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.borke.net/ (prototype: Chrome recommended)

[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 (OSR). Multiple libraries and package managers, among them CRAN, CPAN, WordPress and many others, mirror their data in GitHub 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) provides an automatic OSR categorization system for data science teams and software developers, and (II) establishes visual data exploration (VDE) and topic driven navigation of GitHub organizations for collaborative reproducible research and web deployment.

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 extracted 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 [1],[2]. 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"[3]. 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." Various techniques and interactive interfaces perform VDE from multiple perspectives.

The application spectrum of BitQuery is illustrated by the exploration of the "R universe", a massive collection of all R packages on GitHub including CRAN and Bioconductor. This example shows a great potential of BitQuery as a VA-App which increases the visibility and discoverability of any organization or digital library hosted on GitHub.

Keywords: Software Mining, Clustering Analysis, Visual Analytics

References

[1] L. Borke and W. K. Härdle (2017). Q3-D3-LSA, in W. K. Härdle, H. H. Lu and X. Shen (eds), *Handbook of Big data Analytics*. Springer.

[2] L. Borke and W. K. Härdle (2017b). GitHub API based QuantNet Mining infrastructure in R. *SFB 649 Discussion Paper*, Humboldt Universität zu Berlin.

[3] L. Borke and S. Bykovskaya (2017). GitHub Mining Infrastructure in R. *forthcoming.*