CS276 Information Retrieval and Web Search

Hugh E. Williams

hugh@hughwilliams.com

http://hughewilliams.com

@hughewilliams
Search in Industry

• A few stories with practical illustrations:
  – Understanding users with A/B testing at scale
  – Using data to build systems: clicks, graphs, and their applications
  – The practical intersection of software and people
Search in Industry

• A “flywheel” of software, data, and analytics
• Optimizing the “spinning” of the flywheel
• Ensuring you have more data to build better products
• Ensuring you have better analytics to build better products
UNDERSTANDING USERS WITH A/B TESTING
An A/B test: Larger Images in eBay’s Search
A/B experimentation

• Divide the users into populations
• One (or two) population is the control (or A)
• One or more populations are the tests (or B)
• Collect (vast amounts of) data from each population
• Compute “metrics” from the data, including confidence intervals
• Understand the results
• Make decisions
The A/B Experimentation Flywheel

• An example of a “flywheel” of software, data, and analytics
• Build an experimental product
• Collect data about the new feature
• Compare the new feature to existing features – iterate and improve
Basic Scientific Method

• A/B experimentation must follow basic scientific method. We need to:
  – Have a hypothesis
  – Have a control
  – Have one independent variable
  – Measure statistical significance
  – Draw conclusions
  – Keep track of data
  – Ensure it is repeatable
A/B experimentation...  

• From scientific method follows experimental design:  
  – Writing, agreeing, and approving an hypothesis for testing. Must include the metrics  
  – Understanding the nature and limitations of the control  
  – Trying to minimize the number of independent variables  
  – Getting enough traffic to find signal from noise:  
    • Understanding the exposure of the feature  
    • Deciding how many users see the experiment concurrently  
    • Deciding how long to run the experiment  
    • Deciding on the confidence interval (CI)  
  – Agreeing on the metrics, statistical measures, and CIs  
  – Keeping track of data, and watching the experiment  
  – Ensuring it is actually repeatable
Is it ready for real customers?
A/B experimentation...

• On any given day, Internet companies run many experiments

• It’s very likely each user is in one or more:
  – It could be “more” when they the experiments are probably orthogonal
  – It could be “more” when they are small, short, sanity checks – they do not always worry whether they’re orthogonal
eBay conversion increased by 13%
A/B experimentation...

• It is hard to segment the population:
  – Many users are not logged in
  – Many users use several browsers
  – Many users use several devices
  – (It would be easier to segment in other ways – but it is user effects we want to measure)

• Attributing a result to a change is hard
  – Need to know that the user was affected by the change
  – Need to understand interaction effects caused by changes
  – Metrics are noisy

• Decisions are made with less than 100% confidence
A/B experimentation...

- Test versus control experimentation is challenging:
  - Hard to measure long-term effects
  - Hard to understand “burn-in” time
  - Does not work well for large changes
  - May not work well for very small changes
  - Some effects are seasonal
  - Metrics do not always capture the change
  - Difficult in the modern world of mobile applications, and multi-device users
Taking care with decisions

• Do not use or report the mean average; measure quartiles or percentiles
• There are (almost) always examples where the test is worse than the control
  – What fraction of users have a worse experience?
  – How bad is the worst 5%?
  – What can you learn from failures?
• What can you learn from successes?
What won’t quantitative testing tell you?

• (Usually) An unambiguous picture
• What you do not measure
• Exactly what to do
• Anything qualitative
• When to take risks
• When to be decisive
• How to leap from one mountain to another (more on this later)
CLICKS, GRAPHS, AND THEIR APPLICATIONS
The Product Development Flywheel

• A related “flywheel” of software, data, and analytics
• Working with (vast) data sets to analyze and understand user behaviors
• Building products based on insights and (often) powered by data sets
Example #1: Ranking with Clicks

- For each prior query, sort results by prior “good click” count
- Display results sorted by decreasing “good click” count
- Pros: wisdom of the crowds; simple; distills query intent
- Cons: needs bootstrapping; self-reinforcing prophecy problems; less effective with less data (“tail” problems)
Ranking in the tail: the “Panda Graph”

From Craswell and Szummer, 2007
The argument for data-oriented approaches

- With less data, more complex algorithms outperform simple algorithms
- With more data, simple algorithms improve
- Sometimes, simple algorithms outperform complex algorithms when there is enough data
- (From Brill and Banko, 2001.)
Example #2: Using query-click graphs for query rewriting

- In 2009, eBay’s search engine was literal
- The goal was to make it more intuitive
- Idea: Using query, click, and session data, build query rewrites that improve effectiveness by intuitively including **synonyms** and **structured data**
How do eBay buyers purchase the pilzlampe?

• It turns out, they do one of a few things:
  – Type *pilzlampe*, and purchase
  – Type *pilzlampe*, *pilz lampe*, and purchase
  – Type *pilzlampe*, *pilzlampen*, and purchase
  – Type *pilz lampen*, *pilzlampe*, and purchase
  – ...

How do buyers purchase the pilzlampe?

• From eBay’s large scale data mining and processing:
  – Without human supervision, eBay automatically discovers that *pilz lampe* and *pilzlampe* are the same
  – They also discover that *pilz* and *pilze* are the same, and *lampe* and *lampen* are the same

• From these patterns, they (figuratively) rewrite the user’s query *pilzlampe* as:

  *pilzlampe* OR “*pilz lampe*” OR “*pilz lampen*” OR *pilzlampen* OR “*pilze lampe*” OR *pilzelampe* OR “*pilze lampen*” OR *pilzelampen*
Scale and the tail

• Nothing is easy in the tail or at scale
  – Incorrect strong signals:
    • CMU is not Central Michigan University
    • Mariners is not the same as Marines
  – Context matters
    • Correcting Seattle Marines to Seattle Mariners is (generally) right
    • Denver Nuggets is not Denver in the Jewelry & Watches category at eBay
  – Freshness matters
    • When the iPad launched, eBay corrected iPad to iPod
    • Even the query britney spears needs freshness and is dynamic
  – The tail is sparse:
    • Error prone, and hard to tell when errors have been made
    • Need to use traditional techniques, or need to map a query carefully to another query

• The reality is that blended approaches work best, and they are built using machine learning (see the “Learning to Rank” lecture in a few weeks)
Example #3: Other Applications

• Query-document ranking and query alterations are two prominent areas
• Others include:
  – Adult and spam filtering
  – Query suggestions (related searches)
  – Query autosuggest
  – Creating document snippets / captions / summaries
  – Spike and news-worthiness detection
  – Showing local results
  – ...

HUMANS IN THE LOOP
Where People Matter in Search

• Within the “flywheel”:
  – Analytics to product: decisions, intuition, and innovation
  – Product to data: instrumentation, cleansing, and organization
  – Data to analytics: metrics, measurement, and understanding

• There is more to search than the flywheel...
Example #1: The local vs. global maxima phenomena
Example #2: Intuition using Data
Example #2: Intuition using Data
Example #3: Don’t over-engineer
PS. Enjoyed the lecture? Give something back at http://fiftytwofives.com

QUESTIONS & ANSWERS
Additional Reading
