

Informativeness of tf-idf and Language modelling

Jun Wang

1. INTRODUCTION

Retrieval models form a crucial part of information retrieval. We distinguish mainly two classes: non-probabilistic and probabilistic models. On the non-probabilistic side, tf-idf is the dominant model, and on the probabilistic side, the binary independent retrieval model and language modelling are the main candidates. It's known that both, tf-idf and LM deliver good and stable retrieval quality, and both can be tuned to incorporate collection and application-specific characteristics. In tf-idf-like retrieval, the informativeness (discriminativeness) of terms is a collection-based notion associated with the inverse document frequency (idf). Language modelling (LM) is based on a linear mixture of collection-based and document-based probabilities. The question is how we can associate the idf-based informativeness to LM, and the LM-based probability mixture to tf-idf. Answering this question would contribute to our understanding of what exactly constitutes a retrieval model.

This paper presents research regarding the symmetric formulation (implementation) of tf-idf and LM. This research has been motivated by a number of questions. We ask for a way to isolate the parameters that constitute retrieval models. Can we find informativeness, i.e. an idf-like parameter in LM? Can we find a probability mixture in tf-idf? At last, what exactly constitutes a retrieval model, and what makes a retrieval model good and even superior?

2. TF-IDF MODEL AND LANGUAGE MODELLING

2.1 TF-IDF

For starting the symmetric representing, we look now at the traditional definition of tf-idf. First, we define idf, where the definition can be related to the probability that a term occurs in the collection. In short, this works as follows:

$$\text{idf}(t, c) := -\log \frac{n_D(t, c)}{N_D(c)} = \quad (1)$$

$$= -\log P_D(t|c) = \log \frac{1}{P_D(t|c)} \quad (2)$$

$$\propto -\log P_D(t|\bar{r}) = \log \frac{1}{P_D(t|\bar{r})} \quad (3)$$

The probability $P_D(t|c)$ is an *occurrence* probability. It grows monotonically with $n_D(t, c)$. Thus, the *informativeness* measure $\text{idf}(t, c)$ decreases monotonically with $n_D(t, c)$. The traditional tf-idf definition [2] multiplies a within-document term probability $P_L(t|d)$ with idf. The definition comes usually in the following form:

$$\text{RSV}_{\text{tf-idf}}(d, q) := \sum_{t \in d \cap q} P_L(t|d) \cdot -\log P_D(t|c) \quad (4)$$

Whatever the estimate for $P_L(t|d)$, we can re-write $\text{RSV}_{\text{tf-idf}}$ by moving $P_L(t|d)$ inside the logarithm. Then, we can define the probability $P_{\text{tf-idf}}(t|c, d)$, and rewrite the RSV as follows:

$$P_{\text{tf-idf}}(t|c, d) := (P_D(t|c))^{P_L(t|d)} \quad (5)$$

$$\text{RSV}_{\text{tf-idf}}(d, q) = \sum_{t \in d \cap q} -\log P_{\text{tf-idf}}(t|c, d) \quad (6)$$

$P(t|d)$ estimation can be maximum likelihood ($\frac{n_L(t, d)}{\max_i n_L(t_i, d)}$) or $\frac{n_L(t, d)}{\sum_i n_L(t_i, d)}$, or 2-Poisson estimation (2-Poisson approximation is $\frac{n_L(t, d)}{n_L(t, d) + K}$).

2.2 LM

In the literature (see [1]), language modelling is based on the *linear mixture* of the term occurrence in the collection ($P_L(t|c)$) and the term occurrence in the document ($P_L(t|d)$):

$$P(t|c, d) = \lambda \cdot P_L(t|c) + (1 - \lambda) \cdot P_L(t|d) \quad (7)$$

For language modelling, the retrieval status value is based on the product of term probabilities:

$$P(q|c, d) = \prod_{t \in q} P(t|c, d) \quad (8)$$

Dividing $P(q|c, d)$ by the constant $\prod_{t \in q} \lambda \cdot P_L(t|c)$ yields the retrieval status value RSV_{LM} :

$$RSV_{LM}(d, q) := \sum_{t \in d \cap q} \log \left(1 + \frac{1-\lambda}{\lambda} \cdot \frac{P_L(t|d)}{P_L(t|c)} \right) \quad (9)$$

For relating language modelling to the notion of informativeness, we form the inverse of the argument in the logarithm, to gain an expression where we apply the negative logarithm. Consider the intermediate step in this transformation, where we use $1 = \frac{P_L(t|c)}{P_L(t|d)}$ and $\alpha := \frac{1-\lambda}{\lambda}$.

$$RSV_{LM}(d, q) = \sum_{t \in d \cap q} \log \left(\frac{P_L(t|c)}{P_L(t|c)} + \frac{\alpha \cdot P_L(t|d)}{P_L(t|c)} \right) \quad (10)$$

This leads us to the alternative formulation of the RSV_{LM} . We define the term probability $P_{LM}(t|c, d)$, and apply the negative logarithm.

$$P_{LM}(t|c, d) := \frac{P_L(t|c)}{P_L(t|c) + \alpha \cdot P_L(t|d)} \quad (11)$$

$$RSV_{LM}(d, q) = \sum_{t \in d \cap q} -\log P_{LM}(t|c, d) \quad (12)$$

3. LIGHT MODELS

Having reached this logarithmic form, we find the series that approximates $-\log(1-x)$ (see any math text book):

$$-\log(1-x) = x + \frac{x^2}{2} + \dots + \frac{x^n}{n} + \dots \quad (13)$$

This series leads us to an approximation of the LM retrieval status value, and we denote this approximation as $RSV_{LM-light}$.

$$RSV_{LM-light}(d, q) := \sum_{t \in d \cap q} \frac{n_L(t, d)}{n_L(t, d) + \frac{1}{\alpha} \cdot P(d|c) \cdot n_L(t|c)} \quad (14)$$

The approximation leads us to a slim (“light”) retrieval model that is free of a logarithm. We can illustrate that the sum over the complement $P_{LM}(\bar{t}|c, d)$ is an approximation of the sum over $-\log P_{LM}(t|c, d)$.

$$\begin{aligned} RSV_{LM-light}(d, q) &= \\ &= \sum_{t \in d \cap q} \left(1 - \frac{n_L(t, c)}{n_L(t, c) + \alpha \cdot \frac{1}{P(d|c)} \cdot n_L(t, d)} \right) \end{aligned}$$

Let us relate this finding to tf-idf. Applying the approximation of the logarithm, yields the following form of tf-idf-light:

$$RSV_{tf-idf-light}(d, q) := \sum_{t \in d \cap q} \left(1 - P_D(t|c)^{P_L(t|d)} \right) \quad (15)$$

4. EXPERIMENTAL EVALUATION

For the evaluation, we selected four collections: INEX04, INEX05, TREC-3, and TREC-8. The idea behind this selection is to investigate the models over a fair range of mixed and large-scale data. Experiments were carried out on the full sets of queries and data.

The main findings are: 1. The tf-idf-light approach performs for linear tf estimates consistently better than the genuine

tf-idf approaches, whereas for tf-idf with tf rational, and for LM, the genuine approaches outperform their light comparators. 2. The tf-idf approach with tf rational estimate is overall the superior tf-idf approach, though the tf-idf-light with tf max does well (see INEX04 MAP and P@10, where the tf-idf-light tf max approach yields MAP=27.1% and P@10=37.6%). 3. The genuine LM approach is superior for INEX04, whereas the advantage becomes marginal for INEX05, and tf-idf with tf rational takes the lead for the TREC data.

5. CONCLUSION

In this paper, we present an informativeness-based and symmetric view of LM and tf-idf based on probability mixtures, a new retrieval model with approximation of the logarithm, and a theoretical and experimental analysis of the nature of informativeness and its impact on retrieval quality.

6. REFERENCES

- [1] Bruce Croft and John Lafferty, editors. *Language Modeling for Information Retrieval*. Kluwer, 2003.
- [2] G. Salton, A. Wong, and C.S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620, 1975.

		MAP				P@10			
		INEX04	INEX05	TREC-3	TREC-8	INEX04	INEX05	TREC-3	TREC-8
tf-idf genuine	tf sum	0.1464	0.1362	0.1192	0.1006	0.2294	0.2690	0.2120	0.1080
tf-idf light	tf sum	0.2180	0.1592	0.1216	0.1028	0.3382	0.3310	0.2120	0.1180
tf-idf genuine	tf max	0.1859	0.1758	0.1057	0.0858	0.2971	0.4241	0.2220	0.0900
tf-idf light	tf max	0.2716	0.2032	0.1450	0.1147	0.3765	0.4414	0.2880	0.1680
tf-idf genuine	tf rational	0.2438	0.2774	0.2275	0.2581	0.3412	0.4931	0.4060	0.3180
tf-idf light	tf rational	0.2630	0.2589	0.1875	0.2532	0.3765	0.4862	0.3520	0.3100
LM genuine		0.3484	0.2920	0.1873	0.2436	0.4912	0.5069	0.3620	0.2960
LM light		0.2784	0.2309	0.1619	0.2384	0.3794	0.4552	0.2920	0.2780

Table 1: Retrieval Quality