Axiomatic Evaluation of IR
M2R - MOSIG

Philippe Mulhem
Philippe.Mulhem@imag.fr

Outline

- Evaluation in IR
- Major axioms
- Study of some axioms on DIR model
- Usages of axioms
- Conclusion

Introduction

- In a previous lesson, I said that IR evaluation is only black-box based (Cranfield Paradigm)

- In fact this is not fully true :
  - in 2004, Fang, Tao Tao and Zhai proposed a first way to axiomatize comparisons of IR
    - ➔ at a formula level and not at a system level ➔

Introduction

- Fact : tests shows that very different matching lead to similar evaluations [Fang et al 2011]
  - MAP for short queries

	AP	DOE	FR	ADF	Web	Trec7	Trec8
Piv	0.23	0.18	0.19	0.22	0.29	0.18	0.24
DIR	0.22	0.18	0.18	0.21	0.30	0.19	0.26
BM25	0.23	0.19	0.23	0.19	0.31	0.19	0.25
DFR-PL2	0.22	0.19	0.22	0.19	0.31	0.18	0.26

Introduction

- Ground
  - According to years of research on IR models, the matching between a query and a document *f(d,q)* depends on
    - tf : term frequency of query terms (in query c(w, q) and in document c(w,d))
    - df (or idf) : document frequency (idf(w))
    - document length (|d|)
    - Interaction between these elements

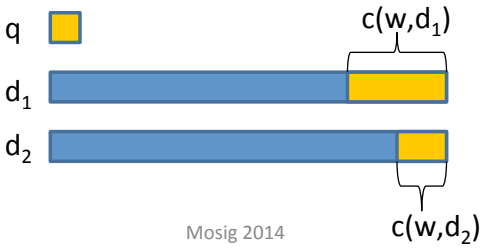
Major axioms [Fang et al, Clinchant & Gaussier]

- Term frequency constraints
  - TFC1, TFC2, TFC3
- Lenght normalization constraints
  - LNC1, LNC2
- Term discrimination contraint
  - TDCspe
- Term Frequency – Lenght constraint
  - TF-LNC

Term Frequency contraints

- TFC1 :
  - Idea : Give a higher score to a document with more occurrences of a query term.
  - Definition [Fang, Tao Tao, Zhai 2004]:

Let q = {w} be a query with only one term w.
Assume $|d_1|=|d_2|$.
If $c(w, d_1) > c(w, d_2)$, then $f(d_1,q)>f(d_2,q)$.

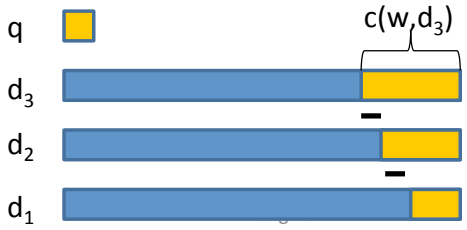


Analytic form

Term Frequency contraints

- TFC2 :
  - Idea : Decreasing of importance of occurrences of a query term.
  - Definition [Fang, Tao Tao, Zhai 2004]:

Let q = {w} be a query with only one term w.
Assume $|d1| = |d2| = |d3|$ and $c(w, d1) > 0$.
If $c(w, d2) - c(w, d1) = 1$ and $c(w, d3) - c(w, d2) = 1$,
then $f(d2, q) - f(d1, q) > f(d3, q) - f(d2, q)$.

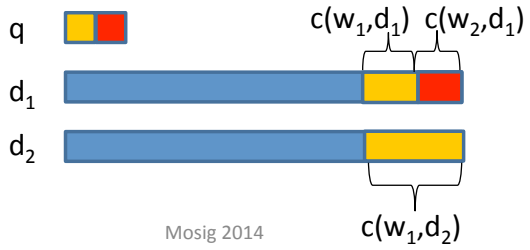


Analytic form

Term Frequency contraints

- TFC3 :
  - Idea : Favor document with more distinct query terms.
  - Definition [Fang, Tao Tao, Zhai 2011]:

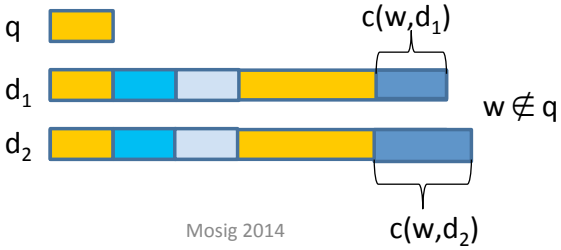
Let q={w1, w2}. Assume $|d_1| = |d_2|$ and $idf(w_1) = idf(w_2)$.
If $c(w_1, d_2) = c(w_1, d_1) + c(w_2, d_1)$
and $c(w_2, d_2) = 0$, $c(w_1, d_1) \neq 0$, $c(w_2, d_1) \neq 0$,
then $f(d_1, q) > f(d_2, q)$.



Lenght Normalization contraints

- LNC1:
  - Idea : Penalize long documents.
  - Definition [Fang, Tao Tao, Zhai 2011]:

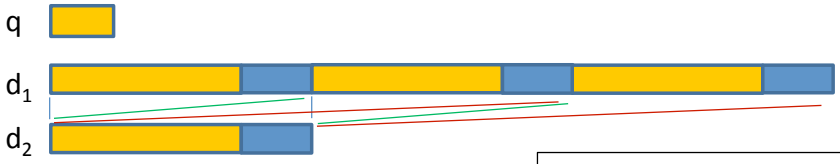
Let q be a query.
If for some word $w \notin q$, $c(w, d_2) = c(w, d_1) + 1$
but for any other word w', $c(w', d_2) = c(w', d_1)$,
then $f(d_1, q) \geq f(d_2, q)$.



Lenght Normalization contraints

- LNC2:
  - Idea : Avoid over-penalizing long documents.
  - Definition [Fang, Tao Tao, Zhai 2011]:

Let q be a query.
$\forall k > 1$, $|d_1| = k \cdot |d_2|$, $c(q, D2) > 0$
and for all terms w, $c(w, d_1) = k \cdot c(w, d_2)$,
then $f(d_1,q) \geq f(d_2,q)$.

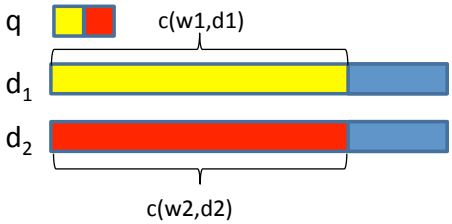


Analytic form

Term Discrimination contraint

- speTDC:
  - Idea : Penalize popular terms in collection.
  - Definition [Clinchant & Gaussier 2011]:

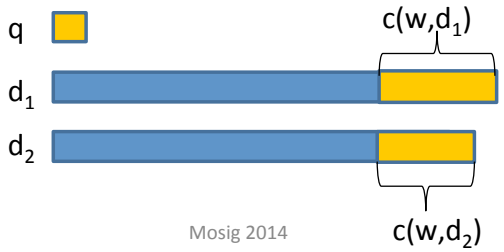
Let $q = \{w_1, w_2\}$.
Assume $|d_1| = |d_2|$, $c(w1, d_1) = c(w2, d_2)$,
$c(w2, d_1) = c(w1, d_2) = 0$,
If $idf(w_1) > idf(w_2)$, then $f(d_1, q) > f(d_2, q)$.



Term Frequency - Lenght contraint

- TF-LNC:
  - Idea : Regularize interaction between tf and length.
  - Definition [Fang, Tao Tao, Zhai 2011]:

> *Let Q = {w} be a query.*
> *If $c(w, d_1) > c(w, d_2)$*
> *and $|d_1| = |d_2| + c(w, d_1) - c(w, d_2)$,*
> *then $f(d_1, q) > f(d_2, q)$.*



Study of some axioms on DIR model

- Consider the DIR model (query likelihod with Dirichlet smoothing)
- Theoretical formula :

$$P(q \mid d) \propto \sum_{w \in V} c(w,q).\ln\left(\frac{c(w,d) + \mu.P(w \mid C)}{|d| + \mu}\right)$$

  - Good from a theoretical point of view, but not from an efficiency point of view
- Implementation using inverted files :

$$f(d,q) = \sum_{w \in d \cap q} c(w,q).\ln\left(1 + \frac{c(w,d)}{\mu.P(w \mid C)}\right) + |q|.\ln\left(\frac{\mu}{\mu + |d|}\right)$$

  … Not obvious… explainations follow…

DIR model

- If $\hat{\theta}_d$ is the Dirichlet smoothed document language model, $\mu \in R^+$

$$\ln P(q|\hat{\theta}_d) =_{rank} \sum_{w \in V} c(w,q).\ln(\frac{c(w,d)+\mu.P(w|C)}{|d|+\mu})$$

$$=_{rank} \sum_{w \in d \cap q} c(w,q).\ln(\frac{c(w,d)+\mu.P(w|C)}{|d|+\mu}) + \sum_{w \in q, w \notin d} c(w,q).\ln(\frac{\mu.P(t|C)}{|d|+\mu})$$

$$=_{rank} \sum_{w \in d \cap q} c(w,q).\ln(\frac{c(w,d)+\mu.P(w|C)}{|d|+\mu}) + \sum_{w \in q, w \notin d} c(w,q).\ln(\frac{\mu.P(t|C)}{|d|+\mu})$$

$$+ \sum_{w \in q} c(w,q).\ln(\frac{\mu.P(t|C)}{|d|+\mu}) - \sum_{w \in q} c(w,q).\ln(\frac{\mu.P(t|C)}{|d|+\mu})$$

DIR model

$$\ln P(q|\hat{\theta}_d) =_{rank} \sum_{w \in d \cap q} c(w,q).\ln(\frac{c(w,d)+\mu.P(w|C)}{|d|+\mu}) - \sum_{w \in d \cap q} c(w,q).\ln(\frac{\mu.P(t|C)}{|d|+\mu})$$

$$+ \sum_{w \in q} c(w,q).\ln(\frac{\mu.P(t|C)}{|d|+\mu})$$

$$=_{rank} \sum_{w \in d \cap q} c(w,q).\ln(1+\frac{c(w,d)}{\mu.P(t|C)}) + \sum_{w \in q} c(w,q).\ln(\frac{\mu}{|d|+\mu})$$

$$+ \sum_{w \in q} c(w,q).\log P(t|C)$$

$$=_{rank} \sum_{w \in d \cap q} c(w,q).\ln(1+\frac{c(w,d)}{\mu.P(t|C)}) + |q|.\ln(\frac{\mu}{|d|+\mu})$$

Study of axiom TFC1 on DIR model

- DIR with q={w}

$$f(d,q) = \ln(1 + \frac{c(w,d)}{\mu.P(w|C)}) + \ln(\frac{\mu}{\mu + |d|})$$

- TFC1 :
  - If > and |d1| = |d2|
  - Then :
    - $\ln(\frac{\mu}{\mu + |d1|}) = \ln(\frac{\mu}{\mu + |d2|})$
    - $\ln(1 + \frac{c(w,d1)}{\mu.P(w|C)}) > \ln(1 + \frac{c(w,d2)}{\mu.P(w|C)})$
    - $f(d1,q) > f(d2,q)$

  → DIR validates TFC1 unconditionnaly

to TFC1

Study of axiom TFC2 on DIR model

- DIR with q={w}

$$f(d,q) = \ln(1 + \frac{c(w,d)}{\mu.P(w|C)}) + \ln(\frac{\mu}{\mu + |d|})$$

- TFC2 :
  - If $c(w,d_3) = c(w,d_2) + 1 = c(w,d_1) + 2 \ \ and \ |d_1| = |d_2| = |d_3|$
  - Then
    - $\ln(\frac{\mu}{\mu + |d_1|}) = \ln(\frac{\mu}{\mu + |d_2|}) \ln(\frac{\mu}{\mu + |d_3|})$
    - ln(1+x) is monotonic increasing for x > 0, but second derivative is $-\frac{1}{x^2} < 0 \quad \forall x \in R$
  - So $f(d_2,q) - f(d_1,q) > f(d_3,q) - f(d_2,q)$
  - DIR validates TFC2 unconditionnaly

to TFC2

Usage of the axioms

- New theorems can be build from the axioms
  - Exemple (from Fang 2011)
    - LNC1 and TF-LNC
      - Let q={w}, and if |d1| < |d2| + c(w,d1) – c(w,d2), and c(w,d1)>c(w,d2)
      - Then f(d1,q) > f(d2,q)

Usage of the axioms

- Existing formulas may be modified to enhance there compatibility with axioms
  - BM25 [Robertson & Walker 1994; Singhal 2001]

$$f(q,d) = \sum_{w \in q \cap d} \ln\left(\frac{N - df(w) + 0.5}{df(w) + 0.5}\right) \times \frac{(k1+1).c(w,d)}{k1.\left((1-b) + b.\frac{|d|}{avdl}\right) + c(w,d)} \times \frac{(k3+1).c(t,q)}{k3 + c(t,q)}$$

    - With k1∈[1.0, 2.0], b=0.75 and k3 ∈[0, 1000]
  - If *df(w)> N/2* then idf part negative
    - TFCs, LNCs and TF-LNC are then only conditionnally validated
      - TFC1 . with |d1|=|d2|=avdl :
        - » $f(\{w\}, d1) = \ln\left(\frac{N - df(w) + 0.5}{df(w) + 0.5}\right) \times \frac{(k1+1).c(w,d1)}{k1 + c(w,d1)}$
        - » If idf negative and, then decreases as c(w,.) increases

Usage of the axioms

- Existing formulas may be modified to enhance there compatibility with axioms
  - Original BM25

$$\mathrm{f(q,d)} = \sum_{w \in q \cap d} \ln\left(\frac{N - df(w) + 0.5}{df(w) + 0.5}\right) \times \frac{(k1+1).c(w,d)}{k1.\left((1-b) + b.\frac{|d|}{avdl}\right) + c(w,d)} \times \frac{(k3+1).c(t,q)}{k3 + c(t,q)}$$

    - With k1∈[1.0, 2.0], b=0.75 and k3 ∈[0, 1000]
  - Modified idf leads to

$$\mathrm{f(q,d)} = \sum_{w \in q \cap d} \ln\left(\frac{N+1}{df(w)}\right) \times \frac{(k1+1).c(w,d)}{k1.\left((1-b) + b.\frac{|d|}{avdl}\right) + c(w,d)} \times \frac{(k3+1).c(t,q)}{k3 + c(t,q)}$$

    - Validates LNC1 unconditionally

Conclusion

- Theoretical evaluation of matching IR function is an important way to a scientific domain.
- Definition of axioms that are validated is a way to achieve that
  - Positive link between axioms validation and quality of results has been shown
  - Are these axioms complete ?
    - Refinements of TDC, speTDC…
    - Axioms are broad principles, more precise analysis introduce perturbations in corpus [Fang Tao Tao Zhai 2011].
    - [Lv & Zhai 2011] added Lower bounding constraints to avoid over-penalizing long documents
    - f(Q,Q) is not maximal …

Work to be done

- Understand the ideas of axiomatic evaluation
- Understand the interest of such theoretical evaluation

References

[Fang et al 2004] Hui Fang and Tao Tao and ChengXiang Zhai, A formal study of information retrieval heuristics, ACM SIGIR 2004, pp. 49-56.

[Fang & Zhai 2005] Hui Fang and ChengXiang Zhai, An exploration of axiomatic approaches to information retrieval, ACM SIGIR 2005, pp. 480-487.

[Fang et al 2011] Hui Fang and Tao Tao and ChengXiang Zhai, Diagnostic Evaluation of Information Retrieval Models, ACM Trans. Inf. Syst. 2011, Vol. 29, N. 2, 1-42.

[Lv & Zhai 2011] Yuanhua Lv and ChengXiang Zhai, Lower-bounding term frequency normalization., *ACM* CIKM 11, pp. 7-16.

[Clinchant & Gaussier 2011] Stéphane Clinchant and Eric Gaussier, Do IR models satisfy the TDC retrieval constraint, *ACM SIGIR 2011*, 1155-1156.

Fang & Zhai tutorials at ICTIR 2013 and SIGIR 2014.