A Frequency-based and a Poisson-based Definition of the Probability of Being Informative

Thomas Roelleke
Department of Computer Science
Queen Mary University of London
thor@dcs.qmul.ac.uk

ABSTRACT

This paper reports on theoretical investigations about the assumptions underlying the inverse document frequency (*idf*). We show that an intuitive *idf*-based probability function for the probability of a term being informative assumes disjoint document events. By assuming documents to be independent rather than disjoint, we arrive at a Poisson-based probability of being informative. The framework is useful for understanding and deciding the parameter estimation and combination in probabilistic retrieval models.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Theory

Keywords

Probabilistic information retrieval, inverse document frequency (idf), Poisson distribution, information theory, independence assumption

1. INTRODUCTION AND BACKGROUND

The inverse document frequency (*idf*) is one of the most successful parameters for a relevance-based ranking of retrieved objects. With N being the total number of documents, and $n(t)$ being the number of documents in which term t occurs, the *idf* is defined as follows:

$$idf(t) := -\log \frac{n(t)}{N},\ 0 <= idf(t) < \infty$$

Ranking based on the sum of the *idf*-values of the query terms that occur in the retrieved documents works well, this has been shown in numerous applications. Also, it is well known that the combination of a document-specific term weight and *idf* works better than *idf* alone. This approach is known as *tf-idf*, where $tf(t,d)$ $(0 <= tf(t,d) <= 1)$ is the so-called *term frequency* of term t in document d. The *idf* reflects the discriminating power (informativeness) of a term, whereas the *tf* reflects the occurrence of a term.

The *idf* alone works better than the *tf* alone does. An explanation might be the problem of *tf* with terms that occur in many documents; let us refer to those terms as "noisy" terms. We use the notion of "noisy" terms rather than "frequent" terms since frequent terms leaves open whether we refer to the document frequency of a term in a collection or to the so-called term frequency (also referred to as within-document frequency) of a term in a document. We associate "noise" with the document frequency of a term in a collection, and we associate "occurrence" with the within-document frequency of a term. The *tf* of a noisy term might be high in a document, but noisy terms are not good candidates for representing a document. Therefore, the removal of noisy terms (known as "stopword removal") is essential when applying *tf*. In a *tf-idf* approach, the removal of stopwords is conceptually obsolete, if stopwords are just words with a low *idf*.

From a probabilistic point of view, *tf* is a value with a frequency-based probabilistic interpretation whereas *idf* has an "informative" rather than a probabilistic interpretation. The missing probabilistic interpretation of *idf* is a problem in probabilistic retrieval models where we combine uncertain knowledge of different dimensions (e.g.: informativeness of terms, structure of documents, quality of documents, age of documents, etc.) such that a good estimate of the probability of relevance is achieved. An intuitive solution is a normalisation of *idf* such that we obtain values in the interval $[0; 1]$. For example, consider a normalisation based on the maximal *idf*-value. Let T be the set of terms occurring in a collection.

$$P_{freq}(t\ is\ informative) := \frac{idf(t)}{maxidf}$$

$$maxidf := \max(\{idf(t)|t \in T\}),\ maxidf <= -\log(1/N)$$

$$minidf := \min(\{idf(t)|t \in T\}),\ minidf >= 0$$

$$\frac{minidf}{maxidf} \leq P_{freq}(t\ is\ informative) \leq 1.0$$

This frequency-based probability function covers the interval $[0; 1]$ if the minimal *idf* is equal to zero, which is the case if we have at least one term that occurs in all documents. Can we interpret P_{freq}, the normalised *idf*, as the probability that the term is informative?

When investigating the probabilistic interpretation of the

SIGIR'03, July 28–August 1, 2003, Toronto, Canada.

normalised *idf*, we made several observations related to disjointness and independence of document events. These observations are reported in section 3. We show in section 3.1 that the frequency-based noise probability $\frac{n(t)}{N}$ used in the classic *idf*-definition can be explained by three assumptions: binary term occurrence, constant document containment and disjointness of document containment events. In section 3.2 we show that by assuming independence of documents, we obtain $1 - e^{-1} \approx 1 - 0.37$ as the upper bound of the noise probability of a term. The value e^{-1} is related to the logarithm and we investigate in section 3.3 the link to information theory. In section 4, we link the results of the previous sections to probability theory. We show the steps from possible worlds to binomial distribution and Poisson distribution. In section 5, we emphasise that the theoretical framework of this paper is applicable for both *idf* and *tf*. Finally, in section 6, we base the definition of the probability of being informative on the results of the previous sections and compare frequency-based and Poisson-based definitions.

2. BACKGROUND

The relationship between frequencies, probabilities and information theory (entropy) has been the focus of many researchers. In this background section, we focus on work that investigates the application of the Poisson distribution in IR since a main part of the work presented in this paper addresses the underlying assumptions of Poisson.

[4] proposes a 2-Poisson model that takes into account the different nature of relevant and non-relevant documents, rare terms (content words) and frequent terms (noisy terms, function words, stopwords). [9] shows experimentally that most of the terms (words) in a collection are distributed according to a low dimension n-Poisson model. [10] uses a 2-Poisson model for including term frequency-based probabilities in the probabilistic retrieval model. The non-linear scaling of the Poisson function showed significant improvement compared to a linear frequency-based probability. The Poisson model was here applied to the term frequency of a term in a document. We will generalise the discussion by pointing out that document frequency and term frequency are dual parameters in the collection space and the document space, respectively. Our discussion of the Poisson distribution focuses on the document frequency in a collection rather than on the term frequency in a document.

[7] and [6] address the deviation of *idf* and Poisson, and apply Poisson mixtures to achieve better Poisson-based estimates. The results proved again experimentally that a one-dimensional Poisson does not work for rare terms, therefore Poisson mixtures and additional parameters are proposed.

[3], section 3.3, illustrates and summarises comprehensively the relationships between frequencies, probabilities and Poisson. Different definitions of *idf* are put into context and a notion of "noise" is defined, where noise is viewed as the complement of *idf*. We use in our paper a different notion of noise: we consider a frequency-based noise that corresponds to the document frequency, and we consider a term noise that is based on the independence of document events.

[11], [12], [8] and [1] link frequencies and probability estimation to information theory. [12] establishes a framework in which information retrieval models are formalised based on probabilistic inference. A key component is the use of a space of disjoint events, where the framework mainly uses terms as disjoint events. The probability of being informative defined in our paper can be viewed as the probability of the disjoint terms in the term space of [12].

[8] address entropy and bibliometric distributions. Entropy is maximal if all events are equiprobable and the frequency-based Lotka law (N/i^{λ} is the number of scientists that have written i publications, where N and λ are distribution parameters), Zipf and the Pareto distribution are related. The Pareto distribution is the continuous case of the Lotka and Lotka and Zipf show equivalences. The Pareto distribution is used by [2] for term frequency normalisation. The Pareto distribution compares to the Poisson distribution in the sense that Pareto is "fat-tailed", i. e. Pareto assigns larger probabilities to large numbers of events than Poisson distributions do. This makes Pareto interesting since Poisson is felt to be too radical on frequent events. We restrict in this paper to the discussion of Poisson, however, our results show that indeed a smoother distribution than Poisson promises to be a good candidate for improving the estimation of probabilities in information retrieval.

[1] establishes a theoretical link between *tf-idf* and information theory and the theoretical research on the meaning of *tf-idf* "clarifies the statistical model on which the different measures are commonly based". This motivation matches the motivation of our paper: We investigate theoretically the assumptions of classical *idf* and Poisson for a better understanding of parameter estimation and combination.

3. FROM DISJOINT TO INDEPENDENT

We define and discuss in this section three probabilities: The frequency-based noise probability (definition 1), the total noise probability for disjoint documents (definition 2). and the noise probability for independent documents (definition 3).

3.1 Binary occurrence, constant containment and disjointness of documents

We show in this section, that the frequency-based noise probability $\frac{n(t)}{N}$ in the *idf* definition can be explained as a total probability with binary term occurrence, constant document containment and disjointness of document containments.

We refer to a probability function as *binary* if for all events the probability is either 1.0 or 0.0. The occurrence probability $P(t|d)$ is binary, if $P(t|d)$ is equal to 1.0 if $t \in d$, and $P(t|d)$ is equal to 0.0, otherwise.

$$P(t|d) \textit{ is binary} : \iff P(t|d) = 1.0 \lor P(t|d) = 0.0$$

We refer to a probability function as *constant* if for all events the probability is equal. The document containment probability reflect the chance that a document occurs in a collection. This containment probability is constant if we have no information about the document containment or we ignore that documents differ in containment. Containment could be derived, for example, from the size, quality, age, links, etc. of a document. For a constant containment in a collection with N documents, $\frac{1}{N}$ is often assumed as the containment probability. We generalise this definition and introduce the constant λ where $0 \leq \lambda \leq N$. The containment of a document d depends on the collection c, this is reflected by the notation $P(d|c)$ used for the containment of a document.

$$P(d|c) \text{ is constant} :\Longleftrightarrow \forall d : P(d|c) = \frac{\lambda}{N}$$

For disjoint documents that cover the whole event space, we set $\lambda = 1$ and obtain $\sum_d P(d|c) = 1.0$. Next, we define the frequency-based noise probability and the total noise probability for disjoint documents. We introduce the event notation *t is noisy* and *t occurs* for making the difference between the noise probability $P(t \textit{ is noisy}|c)$ in a collection and the occurrence probability $P(t \textit{ occurs}|d)$ in a document more explicit, thereby keeping in mind that the noise probability corresponds to the occurrence probability of a term in a collection.

DEFINITION 1. ***The frequency-based term noise probability:***

$$P_{\textit{freq}}(t \textit{ is noisy}|c) := \frac{n(t)}{N}$$

DEFINITION 2. ***The total term noise probability for disjoint documents:***

$$P_{\textit{dis}}(t \textit{ is noisy}|c) := \sum_d P(t \textit{ occurs}|d) \cdot P(d|c)$$

Now, we can formulate a theorem that makes assumptions explicit that explain the classical *idf*.

THEOREM 1. ***IDF assumptions:*** *If the occurrence probability $P(t|d)$ of term t over documents d is binary, and the containment probability $P(d|c)$ of documents d is constant, and document containments are disjoint events, then the noise probability for disjoint documents is equal to the frequency-based noise probability.*

$$P_{\textit{dis}}(t \textit{ is noisy}|c) = P_{\textit{freq}}(t \textit{ is noisy}|c)$$

PROOF. The assumptions are:

$$\forall d : (P(t \textit{ occurs}|d) = 1 \vee P(t \textit{ occurs}|d) = 0) \wedge$$
$$P(d|c) = \frac{\lambda}{N} \wedge$$
$$\sum_d P(d|c) = 1.0$$

We obtain:

$$P_{\textit{dis}}(t \textit{ is noisy}|c) = \sum_{d|t \in d} \frac{1}{N} = \frac{n(t)}{N} = P_{\textit{freq}}(t \textit{ is noisy}|c)$$

□

The above result is not a surprise but it is a mathematical formulation of assumptions that can be used to explain the classical *idf*. The assumptions make explicit that the different types of term occurrence in documents (frequency of a term, importance of a term, position of a term, document part where the term occurs, etc.) and the different types of document containment (size, quality, age, etc.) are ignored, and document containments are considered as disjoint events.

From the assumptions, we can conclude that *idf* (frequency-based noise, respectively) is a relatively simple but strict estimate. Still, *idf* works well. This could be explained by a leverage effect that justifies the binary occurrence and constant containment: The term occurrence for small documents tends to be larger than for large documents, whereas the containment for small documents tends to be smaller than for large documents. From that point of view, *idf* means that $P(t \wedge d|c)$ is constant for all d in which t occurs, and $P(t \wedge d|c)$ is zero otherwise. The occurrence and containment can be term specific. For example, set $P(t \wedge d|c) = 1/N_D(c)$ if t occurs in d, where $N_D(c)$ is the number of documents in collection c (we used before just N). We choose a document-dependent occurrence $P(t|d) := 1/N_T(d)$, i. e. the occurrence probability is equal to the inverse of $N_T(d)$, which is the total number of terms in document d. Next, we choose the containment $P(d|c) := N_T(d)/N_T(c) \cdot N_T(c)/N_D(c)$ where $N_T(d)/N_T(c)$ is a document length normalisation (number of terms in document d divided by the number of terms in collection c), and $N_T(c)/N_D(c)$ is a constant factor of the collection (number of terms in collection c divided by the number of documents in collection c). We obtain $P(t \wedge d|c) = 1/N_D(c)$.

In a *tf-idf*-retrieval function, the *tf*-component reflects the occurrence probability of a term in a document. This is a further explanation why we can estimate the *idf* with a simple $P(t|d)$, since the combined *tf-idf* contains the occurrence probability. The containment probability corresponds to a document normalisation (document length normalisation, pivoted document length) and is normally attached to the *tf*-component or the *tf-idf*-product.

The disjointness assumption is typical for frequency-based probabilities. From a probability theory point of view, we can consider documents as disjoint events, in order to achieve a sound theoretical model for explaining the classical *idf*. But does disjointness reflect the real world where the containment of a document appears to be independent of the containment of another document? In the next section, we replace the disjointness assumption by the independence assumption.

3.2 The upper bound of the noise probability for independent documents

For independent documents, we compute the probability of a disjunction as usual, namely as the complement of the probability of the conjunction of the negated events:

$$\begin{aligned} P(d_1 \vee \ldots \vee d_N) &= 1 - P(\neg d_1 \wedge \ldots \wedge \neg d_N) \\ &= 1 - \prod_d (1 - P(d)) \end{aligned}$$

The noise probability can be considered as the conjunction of the term occurrence and the document containment.

$$P(t \textit{ is noisy}|c) := P(t \textit{ occurs} \wedge (d_1 \vee \ldots \vee d_N)|c)$$

For disjoint documents, this view of the noise probability led to definition 2. For independent documents, we use now the conjunction of negated events.

DEFINITION 3. ***The term noise probability for independent documents:***

$$P_{\textit{in}}(t \textit{ is noisy}|c) := \prod_d (1 - P(t \textit{ occurs}|d) \cdot P(d|c))$$

With binary occurrence and a constant containment $P(d|c) := \lambda/N$, we obtain the term noise of a term t that occurs in $n(t)$ documents:

$$P_{\textit{in}}(t \textit{ is noisy}|c) = 1 - \left(1 - \frac{\lambda}{N}\right)^{n(t)}$$

For binary occurrence and disjoint documents, the containment probability was $1/N$. Now, with independent documents, we can use λ as a collection parameter that controls the average containment probability. We show through the next theorem that the upper bound of the noise probability depends on λ.

THEOREM 2. ***The upper bound of being noisy:*** *If the occurrence $P(t|d)$ is binary, and the containment $P(d|c)$ is constant, and document containments are independent events, then $1 - e^{-\lambda}$ is the upper bound of the noise probability.*

$$\forall t : P_{in}(t \text{ is noisy}|c) < 1 - e^{-\lambda}$$

PROOF. The upper bound of the independent noise probability follows from the limit $\lim_{N\to\infty}(1 + \frac{x}{N})^N = e^x$ (see any comprehensive math book, for example, [5], for the convergence equation of the Euler function). With $x = -\lambda$, we obtain:

$$\lim_{N\to\infty} \left(1 - \frac{\lambda}{N}\right)^N = e^{-\lambda}$$

For the term noise, we have:

$$P_{in}(t \text{ is noisy}|c) = 1 - \left(1 - \frac{\lambda}{N}\right)^{n(t)}$$

$P_{in}(t \text{ is noisy}|c)$ is strictly monotonous: The noise of a term t_n is less than the noise of a term t_{n+1}, where t_n occurs in n documents and t_{n+1} occurs in $n + 1$ documents. Therefore, a term with $n = N$ has the largest noise probability. For a collection with infinite many documents, the upper bound of the noise probability for terms t_N that occur in all documents becomes:

$$\begin{aligned}\lim_{N\to\infty} P_{in}(t_N \text{ is noisy}) &= \lim_{N\to\infty} 1 - \left(1 - \frac{\lambda}{N}\right)^N \\ &= 1 - e^{-\lambda}\end{aligned}$$

□

By applying an independence rather a disjointness assumption, we obtain the probability e^{-1} that a term is not noisy even if the term does occur in all documents. In the disjoint case, the noise probability is one for a term that occurs in all documents.

If we view $P(d|c) := \lambda/N$ as the average containment, then λ is large for a term that occurs mostly in large documents, and λ is small for a term that occurs mostly in small documents. Thus, the noise of a term t is large if t occurs in $n(t)$ large documents and the noise is smaller if t occurs in small documents. Alternatively, we can assume a constant containment and a term-dependent occurrence. If we assume $P(d|c) := 1$, then $P(t|d) := \lambda/N$ can be interpreted as the average probability that t represents a document. The common assumption is that the average containment or occurrence probability is proportional to $n(t)$. However, here is additional potential: The statistical laws (see [3] on Luhn and Zipf) indicate that the average probability could follow a normal distribution, i. e. small probabilities for small $n(t)$ and large $n(t)$, and larger probabilities for medium $n(t)$.

For the monotonous case we investigate here, the noise of a term with $n(t) = 1$ is equal to $1 - (1 - \lambda/N) = \lambda/N$ and the noise of a term with $n(t) = N$ is close to $1 - e^{-\lambda}$. In the next section, we relate the value $e^{-\lambda}$ to information theory.

3.3 The probability of a maximal informative signal

The probability e^{-1} is special in the sense that a signal with that probability is a signal with maximal information as derived from the entropy definition. Consider the definition of the entropy contribution $H(t)$ of a signal t.

$$H(t) := P(t) \cdot - \ln P(t)$$

We form the first derivation for computing the optimum.

$$\begin{aligned}\frac{\partial H(t)}{\partial P(t)} &= -\ln P(t) + \frac{-1}{P(t)} \cdot P(t) \\ &= -(1 + \ln P(t))\end{aligned}$$

For obtaining optima, we use:

$$0 = -(1 + \ln P(t))$$

The entropy contribution $H(t)$ is maximal for $P(t) = e^{-1}$. This result does not depend on the base of the logarithm as we see next:

$$\begin{aligned}\frac{\partial H(t)}{\partial P(t)} &= -\log_b P(t) + \frac{-1}{P(t) \cdot \ln b} \cdot P(t) \\ = & -\left(\frac{1}{\ln b} + \log_b P(t)\right) = -\left(\frac{1 + \ln P(t)}{\ln b}\right)\end{aligned}$$

We summarise this result in the following theorem:

THEOREM 3. ***The probability of a maximal informative signal:*** *The probability $P_{max} = e^{-1} \approx 0.37$ is the probability of a maximal informative signal. The entropy of a maximal informative signal is $H_{max} = e^{-1}$.*

PROOF. The probability and entropy follow from the derivation above. □

The complement of the maximal noise probability is $e^{-\lambda}$ and we are looking now for a generalisation of the entropy definition such that $e^{-\lambda}$ is the probability of a maximal informative signal. We can generalise the entropy definition by computing the integral of $\lambda + \ln P(t)$, i. e. this derivation is zero for $e^{-\lambda}$. We obtain a generalised entropy:

$$\int -(\lambda + \ln P(t))\ d(P(t)) = P(t) \cdot (1 - \lambda - \ln P(t))$$

The generalised entropy corresponds for $\lambda = 1$ to the classical entropy. By moving from disjoint to independent documents, we have established a link between the complement of the noise probability of a term that occurs in all documents and information theory. Next, we link independent documents to probability theory.

4. THE LINK TO PROBABILITY THEORY

We review for independent documents three concepts of probability theory: possible worlds, binomial distribution and Poisson distribution.

4.1 Possible Worlds

Each conjunction of document events (for each document, we consider two document events: the document can be true or false) is associated with a so-called *possible world*. For example, consider the eight possible worlds for three documents ($N = 3$).

world w	conjunction
w_7	$d_1 \wedge d_2 \wedge d_3$
w_6	$d_1 \wedge d_2 \wedge \neg d_3$
w_5	$d_1 \wedge \neg d_2 \wedge d_3$
w_4	$d_1 \wedge \neg d_2 \wedge \neg d_3$
w_3	$\neg d_1 \wedge d_2 \wedge d_3$
w_2	$\neg d_1 \wedge d_2 \wedge \neg d_3$
w_1	$\neg d_1 \wedge \neg d_2 \wedge d_3$
w_0	$\neg d_1 \wedge \neg d_2 \wedge \neg d_3$

With each world w, we associate a probability $\mu(w)$, which is equal to the product of the single probabilities of the document events.

world w	probability $\mu(w)$
w_7	$\left(\frac{\lambda}{N}\right)^3 \cdot \left(1 - \frac{\lambda}{N}\right)^0$
w_6	$\left(\frac{\lambda}{N}\right)^2 \cdot \left(1 - \frac{\lambda}{N}\right)^1$
w_5	$\left(\frac{\lambda}{N}\right)^2 \cdot \left(1 - \frac{\lambda}{N}\right)^1$
w_4	$\left(\frac{\lambda}{N}\right)^1 \cdot \left(1 - \frac{\lambda}{N}\right)^2$
w_3	$\left(\frac{\lambda}{N}\right)^2 \cdot \left(1 - \frac{\lambda}{N}\right)^1$
w_2	$\left(\frac{\lambda}{N}\right)^1 \cdot \left(1 - \frac{\lambda}{N}\right)^2$
w_1	$\left(\frac{\lambda}{N}\right)^1 \cdot \left(1 - \frac{\lambda}{N}\right)^2$
w_0	$\left(\frac{\lambda}{N}\right)^0 \cdot \left(1 - \frac{\lambda}{N}\right)^3$

The sum over the possible worlds in which k documents are true and $N-k$ documents are false is equal to the probability function of the binomial distribution, since the binomial coefficient yields the number of possible worlds in which k documents are true.

4.2 Binomial distribution

The binomial probability function yields the probability that k of N events are true where each event is true with the single event probability p.

$$P(k) := \mathit{binom}(N, k, p) := \binom{N}{k} \; p^k \; (1-p)^{N-k}$$

The single event probability is usually defined as $p := \lambda/N$, i. e. p is inversely proportional to N, the total number of events. With this definition of p, we obtain for an infinite number of documents the following limit for the product of the binomial coefficient and p^k:

$$\begin{aligned} &\lim_{N\to\infty} \binom{N}{k} \; p^k = \\ &= \lim_{N\to\infty} \frac{N \cdot (N-1) \cdot \ldots \cdot (N-k+1)}{k!} \left(\frac{\lambda}{N}\right)^k = \frac{\lambda^k}{k!} \end{aligned}$$

The limit is close to the actual value for $k << N$. For large k, the actual value is smaller than the limit.

The limit of $(1-p)^{N-k}$ follows from the limit $\lim_{N\to\infty}(1 + \frac{x}{N})^N = e^x$.

$$\begin{aligned} \lim_{N\to\infty} (1-p)^{N\text{-}k} &= \lim_{N\to\infty} \left(1 - \frac{\lambda}{N}\right)^{N-k} \\ &= \lim_{N\to\infty} \left(e^{-\lambda} \cdot \left(1 - \frac{\lambda}{N}\right)^{-k}\right) = e^{-\lambda} \end{aligned}$$

Again, the limit is close to the actual value for $k << N$. For large k, the actual value is larger than the limit.

4.3 Poisson distribution

For an infinite number of events, the Poisson probability function is the limit of the binomial probability function.

$$\begin{aligned} \lim_{N\to\infty} \mathit{binom}(N, k, p) &= \frac{\lambda^k}{k!} \cdot e^{-\lambda} \\ P(k) = \mathit{poisson}(k, \lambda) &:= \frac{\lambda^k}{k!} \cdot e^{-\lambda} \end{aligned}$$

The probability $\mathit{poisson}(0, 1)$ is equal to e^{-1}, which is the probability of a maximal informative signal. This shows the relationship of the Poisson distribution and information theory.

After seeing the convergence of the binomial distribution, we can choose the Poisson distribution as an approximation of the independent term noise probability. First, we define the Poisson noise probability:

Definition 4. ***The Poisson term noise probability:***

$$P_{poi}(t \textit{ is noisy}|c) := e^{-\lambda} \cdot \sum_{k=1}^{n(t)} \frac{\lambda^k}{k!}$$

For independent documents, the Poisson distribution approximates the probability of the disjunction for large $n(t)$, since the independent term noise probability is equal to the sum over the binomial probabilities where at least one of $n(t)$ document containment events is true.

$$\begin{aligned} P_{in}(t \textit{ is noisy}|c) &= \sum_{k=1}^{n(t)} \binom{n(t)}{k} \; p^k \; (1-p)^{N-k} \\ P_{in}(t \textit{ is noisy}|c) &\approx P_{poi}(t \textit{ is noisy}|c) \end{aligned}$$

We have defined a frequency-based and a Poisson-based probability of being noisy, where the latter is the limit of the independence-based probability of being noisy. Before we present in the final section the usage of the noise probability for defining the probability of being informative, we emphasise in the next section that the results apply to the collection space as well as to the the document space.

5. THE COLLECTION SPACE AND THE DOCUMENT SPACE

Consider the dual definitions of retrieval parameters in table 1. We associate a collection space $D \times T$ with a collection c where D is the set of documents and T is the set of terms in the collection. Let $N_D := |D|$ and $N_T := |T|$ be the number of documents and terms, respectively. We consider a document as a subset of T and a term as a subset of D. Let $n_T(d) := |\{t|d \in t\}|$ be the number of terms that occur in the document d, and let $n_D(t) := |\{d|t \in d\}|$ be the number of documents that contain the term t.

In a dual way, we associate a document space $L \times T$ with a document d where L is the set of locations (also referred to as positions, however, we use the letters L and l and not P and p for avoiding confusion with probabilities) and T is the set of terms in the document. The document dimension in a collection space corresponds to the location (position) dimension in a document space.

The definition makes explicit that the classical notion of term frequency of a term in a document (also referred to as the within-document term frequency) actually corresponds to the location frequency of a term in a document. For the

space	collection	document
dimensions	documents and terms	locations and terms
document/location frequency	$n_D(t, c)$: Number of documents in which term t occurs in collection c $N_D(c)$: Number of documents in collection c	$n_L(t, d)$: Number of locations (positions) at which term t occurs in document d $N_L(d)$: Number of locations (positions) in document d
term frequency	$n_T(d, c)$: Number of terms that document d contains in collection c $N_T(c)$: Number of terms in collection c	$n_T(l, d)$: Number of terms that location l contains in document d $N_T(d)$: Number of terms in document d
noise/occurrence containment	$P(t\|c)$ (term noise) $P(d\|c)$ (document)	$P(t\|d)$ (term occurrence) $P(l\|d)$ (location)
informativeness conciseness	$-\ln P(t\|c)$ $-\ln P(d\|c)$	$-\ln P(t\|d)$ $-\ln P(l\|d)$

P(informative)
P(concise)
	$\ln(P(t\|c))/\ln(P(t_{min}, c))$
	$\ln(P(d\|c))/\ln(P(d_{min}\|c))$
		$\ln(P(t\|d))/\ln(P(t_{min}, d))$
		$\ln(P(l\|d))/\ln(P(l_{min}\|d))$

Table 1: Retrieval parameters

actual term frequency value, it is common to use the maximal occurrence (number of locations; let *lf* be the location frequency).

$$tf(t, d) := lf(t, d) := \frac{P_{freq}(t\ occurs|d)}{P_{freq}(t_{max}\ occurs|d)} = \frac{n_L(t, d)}{n_L(t_{max}, d)}$$

A further duality is between informativeness and conciseness (shortness of documents or locations): informativeness is based on occurrence (noise), conciseness is based on containment.

We have highlighted in this section the duality between the collection space and the document space. We concentrate in this paper on the probability of a term to be noisy and informative. Those probabilities are defined in the collection space. However, the results regarding the term noise and informativeness apply to their dual counterparts: term occurrence and informativeness in a document. Also, the results can be applied to containment of documents and locations.

6. THE PROBABILITY OF BEING INFORMATIVE

We showed in the previous sections that the disjointness assumption leads to frequency-based probabilities and that the independence assumption leads to Poisson probabilities. In this section, we formulate a frequency-based definition and a Poisson-based definition of the probability of being informative and then we compare the two definitions.

Definition 5. ***The frequency-based probability of being informative:***

$$P_{freq}(t\ is\ informative|c) := \frac{-\ln\frac{n(t)}{N}}{-\ln\frac{1}{N}}$$
$$= \quad -\log_N \frac{n(t)}{N} = 1 - \log_N n(t) = 1 - \frac{\ln n(t)}{\ln N}$$

We define the Poisson-based probability of being informative analogously to the frequency-based probability of being informative (see definition 5).

Definition 6. ***The Poisson-based probability of being informative:***

$$P_{poi}(t\ is\ informative|c) := \frac{-\ln\left(e^{-\lambda}\cdot\sum_{k=1}^{n(t)}\frac{\lambda^k}{k!}\right)}{-\ln(e^{-\lambda}\cdot\lambda)}$$
$$= \quad \frac{\lambda - \ln\sum_{k=1}^{n(t)}\frac{\lambda^k}{k!}}{\lambda - \ln\lambda}$$

For the sum expression, the following limit holds:

$$\lim_{n(t)\to\infty}\sum_{k=1}^{n(t)}\frac{\lambda^k}{k!} = e^{\lambda} - 1$$

For $\lambda >> 1$, we can alter the noise and informativeness Poisson by starting the sum from 0, since $e^{\lambda} >> 1$. Then, the minimal Poisson informativeness is $poisson(0, \lambda) = e^{-\lambda}$. We obtain a simplified Poisson probability of being informative:

$$P_{poi}(t\ is\ informative|c) \approx \frac{\lambda - \ln\sum_{k=0}^{n(t)}\frac{\lambda^k}{k!}}{\lambda}$$
$$= \quad 1 - \frac{\ln\sum_{k=0}^{n(t)}\frac{\lambda^k}{k!}}{\lambda}$$

The computation of the Poisson sum requires an optimisation for large $n(t)$. The implementation for this paper exploits the nature of the Poisson density: The Poisson density yields only values significantly greater than zero in an interval around λ.

Consider the illustration of the noise and informativeness definitions in figure 1. The probability functions displayed are summarised in figure 2 where the simplified Poisson is used in the noise and informativeness graphs. The frequency-based noise corresponds to the linear solid curve in the noise figure. With an independence assumption, we obtain the curve in the lower triangle of the noise figure. By changing the parameter $p := \lambda/N$ of the independence probability, we can lift or lower the independence curve. The noise figure shows the lifting for the value $\lambda := \ln N \approx 9.2$. The setting $\lambda = \ln N$ is special in the sense that the frequency-based and the Poisson-based informativeness have the same denominator, namely $\ln N$, and the Poisson sum converges to λ. Whether we can draw more conclusions from this setting is an open question.

We can conclude, that the lifting is desirable if we know for a collection that terms that occur in relatively few doc-

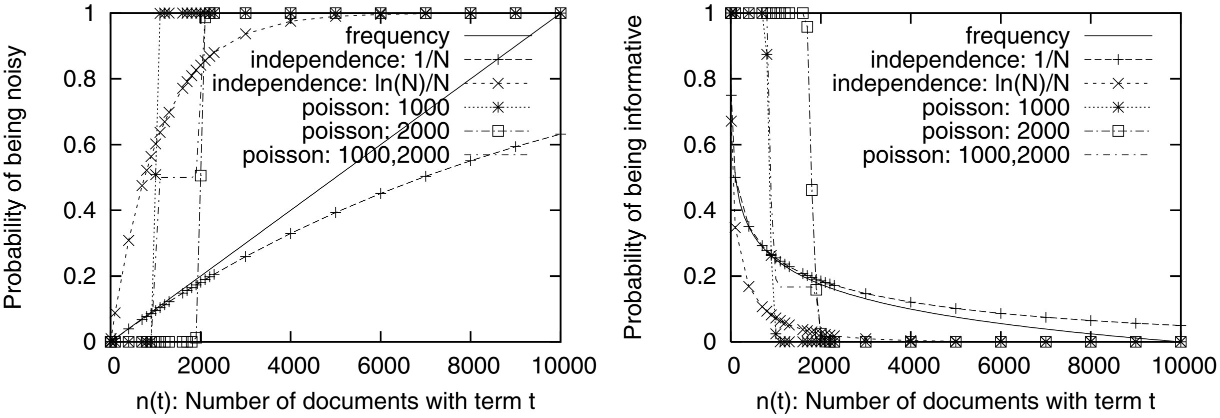

Figure 1: Noise and Informativeness

Probability function		Noise	Informativeness
Frequency P_{freq}	Def	$n(t)/N$	$\ln(n(t)/N)/\ln(1/N)$
	Interval	$1/N \leq P_{freq} \leq 1.0$	$0.0 \leq P_{freq} \leq 1.0$
Independence P_{in}	Def	$1-(1-p)^{n(t)}$	$\ln(1-(1-p)^{n(t)})/\ln(p)$
	Interval	$p \leq P_{in} < 1-e^{-\lambda}$	$\ln(p) \leq P_{in} \leq 1.0$
Poisson P_{poi}	Def	$e^{-\lambda}\sum_{k=1}^{n(t)}\frac{\lambda^k}{k!}$	$(\lambda - \ln\sum_{k=1}^{n(t)}\frac{\lambda^k}{k!})/(\lambda - \ln\lambda)$
	Interval	$e^{-\lambda}\cdot\lambda \leq P_{poi} < 1-e^{-\lambda}$	$(\lambda - \ln(e^{\lambda}-1))/(\lambda-\ln\lambda) \leq P_{poi} \leq 1.0$
Poisson P_{poi} simplified	Def	$e^{-\lambda}\sum_{k=0}^{n(t)}\frac{\lambda^k}{k!}$	$(\lambda - \ln\sum_{k=0}^{n(t)}\frac{\lambda^k}{k!})/\lambda$
	Interval	$e^{-\lambda} \leq P_{poi} < 1.0$	$0.0 < P_{poi} \leq 1.0$

Figure 2: Probability functions

uments are no guarantee for finding relevant documents, i. e. we assume that rare terms are still relatively noisy. On the opposite, we could lower the curve when assuming that frequent terms are not too noisy, i. e. they are considered as being still significantly discriminative.

The Poisson probabilities approximate the independence probabilities for large $n(t)$; the approximation is better for larger λ. For $n(t) < \lambda$, the noise is zero whereas for $n(t) > \lambda$ the noise is one. This radical behaviour can be smoothened by using a multi-dimensional Poisson distribution. Figure 1 shows a Poisson noise based on a two-dimensional Poisson:

$$poisson(k, \lambda_1, \lambda_2) := \pi \cdot e^{-\lambda_1} \cdot \frac{\lambda_1^k}{k!} + (1-\pi) \cdot e^{-\lambda_2} \cdot \frac{\lambda_2^k}{k!}$$

The two dimensional Poisson shows a plateau between $\lambda_1 = 1000$ and $\lambda_2 = 2000$, we used here $\pi = 0.5$. The idea behind this setting is that terms that occur in less than 1000 documents are considered to be not noisy (i.e. they are informative), that terms between 1000 and 2000 are half noisy, and that terms with more than 2000 are definitely noisy.

For the informativeness, we observe that the radical behaviour of Poisson is preserved. The plateau here is approximately at 1/6, and it is important to realise that this plateau is not obtained with the multi-dimensional Poisson noise using $\pi = 0.5$. The logarithm of the noise is normalised by the logarithm of a very small number, namely $0.5 \cdot e^{-1000} + 0.5 \cdot e^{-2000}$. That is why the informativeness will be only close to one for very little noise, whereas for a bit of noise, informativeness will drop to zero. This effect can be controlled by using small values for π such that the noise in the interval $[\lambda_1; \lambda_2]$ is still very little. The setting $\pi = e^{-2000/6}$ leads to noise values of approximately $e^{-2000/6}$ in the interval $[\lambda_1; \lambda_2]$, the logarithms lead then to 1/6 for the informativeness.

The indepence-based and frequency-based informativeness functions do not differ as much as the noise functions do. However, for the indepence-based probability of being informative, we can control the average informativeness by the definition $p := \lambda/N$ whereas the control on the frequency-based is limited as we address next.

For the frequency-based *idf*, the gradient is monotonously decreasing and we obtain for different collections the same distances of *idf*-values, i. e. the parameter N does not affect the distance. For an illustration, consider the distance between the value $idf(t_{n+1})$ of a term t_{n+1} that occurs in $n+1$ documents, and the value $idf(t_n)$ of a term t_n that occurs in n documents.

$$idf(t_{n+1}) - idf(t_n) \quad = \quad \ln\frac{n}{n+1}$$

The first three values of the distance function are:

$$idf(t_2) - idf(t_1) = \ln(1/(1+1)) = 0.69$$
$$idf(t_3) - idf(t_2) = \ln(1/(2+1)) = 0.41$$
$$idf(t_4) - idf(t_3) = \ln(1/(3+1)) = 0.29$$

For the Poisson-based informativeness, the gradient decreases first slowly for small $n(t)$, then rapidly near $n(t) \approx \lambda$ and then it grows again slowly for large $n(t)$.

In conclusion, we have seen that the Poisson-based definition provides more control and parameter possibilities than the frequency-based definition does. Whereas more control and parameter promises to be positive for the personalisation of retrieval systems, it bears at the same time the danger of just too many parameters. The framework presented in this paper raises the awareness about the probabilistic and information-theoretic meanings of the parameters. The parallel definitions of the frequency-based probability and the Poisson-based probability of being informative made the underlying assumptions explicit. The frequency-based probability can be explained by binary occurrence, constant containment and disjointness of documents. Independence of documents leads to Poisson, where we have to be aware that Poisson approximates the probability of a disjunction for a large number of events, but not for a small number. This theoretical result explains why experimental investigations on Poisson (see [7]) show that a Poisson estimation does work better for frequent (bad, noisy) terms than for rare (good, informative) terms.

In addition to the collection-wide parameter setting, the framework presented here allows for document-dependent settings, as explained for the independence probability. This is in particular interesting for heterogeneous and structured collections, since documents are different in nature (size, quality, root document, sub document), and therefore, binary occurrence and constant containment are less appropriate than in relatively homogeneous collections.

7. SUMMARY

The definition of the probability of being informative transforms the informative interpretation of the *idf* into a probabilistic interpretation, and we can use the *idf*-based probability in probabilistic retrieval approaches. We showed that the classical definition of the noise (document frequency) in the inverse document frequency can be explained by three assumptions: the term within-document occurrence probability is binary, the document containment probability is constant, and the document containment events are disjoint. By explicitly and mathematically formulating the assumptions, we showed that the classical definition of *idf* does not take into account parameters such as the different nature (size, quality, structure, etc.) of documents in a collection, or the different nature of terms (coverage, importance, position, etc.) in a document. We discussed that the absence of those parameters is compensated by a leverage effect of the within-document term occurrence probability and the document containment probability.

By applying an independence rather a disjointness assumption for the document containment, we could establish a link between the noise probability (term occurrence in a collection), information theory and Poisson. From the frequency-based and the Poisson-based probabilities of being noisy, we derived the frequency-based and Poisson-based probabilities of being informative. The frequency-based probability is relatively smooth whereas the Poisson probability is radical in distinguishing between noisy or not noisy, and informative or not informative, respectively. We showed how to smoothen the radical behaviour of Poisson with a multidimensional Poisson.

The explicit and mathematical formulation of *idf*- and Poisson-assumptions is the main result of this paper. Also, the paper emphasises the duality of *idf* and *tf*, collection space and document space, respectively. Thus, the result applies to term occurrence and document containment in a collection, and it applies to term occurrence and position containment in a document. This theoretical framework is useful for understanding and deciding the parameter estimation and combination in probabilistic retrieval models. The links between indepence-based noise as document frequency, probabilistic interpretation of *idf*, information theory and Poisson described in this paper may lead to variable probabilistic *idf* and *tf* definitions and combinations as required in advanced and personalised information retrieval systems.

Acknowledgment: I would like to thank Mounia Lalmas, Gabriella Kazai and Theodora Tsikrika for their comments on the as they said "heavy" pieces. My thanks also go to the meta-reviewer who advised me to improve the presentation to make it less "formidable" and more accessible for those "without a theoretic bent". This work was funded by a research fellowship from Queen Mary University of London.

8. REFERENCES

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

[2] G. Amati and C. J. Rijsbergen. Term frequency normalization via Pareto distributions. In *24th BCS-IRSG European Colloquium on IR Research, Glasgow, Scotland*, 2002.

[3] R. K. Belew. *Finding out about*. Cambridge University Press, 2000.

[4] A. Bookstein and D. Swanson. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, 25:312–318, 1974.

[5] I. N. Bronstein. *Taschenbuch der Mathematik*. Harri Deutsch, Thun, Frankfurt am Main, 1987.

[6] K. Church and W. Gale. Poisson mixtures. *Natural Language Engineering*, 1(2):163–190, 1995.

[7] K. W. Church and W. A. Gale. Inverse document frequency: A measure of deviations from poisson. In *Third Workshop on Very Large Corpora, ACL Anthology*, 1995.

[8] T. Lafouge and C. Michel. Links between information construction and information gain: Entropy and bibliometric distribution. *Journal of Information Science*, 27(1):39–49, 2001.

[9] E. Margulis. N-poisson document modelling. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 177–189, 1992.

[10] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, London, et al., 1994. Springer-Verlag.

[11] S. Wong and Y. Yao. An information-theoric measure of term specificity. *Journal of the American Society for Information Science*, 43(1):54–61, 1992.

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