Best Practices for Test Collection Creation and Evaluation Methodologies

Paul Clough
Department of Information Studies
University of Sheffield, UK

Evaluating search

- Evaluation is important for designing and developing effective search systems
- Strong focus on measuring the effectiveness of an IR system
- Test collections have been major evaluation resource in the academic community
- But test collections aren't the be all and end all!

Aims of this best practice report

- Act as a guide for evaluating search systems
  - Necessary resources, procedures and methods
- Aims to "bridge the gap" between
  - Wide range of published research by the academic community
  - Very little material produced for users, administrators and developers of IR systems
- This is a report aimed at practitioners

Test collections

- Test collections provide re-usable resources to evaluate IR systems in particular operational settings
- Typically consist of
  - Collection of documents
  - Set of representative queries (topics)
  - Set of relevance judgments for each topic
  - Evaluation measures

When building test collections ….

- What is the purpose of the evaluation?
- What resources are available to conduct the evaluation?
- What sort of searching is typically conducted on the search engine under test?
- What do you know about the IR system being tested?

Practical guidelines

- Gathering a collection of documents
- Generating a suitable set of queries/topics
  - How do I obtain the queries/topics?
  - How many queries/topics do I need?
- Creating the relevance assessments
  - How do I gather the assessments?
  - Who should do the assessments?
  - How many assessments should be made?
  - What are the assessors expected to do?
  - What about finding missing relevant documents?

Evaluation measures

- Measure of IR system effectiveness
  - Provides simple simulation of user behaviour
- Common measures
  - Precision at fixed ranking (e.g. P10)
  - Mean average precision (MAP)
  - Graded relevance measures (e.g. DCG)
- Comparing results
  - From multiple runs for one system
  - From single runs from different systems
  - Use of significance tests

Two case studies

- Text REtrieval Conference (TREC)
  - Organised by NIST (in the US)
  - Example of how evaluation is done in the academic community
- The UK National Archives (TNA)
  - UK government's official archive (non-academic)
  - Emphasise on developing efficient and cost-effective test collection resources

Summary

- Evaluating search is very important both in academic and commercial contexts
- Evaluation often done using test collections
- This report provides practical guidelines for evaluating in an efficient and cost-effective way
  - Aims to bridge the gap between research in the academic community and the needs of practitioners
- Download: http://www.trebleclef.eu/

The University Of Sheffield.

To Discover And Understand.