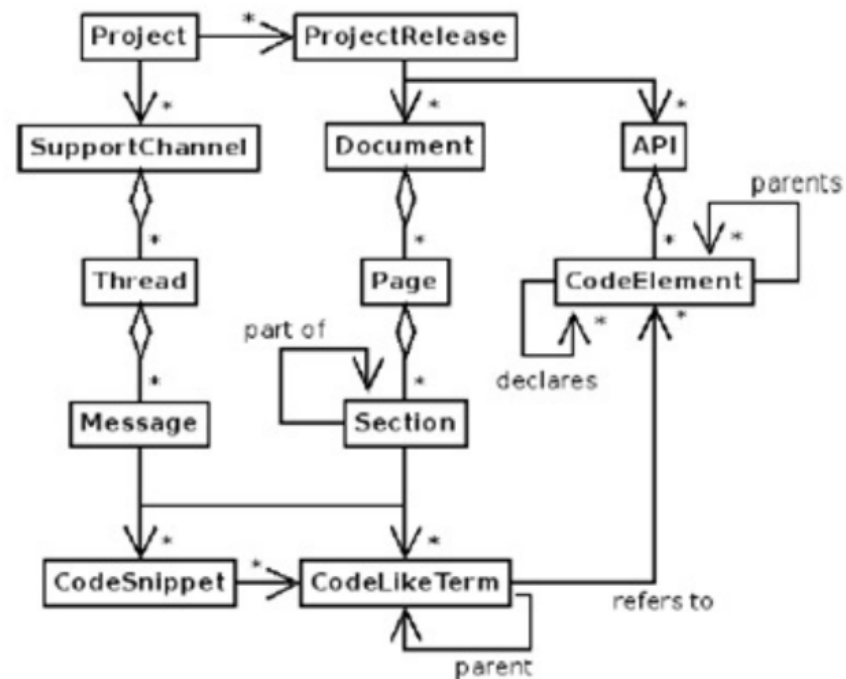


Figure 2. Documentation Meta-Model

e.g. println in SOP to Java.io.PrintStream

Linking Techniques and Challenges

Use of extensible Parsing Infrastructure

e.g. Difference in HTML o/p of Docbook (DocBookParser) + Maven Tool (MavenParser) = DocumentationParser

Further parse the snippets and attempt to link code like terms to code elements

Challenges in linking

Declaration Ambiguity e.g. declaration of methods without their type and package

Overload Ambiguity e.g. overloaded methods

External Reference Ambiguity e.g. reference to external libraries like Java Standard Library or junit

Language Ambiguity e.g. type - HttpClient case - basiclineparser forgetting parameters in call

How to remove ambiguities?



Use of extensible Parsing Infrastructure

e.g. Difference in HTML o/p of
Docbook (DocBookParser) + Maven Tool
(MavenParser) = DocumentationParser

Further parse the snippets and attempt to link code like terms to code elements



Challenges in linking

Declaration Ambiguity e.g. declaration of methods
without their type and package

Overload Ambiguity e.g. overloaded methods

External Reference Ambiguity e.g. reference to external libraries
like Java Standard Library or junit

Language Ambiguity e.g. typo - HttpClient
case - basiclineparser
forgetting parameters in call

How to remove ambiguities?

How to remove ambiguities?

Link Recovery Process

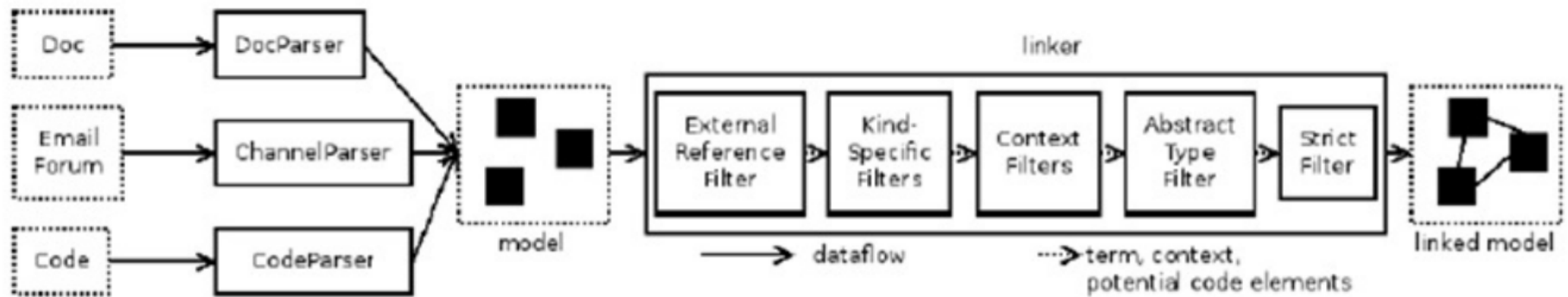


Figure 3. Parsing Artifacts and Recovering Traceability Links

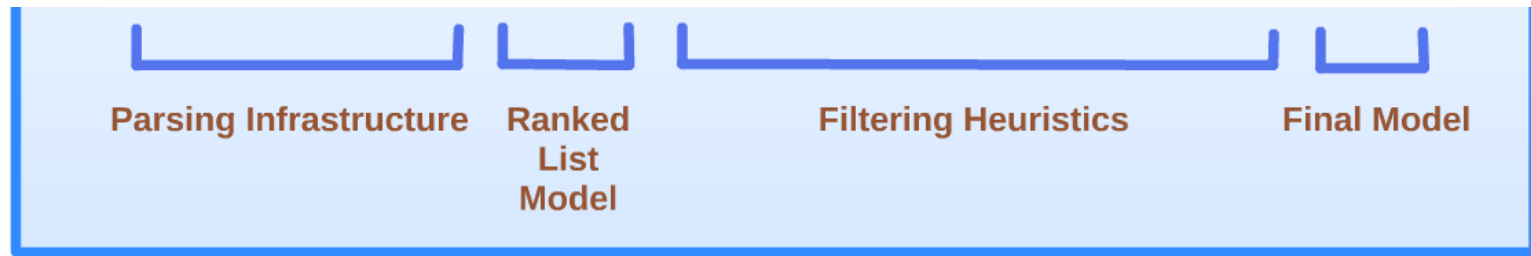
Parsing Infrastructure

Ranked List Model

Filtering Heuristics

Final Model

Link Recovery Process



- ***What is Link Recovery Process***

- Input → code-like-terms
- Associated with → kind (method, field, class)
- Output → ranked list of code elements

- ***Procedure Contains***

- Link to types
- Disambiguate types
- Pass through Filters
- Link misclassified terms

Filtering Heuristics

Filtering is needed if term is referred to more than one code element.

Possible Filters Used;

External reference filter

e.g. HttpClient is system
as well as ~~type~~

Kind specific filter

e.g. Parameter Type mismatch
Parameter Number

Context filter

e.g. toMutableDateTime()
~~Instant.toMutableDateTime~~
~~AbstractInstant.toMutableDateTime~~
ReadableDateTime.toMutableDateTime

Abstract type filter

e.g. More decedents (more abstract)
~~Less decedents (less abstract)~~

Strict filter

e.g. select first
reject ~~rest~~

Evaluation Methodology

Independent variables (What was compared?)

Manual Inspection
vs
RecoDoc Readings



- Exact match
- Similar match
- False -ve
- False +ve

Subjects (What was the data?)

- JodaTime
- HttpComponents
- Hibernate
- Xstream

Dependent variables (What are the units?)

- Precision
- Recall

Table III
RESULTS OF LINK RECOVERY EVALUATION

System	Inspection	RecoDoc	Prec.	Recall
Joda Doc.	807	763 (772)	96.2%	94.5%
Joda Chan.	291	279 (283)	96.5%	95.9%
HC. Doc.	1288	1272 (1273)	98.7%	98.8%
HC. Chan.	266	257 (260)	95.2%	96.6%
Hib. Doc.	361	349 (349)	89.7%	96.7%
Hib. Chan.	265	247 (247)	93.9%	93.2%
XSt. Doc.	175	170 (170)	95.5%	97.1%
XSt. Cha.	267	244 (255)	92.4%	91.4%
Total	3720	3581 (3609)	95.9%	96.3%

Threshold for Prec. and Recall
was set to 90%

Evaluation Results

- ***We considered RecoDoc as wrong when mismatch occurred.***
- ***Evidence of hardness***
 - In 4 systems, out of 300228 code-like-terms, 160970 were methods, those could be linked to average of 16.8 types
- ***Evidence of effectiveness***
 - After filtering, the value 16.8 reduced to 0.7
- ***Filtered out 971577 code-like-terms***

Related Work & Conclusion

Techniques proposed using NLP and Information Extraction

- **Antonoil et al.** applied VSM and probabilistic model to find pages in reference manual.
- **Hipikat** a tool to generate project memory from bug reports, commits, documents etc
- **Xfinder** a tool that matches steps of a tutorial to the code elements that implements each step
- **Dekel and Herbsleb** devised eMoose tool that highlights the method call in the editor when it is called

Conclusion

- Proposed a technique to link code-like-terms to documentation
- Took care of ambiguity and filtering
- In case study, Precision and Recall were 96%
- Bacchelli's comparison found Regex are effective than IR

Techniques proposed using NLP and Information Extraction

- ***Antonoil et al.*** applied VSM and probabilistic model to find pages in reference manual.
- ***Hipikat*** a tool to generate project memory from bug reports, commits, documents etc
- ***Xfinder*** a tool that matches steps of a tutorial to the code elements that implements each step
- ***Dekel and Herbsleb*** devised eMoose tool that highlights the method call in the editor when it is called

Conclusion

- Proposed a technique to link code-like-terms to documentation
- Took care of ambiguity and filtering
- In case study, Precision and Recall were 96%
- Bacchelli's comparison found Regex are effective than IR

Any Questions?



Where can I find **RecoDoc**?

Who is the **Author**?

In which **language** RecoDoc is written?

Press Enter to search.

Python

Barthelemy Dagenais

<http://www.cs.mcgill.ca/~swevo/recodoc/#Cont>

<https://github.com/bartdag/recodoc2>

