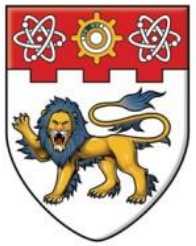


NANYANG TECHNOLOGICAL UNIVERSITY

Mining Technology Landscape from Stack Overflow

Chunyang Chen, Zhenchang Xing

School of Computer Science and Engineering, Nanyang Technological University, Singapore

Motivation

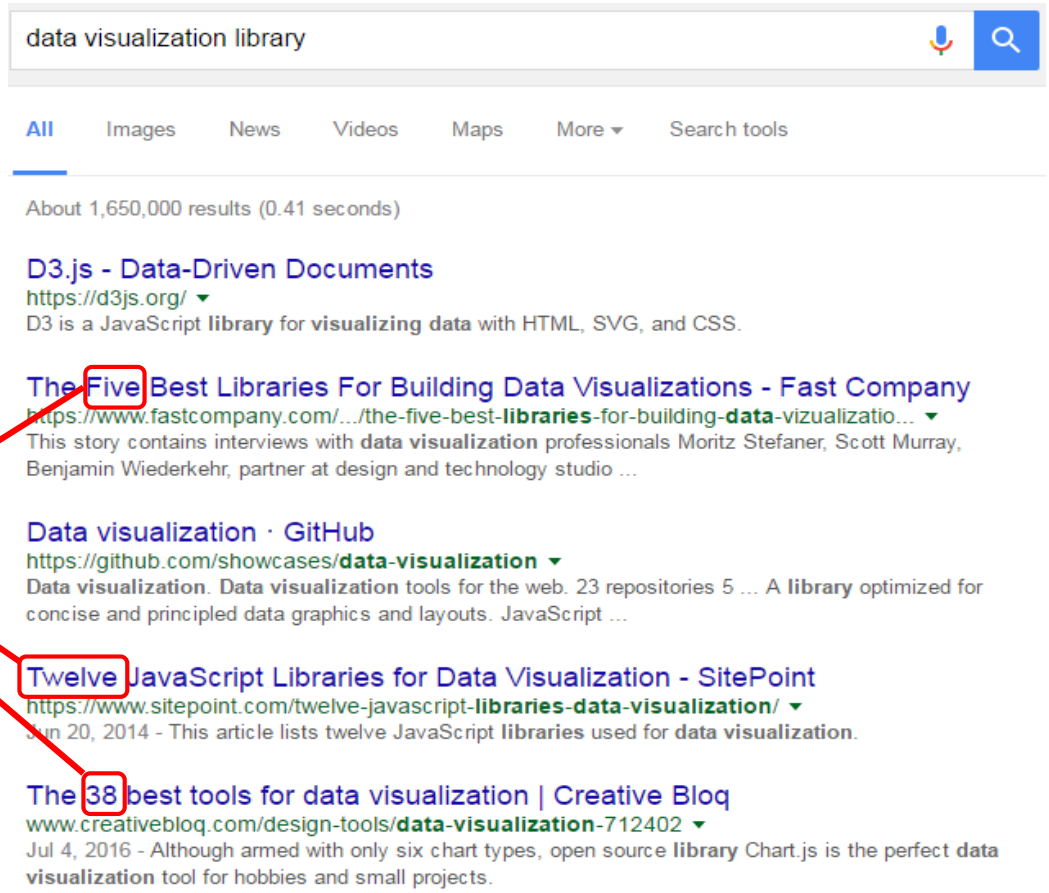
Given a task, most developers need an overview of related technology to understand:

- The scope and challenge of this task;
- The existing method or algorithm for such task;
- The framework, library or related code to help solve the task
- ……

Motivation

How to find technology landscape?

1. Search Google for technology overview



Easy to be lost, biasd and out of date

Motivation

2. They ask in Q&A site such as Stack Overflow

Data Visualization libraries [closed]

1

4

I am currently in the startup of my new project. It's a data visualisation project, where I want to develop an application that can visualise data (no matter where it comes from).

Right now,I am trying to find a visualisation library that I can use. Which one do you recommend?

For me it looks like the main libraries are in javascript(D3.js). I wanted to develop an desktop application, but maybe I should just face it and switch to web based?

I have experience in java,python and C#.

data-visualization

share improve this question

edited Jan 17 at 17:35
Brian Tompsett - 汤莱恩
3,037 ● 11 ● 26 ● 74

asked Feb 17 '14 at 9:32
miniHessel
184 ● 2 ● 19

closed as off-topic by Karl-Johan Sjögren, max taldykin, Paul Collingwood, Doorknob, Andy Feb 28 '14 at 3:10

This question appears to be off-topic. The users who voted to close gave this specific reason:

- "Questions asking us to **recommend or find a tool, library or favorite off-site resource** are off-topic for Stack Overflow as they tend to attract opinionated answers and spam. Instead, describe the problem and what has been done so far to solve it." – Karl-Johan Sjögren, max taldykin, Paul Collingwood, Doorknob, Andy

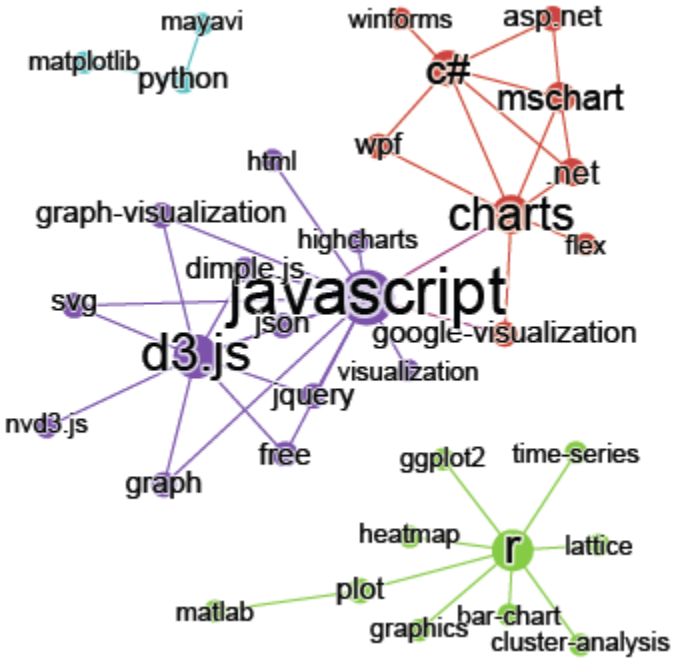
If this question can be reworded to fit the rules in the help center, please edit the question.

Not allowed

How to solve it

- How can we help such information needs automatically and objectively?

  a graph for technology landscape!

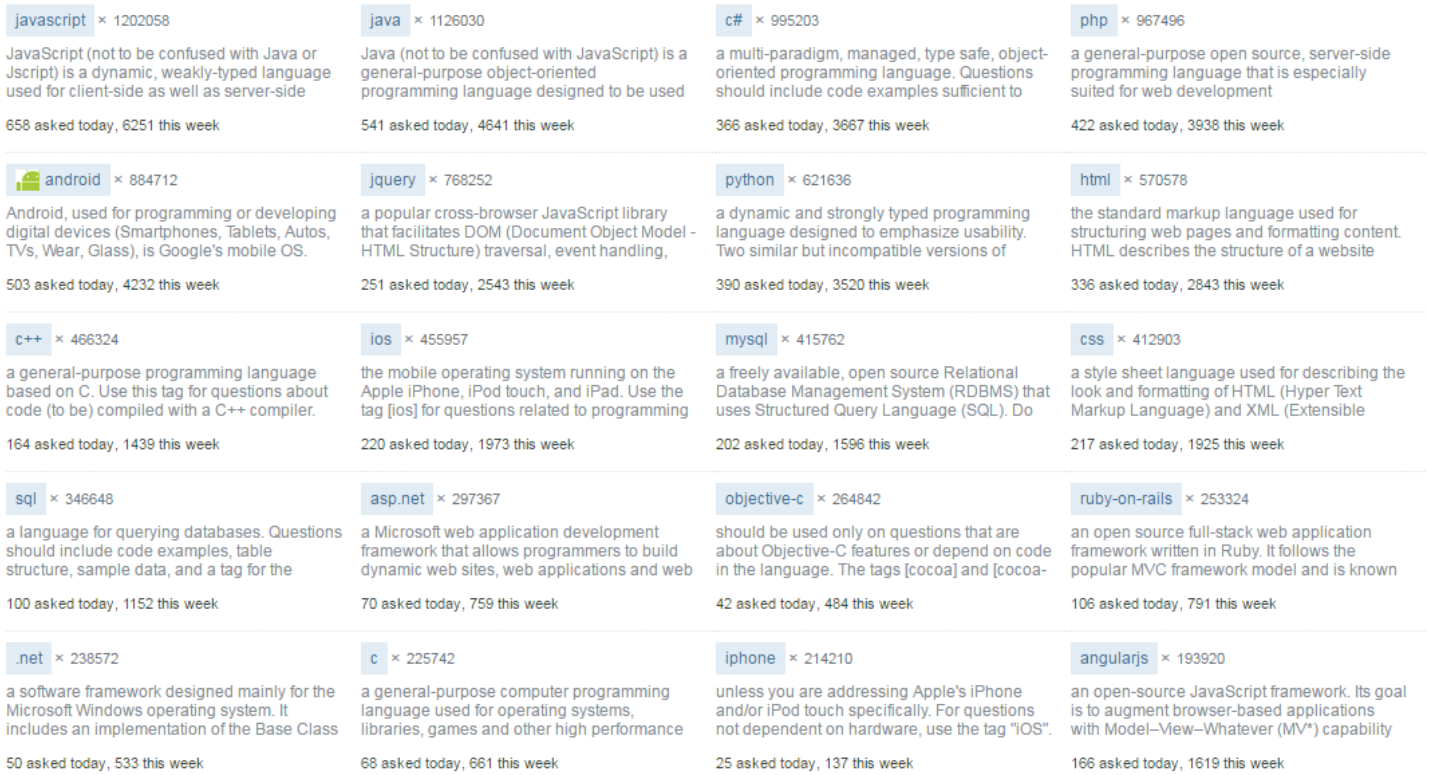


Library	Language	Concept
d3.js	javascript	svg
ggplot2	r	json
matplotlib	c#	graph
highcharts	python	visualization
nvd3.js	matlab	time-series
dimple.js	html	heatmap
flex		plot
mschart		graphics
lattice		bar-chart
mayavi		free

Observation

Stack Overflow covers most programming technologies

- 12m questions, 20m answers, 6m users
- Mine a landscape of technologies automatically from SO.



javascript × 1202058
JavaScript (not to be confused with Java or Jscript) is a dynamic, weakly-typed language used for client-side as well as server-side
658 asked today, 6251 this week

java × 1126030
Java (not to be confused with JavaScript) is a general-purpose object-oriented programming language designed to be used
541 asked today, 4641 this week

c# × 995203
a multi-paradigm, managed, type safe, object-oriented programming language. Questions should include code examples sufficient to
366 asked today, 3667 this week

php × 967496
a general-purpose open source, server-side programming language that is especially suited for web development
422 asked today, 3938 this week

android × 884712
Android, used for programming or developing digital devices (Smartphones, Tablets, Autos, TVs, Wear, Glass), is Google's mobile OS.
503 asked today, 4232 this week

jquery × 768252
a popular cross-browser JavaScript library that facilitates DOM (Document Object Model - HTML Structure) traversal, event handling,
251 asked today, 2543 this week

python × 621636
a dynamic and strongly typed programming language designed to emphasize usability. Two similar but incompatible versions of
390 asked today, 3520 this week

html × 570578
the standard markup language used for structuring web pages and formatting content. HTML describes the structure of a website
336 asked today, 2843 this week

c++ × 466324
a general-purpose programming language based on C. Use this tag for questions about code (to be) compiled with a C++ compiler.
164 asked today, 1439 this week

ios × 455957
the mobile operating system running on the Apple iPhone, iPod touch, and iPad. Use the tag [ios] for questions related to programming
220 asked today, 1973 this week

mysql × 415762
a freely available, open source Relational Database Management System (RDBMS) that uses Structured Query Language (SQL). Do
202 asked today, 1596 this week

css × 412903
a style sheet language used for describing the look and formatting of HTML (Hyper Text Markup Language) and XML (Extensible
217 asked today, 1925 this week

sql × 346648
a language for querying databases. Questions should include code examples, table structure, sample data, and a tag for the
100 asked today, 1152 this week

asp.net × 297367
a Microsoft web application development framework that allows programmers to build dynamic web sites, web applications and web
70 asked today, 759 this week

objective-c × 264842
should be used only on questions that are about Objective-C features or depend on code in the language. The tags [cocoa] and [cocoa-
42 asked today, 484 this week

ruby-on-rails × 253324
an open source full-stack web application framework written in Ruby. It follows the popular MVC framework model and is known
106 asked today, 791 this week

.net × 238572
a software framework designed mainly for the Microsoft Windows operating system. It includes an implementation of the Base Class
50 asked today, 533 this week

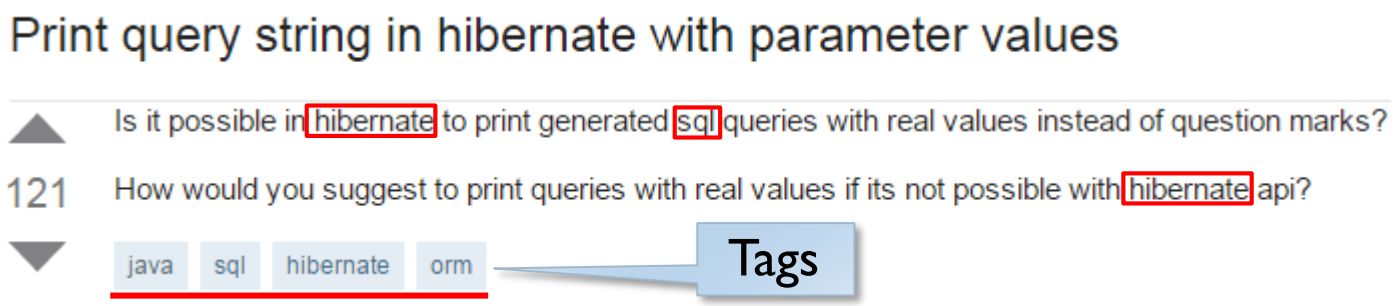
c × 225742
a general-purpose computer programming language used for operating systems, libraries, games and other high performance
68 asked today, 661 this week

iphone × 214210
unless you are addressing Apple's iPhone and/or iPod touch specifically. For questions not dependent on hardware, use the tag "iOS".
25 asked today, 137 this week

angularjs × 193920
an open-source JavaScript framework. Its goal is to augment browser-based applications with Model–View–Whatever (MV*) capability
166 asked today, 1619 this week

Dataset

- Tags in Stack Overflow



  - A tag is a word or phrase that describes the technology or concept of the question.
  - Frequent co-occurring tags indicate potential relations between technologies.
  - There are millions of questions attached with tags.

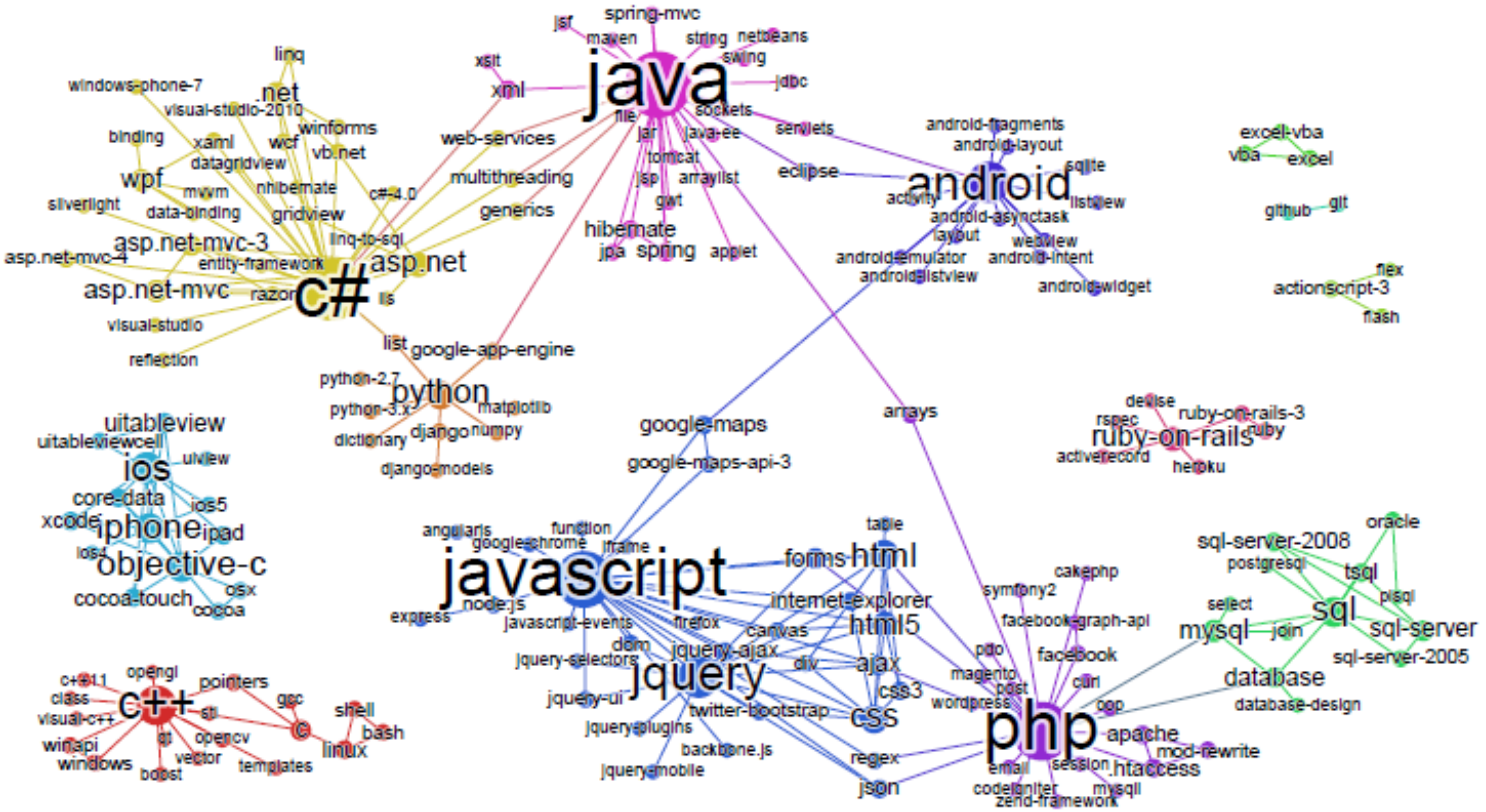
- Hence, we adopt tags in SO to mine technology landscape.

Construct Technology Associative Network

- Association rules mining
  - Minimum support & confidence value
  - Frequent pairs of tags e.g., (java -- spring, c -- pointer, html -- css)
- Build a graph
  - Link tag pairs into a graph, we call it technology associative network (TAN)
- Community detection
  - Cluster nodes in the graph

Technology Associative Network

The overview of important technologies related to programming in Stack Overflow:



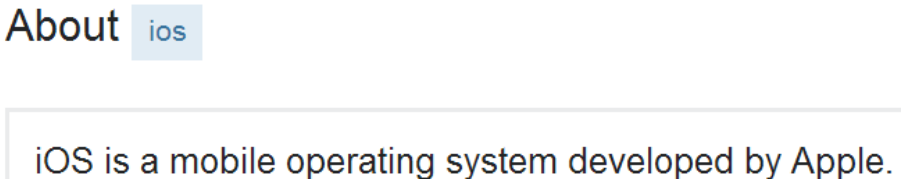
Tag Category

Tags in Stack Overflow belong to different categories:

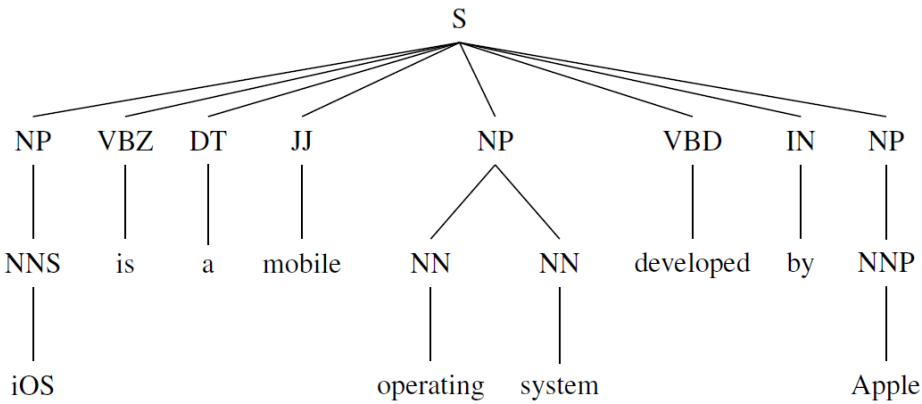
- Java → programming language;
- Eclipse → IDE;
- Binary-search → algorithm;
- Caching → mechanism;
- D3.js → library.

Identify tag category

- Community TagWiki info
  - First sentence

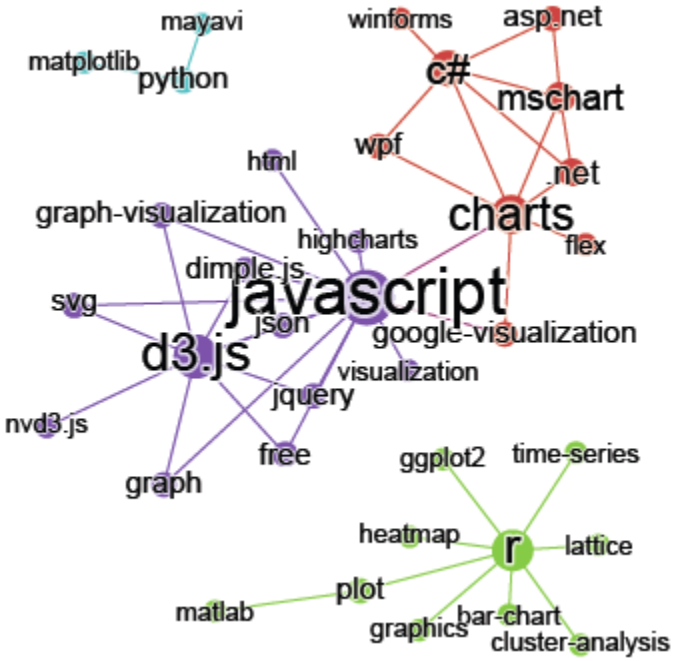


- Get the category of the tag
  - Part-of-Speech tagging and phrase chunking
  - First noun phrase after "be" is category label



Technology Landscape

Combine both the graph and category information to complete the technology landscape:



Library	Language	Concept
d3.js	javascript	svg
ggplot2	r	json
matplotlib	c#	graph
highcharts	python	visualization
nvd3.js	matlab	time-series
dimple.js	html	heatmap
flex		plot
mschart		graphics
lattice		bar-chart
mayavi		free

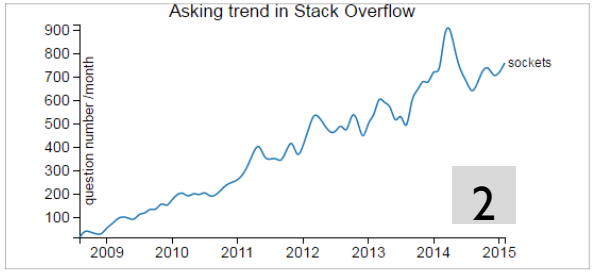
Application

- Website (**TechLand**)
  - https://graphofknowledge.appspot.com

sockets 1

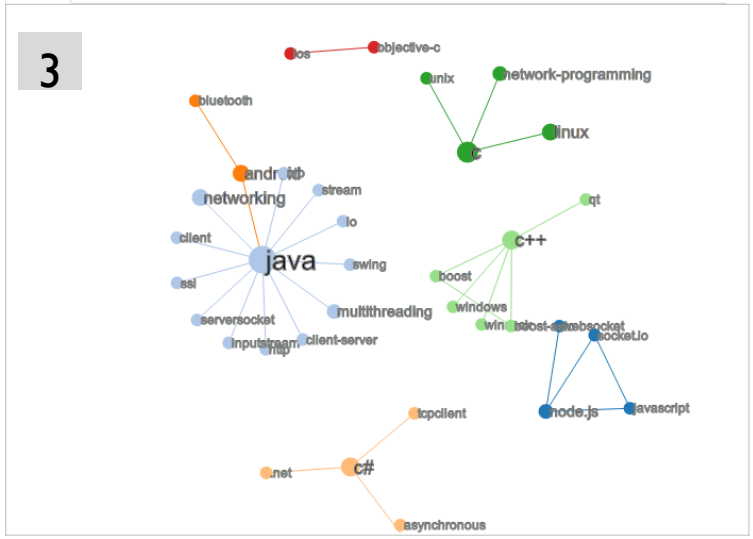
An endpoint of a bidirectional inter-process communication flow. This often refers to a process flow over a network connection, but by no means is limited to such. Not to be confused with websocket (a protocol) or other abstractions (e.g. socket.io).

Learn more »



Total 2009 2010 2011 2012 2013 2014 Like

3



Library

swing
boost
websocket
socket.io
boost-asio
winsock
tcpclient

Top-voted questions: 4

346 What is the difference between a port and a socket?
248 Socket options SO_REUSEADDR and SO_REUSEPORT, how do they differ? Do they mean the same across all major operating systems?
221 What does "connection reset by peer" mean?
142 How much overhead does SSL impose?
114 Can two applications listen to the same port?

Code snippets: 5

```
Java
import java.io.IOException;
import java.net.*;
public class SocketSend {
  public static void main(String args[]) throws IOException {
    sendData("localhost", "hello socket world");
  }

  public static void sendData(String host, String msg) throws IOException {
    Socket sock = new Socket( host, 256 );
    sock.getOutputStream().write(msg.getBytes());
    sock.getOutputStream().flush();
    sock.close();
  }
}
```

Popular links: 6

http://docs.python.org/library/socket.html
http://msdn.microsoft.com/en-us/library/system.net.sockets.socket.aspx
http://developer.android.com/reference/java/net/Socket.html
http://docs.oracle.com/javase/7/docs/api/java/net/Socket.html
http://help.adobe.com/en_US/FlashPlatform/reference/actionscript/3/flash/net/Socket.html

Empirical Study

- RQ1: Can the mined TAN capture the important technologies from a majority of Stack Overflow questions?
- RQ2: How do different mining thresholds affect the size and modularity of the mined TAN?
- RQ3: Are the mined technology associations semantically related?
- RQ4: What are structural properties of the mined TAN?
- RQ5: How do the technology landscape evolve over time?

RQ1: Coverage of Tags and Questions

- If one tag of a question is covered, we count this question as covered
  - The higher minimum support value, the lower coverage

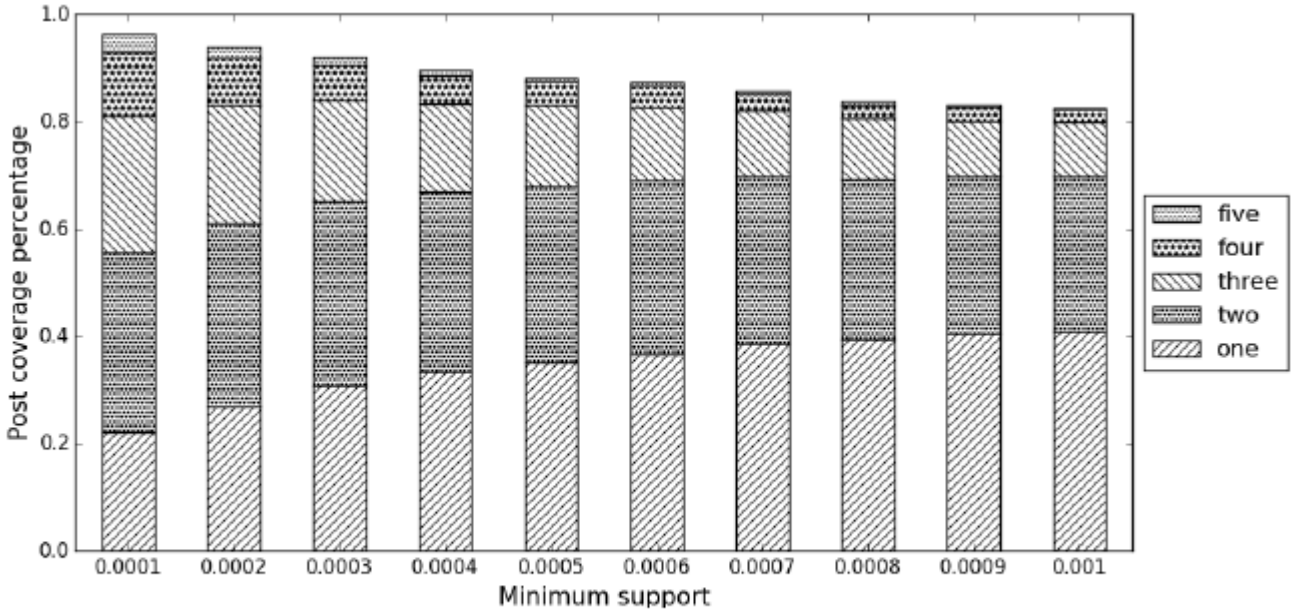
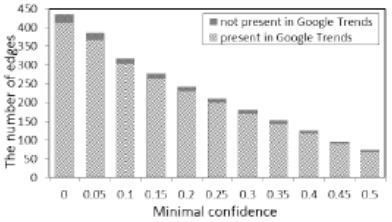


Figure 7: Coverage of questions

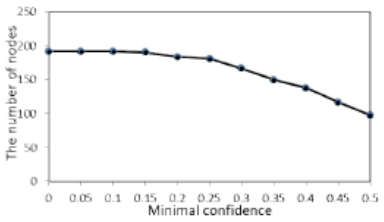
RQ2 & RQ3

- size and modularity of technology community
  - Higher confidence results in sparse graph while lower confidence results in dense graph

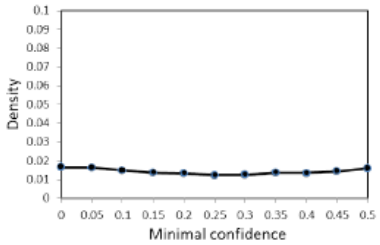
- Semantic distance of technology associations by "Google Distance"
  - As Figure (a) shows, most edges are covered in Google Trends



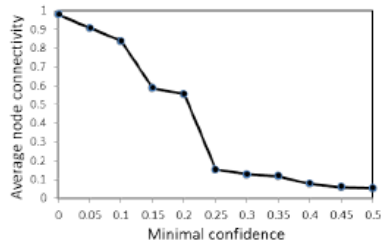
(a) Edge number



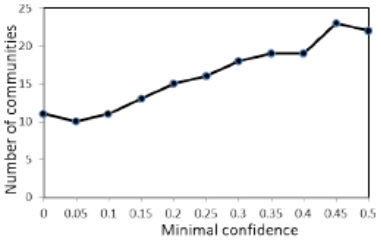
(b) Node number



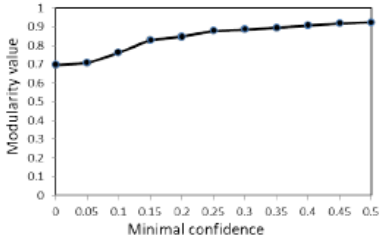
(c) Density



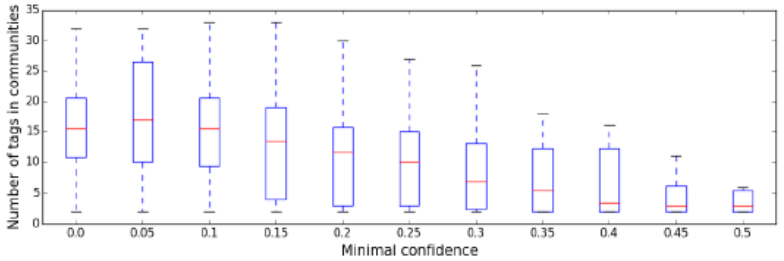
(d) Connectivity



(e) Community number



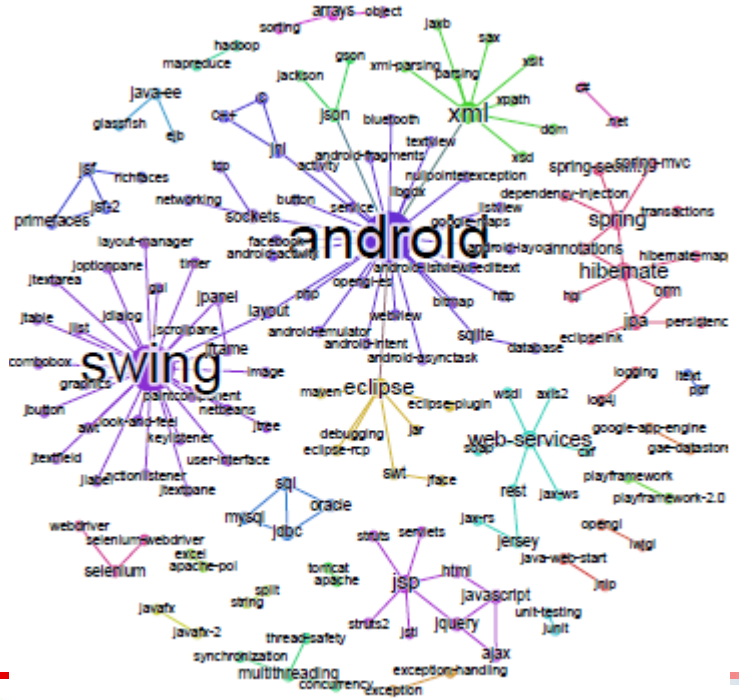
(f) Modularity



(g) Community size distribution

RQ4: network structure

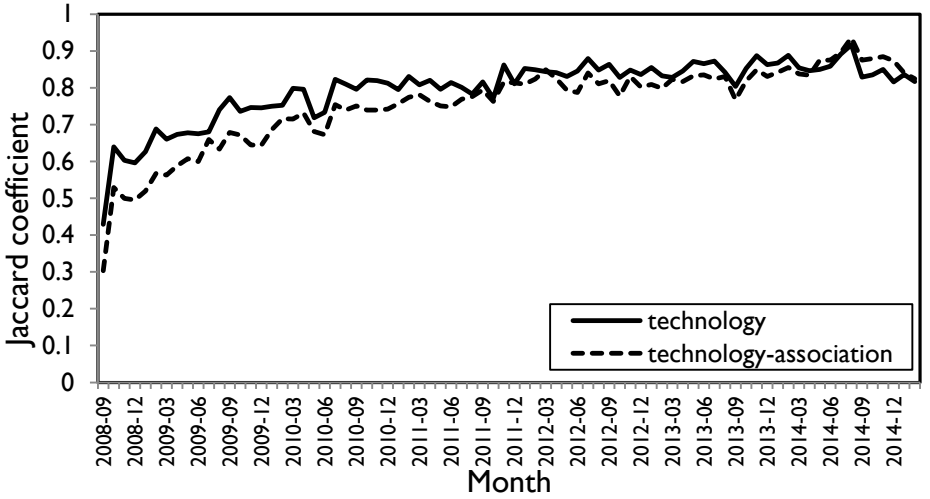
- General & Specific TAN
  - General Technological associative network can capture high-level knowledge.
  - After zooming in, more detailed knowledge emerges

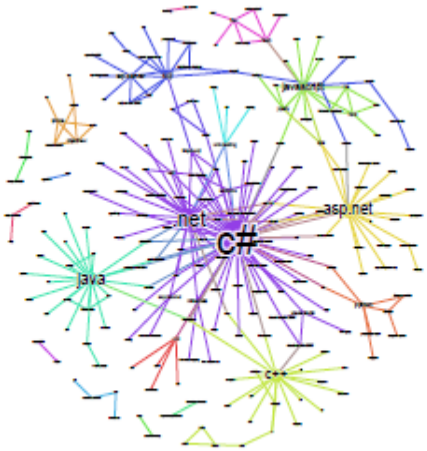


(a) Java (b) Swing

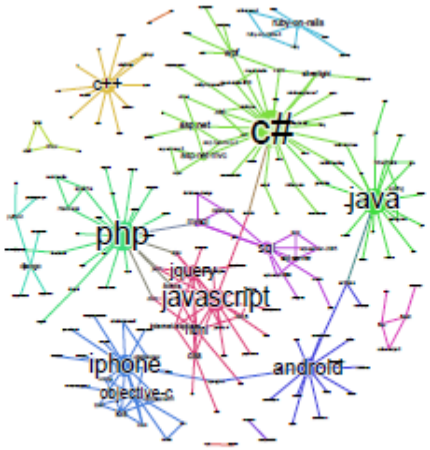
RQ5: Network Evolution

- For each month, draw TAN
  - Their difference is smaller and smaller
  - Different communities begin to emerge and stabilize

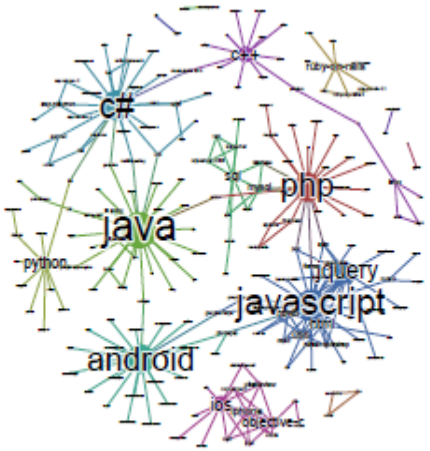




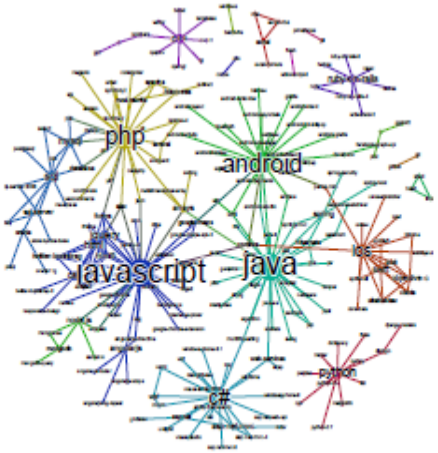
(a) 2009-01



(b) 2011-01



(c) 2013-01



(d) 2015-01

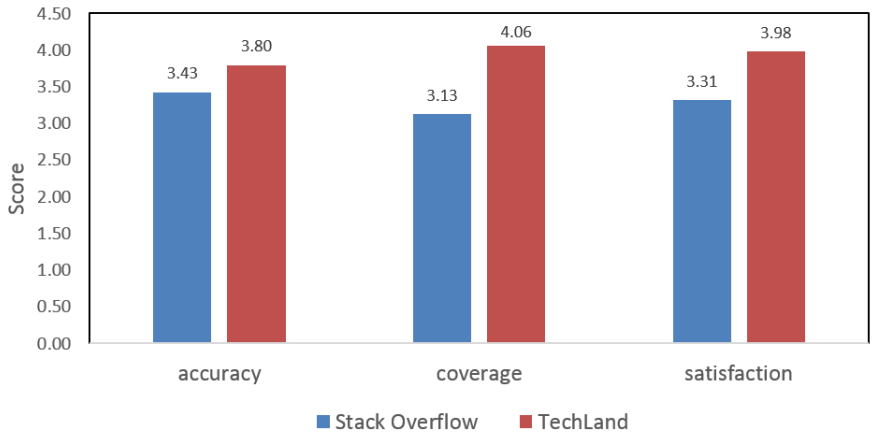
Usefulness of our TechLand

1. Answer questions in Stack Overflow

- 3 types, 9 questions from Stack Overflow;
- 7 PhD students
- Compare original answers and that from our TechLand
- Mark accuracy, coverage, satisfaction by 5-point likert scale

Table 1: 3 types of questions in user study

Type	Question
overview	What are the best overviews for cloud technology
	A good resource for an overview of web technologies
	Technical architecture diagram for an iPhone app
concept	What are the top 3 main concepts in WPF
	What are the core concepts in functional programming
	What are best tools/concepts/things to be a better java programmer
library	What are some good OpenID libraries
	What are your favorite JavaScript libraries/scripts to create tooltips
	What languages and libraries should I use to work with Gmail?



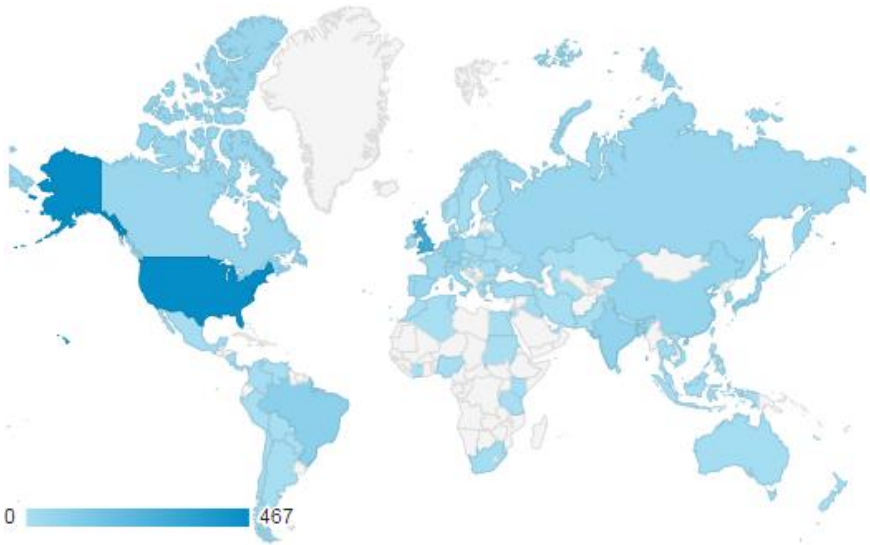
Usefulness of our TechLand

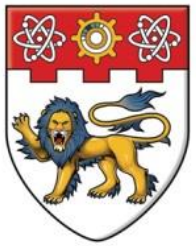
2. Web usage

https://graphofknowledge.appspot.com/

Sessions	Users
3,229	2,220

Page Views	Pages/Session
11,040	3.42





NANYANG TECHNOLOGICAL UNIVERSITY

- Chen, Chunyang, and Zhenchang Xing. "Mining technology landscape from stack overflow." In *Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, p. 14. ACM, 2016.
- Chen, Chunyang, Zhenchang Xing, and Lei Han. "Techland: Assisting technology landscape inquiries with insights from stack overflow." In *2016 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pp. 356-366. IEEE, 2016.

Thanks for listening

Chunyang Chen, Zhenchang Xing

School of Computer Science and Engineering, Nanyang Technological University, Singapore