When no clicks are good news

Carlos Castillo, Aris Gionis, Ronny Lempel, Yoelle Maarek

Yahoo! Research Barcelona & Haifa
Usage mining for search

- **Behavioral signals are useful** to measure performance of retrieval systems.

- Relevant results are:
  - clicked more often,
  - visited for **longer time**,
  - lead to long-term **engagement**,
  - etc.

- However, predicting user satisfaction accurately from search behavior signals is still an **open problem**.
If we satisfy the user by impression, then we observe a lower click-through rate.
Satisfaction by impression
Oneboxes and Direct Displays

Oneboxes\(^1\) and Direct Displays\(^2\) (DD) are

- **Very specific results** answering (mostly) unambiguous queries with a unique answer directly on the SERP

- **Displayed above regular Web results**, due to their high relevance, and in a slightly different format.

- **Typical example:** weather <city name>
  - Test: guess which onebox/DD was served by which search engine:-)

---

1: Google terminology
2: Yahoo! terminology
Increasing number of “by impression” results

- When searching for specific stocks, movie or train schedules, sports results, package tracking (Fedex/UPS), etc.
- To the extreme, what about spell checking, arithmetic operations or currency conversion, addresses, things to do?
The problem

• Click-based metrics for user satisfaction
• For cases where we expect no clicks

• Not only search sessions
  – Any browsing/interaction session
Our proposal

- General method
  - Pick a class of users with a distinctive behavior
  - Study their response to changes
Our proposal

• General method
  • Pick a class of users with a distinctive behavior
  • Study their response to changes

• Specific method
  – Find users who are “Tenacious”
    • reformulate or click, do not let go
  – Measure their abandonment
How to model users?

• Session representation
  – Actions classes: queries and clicks
    • XQCQX means “start, query, click, query, stop”
  – Alternative: reformulation classes

• User representation
  – Frequency of action 3-grams = 15 features in total
  – Tenacity = \((XQQ+XQC)/(XQQ+XQC+XQX)\)
(Preliminary) experiments

- Segment sessions into logical “goals”
- Divide goals in two groups
  - With direct-displays above position 5 (DD)
  - Without (NO-DD)
- Metric
  - Find users with $Tenacity_{NO-DD} \geq 80\%$
  - Measure $\frac{Tenacity_{DD}}{Tenacity_{NO-DD}}$
- Ground truth
  - Ask humans “do you think users querying Q will be satisfied by impression by this DD?”
    - 1=never ... 5=always
Pitbull: editorial vs metric (type “weather”)

Change in the tenacity of tenacious users

Editorial label (satisfaction: 1=never, 5=always)
Pitbull: editorial vs metric (type “weather”)

63% of bad cases
83% precision
Pitbull: editorial vs metric (type “reference”)
Pitbull: editorial vs metric (type “reference”)

71% of bad cases
84% precision

Change in the tenacity of tenacious users
Summary

- Tenacious users can be used to identify bad DDs
- General method: usage mining on classes of users
  - Shoppers
  - Smart searchers
  - Click-a-lots / explorers
  - Leaders
  - Poodles?
  - etc.
- General/shared taxonomy of users?
Thank you!

chato@yahoo-inc.com