Query difficulty estimation

Query difficulty estimation

Attempt to estimate the quality of search results for a query from a given collection of documents in the absence of user relevance feedback

Motivations

Understanding what constitutes an inherently *difficult* query is important

Even for good systems, the quality for some queries can be very low

Benefits

Can inform users that it is a difficult query

=> they can then
remodel/reformulate query
or submit query elsewhere

Benefits

Can inform system that it is a difficult query

=> can then adopt a different strategy:

- query expansion
- log mining
- incorporate collaborative filtering or other evidence

Benefits

Can inform system administrator that it is a difficult query

=> improve collection

Benefits

Can also help within specific IR domains

e.g. Merging results in distributed IR

Robustness problem in IR

Most IR systems exhibit large variance in performance in answering users' queries

Many causes:

- the query itself (ambiguous terms)
- vocabulary mismatch problem
- *missing content* queries

Robustness problem in IR

Many issues with types of failures in queries

- failure to recognise all aspects in the query
- failure in pre-processing
- over-emphasis on a particular aspect/term
- query needs expansion
- need analysis to identify intended meaning of query (NLP)
- need better understanding of proximity relationship among terms

TREC robust track

-50 of the most difficult topics from previous TREC runs selected

-New measures of performance adopted to explicitly measure robustness

-Human experts asked to categorise topics/queries as *easy, medium, hard*

-Low correlation between humans and systems (PC = 0.26)

-Also, relatively low correlation among humans (PC = 0.39)

More recent work illustrating the same phenomena

TREC robust track – prediction task

Systems were challenged to predict the difficulty of a query and then perform retrieval.

The predicted values were compared to actual values

Very poor prediction ability; many systems exhibited negative correlation

Difficulty across collections?

-Difficult query for collection 1 may not be as difficult for collection 2

-However, relative difficulty largely maintained

Recap: some basic concepts

Precision measures:

Precision at k (P@k)

AP(q)

Retrieval task:

Given q and D, retrieve Dq (result list)

Goal:

Wish to estimate the quality/usefulness/retrieval performance of Dq in satisfying the user's information need represented as q

=> Predict AP(q) when no relevance information provided

Recap: some basic concepts

Quality of performance indicator can be measured by comparing AP(q) with estimated AP (q)

Can measure correlation using:

- Pearson
- Spearman rank

Query difficulty estimation ?

Approaches?

Approaches?

Can be categorised as:

- pre-retrieval approaches
  - estimate difficulty without running the system

- post-retrieval approaches
  - run the system against query and examine results

Linguistic approaches (pre-retrieval)

-Use some NLP approaches to analyse query

-Use external sources of information to identify ambiguity etc.

Most linguistic features do not correlate well with performance

Statistical approaches (pre-retrieval)

-Take into account the distribution of the query term frequencies in the collection

- e.g., consider idf and icf of terms

-Take into account *specificity* of terms

-Queries containing non-specific terms are considered difficult

Statistical approaches (pre-retrieval)

Term relatedness

If query terms co-occur frequently in collection, we expect good performance

Mutual information or Jaccard coefficient etc. can be used

Statistical approaches (pre-retrieval)

Query scope

what percentage of documents contain at least one query term, if a lot then this is probably a difficult query

Simplified query scope

Measures difference between language model of collection with language model of query

Approaches
(post retrieval-retrieval)

Three main categories:

- Clarity measures
- Robustness
- Score analysis

Clarity
(post retrieval-retrieval)

Attempts to measure the coherence in the result set

The language of the result set should be distinct from the rest of the collection

Compare language model induced from answer set and one induced from the corpus/collection

Related to the cluster hypothesis

Robustness
(post retrieval-retrieval)

Explores robustness of system in the face of perturbations to:

i) Query
   Overlap between query and sub-queries. In difficult queries some terms have little or no influence
ii) Documents
   Compare system performance against collection C and some modified version of C
iii) Retrieval performance
   Submit same query to many systems over same collection; divergence in results tells us something about difficulty of query

Score analysis
(post retrieval-retrieval)

Analyse score distributions in returned ranked list:

- difficulty can be measured based on distribution of values; is cluster hypothesis supported?

- can look at distribution of scores in answer set and document set and attempt to gauge difficulty

- relatively simple measures shown to be effective

Exercises

We have seen many alternative approaches to predicting difficulty; can you identify an approach to combining them to make another prediction approach?

In this class, we have considered prediction of difficulty of queries in adhoc retrieval. Can you identify approaches that may of use in:

i) Predicting a difficult ‘user’ in collaborative filtering
ii) Predicting whether a query expansion technique has improved the results