Microsoft

The Data Stack in Information Retrieval

Omar Alonso

Disclaimer

The views, opinions, positions, or strategies expressed in this talk are mine and do not necessarily reflect the official policy or position of Microsoft.

IR community and data

Data is the new oil

Database community

Beckman report [CACM 2016]

Where is IR?

Big data

Small data

Utility value

People

Engineering and research roles

Full stack developer

Data modeling, business logic, API, user interface, and user experience
File system, testing
LAMP (Linux, Apache, MySQL, PHP); AWS/MongoDB, NodeJS; Azure

Data scientist

Data analysis and mining, experimentation
New tools: R, MapReduce-like, scripting

Data stack in IR

Ingestion and processing of raw data

Crawling, data cleaning, near-duplicate detection, etc.

Annotation for augmenting the ingested data

Named-entity detection, information extraction, metadata generation, content classification, etc.

Indexing and ranking of high quality content

Efficient data structures, feature engineering, etc.

Behavioral data for capturing user activity

Search query logs, clicks, link sharing, etc.

Experimentation and analysis infrastructure for evaluation and exploration

Sampling, interleaving, A-B testing, crowdsourcing, etc.

Other dimensions

Search

Desktop
Enterprise
Web
Mobile
Cloud

Computing power, data size, and relevance

Data driven IR

Design and modeling

Exploratory data analysis

Experiment and prototype often

Numbers

"Getting numbers is easy; getting numbers you can trust is hard" R. Kohavi

Avoid HIPPO (Highest Paid Person's Opinion)

No p-hacking

Debugging

Real-world systems

Scalability & availability 24x7

Challenges and opportunities

Great time to work on IR and related stuff

R&D at different levels of the stack

Problems, metrics, and user satisfaction

Work at different levels, using different techniques at each step

Work with all sorts of [small | medium | large] data

Spend more time looking at data

Potential for new information seeking systems

Thanks!

We are hiring ☺

Email: omalonso@microsoft.com
Twitter: @elunca