Mining Technology Landscape from Stack Overflow

Chunyang Chen, Zhenchang Xing
School of Computer Science and Engineering, Nanyang Technological University, Singapore
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- Easy to be lost, biased and out of date
2. They ask in Q&A site such as Stack Overflow

Data Visualization libraries [closed]

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#.

Motivation

Not allowed
How to solve it

- How can we help such information needs automatically and objectively?
  - a graph for technology landscape!

<table>
<thead>
<tr>
<th>Library</th>
<th>Language</th>
<th>Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>d3.js</td>
<td>javascript</td>
<td>svg</td>
</tr>
<tr>
<td>ggplot2</td>
<td>r</td>
<td>json</td>
</tr>
<tr>
<td>matplotlib</td>
<td>c#</td>
<td>graph</td>
</tr>
<tr>
<td>highcharts</td>
<td>python</td>
<td>visualization</td>
</tr>
<tr>
<td>nvd3.js</td>
<td>matlab</td>
<td>time-series</td>
</tr>
<tr>
<td>dimple.js</td>
<td>html</td>
<td>heatmap</td>
</tr>
<tr>
<td>flex</td>
<td></td>
<td>plot</td>
</tr>
<tr>
<td>mschart</td>
<td></td>
<td>graphics</td>
</tr>
<tr>
<td>lattice</td>
<td></td>
<td>bar-chart</td>
</tr>
<tr>
<td>mayavi</td>
<td></td>
<td>free</td>
</tr>
</tbody>
</table>
Observation

Stack Overflow covers most programming technologies

- 12m questions, 20m answers, 6m users
- Mine a landscape of technologies automatically from SO.
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
The overview of important technologies related to programming in Stack Overflow:
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
Combine both the graph and category information to complete the technology landscape:

<table>
<thead>
<tr>
<th>Library</th>
<th>Language</th>
<th>Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>d3.js</td>
<td>javascript</td>
<td>svg</td>
</tr>
<tr>
<td>ggplot2</td>
<td>r</td>
<td>json</td>
</tr>
<tr>
<td>matplotlib</td>
<td>c#</td>
<td>graph</td>
</tr>
<tr>
<td>highcharts</td>
<td>python</td>
<td>visualization</td>
</tr>
<tr>
<td>nvd3.js</td>
<td>matlab</td>
<td>time-series</td>
</tr>
<tr>
<td>dimple.js</td>
<td>html</td>
<td>heatmap</td>
</tr>
<tr>
<td>flex</td>
<td></td>
<td>plot</td>
</tr>
<tr>
<td>mschart</td>
<td></td>
<td>graphics</td>
</tr>
<tr>
<td>lattice</td>
<td></td>
<td>bar-chart</td>
</tr>
<tr>
<td>mayavi</td>
<td></td>
<td>free</td>
</tr>
</tbody>
</table>
Application

Website (TechLand)

https://graphofknowledge.appspot.com

Top-voted questions:
1. What is the difference between a port and a socket?
2. Socket options SO_REUSEADDR and SO_REUSEPORT, how do they differ? Do they mean the same across all major operating systems?
3. What does “connection reset by peer” mean?
4. How much overhead does SSL impose?
5. Can two applications listen to the same port?

Library

Code snippets:

Popular links:

http://docs.python.org/library/socket.html
http://docs.oracle.com/javase/7/docs/api/java/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
RQ4: network structure

- General & Specific TAN
  - General Technological associative network can capture high-level knowledge.
  - After zooming in, more detailed knowledge emerges
RQ5: Network Evolution

- For each month, draw TAN
  - Their difference is smaller and smaller
  - Different communities begin to emerge and stabilize
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>
<thead>
<tr>
<th>Type</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>overview</td>
<td>What are the best overviews for cloud technology</td>
</tr>
<tr>
<td></td>
<td>A good resource for an overview of web technologies</td>
</tr>
<tr>
<td></td>
<td>Technical architecture diagram for an iPhone app</td>
</tr>
<tr>
<td>concept</td>
<td>What are the top 3 main concepts in WPF</td>
</tr>
<tr>
<td></td>
<td>What are the core concepts in functional programming</td>
</tr>
<tr>
<td></td>
<td>What are best tools/concepts/things to be a better java programmer</td>
</tr>
<tr>
<td>library</td>
<td>What are some good OpenID libraries</td>
</tr>
<tr>
<td></td>
<td>What are your favorite JavaScript libraries/scripts to create tooltips</td>
</tr>
<tr>
<td></td>
<td>What languages and libraries should I use to work with Gmail?</td>
</tr>
</tbody>
</table>

![Table showing scores for accuracy, coverage, and satisfaction for Stack Overflow and TechLand](chart.png)
2. Web usage

<table>
<thead>
<tr>
<th>Sessions</th>
<th>Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>3,229</td>
<td>2,220</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Page Views</th>
<th>Pages/Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>11,040</td>
<td>3.42</td>
</tr>
</tbody>
</table>

https://graphofknowledge.appspot.com/

Thanks for listening

Chunyang Chen, Zhenchang Xing
School of Computer Science and Engineering, Nanyang Technological University, Singapore