

SODA: The Stack Overflow Dataset Almanac

Nicolas Latorre, Roberto Minelli, Andrea Mocci, Luca Ponzanelli, Michele Lanza
REVEAL @ Faculty of Informatics — Università della Svizzera italiana (USI), Switzerland

Abstract—Stack Overflow has become a fundamental resource for developers, becoming the de facto Question and Answer (Q&A) website, and one of the standard unstructured data sources for software engineering research to mine knowledge about development.

We present SODA, the *Stack Overflow Dataset Almanac*, a tool that helps researchers and developers to better understand the trends of discussion topics in Stack Overflow, based on the available tagging system. SODA provides an effective visualization to support the analysis of topics in different time intervals and frames, leveraging single or co-occurrent tags. We show, through simple usage scenarios, how SODA can be used to find interesting peculiar moments in the evolution of Stack Overflow discussions that closely match specific recent events in the area of software development. SODA is available at <http://rio.inf.usi.ch/soda/>

I. INTRODUCTION

Stack Overflow is an example of online resource that has become fundamental in the daily working life of developers [1], providing “archives with millions of entries that contribute to the body of knowledge in software development” [2], with more than 6,000 questions asked every day [3], [4]. Because of its prominent role, research leveraged Stack Overflow for many purposes, e.g., to understand how developers use APIs [5], to construct recommender systems [6], [7], or to model and improve post quality [3], [4], [8]. Essentially, Stack Overflow has enriched the field of mining software repositories with unstructured data coming from a fundamentally novel source, produced outside the development process, but consumed and fundamentally exploited inside it.

While being a valuable source for research, we believe that Stack Overflow data is important *per se*, beyond the unstructured contents of its discussions. In fact, Stack Overflow provides evidence for social aspects of development, e.g., what are the important topics for developers, and it could be used to understand popularity trends, like changes in preferred programming languages. As with any large dataset, however, a manual analysis is hard if not impossible. According to Ware [9], visualization is the preferred way of getting acquainted with large data. Inspired by this point of view, we investigate how it is possible to visualize, in an effective way, the large data set provided by Stack Overflow.

As a first step, we present SODA, a tool to visually explore the evolution of tags in Stack Overflow discussions. SODA works by analyzing the last official Stack Overflow dump¹, and storing in a local database the discussions metadata, essentially excluding their contents. The tool provides an intuitive visualization that helps to understand, in a given period of

time, how much a certain tag was used in discussions with respect to its peak occurrence. SODA also shows the evolution of tag occurrences, co-occurrences, and trends through charts.

As a proof of concept, we show how SODA can be used to discover interesting trending stories on discussion topics in Stack Overflow. For example, we show how to detect and visualize abandoned tags, and how to determine trends in the popularity of programming languages.

II. VISUALIZATION PRINCIPLES

Figure 1 shows the main user interface of SODA, composed of three parts: i) The *toolbar* on the top, which shows basic statistics about the current view, provides buttons to manipulate it, and shows additional visualizations, i.e., line charts; ii) The *tag view* in the middle, which provides the main visualization of SODA, essentially a square-packing layout that shows the distribution of tags for discussion in a given time interval; iii) the *timeline slider* and the *player* on the bottom.

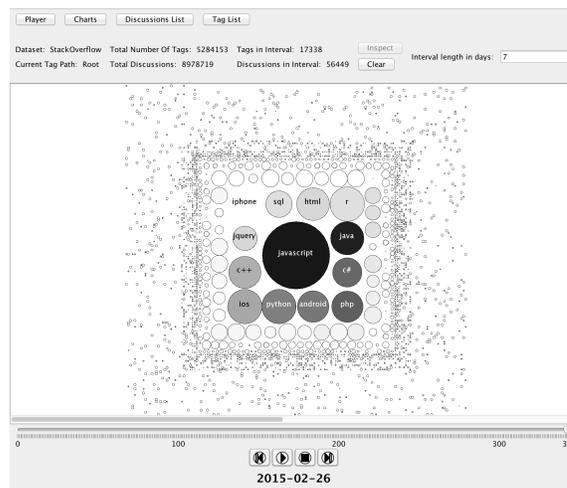


Fig. 1. The Main User Interface of SODA

The user can also choose the length of the interval for the main visualization, using the input box on the right-most part of the toolbar, with the minimum granularity of one day. In the example in Figure 1, it is set to 7 days, i.e., a week.

The tag view shows, through a compact zoomable visualization, the distribution of discussions tagged with specific tags in the time period selected from the timeline slider (i.e., the last week in Figure 1). The user can select any week in the history of Stack Overflow to show the distribution of tagged discussions, or use the player to visualize an animated evolution of the view.

¹<https://archive.org/details/stackexchange>

Figure 2 depicts the main view: Each tag is represented by a colored circle inside a square, displayed only for the sake of this explanation, but normally omitted from the view.

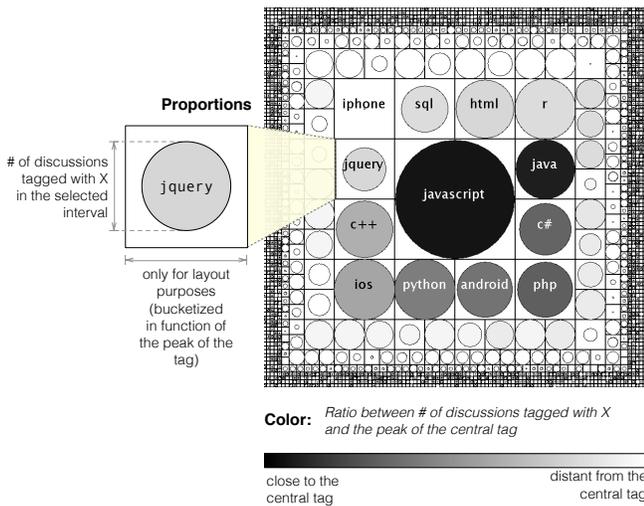


Fig. 2. The Main Visualization Explained

Layout. The layout is computed in multiple steps. First, SODA computes the occurrences of tags in discussions in each period according to the selected time interval size. Then, it determines the peak occurrence, *i.e.*, the maximum amount of discussions tagged with a given tag among all the time intervals, and sorts tags accordingly. Tags are then placed with a square-packing layout, starting from the most popular tag on the center and putting the remaining around it, at each round splitting the size by a half.

Consider again a time interval of a week. Figure 2 shows that `javascript` reached the peak as the most-discussed tag per week ever, meaning that there exists no other tag that in any week had more corresponding tagged discussions.

The size of the internal circle is proportional to the actual occurrence of discussions tagged with `javascript` in a specific week. Moreover, the circle is colored with a 30-level greyscale computed in function of the ratio between the occurrence of discussions tagged with a specific tag and the peak of the most popular tag per week (*i.e.*, `javascript`). The darker the color, the closer the number of discussions in that week is to the peak of the central tag.

For example, Figure 2 shows how discussions tagged with `ios` and `python` are closer to their relative peak (expressed by the size of the circle representing them), but their total amount of discussions is lower than the one reached by `javascript` (since their color is light grey).

Many tags in the outer layers correspond to circles perfectly inscribed in their assigned square, but they are essentially filled with white or very light shades of gray. This case corresponds to the fact that these tags are being discussed as much as in their peak, but the absolute amount of discussions is significantly lower than the most popular tag `javascript`.

Tag Co-occurrences. The user can select a tag and discover the tags the co-occur with it by pressing the a key. If the user double clicks on a tag, SODA changes the view by considering only the tags that co-occur with the selected tag, and recomputing the view accordingly. On the toolbar, SODA changes the actual *tag path* to keep track of the selected co-occurrent set of tags that correspond to the actual view.

Figure 3 depicts both the way SODA shows the tags that ever co-occurred with `javascript`, that include both popular and less popular tags.

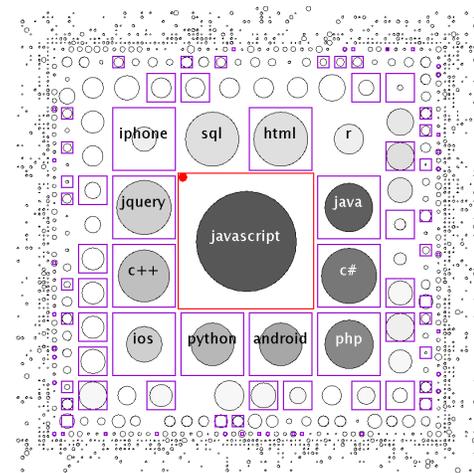


Fig. 3. Visualizing the Tags Co-occurring with javascript

Figure 4 shows the recomputed main visualization when the user double clicks on `javascript`, showing the set of tags (and the corresponding trends) for the co-occurrent tags only. The most popular co-occurrent tag is `jquery`, followed by tags like `html` and `angularjs`.

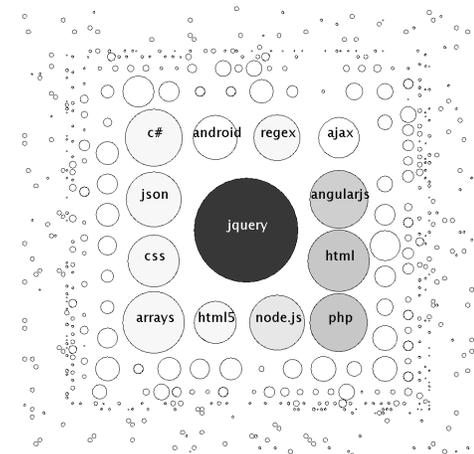


Fig. 4. Visualizing the Tags Co-occurring with javascript

Charts. SODA also provides means to depict how a set of selected tags co-evolve by means of line, bar, and pie charts. Such charts can be used to determine the actual amount of discussions tagged with a given set of tags.

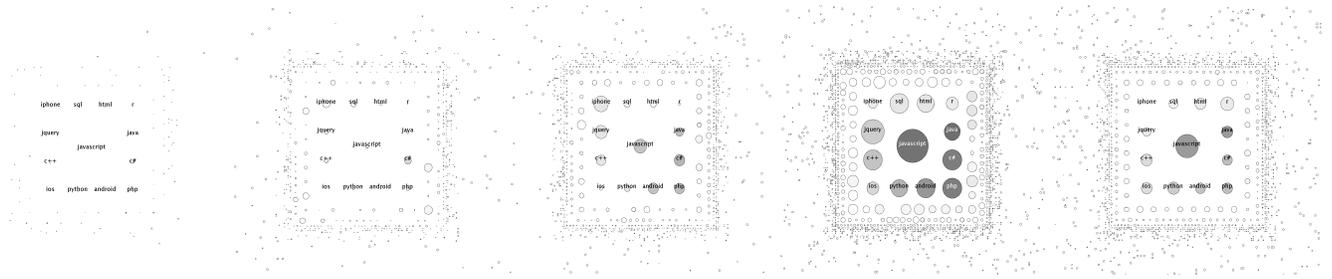


Fig. 5. Samples for the Evolution of Trending Tags per Week

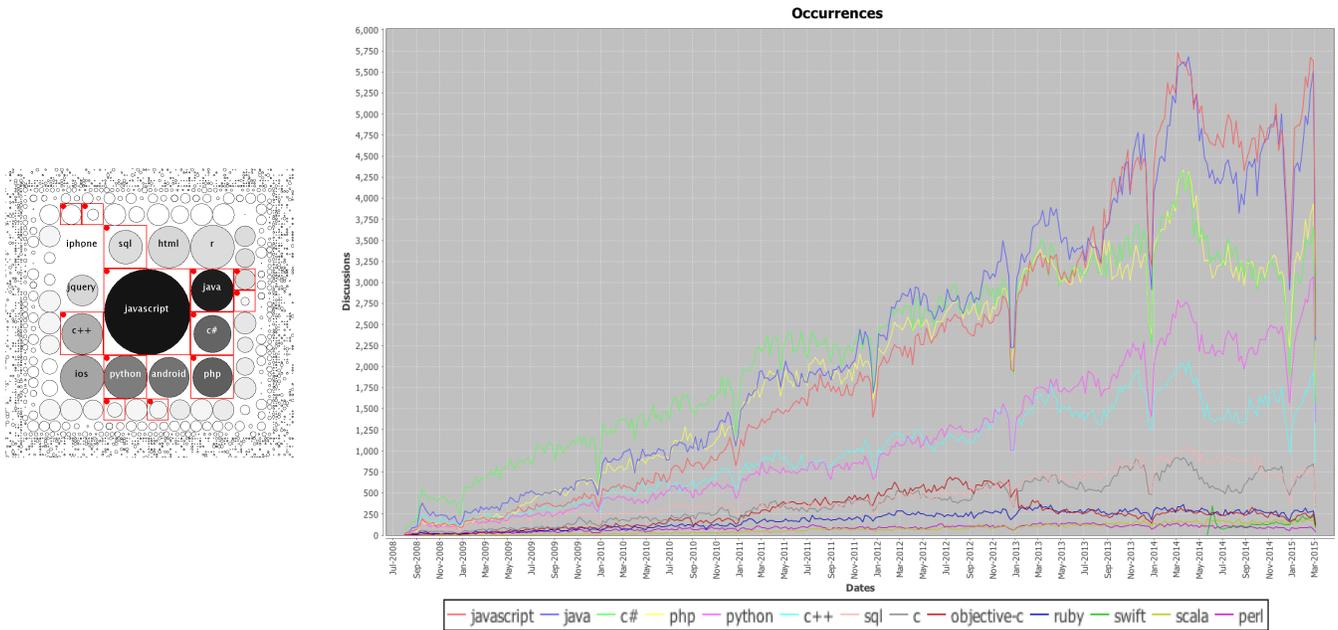


Fig. 6. Most Popular Programming Languages Tags and Corresponding Line Chart

Replaying the Evolution. The player on the bottom of the main window can be used to automatically replay the evolution of the view. Figure 5 shows five snapshots of the SODA visualizations, in chronological order, taken from the first to the last week of the Stack Overflow data dump.

Summing up. In the next section we will use the main visualization and these charts to tell evolutionary stories of the discussion trends in Stack Overflow.

III. STORIES IN STACK OVERFLOW TRENDS

We determined some interesting stories of development trends by looking at the main visualization provided by SODA. We report two of them, the *evolution of programming language popularity* and the effects of the *introduction of iOS*.

Popularity of Programming Languages. Figure 6 shows the most popular programming language tags as located in the main visualization of SODA corresponding to the last week in the Stack Overflow dump. The remaining popular tags either correspond to development platforms (e.g., `ios` and `android`) or specific frameworks (e.g., `jquery`).

From the corresponding line chart, one can spot that most of the tags are constantly growing during the whole history of Stack Overflow, which is a simple consequence of the increasing importance of this online resource.

The popularity rank between languages is also pretty much stable, with a couple of exceptions:

- around May 2012, `python` starts to become more popular than `c++`, a fact that is more evident after 2014;
- `c#` and `php` are initially the most popular languages in Stack Overflow, but they start a relative decline in favor of `java` and `javascript`, which nowadays are the most popular discussed languages;
- around January 2013, `objective-c` has a relative drop in popularity;
- recently, `javascript` is becoming extremely popular, even more than `java`.

Another interesting fact that is pretty much evident is the birth of `swift`², with a considerable spike during the first months after its introduction and a slow increase of popularity.

²See <https://developer.apple.com/swift/>

The introduction of iOS. Figure 7 depicts the last week of the Stack Overflow Dump. Almost every tag around the center is still active, except one in the upper left corner, *i.e.*, `iphone`. If a tag has ever been popular, meaning that it reached a significant (weekly, in this case) peak in the history of Stack Overflow, it occupies a significant square in the visualization. If for some reason it became less popular, that fact would result (in more recent times) in almost-empty locations. This is the case of the `iphone` tag: After Apple named the iPhone operating system as “iOS”, developers abandoned the `iphone` tag and started to adopt a new tag: `ios` (Figure 7).

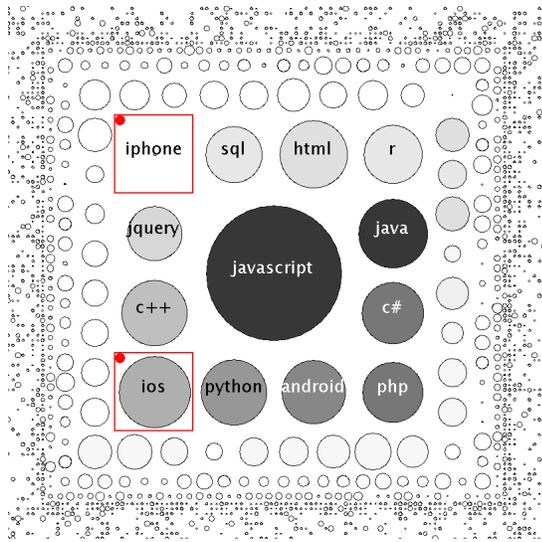


Fig. 7. The Abandoned `iphone` Tag

Figure 8 shows a line chart depicting the number of discussions, per week, tagged with `iphone` or `ios`. It can be seen the slow, initial evolution of the latter tag that slowly became predominant, up to the present time, where the former is essentially unused.

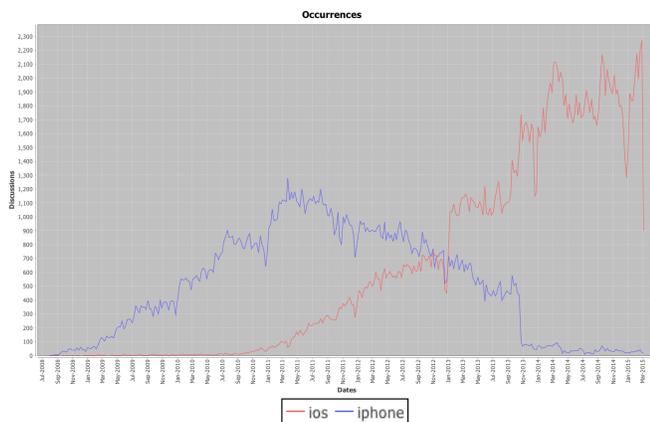


Fig. 8. The Line Chart for `ios` and `iphone` Tags

IV. RELATED WORK

Our approach can be framed in the context of studies and research about social media resources in software engineering [10]. For example, Treude and Storey [11] found that basic tagging mechanisms can be effectively used to bridge the gap between technical and social aspects of software development (*e.g.*, to support informal processes).

Another work by Treude *et al.* [2] is one of the first conceptual explorations of the Stack Overflow data, providing a general categorization of tags (*e.g.*, programming language, framework and homework) and question types (*e.g.*, instructions and unexpected behaviors). Many research approaches in software engineering have leveraged the Stack Overflow dataset, including its tagging system. For example, Xia *et al.* [12] presented a technique to store, pre-process, analyze and recommend tags, using Stack Overflow as one of the possible case studies.

Another set of related work involves studies about popularity and trends in software engineering. Achananuparp *et al.* presented an analysis to capture real-time information about trends and topics in software related microblogs (*e.g.*, Twitter), leveraging also visualization techniques [13]. In a different context, Bissyandé *et al.* [14] studied the popularity of programming languages considering around 100,000 open-source projects.

V. CONCLUSION

We presented SODA, the Stack Overflow Dataset Almanac, a tool to visualize the evolution of the trends in the Stack Overflow dataset based on tags. By choosing a time interval, SODA determines the topics that reached a peak in that interval, depicts it with a novel visualization, and enables to replay the evolution of discussion topics. SODA also visualizes the co-occurrent tags and generate trending charts for a selected set of tags. We used SODA to explore the discussion tags in the history of Stack Overflow, and we reported a couple of interesting stories related to development topics.

A. Limitations and Future Work

SODA is based on the assumption that discussion tags can be roughly mapped to topics of a discussion. This is true with a certain level of approximation. In the future, we plan to use natural language processing methods like *Latent Dirichlet Association (LDA)* [15] to model discussion topics in a more precise way. We also plan to explore different visualizations to depict different aspects of the data, like the ones related to developer demographics, comments and replies, and more importantly, aspects of discussion contents like the presence of different structured fragments like code. We also plan to consider potential filters that may improve data quality or support specific analyses, *e.g.*, by supporting selection or removal of questions without answers.

ACKNOWLEDGMENTS

We gratefully acknowledge the financial support of the Swiss National Science foundation for the project “HI-SEA” (SNF Project No. 146734).

REFERENCES

- [1] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann, "Design lessons from the fastest Q&A site in the west," in *Proc. of CHI 2011 (29th Conference on Human factors in computing systems)*. ACM, 2011, pp. 2857–2866.
- [2] C. Treude, O. Barzilay, and M.-A. Storey, "How do programmers ask and answer questions on the web? (nier track)," in *Proceedings of ICSE 2011 (33rd International Conference on Software Engineering)*, ACM, Ed., 2011, pp. 804–807.
- [3] L. Ponzanelli, A. Mocchi, A. Bacchelli, and M. Lanza, "Understanding and Classifying the Quality of Technical Forum Questions," in *Proceedings of QSIC 2014 (14th International Conference on Quality Software)*. IEEE CS Press, 2014, pp. 343–352.
- [4] L. Ponzanelli, A. Mocchi, A. Bacchelli, M. Lanza, and D. Fullerton, "Improving low quality stack overflow post detection," in *Proceedings of ICSME 2014 (30th International Conference on Software Maintenance and Evolution, Industrial Track)*, 2014, pp. pp. 541–544.
- [5] K. Bajaj, K. Pattabiraman, and A. Mesbah, "Mining questions asked by web developers," in *Proceedings of MSR 2014 (11th Working Conference on Mining Software Repositories)*. ACM, 2014, pp. 112–121.
- [6] L. Ponzanelli, G. Bavota, M. D. Penta, R. Oliveto, and M. Lanza, "Mining StackOverflow to Turn the IDE into a Self-confident Programming Prompter," in *Proceedings of MSR 2014 (11th Working Conference on Mining Software Repositories)*. ACM, 2014, pp. 102–111.
- [7] J. Cordeiro, B. Antunes, and P. Gomes, "Context-based Recommendation to Support Problem Solving in Software Development," in *Proceedings RSSE 2012, (3rd International Workshop on Recommendation Systems for Software Engineering)*. IEEE Press, 2012, pp. 85–89.
- [8] D. Correa and A. Sureka, "Chaff from the Wheat : Characterization and Modeling of Deleted Questions on Stack Overflow," in *Proceedings of WWW 2014 (23rd international conference on World Wide Web)*. ACM, 2014.
- [9] C. Ware, *Information Visualization: Perception for Design*, 2nd ed. Morgan Kaufmann, 2004.
- [10] A. Begel, R. DeLine, and T. Zimmermann, "Social media for software engineering," in *Proceedings of FOSE 2010 (FSE/SDP Workshop on Future of Software Engineering Research)*, 2010, pp. 33–38.
- [11] C. Treude and M.-A. Storey, "How tagging helps bridge the gap between social and technical aspects in software development," in *Proceedings of ICSE 2009 (31st ACM/IEEE International Conference on Software Engineering)*, 2009, pp. 12–22.
- [12] X. Xia, D. Lo, X. Wang, and B. Zhou, "Tag recommendation in software information sites," in *Proceedings of MSR 2013 (10th Working Conference on Mining Software Repositories)*, 2013, pp. 287–296.
- [13] P. Achananuparp, I. Lubis, Y. Tian, D. Lo, and E.-P. Lim, "Observatory of trends in software related microblogs," in *Proceedings of ASE 2012 (27th IEEE/ACM International Conference on Automated Software Engineering)*, 2012, pp. 334–337.
- [14] T. Bissyande, F. Thung, D. Lo, L. Jiang, and L. Reveillere, "Popularity, interoperability, and impact of programming languages in 100,000 open source projects," in *Proceedings of COMPSAC 2013 (37th IEEE Annual Computer Software and Applications Conference)*, 2013, pp. 303–312.
- [15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.