BitQuery[1] - a GitHub API driven and D3 based search engine for open source repositories

L. Borke[a] and S. Bykovskaya[b], [1]http://bitquery.de/ (best viewed in Chrome)

[a]*Humboldt-Universität zu Berlin,* [b]*Lomonosov Moscow State University*

With the growing popularity of GitHub, the largest host of source code and collaboration platform in the world, it has evolved to a Big Data resource offering a variety of open source repositories. Multiple libraries and package managers, among them CRAN, PHP, NPM and many others, mirror their data in GitHub repositories and organizations, while a growing number of developers are releasing their packages directly on GitHub. We present BitQuery, a new GitHub API driven search engine which (I) queries the repository metadata via text mining and clustering methods, thus providing an automatic categorization system for software developers (II) establishes visual data exploration and topic driven navigation of GitHub repositories for more transparency in source code management (III) promotes discoverability and technology transfer for data science and project teams.

The BitQuery architecture consists of three abstraction layers, following the ETL paradigm (Extract, Transform, Load), or, equivalently, the visual analytics approach (data management, analysis, and visualization). First, the information is queried via the GitHub API based parser layer. Next, the Smart Data layer transforms Big Data into value, processing the data semantics and metadata via dynamic calibration of metadata configurations, text mining (TM) models and clustering methods. One of the examined TM models is the latent semantic analysis (LSA) technique which measures semantic relations and allows dimension reduction. Both layers were implemented via several novel R packages, forming the basis of a self-contained "GitHub Mining infrastructure in R"[1].

Thus derived Smart Data is loaded into the web-based visual analytics application (VA-App) realized via the D3-3D Visu layer, which is powered by two JavaScript libraries for producing dynamic data visualizations in web browsers: D3.js and Three.js. The D3-3D Visu layer was designed in full compliance with the so-called visual information seeking mantra: "Overview first, zoom and filter, and then details-on-demand." The highly customizable graph layout performs visual software mining from multiple perspectives.

The application spectrum of BitQuery consists of three different layout types: Special, CRAN and GitHub edition [2]. The following metadata types are supported: JSON, YAML, DCF and Markdown. This demonstrates a great potential of BitQuery as a VA-App which increases the visibility and discoverability of any organization or digital library hosted on GitHub.

References

[1] L. Borke (2017). Dynamic Clustering and Visualization of Smart Data via D3-3D-LSA. *Doctoral thesis*. Humboldt University of Berlin.

[2] L. Borke and S. Bykovskaya (2017). BitQuery: a GitHub API driven and D3 based search engine for open source repositories. Available at http://bitquery.de.