Oliver Keyes

Geolocation • Content consumption • R • Hadoop • C++ • Java

Apartment 3
47 Park Vale Avenue
Boston, MA 02134 United States
(415) 528-9024

ironholds@gmail.com
http://ironholds.org/
https://github.com/Ironholds/

Summary

A driven and innovative researcher for a top-10 website, currently working as the lead traffic analyst for the Wikimedia Foundation. Research specialisations include geographic analysis, temporal analysis, and understanding the implications of the growing mobile web. Technical specialisations include geolocation, session reconstruction, and the use of high-volume, high-throughput, Hadoop-based systems in research.

Professional contributions

2013-present Research Analyst for the Wikimedia Foundation. In this role, my work has included:

- Pioneering the use of Hadoop by our research team. When I joined the team, the system was in 'alpha': I provided extensive testing, reduced the barrier to usage by building a standardised R connector to the databases on top of RJDBC, demonstrated the first use of the system to provide standardised, regular reports to our product team about key consumption metrics, and built a host of Java-based UDFs to reduce the barrier to efficient use of the system.
- Defining key content consumption metrics, most crucially 'pageviews' and 'sessions'. This involved extensive testing of fingerprinting methods, a deep dive into academic work around session reconstruction, and produced both an academic paper and an open-source library (see below). Both definitions are now consistently used to drive product's understanding of our user base.
- Studying the geographic and cultural diversity of our user base, and how geographic location, device and connection types, and experience levels impact contribution rates.
- Providing support to our Product Development department in understanding our userbase through providing both automated and ad-hoc summary datasets and visualisations

2011-2013 Community Liaison for the Wikimedia Foundation. This role was akin to that of a business analyst, consisting largely of acting as a conduit between the product Development team and our stakeholders. Particular highlights were:

- Building a relationship between our product managers and users, helping both sides empathise with the other's conflicting needs and priorities
- Dramatically expanding the breadth and depth of our interactions with users, hosting weekly, live discussions between engineering teams and the users dependent on their work.

Academic contributions

Papers

2015 Halfaker, Aaron; **Keyes, Oliver**; Kluver, Daniel; Thebault-Spieker, Jacob; Nguyen, Tien; Shores, Kenneth, 'User Session Identification Based on Strong Regularities in Inter-activity Time'. *Proceedings of the International World Wide Web Conference* (accepted)

2015 Sen, Shilad; Ford, Heather; Musicant, David; Graham, Mark; **Keyes, Oliver**; Hecht, Brent, 'Barriers to the Localness of Volunteered Geographic Information'. *Proceedings of the International Conference on Computer-Human Interaction* (accepted)

2013 Halfaker, Aaron; **Keyes, Oliver**; Taraborelli, Dario, 'Making Peripheral Participation Legitimate: Reader Engagement Experiments in Wikipedia'. *Proceedings of the 2013 Conference on Computer-Supported Co-operative Work* (published)

Talks

2015 Keyes, Oliver; '[vectorisation needed]: Analysing Traffic to a Top-10 Website with R'. *Effective Applications of the R Language (EARL) 2015* (submitted)

2014 Keyes, Oliver; 'Big in Japan: Combating Systemic Bias Through Mobile Editing'. *Wikimania 2014* (presented)

Open-Source contributions

A strong proponent of open-source software development and the copyleft movement, I release both datasets and codebases publicly wherever possible. Highlighted releases include:

- openssl, a cryptographic library developed in collaboration with Jeroen Ooms. Openssl makes the OpenSSL cryptographic library available in R, providing vectorised, rigorously-tested cryptographic hash generation several orders of magnitude faster than the existing digest package. R frontend, C backend.
- urltools, a URL-related toolkit for R. Provides vectorised, efficient versions of URLencode and URLdecode while fixing several bugs in their existing implementations; also includes a URL parser and component retrieval/setting operations. R frontend, C++ backend.
- reconstructr, a generalised, mutable library for reconstructing user sessions and calculating a host of common metrics from the resulting dataset, such as session length and time-on-page. R frontend, C++ backend.
- rgeoip, a client library for the MaxMind geolocation binaries. Capable of geolocating a million IP addresses in six seconds; R frontend, C++ backend.