IR Models: Foundations and Relationships

Thomas Roelleke
Queen Mary, University of London
thor@eecs.qmul.ac.uk

ABSTRACT

In IR research it is essential to know IR models. Research over the past years has consolidated the foundations of IR models. Moreover, relationships have been reported that help to use and position IR models. Knowing about the foundations and relationships of IR models can significantly improve building information management systems.

The first part of this tutorial presents an in-depth consolidation of the *foundations* of the main IR models (TF-IDF, BM25, LM). Particular attention will be given to notation and probabilistic roots. The second part crystallises the *relationships* between models. Does LM embody IDF? How "heuristic" is TF-IDF? What are the probabilistic roots? How are LM and the probability of relevance related? What are the components shared by the main IR models?

After the tutorial, attendees will be familiar with a consolidated view on IR models. The tutorial will be illustrative and interactive, providing opportunities to exchange controversial issues and research challenges.

1. IR MODELS: FOUNDATIONS (90 MINS)

The introduction of the traditional strands and instances of IR models includes: TF-IDF, VSM (Vector-Space Model), G-VSM (Generalised VSM), PRF (Probability of Relevance Framework), BIR (Binary Independence Retrieval, RSJ (Robertson-SparckJones) weight, Laplace-like estimation of probabilities), Probabilistic Inference Networks (PIN), Poisson and 2-Poisson, BM25, DFR (Divergence from Randomness), and LM.

This part will consolidate the main foundations following from [15] (probabilistic retrieval, BIR model), [5] (missing relevance), [20, 22, 21, 11] (PIN), [13] (2-Poisson, foundation for the BM25 TF quantification $\mathrm{tf}/(\mathrm{tf}+K)$), [4] (IDF is Poisson approximation, aspects of burstiness), [3] (DFR), [12, 8, 9] (LM).

Moreover, this part will emphasises that IR models are consequent applications of basic techniques from probability theory: Bayes' Theorem, Total Probability Theorem, probability mixtures, and divergence measures (e.g. KL-divergence, mutual information).

Particular emphasis will be given to the issue of event spaces [16, 18, 10], and a notation that is consistent across different models.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

Finally, the first part will refer to general approaches to IR such as logical IR [23, 25] and PIN's [20].

2. IR MODELS: RELATIONSHIPS (90 MINS)

The second part of the tutorial is dedicated to the relationships between retrieval models. In addition to the aim to look across the boundaries of single models models, i.e. different from "single-model" tutorials, the relationships between models is a differentiator to previous tutorials presented by Hugo Zaragoza, Stephen Robertson, Djoerd Hiemstra, Victor Lavrenko, and Donald Metzler.

There are several relationships between the models, and for selected relationships the tutorial will outline the mathematical proofs while keeping the formalisms minimal. We will look at the following relationships:

- BIR and TF-IDF: BIR can be used to "explain" TF-IDF; essentially, this views IDF as an approximation of the BIR/RSJ weight for the case of missing relevance, [14, 6].
- TF-IDF, LM and Poisson: There are relationships that follow from being precise about the event spaces [16, 18, 10], e.g. probabilities based on document frequencies of terms, or based on token occurrences of terms.
- PRF, TF-IDF and LM: Recently, [2], "Towards a better understanding of the relationship between probabilistic models in IR", investigated some controversial aspects regarding the relationship between PRF (probabilistic odds), TF-IDF and LM.
- PRF and LM: There is an early view on the relationship between LM (language modelling) and the PRF (probability of relevance framework) [9] which is addressed and criticised in [10].
- PIN's and TF-IDF and LM: A relationship between LM and PIN's has been pointed out in [11]; this relationship builds upon the earlier work [20, 22, 21] discussing the relationship between PIN's and TF-IDF; overall, this marks the total probability theorem and PIN's as a binding link between TF-IDF and LM.
- Logical IR, VSM and G-VSM: Expanding [25], we can go as far as relating the G-VSM to the total probability theorem, a relationship that seamless leads to combinations of geometric and probabilistic IR [24].
- Information theory: On the information-theoretic side, there are controversial ways to relate IR model to concepts of information theory [1, 14].

- Axiomatic approach: Models can be characterised by axioms/constraints [7] the models do or do not satisfy.

Regarding technical aspects covered, the tutorial makes explicit that "there is as much LM in TF-IDF as there is TF-IDF in LM". TF-IDF has probabilistic roots. The duality of TF-IDF and LM marks TF-IDF as a model, not just as a weighting scheme in the VSM; the VSM is a "framework" to express models [17].
Also, the tutorial looks at the statement "in LM, the estimate $P(t) := \mathrm{df}(t)/(\sum_{t'} df(t'))$ is wrong; $P(t)$ should be based on the term frequency". Moreover, the tutorial will review that LM is based on $P(q|d)/P(q)$, how this interpretation justifies the relationship between LM and the probability of relevance $P(r|d,q)$ [9]. Finally, the tutorial will include conducive interpretations of the renown BM25 TF quantification $\mathrm{tf}/(\mathrm{tf}+K)$ [3, 26, 19]. After the tutorial, the participants will have their personal view on statements such as "we know *that* LM works, but we do not know *why*"; "TF-IDF is intuitive, LM is not"; "TF-IDF is heuristic, whereas LM has a probabilistic semantics".
In summary, the tutorial is structured as follows:

IR Models: Foundations
TF-IDF, Vector-space Model
PRF, BIR, 2-Poisson, BM25
LM, Relevance
DFR, KL-Divergence
Entropy, Mutual Information
IR Models: Relationships
PIN's, TF-IDF and LM [20, 11]
IDF and BIR [14, 6]
Event spaces [16, 18, 10]
Model axioms/constraints [7]
LM and PRF/BM25: related? [9, 2]
LM and TF-IDF: siblings! [19]

Thomas Roelleke is a senior lecturer at Queen Mary, University of London. His research expertise lies in IR Models & Theory and Probabilistic DB+IR. Currently, Thomas is a visiting scientist at Yahoo Research, Barcelona Media. Thomas has presented IR model tutorials at German summer schools, taught MSc modules on Foundations of IR, and published several IR model & theory papers.

3. REFERENCES

[1] Akiko Aizawa. An information-theoretic perspective of tf-idf measures. *Information Processing and Management*, 39:45–65, January 2003.

[2] R. Aly and T. Demeester. Towards a better understanding of the relationship between probabilistic models in ir. In *Advances in Information Retrieval Theory: Third International Conference, ICTIR, 2011, Bertinoro, Italy, September 12-14, 2011, Proceedings*, volume 6931, pages 164–175. Springer-Verlag New York Inc, 2011.

[3] Gianni Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transaction on Information Systems (TOIS)*, 20(4):357–389, October 2002.

[4] K. Church and W. Gale. Inverse document frequency (idf): A measure of deviation from Poisson. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 121–130, 1995.

[5] W.B. Croft and D.J. Harper. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35:285–295, 1979.

[6] Arjen de Vries and Thomas Roelleke. Relevance information: A loss of entropy but a gain for IDF? In *ACM SIGIR*, pages 282–289, Salvador, Brazil, 2005.

[7] Hui Fang and ChengXiang Zhai. An exploration of axiomatic approaches to information retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 480–487, New York, 2005. ACM.

[8] Djoerd Hiemstra. A probabilistic justification for using tf.idf term weighting in information retrieval. *International Journal on Digital Libraries*, 3(2):131–139, 2000.

[9] John Lafferty and ChengXiang Zhai. *Probabilistic Relevance Models Based on Document and Query Generation*, chapter 1. Kluwer, 2003.

[10] Robert W. P. Luk. On event space and rank equivalence between probabilistic retrieval models. *Inf. Retr.*, 11(6):539–561, 2008.

[11] Donald Metzler and W. Bruce Croft. Combining the language model and inference network approaches to retrieval. *Information Processing & Management*, 40(5):735–750, 2004.

[12] J.M. Ponte and W.B. Croft. A language modeling approach to information retrieval. In W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281, New York, 1998. ACM.

[13] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In W. Bruce Croft and C. J. van Rijsbergen, editors, *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, London, et al., 1994.

[14] S.E. Robertson. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation*, 60:503–520, 2004.

[15] S.E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146, 1976.

[16] Stephen Robertson. On event spaces and probabilistic models in information retrieval. *Information Retrieval Journal*, 8(2):319–329, 2005.

[17] Thomas Roelleke, Theodora Tsikrika, and Gabriella Kazai. A general matrix framework for modelling information retrieval. *Journal on Information Processing & Management (IP&M), Special Issue on Theory in Information Retrieval*, 42(1), 2006.

[18] Thomas Roelleke and Jun Wang. A parallel derivation of probabilistic information retrieval models. In *ACM SIGIR*, pages 107–114, Seattle, USA, 2006.

[19] Thomas Roelleke and Jun Wang. TF-IDF uncovered: A study of theories and probabilities. In *ACM SIGIR*, pages 435–442, Singapore, July 2008.

[20] H. Turtle and W. B. Croft. Inference networks for document retrieval. In J.-L. Vidick, editor, *Proceedings of the 13th International Conference on Research and Development in Information Retrieval*, pages 1–24, New York, 1990. ACM.

[21] H. Turtle and W.B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, 1991.

[22] H.R. Turtle and W.B. Croft. Efficient probabilistic inference for text retrieval. In *Proceedings RIAO 91*, pages 644–661, Paris, France, 1991.

[23] C. J. van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 29(6):481–485, 1986.

[24] Cornelis J. van Rijsbergen. *The Geometry of Information Retrieval*. Cambridge University Press, 2004.

[25] S.K.M. Wong and Y.Y. Yao. On modeling information retrieval with probabilistic inference. *ACM Transactions on Information Systems*, 13(1):38–68, 1995.

[26] Hengzhi Wu and Thomas Roelleke. Semi-subsumed events: A probabilistic semantics for the BM25 term frequency quantification. In *ICTIR (International Conference on Theory in Information Retrieval)*. Springer, 2009.