Research Papers Review

Name : Kweku E. Acquaye

Topic : Effective Generalised IR Rules

Starting Reference Paper : *A formal study of information retrieval heuristics (H. Fang, T. Tao, and C. Zhai, SIGIR, 2004)*[1]

Second Reference Paper : Diagnostic Evaluation of Information Retrieval Models7 (H. Fang, T. Tao, and C. Zhai, ACM Transactions on Information Systems, Vol. V, No. N, October 2010, Pages 1–46.

Research Paper 1

A formal study of information retrieval heuristics[1]
- H. Fang, T. Tao, and C. Zhai, SIGIR, 2004

- The problem being addressed:
  - ➢ identifying retrieval performance handicaps
- How it was addressed:
  - ➢ *heuristics* that work *empirically* to indicate good retrieval performance
  - ➢ checked against variety of retrieval formulas
    - ✓ vector space model (pivoted normalization)[2]
    - ✓ classic probabilistic retrieval model, (Okapi)[3]
    - ✓ the language modelling approach (Dirichlet prior smoothing)[4].

The 6 Constraints

- TFC1 - Term Frequency Constraint 1

  Formal definition:

  Let $q = \{w\}$ be a query with only one term w.
  Assume $|d1| = |d2|$.
  if $c(w, d1) > c(w, d2)$,
  then $f(d1, q) > f(d2, q)$.

  Intuition: to favour and give higher scores to documents with more occurrence of a query term.

- TFC2 - Term Frequency Constraint 2

  Formal definition:

  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 $f(d2, q) - f(d1, q) > f(d3, q) - f(d2, q)$.

  Intuition: Basically, for equal length documents, the document with more distinct query terms has higher score.

- TDC - Term Discrimination Constraint

  Formal definition:

  Let q be a query and $w1, w2 \in q$ be 2 query terms.
  Assume $|d1| = |d2|$, $c(w1, d1)+c(w2, d1) = c(w1, d2)+c(w2, d2)$.
  if $idf(w1) \geq idf(w2)$ and $c(w1, d1) \geq c(w1, d2)$,
  then $f(d1, q) \geq f(d2, q)$.

  Intuition: This is to mitigate the impact of TF and IDF.

- LNC1 – Length Normalization Constraint 1

  Formal definition:

  Let q be a query and $d1$, $d2$ be 2 documents.
  If for some word $w' /\in q$, $c(w', d2) = c(w', d1) + 1$
  but for any query term w, $c(w, d2) = c(w, d1)$,
  then $f(d1, q) \geq f(d2, q)$.

  Intuition: The intuition here is to favour shorter length documents over longer documents with the same term frequency.

The 6 Constraints (continued)

- LNC2 – Length Normalization Constraint 2

Formal definition:

Let q be a query. ∀k > 1, if d1 and d2 are documents such that |d1| = k ▪ |d2| and for all terms w,

c(w, d1) = k ▪ c(w, d2),

then f(d1, q) ≥ f(d2, q).

Intuition: To avoid penalising long documents too much.

- TF-Length Constraint

Formal definition:

Let q = {w} be a query with only one term w.

if c(w, d1) > c(w, d2) and

|d1| = |d2| + c(w, d1) − c(w, d2),

then f(d1, q) > f(d2, q).

Intuition: To hold steady the relation between term frequency and document length.

- What are the results:
  - no single retrieval formula satisfies all constraints/heuristics
  - some formulas violate more constraints/heuristics than others
  - some formulas violate some constraints/heuristics more “seriously” than others
- My views:
  - My expertise level not sufficient to be critical, however:
  - constraints/heuristics not exhaustive; why 6 and not 10?
  - constraints/heuristics probably skewed to authors bias, what is the yardstick of ultimate objectivity?

Further reading:

- An exploration of axiomatic approaches to information retrieval (H. Fang, C. Zhai, SIGIR, 2005)[5].
- An exploration of proximity measures in information retrieval (T. Tao, C. Zhai, SIGIR, 2007)[6].
- Diagnostic evaluation of Information Retrieval models (H. Fang, T. Tao, and C. Zhai, ACM, 2010)[7].

Research Paper 2

Diagnostic Evaluation of Information Retrieval Models[7]

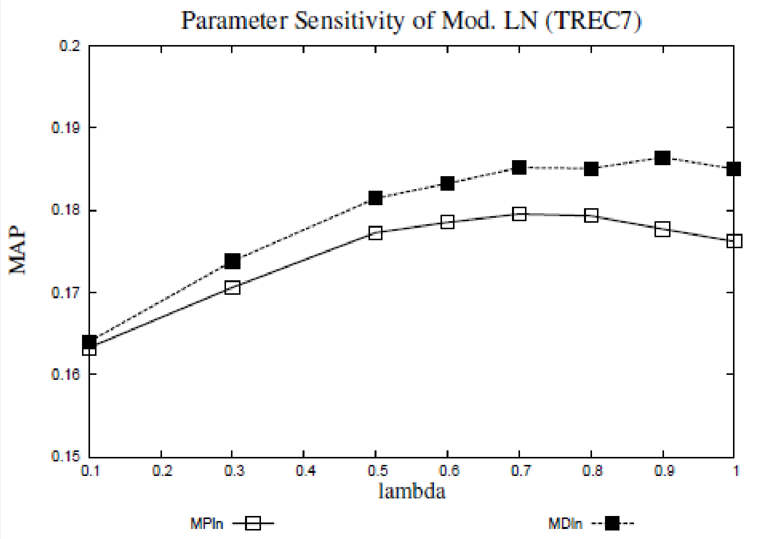
- H. Fang, T. Tao, and C. Zhai, ACM Transactions on Information Systems, Vol. V, No. N, October 2010, Pages 1–46.

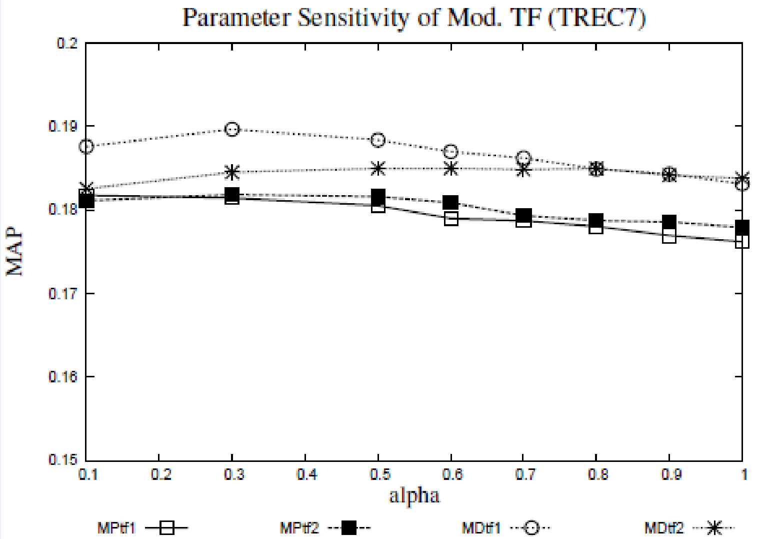
- The problem being addressed:
  - ➢ improving optimal retrieval performance

- How it is being addressed:
  - ➢ 7 constraints – same heuristics as before + 1
  - ➢ checked against variety of retrieval formulas
  - ➢ diagnostic test developed
  - ➢ modify and improve retrieval functions:
    1. Improving Length Normalization
    2. Improving TF Implementations
    3. Combining modified TF and LN implementations
  - ➢ evaluate on 8 representative datasets

- What are the results:
  - ➢ improvements in modified algorithms by means of the mean average precision (MAP) measure
  - ➢ modified algorithms outperform originals
  - ➢ provides guidance for improving existing retrieval functions

- My views:
  - ➢ Again, my expertise level is not sufficient to be critical (still learning), however:
  - ➢ the 8 representative datasets used for evaluation was drawn up by the authors, to eliminate implicit bias it should have been constituted by neutral expert third party.





* Graph illustrations reproduced from H. Fang et al[4].

References

1. Hui Fang, Tao Tao, and ChengXiang Zhai (2004), *A formal study of information retrieval heuristics*, SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, July 2004, pp. 49-5, https://doi.org/10.1145/1008992.1009004
2. G. Salton, and C. Buckley (1988), *Term-weighting approaches in automatic text retrieval*, Information Processing and Management, **vol. 24**, pp. 513–523.
3. J. Lafferty, and C. Zhai (2003), *Probabilistic relevance models based on document and query generation*, In W. B. Croft and J. Lafferty, editors, Language Modeling and Information Retrieval, Kluwer Academic Publishers.
4. C. Zhai, and J. Lafferty (2001), *A study of smoothing methods for language models applied to ad hoc information retrieval*, In Proceedings of SIGIR’01, pp. 334–342.
5. Hui Fang, and ChengXiang Zhai (2005), *An exploration of axiomatic approaches to information retrieval,* SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, August 2005, pp. 480–487, https://doi.org/10.1145/1076034.1076116
6. Tao Tao, and ChengXiang Zhai (2007), *An exploration of proximity measures in information retrieval,* SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, July 2007, pp. 295–302, https://doi.org/10.1145/1277741.1277794
7. Hui Fang, Tao Tao, and ChengXiang Zhai (2010), *Diagnostic Evaluation of Information Retrieval Models,* ACM Transactions on Information Systems, **vol. V**, No. N, October 2010, pp. 1–46, http://sifaka.cs.uiuc.edu/czhai/pub/tois-diag.pdf

Thank you !