

# 代码搜索技术的研究和实践

### 张洪宇 (Hongyu Zhang)

微软亚洲研究院 (Microsoft Research, Asia)

2016.7

## Programming is hard...

- Programming is largely a manual effort, which is tedious and time consuming.
- Developers often wonder what code to write in order to complete a programming task
  - Unfamiliar problems
  - Unfamiliar APIs
  - ...



• In a survey conducted at Microsoft, 67.6% respondents mentioned that they encountered problems in learning APIs.

(M.P. Robillard, "What makes APIs hard to learn? Answers from developers," IEEE Softw., vol. 26, no. 6, pp. 27–34, 2009)

• Vision: Improving Developer Productivity by Code Search

### Searching for API Usage

Question: how to reuse an API? Our work: Recommending sample code about the API usage

- Extract API usage information from a software repository
- Support C/C++/C#/Java/JavaScript languages.
- Support native and third-party APIs.
- Support a variety of data sources (including GitHub, MSDN, and Microsoft initiatives like Office APIs, Azure IOT, Universal Windows Platform, etc.)



Hongyu Zhang, Anuj Jain, Gaurav Khandelwal, Chandrashekhar Kaushik, Scott Ge and Wenxiang Hu, Bing Developer Assistant – Improving Developer Productivity by Recommending Sample Code, To appear: *FSE 2016*, industry track, Seattle, WA, USA, November 2016.

#### Static analysis of a program:

- Parse the source code and header files into an AST
- Performs an in-order traversal of the AST
- Process each node according to its type
- Challenges:
  - Third-party APIs
  - Type inference for weakly typed languages



An Example of an AST

Algorithm 1 Analysis of C/C++ files
1: $S \leftarrow One \ Source \ File$
2: $H \leftarrow Head \ files$
3: $AST \leftarrow \emptyset$
4:
5: function MAIN
6: $Sourcefile \leftarrow S$
7: $Headfiles \leftarrow H$
8: $AST \leftarrow PARSER(Sourcefile, Headfiles)$
9: for $Node \in AST$ root's Children do
10: $VISIT(Node)$
11: end for
12: end function
13:
14: function $VISIT(Node)$
15: <b>if</b> $Node \in S$ <b>then</b>
16: switch $Node$ do
17: $\mathbf{case} \ FunctionCall$
18: $PROCESS\_FUNCALL(Node)$
19: <b>case</b> VariableDeclaration
20: $PROCESS_VARDECL(Node)$
21: <b>case</b> Otherkindsofnodes
22:
23: for $Child \in Node's \ Children \ do$
24: $VISIT(Child)$
25: end for
26: end if
27: end function
28:
29: function $PROCESS_FUNCALL(Node)$
30: $Callee \leftarrow \text{EXTRACT\_CALLEE}(Node)$
31: $Decl \leftarrow Find\_Declaration(Callee)$
32: if $Decl \neq \emptyset$ then
33: $File = WHERE\_IS\_FROM(Decl)$
34: <b>if</b> $File \neq S$ <b>then</b>
35: $LibName = WHICH\_LIBRARY(File)$
36: $LocationInfo = GET\_LOCATIONINFO(Node)$
$37: SAVE\_DATA(Callee, LocationInfo, LibName)$
38: end if
39: end if
40: end function

### Extracted API Usage Data

- We construct a large-scale codebase by crawling projects from MSDN and GitHub: 65,253 projects, 437 GB, 3.5 million source code files.
- In total, 78 million code snippets are obtained.
- The API usage data and sample code are stored in and managed by Microsoft Azure Table.

Language	#Projects	#Files	<b>#Unique APIs</b>	#Code Snippets
C/C++	12,162	730,084	2,980,543	43,944,234
C#	26,322	907,632	698,651	12,336,251
JavaScript	17,115	453,449	21,280	4,205,998
Java	9,654	1,398,695	68,958	17,679,077
Total	65,253	3,489,860	3,769,432	78,165,560

The statistics of the obtained API usage data

## Bing Developer Assistant (BDA) <a href="http://aka.ms/devassistant">http://aka.ms/devassistant</a>

- BDAWPFApplication Microsoft Visual Studio View Project Build Debug Edit Team Jools Architecture Test Analyze Window Help Hussenson Code Samples MainWindow.xaml MainWindowxaml.cs\* 😐 readme.bt BDAWPFApplication BDAWPFApplication.MainWindow - @ MainWindow() . ÷ 16 5p - namespace BDAWPFApplication 17 18 19 1 /// <summary> 20 /// Interaction logic for MainWindow.xaml 21 /// </summary> 2 references 22 public partial class MainWindow : Window 23 **O** references 24 public MainWindow() 25 26 InitializeComponent(); 27 var doc = new HtmlDocument(); 28 doc.1 29 How do I ... (Ctrl+Shift+F1) void HtmlDocument.Load(string path) (+ 10 overloads) Loads an HTML document from a file. & CheckSum 30 CreateAttribute (A) Code Sample (Powered by bong) Search More 31 CreateComment 32 https://github.com/JabbR/JabbR/tree/dev// G CreateElement 4 9 00 Was this helpful? labbR/ContentProviders/ G CreateNavigator DictionaryContentProvider.cs @ CreateTextNode DeclaredEncoding private Task<PageInfo> ExtractFromResponse(ContentProviderHttpRequest request) DetectEncoding return Http.GetAsync(request.RequestUri).Then(response => DetectEncodingAndLoad DetectEncodingHtml var pageInfo = new PageInfo(); € DocumentNode using (var responseStream = response.GetResponseStream()) € Encoding @ Equals var htmlDocument = new HtmlDocument(); htmlDocument.Load(responseStream); GetElementbyld GetHashCode var title = htmlDocument.DocumentNode.SelectSingleNode("//meta[@property='og:title']"); GetType var imageURL = htmlDocument.DocumentNode.SelectSingleNode("//meta ( Load [@property='og:image']'): O LoadHtml pageInfo.Title = title != null ? title.Attributes["content"].Value : String.Empty; 129 % pageInfo.ImageURL = imageURL != null ? imageURL.Attributes["content"].Value : OptionAddDebuggingAttributes Error List Output Find Symbol Results Package @ OptionAutoCloseOnEnd String.Empty; 10 Ch 18 leady Ln 28 Col 18 INS Publisi
- Client side: an extension of Microsoft Visual Studio.
- Server side: Microsoft Azure servers located around the world.
- The APIs provided by IntelliSense of VS trigger the backend BDA service.
- The returned sample code is displayed within Visual Studio.

# Technology Transfer

- Transferred to Bing Developer Assistant (BDA) in 2015.
- Status Updates (as of July 2016):
  - Received more than 450K downloads
  - ~2.1 million queries per month
  - Build 2016 presence
  - New release of BDA on July 13, 2016 (with C/C++ support)

Available at: http://aka.ms/devassistant





Developer Assistant nov 🗙 田公 blogs.msdn.microsoft.com/visualstudio/2016/07/13/developer-assistant-supports-cpp Application Executive Bloggers Visual Studio Languages Platform Lifecycle Development Managemen The Visual Studio Blog **Visual Studio** The official source of product insight from the Visual Studio Engineering Team Download Visual Studio (→) Developer Assistant now supports C++ \*\*\*\*\* July 13, 2016 by Visual Studio Blog # 0 Comments



Today we are happy to announce a major update to Developer Assistant! Developer Assistant now offers contextually aware web powered solutions for C++.

Developer Assistant for Visual Studio is a productivity plugin that brings the combined power of Bing search capabilities and your development environment to solve your day-to-day developer problems. With the addition of C++, we are opening up new possibilities for millions of C++ developers on Visual Studio.

We are now expanding all the 3 important components of Developer Assistant to C++

- Code Samples access in your IDE environment
- · Project search from popular sources
- Bing powered contextual search

Visual Studio Blog

Visual C++ Team Blog









#### **Related Resources**

Visual Studio Product Website Visual Studio Developer Center

#### Getting Started Resources

Write, Navigate, Fix your Code Debug, Profile, Diagnose your Code

## Searching for Reusable Code

We propose **CodeHow**, a method for searching reusable code based on free-form queries

- Given a user query, CodeHow searches codebases and returns the relevant code snippets that match the query.
- Consider both text similarity and program semantics in code retrieval
- Build the tool on top of ElasticSearch and Microsoft Azure
- Indexed ~50K C#/Java/C++/VB projects collected from GitHub and Codeplex



Fei Lv, Hongyu Zhang, Jian-guang Lou, Shaowei Wang, Dongmei Zhang, and Jianjun Zhao, "<u>CodeHow: Effective Code Search based on API</u> <u>Understanding and Extended Boolean Model</u>", in *Proc. ASE 2015*, Lincoln, Nebraska, Nov 2015.

# Natural Language Query - Query Formulation

- Text Similarity Query *qtext* 
  - For retrieving code snippets that match the query in terms of text similarity
- API Query  $q_{api_i}$ 
  - Obtain the top k APIs that are potentially relevant to the query
  - For retrieving code snippets that contain the APIs

• Expanded Query: incorporate both API similarity and text similarity

$$q_{expand} = (q_{api_1}, q_{api_2}, \dots, q_{api_k}, q_{text})$$

### Natural Language Query - Retrieval

• Retrieve code snippets based on the similarity between the expanded query *q*<sub>expand</sub> and a code snippet *d*:

$$sim(q_{expand}, d) = \sum_{i=1}^{k} sim(q_{api_i}, d) + sim(q_{text}, d)$$

The similarity value is computed using Extended Boolean model:

$$sim(q_{or},d) = \left(\frac{w_{t_1,q}^p w_{t_1,d}^p + w_{t_2,q}^p w_{t_2,d}^p + \dots + w_{t_n,q}^p w_{t_n,d}^p}{w_{t_1,q}^p + w_{t_2,q}^p + \dots + w_{t_n,q}^p}\right)^{\frac{1}{p}}$$
  
$$sim(q_{and},d) = 1 - \left(\frac{w_{t_1,q}^p (1 - w_{t_1,d})^p + \dots + w_{t_n,q}^p (1 - w_{t_n,d})^p}{w_{t_1,q}^p + w_{t_2,q}^p + \dots + w_{t_n,q}^p}\right)^{\frac{1}{p}}$$

Fei Lv, Hongyu Zhang, Jian-guang Lou, Shaowei Wang, Dongmei Zhang, and Jianjun Zhao, "<u>CodeHow: Effective Code</u> <u>Search based on API Understanding and Extended Boolean Model</u>", *Proc. ASE'15*, Lincoln, USA, Nov 2015.



←

₽.

#### Natural Language Query

#### Query Examples:

red black tree get free space for disk drive parse an xml document compute standard deviation quick sort MD5 hash how to save an image as jpeg how to export to excel

red black tree	Search
Code Search / API Usage	
example: how to save image, more	Enter your query here
About 48091 results (2.037 seconds)	Enter your query here
Source: RedBlackTreeTest.cs	
<pre>//Method Name: GetTestTree //Reparameters: int noofItems</pre>	Returned results
,,,	
	Icode source file project link license
<pre>var redBlackTree = new RedBlackTree<int, string="">(); for (var i = 0; i &lt; noOfItems; i++)</int,></pre>	(code, source file, project link, license
<pre>var redBlackTree = new RedBlackTree<int, string="">(); for (var i = 0; i &lt; noOfItems; i++) {</int,></pre>	(code, source file, project link, license
<pre>var redBlackTree = new RedBlackTree<int, string="">(); for (var i = 0; i &lt; noOfItems; i++) {     redBlackTree.Add(i, i.ToString()); }</int,></pre>	(code, source file, project link, license
<pre>var redBlackTree = new RedBlackTree<int, string="">(); for (var i = 0; i &lt; noOfItems; i++) {     redBlackTree.Add(i, i.ToString()); }</int,></pre>	(code, source file, project link, license
<pre>var redBlackTree = new RedBlackTree<int, string="">(); for (var i = 0; i &lt; noOfItems; i++) {     redBlackTree.Add(i, i.ToString()); } GitHub Link License: GPL</int,></pre>	(code, source file, project link, license RedBlackTreeTest .GetTestTr
<pre>var redBlackTree = new RedBlackTree<int, string="">(); for (var i = 0; i &lt; noOfItems; i++) {     redBlackTree.Add(i, i.ToString()); } GitHub Link License: GPL Source: RedBlackTreeTest.cs</int,></pre>	(code, source file, project link, license RedBlackTreeTest .GetTestTr
<pre>var redBlackTree = new RedBlackTree<int, string="">(); for (var i = 0; i &lt; noOfItems; i++) {     redBlackTree.Add(i, i.ToString()); } SitHub Link License: GPL Source: RedBlackTreeTest.cs //Method Name: GetTestTree</int,></pre>	(code, source file, project link, license RedBlackTreeTest.GetTestTr
<pre>var redBlackTree = new RedBlackTree<int, string="">(); for (var i = 0; i &lt; noOfItems; i++) {     redBlackTree.Add(i, i.ToString()); } BitHub Link License: GPL Source: RedBlackTreeTest.cs //Method Name: GetTestTree var redBlackTree = new RedBlackTree<int, string="">();</int,></int,></pre>	(code, source file, project link, license RedBlackTreeTest .GetTestTr
<pre>var redBlackTree = new RedBlackTree<int, string="">(); for (var i = 0; i &lt; noOfItems; i++) {     redBlackTree.Add(i, i.ToString()); } GitHub Link License: GPL Source: RedBlackTreeTest.cs //Method Name: GetTestTree var redBlackTree = new RedBlackTree<int, string="">(); for (var i = 0; i &lt; 100; i++) </int,></int,></pre>	(code, source file, project link, license RedBlackTreeTest .GetTestTr
<pre>var redBlackTree = new RedBlackTree<int, string="">(); for (var i = 0; i &lt; noOfItems; i++) {     redBlackTree.Add(i, i.ToString()); } BitHub Link License: GPL Source: RedBlackTreeTest.cs //Method Name: GetTestTree var redBlackTree = new RedBlackTree<int, string="">(); for (var i = 0; i &lt; 100; i++) {     redBlackTree.Add(i, i.ToString()); } </int,></int,></pre>	(code, source file, project link, license RedBlackTreeTest .GetTestTr

GitHub Link License: GPL

RedBlackTreeTest .GetTestTree

Source: RedBlackTreeTest.cs

#### Searching for API Usage

- Question: how to use an API? ٠
- Our work: Recommending sample code about the API usage
  - Extract API usage information from a ٠ software repository
  - Support C/C++/C#/Java/JavaScript ٠ languages.
  - Support native and third-party APIs. ٠
  - Support a variety of data sources (including GitHub, MSDN, and Microsoft initiatives like Office APIs, Azure IOT, Universal Windows Platform, etc.)

		£	
A	PI Usage Dat	ta (Azure Table)	-
û,	Û.	0	0
unction	Class	Structure	-
Û	Û	Û	Û
	API Usage	e Extraction	
	1	}	
interr	mediate Repr	resentation (ASI	(2)
û.	Û	<u>.</u>	Û
lang	Roslyn	TOL	19
۵. I		Ŷ	企
Softwa	re Repositor	v 10/0++/0#/Jav	1 1

#### Searching for Reusable Code

We propose CodeHow, a method for searching reusable code based on free-form queries

Local Projects

- Given a user query, CodeHow searches codebases and returns the relevant code snippets that match the query.
- Consider both text similarity and program semantics in code retrieval
- Build the tool on top of ElasticSearch and Microsoft Azure
- Indexed ~50K C#/Java/C++/VB projects collected from GitHub and Codeplex

How to use SalConnection.Open Ditien Source Picket) How to use SSL.accept How to compute MD5 Hash How to convert a string to int



Fei Ly, Hongyu Zhang, Jian-guang Lou, Shaowei Wang, Dongmei Zhang, and Jianjun Zhao, "CodeHow: Effective: Code Search based on API Understanding and Extended Boolean Mode?", in Proc. ASE 2015, Lincoln, Nebraska, Nov 2015.

#### Technology Transfer

- Transferred to Bing Developer ٠ Assistant (BDA)
- Status Updates (as of July 2016): ٠
  - Received more than 450K downloads
  - ~2.1 million gueries per month
- Build 2016 presence
- New release of BDA on July 13, 2016

#### Available at: http://aka.ms/devassistant

Collaborators: **Bing Tech Segment** > bing Hyderabad



The Visual Studio Blog

and it has a reading that it is not the

11 diversity of himself and

the procession of the 1 Kip back

1. A deal sector in a section.

- And Intelligence in such that the

A REAL PROPERTY.

Developer Assistant now supports C+ I

the same law to be which descent many and the shift of the same sector and the same to same the

https://blogs.msdn.microsoft.com/visualstudio/2016/07/13/developer-assistant-supports-cpc/

E france familie 7 B - 6 ------

4



00 = 2

Visual Studio

Distant West Dates (

a 🛙

Sainted Townson,

Getting Marked Baserieses

the later

#### 0 a Natural Language CodeHow Query THE MARK THE Code Selects ( Mr. Marco stample. Now hit has a larger but Query Examples: Enter your query here Alard \$2271 reads (2027 interest red block tree Suce Addies Inclusion get free space for disk drive daried time torcestill parse an xml document Returned results that similarly into the William compute standard deviation (code, source file, project link, license) or restigition a see instighting let, through for the links incontribute line quick sort MD5 hash Patron Man, April, 1, 1941 (1991) how to save an image as jpeg how to export to excel Simplifi Lipper GPL Building's Tree famil Garfree Tree Inches Providing Constant in Sector and Sectories restanting - --- tolligentres.int. f-tapit() New Color 2 is No. 2 is same local relation with Links and Denier Low Lineman GPL Rettlied Tree liest Gerfleet Tree Same Institut Install



# Thanks!

Hongyu Zhang Microsoft Research No.5 Danling Street Beijing 100080, China Email: honzhang@microsoft.com URL: http://hongyujohn.github.io/