AB testing / Search / Recommendations

Roman Chyla and the ADS Team

ADS Users Group Meeting, 19-20 Nov. 2020
Overview

- **A/B Testing**
  - To measure impact of changes on user behaviour

- **Recommendations**
  - First real-life experiment

- **Search Engine**
  - Many small updates
  - One big: upgrade of the codebase
A/B Overview
A/B Testing

- Deploy multiple versions to the client (UI)
  - Users are unaware of participating in an experiment
  - “Experiment with ID: 4Sq1k0daScCOQKa1IBX9AA is on variant: 2”
  - Potentially running multiple experiments simultaneously
- Measuring user behaviour
  - Tracking actions/events
  - Plans for the future: analyze user behaviour across time
    - Step 0: searched
    - Step 1: clicked $ith$ result (and spent X seconds reading)
    - Step 2: changed search parameters (unhappy with results, went back/left)
  - The time as a very important but hard to capture dimension
Recommendations

The completed SDSS-IV extended baryon oscillation spectroscopic survey: growth rate of structure measurement from anisotropic clustering analysis in configuration space between redshift 0.6 and 1.1 for the emission-line galaxy sample
Tamone, Amélie; Raichoor, Anandi; Zhao, Cheng ...

The Completed SDSS-IV extended Baryon Oscillation Spectroscopic Survey: exploring the halo occupation distribution model for emission line galaxies
Avila, S.; Gonzalez-Perez, V.; Mohammad, F.G. ...

The completed SDSS-IV extended baryon oscillation spectroscopic survey: geometry and growth from the anisotropic void-galaxy correlation function in the luminous red galaxy sample
Nadathur, Seshadhri; Woodford, Alex; Percival, Will J. ...

Removing imaging systematics from galaxy clustering measurements with Obiwan: application to the SDSS-IV extended Baryon Oscillation Spectroscopic Survey emission-line galaxy sample
Kong, Hui; Burleigh, Kaylan J.; Ross, Ashley ...

Beware of commonly used approximations. Part I. Errors in forecasts
Bellomo, Nicola; Bernal, Jose Luis; Scoirto, Giulio ...

A brief review on cosmological analysis of galaxy surveys with multiple tracers
Wang, Yuting; Zhao, Gong-Bo;}

Open Cluster Chemical Homogeneity throughout the Milky Way
Poovelli, Vijith Jacob; Zasowski, G.; Hasselquist, S. ...

The far infrared and optical properties of the Nubeculae H II region
Ezawa, Hiroshi; Zasowski, G.; Da Costa, G. S. ...
Recommendations

- Second “real” experiment
  - Based on existing implementation of the search (plus one new operator)
    - similar(user) AND entdate:[NOW-5DAYS TO *]
    - trending(user) AND ....
  - Interesting discussion leading up to the experiment
    - Do we have enough users?
    - Do people even see/care for the area under the search bar?
    - Can we generate meaningful recommendations?
  - Answers are pretty clear-cut
    - Perhaps not what some hoped for, but it is quite nice to have "hard" data
"similar()" is the leader - End the experiment

Deploy "similar()" which has a 100% probability to be best. Learn more.

**ACTIVE VISITORS**

178

**COLLECTED SESSIONS**

817,465

**DAYS**

21

**STARTED MANUALLY:**

Tue, Oct 20, 2020, 6:02 PM EDT

**EXPIRATION SCHEDULED:**

Mon, Jan 18, 2021, 5:02 PM EST

---

**Variant**

**Original**
- 10 active visitors
- 33,099 Experiment Sessions
- 0 Experiment Conversions
- 0.00% Calculated Conversion Rate

**similar()**
- 81 active visitors
- 389,779 Experiment Sessions
- 295 Experiment Conversions
- 0.08% Calculated Conversion Rate

**trending()**
- 87 active visitors
- 388,113 Experiment Sessions
- 172 Experiment Conversions
- 0.04% Calculated Conversion Rate

---

**OPTIMIZE ANALYSIS**

**Probability to be Best**
- 100% for similar()
- 100% for trending()

**Probability to beat Original**
- 100% for similar()
- 100% for trending()

**Modeled Conversion Rate**
- Baseline

**Modeled Improvement**
- 922,337,203,665,478% to 922,337,203,665,478%
Search *(from the last year)*

- **Significant changes to relevancy computation**
  - This was lots of fun
  - Special thanks to Kelly and Alberto

- **New algorithm resembles old Classic**
  - We don’t know if it is good enough!
    - We like it though
    - And users may not actually care (wonderful example of too much ado about nothing)
  - Examples to illustrate the problem [1], [1b], [2], [3]
    - Relevancy in ADS Classic
    - Final score computation in SOLR
    - Picking appropriate weights
    - Avoiding double counting
Current ADS (default behaviour)
Current ADS (default behaviour II.)
When OR is the default operator
When OR+relevancy is the default

The first 6 results exactly the same as ADS default

← 383 vs 2139303
Hopefully in next few months...

- We gather enough evidence to convince ourselves that the default search scoring can (and must) evolve
  - And then start learning to rank better
- Also need to develop better tools/methodology for analyzing data we are collecting
  - Across users/time periods
  - In-house aggregation; analytical engine