Inf Retrieval (2014) 17:153–176
DOI 10.1007/s10791-013-9225-4

A nonparametric term weighting method for information retrieval based on measuring the divergence from independence

İlker Kocabaş · Bekir Taner Dinçer · Bahar Karaoğlan

Received: 6 September 2012 / Accepted: 16 May 2013 / Published online: 24 May 2013

Abstract In this article, we introduce an out-of-the-box automatic term weighting method for information retrieval. The method is based on measuring the degree of divergence from independence of terms from documents in terms of their frequency of occurrence. Divergence from independence has a well-establish underling statistical theory. It provides a plain, mathematically tractable, and nonparametric way of term weighting, and even more it requires no term frequency normalization. Besides its sound theoretical background, the results of the experiments performed on TREC test collections show that its performance is comparable to that of the state-of-the-art term weighting methods in general. It is a simple but powerful baseline alternative to the state-of-the-art methods with its theoretical and practical aspects.

Keywords Information retrieval · Nonparametric index term weighting · Statistical dependence · Pearson's Chi-Square statistics

1 Introduction

Any text-based IR system has two integral parts: (1) a term weighting method, such as TFxIDF, and (2) a retrieval model (or ranking model), such as the vector space model. Words constituting the text do not contribute evenly to the informative content of the

İ. Kocabaş (✉) · B. Karaoğlan
International Computer Institute, Ege University, Bornova, Izmir, Turkey
e-mail: ilker.kocabas@ege.edu.tr

B. T. Dinçer
Department of Statistics, Muğla University, Mugla, Turkey
e-mail: dtaner@mu.edu.tr

B. T. Dinçer
Department of Computer Engineering, Muğla University, Mugla, Turkey

document. In this context, the research question of interest in term weighting can be stated as: *which terms are related to which documents with respect to informative content?* We base our answer to this question on a frequentist point of view by using the simple assumption that in a document written in natural language, some words are used due to grammatical necessity (i.e., semantically nonselective or function words), while some are used to reflect the document's contents (i.e., semantically selective or content bearing words). Given a collection of documents, this implies that function words are supposed to appear in every document in contrast to content bearing words (keywords), provided that the collection is composed of documents with different topics. This assumption directly leads to a well-known method of distinguishing keywords from function words, called the *Inverse Document Frequency*-IDF (Sparck Jones 1972; Robertson and Sparck Jones 1976).

In order to identify the terms that contribute to the informative contents of documents, every frequentist approach to term weighting requires making a second assumption about language use which may be stated as: there is a *causal* relation between frequency of a word occurrence and its contribution to informative content. The question then arises on how to model such a relation, in order to quantify the contribution of a word to informative content based on its frequency of occurrence. On this account, Luhn (1958) states that "the frequency of word occurrence in an article furnishes a useful measurement of word significance", and claims that "keywords tend to occur at mid-range frequencies in documents, rather than at low or high frequency ranges." If Luhn's claim holds perfectly true, then keywords would completely be distinguished from function words by means of a method utilizing *within document term frequencies*. Unfortunately, this is not the case in practice.

We can further make a third assumption so as to have a better model: Keywords have frequency distributions different from that of function words in the population of documents. On this account, Harter (1975a, b) claims that both keywords and function words follow a Poisson distribution in the population of documents but with different means. If Harter's claim holds perfectly true, then keywords would completely be discriminated from function words by means of a method utilizing *term frequency distributions*. Again as in Luhn's claim, this is not the case either. In the same line of research, a recent answer comes from Amati and van Rijsbergen (2002). They claim, in principle, that a semantically selective word occurs in a semantically related content with a frequency different from the frequency of the word in common use, which can be determined by a model of randomness, such as Poisson distribution, e.g. PL2.

Our *divergence from independence* (DFI) hypothesis which brings another point of view to this discussion is that a semantically selective word occurs in a semantically related content with a frequency different from the frequency of a word in common use, which can be determined by the saturated model of independence. According to the notion of independence, if the ratio of the frequencies of two particular words remains constant across all documents, then the occurrences of those words are said to be independent from the documents. Assuming that the amount of contribution of a word to the information content of a document is proportional to the observed frequency of the word in the document, we can further say that both words evenly contribute to the information content of every document. As the function words are the ones which make equal contribution to the contents of every document, one can therefore discriminate the keywords from the function words by measuring the divergence of the observed frequency of each term in the documents from the frequency expected under independence.

With a system using DFI term weighting functions, we participated in TREC 2009, 2010, and 2012 (Dinçer et al. 2009, 2010; Dinçer 2012). Even though the operational

settings of indexing was not ideal for DFI in TREC 2009 (i.e., stop-word elimination was applied to the ClueWeb09-T09B data set), average retrieval performance with respect to other systems has been achieved. On the other hand, in our system run on the TREC 2010 Web track data set, stop-words were preserved and two supplementary techniques were used: (1) *n*-gram phrase matching and (2) spam-page filtering. The official TREC results show that it is one of the best performing runs (Clarke et al. 2010). In this respect, this article presents an in-depth description of how the concept of independence can be exploited in term weighting, and the retrieval performance comparisons with the state of-the-art term weighting methods/models using past TREC test collections.

The organization of this article is as follows. The previous works related to DFI term weighting are summarized in the "Related Work" section. "Models of independence and measures of dependence" and "Term weighting based on DFI" are given afterwards. The design of experiments, the experimental results, a short discussion, and the conclusions are presented subsequently.

2 Related works

Early approaches to the term weighting problem, from the viewpoint of statistics, date back to the pioneering work of Maron and Kuhns (1960). Since then, several attacks have been made, including the works of Damerau (1965), Bookstein and Swanson (1974), Robertson and Sparck Jones (1976), Cooper and Maron (1978), Croft and Harper (1979), Fuhr (1989), Margulis (1992), Wong and Yao (1995), Ponte and Croft (1998), and Amati and van Rijsbergen (2002). Among all, Harter (1975a, b) is the first researcher who introduced the notion of *eliteness*, a notion that gave rise to the most successful probabilistic approaches to term weighting problem. According to Harter, there are "specialty words" and "non-specialty words". Specialty words are the ones which occur densely in "elite" documents whose informative contents are composed of the meaning conveyed by those speciality words. In contrast, non-specialty words are the ones which occur randomly, and which do not contribute to the contents of the document. Harter claims that specialty words differ from non-specialty words in distribution on a collection of documents, and both the specialty and non-specialty words follow a Poisson distribution with different means λ_1 and λ_2, respectively, where $\lambda_1 > \lambda_2$.

In spite of the fact that the original "2-Poisson model" of Harter is plausible in theory, and justified on particular samples, it does not perform well in practice. It is refined by Robertson et al. (1981), and generalized by Robertson and Walker (1994) as a series of successful implementations called BMs (e.g., BM25). In the same line of research, Amati and van Rijsbergen (2002) developed a sophisticated probabilistic model/framework, called the *divergence from randomness*-DFR, by further refining the notion of eliteness on the basis of the semantic information theory (Hintikka 1970) and the Popper's (1995) notion of informative content. Finally, in their recent work, Clinchant and Gaussier (2010) incorporate the *burstiness* notion (Church 1995) with the notion of eliteness so as to result in better and simpler information-based methods of term weighting, such as the method with a priori *log-logistic* distribution called LGD.

Among all the previous works in this line of research, DFI-based term weighting is mostly related to DFR-based term weighting. Randomness in occurrence characterizes non-speciality words within the Harter's paradigm, and it is quantified in the context of DFR by means of a *basic model of randomness*, which corresponds to a Poisson distribution in the paradigm. In order to model the actual (chance) distribution of frequency of

word occurrence in documents on the population, Amati and van Rijsbergen (2002) examine several probability density functions, such as Hyper-Geometric, Bose-Einstein, etc., as well as Poisson distribution. They claim that speciality words are the ones which occur in documents with a frequency different from the frequency suggested by the assumed model of randomness; and complementarily, non-speciality words are the ones which occur in documents with a frequency that can be attributed to chance under the assumed model of randomness.

On the other hand, statistical independence takes the place of randomness in the context of DFI, such that non-speciality words are characterized by independence in occurrence rather than randomness. In this respect, DFI is the nonparametric counterpart of DFR. From the viewpoint of statistics, a term weighting method based on DFR or DFI can be considered as a method/procedure of (inductive) inference from observations to population, where observations are observed frequencies of words in documents and population is the collection of frequencies that could be generated by a model of randomness or independence. As an inferential procedure, term weighting based on DFR is of parametric type in contrast to DFI, since it is necessary to make an assumption about the precise shape/form of the actual distribution of frequency of word occurrence in the population of documents in order to define what is random (Wolfowitz 1942; Bradley 1968). It is worth mentioning in particular that DFR models are also qualified as nonparametric, but the term "nonparametric" means *no-parameter* or *parameter-free* in this context. On this account, it can be said that DFI models are nonparametric in both senses.

3 Measures of divergence from independence

Divergence from independence (DFI) implies statistical dependence. In brief, statistical dependence refers to a relation between two random variables that makes one of the random variables less or more probable to take on a value when the value of the other one is given. To measure the degree of statistical dependence between two random variables, there needs to be a model of independence that expresses the relation between the random variables in terms of probability models. The model of independence that is used in the study presented in this article for the purpose of term weighting is referred to as the *saturated* model in statistics.

3.1 Saturated model of independence

Suppose that the cells of a $r \times c$ contingency table contain densities or proportions, p_{ij} for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$. Then, the marginal density for row i is given by $p_i = \sum_j p_{ij}$ and the marginal density for column j is given by $p_j = \sum_i p_{ij}$. The bivariate distribution of the categorical random variables X and Y can be characterized in terms of probability models relating cell densities. Multiplication rule of probability states that if two random variables are independent of each other in distribution, the bivariate probability distribution of the random variables, $f_{X,Y}(x, y)$, can be expressed as the product of their marginal probability distributions: $f_{X,Y}(x, y) = f_X(x) \times f_Y(y)$, implying that the probability of co-occurrence of two events, one of which is defined in the sample space of X (i.e., $X = i$) and the other one is defined in the sample space of Y (i.e., $Y = j$), can be expressed as the product of the marginal probabilities of occurrences of individual events:

$$Pr(X = i, Y = j) = Pr(X = i) \times Pr(Y = j),$$

where $Pr(X = i, Y = j)$ corresponds to the cell density p_{ij}, $Pr(X = i)$ the marginal density p_i for row i, and $Pr(Y = j)$ the marginal density p_j for column j, for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$.

Under independence, the observed density in the cell (i, j), tf_{ij}/N, is therefore expected to be equal to the product of the observed marginal density for row i, TF_i/N, and the observed marginal density for column j, D_j/N:

$$\frac{tf_{ij}}{N} = \frac{TF_i}{N} \times \frac{D_j}{N},$$

for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$.

Among all, the simplest model of independence is the *equal cell probability model*,

$$p_{ij} = \frac{1}{rc},$$

implying that all $r \times c$ possible events are equally likely. Other models which assume constant densities across rows and constant densities down columns are given by

$$p_{ij} = p_i \cdot \frac{1}{c} \quad \text{and} \quad \mathrm{p_{ij}} = \frac{1}{\mathrm{r}} \cdot \mathrm{p_j}.$$

In the *constant column density model*, the marginal density for each column is $p_j = 1/c$. Also in the *constant row density model*, the marginal density for each row is $p_i = 1/r$.

Multiplication rule of probability may be expressed in terms of the three simple models, such that

$$p_{ij} = \left[\frac{1}{rc}\right]\left[\frac{p_i}{1/r}\right]\left[\frac{p_j}{1/c}\right].$$

The first term represents the density for cell (i, j) under the constant cell density model. The second term represents the ratio of the marginal density for row i to the marginal density under the constant row marginal model. The third term represents the ratio of the marginal density for column j to the marginal density under the constant column marginal model.

This independence model can also be interpreted as the product of an average effect $[1/rc]$, a row effect $[r \cdot p_i]$, and a column effect $[c \cdot p_j]$. Under independence, the row effect and the column effect account for all the variation in cell densities.

If this independence model does not hold, cell densities can be expressed by including a residual as given by

$$p_{ij} = \left[\frac{1}{rc}\right]\left[\frac{p_i}{1/r}\right]\left[\frac{p_j}{1/c}\right]\left[\frac{p_{ij}}{p_i p_j}\right].$$

The last term $p_{ij}/(p_i p_j)$ is the residual, and accounts for the variation in cell densities due to the statistical dependence between the row and column categories.

This model of independence is called the *saturated model*, because it *perfectly fits* any given contingency table:

$$\frac{tf_{ij}}{N} = \left[\frac{1}{rc}\right]\left[\frac{TF_i/N}{1/r}\right]\left[\frac{D_j/N}{1/c}\right]\left[\frac{tf_{ij}/N}{(TF_i/N)(D_j/N)}\right],$$

for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$.

The degree of divergence of the residual from the value 1 gives the magnitude and direction of the divergence from independence in cell densities. For any term t_i $(i = 1, 2, \ldots, r)$ and document d_j $(j = 1, 2, \ldots, c)$, the degree of divergence from independence in the observed frequency of occurrence of t_i in d_j can therefore be measured as

$$dfi_{ij} == \frac{tf_{ij}/N}{TF_i/N \cdot D_j/N} - 1 = \frac{tf_{ij}}{e_{ij}} - 1 = \frac{tf_{ij} - e_{ij}}{e_{ij}}, \quad (1)$$

where $e_{ij} = (TF_i \cdot D_j)/N$. This is the measure of statistical dependence that we used, in principle, for the purpose of term weighting.

The residual of the saturated model of independence can also be regarded as the ratio of the relative frequency of occurrence of term t_i in a document d_j to the relative collection wide frequency of that term: (tf_{ij}/D_j) / (TF_i/N), which expresses the DFI hypothesis in the true sense: "to what degree does the frequency of occurrence of a term in a document diverge from the frequency of occurrence of the term in common use?" The DFI hypothesis assumes that a function word has a relative frequency ratio that is equal to 1. This means that a function word is a term whose collection frequency distributes on documents proportional to the length of the documents: then for a function word t_i

$$tf_{ij} = TF_i \frac{D_j}{N},$$

for $j = 1, 2, \ldots, c$.

3.2 Normalized chi-squared distance from independence

Saturated model of independence is the underlying model of the Pearson's chi-squared test of independence (Agresti 2002). The test statistic used in chi-squared tests is given by

$$G^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(tf_{ij} - e_{ij})^2}{e_{ij}}.$$

Since Pearson's chi-square statistic is simply the sum of squares of the standardized degree of divergence from independence in observed cell frequencies, each component of G^2 is usually referred to as the *normalized chi-squared distance* from independence. In addition to the DFI formula given in (1), both standardized and chi-squared distance from independence can be used as a measure of DFI.

In general, for any experiment whose sample space can be divided into two disjoint subsets, *success* refers to the event that an outcome from the subset of interest occurs, and *failure* refers to the event that an outcome from the other subset occurs.

Suppose that outcome of an experiment is the classification of a given observation into one of the $r \times c$ cell categories in a contingency table, such that the sample space of the experiment is composed of $r \times c$ outcomes: classifying a given observation into the first cell category, classifying the same observation into the second cell category, and so on. Let *success* be the classification of an observation into a particular cell category, and *failure* be the classification of the same observation into one of the other $(r \times c) - 1$ cell categories. Then, cross-classification of a document collection of size N with respect to the two categorical random variables X and Y can be modeled as a random experiment, a series of N identically and independently repeated classification over $r \times c$ cell categories.

A Bernoulli trial is an experiment that has dichotomous outcomes such as "head" which is usually referred to as *success*, and "tail" which is usually referred to as *failure* in the

case of throwing a coin. If we denote the probability of success in a Bernoulli trial by p and the number of successes in a series of N identically and independently repeated Bernoulli trial by k, then the expected number of successes $E(k)$ is given by $N \times p$. For the case of the cross-classification of a document collection of size N, the expected number of observations in cell category (i, j), $E(tf_{ij})$ is therefore given by $N \times p_{ij}$, where $N \times p_{ij} = N \times (p_i \times p_j) = e_{ij}$ under independence, for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$.

Under this random experiment design, the difference of the ratio $tf_{ij}/E(tf_{ij})$ from the value one gives the direction and magnitude of the degree of *divergence from randomness* in the observed cell frequency tf_{ij}. A given actual bivariate distribution of the random variables X and Y in the population determines the probability of classifying an observation in cell category (i, j), p_{ij}, and one would therefore measure the degree of divergence from randomness in cell frequencies. On the other hand, since the probability of classifying an observation in cell category (i, j) can be expressed under independence as the product of the marginal probability of classifying the observation in row category i and the marginal probability of classifying the observation in column category j, $p_i \times p_j$, we can calculate the degree of divergence from independence in cell frequencies without knowing or making any assumption about the actual bivariate distribution of the random variables X and Y on the population. This is the reason why divergence from independence is considered the nonparametric counterpart of divergence from randomness, from the viewpoint of statistics.

It can be shown that the number of successes k in a series of N identically and independently repeated Bernoulli trial follows a *binomial* distribution with mean $\mu = Np$ and variance $\sigma^2 = Np(1 - p)$. Let K be a binomial random variable. Then, it can also be shown that the random variable $Z = (K - \mu)/(\sqrt{\sigma^2})$, which is a transformation of the binomial random variable K, follows a binomial distribution with zero mean and unit variance. This transformation, which is commonly referred to as the *standardization* in statistics (i.e., standard scores or z-scores), transforms any given set of data to a set of data with zero mean and unit variance, while preserving the shape of the distribution of the original data.

In consequence, the standardized frequency in cell (i, j),

$$z_{ij} = \frac{tf_{ij} - Np_{ij}}{\sqrt{Np_{ij}(1 - p_{ij})}},$$

follow a binomial distribution with zero mean and unit variance, for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$.

Now suppose, without loss of generality, that we have a contingency table with only two cells. Let tf be the number of observations in the first cell, N be the total number of observations, and p be the probability of classifying an observation in the first cell. Then, the degree of divergence from randomness in the first cell and the degree of divergence from randomness in the second cell are respectively given by

$$\frac{tf}{Np} - 1 \quad \text{and} \quad \frac{N - tf}{N(1 - p)} - 1,$$

where the expected frequency $E(tf)$ for the first cell is Np, and the expected frequency $E(N - tf)$ for the second cell is $N(1 - p)$. The total degree of divergence from randomness in this contingency table with two cells can be calculated as

$$\frac{(tf - Np)^2}{Np} + \frac{(N - tf - N(1-p))^2}{N(1-p)} = \left(\frac{tf - Np}{\sqrt{Np(1-p)}}\right)^2 = z^2.$$

This suggests that the total degree of divergence from randomness in a given $(r \times c)$ contingency table can be measured as the sum of squares of the standardized cell frequencies:

$$\sum_{i=1}^{r}\sum_{j=1}^{c} z_{ij}^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(tf_{ij} - Np_{ij})^2}{Np_{ij}(1-p_{ij})}.$$

Now consider the case of the cross-classification of a large document collection, where classifying an observation into any cell category is a rare event, i.e., p_{ij}s are small in magnitude. In such a case, the mean and the variance of the binomial distribution associated with each cell would approximately be equal to each other: $Np_{ij} \approx Np_{ij}(1 - p_{ij})$ for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$. Thus, as N goes to infinity and success probabilities approach to zero, the total degree of divergence from randomness in cell frequencies can be approximated as

$$\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(tf_{ij} - Np_{ij})^2}{Np_{ij}}.$$

It follows that the total degree of divergence from independence can be calculated as

$$G^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(tf_{ij} - e_{ij})^2}{e_{ij}},$$

where $e_{ij} = N(p_i p_j) = Np_{ij}$ under independence. Each component of G^2, which measures the normalized chi-squared distance from independence in cell frequencies,

$$z_{ij}^2 = \frac{(tf_{ij} - e_{ij})^2}{e_{ij}}, \tag{2}$$

can be used for measuring the degree of DFI in cell frequencies as well as the standardized distance,

$$z_{ij} = \frac{tf_{ij} - e_{ij}}{\sqrt{e_{ij}}}. \tag{3}$$

3.3 Term frequency normalization

Term frequency normalization is an important issue for the state-of-the-art term weighting methods, and it has shown that it has a significant impact on retrieval performance (Singhal et al. 1996; He and Ounis 2003, 2005).

The ratio of the observed cell frequency tf_{ij} to the frequency expected under independence can be rewritten as

$$\frac{tf_{ij}}{(TF_i D_j)/N} = \frac{tf_{ij}}{(TF_i D_j)/c\overline{D}} = \frac{tf_{ij} \times (\overline{D}/D_j)}{TF_i/c} = \frac{tfn_{ij}}{\lambda_i},$$

where tfn_{ij} is the normalized frequency of term t_i, as defined in the work of Amati and van Rijsbergen (2002), λ_i is the sample estimate of the population mean frequency of

occurrence of term t_i in a document, and $\overline{D}$ is the average length of a document in the collection given. This shows that DFI term weighting models implicitly apply term frequency normalization, which fosters our proposed DFI measures further.

4 Scoring documents: an information theoretic approach

Our past TREC experiences on DFI-based term weighting suggest that the way how to aggregate DFI measurements over all terms of a given query for getting the final scores of documents is one of the factors that largely influences retrieval performance. The document scoring function which best accords, by design, with the DFI measures in (1), (2), and (3) is the one that is stemmed from Shannon's *information theory* (1949). On the basis of the theory of probability, Shannon ascribes a measure to the information conveyed by a signal s through a channel from a "source" to a "destination" as $\mathrm{I}(s) = -\log_2[Pr(s)]$, where $Pr(s)$ denotes the probability of observing signal s in the channel. The information that is conveyed by a signal, $\mathrm{I}(s)$, increases as the probability of occurrence of that signal in the channel, $Pr(s)$, decreases. This measure of information quantifies the *self* information value of a signal in terms of the number of "bits" required to represent that signal within the finite set of signals to which the signal belongs.

Assume that a message M of length N is a series of N signals each of which is identically and independently emitted from a source to a destination. Cross-classification of a document collection into $r \times c$ cell categories can then be modeled as a transmission. The set of $r \times c$ cell categories described herein corresponds to the finite set of $r \times c$ signals, where the probability of occurrence of signal s_{ij} in the channel is equal to the probability of classifying an observation into the cell category (i, j), p_{ij}, for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$. Accordingly, in such a transmission, each signal s_{ij} is expected to occur Np_{ij} times in the channel.

In this information theoretic model, emission of a signal refers to a Bernoulli trial, where success corresponds to the emission of the signal s_{ij} of interest, and failure the emission of one of the other $(r \times c) - 1$ signals. Thus, the emission probability p_{ij} associated with each signal s_{ij} is again determined by a binomial distribution with mean Np_{ij} and variance $Np_{ij}(1 - p_{ij})$, for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$.

On the basis of the model given, self-information value of a signal s_{ij} can be expressed as

$$-\mathrm{I}(\mathrm{s_{ij}}) = \log_2(\mathrm{p_{ij}}) = \log_2\left[\frac{1}{\mathrm{rc}}\right] + \log_2\left[\frac{\mathrm{p_i}}{1/\mathrm{r}}\right] + \log_2\left[\frac{\mathrm{p_j}}{1/\mathrm{c}}\right] + \log_2\left[\frac{\mathrm{p_{ij}}}{\mathrm{p_i p_j}}\right],$$

for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$. This is one of the log-linear forms of the saturated model of independence which would in general provide a better fit to the log-transformed cell densities than that of the original model to the raw cell densities (Hoaglin et al. 1983). In this case, under independence, the residual $\log_2[p_{ij}/(p_i \cdot p_j)]$ is necessarily 0 instead of 1, and itself gives the magnitude and direction of DFI in self-information. It follows that, for any term $t_i (i = 1, 2, \ldots, r)$ and document $d_j (j = 1, 2, \ldots, c)$, the degree of DFI in self-information is given by

$$ldfi_{ij} = \log_2\left[\frac{tf_{ij}}{e_{ij}}\right] = \log_2\left[\left(\frac{tf_{ij}}{e_{ij}} - 1\right) + 1\right] = \log_2[dfi_{ij} + 1], \tag{4}$$

for $(dfi_{ij} > 0)$ and zero elsewhere, where "*ldfi*" stands for "log-DFI" and *dfi* is the measure of DFI given in (1). In a general sense, the lesser the probability of observing a degree of

DFI in tf_{ij} equal to dfi_{ij} the higher is the amount of information on the content of d_j that we would get by observing term t_i, and vice versa.

As a result, given a query q, the simplest information theoretic, DFI-based document scoring function can be formulated as

$$\sum_{t_i \in q \cap d_j} qtf_i \times ldfi_{ij},$$

for $(tf_{ij} - e_{ij}) > 0$ and zero elsewhere, where qtf_i is the frequency of term t_i in q. This document scoring function basically measures the contribution of document d_j to the information on q under the assumption of independence.

4.1 On the two sources of information

The fact that the measures of DFI presented in this article are defined from a cell-centric viewpoint suggests that they utilize only the first source of information, "within document term frequencies". In scoring documents with DFI, in order to take into account the second source of information, "term frequency distributions", the measured degree of DFI in the self-information value of each query term needs to be weighted in accordance with the contribution of the terms to *total inertia*.

The total inertia can be thought of as a measure of the magnitude of the total row squared deviations from independence or equivalently the magnitude of the total column squared deviations. The total inertia associated with the system represented by a given data matrix of frequencies is given by G^2/N, where the contribution of a term t_i to the total inertia (CTI) is

$$G_i^2/N = \frac{1}{N} \sum_{j=1}^{c} \frac{(tf_{ij} - e_{ij})^2}{e_{ij}},$$

for $i = 1, 2, \ldots, r$. Under independence, contribution of each term to total inertia is necessarily zero.

Under independence, the collection frequency TF_i of t_i is expected to be distribute on documents proportional to the lengths of documents D_js, resulting in a zero contribution to total inertia, $G_i^2/N = 0$. This means, on the basis of the DFI hypothesis, that a term of which the use in documents is due to a reason other than serving to impart knowledge would have no contribution to total inertia. Thus, the weight that would be assigned by means of G_i^2/N for term t_i is expected to be close to zero in magnitude when the use of t_i in documents is of function type: otherwise, a weight that is greater than zero is expected.

As a result, to take into account the second source of information in the sense of term specificity, one can use CTI as a weighting factor in scoring documents:

$$\text{DFI} \times \text{CTI}(t_i) = ldfi_{ij} \times \log_2[G_i^2],$$

where G_i^2/N is equivalent to G_i^2 with respect to document ranking.

4.2 A DFI × IDF weighting scheme

Essentially, the function of CTI component in DFI × CTI weighting scheme is the same as the function of the IDF component of TF × IDF weighting scheme (Salton and Buckley

1988): both are a measure of term specificity. Thus, it is also possible to use IDF instead of CTI for scoring documents based on DFI.

Robertson (2004) states that "a query term which occurs in many documents is not a good discriminator, and should be given less weight than one which occurs in few documents". Although both IDF and CTI are a measure of term specifity, they are different in that the weight of a term occurred in whole documents of a collection is measured zero by IDF. In measuring term specificity based on CTI, a term may be given a zero weight only if its frequency of occurrence is independent of documents. Otherwise, a weight that is greater than zero will be given, no matter how many documents the term occurs in.

In theory, this property of CTI provides a certain stability in measuring term specificity. Suppose that the term of interest is "and". Since it is likely that the term "and" occurs in every document, IDF would fail to identify its specificity to a document that could provide information on the linguistic properties of "and" as a conjunction, whereas CTI would not, to a certain extent.

Besides, IDF suffers, at least in theory, from a certain weakness for which CTI can also be a remedy. Church and Gale (1995) state this weakness as "… *favoring extremely rare words, no matter how they are distributed.*"

4.3 The relation between DFR and DFI

Amati and van Rijsbergen (2002) express the DFR term weighting model in terms of probability models as "the weight of a term in a document is a function of two probabilities $Prob_1$ and $Prob_2$": $w = (1 - Prob_2) \times \ - \log_2(Prob_1)$, where $Prob_1$ refers to the conditional probability of the frequency of occurrence of a term in a document to the whole document collection, and $Prob_2$ refers to the same probability but conditional to the elite set. In the DFR framework, it is assumed that elite set of a term is the set of all documents in which the term occurs. In its original definition, $Prob_1$ is determined by a basic randomness model, a probability distribution function with mean λ_1, and $Prob_2$ is determined by a probability distribution function, which is different from the basic randomness model assumed, with mean λ_2.

In the family of Pareto distributions, a Type-I Pareto distribution has two parameters (Arnold 1983), a (necessarily positive) minimum possible value of K, $k_{min} > 0$, and a positive valued parameter $\xi > 0$ (i.e., Pareto index), and it is given by

$$Pr(K \leq k | k_{min}) = 1 - \left(\frac{k_{min}}{k}\right)^{\xi},$$

for $k \geq k_{min}$, and zero elsewhere. A Type-I Pareto distribution with $k_{min} = Np_{ij}$ and $\xi = 1$ can be considered as the characteristic distribution function of the degree of divergence from randomness in observed cell frequencies tf_{ij} for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$:

$$Pr[DFR(tf_{ij})] = Pr(K \geq tf_{ij} | Np_{ij}) = \frac{Np_{ij}}{tf_{ij}},$$

for $tf_{ij} \geq Np_{ij}$, and zero elsewhere. Assuming that the minimum possible frequency of occurrence of a function word in a document is k_{min}, the probability of being a function word for term t_i in document d_j can therefore be calculated under independence as

$$Pr(K \geq tf_{ij}|e_{ij}) = \frac{e_{ij}}{tf_{ij}} = Pr(dfi_{ij}),$$

for $tf_{ij} \geq e_{ij}$, and zero elsewhere. The degree of divergence from independence in self-information can thereby be rewritten as

$$ldfi_{ij} = \log_2\left[\frac{tf_{ij}}{e_{ij}}\right] = \log_2\left[\frac{1}{Pr(dfi_{ij})}\right] = -\mathrm{I}(dfi_{ij}).$$

Suppose that the basic randomness model is a Type-I Pareto distribution with $k_{min} = \lambda_1$ and $\xi = 1$:

$$Prob_1 = Pr(K \geq tf_{ij}|\lambda_1) = \frac{\lambda_1}{tf_{ij}}.$$

Then, the informative content of a term t_i in document d_j, which is defined by Amati and van Rijsbergen (2002) as $\mathrm{Inf}_1(tf_{ij}) = -\log_2(Prob_1)$, can be expressed as the degree of divergence from randomness in self-information:

$$\mathrm{Inf}_1(tf_{ij}) = -\log_2\left[\frac{\lambda_1}{tf_{ij}}\right],$$

where $\lambda_1 = Np_{ij}$ for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$.

The total information contained in a message $\mathbf{M}$ of N independently and identically emitted signals from a source with emission probability p_{ij} is given by

$$\mathrm{I}(\mathbf{M}) = \sum_{i,j} Np_{ij} \times -\log_2(p_{ij}).$$

This also gives the total information that would be contained in a document collection $\mathbf{D}$ of length N under randomness or a pure chance distribution. The observed total information in a given document collection can be calculated as

$$\mathrm{I}(\mathbf{D}) = \sum_{i,j} N\frac{tf_{ij}}{N} \times -\log_2\left[\frac{tf_{ij}}{N}\right] = \sum_{i,j} tf_{ij} \times -\log_2\left[Pr(tf_{ij})\right].$$

It follows that the total degree of divergence from randomness in self-information is given by

$$\begin{aligned}
\mathrm{I}(\mathrm{DFR}(\mathbf{D})) &= \sum_{i,j} DFR(tf_{ij}) \times -\log_2\left(Pr[DFR(tf_{ij})]\right) \\
&= \sum_{i,j}\left(\frac{tf_{ij}}{Np_{ij}} - 1\right) \times -\log_2\left[\frac{Np_{ij}}{tf_{ij}}\right] \\
&= \sum_{i,j}\left(1 - \frac{tf_{ij}}{Np_{ij}}\right) \times \log_2\left[\frac{Np_{ij}}{tf_{ij}}\right] \\
&\equiv \sum_{i,j}\left(1 - \frac{Np_{ij}}{tf_{ij}}\right) \times -\log_2\left[\frac{Np_{ij}}{tf_{ij}}\right] \\
&= \sum_{i,j}\left(1 - \frac{\lambda_1}{tf_{ij}}\right) \times -\log_2\left[\frac{\lambda_1}{tf_{ij}}\right],
\end{aligned}$$

where the third line is rank equivalent to the fourth line.

Thus, assuming a Type-I Pareto distribution for $Prob_2$ with $k_{min} = \lambda_2 = \lambda_1, \mathrm{I}(\mathrm{DFR}(\mathbf{D}))$ can be rewritten as

$$\mathrm{I}(\mathrm{DFR}(\mathbf{D})) = \sum_{i,j}(1 - Prob_2) \times -\log_2(Prob_1).$$

In here,

$$Pr(K \leq tf_{ij}|\lambda_1) = 1 - \frac{\lambda_1}{tf_{ij}} = 1 - Prob_2 = 1 - Prob_1$$

gives the probability of observing term t_i in document d_j less than λ_1 times. In one sense, it refers to the risk of accepting the term t_i as a descriptor of the document d_j. The odds in favor of accepting t_i as a descriptor of d_j is given by

$$\frac{Prob_1}{1 - Prob_1}.$$

Given a pair of terms in a particular document, the odds ratio gives the risk of accepting one of the terms relative to the other as the descriptor of the document, i.e., relative risk.

This suggests, as a result, that DFR models can be reduced to DFI models under independence. To give a concrete example, consider the DFR model PL2 which assumes a Poisson distribution with mean $\mu = \lambda_1$ for the basic randomness model, and a probability distribution based on Laplace law of succession for $Prob_2$. Then,

$$Prob_2 = Pr(tf_{ij} + 1|tf_{ij}) = \frac{tf_{ij}}{tf_{ij} + 1},$$

which in fact gives the probability of observing a term t_i that was already occurred tf_{ij} times in document d_j, one more time, under a Type-I Pareto distribution with $k_{min} = \lambda_2 = tf_{ij}$ and $\xi = 1$. This suggests that PL2 can be reduced to a nonparametric model based on DFI under independence.

5 The DFI retrieval functions

This study proposes three measures of statistical dependence for the purpose of term weighting:

$$dfi_{ij} = \frac{tf_{ij} - e_{ij}}{e_{ij}}$$

based on the saturated model of independence;

$$z_{ij} = \frac{tf_{ij} - e_{ij}}{\sqrt{e_{ij}}} \quad \text{and} \quad z_{ij}^2 = \frac{(tf_{ij} - e_{ij})^2}{e_{ij}}$$

based on standardization and normalized chi-squared distance, respectively, for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$.

Based on a measure of statistical dependence, the degree of relevance of a document d_j to a given query q can be measured as

$$\mathrm{R}(q, d_j) = \sum_{t_i \in q \cap d_j} qtf_i \times w_{ij}, \tag{5}$$

for $(tf_{ij} - e_{ij}) > 0$ and zero elsewhere. In here, w_{ij} represents a possible weighting scheme for weighting query term t_i with respect to document d_j based on DFI. In this study, we examined 9 possible DFI weighting schemes as listed in Table 1.

6 Experimental design and materials

In our experiments, TERRIER retrieval platform version 3.0 (Ounis et al. 2007) is used to index and search TIPSTER disks 1 & 2, TREC disks 4 & 5, and Clueweb09-T09B data sets. TIPSTER disks 1 & 2 consist of about 740,000 documents from the Wall Street Journal, the Federal Register, the Associated Press, Department of Energy abstracts, the Computer Select disks of Ziff-Davis, which are used in the ad hoc tracks of TREC 1 through 3. TREC disks 4 & 5 consist of about 550,000 documents from the Financial Times, the Congressional Record of the 103^{rd} Congress, the Federal Register, the Foreign Broadcast Information Service, and the Los Angeles Times, which are first used in TREC 6 ad-hoc track. They are also used in the ad-hoc tracks of TREC 7 & 8 and the robust tracks of TREC 2003 and TREC 2004 but without the Congressional Record (about 28,000 documents). Clueweb09-T09B consists of over 50 million (English) WEB pages crawled from the Internet in between January 2009 and February 2009, which is used in the WEB tracks of TREC 2009, 2010, and 2011 as "Category B" document collection.

During indexing and searching, terms are stemmed by the TERRIER's implementation of Porter's stemmer but no stop-word elimination is applied to the documents: DFI models depend on the lengths of documents and stop-word elimination degenerates documents in length, arbitrarily. Nevertheless, performance evaluations also carried out on the stop-word eliminated versions of TIPSTER disks 1 & 2 and TREC disks 4 & 5 document collections for the comparison.

During indexing, a cascaded term filter is applied to the TERRIER's default term validity checker, which can be expressed in a regular expression as

$[\mathtt{a}-\mathtt{zA}-\mathtt{Z}]+\backslash\backslash' * \backslash\backslash\& * [\mathtt{a}-\mathtt{zA}-\mathtt{Z}]+[\mathtt{0}-\mathtt{9}]\{\mathtt{0},\mathtt{4}\}|[\mathtt{0}-\mathtt{9}]\{\mathtt{1},\mathtt{4}\}.$

This regular expression accepts such symbol sequences as valid terms which are composed of only letters, or an apostrophe surrounded by letters (e.g. "John's"), or an

Table 1 DFI term weighting schemes

	Abbrv.	Scheme (w_{ij})	Description
I	DFIB	$\log_2(idf_{ij} + 1)$	Saturated model of independence
II	DFIZ	$\log_2(z_{ij} + 1)$	Standardization
III	DFIC	$\log_2\left(z_{ij}^2 + 1\right)$	Normalized chi-squared distance
IV	DFIB-CTI	(I) $\times$ $\log_2(G_i^2)$	DFIB $\times$ contribution to total inertia
V	DFIZ-CTI	(II) $\times$ $\log_2(G_i^2)$	DFIZ $\times$ contribution to total inertia
VI	DFIC-CTI	(III) $\times$ $\log_2(G_i^2)$	DFIC $\times$ contribution to total inertia
VII	DFIB-IDF	(I) $\times$ $\log_2(c/c_i)$	DFIB $\times$ inverse document frequency
VIII	DFIZ-IDF	(II) $\times$ $\log_2(c/c_i)$	DFIZ $\times$ inverse document frequency
IX	DFIC-IDF	(III) $\times$ $\log_2(c/c_i)$	DFIC $\times$ inverse document frequency

ampersand surrounded by letters (e.g. "AT&T"), or letters followed by numbers of maximum 4 digit long (e.g. "PS2", "BM25", "TREC2004", etc), or numbers of maximum 4 digit long.

In each track of TREC, a set of 50 topics is used to measure and compare the retrieval performances of the runs submitted to the track, which cumulatively sum up to 550 topics over 11 tracks: the topics 51–200 of TREC 1, 2, and 3 ad hoc tracks; the topics 301–450 of TREC 6, 7, and 8 ad hoc tracks; the topics 601–700 of TREC 2003 and 2004 robust tracks; and the topics 1–150 of TREC 2009, 2010 and 2011 WEB tracks. But, there are actually 545 effective topics in total due to the lack of success in discovering relevant documents to some topics at the time of construction. The topics with no relevant documents are the topic 672 of TREC 2004 robust track and the topics 19, 20, 95, and 100 of TREC 2009 WEB tracks.

While scoring a document, TERRIER retrieval platform ignores the terms that have low IDF (i.e., it ignores a term if the marginal frequency of the term is greater than the number of documents in a given document collection: $TF_i > c$) by default. This feature is reasonable for the short and long queries, so it is left intact. However, for the very short queries, it would be either ineffective whenever such a query term is not present or dramatically effective in the negative sense whenever most of the query terms would be ignored in scoring, such as topic 42 "the music man", or topic 70 "to be or not to be that is the question", or topic 92 "the wall". Thus, it is not used for very short queries.

Every term weighting method is allowed to return a result set of maximum 1,000 documents from a corresponding data set to an associated query. For the TREC 2009, 2010, and 2011 Web track topics, spam-page filtering (Cormack et al. 2010) is applied to the result sets returned by each term weighting model, as given by

$$\mathrm{sds}(q, d_j) = [\delta \times \mathrm{R}(q, d_j)] + [(1 - \delta) \times (\phi \times \mathrm{R}(q, d_j))],$$

where $0 \leq \delta \leq 1$ is a weighting factor (where $\delta = 1$ means no spam-page filtering) and $0 \leq \phi \leq 0.50$ is the percent of spammness calculated for document d_j in Cormack et al. (2010): the documents with $\phi > 0.50$ are assumed not spam in this study, where $\phi = 0$ refers to 100 % chance of being a spam document and $\phi = 0.50$ means 50 % chance of being a spam document. For example, $\delta = 0.5$ and $\phi = 0$ yields the half of the original score of document d_j, and if $\phi = 0.50$ (implying 50 % chance of being a spam document), then the original score is left intact at $\delta = 0.5$. The resultant $\mathrm{sds}(q, d_j)$ scores are then used to re-rank the result sets. The base run of each term weighting method which employs no spam-page filter is allowed to return a result set of maximum 10,000 documents. Spam-page filtering is then applied to that result sets of size 10,000, and the resultant re-ranked result sets are reduced to 1,000 in size.

The retrieval performance of DFI is compared with the retrieval performances of 5 standard term weighting methods as implemented in TERRIER. We follow the TERRIER's method naming convention for those weighting methods. Out of the DFR-based term weighting methods, there are 3 methods considered: "BM25", "H-LM", and "D-LM". "BM25" is the OKAPI's BM25 (Robertson et al. 1999). "H-LM" is the Hiemstra's language model (Hiemstra 2000). "D-LM" is also, a language model but it uses Bayesian smoothing with Dirichlet Prior (Zhai and Lafferty 2004). The remaining 2 are DFR-based methods (Amati and van Rijsbergen 2002; Macdonald et al. 2005; He and Ounis 2003): "DFree" and "PL2". "DFree" is a parameter-free DFR method that assumes a Hypergeometric term distribution. "PL2" is a parametric DFR method that assumes a Poisson term distribution as

a basic randomness model with Laplace after-effect, and tailored to the tasks that require early precision.

In the experiments, we measure retrieval performance on the out-of-the-box runs of those 5 term weighting methods with the default values of parameters as defined in TERRIER, since DFI models are of out-of-the-box type, by design. One may therefore obtain different performance scores than the ones presented in this article by tuning the values of necessary parameters for individual data sets, or by changing the operational settings of the experiments.

For the methods based on parametric models, change of operational settings, such as the change of document collection or query, would in general invalidate the tuned values of parameters and hence causes a reduction in their expected retrieval performance. The benefit of DFI approach is, in this respect, that a nonparametric model is expected to be robust against such changes (Sect. 8). In addition to the out-of-the-box runs of "BM25" and "PL2", the runs with the best parameter values for the TIPSTER disks 1 & 2 and the TREC disks 4 & 5 document collections, which are obtained by optimizing MAP using a simulated annealing process[1], are also taken into account to have strong baseline performance scores in comparisons. The best b parameter values of "BM25" are respectively 0.3277 and 0.3444, and the best c parameter values of "PL2" are 4.607 and 9.150. For the Clueweb09-T09B document collection, we used the best parameter values obtained on "WT10G, TREC9-10 Web Tracks": $b = 0.2505$ and $c = 12.33$. The best runs are denoted by "BM25B" and "PL2B" in the presentation of the experimental results.

7 Experimental results

The runs of the term weighting methods under consideration are evaluated using two performance measures: Mean Average Precision ("MAP") and the total number of relevant documents retrieved ("RR") over all queries.

The results of the experiments performed are presented in two parts. First, DFI-based models are compared with the 5 standard term weighting models using test collections with stemming but applying no stop-word elimination. Second, to see whether stop-word elimination affects the retrieval performance, the same analysis is repeated by applying stop-word elimination to TIPSTER disks 1 & 2 and TREC disks 4 & 5 document collections.

7.1 Part I

Table 2 lists the observed performance scores of the runs of the weighting models under consideration for both the *very short* (topic only) versions and the *short* (topic and description) versions of the (first) set of ad-hoc topics 51-200 (denoted by TS-I) and the (second) set of ad-hoc topics 301-450&601-700 (TS-II).

Figure 1 shows the multiple comparisons after Friedman's test (Hollander and Wolfe 1999) based on MAP scores for the very short versions of TS-I and TS-II. Friedman's test is the nonparametric counterpart of the balanced two-way ANOVA test: it tests only for row effect (i.e., methods) after adjusting for possible column effects (i.e., topics). It refers to the balanced one-way ANOVA with homogenous blocks. In this respect, it only tests the run effect or the topic effect, at a time, which in fact makes it more appropriate for the

[1] TERRIER official site: http://terrier.org/docs/v3.5/trec_examples.html.

Table 2 Performance scores of the runs of the term weighting models for both the **very-short** versions and the **short** versions of TS-I and TS-II

Model	Very-short queries				Short queries			
	TS-I		TS-II		TS-I		TS-II	
	MAP	RR	MAP	RR	MAP	RR	MAP	RR
DFIB	0.1876	16722	0.2273	9999	0.2235	18930	0.2689	10694
DFIC	0.1899	17020	0.2415	10068	0.2035	18766	0.2605	10570
DFIZ	0.1913	17092	0.2440	10170	0.2027	18659	0.2600	10458
B-CTI	0.1900	16859	0.2285	10058	0.2265	19130	0.2719	10787
C-CTI	0.1912	17125	0.2421	10094	0.2067	18986	0.2627	10680
Z-CTI	0.1928	17182	0.2450	10198	0.2062	18906	0.2630	10590
B-IDF	0.1887	16978	0.2198	9721	0.2197	19654	0.2578	10738
C-IDF	0.2187	18503	0.2417	10181	0.2377	20764	0.2733	11095
Z-IDF	**0.2222**	**18703**	0.2442	10293	**0.2392**	**20831**	**0.2764**	**11159**
BM25	0.1810	16037	0.2328	9990	0.1493	14003	0.1803	6409
BM25B	0.2042	17116	0.2479	10289	0.2356	19433	0.2705	10669
PL2	0.1522	14651	0.2169	9604	0.0916	12121	0.0800	4472
PL2B	0.2117	18325	**0.2539**	**10409**	0.2356	19480	0.2612	10229
D-LM	0.2202	18622	0.2416	10096	0.1735	16392	0.2321	9587
H-LM	0.1791	16351	0.2148	9640	0.1685	16018	0.2398	9846
DFRee	0.2170	17903	0.2478	10314	0.2316	19154	0.2718	10449

The best score on each performance measure is bold faced. The total number of relevant documents available for TS-I is 37,836 and for TS-II 17,733

multiple comparison of the results of IR experiments than ANOVA (Hull 1993). The multiple comparisons shown in Fig. 1 use Tukey's honestly significant difference criterion (Tukey's HSD), which is based on the Studentized range distribution.

For the very short version of the TREC 1, 2, and 3 ad-hoc topics, 51-200 (TS-I), it appears that "Z-IDF" model has a retrieval performance that is not significantly different from that of the baseline runs with best parameter values, "BM25B" and "PL2B". For this set of 150 topics, IDF component seems to contribute to the retrieval performance of DFI models significantly higher than that of CTI component, and when CTI component considered separately it does not result in a significant effect on the retrieval performance of DFI models.

For the very short version of the TREC 6, 7, and 8 ad-hoc topics and TREC 2003 and 2004 robust track topics, 301-450&601-700 (TS-II), the case is slightly different. For this set of 250 topics, "DFIZ" model itself has a retrieval performance that is not significantly different from that of the baseline runs, and neither IDF nor CTI component has a significant effect on the retrieval performance of "DFIZ" model as well as "DFIB" and "DFIC" models. Moreover, in contrast to TS-I, the effect of CTI component and the effect of IDF component are also not significantly different from each other for this topic set.

For the short versions of TS-I and TS-II, the multiple comparisons are given in Fig. 2. The case of short queries (topic and description) is similar to the case of the very short queries (topic only): DFI models have a retrieval performance that is not significantly different from that of baseline runs.

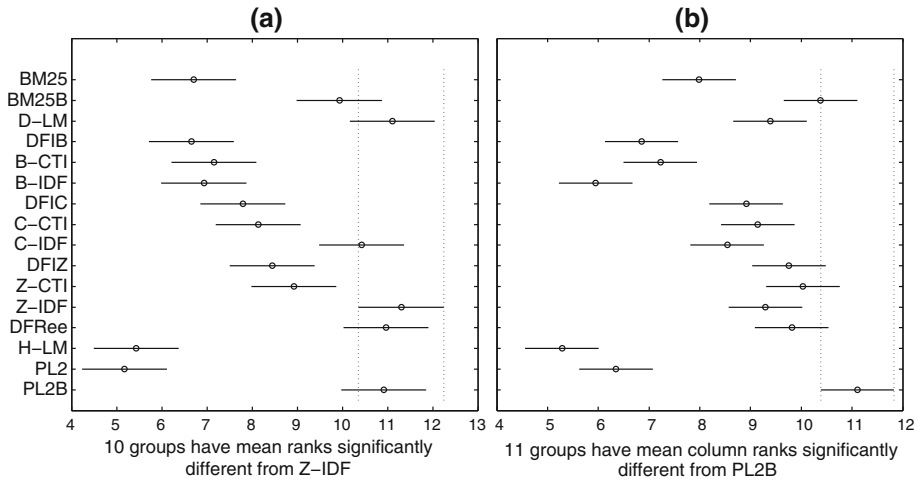

Fig. 1 Multiple comparisons after Friedman's test based on MAP scores for the **very short** versions of TS-I (**a**) and TS-II (**b**). In **a**, the mean rank of each method over 150 topics is marked by a circle, and the associated 95 % CI is depicted by a *horizontal solid line* crossing the *circle*. The *vertical dashed lines* emphasizes the ends points of the 95 % CI of "Z-IDF". Non-overlapping CIs indicates significant differences between the associated methods: 10 methods have mean ranks significantly different from "Z-IDF". Similarly, **b** shows the multiple comparison over the 250 topics in TS-II

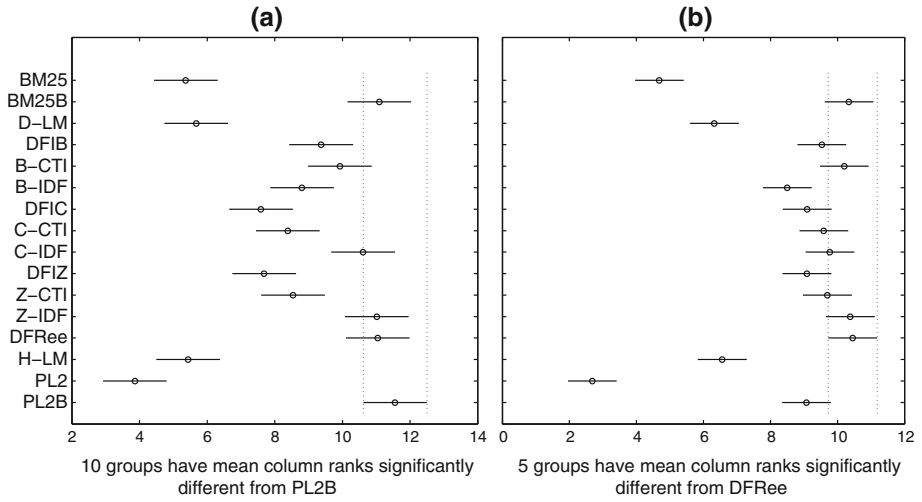

Fig. 2 Multiple comparisons based on MAP scores for the **short** versions of TS-I (**a**) and TS-II (**b**)

For the TREC 2009, 2010, and 2011 Web track topics 1-150 (TS-III), the observed performance scores are given in Table 3 for the varying values of spam-page filtering weight δ. Multiple comparison is given in Fig. 3.

Relating all the analyses performed, it would appear, as a result, that nonparametric DFI models perform as well as parametric models. The DFI model "Z-IDF" seems to be a good

Table 3 Performance scores of the runs of the term weighting models on TS-III for $\delta = 1$ (no spam-page filtering), $\delta = 0.25$, and $\delta = 0.5$

Model	$\delta = 1$		$\delta = 0.25$		$\delta = 0.5$	
	MAP	RR	MAP	RR	MAP	RR
DFIB	0.0972	4826	0.0968	4493	0.0928	4140
DFIC	0.1692	6118	0.1978	6016	0.1906	5313
DFIZ	0.1698	6137	**0.1981**	**6033**	**0.1911**	5335
B-CTI	0.0967	4815	0.0962	4489	0.0921	4135
C-CTI	0.1679	6105	0.1970	6005	0.1896	5309
Z-CTI	0.1687	6128	0.1972	6022	0.1900	**5337**
B-IDF	0.0911	4738	0.0885	4334	0.0855	4034
C-IDF	0.1663	6154	0.1912	5975	0.1852	5239
Z-IDF	0.1664	**6168**	0.1909	5955	0.1851	5238
BM25	0.1626	5520	0.1563	4880	0.1543	4697
BM25B	**0.1789**	5820	0.1868	5162	0.1844	4939
PL2	0.1366	5378	0.1452	5244	0.1373	4660
PL2B	0.1711	6074	0.1953	5790	0.1886	5182
D-LM	0.1608	6019	0.1917	5949	0.1845	5239
H-LM	0.0902	4672	0.0905	4426	0.0862	4009
DFRee	0.1717	5670	0.1710	5262	0.1669	4902

The total number of relevant documents available for TS-III is 8,754

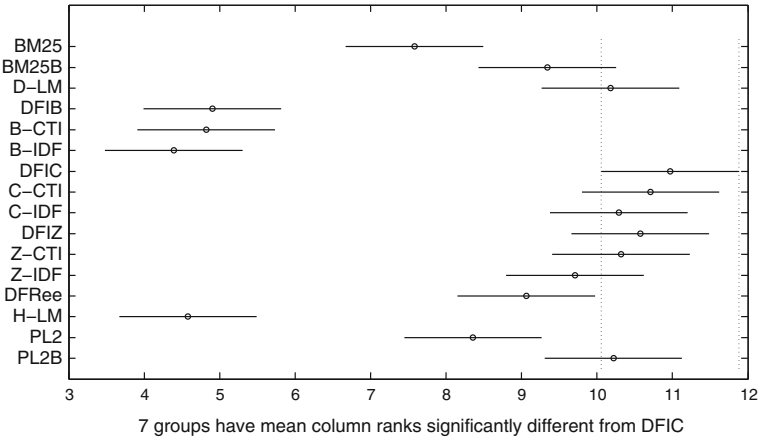

Fig. 3 Multiple comparisons based on MAP scores for TS-III, using spam-page filtering applied result sets with $\delta = 0.25$

choice among the alternatives for well-structured, controlled document collections like the ones in TIPSTER disks 1 & 2 and TREC disks 4 & 5. On the other hand, the DFIZ or DFIC model with or without IDF/CTI component seems to be a good choice for non-controlled document collections like the collection of Web pages, after applying spam-page filtering.

7.2 Part II

This part presents the evaluation of the runs of the term weighting methods using the stop-word eliminated versions of TIPSTER disks 1 & 2 and TREC disks 4 & 5. Table 4 lists the observed performance scores for both the very short versions and the short versions of TS-I and TS-II.

For the term weighting models, "PL2B" and "BM25B", the parameter values are tuned based on MAP for very short versions of topic sets using the stop-word eliminated versions of document collections. Thus, the observed MAP scores are the maximum ones for those runs.

Together with the results of the multiple comparisons given in Fig. 4, it can be said, all in all, that applying stop-word elimination does not result in a significant effect on the retrieval performance of DFI models. DFI models are still comparable with the baseline runs in retrieval performance.

It is seen clearly that the CTI component has a significant contribution to the retrieval performance of DFI models, especially for the model DFIZ. Interestingly, in contrast to CTI, IDF component results in a consistent negative effect on retrieval performance.

8 Discussion

In the context of statistical inference, the advantage of a nonparametric procedure is that, even when the parametric assumptions about the distribution of the sampled population

Table 4 Performance scores of the runs of the term weighting models for both the **very short** versions and the **short** versions of TS-I and TS-II using the stop-word eliminated versions of the corresponding document collections

Model	Very-short queries				Short queries			
	TS-I		TS-II		TS-I		TS-II	
	MAP	RR	MAP	RR	MAP	RR	MAP	RR
DFIB	0.1834	16780	0.2311	10084	0.1966	17966	0.2578	10465
DFIC	0.2085	17829	0.2417	10090	0.1992	18340	0.2560	10361
DFIZ	0.2111	17957	0.2441	10202	0.1988	18347	0.2564	10337
B-CTI	0.1879	16975	0.2330	10135	0.2030	18201	0.2633	**10575**
C-CTI	0.2113	17993	0.2426	10114	0.2036	18592	0.2590	10429
Z-CTI	0.2136	18119	0.2452	10234	0.2032	18577	0.2592	10454
B-IDF	0.1511	16111	0.2218	9691	0.1328	16151	0.1955	9316
C-IDF	0.1810	17549	0.2388	10098	0.1685	17949	0.2261	9923
Z-IDF	0.1839	17751	0.2411	10208	0.1710	18126	0.2293	10019
BM25	0.2125	18116	0.2356	10028	**0.2306**	**19543**	0.2611	10495
BM25B	**0.2306**	**18902**	0.2483	10230	0.2249	19046	0.2595	10334
PL2	0.2050	17620	0.2243	9734	0.2242	19178	0.2520	10272
PL2B	0.2227	18499	**0.2523**	**10345**	0.2210	18954	0.2567	10212
D-LM	0.1861	17343	0.2301	9806	0.1293	13980	0.1720	8119
H-LM	0.1666	15995	0.2202	9748	0.1451	14462	0.1975	8883
DFRee	0.2157	17979	0.2491	10286	0.2163	18646	**0.2648**	10411

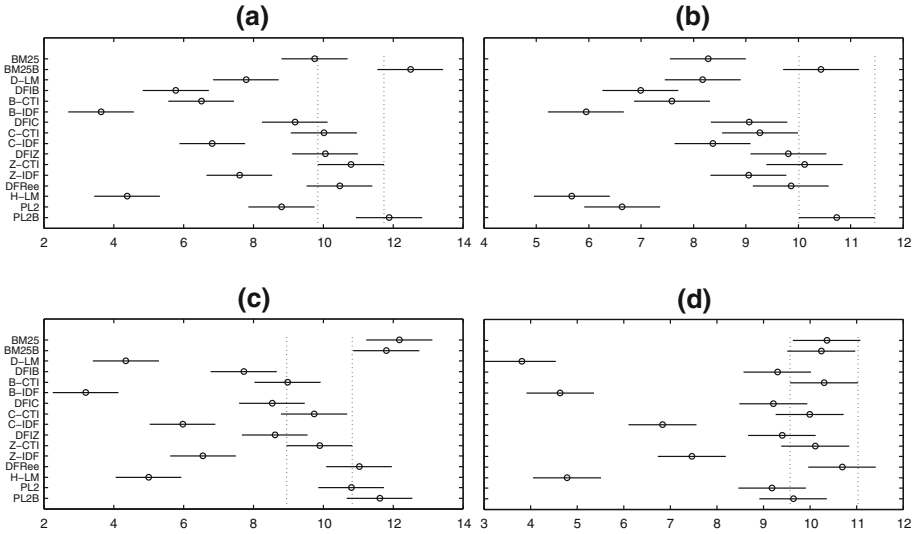

Fig. 4 Multiple comparisons using the stop-word eliminated version of the document collections in TIPSTER disks 1 & 2 and TREC disks 4 & 5: **a** for the very-short version of TS-I, **b** for the very-short version of TS-II, **c** for the short version of TS-I, and **d** for the short version of TS-II

hold perfectly true, a nonparametric procedure is only slightly less powerful than its parametric counterpart. But, if the parametric assumptions are failed to hold, only the nonparametric test procedures are valid (Mosteller and Tukey 1977). This implies that a parametric term weighting model should outperform its nonparametric counterpart on the document collections in which the frequency of term occurrence perfectly follows the distribution assumed by the parametric model. For example PL2 assumes a Poisson term distribution, and it should outperform DFI models on those document collections where the actual frequency of term occurrence follows a Poisson distribution. On the other hand, it also implies that nonparametric term weighting methods tend to approach and stay close to their parametric counterparts in retrieval performance. A nonparametric term weighting model is, in this respect, expected to outperform its parametric counterpart on a document collection in which the frequency of term occurrence imperfectly follows the assumed parametric distribution. As a result, it is reasonable to expect that DFI models will, on average, tend to approach and stay close to the parametric models in average retrieval performance over a large set of document collections. The fact that term distribution is likely to vary on the different document collections suggests that it is hard, if it is not impossible, to single out a parametric model whose assumptions hold perfectly true for all document collections. Change of operational settings, such as the change of document collection or query, would in general invalidate parametric models, and hence cause a reduction in their expected retrieval performance, whereas, in contrast, a nonparametric model is expected to be robust against such changes.

9 Conclusion

We present a nonparametric term weighting method based on measuring the *divergence from independence*-DFI, upon which we build an information theoretic document scoring

function. We propose here three basic measures of DFI: (1) "DFIB" based on the saturated model of independence, (2) "DFIZ" based on standardization, and (3) "DFIC" based on the normalized chi-squared distance.

The objective of this study is first to introduce nonparametric term weighting methods, and second to show that the theoretical expectation that nonparametric term weighting methods tend to approach and stay close to parametric methods in average retrieval performance over a large set of topics and document collections holds true in practice. The empirical results overall confirms this expectation to be true, at least in the body of the data at hand. In this regard, the DFR model "DFRee", for example, assumes a hyper-geometric distribution for the term distribution. There is no empirical evidence in support for the hypothesis that assuming a hyper-geometric distribution, relative to making no assumption (i.e., DFI models), for the actual term distribution significantly improves retrieval performance. The same is true for the DFR model "PL2", which assumes a Poisson term distribution.

Here, we also propose a DFI-based measure of term specificity, CTI (contribution of terms to total inertia), as an alternative to the well-known term specificity measure, called the IDF, inverse document frequency, and examine the retrieval performance of the corresponding DFI $\times$ CTI weighting scheme in comparison with the DFI $\times$ IDF weighting scheme, to see (1) whether CTI contributes to the retrieval performance of DFI models, and (2) whether there is a significant difference between CTI and IDF in contribution to the performance of DFI models. The empirical results show that IDF and CTI are indifferent in contribution to retrieval performance of DFI models.

As a result, we conclude that DFI-based term weighting promises a new direction in IR research, as being a simple but powerful baseline alternative to the state-of-the-art parametric models.

Acknowledgments Authors are thankful to anonymous reviewers for their valuable comments and advices that make this a better paper, and also to Craig Macdonald, Giambattista Amati, and Iadh Ounis for their kind helps. Index term weighting by DFI is developed under the project titled "Design of A Statistical Information Retrieval System", and supported by TUBITAK, The Scientific and Technological Research Council of Turkey, with Project No:107E192. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

Agresti, A. (2002). *Categorical data analysis*. New Jersey: Wiley-Interscience .

Amati, G., van Rijsbergen, C. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS), 20*(4), 357–389.

Arnold, B. C. (1983). *Pareto distributions*. Fairland, Maryland: International Cooperative Publishing House.

Bookstein, A., Swanson, D. (1974). Probabilistic models for automatic indexing. *Journal of the American Society for Information Science (JASIS)*,25, 312–318.

Bradley, J. V. (1968). *Distribution free statistical tests*. Englewood Cliffs, NJ: Prentice Hall

Church, K. W. (1995). One term or two? In: SIGIR'95: Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval (pp 310–318). Seattle, US.

Church, K. W., Gale, W. (1995). Inverse document frequency (IDF): A measure of deviations from Poisson. In: D. Yarowsky, & K. Church (Eds.), *Proceedings of the ACL 3rd workshop on very large corpora, ACL, MIT* (pp 121–130).

Clarke, C. L. A., Craswell, N., Soboroff, I., & Cormack, G. V. (2010). Overview of the trec 2010 web track. In: Proceedings of the 19th text retrieval conference (TREC'10), Gaithersburg, MD, USA.

Clinchant, A., Gaussier, E. (2010). Information-based models for ad hoc ir. In: Proceeding of the 33rd annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'10), (pp 234–241).

Cooper, W., & Maron, M. (1978). Foundations of probabilistic and utility-theoretic indexing. *Journal of Association for Computing Machinery*,25, 67–80.
Cormack, G. V., Smucker, M. D., & Clarke, C. L. A. (2010). Efficient and effective spam filtering and re-ranking for large web datasets URL http://arxiv.org/abs/1004.5168, retrieved from http://arxiv.org/abs/1004.5168v1, 1004.5168.
Croft, W. B., & Harper, D. J. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation, 35*(4), 285–295.
Damerau, F. (1965). An experiment in automatic indexing. *American Documentation*,16, 283–289.
Dinçer, B. T. (2012). Irra at trec 2012: Index term weighting based on divergence from independence model. In: Proceedings of the 21th text retrieval conference (TREC'12), Gaithersburg, MD.
Dinçer, B. T., Kocabaş, I., & Karaoğlan, B. (2009). Irra at trec 2009: Index term weighting based on divergence from independence model. In: Proceedings of the 18th text retrieval conference (TREC'09), Gaithersburg, MD.
Dinçer, B. T., Kocabaş, I., & Karaoğlan, B. (2010). Irra at trec 2010: Index term weighting based on divergence from independence model. In: Proceedings of the 19th text retrieval conference (TREC'10), Gaithersburg, MD.
Fuhr, N. (1989). Models for retrieval with probabilistic indexing. *Journal of Information Processing and Management, 25*(1), 55–72.
Harter, S. (1975). A probabilistic approach to automatic keyword indexing. Part I: On the distribution of specialty words in a technical literature. *Journal of the American Society for Information Science (JASIS), 26*, 197–216.
Harter, S. (1975). A probabilistic approach to automatic keyword indexing. Part II: An algorithm for probabilistic indexing. *Journal of the American Society for Information Science (JASIS), 26*, 280–289.
He, B., & Ounis, I. (2003). A study of parameter tuning for term frequency normalization. In: Proceedings of the 12th international conference on information and knowledge management, New Orleans, LA.
He, B., & Ounis, I. (2005). Term frequency normalisation tuning for BM25 and DFR model. In: Proceedings of the 27th European conference on information retrieval (ECIR'05) (pp. 200–214).
Hiemstra, D. (2000). A probabilistic justification for using tf.idf term weighting in information retrieval. *International Journal on Digital Libraries, 3*(2), 131–139.
Hintikka, J. (1970). On semantic information. In: J. Hintikka, & P. Suppes (Eds.), *Information and inference* (pp. 3–27). Dordrecht: Synthese Library.
Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.) (1983). Understanding robust and exploratory data analysis. Wiley series in probability and mathematical statistics. Wiley-Interscience
Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods*. Hoboken, NJ: Wiley
Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In: Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'93), (pp 329–338).
Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2(2):159–165, doi: http://dx.doi.org/10.1147/rd.22.0159
Macdonald, C., He, B., Plachouras, V., & Ounis, I. (2005). University of Glasgow at TREC 2005: Experiments in terabyte and enterprise tracks with terrier. In: Proceedings of TREC 2005.
Margulis, E. (1992). N-poisson document modelling. In: Proceedings of the 15th International ACM SIGIR conference on research and development in information retrieval (ACM–SIGIR'92) (pp 177–189).
Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing, and information retrieval. *Journal of the Association for Computing Machinery, 7*(3), 216–244.
Mosteller, F., & Tukey, J. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.
Ounis, I., Lioma, C., Macdonald, C., & Plachouras, V. (2007). Research directions in Terrier. Novatica/UPGRADE special issue on web information Access, Ricardo Baeza-Yates et al (Eds.), Invited paper.
Ponte, J., & Croft, B. (1998). A language modeling approach in information retrieval. In: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'98). (pp 275–281).
Popper, K. (1995). *The logic of scientific discovery*. London: Routledge.
Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation, 60*(5), 503–520.
Robertson, S., & Walker, S. (1994). Some simple 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 (SIGIR'94) (pp 232–241).
Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science (JASIS)*,27, 129–146.

Robertson, S. E., van Rijsbergen, C. J., & Porter, M. (1981). Probabilistic models of indexing and searching. In: S. E. Robertson, C. J. van Rijsbergen, & P. Williams (Eds.), *Information retrieval research, chap 4* (pp. 35–56). Oxford: Butterworths.

Robertson, S. E., Walker, S., & Beaulieu, M. (1999). Okapi at TREC-7: Automatic ad hoc, filtering, VLC and interactive track. In: The seventh text REtrieval conference (TREC-7), NIST Special Publication 500:242 (pp 253–264).

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513–523.

Shannon, C. E. (1949). The mathematical theory of communication. In: C. E. Shannon, & W. Weaver (Eds.), *The mathematical theory of communication* (pp. 3–91). Urbana: The University of Illinois Press.

Singhal, A., Buckley, C., Mitra, M., & Mitra, A. (1996). Pivoted document length normalization. In: Proceedings of the 19th international ACM SIGIR conference on research and development in information retrieval (SIGIR'96), (pp 21–29).

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11–21.

Wolfowitz, J. (1942). Additive partition functions and a class of statistical hypotheses. *Annuals of Statistics*, 13, 247–279.

Wong, S., & Yao, Y. (1995). On modeling information retrieval with probabilistic inference. *ACM Transactions on Information Systems, 16*, 38–68.

Zhai, C., Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS), 22*(2), 179–214.