Evaluation in IR

Introduction to Information Retrieval
Informatics 141 / CS 121
Donald J. Patterson

Content adapted from Hinrich Schütze
http://www.informationretrieval.org
Evaluation in IR

Outline

- Intro to Evaluation
- Standard Test Collections
- Evaluation of Unranked Retrieval
- Evaluation of Ranked Retrieval
- Assessing relevance
- Broader perspectives
- Result Snippets
Intro to Evaluation

- There are many implementation decisions to be made in an IR system
  - Crawler
    - Depth-first or breadth-first?
  - Indexer
    - Use zones?
    - Which zones?
    - Use stemming?
    - Use multi-word phrases? Which ones?
Intro to Evaluation

- There are many implementation decisions to be made in an IR system
  - Query
    - Ranked Results?
    - PageRank?
    - Which formula do we use in the TF Matrix?
    - Should we use Latent Semantic Indexing?
  - How many dimensions should we reduce?
Intro to Evaluation

- There are many implementation decisions to be made in an IR system
- Results
  - How many do we show?
  - Do we show summaries?
  - Do we group them into categories?
  - Do we personalize the rankings?
  - Do we display graphically?
Intro to Evaluation

• How can we evaluate whether we made good decisions or not?
• Measure them
Measures for a search engine

- How fast does it index?
  - Number of documents per hour
  - Average document size
- How fast does it search
  - Latency as a function of index size
- Expressiveness of query language
  - Ability to express complex information needs
  - Speed on complex queries
Measures for a search engine

- We can measure all of these things:
  - We can quantify size and speed
  - We can make this precise
- What about user happiness?
  - What is this?
  - Speed of response/size of index are factors
  - But fast, useless answers won’t make a user happy
- Need to quantify user happiness also.
Measuring user happiness

• Issue: Who is the user we are trying to make happy?
  • It depends.
Measuring user happiness

• Issue: Who is the user we are trying to make happy?
• Web engine:
  • The user finds what they want.
  • Measure whether or not they come back.
Measuring user happiness

- Issue: Who is the user we are trying to make happy?
  - eCommerce Site
    - User finds what they want
    - Are we interested in the happiness of the site?
    - Are we interested in the happiness of the customer?
    - Measure the $$ of sales per user
    - Measure number of transactions per user
    - Measure time to purchase
    - Measure conversion rate (lookers -> buyers)
Measuring user happiness

- Issue: Who is the user we are trying to make happy?
  - Enterprise site
    - Are the users “productive”?
    - Measure time savings when using site
    - Measure “things accomplished”
    - careful about confounding factors
  - Measure how much a user utilizes the site’s features
Measuring user happiness

- Can we measure happiness?
- Do we want to measure happiness?
- What are some proxies for happiness?
  - Relevance of search results
    - How do we measure relevance?
Measuring Relevance Instead

- What do we need to measure relevance?
  - A document collection, a test corpus
  - A set of queries, benchmark queries
  - A set of answers, a gold standard
    - i.e., Document, d, {is, is not} relevant to query q
  - Alternatives to binary exist, but atypical
- Cross-validation methodology
  - Parameter tuning
Information need

• Remember the user has an information need
  • not a query
• Relevance is assessed relative to the information need, not the query
  • e.g., I am looking for information on whether drinking red wine is more effective than eating chocolate at reducing risk of heart attacks
• Query: red wine heart attack effective chocolate risk
• Does the document address the need, not the query
Evaluation in IR

Relevance benchmarks

- TREC - National Institute of Standards and Testing (NIST) has run a large IR test bed for many years
- Reuters and other benchmark document collections
- Retrieval tasks which are specified
  - sometimes as queries
- Human experts mark, for each query and for each document
  - Relevant or Irrelevant