



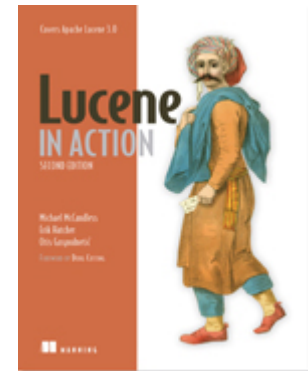
Search Analytics

What? Why? How?

Otis Gospodnetić – Sematext International

About Otis Gospodnetić

- Member: Apache Lucene, Solr, Nutch, Mahout
- Author: Lucene in Action 1 & 2
- Entrepreneur: Sematext, Simpy



About Sematext

Consulting, development, support:

- **Search** (Lucene, Solr, Elastic Search...)
- **Big Data** (Hadoop, HBase, Voldemort...)
- **Web Crawling** (Nutch)
- **Machine Learning** (Mahout)

Agenda

- **Intro: Otis & Sematext - DONE**
- **What**
- **Why**
- **Reports**

What is Search Analytics?

- Input: queries and clicks
 - Output: reports – over time
 - Next: actions
-
- The means, not the goal
 - Ongoing, not one-off

Search Analytics and SEO

- Not the same
- SA can help with SEO

Search vs. Web Analytics

- User intent and information needs vs. inferring
- Hand in hand
- Ideally you can relate data from both or even unify it

Why Search Analytics?

- Measure and monitor everything
- Supports (re)design, navigation choices
- Helps with content acquisition & enhancement
- Improve search experience
- Mula

Report Groups

- Failures vs. non-failures
- Actionable vs. non-actionable

Failures

Be aware of failures, but don't be one.

- Zero hits
- Low query CTR
- High search exit rate
- Irrelevant results
- Over N refinements

Report: Zero Hit Queries

- Popular queries; percentage, not raw count
- Misspellings?
- Synonyms?
- No matching content?
- Need (different) tagging?
- Bad analysis?
- Multilingual issue?

Report: Zero Hit Queries (cont.)

- Use Query Spellchecker (aka DYM)
- Using AutoComplete
- Using DYM ReSearcher
- Designing No Results page

Report: High Exit Rate Queries

- Disappointed, frustrated users
- Major revenue loss, no second chance
- Marriage with Web Analytics
- Relevance bad?
- Default ordering bad?
- Titles of hits need adjusting?
- Search terms highlighting looking bad?
- Bad thumbnails? Need thumbnails?

Report: Irrelevant Result Queries

- Manual: top N hits of top N queries
 - Judge relevance of each hit and assign score
 - Per-query score: sum scores of top N hits
 - Cumulative top N query score: sum per-query scores
-
- Automated: Mean Reciprocal Rank (MRR)

Report: Total Queries

- Search vs. navigation/browsing
- Search vs. overall site usage
- Related report: % of visits with search
- Segment: new users vs. return users, etc.

- Questions: do you count paging? Facet selection? Re-sorting?

Report: Total Distinct Queries

- What's distinct? Car vs. Cars
- $\# \text{ Total Queries} / \# \text{ Distinct Queries} = \text{Avg. } \#$
- Tied to performance and query cache utilization
- Extension: Total distinct words in queries

Report: Words Per Query

- Informative, slowly changing, not terribly actionable
- Can affect search box size
- Use AutoComplete if queries are long

Report: Top Queries

- User intent and information needs
- Ensure good results
- Calculate MRR for top N queries
- Calculate MRR for each top N query
- Compare to global MRR

Report: Top Queries (cont.)

- New top queries – new trend? New demand?
- Best Bets (aka Query Elevation in Solr)
- Expose before search is needed
- Seasonality – hour of day, day of the week, etc.
 - Adjust content presentation and availability (e.g. week vs. weekend, business vs. personal)
 - Anticipate demand in the next cycle

Clickstream Analysis

- Query analysis is not a complete story:
 - Queries
 - Clicks
 - (Trans)action

Query and Hit Valuation

- Query: by popularity (count)
- Query: by CTR
- Query: by subsequent (trans)action count/pct.

- Hit: by click count
- Hit: by subsequent (trans)action count/pct.

Query and Hit Valuation (cont.)

- Maximize:
 $\text{pop}(\text{query}) + \text{ctr}(\text{query}) + \text{action}(\text{query})$
 $\text{clicks}(\text{hit}) + \text{action}(\text{hit})$
- Failures:
high $\text{pop}(q)$, yet low $\text{ctr}(q)$
high $\text{pop}(q)$, high $\text{ctr}(q)$, yet low $\text{action}(q)$
- Integration with backend required

Report: Low CTR Queries

- Popular queries; percentage, not raw count
- Relevance bad?
- Default ordering bad?
- Titles of hits need adjusting?
- Search terms highlighting looking bad?
- Bad thumbnails? Need thumbnails?

Report: Queries with Most Clicks

- i.e. Queries with Highest CTR
- Informative? Yes
- Actionable? Somewhat: expose relevant content outside of search

Search Session

- Search activity aimed at satisfying a specific information need in a some limited amount of time.
- i.e. it's very fuzzy

Interesting Search Sessions

- More than N queries in M minutes
- Sessions that end in a failure
- Sessions for specific type of info (e.g. person name, product name, event)

Segmentation

- Searches that resulted in conversion vs. not
- Search metrics for
- New vs. returning visitors
- English vs. French vs. Spanish vs. ...
- Chrome vs. IE
- ...

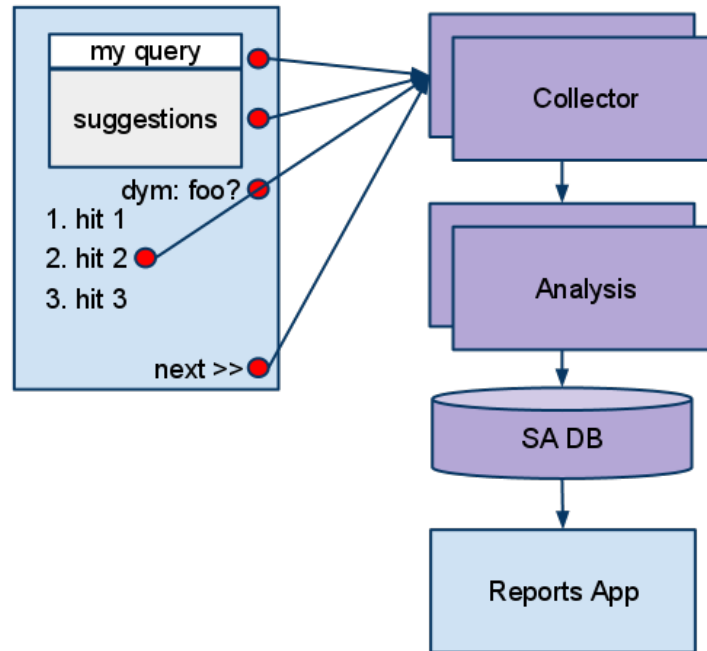
More SA Reports/Questions

- % of queries from DYM vs. AC vs. typed
- Most common queries per clicked hit
- Which hits are generally popular?
- Which hits are trending up?
- Are there docs that are never ever clicked on?
- Average number of queries per session
- Breakdown of queries by number of hits

More SA Reports/Questions

- Breakdown of queries by latency
- Frequently used facets or sort criteria
- Avg number of clicks per query
- Time spent on site before/after searching
- Search initiation pages
- How deep into SERPs are people drilling?
- Are too many clicks on pages other than 1st?
- ...

Data Collection



Contact

- sematext.com
- blog.sematext.com
- [@sematext](https://twitter.com/sematext)
- [@otisg](https://twitter.com/otisg)
- otis@sematext.com

