Semi-Subsumed Events: A Probabilistic Semantics of the BM25 Term Frequency Quantification

Hengzhi Wu and Thomas Roelleke

Queen Mary, University of London
{hzwoo,thor}@dcs.qmul.ac.uk

Abstract. Through BM25, the asymptotic term frequency quantification TF $=$ $\text{tf}/(\text{tf}+K)$, where tf is the within-document term frequency and K is a normalisation factor, became popular. This paper reports a finding regarding the meaning of the TF quantification: in the triangle of independence and subsumption, the TF quantification forms the altitude, that is, the middle between independent and subsumed events. We refer to this new assumption as semi-subsumed. While this finding of a well-defined probabilistic assumption solves the probabilistic interpretation of the BM25 TF quantification, it is also of wider impact regarding probability theory.

1 Introduction and Motivation

The BM25 TF quantification/normalisation of the form $\text{tf}/(\text{tf}+K)$ where tf is the total within-document frequency and K is a normalisation parameter (includes the pivoted document length) is renown for superior retrieval quality, outperforming by far the bare total count tf or a maximum-likelihood estimate of the form tf/N_d where N_d is the document length. The total tf corresponds to an *independence* assumption. That is each occurrence of the same term is treated as independent. This assumption is wrong, as the success of BM25 proves.

If a term occurs in a document, and let the initial probability for this occurrence be $P(t|c) = 1/100$ (for example, t occurs in 1% of the documents), then the probability that it occurs again further on in the same document is greater than the initial probability. In other words, the occurrence of an event depends on previous occurrences.

The core contribution of this paper is the notion "semi-subsumed", a probabilistic assumption precisely half-way between independent and subsumed. Probabilistic assumptions are essential in large-scale applications of probabilistic reasoning. Often, the classical assumptions disjointness, independence or subsumption are assumed for events since otherwise the probabilistic reasoning is computationally too expensive. In this paper we focus mainly on the theory around semi-subsumed events, and the effect and application of assuming events to be semi-subsumed in more general probabilistic frameworks such as probabilistic inference networks (PIN) is topic of future research.

2 TF-IDF and BM25

The BM25 TF quantification can be viewed as an approximation of the 2-Poisson model ([2]); this is a probabilistic semantics, however, this paper contributes what can be seen as an intuitive assumption.

This section reviews the probabilistic interpretation of TF-IDF. Let $\text{tf} := n_L(t,d)$ denote the within-document term frequency, i.e. the number of *locations* at which term t occurs in document d; similarly, let $\text{df}(t,c) := n_D(t,c)$ denote the number of *documents* containing term t in collection c; $N_D(c)$ is the total number of documents. The notation allows for a consistent representation of the dimensions used in document retrieval ([4]). Then, TF-IDF and BM25 are defined as follows:

$$P_D(t|c) := n_D(t,c)/N_D(c) \tag{1}$$

$$\text{idf}(t,c) := -\log P_D(t|c) \tag{2}$$

$$\text{RSV}_{\text{TF-IDF}}(d,q,c) := \sum_{t \in d \cap q} \text{tf}(t,d) \cdot \text{idf}(t,c) \tag{3}$$

$$\text{RSV}_{\text{BM25}}(d,q,r,\bar{r}) := \sum_{t \in d \cap q} \frac{\text{tf}(t,d)}{\text{tf}(t,d)+K} \cdot w_t \tag{4}$$

$P_D(t|c)$ is the *document-based* term probability, and $idf(t,c)$ is the negative logarithm of this probability. The term weight w_t is the binary independence weight (based on the probabilities $P(t|r)$ and $P(t|\bar{r})$ that t occurs in relevant and non-relevant documents). The $\text{idf}(t,c)$ can be viewed as an approximation of $w_t = -\log 1/P(t|\bar{r})$ for missing relevance, and this constitutes the close relationship of TF-IDF and BM25 ([1, 3]).

To demonstrate how TF-IDF/BM25 relate to $P(d|q)$ and an assumption for subsequent term events, the next equation forms the exponent of $\text{RSV}_{\text{TF-IDF}}$.

$$\exp(\text{RSV}_{\text{TF-IDF}}) = \prod_{t \in d \cap q} \left(\frac{1}{P_D(t|c)} \right)^{\text{tf}(t,d)} \tag{5}$$

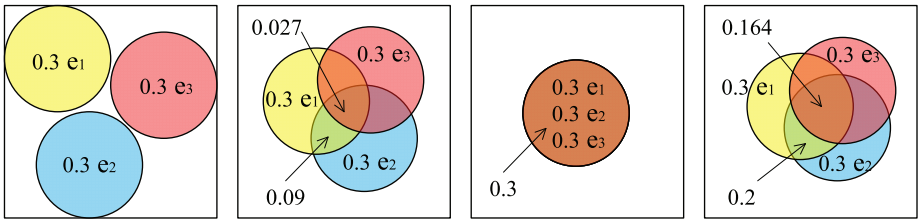
This transformation shows that "naive" TF-IDF involves the expression $P_D(t|c)^{\text{tf}(t,d)}$. $P_D(t|c)$ is the document-based term probability, and the exponent means that "naive" TF-IDF assumes the occurrences of t to be independent events.

The BM25 TF component can be viewed as proposing $P_D(t|c)^{\text{tf}/(\text{tf}+K)}$ to be the term probability, and this probability is significantly greater than $P_D(t|c)^{\text{tf}}$, i.e. the BM25 suggestion is that the probability of subsequent term occurrences is greater than the probability for independent occurrences. The next section shows that this corresponds to assuming subsequent occurrences of a term to be *semi-subsumed* events.

3 Semi-Subsumed Events

Figure 1 illustrate the assumption "semi-subsumed" for three occurrences of an event. Semi-subsumed events overlap more than independent events do, but the overlap is less than for fully subsumed events. For example, given the single event probability $P(e) = 0.3$, for independent occurrences, $P(e_1 \wedge e_2) = 0.3^2 = 0.09$, whereas for subsumed occurrences $P(e_1 \wedge e_2) = 0.3^{2 \cdot 2/3}$.

The independence-subsumption triangle in Figure 2 shows the justification and meaning of the exponent for semi-subsumed events. The left edge of the triangle corresponds to independence, i.e. $P(t|c)^n$ for n occurrences of t, and the right edge corresponds to subsumption, i.e. $P(t|c)$ for any occurrence of t. The rows correspond to



(a) Disjoint (b) Independent (c) Subsumed (d) Semi-Subsumed

Fig. 1. Probabilistic assumptions: three event occurrences

frequencies. The values $\frac{n}{1} \ldots \frac{n}{n}$ in row n correspond to exponents, reflecting independence for $n/1 = n$ and subsumption for $n/n = 1$. The centre column (altitude) is half-way between independence and subsumption. Consequently, $n/(n+1)/2$ is half-way between independence and subsumption, and this leads to the probabilities for independent, semi-subsumed, and subsumed term occurrences:

	independent occurrences	semi-subsumed	subsumed occurrences
1		$\frac{1}{1}$	
2	$\frac{2}{1}$	$\frac{2}{3/2}$	$\frac{2}{2}$
3	$\frac{3}{1}$	$\frac{3}{4/2}$	$\frac{3}{3}$
4	$\frac{4}{1}$ $\frac{4}{2}$		$\frac{4}{3}$ $\frac{4}{4}$
5	$\frac{5}{1}$ $\frac{5}{2}$	$\frac{5}{6/2}$	$\frac{5}{4}$ $\frac{5}{5}$
...	...	...	...
n	$\frac{n}{1}$ $\frac{n}{2}$ $\frac{n}{3}$	$\frac{n}{(n+1)/2}$	$\frac{n}{n-2}$ $\frac{n}{n-1}$ $\frac{n}{n}$

Fig. 2. Independence-Subsumption Triangle (IST)

Independent term occurrences	$P(t\|c)^n = P(t\|c)^{n/1}$
Semi-subsumed term occurrences	$P(t\|c)^{2n/(n+1)}$
Subsumed term occurrences	$P(t\|c)^1 = P(t\|c)^{n/n}$

The triangle in figure 2 and the table above underline how the notion of semi-subsumed events fits "neatly" into the traditional assumptions. The next section shows in a formal proof how the BM25 TF relates to the notion of semi-subsumed events.

4 BM25 TF: Subsequent Term Occurrences are Semi-subsumed Events

The relationship between the BM25 TF and the notion of semi-subsumed events is not directly evident. Therefore, we prove now formally that the BM25 TF quantification assumes semi-subsumed term occurrences.

For an event occurring n times, $2n/(n+1)$ is the value in the altitude of the IST. This value is not equal to the BM25 TF quantification $\text{tf}/(\text{tf} + K)$. The common rewriting

$(\text{tf}/K)/(\text{tf}/K+1)$ helps to establish the relationship between BM25 TF and semi-subsumed.

Theorem 1. *The BM25 TF quantification assumes the occurrences of a term to be semi-subsumed events, i.e. the subsequent occurrence of a term is more likely than if the occurrences were independent, and it is less likely than if they were subsumed.*

Proof. The probability for semi-subsumed events is $P(t|c)^{2n/(n+1)}$.

Set $n := \text{tf}/K$, i.e. n is the normalised term frequency, where K is a normalisation factor (usually involving the pivoted document length). Then, the following equation holds:

$$P_D(t|c)^{2\cdot\text{tf}/(\text{tf}+K)} = P_D(t|c)^{2n/(n+1)} \tag{6}$$

The logarithmic form is $\sum_{t\in d\cap q} 2\cdot\text{tf}/(\text{tf}+K)\cdot\text{idf}(t,c)$. The constant 2 does not affect the ranking.

This proof finalises the contribution of this paper: The BM25 TF quantification assumes subsequent term occurrences (of the same term) to be semi-subsumed events.

5 Summary and Outlook

This paper introduced and discussed "semi-subsumed events". Semi-subsumed events overlap more than if the events were independent, and less than if they were subsumed. For the document-based, collection-wide term probability, $P_D(t|c)^n$ assumes *independence* of n occurrences of t, $P_D(t|c)^1$ assumes *subsumption*, and $P_D(t|c)^{2n/(n+1)}$ assumes *semi-subsumption*. The impact of semi-subsumed events is potentially beyond explaining the BM25 TF quantification. The wider impact is two-fold: on one hand the assumption semi-subsumed helps the theoreticians to develop probabilistic models with a precise semantics; on the other hand, making assumptions is essential for the pragmatic engineers to succeed in large-scale probabilistic reasoning. Regarding the semantics of probabilistic models, in many applications, there seems to be a "law of the series". The Dirichlet distribution and the Laplace law of succession address this law of the series, and future research is to relate Dirichlet and Laplace to semi-subsumed events. Also, the mid-point between disjoint and independent, i.e. semi-disjoint, is a special assumption and will be discussed in future work.

Acknowledgments. We would like to thank the reviewers for their helpful and inspiring comments and suggestions.

References

1. S. Robertson. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation*, 60:503–520, 2004.
2. S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. *ACM SIGIR*, pages 232–241, 1994.
3. S. E. Robertson, S. Walker, and M. Hancock-Beaulieu. Large test collection experiments on an operational interactive system: Okapi at TREC. *IP&M*, 31:345–360, 1995.
4. T. Roelleke, T. Tsikrika, and G. Kazai. A general matrix framework for modelling information retrieval. *IP&M, Special Issue on Theory in Information Retrieval*, 42(1), 2006.