IRRA at TREC 2012: Divergence From Independence (DFI)

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

1 Introduction

IRRA (IR-Ra) group participated in the 2012 Web track, with a system implementing a non-parametric term weighting method based on measuring the *divergence from independence* (DFI). This is the third year of participation for IRRA group, following the participations in TREC 2009 and 2010 Web tracks. In this year, the aim is to evaluate a new DFI-based term weighting model developed on the basis of Shannon's information theory (Shannon, 1949), along with the evaluation of a heuristic approach that is expected to provide early precision when used together with DFI term weighting.

The TERRIER retrieval platform version 3.0 (Ounis et al., 2007) is used to index and search the ClueWeb09-T09B[1] data set ("Category B" data set), a subset of about 50 million Web pages in English. During indexing and searching, terms are stemmed (Porter's stemmer as implemented in TERRIER) but not stopped. The result sets are filtered using the fusion of two spam-page lists provided by Cormack et al. (2010) for ClueWeb09 document collection.

2 Measures of Divergence From Independence

There are three basic measures of divergence from independence, each of which arises from different bases:

$$\mathrm{sz}_{ij} = \frac{tf_{ij} - e_{ij}}{e_{ij}} \tag{1}$$

based on *saturated model of independence*,

$$\mathrm{z}^2_{ij} = \frac{(tf_{ij} - e_{ij})^2}{e_{ij}} \tag{2}$$

based on *normalized chi-squared distance* from independence, and

$$\mathrm{z}_{ij} = \frac{tf_{ij} - e_{ij}}{\sqrt{e_{ij}}} \tag{3}$$

based on *standardization*, for $(tf_{ij} - e_{ij}) > 0$ and zero elsewhere. In here, tf_{ij} is the frequency of term i in document j ($i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$, where r and c are respectively the number of unique terms and the number of documents in the document collection given). The pivotal component of the formulae given above is e_{ij}, which gives the *expected frequency* of term i in document j under independence:

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

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

where TF_i is the collection frequency of term i ($TF_i = \sum_j tf_{ij}$), D_j is the length of document j ($D_j = \sum_i tf_{ij}$), and N is the size of the collection ($N = \sum_i \sum_j tf_{ij}$).

Although these three forms of the measure of DFI are nearly indistinguishable from each other (note especially that the later two are in fact arithmetically equal to each other), they have different retrieval performances in practice due to the differences in inherent behavior in weighting terms, as shown in the following section.

In brief, the DFI measure based on standardization in (3) is good at tasks that require high recall and high precision, especially against short queries composed of a few words as in the case of Internet searches. For the tasks that require high recall against long queries, the DFI measure based on saturated model of independence in (1) is usually the best one among all. The DFI measure in (2), which is based on the normalized chi-squared distance from independence, can be used for tasks that require high precision, against both short and long queries.

3 The Heuristic Approach for Early Precision

In this year, IRRA runs employ a heuristic approach primarily for early precision, and, to some extent, for filtering documents with no content of any use to the users, i.e., in one sense, spam-page or junk-page filtering. Note that in here, this did not affect the operational settings of the IRRA runs submitted to TREC 2012 Web track: the result sets yielded from IRRA runs with DFI term weighting are filtered using the fusion of two spam-page lists provided by Cormack et al. (2010). Experiments on past Web track topics (TREC 2009, 2010, and 2011) shown that spam-page filtering improves the retrieval performance of DFI-based term weighting, no matter whether or not the heuristic approach is applied.

The rationale utilized in here for providing early precision can best be explained as follows. Clinchant and Gaussier (2010) define four *form conditions* of being an ideal retrieval function, by extending the formalism introduced by Fang et al. (2004). This formalism postulates that "h [retrieval function] should be an increasing, concave function of the term frequency, the concavity ensuring that the increase in the retrieval score will be smaller for larger term frequencies". This means that, if this hypothesis of being an ideal retrieval function holds true, given a pair of retrieval functions, the better is one which yields a sharper increase in the term weight value than the other for terms with low frequency.

They compare the behavior of the LGD model of term weighting introduced in their original work (Clinchant and Gaussier, 2010) to the behavior of the DFR model introduced by Amati and van Rijsbergen (2002), and show that LGD model has, in theory, a sharper increase than DFR model for low frequency terms. The LGD model of document scoring is actually based on the DFR framework, and defined, in principle, by Clinchant and Gaussier (2010) as

$$LGD(q, d_j) \quad = \quad \sum_{t_i \in q \cap d_j} qtf_i \times -\log_2 \left(\frac{\lambda_i}{tf_{ij} + \lambda_i} \right).$$

To show that the LGD model has a better term weighting behavior than the DFR model, they perform an analysis defined as "we used a value of 0.005 for λ and computed the term weight obtained for term frequencies varying from 0 to 15". The same analysis is repeated here for the comparison of DFI with LGD, where $e_{ij} = \lambda = 0.005$. The term weights that are yielded from the basic measures of DFI and LGD for varying term frequencies (tf_{ij}) from 0 to 15 are listed in Table 1, and the corresponding plot is shown in Figure 1. As seen, the logarithm of the DFI measure based on standardization ("LogZ": $\log_2(z_{ij} + 1)$) has a similar behavior to that of LGD in general. Whereas, in contrast, the DFI measure based on the saturated model of independence ("LogSZ": $\log_2(sz_{ij} + 1)$) and the measure based on the normalized chi-square distance ("LogZ2": $\log_2(z_{ij}^2 + 1)$) both have a sharper increase in the term weight value than those two functions for the low term frequencies.

This suggests that DFI measures, as a retrieval function, is at least as well-behaved in weighting low term frequencies as LGD. If the hypothesis of being an ideal retrieval function holds true, it would, then, be reasonable to expect that DFI will show a retrieval performance noticeably higher than that of LGD. However, the experiments performed on past TREC Web track topics failed to verify this expectation: there is a difference between DFI and LGD in retrieval performance but not that much the discrepancy the analysis suggested in magnitude. At first glance, it is possible to interpret the experimental results

tf_{ij}	z_{ij}	sz_{ij}	LGD	LogZ	LogSZ	LogZ2
0	-0	-1	0.00	0.00	0.00	0.00
1	14	199	5.30	3.81	7.64	7.63
2	28	399	5.99	4.82	8.64	9.64
3	42	599	6.40	5.40	9.23	10.81
4	56	799	6.69	5.82	9.64	11.64
5	71	999	6.91	6.14	9.96	12.28
6	85	1199	7.09	6.41	10.23	12.81
7	99	1399	7.24	6.63	10.45	13.26
8	113	1599	7.38	6.82	10.64	13.64
9	127	1799	7.50	6.99	10.81	13.98
10	141	1999	7.60	7.14	10.97	14.29
11	155	2199	7.70	7.28	11.10	14.56
12	170	2399	7.78	7.41	11.23	14.81
13	184	2599	7.86	7.52	11.34	15.04
14	198	2799	7.94	7.63	11.45	15.26
15	212	2999	8.01	7.73	11.55	15.46

Table 1: Term weights yielded from log-logistic weighting function (LGD) and the DFI weighting functions at $\lambda = e_{ij} = 0.005$ for varying term frequencies (tf_{ij}) from 0 to 15. $LGD = \log_e(\lambda/(\lambda + tf_{ij}))$; $LogZ = \log_2(z_{ij} + 1)$, $LogSZ = \log_2(sz_{ij} + 1)$, and $LogZ2 = \log_2(z_{ij}^2 + 1)$ for $(tf_{ij} - e_{ij}) \geq 0$ and zero elsewhere.

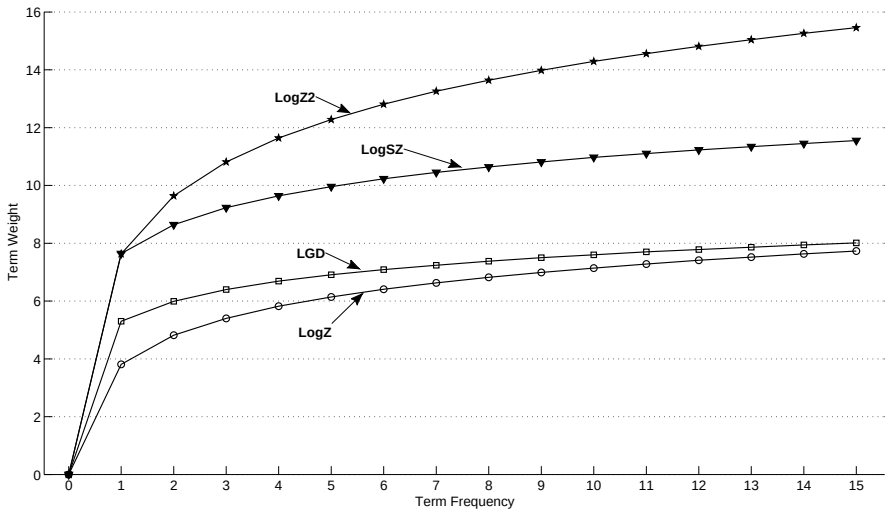

Figure 1: Plot of the term weights yielded from the basic DFI weighting functions and LGD for varying term frequencies (tf_{ij}) from 0 to 15.

as an evidence that counts against the hypothesis of being an ideal retrieval function, but a "bug" in the code of the TERRIER's implementation of evaluation revealed that this is not the case. In those experiments, TERRIER is used only to get the result sets against a given set of topics, and the evaluation of the result sets is made by using a different software. The evaluation results that seem to count against the hypothesis are yielded from that evaluation software. In one of the experiments, the TERRIER's evaluation is used just to make a quick look at the obtained results, and the MAP score the TERRIER calculated for DFI was too high to be true (a MAP score of about 0.26, while the scores calculated using the evaluation software are always below 0.2). It is figured out that TERRIER accidently considers the junk documents (i.e., a minus relevance value in the qrels files of Web track) as if they were relevant. This bug sent a signal that the hypothesis of being an ideal retrieval function may still hold true, and most importantly that the result sets from IRRA system are filled up by junk documents especially at the top ranks: thus, it leaded to the heuristic applied to IRRA 2012 system for early precision.

The heuristic approach or the tweak the IRRA system employed, which simply utilizes the structural regularities of junk documents, is basically a supplementary function that modifies term weights yielded from DFI weighting models so as to suppress junk documents and promote relevant documents towards

higher ranks, and given by

$$\Lambda_{ij} \quad = \quad \alpha_{ij}^{3/4} \times \beta_{ij}^{1/4}$$

where

$$\alpha_{ij} = \frac{D_j - tf_{ij}}{D_j} \qquad \text{and} \qquad \beta_{ij} = \frac{2}{3} \times \frac{(tf_{ij}+1)}{tf_{ij}}.$$

The first component of the function, α_{ij} promotes those documents in which the query term i does not fill up the whole document in frequency (i.e., fake frequency boosting), and hence it favors terms with low frequency relative to the document length. The second component, β_{ij} favors, in contrast, the terms that have high frequency in magnitude irrespective of the length of documents. The multiplicative constant, 2/3, is determined by try-and-fail experiments. Thus, the function as a whole favors those terms whose frequencies are not so high to be considered fake while high enough in its own right to be considered content bearing.

At the time of integration, the influence of the function to the resulting document scores is adjusted, for now, roughly by taking arithmetic roots of the components: $\alpha_{ij}^{3/4}$ and $\beta_{ij}^{1/4}$. This setting seems to work well in providing early precision against the past three sets of TREC Web topics (1-150) on ClueWeb09 category B data set, without harming significantly the level of recall that could be obtained when it is not applied. It is also observed that the function can be used to provide early precision on the collections of well-structured and coherent documents, such as TIPSTER disks 1 & 2 and TREC disks 4 & 5. This is the first attempt to form the function that modifies term weights from DFI to provide early precision, and it is still in progress.

4 An Information-Theoretic Weighting Model based on DFI

The new DFI-based information theoretic term weighting model the IRRA system employed in this year is given by

$$\Delta(I_{ij}) = \left[(tf_{ij}+1) \times \log_2 \left(\frac{(tf_{ij}+1)}{\sqrt{e_{ij}^+}} \right) \right] - \left[tf_{ij} \times \log_2 \left(\frac{tf_{ij}}{\sqrt{e_{ij}}} \right) \right],$$

where

$$e_{ij}^+ \quad = \quad \frac{(TF_i+1) \times (D_j+1)}{N+1}.$$

It simply calculates the amount of increase in total information we would get by observing term i in document j one more time (i.e., $tf_{ij}+1$), given that it occurs tf_{ij} times in document j.

This term weighting model is directly stemmed from the hypothesis of being an ideal retrieval function, assuming that low frequency corresponds to high amount of information, and vice versa. If the assumption holds true, the amount of increase in total information will be higher for smaller term frequencies, and vice versa, and hence the model will assign weights to terms in a way that the hypothesis states. This weighting function can be thought of as the information-theoretic realization of the hypothesis of being an ideal retrieval function itself based on DFI. In the model, DFI is used only in measuring the amount of information yielded from each occurrence of a term in a document; the realization is accomplished by means of the substraction of the two quantities of total information enclosed by brackets. Thus, this realization is independent of whether DFI itself, as a retrieval function, is in accordance with the hypothesis: one may use an other measure of information, instead of the one based on DFI, and put the hypothesis into practice by means of that measure, i.e., by replacing e_{ij}/tf_{ij} in $-log_2(\cdot)$ by an other measure of the probability of occurrence of a term in a document.

5 Run Descriptions

irra12a : This is the run of the system developed for high recall, which utilizes the DFI formula given in (2), i.e., chi-squared distance form independence. Among all, this run is most suitable for tasks that require high recall.

irra12b : This is the run for early precision, which employs the DFI formula given in (3). This run is most suitable for tasks that require high precision, with a slight decrease in recall compared to irra12a.

irra12c : This is the run of the system that uses a term weighting model developed on the basis of the Shannon's information theory. It employs the DFI formula given in (1), and is based on measuring the amount of contribution of observing a term one more time in a document to the total information we will get. This run is one which provides early precision but with the cost of a moderate loss in recall compared to the other runs.

The term weighting formula the system run "irra12a" employed is given by

$$w_{ij} \quad = \quad \log_2 \left(\frac{(tf_{ij} - e_{ij})^2}{e_{ij}} + 1 \right)$$

and the term weighting formula for "irra12b" is given by

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

and for "irra12c" it is given by

$$w_{ij} \quad = \quad \Delta(I_{ij}).$$

Given a query with k terms, the score of document j is given by

$$s_j = \sum_{i=1}^{k} qtf_i \times w_{ij} \times \Lambda_{ij},$$

where qtf_i is the frequency of term i in the query.

Finally, the spam-page filtering is applied to the result sets yielded from IRRA runs as given by

$$s_j^p = (0.75 \times s_j) + (0.25 \times (ss \times s_j)),$$

where ss is the percent of spammness associated with document j, i.e., the score in the original fusion file divided by 100. The resultant s_j^p scores are then used to re-rank the result sets. Note that spam filtering does not remove documents from result sets; it just promotes them towards top of the ranking or suppress down to the ranking so that it preserves the level of recall achieved via the original scoring function. Approaches based on removing documents from result sets, which is suggested in Cormack et al. (2010), has not been attempted yet.

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.

B. C. Arnold. *Pareto Distributions*. International Cooperative Publishing House, Fairland, Maryland, 1983.

A. Clinchant and E. Gaussier. 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)*, pages 234–241, 2010.

Gordon V. Cormack, Mark D. Smucker, and Charles L. A. Clarke. Efficient and Effective Spam Filtering and Re-ranking for Large Web Datasets. April 2010. URL `http://arxiv.org/abs/1004.5168`. Retrieved from http://arxiv.org/abs/1004.5168v1.

H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information retrieval (ACM SIGIR04)*, pages 49–56, 2004.

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.

C. E. Shannon. The mathematical theory of communication. In Claude E. Shannon and Warren Weaver, editors, *The Mathematical Theory of Communication*, pages 3–91. The University of Illinois Press, 1949.