MANAGEMENT OF UNCERTAINTY AND IMPRECISION IN MULTIMEDIA INFORMATION SYSTEMS: INTRODUCING THIS SPECIAL ISSUE

CHARLES M.E.E. PETERS
Department of Computer Science
University of Nijmegen
Nijmegen, The Netherlands
charlesp@cs.kun.nl

CORNELIS H.A. KOSTER
Department of Computer Science
University of Nijmegen
Nijmegen, The Netherlands
kees@cs.kun.nl

Received (April 2002)
Revised (August and November 2002)

Abstract
This paper discusses the notion of Uncertainty, which has a prominent place in the theory and experimental practice of modern Physics. It argues that the awareness of Uncertainty may also be of tremendous importance to the field of Information Retrieval, and in particular Text Categorization.

As an application of Uncertainty in Text Categorization, a new criterion for Term Selection is described, which is based on the Uncertainty in Term Frequency across categories. This criterion allows to distinguish between low-quality (or “noisy”) and high-quality (“stiff”) terms.

We describe an experiment investigating the effect of eliminating noisy and stiff terms in the context of text classification. In the experiment we applied the Rocchio and Winnow classification algorithms to a collection of newspaper items, a mono-classified subset of the well-known Reuters 21578 corpus.

This investigation shows that both the local elimination of noisy terms and the global elimination of stiff terms can be used for Term Selection in Text Categorization.

Keywords: Uncertainty, Term Selection, Text Categorization, noisy terms, stiff terms

1. Introduction

Most Information Retrieval applications are based on some form of scoring. This holds in particular for Text Categorization (also known as text classification; for an overview see 17), where documents are to be assigned to those categories
for which they obtain the highest score from some categorization algorithm. The score is usually some real number, representing the degree of its similarity to other documents in that category rather than a true probability. This makes it hard to compare scores given by different algorithms. Furthermore, usually no variance in the score is computed, neither is there any other indication of the reliability of that score.

This is in great contrast to other fields like Physics, which have long relied on the adage “to measure is to know”, but have recently come to realize that it is also important to quantify how much you do not know. It is our conviction that measurements made in Information Retrieval can benefit greatly by embracing the notion of Uncertainty as it is used in modern measurement theory.

1.1. Overview

We first provide a short introduction to theory of measurement and uncertainty, and then (in section 2) demonstrate its relevance to Information Retrieval in general and Text Categorization in particular. We derive a new term selection criterion in section 3 and describe some experiments using the Rocchio and Winnow Algorithms in classifying a collection of newspaper items (section 4) to investigate the effectiveness of uncertainty-based term selection, discovering two viable term selection criteria: local noise elimination (section 5) and stiff term elimination (section 6). In the last section, we point out some other potential applications of Uncertainty in Information Retrieval.

2. General aspects of uncertainty

Although in Informatics in general, and in Information Retrieval in particular, there appears to be ample scope for application of Uncertainty Theory, this theory has been studied mostly in metrology, the philosophy of measurement in Physics. It is therefore instructive to look at metrology as an example, in particular at the process in Physics called measurement. After that, we shall look at the more general aspects of uncertainty, its mathematical representation and possible technical methods for increasing the information and reducing uncertainty in the quantifiable case.

2.1. Measurement and uncertainty

Apart from the fundamental natural constants reflecting the atomic nature of the world, as known from Quantum mechanics, and the implicit fundamental Heisenberg uncertainty relation caused by the indistinguishable character of the elementary particles, called spin statistics, even classical real-world Physics has to take the concept of uncertainty into account. The reason is evident: since there will always be missing information, there will be uncertainty in the quantification. Hence every measurement represents essentially an estimate. Consequently, a measurement
Uncertainty and Term selection in Text Categorization

The process should give as its result not only the measured value, an estimate \( v \) of the “real” value, but also an uncertainty \( u(v) \) in this value.

The uncertainty delimits, together with the measured value, the interval in which the “real” value is expected to lie. From an information-theoretic point of view, the measurement value represents the information gained about the object under investigation and the uncertainty represents the missing information about the object. This interpretation is best illustrated by looking at the entropy (the amount of missing information) of the normal distribution:

\[
\int_{-\infty}^{\infty} \left( -\frac{(x-v)^2}{2u^2(v)} - \log u(v) - \frac{\log \pi}{2} - \frac{\log 2}{2} \right) e^{-\frac{(x-v)^2}{2u^2(v)}} dx = 2 \log u(v) + \log \pi + \log 2 + 1
\]

This is clearly only dependent on the uncertainty. In general, however, the use of the normal distribution is not appropriate since important boundary conditions may exist on the possible values of the measurand. The most common of these are: strict positiveness of the measurand or boundedness of the measurand between a minimal and a maximal value. Such constraints provide important additional information, making other distributions than the normal distribution more appropriate.

For the associated distribution function, representing the measurement result in the presence of such boundary conditions, the worst-case distribution should be taken for the situation under investigation. This worst-case distribution is determined by demanding stationarity of the entropy of the distribution under the given information and the constraints. The normal distribution is appropriate when no boundary conditions are given, and furthermore serves as an interesting case for illustrative purposes.

In case only the mean and standard deviation are given, with no additional constraints, two types of uncertainty can be distinguished:

1. **TYPE A** – statistically determined uncertainties. In case it is certain that the measured value is single-valued and independent of external influence parameters (for example a controlled temperature), the standard deviation itself is taken as the uncertainty and the corresponding distribution is taken as an approximation of the representation. In this case the measurement result is given by the estimated value (the mean) and the uncertainty in this estimate.

2. **TYPE B** – All other cases, i.e. if the uncertainty can not be estimated purely by statistical means but it is known that the measurand has some natural variability or is significantly influenced by external parameters and the measurement result is a mean value with a standard deviation. In this case the uncertainty is two times the statistical uncertainty, as propagated by all influencing factors causing uncertainty, in the mean. The worst-case
distribution is then given by a distribution with a standard deviation equal to the estimated uncertainty. This distribution can be interpreted as the probability density for the real mean of the measurement. The measurement result then comprises: the estimated mean, the estimated uncertainty in the mean and the estimated standard deviation.

In both these cases, the two-times-the-uncertainty interval is interpreted as the 95% confidence interval within which the estimated value is expected to lie. For this reason the uncertainty is often stated with a so called k-factor of two $^1$.

In some cases a k-factor of two is not applicable, because it would imply that the specified interval can cross the boundary conditions, so that there are real values possible lying outside the boundaries, which was stated to be impossible.

2.2. An example from Physics

Constraints on measurands are mostly implied by the measurement method. To give an obvious example from Physics, the length of a rod is dependent on temperature. So for measuring its length, its temperature has to be measured and controlled. The measurement of the temperature implies an uncertainty in the temperature. Its contribution to the uncertainty in the length of the rod has to be taken into account in the uncertainty in the measurement result of the length of the rod.

The fact that the rod’s length $l$ certainly is greater than zero, $l > 0$, implies for the worst-case distribution a second constraint. The associated worst-case distribution is the so called Pierson III distribution. In Physics, length measurement techniques may be so accurate (with relative uncertainties in the order of $u(l) = 10^{-6} - 10^{-12}$) that the normal distribution is a nearly perfect approximation, and the boundary conditions give hardly any additional information.

In general all measurements can be represented as

$$\{ \text{reference standard} \mid \text{measurement method} \mid \text{measurand} \}$$

with:

- reference standard: an object providing the property to be measured with known value as a reference.
- measurand: the object of which the value of the property is unknown but is to be determined.
- measurement method: an instrument constructed using:
  - knowledge about the so called influence parameters of the property
  - knowledge about the physical laws governing the dependency between the property and the influence parameters

to produce a quantitative estimate of the measurand in terms of the reference standard.
However, before such a measurement can be performed, many processes have already taken place, mostly subjective decisions. These are the cause of qualitative uncertainties.

2.3. Qualitative aspects of Uncertainty

Uncertainties in metrology, the realization of the fundamental constants and the SI-system, have in the last decades been greatly reduced. In measuring a mass, uncertainties in the order of $10^{-9}$ are normal; for electrical standards and length, the other three fundamental standards, meter, time and Ampere, can be realized with even smaller uncertainties. Why is that the case?

Physics, studying and describing nature, may be complicated, but nature itself is simple, based on relatively simple structures or laws. Essential to its understanding is the concept of entropy, or missing information. In this consideration one should take into account that Physics tries to describe, represent and manipulate (technology) the “real-world object”, or better “our human ability to observe them and manipulate them”: two quite different concepts, the objective real world and our subjective observations and our ability to observe . realized with relative uncertainties. Without the invention of the microscope, bacteria would still exist, but we wouldn’t know anything about them. So, acceptance of the concept of a real-world and acceptance of our restricted ability to observe it is the basis for reducing uncertainty and increasing observability. Given the fact that Physics deals with observability, a second factor comes into play, the fact which distinguishes the human species from other species: a common memory called speaking and (more recently) writing and counting.

The concept of representation is an other source of uncertainty. At the moment an observable property is represented by a symbol, information is lost. Hardly anybody realizes when shopping that the Kilo is a piece of platinum, rather arbitrarily but conveniently chosen, which is kept at the Paviljon de Breteuille at the BIPM in Sevre, France, as a “mondial reference standard”.

Implicitly, the concept of a property and its proper choice is quite important, whether quantifiable or not. Certainly the shine of gold, caused by its physical properties, but in general it is economically preferred to weigh its mass in order to estimate its value. A property chosen so well that after 4500 years it is so commonly used in daily life that nobody notices it.

In general, subjective choices are made about the properties that people want to know, neglecting all other information. Another source of uncertainty.

In order to be able to make a measurement, a (sufficiently objective) observable property has to be found and a standard to collectively refer to. Since such a standard has to be produced as an object, it certainly also has some uncertainty, since it is produced using measurement techniques in its construction.

Further, there is only one mondial base kilo in Sevre, so every country has to have its own standard, which has to be compared to the kilo at Sevre, and every kilo in the country, even a kitchen balance, to this national kilo. This is done by
Uncertainty and Term selection in Text Categorization

a measurement process called calibration. But in each calibration in the chain the uncertainty increases with a factor of 10.

Back to the question why physical measurements are now so extremely accurate, with such a small uncertainty (something that, at the moment, can not be said of Information Retrieval). Remembering that uncertainty and entropy deal with missing information, the quite amusing but seemingly useless story above gives the answer.

When choosing the right properties, the biggest lack of information is the lack of knowledge of the (mathematical) relations between them. Physics has only reduced uncertainties to the order presently reached by 4500 years of observing, studying and recording exactly these relations, mostly for economic or technological reasons.

The biggest gain in information can be obtained by gaining knowledge about the relation between things we observe and therefore reducing the amount of missing information, i.e. the uncertainty. The astonishing small relative uncertainties, ranging from $10^{-9}$ to $10^{-14}$ in measurements in Physics, or even better, are not just a consequence of better measurement procedures. They are a consequence of the reduction in uncertainty achieved by reducing the amount of missing information using mathematical knowledge about relations between observations and the mathematical laws governing them.

The technological revolution is not the origin of our knowledge revolution. The knowledge, i.e. the possibility to reduce uncertainty and control even electrons, is the necessary factor allowing technology to make progress.

2.4. The representation of quantitative uncertainties

When first confronted with uncertainty, most Physicists immediately recognize the concept of distribution function. A quite general reaction is: simply statistics!

The above section should show that that’s not all. The distribution functions are representations to be used mathematically, but some uncertainties can not be measured, neither statistically nor physically, but have to be estimated by human experience, because they are not measurable, i.e. quantifiable by an instrument.

While symbolic logic is essential in the analytical process of reasoning and gaining knowledge, one should realize that the acceptance of uncertainty in real-world observations implies the absence of equality as an objective concept. So boolean and symbolic logic do not provide the means to represent them, they can only be used after a decision has been made based on the observations made.

However, measuring, estimating and uncertainty are concepts involved in obtaining the information for taking this decision. Then even when the observation is made, there is a difference between deciding that a conclusion is wrong and the intermediate state called undecidable. Since equality is possible, we have to try equivalence relations.

There is a beautiful and powerful mathematical vehicle to do this, called the Principle of least Action in classical physics. In thermodynamics it is generalized to the Boltzmann objectivity principle, more generally known as the Second law of
Thermodynamics. It states that the entropy, as a measure of missing information, has to be extremized in order to be as objective as possible. This is done using so called partition functions and probability distributions (for the relation between them see \(^{15}\)), in mathematics they are simply called distribution functions. In Informatics the measure for the entropy is called the Shannon measure.

They have nothing to do with statistics, but even there they have proven their value. They only obtain statistical or probabilistic meaning after we have assigned this probabilistic property to them. They are an important tool in statistics, but also important as charge and mass distributions in classical physics or as eigenfunctions in quantum mechanics. And in these cases there is nothing statistical about them at all.

Distribution functions are an impressive and powerful mathematical tool for describing information about systems in using the mathematical theory of functions and even something called knowledge, for the simple reason that the representation by distribution functions can incorporate implicitly or explicitly the associated uncertainty.

In Information Retrieval, many properties are observed and quantified, such as term frequencies, document frequencies, scores and “relevance”, so the theory of uncertainty could take a prominent position. In this article we focus only on scores (expressing the probability of a given document belonging to a class) as obtained in text categorization, in which usually not even the variance is stated in publications.

Except from some variational methods no functional theory will be touched in this paper, apart from those mathematical and information theoretic tools necessary for deriving “worst-case distributions” as information representations for quantifiable relative term frequencies, or other frequencies.

2.5. Getting information from boundary conditions

In this section we use the mathematical technique of variational calculus, with the Shannon entropy measure and Lagrange multipliers, for deriving worst-case distribution function representations for information.

First some concepts:

1. Stationarity: the determination of the extremal paths of an integral given a certain measure using variational calculus, in this case the Shannon entropy measure.

2. Lagrange method: a way to implement a priori boundary conditions, physical laws, structure or other constraint information into the measure chosen.

3. Distribution function: the worst-case solution for the a priori solution above, obtained by solving the constraint information using the observed data. Whereas the above are a priori, this step is a posteriori.

Considering an uncertainty as an estimate, its value only makes sense if its accuracy is known, or better its uncertainty, because the accuracy is a subjective decision,
the uncertainty is its quantification. In this sense each measurement or quantization result consists of a value and an uncertainty, even counted ones.

Humans and even computers make errors which have to be estimated. These estimates, consisting mostly of a value with an uncertainty, are not simply normal distributions in which the value is the mean and the uncertainty two times the standard deviation. Often there is additional information constricting the possible values and uncertainties. Also uncertainties are propagated in a system in where there are dependencies like the laws of Physics or Economics or boundary conditions. How to use this information is difficult with only a value and an uncertainty. A representation by worst-case distribution functions is more appropriate in the a priori phase of analysis. In this sense, a priori means before the decisions are taken, a posteriori means after the data is available, in between there is again a gap.

Now we turn to the mathematical method to obtain worst-case, or most objective, information representations using distribution representations.

### 2.6. Shannon Information Measure, Entropy and Stationarity

In order to find a distribution function $P(v,u)$ representing the measurement result by maximizing the amount of missing information, a measure for that missing information, the entropy, is necessary.

In this paper the Shannon information measure:

$$E = \int P(x) \ln P(x) \, dx.$$ 

is used as the entropy measure.

In general the problem of finding the representing function over a space $X$, the space of all possible states called state space, with information (measurement result) $M$, is resolved by extremizing the integral (summing) over $x \in X$, under the constraint that certain relations of the form $F_i = 0$ are fulfilled for the measurement result on state space $X$. These constraints must be of the form

$$F_i = \int_X f_i(x, P(x)) \, dx = 0,$$

(1)

otherwise they would not contain information about $P(x)$. In general, but not always, as in the case of auto-correlated measurements, they even are of the form,

$$F_i = \int_X f_i(x) P(x) \, dx = 0.$$ 

(2)

Now we can implement the constraints, by adding these zero terms to the entropy, obtaining

$$E = \int \left( \kappa P(x) \ln P(x) + \sum_i \lambda_i f_i(x, P(x)) \right) \, dx.$$

Extremization is done by means of variational calculus. The function $P(x)$, which is the yet unknown solution, is replaced by another function,

$$P(x, \alpha) = P(x) + \alpha \Psi(x),$$
where $\Psi(x)$ an arbitrary function on $x$. Denoting the variation of $E$ by $\delta E = \frac{d}{dx} E$, from the normal rules of calculus it is clear that on an extremum of $E$, $\delta E = \frac{d}{dx} E = 0$. This implies that $\forall \Psi(x)$,

$$
\delta E = \delta \int P(x, \alpha) \ln P(x, \alpha) + \sum_i \lambda_i f_i(x, P(x, \alpha)) dx = 
$$

$$
\int \frac{d}{d \alpha} P(x, \alpha) \ln P(x, \alpha) + \sum_i \lambda_i f_i(x, P(x, \alpha)) \frac{d P(x, \alpha)}{d \alpha} dx = 
$$

$$
\int \kappa - \ln P(x, \alpha) + \sum_i \frac{d}{d P(x, \alpha)} \lambda_i f_i(x, P(x, \alpha)) \frac{d P(x, \alpha)}{d \alpha} dx = 
$$

$$
\int \kappa - \ln P(x, \alpha) + \sum_i \frac{d}{d P(x, \alpha)} \lambda_i f_i(x, P(x, \alpha)) \Psi(x), dx = 0.
$$

The $\lambda_i$ constants have to be determined later by the constraints on $P(x)$ and the normalization of $P(x)$. This can only hold when

$$
\kappa - \ln P(x) + \sum_i \frac{d}{d P(x)} \lambda_i f_i(x, P(x)) = 0 \quad (3)
$$

For the most common case that $\lambda_i f_i(x, P(x)) = \lambda_i f_i(x) P(x)$, this reduces to

$$
\kappa - \ln P(x) + \sum_i f_i(x) = 0
$$

or

$$
\ln P(x) = -\kappa - \sum_i \lambda_i f_i(x) \quad (4)
$$

so

$$
P(x) = \exp \sum_i \lambda_i f_i(x) = \prod_i \exp(\lambda_i f_i(x_i)) \quad (5)
$$

which has to be normalized by solving the normalization condition

$$
\int_X P(x) dx = 1
$$

for $\kappa$ after the determination of the $\{\lambda_i\}$. The $\lambda_i$ are determined by integrating the $\{F_i\}$ explicitly, giving a set of max ($i$) constraint equations for max ($i$) unknown $\lambda_i$, which can be solved either analytically in general or numerically in a specific case.

A special type of constraint is given by the demand that $P(x)$ is only defined on a certain domain $D = a, b$. Given the fact that a $P(x)$ is only defined for $P(x) > 0$ and (7), use a type (4) constraint and $P(a) = P(b) = 0$. So the constraint is formulated by introducing a simple pole at $a$ and $b$ using the ln $x$ function. For example, if the domain is $D = [a, \infty)$, i.e. $x > a$ the constraint becomes

$$
\int_a^{\infty} \ln(x - a) P(x) dx = 0.
$$
Again a Lagrange multiplier has to be used, so at least one of the measured quantities has to be used to determine this multiplier.

2.7. Distribution functions for information representation

As an example, some of the most commonly known distribution functions as worst-case representations given some data will be derived here, the Poisson distribution and the Gaussian or normal distribution. Additionally the two most commonly used distribution functions in this paper, the Pierson type III and the Beta distribution, will be described. For these distributions the boundary conditions and constraints are given, as well as their domain of applicability. Derivation is left to the user. In general, the reader is referred to Abramowitz and Stegun Chapter 26 tables of other distributions and their domains, as well for additional properties of the distributions.

2.7.1. The Poisson distribution

The case where the only information available about a quantity is the mean \( \bar{x} \) and the fact that \( x \) is real and \( x > 0 \), is treated here. The single constraint is given by

\[
\int_0^{\infty} (x - \bar{x}) P(x) \, dx = 0
\]

and of the form (4), so solution (7)

\[
P(x) = \exp(\kappa - \lambda(\bar{x} - x))
\]

can be used. The constraint reads

\[
\int_0^{\infty} (x - \bar{x}) e^{\kappa-(x-\bar{x})} \xi \, dx = \left( -\frac{(e^\kappa \bar{x} \xi - e^\kappa)}{\xi^2} = 0 \right),
\]

which has the solution

\[
\xi = \frac{1}{\bar{x}}.
\]

The normalization condition is then given by

\[
\int_0^{\infty} e^{\kappa-x} \bar{x} \, dx = (e^{\kappa+1} \bar{x} = 1)
\]

with solution

\[
\kappa = \log \left( \frac{1}{\bar{x}} \right)
\]

In this way the well-known Poisson distribution

\[
P(x) = \frac{\exp(-\frac{x}{\bar{x}})}{\bar{x}}
\]

is obtained. This shows the Poison density to be the worst-case distribution representation when the information available is restricted to only the mean and the fact that \( x > 0 \).
2.7.2. The Gaussian or normal distribution

As a second example, the case is treated where the only information obtained from a measurement is the mean value \( \bar{x} \) of the data set and its standard deviation \( \sigma \). It is shown here that the normal or Gaussian density is the worst-case information representation for the information contained about the value of \( x \) in this data set. When mean \( \bar{x} \) and standard deviation \( \sigma \) are known, the conditions become

\[
\int_{-\infty}^{\infty} (x - \bar{x}) P(x) \, dx = 0
\]

and, with the uncertainty in the mean given by

\[
u(\bar{x}) = \frac{\sigma}{\sqrt{N - 1}}
\]

the second constraint becomes

\[
\int_{-\infty}^{\infty} ((x - \bar{x})^2 - \nu(\bar{x})^2) P(x) \, dx = 0,
\]

Since these are both of type (4), (7) can be used again, leading to

\[
P(x) = \exp (1 - \xi_x)(x - \bar{x}) - \xi_\sigma (\sigma^2 - (x - \bar{x})^2).
\]

The first constraint gives the equation

\[
\int_{-\infty}^{\infty} (x - \bar{x}) e^{\xi_x (\sigma^2 - (x - \bar{x})^2) - \xi_\sigma (x - \bar{x}) + \kappa} \, dx = -\frac{\sqrt{\pi} \xi_x e^{\frac{4 \xi_x^2 \xi_\sigma^2 + 4 \xi_x \xi_\sigma + \xi_\sigma^2}{4 \xi_\sigma^2}}}{2 \xi_\sigma^{3/2}} = 0
\]

which has the solution \( \xi_x = 0 \). The second constraint gives the equation

\[
\int_{-\infty}^{\infty} \left(\sigma^2 - (x - \bar{x})^2\right) e^{\xi_x (\sigma^2 - (x - \bar{x})^2) + \kappa} \, dx = \frac{\left(\frac{\sqrt{\pi} (4 \xi_x^2 \xi_\sigma^2 - \xi_\sigma^2)}{\sqrt{\xi_\sigma^2}} - 2 \sqrt{\pi} \sqrt{\xi_\sigma}\right) e^{\frac{4 \xi_x^2 \xi_\sigma^2 + 4 \xi_x \xi_\sigma + \xi_\sigma^2}{4 \xi_\sigma^2}}}{4 \xi_\sigma^2} = 0
\]

with solution

\[
\xi_\sigma = \frac{1}{2 \sigma^2}.
\]

The normalization condition

\[
\int_{-\infty}^{\infty} e^{\frac{x^2 - (x - \bar{x})^2}{2 \sigma^2} + \kappa} \, dx = \sqrt{\frac{\pi}{2}} e^{\frac{\kappa}{2} + \frac{1}{2}} |\sigma| = 1
\]

then fixes \( \kappa \)

\[
\kappa = \frac{-2 \log \sigma + \log \pi + \log 2 + 1}{2}
\]
This shows that in cases where only the mean and the uncertainty are known, the gaussian or normal distribution

\[ N(x) = \frac{e^{-\frac{x^2-2x\bar{x}+\bar{x}^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \]

is the worst-case information representation.

2.7.3. The Pierson type III distribution

The Poisson distribution obtained above is a special case of the more general Pierson type III distribution. This is the distribution for the case that \(-\infty < \alpha < x < \infty\) and the mean \(\bar{x}\) and standard deviation \(\sigma\) are known. The constraints are given by the domain boundary condition

\[ \int_{-\infty}^{\infty} \ln(x-\alpha)P(x)\,dx = 0 \]

and one of the two constraints

\[ \int_{-\infty}^{\infty} (x-\bar{x})P(x)\,dx = 0 \]

or

\[ \int_{-\infty}^{\infty} ((x-\bar{x})^2-\sigma^2)P(x)\,dx = 0. \]

To solve for the two Lagrange multipliers, both these constraints are used. The solution is

\[ P(x) = \frac{(x-\alpha)^{(p-1)}}{\Gamma(p)\beta^p} e^{-\frac{x-\alpha}{\beta}} \]

with \(\alpha\) the given constant boundary and the constants \(p\) and \(\beta\) defined by

\[ \beta = \frac{\sigma^2}{\bar{x} - \alpha} \]

\[ p = \frac{(\bar{x} - \alpha)^2}{\sigma^2}. \]

Since the distribution is implicitly asymmetric the distribution is skew. [Abr.&Steg. 26.2.31]

2.7.4. The Beta distribution

One of the most important distributions in this paper is that of term frequencies in a document. The mean frequency \(\bar{f}\) and uncertainty \(u(\bar{f})\) are measured for all terms. Such frequencies always fulfill

\[ 0 < \bar{f} = \frac{n}{N} < 1 \]
where \( n \) is the term count and \( N \) the total number of terms in the document. We need a distribution that fulfills these constraints. The boundary conditions are implemented using

\[
\int_0^1 \ln(x) P(x) \, dx = 0
\]

and

\[
\int_0^1 \ln(1 - x) P(x) \, dx = 0.
\]

The associated Lagrange multipliers are solved for using the constraints

\[
\int_0^1 (\bar{f} - x) P(x) \, dx = 0
\]

and

\[
\int_0^1 ((x - \bar{f})^2 - u(\bar{f})^2) P(x) \, dx = 0.
\]

The solution is the Beta distribution

\[
P(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a, b)}
\tag{7}
\]

with

\[
a = -\bar{f} \left( \frac{\sigma^2 + \bar{f}^2 - \bar{f}}{\sigma^2} \right)
\]

\[
b = \left( \bar{f} - 1 \right) \left( \frac{\sigma^2 + \bar{f}^2 - \bar{f}}{\sigma^2} \right)
\]

and

\[
B(a, b) = \frac{\Gamma(a) \Gamma(b)}{\Gamma(b + a)}
\]

the so-called Beta function.

The parameters \( \bar{f} \) and \( \sigma \) are respectively the mean relative frequency and its uncertainty.

The entropy of this distribution is given by

\[
(-b - a + 2) \psi_0(b + a) + (b - 1) \psi_0(b) - \log B(a, b) + (a - 1) \psi_0(a)
\]

The occurrence of both parameters in the entropy is caused by information coupling, so less information is lost than in the normal distribution.

For this distribution to make sense \( a > 1 \) and \( b > 1 \), ergo the conditions (assuming \( 0 < \bar{f} \ll 1 \)),

\[
\sigma < \bar{f} \frac{\sqrt{1 - \bar{f}}}{\sqrt{\bar{f} + 1}} \approx \bar{f}
\]

and

\[
\sigma < (1 - \bar{f}) \frac{\sqrt{\bar{f}}}{\sqrt{2 - \bar{f}}} \approx \sqrt{\bar{f}} / 2
\]
have to be fulfilled. Otherwise, the boundary conditions are violated \( \beta(x) = -\infty \) and consequently the measurement provides non-information or garbage. The associated information is consequently classified as noise and has to be removed from the measurement data. Implicitly this distribution in general is skew, with skewness \( \frac{\kappa^2}{\kappa} \).

2.8. The Beta distribution in text categorization

The beta distribution is a rather fundamental distribution in text categorization, as can be concluded from the fact that the so called Feller-Pareto distribution is obtained by a simple parameter transformation, \( x \rightarrow \frac{1}{1+u} \).

This Pareto distribution is clearly an important distribution in text categorization, as shown by its relevance in the renormalization problem in text categorization \(^7\) Under some more specific assumptions it even generalizes to Zipf’s law as treated by Mandelbrot \(^12\). In this last form it demonstrates the usefulness of the physical concept of least action, which is simply entropy extremization, as a principle in nature and in text categorization.

People tend to use language in the most efficient way. Short words are used for most common concepts, long and complicated words for the uncommon complex ones. Zipf’s law introduces as a measure for the complexity of a word its frequency rank. The higher the rank, the lower its frequency of use. One could in a sense state that in subjective terms humans use words in communication quite effectively, in accordance with the least action principle in Physics.

So there is structure, there are laws governing this efficiency in human communication. Most are (yet) not quantifiable. The way to change that: gaining knowledge “the value observed plus the relations already known” but we have to acknowledge and assert the missing information in the observation called “uncertainty” before reasoning about it.

3. Uncertainty in Text Categorization

In Text Categorization three abstraction levels have to distinguished: The category level, a class whose objects are documents; the documents, whose objects are terms; and the term level, since the terms themselves are representations. One should notice that a document by itself is already a representation. In the process of representing what the document is about, already a lot of information is lost. The further abstraction of subjective categorization also looses information.

On the category level, some of the causes of uncertainties are:

1. to begin with, the fact that the train (and test) set have been labeled manually by human judges, causing
   
   • subjectivity in the assignation of labels (inter-judge variability)
   • quantization error (esp. in multi-classification) because labels have to be either present or absent, no gradual judgements are allowed
• human error and variability (friday judgements may be different from monday judgements)

2. uncertainty in measuring class properties on the document level of the train set: In how far is the train set representative for the test set?

3. variability in class properties: there will be categories with several different disjunct subcategories, with highly varying relative term frequencies for the distinct subcategories

4. uncertainty due to the classification algorithm: different algorithms may be better in different situations.

On the document level:

1. polymorphism: the meaning of the terms themselves, i.e. their information content, may depend on the context in which they are used.

2. synonymy: different terms with the same information content can be used to represent the category properties.

Although these concepts clearly enhance the beauty and readability of a text, for the computer they imply useless loss of information. Hence additional uncertainty.

Then on the term level:

1. again human error, for example typing errors, the degree or ability of mastering a foreign language with sufficient sophistication or a quite common condition called dyslexia.

2. the uncertainties in the quantification or resolution of the properties themselves by the restrictions posed by the chosen measurement process itself.

3. certain terms will not get enough information from the train set, causing lack of resolution, they will be useless.

4. on the other hand, a lot of terms are so commonly used that the quality of their quantization on the global level is nearly perfect, hence useless. Classification demands distinguishability of terms.

Uncertainty theory demands an analysis of all sources of uncertainty and their interaction. Addressing just one issue, we shall show how to estimate the uncertainties in the relative term frequencies. So this paper restricts itself to the uncertainties in the quantification on the term level. Two aspects will be considered, the quality of the term properties and the selection of proper terms distinguishing the classes, called term selection (TS)

Furthermore, we will show how constraints, or better boundary conditions, on the quantified information obtained from a training set can be used to provide additional information about the term data quality. In this way invalidly quantified information can be detected and removed.
3.1. Term frequency measurement

Let the relative term frequency of term \( k \) in document \( i \) be given by

\[ f_{ki} = \frac{n_{ki}}{N_i} \]  

(8)

in which \( n_{ki} \) stands for the number of occurrences of term \( t_k \) in document \( d_i \) and \( N_i \) for the total number of terms in document \( d_i \). Both of these quantities are natural numbers. The category-dependent term frequency for term \( k \) in category \( j \) is given by the mean

\[ TF_{jk} = \frac{1}{C_j} \sum_{d_i \in T_j} f_{ki} \]  

(9)

in which \( T_j \) is the set of training documents for category \( j \) and \( C_j \) the total number of documents in this train set. Since the document term frequency has a natural variability, it has a standard deviation around this mean which is given by

\[ s_{jk} = \sqrt{\frac{1}{C_j} \sum_{d_i \in T_j} (f_{ki} - TF_{jk})^2} \]  

(10)

Usually, the uncertainty is chosen to be based on two times the standard deviation equivalent. However, since in this paper the distribution function representation approach is used, the standard deviation itself will be used as the base of the uncertainty. The uncertainty in the mean is then estimated by the reliability in the estimated mean, i.e.

\[ u(TF_{jk}) = \frac{s_{jk}}{\sqrt{C_j}} \]  

(11)

The measurement result is now given by the set of associations

\[ TM_j = \{ [t_k | [TF_{jk}, u(TF_{jk})]] \} \]

for category \( j \). The total information over all categories is given by the set

\[ TM^* = \{ \{TM_j\} \} \]

3.2. Noise and information

As mentioned above, the estimated term frequencies are a priori probabilities of the terms in a class-dependent manner. They are certainly valued between zero and one. For this type of boundary condition, under the assumption that the “real” quantity is single-valued, and with known mean estimate and uncertainty, the worst-case distribution is the Beta distribution as derived before.

\[ P(x|a, b) = \frac{(1-x)^{b-1} x^{a-1}}{B(a, b)} \]
In terms of the estimated mean and its uncertainty, the parameters $a$ and $b$ are given by

$$ a = - \frac{m (u(m)^2 + m^2 - m)}{u(m)^2} $$

and

$$ b = \frac{(m - 1) (u(m)^2 + m^2 - m)}{u(m)^2} $$

where $m$ is the estimated mean and $u(m)$ the uncertainty.

It is clear that the boundary conditions on the mean imply boundary conditions on the uncertainty. One is trivial: $u(m) > 0$. The other one is given by the fact that $a > 1$ and $b > 1$ must hold, since otherwise the distribution function diverges at zero, giving infinite probabilities. Solving $a$ and $b$ for this boundary condition gives

$$ u(m) < m \sqrt{-\frac{m - 1}{m + 1}} $$

and

$$ u(m) < (1 - m) \sqrt{-\frac{m}{m - 2}} $$

or

$$ u(m) < \min \left( m \sqrt{-\frac{m - 1}{m + 1}}, (1 - m) \sqrt{-\frac{m}{m - 2}} \right) $$
with “min” the usual minimization function. We shall call $k = m/u(m)$ the k-factor or “quality” of the data.

3.3. 

**Noisy term elimination**

The smaller $k$ is, the greater the chance that the data is noise. The bigger $k$, the more certain it is that the value really contains information about a term. So $k$ can be considered as a measure for the quality (not the significance) of the information.

It is advisable to remove the noise from the data before processing it further, because otherwise noise may propagate through the entire process. In particular, a learning process which is amplifying significant data and suppressing insignificant data cannot itself distinguish between noise and information. Unless the learning process inherently filters noise, noise will then be randomly amplified or suppressed.

4. Uncertainty and term selection

The automatic classification of documents (also known as Text Categorization) presents a form of machine learning where features abound. By far the most of these have no value for the classification and should be eliminated. Even though some learning algorithms are more robust against spurious features than others, the performance of the algorithms deteriorates with larger numbers of features (called terms in Information Retrieval). That is why term selection is an important issue in many Information Retrieval applications.

For the further development of the Linguistic Classification System LCS in the PEKING project, which uses phrases rather than (key)words as terms, we need a term selection technique which is applicable to linguistic phrases as well as to single keywords. We cannot use a conventional stop list, because a list of stop phrases makes little sense. Furthermore, for performance reasons we need the term selection to make a drastic reduction in the number of terms that have to be considered by the classification algorithm, for as far as this is possible without loss in accuracy. The fewer the terms, the lower the cost of classification in time and memory.

We have experimented with the use of uncertainty in the term frequency as a criterion for term selection, selecting or eliminating each term $k$ on the basis of its quality $\frac{TF_k}{TF_k + F_{<k}}$.

4.1. 

**Experimental approach**

In we have reported the results of a large scale experiment, comparing six Term Selection methods comparing Uncertainty-based term selection (UC) with five other Term Selection methods, using three different corpora with very different characteristics. In this paper we shall use only one corpus (the Reuters mono corpus), since the previous experiment has shown that the results for different corpora were very similar.

http://www.cs.kun.nl/peking
4.1.1. The corpus

The Reuters mono corpus is a selection of 9090 documents from the well-known Apte subset of the Reuters 21578 corpus, split into a train set of 6519 and a test set of 2574 documents, which are mono-classified in the sense that each document belongs to precisely one class (topic). The documents are short newspaper articles, unevenly distributed over 66 classes. The largest class comprises about 25% of the documents.

We used all words of the text as terms. The only pre-processing applied was the replacement of capital letters by the corresponding small letters (decapitalization) and the removal of interpunction and some special characters. In particular, no stemming was performed and no stop list was applied. We relied on the term selection to eliminate redundant terms.

4.1.2. Two algorithms

The well-known Rocchio algorithm is the typical workhorse for document retrieval and classification in Information Retrieval. It is based on the Vector Space model and computes a classifier essentially as a vector of term weights.

The less well-known Winnow is a heuristical learning algorithm in the Perceptron family, heuristically computing a classifier while iterating over the training documents in order to obtain an optimal linear separator between relevant and non-relevant documents (for the underlying theory see ).

In terms of the admirable exposition given in , what we have implemented is the Balanced Winnow algorithm with the thick threshold heuristic. We have made a slight improvement to its speed of convergence by using normalized term strengths to initialize the Winnow weights.

4.1.3. The measure of Accuracy

The measure of Accuracy used is the F1 measure, a kind of average of Precision and Recall defined by

\[ F1 = \frac{2}{ \frac{1}{\text{Precision}} + \frac{1}{\text{Recall}} } \]

which is micro-averaged over the classes, meaning that the number of relevant documents selected (RS) are summed over all classes, and similarly for RNS, NRS and NRNS. This measure favours the large classes over the small ones, but it can be argued that it gives a useful overall view. Using the F1-value as Accuracy measure attaches equal importance to Precision and Recall.

4.1.4. Mono- or multi-classification?

Even though all documents are mono-classified, we make no use of the fact that each test document should belong to precisely one class but rather allow it to be assigned to several classes (between zero and three). Not using this information lowers the Accuracy achieved somewhat, because not all information is used, but
it actually provides a harsher test, which is also relevant to multi-classification. Thus, although we are training 66 independent classifiers we also take into account the interaction between them (by looking at the combined set of terms and by multi-classifying the test data).

5. Local noise elimination

We shall first investigate the use of the uncertainty-based noise removal criterion as a local (i.e., category-dependent) term selection criterion.

Local Noise removal consists in the elimination, for each category separately, of all terms whose quality is less than some value Qmin. The relation between Qmin and the Accuracy for Winnow and Rocchio is illustrated by the following graphs:

After an initial plateau, the Accuracy of both algorithms degrades. Winnow first shows a slight improvement in Accuracy for Qmin around 1.3. The graphs (and those in 13) show that local noise reduction with Qmin at most 2 is a reliable, although not very aggressive, form of term selection, with an expected reduction of the number of terms by 80%.

6. Stiff term elimination

Global term selection consists in the removal of undesirable terms from all doc-
Uncertainty and Term selection in Text Categorization

Documents, in a way which depends on the corpus but not on the particular category. The simplest form of global term selection is based on the overall document frequency DF or term frequency TF of terms. These are very simple corpus statistics, which can be collected quite efficiently, and both TF and DF have repeatedly been shown to capable of reducing the number of terms without loss of accuracy (see 19 or the section on Feature selection in 17).

In this case we are interested in the use of Uncertainty for global Term Selection. The uncertainty in the frequency of terms will have to be computed taking the whole corpus as one category. We shall investigate the elimination of global high-quality terms: those terms that occur most reliably across all documents. We shall call such terms the *stiff terms* (after a perceived mechanical analogon), which are eliminated by imposing a maximum value Qmax on the term quality.

For the Reuters Mono corpus, the relation between Accuracy and maxQ is illustrated by the following graphs:

As Qmax is lowered, Winnow gently deteriorates and then collapses. Rocchio achieves a slight improvement around Qmax = 15 before collapsing. Winnow obviously is quite good at removing stiff terms, whereas Rocchio can use some help. Unfortunately, the optimal region for stiff term elimination depends somewhat on the corpus (see 13), so that some manual tuning may be required.

The last graph shows the number of terms eliminated. There are not many stiff
terms that can safely be discarded (only a hundred by the time the collapse sets in), and it is remarkable that the removal of so few terms has such an effect. Obviously their presence is very disturbing.

The stiff terms are analogous to the stop words in classical Retrieval, but they are corpus-specific, and based on statistical rather than linguistical considerations.

7. Conclusion

We have described the notion of Uncertainty, and discussed its prominent place in the theory and experimental practice of modern Physics. We have argued that the awareness of Uncertainty may also be of tremendous importance to the field of Information Retrieval, and in particular Text Categorization.

The experiments reported here and in 13 show that the elimination of local noisy terms is a reasonably effective Term Selection method (UC). It is controlled by one parameter Qmin which is independent of the corpus and therefore easy to install as default.

In this paper also a new global term selection method was introduced, Stiff Term Elimination (STE), which eliminates those terms which are evenly distributed across all documents. For classification algorithms which are sensitive to stiff terms (such as Rocchio), this may improve Accuracy. Stiff Term Elimination is analogous to the traditional stop word elimination, but it is independent of the use of words as terms and therefore applicable with other representations like phrases.

We foresee many more applications of Uncertainty and measurement theory to Information Retrieval. We are currently investigating more serious applications: measuring uncertainty in document scores, measuring entropy and uncertainty in class profiles and developing more objective techniques for comparing the quality of categorization algorithms.

The possibilities to improve Text Categorization by better term selection appear rather limited, to improve the resolution further we have to attack the more restrictive qualitative uncertainties. The entropy may be further reduced using qualitative information, for instance the structural relations in the syntax and between terms in a text. These are the concepts closest to physical laws.

A lot can be gained by resolving the fuzzyness caused by polymorphy and synonymy. This implies resolution of structure in the document and analyzing interrelationships between terms in phrases.

Another important factor is the fuzzyness caused by boolean class labeling, but this can also be tackled by a feedback mechanism in the learning process. It should become possible to gauge and calibrate classification systems and measure the quality of labeled data sets.