Visualizing the ecosystem of Computer Science: A network approach through Stack Overflow

Konstantinos Papadopoulos
Department of Computer Science
Boston University
konpap94@bu.edu

Yue Lei
Department of Computer Science
Boston University
tigerlei@bu.edu

Abstract
In this paper we examined and visualized the relationship of technology-tags in the field of Computer Science. Using public data from user posts made on Stack Overflow from 2010 to 2015 we computed and analysed a similarity network for the most popular technologies and their relationship and communities within them. The visualization of our network can be found at http://bit.ly/1TbZAyz

Keywords: Data Mining, Network Analysis, Technology Similarity, Stack Overflow

----------***----------

Introduction

Computer science is a field that is evolving very rapidly. Year to year, new technologies, methodologies, libraries and techniques are being introduced at a rate that may be hard to follow. As a result, professionals, hobbyists and students introduced to the field can be significantly overwhelmed by the constantly changing relationships among those technologies.

Introducing a simpler way of viewing and clustering technologies within the field of computer science would help people get a clearer picture of the application and trends of the technologies people are using. As an example, someone being introduced to Python for the first time might understand the syntax and basic applications, but won't be able to recognize how Python interacts and fits in the grand scheme of technologies.

For the characterization of such technologies we used Stack Overflow as a source of data directly related to the field trends. Stack Overflow is a popular internet platform for people to ask and answer technologically relevant questions. We used their API to gather the tags of each post and construct a similarity network based on the frequency two tags are seen together.

We chose to visualise our results by constructing a node map using the network analysis library *networkX* and the visualization package *Gephi*. The visualization contains each note as the tag and all the vector-nodes as the related technologies.

Methods and Techniques

We constructed a network out of the 1000 most popular[1] tags and examined the characteristics of our network. Each network node reflects a technology tag and each connection represents a directional dependency to other tags. In order to measure the dependency we computed the conditional probability of each tag appearing with another.

$$d_{1,2} = Dependency_{(tag1->tag2)} = P(tag2|tag1)$$

This metric implies that each tag combination has 2 relationships that it can be characterized with. One is the relationship of tag1 to tag2 and, conversely, the relationship of tag2 to tag1. We chose dependency over the jaccard distance of two tags due to hub nodes being connected to so many other nodes, and, as a result, the strength of the connection becomes diluted.

As an example, python's connection to pandas has an index of 0.08 (8% of all python posts are about pandas) but pandas' connection to python is 0.9, meaning that 90% of all pandas posts also include the "python" tag.

$$Dependency_{(python->pandas)} = 0.08$$

$$Dependency_{(pandas->python)} = 0.9$$

In order to consider a connection between two tags as valid, we set a threshold of d(1,2) > 5%. To construct the network plot of the most important technologies we first identified the hub nodes in the network by measuring degree centrality.

Finally, we identified communities within our network model by using the Louvain method[1], a community detection algorithm in which node inter-cluster node modularity is maximized.

Datasets and Experiments

For our dataset, we downloaded the question data from Stack Overflow on 2015 through their given API. This included all the posts made on the website throughout that year. We used the Stack Overflow posts as a sample for the field. We chose Stack Overflow due to its relevance in the developer world. We believe that it's a good indication of the developing trends and technologies. Another reason we chose Stack Overflow is because they provide a good API that we can use and their data has a timestamp that we can adopt to explore and characterize the evolution and the similarity of the post tags.

StackOverflow provides their data in JSON format. Each JSON object provides various information about the attributes of each post. Out of all the attributes, we filtered the once that include information about the creation date of the post and the tags the users submitted when creating each post. Out of the 10GB of initial data, our parsing algorithm filtered the dataset down to 168MB. The total length of the set of tags used in all posts exceeded 30,000 tags. We chose to narrow down our target to the 1,000 most encountered tags.

Results and Discussion

We constructed a visualization that accurately describes the relationship between technology tags based on the frequency two tags are seen together. Our network contained 1000 nodes and 6260 edges. The average shortest path length was 2.6 for unweighted and 6.3 for weighted computations. That means that given 2 random tags as a source and a destination, it would take on average 6 to 7 path lengths to reach the destination given a weighted and directional environment but only 2-3 path lengths given a non weighted, directional environment .

That translates into a network that is tightly connected yet directionally separate. We found the reciprocity of the network to be 0.2, meaning that given a directional connection of two nodes, there's a 20% chance the connection is a two-way connection (both nodes point at each other.)

With that in mind, we can assert that the hub nodes are connected to many technologies that are highly dependent on them and that establishes a clear community unfolding within the network. We identified 9 key communities structured around the following hub nodes:

Javascript, Java, Android, Python, C++/Linux, C#, php/SQL, iOS and misc.

Conclusion

Given our findings, we have created an intuitive visualization of technologies and their weighted, directional interaction. Stack Overflow proved to be a promising platform for monitoring relationships of technologically relevant tags. The visualization of our graph can be found at http://bit.ly/1TbZAyz

References

1. Blondel, V. D., Guillaume, J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Journal of Statistical Mechanics: Theory and Experiment, 2008*(10). doi:10.1088/1742-5468/2008/10/p10008