UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Ranking Models

Experimental Methods

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br

One problem

***f*(*q*, *d*)**

Many solutions

Similarity-based models: $f(q,d) = \mathrm{sim}(q,d)$
- Vector space models

Probabilistic models: $f(d,q) = p(R=1|d,q)$
- Classic probabilistic models
- Language models
- Information-theoretic models

Many solutions

Extended models
- Beyond bags-of-words
- Beyond lexical matching
- Beyond queries

Machine-learned models
- Beyond single features

Why evaluate?

Lots of alternative solutions
- Which one to choose?
- How to improve upon them?

Evaluation enables an informed choice
- Rigor of science
- Efficiency of practice

Why evaluate?

IR as an applied scientific discipline
- Experimentation is a critical component

IR has become plagued with weak experimentation
- Outsiders think of IR as non-scientific
- Minor improvements vs. weak baselines
- Difficulty in defining the “state-of-the-art”

Why evaluate?

Convince others
- Reviewers, other researchers, funders
- Company VPs, investors, clients

Convince yourself
- ”If you can’t measure it, you can’t improve it”
- Evaluation guides meaningful research directions

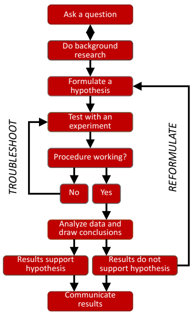
What to evaluate?

Three fundamental types of IR research
- Systems (efficiency)
- Methods (effectiveness)
- Applications (user utility)

Evaluation plays a critical role for all three
- Our primary focus is on “methods” research

How to evaluate?

Scientifically, of course!



Asking questions

What problem are you trying to solve?
- Or in IR parlance, what task?

Hard to solve an ill-defined task!
- Is it a well-known task? Review the literature!
- Is it unlike anything done before?

Asking (new) questions

Characterize the task
- How is the system used?
- What are the inputs? Outputs?
- How do you define success?

Formulating hypotheses

A hypothesis must be falsifiable
- Ideally concerning an isolated component
  e.g., *smoothing improves language modeling*

It either holds or does not…
- … with respect to the considered data (scope)
- … perhaps under certain conditions (extent)

Performing experiments

Key components
- Experimental setup
- Analysis of results

Key concern: ***reproducibility***
- Must specify each and every detail needed for reproducing our method and the experiment

Experimental setup

Research questions

Evaluation methodology

Evaluation benchmarks

Reference comparisons

Parameter tuning

Research questions

Methods are not devised arbitrarily
- We always have a hypothesis (whether implicit or explicit) for why our work should improve
- Even the best results are useless if nobody understands what you are trying to solve

So, spell out your research questions!

Evaluation methodology

Feedback
- Implicit
- Explicit

Mode
- Retrospective
- Prospective

Feedback acquisition

We want to know
- What users consider relevant

We can observe
- What users tell us (explicit feedback)
- What users do (implicit feedback)

These are *noisy* measurements

Evaluation mode

Prospective experiments
- How well can we predict future preferences?

Benchmarked using live user interactions
- Poorly reproducible
- Highly realistic

Evaluation mode

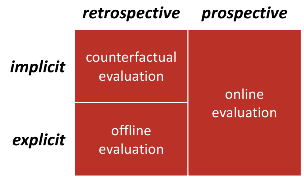
Retrospective experiments
- How well can we predict (hidden) past preferences?

Benchmarked using static test collections
- Highly reproducible
- Poorly realistic

Evaluation methodology

Feedback
- Implicit
- Explicit

Mode
- Retrospective
- Prospective



Public test collections

Text REtrieval Conference
- TREC has collections on Web, blog, tweet, video, question-answering, legal documents, medical records, chemicals, genomics, … search
  http://trec.nist.gov/tracks.html
  http://trec.nist.gov/data.html

You can build your own

Three core components
- A corpus of documents
- A set of users' queries
- A map of users' relevance assessments

You can build your own

Document corpus
- Go crawl it!

Queries
- The more the better (e.g., at least 50)
- Representative of the population (e.g., from a log)

Relevance judgments

How to judge relevance?

Who does it?
- Hired judges? Volunteers? Experts?

What are the instructions?
- Short queries? Long narratives?

What is the level of agreement?
- Redundancy to counter subjectivity

What to judge for relevance?

Exhaustive assessment is not practical
- Alternative: document sampling

Stratified sampling via pooling
- Top k results from m rankers merged
- Unique (up to km) results submitted for judgment

Generally robust for evaluating new rankers

Reference comparisons (aka baselines)

My method achieves 0.9 precision
- Meaningless without a reference comparison
- Rephrasing: is it better or worse?

Choice of baseline depends on the hypothesis
- Key question: what are you trying to show?

Choosing baselines

Vanilla baselines
- Have the proposed effect turned off
  e.g., language modeling without smoothing

Competing baselines
- Exploit the proposed effect in a different manner
  e.g., alternative smoothing technique

Choosing baselines

Try to stay "within the same framework"
- In our smoothing example: language modeling
- Should we compare to a vector space model?

Aim for the state-of-the-art
- In our case, Dirichlet smoothing

What if no baseline exists (e.g., for new tasks)?

Parameter tuning

Your method may have parameters
- Your baselines may also have parameters
  e.g., μ for Dirichlet smoothing

Which parameters need tuning?
- Which can stay fixed?
- How to tune?

Analysis of results

Measure, compare, slice and dice results
- Helps prove (or disprove) your hypotheses
- Demonstrates how your methods or systems compare against the existing state-of-the-art
- Provides fundamental insights into the underlying research problems being addressed

Evaluation metrics

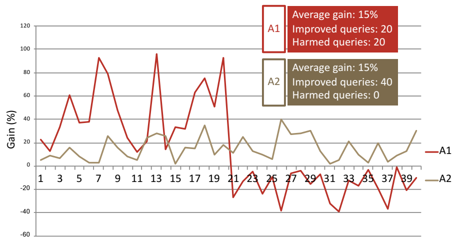
General form: $\Delta(R, G)$
- R: ranking produced by model f for query q
- G: ground-truth produced for query q

Metrics should be chosen according to the task
- Web search (precision) vs. legal search (recall) (more on next class)

Results significance

Effectiveness varies across queries
- Large average improvement may not be consistent
- Might improve a lot on some queries, hurt on many

Variable effectiveness



Results significance

Effectiveness varies across queries
- Large average improvement may not be consistent
- Might improve a lot on some queries, hurt on many

Improvements should be tested for significance
- Statistical significance (see next class)
- Practical significance

Deeper analyses

My method beats the baseline…
- *… phew, let's call it a victory and go home!* **#NOT**

Deeper analyses may provide further insights
- Why the method works
- When the method works
- And when it doesn't!

Deeper analyses

Parameter sensitivity analysis
- How sensitive is the method to its parameters?

Breakdown analysis
- How does it perform for different queries?

Failure analysis
- What are the main reasons for failure?

Summary

Experimentation drives search innovation
- Experiments should be economically practical
- Experiments should be scientifically rigorous
- Experiments should be reproducible
- Experiments should provide insights

References

Experimental methods for information retrieval
Metzler and Kurland, SIGIR 2012

Introduction to Information Retrieval, Ch. 8
Manning et al., 2008

Search Engines: Information Retrieval in Practice, Ch. 8
Croft et al., 2009

UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Coming next…

Offline Evaluation

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br