IRRA at TREC 2010: Index Term Weighting by Divergence From Independence Model

Bekir Taner Dinçer
Department of Statistics
Muğla University
dtaner@mu.edu.tr

İlker Kocabaş
International Computer Inst.
Ege University
ilker.kocobas@ege.edu.tr

Bahar Karaoğlan
International Computer Inst.
Ege University
bahar.karaoglan@ege.edu.tr

1 Introduction

IRRA (IR-Ra) group participated in the 2010 Web track. In this year, the major concern is to examine the effect of supplementary methods on the effectiveness of the new nonparametric index term weighting model, *divergence from independence* (DFI).

Every written text document contains words, but the words used in individual documents may differ due to many divergent (latent) factors, such as topic, author, style, etc. Some words should be intentionally used by authors, in order to compose the information contents of documents, while some words are used due to the grammatical rules. The former set of words is commonly referred to as the *keywords* or the *content bearing words*, and the later ones are referred to as the *function words* or the *stop words*. Since the function words are used due to the grammatical rules, they should appear, less or more, but in almost all documents, irrespective of (or independently from) the information contents of documents. It is, therefore, reasonable to expect the function words be distributed proportionally to the lengths of documents. On the other hand, since the content bearing words are intentionally used by the authors, their frequency distributions must be affected, and hence should differ from the frequency distributions of the function words on a collection of documents.

The content bearing words of a document can be identified by measuring the divergence from independence. According to the DFI model, if the ratio of the frequencies of two different words remains constant for all documents, the occurrences of those words in documents are said to be independent from the documents. Assume that the magnitude of the contribution of a word to the information content of a particular document is proportional to the observed frequency of the word on that document. Then, it can be said that both words contributes to the information contents of all documents, equally. However notice that an equal contribution to the information contents of all documents actually implies no contribution. Such words can only be the words that are used due to a particular reason/rule, such as grammar; because otherwise, a word could not appear in all documents having different information contents.

In analogy, the use of HTML tags in Web pages is a good basis to exemplify the independence notion. Since the function words can appear in all documents, not because of their contribution to the information contents of documents, but because of the grammatical rules, they can be thought of as the HTML tags. For instance, every Web page contains exactly two “html” tags and two “body” tags, so the ratio of the frequencies of the “html” and the “body” tags remains constant for all Web pages. According to the independence model, this suggests that the occurrence of “html” tag relative to the “body” tag does not depend on the Web pages, and that the “html” and the “body” tags contribute to the information content of each Web page, equally. It is already known that the HTML tags are used by design, independently from the information contents of the Web pages. But the point in here is that, by using the independence model, this property of HTML tags can be related to their observed frequency distributions on the Web pages, and thereby, it can be recovered without any external knowledge. This definition of independence is easy to understand, but hard to use in practice. In order to use it in practice, it is necessary to measure the degree of independence/dependence between a word and a document, individually. In fact, for each pair of word and document, the independence model can suggest the frequency expected under independence. This enable us to decide whether a particular word is independent from a given document.

If the observed frequency of a word in any given document is equal to the frequency suggested by the independence model, then the word is said to be independent from the document.

The DFI model of term weighting is closely related to the *divergence from randomness* (DFR) model introduced by Amati and Van Rijsbergen (2002). But they are different in that, in the DFR model, it is assumed that the important terms of a document are the terms whose frequencies diverge from the frequency suggested by a basic randomness model, such as *Poisson*, *Hyper-Geometric*, *Bose-Einstein* etc., whereas in the DFI model, it is assumed that the important terms are the terms whose frequencies diverge from the frequency suggested by the independence model. Harter (1975a,b) is the first researcher who introduces the paradigm used in both the DFR and the DFI. According to this paradigm, there are "speciality words" and "nonspeciality words". Speciality words are the words that occur densely in an "elite set" of documents whose informative contents are composed of the meanings represented by that words. In contrast, nonspeciality words are the words that appear in documents, *randomly*, and hence, they are the words that do not contribute to the information contents of documents. Speciality words are assumed to differ from nonspeciality words in distribution on a collection of documents. Harter claims that both the speciality and the nonspecialty words follow a Poisson distribution, but with different means, λ_1 and λ_2, respectively, where $\lambda_1 > \lambda_2$.

In the DFR model, it is assumed that a speciality word is the word whose frequency distribution diverges from the basic randomness model, while a nonspeciality word is the word whose frequency distribution follows the basic randomness model. The basic randomness models that are considered in the work of Amati and Van Rijsbergen (2002) include, but not limited to the Poisson distribution: many probability density functions are examined, such as Hyper-Geometric, Bose-Einstein, etc. On the other hand, in the DFI model, it is assumed that a speciality word is the word whose within document frequency diverges from the frequency suggested by the independence model, while a nonspeciality word is the word whose within document frequency follows the frequency suggested by the independence model.

In essence, the DFI model of term weighting replaces the notion of randomness with the notion of independence. This means that the DFI model is the nonparametric counterpart of the DFR model. In every index term weighting model, a given term is weighted by means of a statistic (the weighting function), which is derived from the data at hand, i.e., the document collection. A nonparametric statistic, or rather a nonparametric method/procedure, can be defined by what it is not. Traditional statistical (hypothesis testing) methods are based on parametric assumptions such that the population of data can be generated by some well-known family of distributions, such as *normal*, *exponential*, *Poisson*, and so on. Each of these distributions has one or more parameters (e.g. the normal distribution has mean μ and variance σ^2), at least one of which is presumed unknown and must be inferred from a sample of data drawn from the population. In the literature of statistics, Wolfowitz (1942) first coined the term nonparametric: "We shall refer to this situation [where a distribution is completely determined by the knowledge of its finite parameter set] as the parametric case, and denote the opposite case, where the functional forms of the distributions are unknown as the nonparametric case". In addition, Bradley (1968) mentions that "The terms nonparametric and distribution-free are not synonymous ... Popular usage, however, has equated the terms ... Roughly speaking, a nonparametric test is one which makes no hypothesis about the value of a parameter in a statistical density function, whereas a distribution-free test is one which makes no assumptions about the precise form of the sampled population.". It can be confusing to understand what is implied by the word "nonparametric". However as a rule of thumb, it is enough to know that the adjective "nonparametric" qualifies not only the statistic, but also the (statistical/inductive) inference process in which the statistic is used. In here, the "inference" corresponds to decide whether or not the observed frequency of a term in a document diverges from independence/randomness.

The DFR model of term weighting necessitates a hypothesis about the functional form of the frequency distributions of terms on the document collection in use, in order to define what is random. In the context of statistical inference, this suggests that the DFR model is of the parametric type. In contrast, in the DFI model, the amount of divergence from independence is measured based on the Pearson's Chi-Square statistic. The fact that the Pearson's Chi-Square statistic is of the nonparametric type (Conover, 1999) suggests that the DFI model of term weighting is of the nonparametric type. On the other hand, it should also be noted that the DFR model of term weighting is also qualified as a nonparametric model in the original work of Amati and Van Rijsbergen (2002), where the term "nonparametric" means no-parameter or parameter-free; and "parameter-free models are meant to be models that do not contain

parameters that are learned from relevance feedback"[1]. In this respect, it can be said that the DFI model is nonparametric in both senses.

The TERRIER retrieval platform (Ounis et al., 2007) is used to index and search the ClueWeb09-T09B[2] data set ("Category B" data set), a subset of about 50 million Web pages in English. During indexing and searching, terms are stemmed but not stopped.

2 DFI Formula

The DFI formula used in IRRA runs is given by

$$DFI_{ij} \quad = \quad log_2\left(\frac{tf_{ij} - e_{ij}}{\sqrt{e_{ij}}} + 1\right), \tag{1}$$

where tf_{ij} is the frequency of term i in document j, and e_{ij} is the expected frequency of term i in document j. Under independence, expected frequency, e_{ij} is given by

$$e_{ij} = TF_i \frac{D_j}{N}$$

where TF_i is the collection frequency of term i, D_j is the length of document j, and N is the collection size in terms of words. Roughly speaking, under independence, the collection frequency of term t_i (TF_i) should be distributed on documents, proportionally to the proportion of the length of each document (D_j/N). That is, $\sum_j D_j = N$ and $\sum_j TF_j/N = 1$, so $\sum_j e_{ij} = TF_i$, meaning that $\sum_j (tf_{ij} - e_{ij}) = 0$.

3 Run Descriptions

IRRA runs use TFxIDF weighting scheme where the DFI formula in Equation 1 is used in place of TF component and IDF is the BM25 IDF (Robertson et al., 1981).

irra10b : This is the base run of the system developed for high recall and average precision. It uses spam filtering and phrase searching based on n-grams.

irra10hp : This is the run of the system developed for high precision. The goal of this run is to maximize precision on topics that the system performs well.

irra10rob : This is the run of the system developed for robustness in retrieval performance, i.e., it is the system that is expected to return predictable results to every topic.

Acknowledgement

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

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

J. V. Bradley. *Distribution Free Statistical Tests*. Prentice Hall, Englewood Cliffs, NJ, 1968.

W. J. Conover. *Practical Nonparametric Statistics*. Wiley, New York, 3^{rd} edition, 1999.

[1] Personal contact to Amanti and Van Rijsbergen.

[2] http://boston.lti.cs.cmu.edu/Data/clueweb09/

S.P. Harter. 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*, 26: 197–216, 1975a.

S.P. Harter. A probabilistic approach to automatic keyword indexing. Part II: An algorithm for probabilistic indexing. *Journal of the American Society for Information Science*, 26:280–289, 1975b.

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

I. Ounis, C. Lioma, C. Macdonald, and V. Plachouras. Research directions in Terrier. *Novatica/UPGRADE Special Issue on Web Information Access, Ricardo Baeza-Yates et al. (Eds), Invited Paper*, 2007.

S. Robertson and S. Walker. Some simple approximations to the 2-Poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, 1994.

S. E. Robertson, C. J. Van Rijsbergen, and M. Porter. Probabilistic models of indexing and searching. In S. E. Robertson, C. J. van Rijsbergen, and P. Williams, editors, *Information Retrieval Research*, chapter 4, pages 35–56. Butterworths, Oxford, UK, 1981.

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