# Query characterisation in Information Retrieval: performance, difficulty, uncertainty and rank fusion

Alejandro Bellogín Kouki
Universidad Autónoma de Madrid
alejandro . bellogin @ uam . es

June 18, 2008

# Introduction

## Important question

What is uncertainty and how can it be measured

## Objectives

- State-of-the-art in query characterisation
- Different techniques for modeling uncertainty
- Dynamic rank fusion?

# Utility

- User: rephrasing the query.
- Retrieval system: retrieval consistency (distinguishing poorly performing queries based on performance prediction techniques). The retrieval system can invoke alternative retrieval strategies for different queries (query expansion or different ranking functions).
- System administrator: identify difficult queries for the search engine, and expand the collection of documents to better answer. Simple evaluation.
- Distributed information retrieval: decide which search engine to use.

# Definitions

- Query clarity
- Query difficulty
- Query scope
- Ranking robustness
- Query hardness

# Definitions

## Query clarity

Cronen-Townsend *et al.* [CTZC02] defined the **query clarity** as a degree of (the lack of) the **query ambiguity**.

- Query ambiguity as *the degree to which the query retrieves documents in the given collection with similar word usage*.
- Degree of dissimilarity between the language usage associated with the query and the generic language of the collection as a whole.
- It is equivalent to the relative entropy, or Kullback-Leibler divergence, between the query and collection language models.
- System-independent
- Formulation:

$$\begin{aligned} P(w|Q) &= \sum_{D \in R} P(w|D)P(D|Q), \quad P(Q|D) = \prod_{q \in Q} P(q|D) \\ P(w|D) &= \lambda P_{ml}(w|D) + (1 - \lambda) P_{coll}(w) \\ \text{clarity score} &= \sum_{w \in V} P(w|Q) \log_2 \frac{P(w|Q)}{P_{coll}(w)} \end{aligned}$$

with $w$ any term, $Q$ the query, $D$ a document or the model, $R$ is the set of documents that contain at least one query term, $P_{ml}(w|D)$ is the relative frequency of term $w$ in document $D$, $P_{coll}(w)$ is the relative frequency of the term in the collection as a whole, $\lambda$ is a parameter (in their work, $0.6$) and $V$ is the entire vocabulary.

# Definitions

## Query difficulty

In [ACR04], Amati *et al.* proposed the notion of **query difficulty** to predict query performance.

- Amount of information $\mathrm{Info}_{\mathrm{DFR}}$ gained after a first-pass ranking: if there is a significant divergence in the query-term frequencies before and after the retrieval, then the authors make the hypothesis that this divergence is caused by a query which is easy-defined
- $\mathrm{Info}_{\mathrm{DFR}}$ is defined as

$$\mathrm{Info}_{\mathrm{DFR}} = \sum_{t \in Q} -\log_2 \mathrm{Prob}(\mathrm{Freq}(t|\mathrm{TopDocuments})|\mathrm{Freq}(t|\mathrm{Collection}))$$

- System-dependent

# Definitions

## Query scope

Plachouras *et al.* [POvRC03, HO04, MHO05] defined the **query scope** as a measure of the **specificity** of a query: $-\log\left(N_Q/N\right)$, where $N_Q$ is the number of documents containing at least one of the query terms, and $N$ is the number of documents in the whole collection.
The authors found that query scope is effective in inferring query performance for short queries in ad-hoc text retrieval.

## Example (Application to Dempster-Shafer)

Another application of query scope can be found in [VI05], where it is used for assigning a measure of uncertainty to each source of evidence (in their work these sources were content analysis and link structure analysis) and then applying Dempster-Shafer's theory of evidence.

# Definitions

## Ranking robustness

**Ranking robustness** [ZC06] refers to a property of a ranked list of documents that indicates how stable the ranking is in the presence of **uncertainty** in the ranked documents.

- It is inspired by a general observation in noisy data retrieval that the degree of ranking robustness against noise is positively correlated with retrieval performance.
- Regular documents also contain *noise* if we interpret noise as *uncertainty*.
- This robustness score performs better than or at least as good as the clarity score.
- Collection-dependent

# Definitions

## Query hardness

Definition of the **query hardness** by Aslam and Pavlu in [AP07]: results returned by multiple retrieval engines will be relatively similar for *easy* queries but more diverse for *difficult* queries.
They distinguish two notions of query hardness:

System query hardness difficulty of a query for a given retrieval system run over a given collection. To capture the difficulty of the query for a particular system, run over a given collection. It is system-specific.

Collection query hardness difficulty of a query with respect to a given collection. Capturing the inherent difficulty of the query (for the collection) and perhaps applicable to a wide variety of typical systems. It is independent of any specific retrieval system.

# Different models

- Numeric representations:
  - Probability measures
  - Dempster-Shafer belief functions
  - Possibility measures
  - Ranking functions
- Nonnumeric representations:
  - Plausability measures

## Why so many models?

- Either one event is more probable than the other, or they have equal probability. It is impossible to say that two events are comparable in likelihood.
- The numbers are not always available

# Fuzzy representation

## Fuzzy models applied to IR

- Extended boolean models: fuzzy document representation
- Extended Boolean models: fuzzy extensions of the query language
- Fuzzy Thesauri of terms
- Fuzzy Clustering of Documents

# Dempster-Shafer representation I

## Using this theory

- The set of elements in which we are interested is called the frame of discernment
- When two bodies of evidence are defined in the same frame of discernment, we can combine them using Dempster's combination rule, under the condition that the two bodies are independent of each other.
- The rule of combination of evidence returns a measure of agreement between two bodies of evidence.

## In IR I [VI05]

- The frame of discernment is the set of Web documents in the collection
- The scoring functions (content analysis, link structure analysis) are the bodies of evidence that will be combined into a single body of evidence in the frame of discernment.
- Vassilis and Iadh found that Dempster-Shafer theory of evidence is not effective in significantly improving precision (quality of the sources of evidence or method?)

# Dempster-Shafer representation II

## In IR II [Lal98]

Lalmas uses the Dempster-Shafer theory to express a four-featured model in two steps:

- The initial Dempster's theory: to represent structure and significance
- The refinement function (Shafer): a possible method for representing partiality and uncertainty

Lalmas proposes:

- The different representations of the document capture the partiality of information.
- The transformed documents are not actual documents, but consist of more exhaustive representations of the original document.
- The transformation may be uncertain.
- A document that requires less transformations than another one is usually more relevant to the query than the other document.

# Performance prediction

## Measuring the quality of the performance prediction methods

Compare the rankings of queries based on their actual precision (such as MAP) with the rankings of the same queries ranked by their performance scores

Classification:

- Based on necessity of retrieval results
    - Non-retrieval
    - Pre-retrieval
    - Post-retrieval
- Based on training
    - Trained predictors
    - Untrained predictors

# Non-retrieval approaches

Mothe *et al.* [MT05] extract 16 features of the query and study their correlation with respect to recall and average precision. In this study they used TREC 3, 5, 6 and 7 as datasets.

- The only positively correlated feature is the number of proper nouns
- Many variables do not have significant impact on any evaluation measure. Only the more *sophisticated* features appear more than once
- The only two variables found correlated in more than one TREC campaign are the average syntactic links span (for precision) and the average polysemy value (for recall)

# Pre-retrieval approaches I

## Basics

- These predictors do not rely on the retrieved document set.
- The efficiency is often high since the performance score can be computed prior to the retrieval process.
- These predictors generally have a low prediction accuracy since many factors related to retrieval effectiveness are not exploited

# Pre-retrieval approaches II

## Examples

- IDF-related:
    - He and Ounis [HO04] proposed a predictor based on the standard deviation of the IDF of the query terms.
    - Plachouras [PHO04] represented the quality of a query term by Kwok's inverse collection term frequency.
- Diaz and Jones [DJ04] have tried time features for prediction (together with clarity scores improves prediction accuracy).
- Kwok *et al.* [KGSD04] built a query predictor using support vector regression.
- He and Ounis [HO04] proposed the notion of query scope

# Post-retrieval approaches I

## Basics

- These predictors make use of retrieved results in some manner.
- Techniques in this category provide better prediction accuracy.
- Computational efficiency can be an issue for many of these techniques.

# Post-retrieval approaches II

## Examples

- Jensen *et al.* [JBG$^+$05] trained a regression model with manually labeled queries using visual features, such as titles and snippets
- Elad Yom-Tov *et al.* [YTFCD05] proposed a histogram-based predictor and a decision tree based predictor (features: the document frequency of query terms and the overlap of top retrieval results between using the full query and the individual query terms).
- Clarity score
- Amati [ACR04] proposed to use the KL-divergence (as one possible probabilistic model) between a query term's frequency in the top retrieved documents and the frequency in the whole collection
- He and Ounis [HO04] proposed a simplified version of the clarity score where the query model is estimated by the term frequency in the query
- Carmel *et al.* [CYTDP06] found that the distance measured by the Jensen-Shannon Divergence (JSD) between the retrieved document set and the collection
- Vinay *et al.* [VCMFW06] propose four measures to capture the geometry of the top retrieved documents for prediction. The most effective measure is the sensitivity to document perturbation
- Kwok *et al.* [KGDD05] suggest predicting query performance by retrieved document similarity.
- Grivolla *et al.* [GJM05] calculate the entropy and pairwise similarity (of the set of the K top-ranked documents for a query)
- Diaz [Dia07] proposes a technique called spatial autocorrelation (degree to which the top ranked documents receive similar scores by spatial autocorrelation)
- Zhou *et al.* [ZC07] defined Weighted Information Gain (WIG) (it measures the change in information about the quality of retrieval (in response to query $Q_i$) from an imaginary state that only an average document is retrieved to a posterior state that the actual search results are observed) and Query Feedback (QF) (it measures the degree of corruption that arises when $Q$ is transformed to $L$, output of the channel when the retrieval system is seen as a noisy channel).

# First results I

| Queries | Pearson | Spearman | Kendall |
|---|---|---|---|
| All | Proper nouns (0.2305), hyponymy (−0.1808), polysemy (−0.1933), normalized polysemy (−0.2799) | Proper nouns (0.2103), polysemy (−0.2089), normalized polysemy (−0.2506) | Proper nouns (0.1726), polysemy (−0.1414), normalized polysemy (−0.1685) |
| TREC 8 | Proper nouns (0.2857), syntactic depth (−0.1201) | Proper nouns (0.3360), syntactic depth (−0.0275) | Proper nouns (0.2772), syntactic depth (−0.0211) |
| TREC 9 | Proper nouns (0.2978), hyponymy (−0.3084), normalized polysemy (−0.3218) | Normalized polysemy (−0.3445), normalized hyponymy (−0.3099) | Normalized hyponymy (−0.2177), normalized polysemy (−0.2276) |
| TREC 2001 | Acronyms (0.3626) | Acronyms (0.2814) | Acronyms (0.2320) |

Table: Linguistic features found statistically significant correlated with average precision (correlation in parenthesis, the greater absolute value, the more dependance between variables)

# First results II

| Queries | Pearson | Spearman | Kendall |
|---|---|---|---|
| All | SCS (0.2615), clarity (−0.2154) | SCS (0.3519), clarity (−0.3005) | SCS (0.2361), clarity (−0.2003) |
| TREC 8 | Scope (0.4771), SCS (0.6037) | Scope (0.3248), SCS (0.4919), clarity (−0.3268) | Scope (0.2640), SCS (0.3339), clarity (−0.2327) |
| TREC 9 |  | SCS (0.4402) | SCS (0.3011) |
| TREC 2001 | Clarity (−0.4822) | Clarity (−0.4452) | Clarity (−0.3004) |

Table: Non-linguistic features found statistically significant correlated with average precision (correlation in parenthesis, the greater absolute value, the more dependance between variables)

# No training data

- IDF-related features as predictors: standard deviation of the IDF of the query terms [HO04], Kwok's inverse collection term frequency [PHO04]
- Related to the ideas in the clarity score technique:
    - KL-divergence between a query term's frequency in the top retrieved documents and the frequency in the whole collection [ACR04]
    - Simplified version of the clarity score (query model is estimated by the term frequency in the query) [HO04]
    - Percentage of documents that contain at least one query term in the collection (query scope) [HO04]
    - Clarity scores extended to include time features [DJ04])
    - Predict query performance by retrieved document similarity [KGDD05].

# With training data

- Histogram-based predictor and a decision tree based predictor (features: document frequency of query terms and the overlap of top retrieval results between using the full query and the individual query term) [YTFCD05]
- Using support vector regression (features: the best three terms in each query, their log document frequency and their corresponding frequencies in the query) [KGSD04]
- Regression model with manually labeled queries to predict precision at the top 10 documents (visual features from a surrogate document representation of retrieved documents) [JBG$^+$05]

# Experiments

## Prospective experiments

1. Comparison between linguistic and non-linguistic performance predictors, correlations found between average precision and these predictors.
2. Implementation of clarity-driven personalisation model.
3. Given a set of queries for testing, they are clustered according to their clarity value, and these clusters are used to discriminate which scores have to be taken into account when the source distribution is being build.
4. Use the clarity score to weight each source according to the clarity each one assigns to the query.

# More results

| Method | TREC 8 | TREC 9 | TREC 2001 |
|---|---|---|---|
| Normal | 0.3734 | 0.1928 | 0.3273 |
| $\text{Clarity}_M^B$ | 0.4566 | 0.2511 | 0.3577 |
| $\text{Clarity}_M^W$ | 0.2942 | 0.1342 | 0.2848 |

MAP for different normalisation methods. The separation when clarity is used is given by the median value of all the query clarities involved in each track. If the superscript is *B* the cluster used in the normalisation is fromed with the less ambiguous queries (greater clarity value).

# Future work I

- Define a more general clarity score
- New paradigm for the query difficulty prediction: given a normalised distribution (from a source or a set of sources) for a given query, infer the difficulty of that query.
- Check how the clarity score behaves with a dynamic collection, like in the voting-like experiment.

# Future work II

- Improve the clarity-driven personalisation model.
- Use ambiguity predictors in order to measure similarity between users (folksonomies, user profiles, creating groups of users).
- Combine different predictors linearly or with the aid of genetic algorithms.

# Conclusions

- A lot of works have tried to resolve the problem of query characterisation in order to improve the system performance.
- This problem is far from being completely solved.
- Community is using these techniques and applying them in different fields.
- Our opinion is that it is very promising in metasearch area.
- Some baselines have been found (i.e. clarity score), but may be a change of paradigm is needed (fuzzy models, or more specialised vague modeling approaches).

# Thank you

Gracias!

Giambattista Amati, Claudio Carpineto, and Giovanni Romano.
Query difficulty, robustness, and selective application of query expansion.
*Advances in Information Retrieval*, pages 127–137, 2004.

Javed A. Aslam and Virgiliu Pavlu.
Query hardness estimation using jensen-shannon divergence among multiple scoring functions.
In *ECIR*, pages 198–209, 2007.

Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft.
Predicting query performance.
In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306, New York, NY, USA, 2002. ACM.

David Carmel, Elad Yom-Tov, Adam Darlow, and Dan Pelleg.
What makes a query difficult?

In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 390–397, New York, NY, USA, 2006. ACM.

Fernando Diaz.
Performance prediction using spatial autocorrelation.
In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 583–590, New York, NY, USA, 2007. ACM.

Fernando Diaz and Rosie Jones.
Using temporal profiles of queries for precision prediction.
In *SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 18–24. ACM Press, 2004.

Grivolla, Jourlin, and De Mori.
Automatic classification of queries by expected retrieval performance.

In *Predicting Query Difficulty - Methods and Applications, SIGIR 2005*, 2005.

Ben He and Iadh Ounis.
Inferring query performance using pre-retrieval predictors.
In *String Processing and Information Retrieval, SPIRE 2004*, pages 43–54, 2004.

Eric C. Jensen, Steven M. Beitzel, David Grossman, Ophir Frieder, and Abdur Chowdhury.
Predicting query difficulty on the web by learning visual clues.
In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 615–616, New York, NY, USA, 2005. ACM.

K. L. Kwok, L. Grunfeld, N. Dinstl, and P. Deng.
Trec 2005 robust track experiments using pircs.
In *Online Proceedings of 2005 Text REtrieval*, 2005.

K. L. Kwok, L. Grunfeld, H. L. Sun, and P. Deng.

Trec 2004 robust track experiments using pircs.
In *Online Proceedings of 2004 Text REtrieval*, 2004.

Mounia Lalmas.
Information retrieval and dempster-shafer's theory of evidence.
In *Applications of Uncertainty Formalisms*, pages 157–176.
Springer-Verlag, 1998.

Craig Macdonald, Ben He, and Iadh Ounis.
Predicting query performance in intranet search.
In *Predicting Query Difficulty - Methods and Applications, SIGIR 2005*, 2005.

Josiane Mothe and Ludovic Tanguy.
Linguistic features to predict query difficulty.
In *Predicting Query Difficulty - Methods and Applications, SIGIR 2005*, 2005.

V. Plachouras, B. He, and I. Ounis.

University of Glasgow at TREC2004: Experiments in Web, Robust and Terabyte tracks with Terrier.
In *Proceeddings of the 13th Text REtrieval Conference (TREC 2004)*, 2004.

Vassilis Plachouras, Iadh Ounis, Cornelis J. van Rijsbergen, and Fidel Cacheda.
University of glasgow at the web track: Dynamic application of hyperlink analysis using the query scope.
In *TREC*, pages 646–652, 2003.

Vishwa Vinay, Ingemar J. Cox, Natasa Milic-Frayling, and Ken Wood.
On ranking the effectiveness of searches.
In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 398–404, New York, NY, USA, 2006. ACM Press.

Plachouras Vassilis and Ounis Iadh.

Dempster-shafer theory for a query-biased combination of evidence on the web.
*Information Retrieval*, 8(2):197–218, April 2005.

Elad Yom-Tov, Shai Fine, David Carmel, and Adam Darlow.
Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval.
In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 512–519, New York, NY, USA, 2005. ACM Press.

Yun Zhou and Bruce W. Croft.
Ranking robustness: a novel framework to predict query performance.
In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 567–574, New York, NY, USA, 2006. ACM.

Yun Zhou and Bruce W. Croft.

Query performance prediction in web search environments.
In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 543–550, New York, NY, USA, 2007. ACM.