Robust track: Failure analysis, evaluation, and approaches

Jacques Savoy, Samir Abdou
University of Neuchatel, Switzerland
iiun.unine.ch

Context

The robust track focus on *difficult* topics.

Our prior feelings …

Why do we need to worry for just a few cases? (perhaps 1% to 5%?)

1) Because our search systems are based on a set of very reasonable assumptions in IR

IR assumptions

- Stopword list (463 words)
  words with no "meaning" (the, in, is)
- Stemming
  - Inflections (cats → cat)
  - Derivational (reliability → reliable)
- tf: more importance to frequent terms
- df: less weight to terms appearing in many documents
- Prefer short documents

Context

The robust track focus on *difficult* topics.

Our prior feelings …

Why do we need to worry for just a few cases? (perhaps 1% to 5%?)

1) Because our search systems are based on a set of very reasonable assumptions in IR
2) Recent IR models demonstrate high performance

IR Models

Probabilistic
- Okapi
- GL2 (*Divergence from Randomness*)
- Language Model (LM)

Vector-space
- Lnu-ltc
- *tf·idf*

Various evaluation campaigns demonstrate that such approaches are really effective.
But which one performs the best?

French Evaluation (91 queries)

Evaluation on the entire collection and with the test set

Model	MAP	
Query	T	TD
Okapi	**0.3969**	**0.4816**
GL2	0.3742*	0.4714*
LM	0.3611*	0.4535*
Lnu-ltc	0.3669*	0.4518*
tf·idf	0.2447*	0.2988*

Mean average precision

The mean is a good way to summarize a sample of values.

But it hides irregularities between queries.

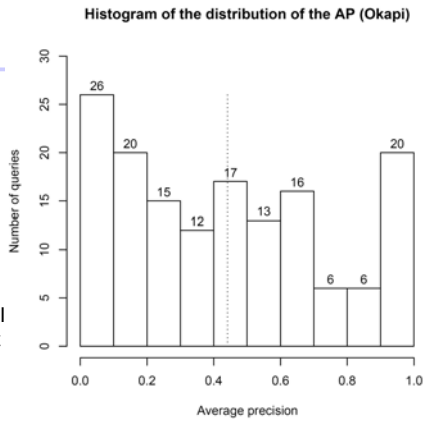
MAP

A single value MAP: 0.4412 or ...

For 20 queries, the perfect answer.

For 26, Okapi "fails".

Having 1 or 2 rel item(s) does not mean that the query is "hard"



Average precision

Now look at the meaning of a *single AP* (the performance achieved by a given query)

Topic #71 « Vegetables, Fruit and Cancer » (three relevant documents)

From T and applying PRF, AP varies from 0.6759 to 0.4175

What is the meaning

a) of these values for the user?

b) the difference between them?

AP: Does a real user see the difference?

rank	Okapi (A)		Okapi & PRF (B)	
1	R	1/1	nR	
2	R	2/2	R	1/2
3	nR		R	2/3
...	nR		nR	
35	nR		R	3/35
...	nR		nR	
108	R	3/108	nR	
	AP =	0.6759	AP =	0.4175
				-38.2%

MAP or GMAP

With the MAP:

If the AP of topic A increases from 0.6 to 0.62,

we have the same effect on the MAP

if AP of topic B increases from 0.02 to 0.04.

Here we prefer the second case (improvement over difficult queries).

MAP or GMAP

And the geometric mean (GMAP)?

$$\text{MAP} = \frac{1}{n}\sum_{i=1}^{n} AP_i$$

$$\text{GMAP} = \sqrt[n]{\prod_{i=1}^{n} AP_i} = e^{\frac{1}{n}\cdot\sum_{i=1}^{n} log(AP_i)}$$

The idea: improvement of poor queries has a greater impact on the GMAP (emphasis on AP close to 0.0)

Strong correlation between both measures (r=0.96).

French Evaluation (91 queries)

Model	MAP		GMAP	
Query	T	TD	T	TD
Okapi	**0.3969**	**0.4816**	**0.2121**	**0.3534**
GL2	0.3742	0.4714	0.1833	0.3316
LM	0.3611	0.4535	0.1745	0.3079
Lnu-ltc	0.3669	0.4518	*0.1941*	*0.3291*
tf·idf	0.2447	0.2988	0.0944	0.1606

MAP or GMAP?

Consider the difference between T and TD topic formulation

Topic #200: the largest improvement when considering GMAP
- T, Okapi AP: 0.0001
- TD, Okapi AP: 0.0264

but with T or TD Prec@10 is still 0.0
(first rel. item: position 65 with TD)

Failure analysis

1. System flaws
2. Topic intrinsic difficulty

It is difficult to know in advance if a given topic is "easy" or "hard".
Our definition:
A hard topic → Prec@10 = 0

Example (spelling error)

Topic #200 best AP: 0.0001
«Innondationeurs en Hollande et en Allemagne»
«Flooding in Holland and Germany»

The query is «holland», «allemagn»
with df = 244 and 8,174

Spelling error («Innondations»)

Topic #46, «Iraq» instead of «Irak»

Example (stoplist)

Topic #91 best AP: 0.0012
« AI en Amérique latine »
« AI in Latin America »

The query is «ameriqu», «latin»
with df = 2,518 and 1,353

«AI» (viewed as «ai») is included in the French stopword list (it is a verbal form of the verb "to have").

«AI» means also «ad interim» and it is the acronym of a Swiss social insurance

Example (stoplist)

Similar problem in English
- IT engineer → it engineer → engineer
- vitamin A → vitamin a → vitamin
- US citizen → us citizen → citizen

Why WestLaw® uses only one stopword?

Example (stemming)

Topic #117 best AP: 0.0193
«Elections parlementaires européennes»
«European Parliament Elections»

The forms «Europe» and «européennes» return the same stem, but not the forms «parlement» (parliament) and «parlemantaires» (parliamentary) that are indexed under two distinct stems.

Example (specificity)

Topic #51 best AP: 0.0379
«Coupe du monde de football»
«World Soccer Championship»

Many articles with «Coupe du monde de football» in the title + short documents. All are irrelevant.

The descriptive part specifies what the user wants (e.g., the final result).

#120 «Edouard Balladur» (0.0133),
#156 «Trade Unions in Europe» (0.0114), …

Improvement using…

Pseudo-relevance feedback?

Without looking at the first k=5 documents,
we assume that they are relevant.

$$Q' = \alpha \cdot Q + \beta \cdot \frac{1}{k} \sum_{i=1}^{k} D_i$$

PRF Evaluation (91 queries)

Model	Okapi	GL2	LM
Simple	0.3969 (11)	0.3752 (14)	0.3611 (14)
& PRF	3 doc/ 10 terms 0.4058 (**10**)	5 doc/ 30 terms 0.4029 (15)	5 doc/ 10 terms 0.4137 (15)

Pseudo-relevance feedback

If we have relevant documents in the first positions, it will improve the retrieval effectiveness.

But poor topics (by definition) do not have such pertinent information!

Pseudo-relevance feedback

Not always! For example, Topic #95
« Conflict in Palestine »
with 117 relevant documents

Okapi: AP = 0.059 and
Prec@10 = 0

Okapi + PRF: AP =0.1383 (+130%)

Improvement using…

Data fusion

If one search model fails for a given query, another may provide a better answer.
We combined Okapi, GL2, and LM.

- Round-robin
- Sum RSV
- Normalize RSV
- Z-score

Data Fusion Eval. (91 queries)

IR Model	Okapi	GL2	LM
Single	0.3969 (11)	0.3752 (14)	0.3611 (14)
& PRF	0.4058 (**10**)	0.4029 (15)	0.4137 (15)

Data fusion	Single	& PRF
Round-Robin	0.3845 ↓(15)	0.4205 (16)
SumRSV	0.3851 ↓(16)	0.4313 (18)
Normalize	0.3825 ↓(16)	0.4385 (16)
Z-score	0.3822 ↓(16)	**0.4392** (15)

Improvement using…

Other document collections?

But thematic, time and cultural differences
- *Financial Times* vs. *The Sun*
- News from 1994-95 vs. 2006
- Freely available vs. $, £ or €
- *Le Devoir* (Montreal), *Le Monde* (France) or *Le Temps* (CH)

Improvement using…

Cultural difference

Mobile phone? (Topic #155)

- « *Natel* » in Switzerland
- « *Cellulaire* » in Quebec
- « *Téléphone portable* » in France
- « *Téléphone mobile* » in Belgium

Improvement using…

We have used YAHOO! FRANCE
- Send the T query
- Extract the top k=10 snippets
- Add them to the query (k=100 + 40 pages in [Kwok *et al.* TREC 2004])

Mean query size increases
- T: 2.91 distinct search terms
- TD: 7.51 distinct search terms
- T+Yahoo: 112.56 distinct search terms

Web – Evaluation (91 queries)

Query	T	TD	T-Yahoo	TD-Yahoo
Okapi	0.3969	0.4816	0.4160	0.4354
+PRF 3/15	0.4014 +1.1%	0.4993 +3.7%	**0.4217*** +1.4%	**0.4411*** +1.3%
+PRF 5/15	**0.4141** +4.3%	**0.5035*** +4.5%	0.4198 +0.9%	0.4393 +0.9%

Summary

MAP and number of queries with P@10=0

Query	T	TD
Okapi	0.3969 (11)	0.4816 (5)
+PRF (5/15)	0.4141 (11)	
+ data fusion	**0.4392** (15)	
Yahoo	0.4160 (**8**)	
Yahoo+PRF	0.4217 (9)	

Conclusion: query-by-query

- Robust is a real concern (practical / to improve the MAP)
- The mean hides irregularities
- Hard topics: why?
  - System flaws
  - Topic intrinsic difficulty

Conclusion: Measurement

- The MAP is perhaps not the best measure
- Geometric MAP knows also some problems
- A measure that the user may understand, P@10, or GS@10 (see Hummingbird's paper, CLEF-2006)?

Conclusion: How?

- Query expansion using
  - Pseudo-relevance feedback?
  - Data fusion (Z-score)?
  - Other text collections?
  - The Web (via search engines or specific web site), interesting for short queries