IRRA at TREC 2009: Index Term Weighting based on 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 2009 Web track (both adhoc task and diversity task) and the Million Query track. In this year, the major concern is to examine the effectiveness of a novel, nonparametric index term weighting model, *divergence from independence* (DFI).

The notion of independence, which is the notion behind the well-known statistical exploratory data analysis technique called the *correspondence analysis* (Greenacre, 1984; Jambu, 1991), can be adapted to the index term weighting problem. In this respect, it can be thought of as a qualitative description of the *importance* of terms for documents, in which they appear, importance in the sense of contribution to the information contents of documents relative to other terms. According to the independence notion, if the ratios of the frequencies of two different terms are the same across documents, they are independent from documents. For example, each Web page contains a pair of "html" and a pair of "body" tags, so that the ratio of frequencies of these tags is the same across all Web pages, indicating that the "html" and "body" tags are independent from Web pages. They are used by design, irrespective of the information contents of Web pages. On the other hand, some tags, such as "image", "table", which are also independent from Web pages, may occur less or more in some pages than the expected frequencies suggested by the independence model; so, their associated frequency ratios may not be the same for all Web pages. However, it is reasonable to expect that, if the pages are not about the tags' usage, such as a "HTML Handbook", frequencies of those tags should not be significantly different from their expected frequencies: they should be close to the expectation, i.e., in a parametric point of view, their observed frequencies on individual documents should be attributed to chance fluctuation. Although this tag example is helpful in exemplifying the use of independence notion, it is obvious that the tags are artificial, and so, governed by some rules completely different from the rules of a spoken language. Nonetheless, some words, like the ones in a common "stopwords list", appear in documents, not because of their contribution to the information contents of documents, but because of the grammatical rules. On this account, such words can be modeled as if they were tags, because they are independent from documents in the same manner. Their observed frequencies in individual documents is expected to fluctuate around their frequencies expected under independence, as in the case of tags. Content bearing words are, therefore, the words whose frequencies highly diverge from the frequencies expected under independence. The results of the TREC experiments about IRRA runs show that the independence notion promises a natural basis for quantifying the categorical relationships between the terms and the documents.

The TERRIER retrieval platform (Ounis et al., 2007) is used to index and search the ClueWeb09-T09B[1] data set, a subset of about 50 million Web pages in English (TREC 2009 "Category B" data set). During indexing and searching, terms are stemmed and a particular set of stop words[2] are eliminated.

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

[2] The set of stop words bundled in TERRIER + all numbers + number-word mixtures.

2 Data Organization

In statistics, the raw input to multivariate data analysis is usually a $r \times c$ (r rows by c columns) rectangular array of real numbers called a *data matrix* given in Figure 1.

$$\mathbf{X} = \begin{array}{c|ccccc|c} & \multicolumn{5}{c|}{\text{Variables (Documents)}} & \\ \text{Objects (Terms)} & 1 & 2 & 3 & \dots & c & \text{Totals} \\ \hline 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1c} & x_{1\cdot} \\ 2 & x_{21} & x_{22} & x_{23} & \dots & x_{2c} & x_{2\cdot} \\ 3 & x_{31} & x_{32} & x_{33} & \dots & x_{3c} & x_{3\cdot} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r & x_{r1} & x_{r2} & x_{r3} & \dots & x_{rc} & x_{r\cdot} \\ \hline \text{Totals} & x_{.1} & x_{.2} & x_{.3} & \dots & x_{.c} & x_{..} \end{array}$$

Figure 1: The data matrix for a set of documents.

The data matrix represents the observations made on r objects each of which is characterized with respect to c variables. The observed values of those variables may represent the measurement of a quantity or a numerical code for a classification scheme. The *objects* may be an individual or a unit, and hence the *variables* will be the attribute, characteristics, response or item of these objects. This multidimensional data are represented by $r \times c$ matrix $\mathbf{X}$. The elements of $\mathbf{X}$ represent the observed values and denoted by x_{ij} where $i = 1, 2, \dots, r$ and $j = 1, 2, \dots, c$; the marginal total of row i is denoted by $x_{i.}$ and is given by $x_{i.} = \sum_{j=1}^{c} x_{ij}$; similarly $x_{.j}$ denotes the marginal total of column j and is given by $x_{.j} = \sum_{i=1}^{r} x_{ij}$; finally, $x_{..}$ denotes the grand total and is given by $x_{..} = \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij}$. Any given set of documents can be formed into a $r \times c$ data matrix whose columns represent documents, rows represent terms, and cells contain the frequency of each term in the associated documents.

3 Divergence From Independence – DFI

Given a *term-by-document* data matrix $\mathbf{X}$, the notion of *independence* can simply be explained by *odds ratios*. The ratio $x_{ij}/x_{\ell j}$ measures the odds of being in row i relative to row ℓ given column j. The ratio $x_{ik}/x_{\ell k}$ measures the odds of being in row i relative to row ℓ given column k. The odds ratio is the ratio of the two sets of odds and is given by

$$\frac{x_{ij}/x_{\ell j}}{x_{ik}/x_{\ell k}}$$

The odds ratio is necessarily 1, if the independence assumption holds. Under independence, the odds of being in row i relative to row ℓ do not depend on the column.

DFI is closely related to the *divergence from randomness* (DFR) model introduced by Amati and Van Rijsbergen (2002), but they are different in that, in DFR, it is assumed that the important terms of a given document are those words whose frequencies diverge from the frequencies suggested by a basic randomness model, such as *Poisson*, *Hyper-Geometric*, *Bose-Einstein* etc, whereas in DFI, it is assumed that the important terms of a given document are those words whose frequencies diverge from the frequencies suggested by the independence model.

Harter (1975a,b) is the first researcher who uses the Poisson distribution for weighting index terms. In essence, the notion behind DFR, as well as DFI, is the notion behind the Harter's approach. In the Harter's approach, words are classified into two groups, namely the "speciality words" and the "nonspeciality words". Speciality words, or the content bearing words, are the words that occur densely in an "elite set" of documents, whose informative contents are actually related to the meanings of that words, whereas nonspeciality words are those words whose frequencies distribute on documents, *randomly*. In this point of view, speciality words should differ from the nonspeciality words in distribution on a collection of documents. Harter argues that both the speciality and the nonspecialty words follows a Poisson distribution, but with different means, λ_1 and λ_2, respectively, where $\lambda_1 > \lambda_2$. According to the DFR model, speciality words are those words whose within document frequencies diverge from the frequencies suggested by the basic randomness model, whereas nonspeciality words are the words that follow the basic randomness model. According to the DFI model, speciality words are those words whose within document frequencies depend on the documents, whereas nonspeciality words are those words whose within document frequencies are independent from the documents. The difference is that, DFI replaces the notion of randomness with the notion of independence. In DFI, amount of divergence from independence is measured as *chi-square* distance. The fact that the Pearson's chi-square statistic is of the *nonparametric* type (Conover, 1999) suggests that the proposed index term weighting model is the nonparametric counterpart of the model introduced by Harter. The major advantage of a nonparametric index term weighting model is that, it does not require a hypothesis about the functional form of the term frequency distributions on document population, such as Poisson distribution: that is, to decide whether a particular term is independent from a given document, the proposed model does not require any external reference[3].

3.1 Weighting Models based on DFI

DFI quantifies the categorical relationships between the terms and the documents, and hence, it basically corresponds to the TF component of the well-known TF×IDF weighting scheme (Salton and Buckley, 1988), where TF stands for the *term frequency* and IDF stands for the *inverse document frequency*. In contrast to TF, IDF is a collection dependent factor, which identifies the terms that concentrates in a few documents of the collection. Salton and Buckley (1988) state that "the *term discrimination* considerations suggest that ... the best terms should have high term frequencies but low overall collection frequencies". According to the DFI notion, this statement can be such that "... the best terms should have high DFI scores but low overall collection frequencies". A trivial weighting model based on the TF×IDF scheme can therefore be given as

$$w_{ij} = DFI_{ij} \times IDF_i$$

The amount of divergence from independence, DFI_{ij} of a given term t_i $(i = 1, 2, \ldots, r)$ in a particular document d_j $(j = 1, 2, \ldots, c)$ is measured as the difference of the frequency of the term in that document (x_{ij}) from the expected frequency (e_{ij}) suggested by the independence model, given by

$$e_{ij} = x_{i.}\frac{x_{.j}}{x_{..}}$$

where $x_{i.}$ corresponds to the total frequency of t_i in collection (i.e. $x_{i.} = \sum_{j=1}^{c} x_{ij}$), and $x_{.j}$ corresponds to the length of document d_j (i.e., $x_{.j} = \sum_{i=1}^{r} x_{ij}$). The intuition behind expected frequencies is as follows. Under independence, marginal (total) frequencies of terms, $x_{i.}$'s should be distributed on documents, proportionally to the length of documents $(x_{.j}/x_{..})$, i.e., $\sum_j x_{.j} = x..$ and so $\sum_j x_{.j}/x_{..} = 1$; thus $\sum_j e_{ij} = x_{i.}$. For any term t_i and document d_j, the amount of divergence from independence, DFI_{ij} can be measured as chi-square distance, given by

$$DFI_{ij} = \frac{(x_{ij} - e_{ij})^2}{e_{ij}}$$

Notice that, in this formulation, it is not possible to distinguish whether the frequency of a particular term is below or above the expected frequency, i.e., it is not possible to determine the direction of interaction between terms and documents. On the other hand, it is a fact that terms whose frequencies are above the expected frequencies are the terms that have a categorical relationship with the documents, in a positive sense, whereas terms whose frequencies are below the expected frequencies are the terms that have also a categorical relationship with the associated documents, but in a negative sense. A term weighting method is supposed to identify documents that have a positive categorical relationships with the given query terms; thus, it is necessary to make the direction of interaction explicit, as given by

$$DFI_{ij} \quad = \quad \frac{x_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$

[3] DFR is also qualified as a nonparametric model in the work of Amanti and van Rijsbergen. But they mentioned that, in IR, the term "nonparametric" has a different meaning than in statistics (personal contact). In IR, nonparametric means to "parameter-free" models, and parameter-free models are meant to be models that do not contain parameters that are learned from relevance feedback. On this account, DFI model is a nonparametric model in both sense.

In practice, for a given term t_i, its associated weight w_{ij} with document d_j should be taken as 0, when $DFI_{ij} \leq 0$. In consequence, given a query q_k $(k = 1, 2, \ldots)$ with p terms, score (s_{kj}) of a given document d_j can be calculated as

$$s(q_k, d_j) = \sum_{i}^{p} x_{ik} \times w_{ij}$$

where x_{ik} is the frequency of term t_i in query q_k.

3.2 DFI Formulae used in IRRA runs

The basic DFI formula used in IRRA runs is given by

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

Use of a log-transformation is attractive for two reasons. First, it provides a multiplicative model for document scoring (i.e., $s = log(w_1) + log(w_2) = log(w_1 \times w_2)$), instead of an additive model (i.e., $s = w_1 + w_2$). Given a query q_k with p terms, in the additive model, scores of two different documents can be the same, when the sets of term weights associated with individual documents sum up to the same total, while in contrast, in the multiplicative model, different sets of weights associated with different documents might result in different scores. In other words, given a query q_k with p terms, for every document d_j, an additive model produces the same document score with different sets of term weights, satisfying the equation given by $w_{1j} + w_{2j} + \ldots + w_{pj} = C$, while a multiplicative model produces the same document score with different sets of term weights, satisfying the equation given by $w_{1j} \times w_{2j} \times \cdots \times w_{pj} = C$. The multiplicative model can therefore provide more discrimination power than the additive model. Second, it is a fact that it is possible to obtain a multiplicative model of document scoring by simply multiplying the term weights (i.e., $s(q_k, d_j) = \prod x_{ik} \times w_{ij}$), but in this time, the benefit of using a *power transformation* is lost. As in the case of DFI, if the symmetry of the main body of the data around the center is desired but skewness in the tails is relatively unimportant, then a log-transformation should be used.

Notice that the DFI formula given in Equation 1 also applies a power transformation to the expected frequencies ($\sqrt{e_{ij}}$) in the denominator, namely a square-root transformation which provides the symmetry in the tails of the distribution, i.e., providing the symmetry in the tails of the expected frequency distributions of terms in individual documents. In consequence, the following DFI formulae can also be considered:

$$\begin{aligned} DFI_{ij} &= \frac{x_{ij} - e_{ij}}{e_{ij}}, \\ DFI_{ij} &= log_2\left(\frac{x_{ij} - e_{ij}}{e_{ij}}\right), \end{aligned} \tag{2}$$

However it should be noted that these formulae do not measure the divergence from independence in the true sense; hence, they actually suffer from the lack of a well-defined theoretical basis, though the results of the experiments performed on the past TREC collections shown that DFI formula given in Equation 2 is the superior on long queries (i.e., Title + Description + Narrower).

3.3 The IDF component

IRRA runs, which were submitted to all TREC tracks, employ an IDF formulation derived based on the DFI. However, for the IDF component, the following standard IDF formulations could also be used:

$$\begin{aligned} IDF_1 &= log_2(c/c_i) \\ IDF_2 &= log_2((c - c_i + 0.5)/(c_i + 0.5)), \quad \text{(BM25 IDF)} \\ IDF_3 &= log_2((c + 1)/(c_i + 0.5)), \quad \text{(DFR IDF)} \end{aligned}$$

where c is the total number of documents in the collection and c_i is the number of documents that contain term t_i. IDF_1 is attributed to Jones (1972); IDF_2 is attributed to Robertson et al. (1981), and Robertson and Walker (1994). Empirical results suggest that the optimal weighting model is the model that includes IDF_2 coupled with the DFI formula given in the Equation 1. Nevertheless, other IDF formulae could also be used. Empirical results revealed that the differences in contribution to the overall retrieval performance between the IDF formulae are negligible, compared to the differences between the DFI formulae.

4 Run Descriptions

IRRA runs are of the *content-only* type: none of the information in Web pages, like Meta information, Link information, etc., was distinguished and utilized. In addition, during searching, none of the common supplementary methods, such as parsing rules, phrase processing, query processing, query expansion, relevance feedback, thesaurus, WordNets is used. IRRA runs are pure, out-of-the-box DFI runs.

irra1a : TF component is a derivation of the DFI formula given in Equation 1 and IDF component is also a DFI based formula developed as an alternative to the existing IDF formulae.

irra2a : DFI formula given in Equation 2 and IDF used in "irra1a".

irra3a : DFI formula given in Equation 1 and IDF used in "irra1a".

These are the runs that were submitted to the *Web track adhoc task*. The runs submitted to the *diversity task* are based on the same strategies used in adhoc task, with a difference. For diversity task, the result sets are filtered such that they consist of at most two Web pages from the same host (URL), and labeled as "irra1d", "irra2d" and "irra3d", respectively. For MQ track, both strategies are used, and the runs are labeled as "irra1mqa", "irra1mqd", "irra2mqa", "irra2mqd", and "irra3mqd', where "a" and "d" indicate that the associated run employs adhoc strategy and diversity strategy, respectively.

5 Results

In this year, the major concern is to verify that the DFI is a valid basis for weighting index terms.

5.1 Adhoc Task

The estimated performance scores of adhoc runs over 49 topics[4] are given in the Table 1 and the Table 2.

	statAP Estimates				
	MAP	nDCG	R-Prec	Prec30	SD(AP)
irra1a	0.1530	0.2924	0.2399	0.3316	0.006855
irra2a	0.1100	0.2290	0.1798	0.3182	0.007478
irra3a	0.1557	0.2954	0.2467	0.3321	0.006851
best	0.4392	0.6215			
median	0.1570	0.3016			
worst	0.0042	0.0202			

Table 1: Estimated performance scores of IRRA adhoc runs based on statAP.

	MTC Estimates							
run	eMAP	eR-Prec	eP5	eP10	eP15	eP20	eP30	eP100
irra1a	0.0375	0.0968	0.2167	0.2780	0.3017	0.3122	0.3401	0.3366
irra2a	0.0274	0.0860	0.2811	0.2756	0.2865	0.2793	0.2846	0.2865
irra3a	0.0379	0.0971	0.2117	0.2810	0.3027	0.3197	0.3399	0.3420

Table 2: Estimated performance scores of IRRA adhoc runs based on MTC.

[4] For topic 20, none of the participating runs had returned any relevant document.

It appears that the runs “irra1a” and “irra3a” are compatible in performance. Although they show relatively higher performance than “irra2a”, both “irra1a” and “irra3a” seems to be the median runs on the average. However, the observed *statAP* scores of “irra1a” on topic 12 and 26 are the highest scores that could be achieved by the submitted runs. Similarly, “irra2a” has the best scores at topic 34 and 45; “irra3a” has the best score at topic 12. Observed statAP scores of “irra1a” are above the median scores of 24 topics; “irra2a” is above median on 14 topics; “irra3a” is above median on 26 topics. On the other hand, “irra2a” is also said to be compatible in performance with the other IRRA runs, based on the *Prec30* measure. MTC results, which is given in Table 2, also lead to the similar conclusions. Based on the MTC statistics, relatively low confidence value between “irra1a” and “irra3a” ($P(\Delta MAP < 0) = 0.7670$) indicates high uncertainty in the final ranking of these two runs. This also suggests that they are compatible in performance. However, it should be noted that, based on *eP5* (expected precision at 5 documents), “irra2a” is the best run, and that its performance very stable across different measurement depths. Its eP5, eP10, eP15, eP20, eP30 and eP100 scores are approximately equal, while the performance scores of others tend to increase by measurement depth.

It is a fact that the average performance is a valid criterion to decide that one system is better than other. But it is also a fact that making decisions solely based on this criterion may well be misleading. High average performance is the necessary, but not the sufficient condition for being an effective IR system. A univariate data analysis employs univariate statistics, and a univariate statistic, such as raw averaging, may oversimplify, even hide the performance differences between retrieval strategies. In contrast to univariate data analysis, *multivariate data analysis* is concerned with the joint nature of measurements, and it refers to analyze the data in a manner that takes into account the relationships among measurements. It is, therefore, better if a univariate analysis is supported by the multivariate data analysis techniques, before making the final decision about the effectiveness of the retrieval strategies under consideration.

To visually inspect the mutual performance relationships among different retrieval strategies across a given set of topics, one can use the *principal components analysis* (PCA), a statistical, multivariate, exploratory data analysis technique. How PCA is applied to the results of retrieval experiments can be found in the work of Dinçer (2007)[5]. PCA is basically a dimension reduction technique. Since higher dimensional spaces are difficult to inspect, it becomes necessary to reduce them into lower dimensions. Roughly speaking, in a test collection with many topics, groups of topics are often performed similarly by the retrieval systems. One reason for this is that more than one topic might be measuring the same driving characteristic governing the behavior of the retrieval systems. In many systems there are only a few such driving forces. But an abundance of instrumentation enables us to measure dozens of topics. When this happens, you can take advantage of this redundancy of information. Thereby, you can simplify the problem by replacing a group of topics with a single new “meta topic”. PCA is simply a method for achieving this simplification. In PCA, principal components correspond to those “meta topics”. Each principal component is a linear combination of the original topics. All the principal components are orthogonal to each other; so, opposed to the original topics, there is no redundant information. The first principal component is a single axis in space. When you project each measured performance score on that axis, the resulting values form a new meta topic. And the variance of this meta topic is the maximum among all possible choices of the first axis. The second principal component is another axis in space, perpendicular to the first. Projecting the performance scores on this axis generates another new meta topic. The variance of this meta topic is the maximum among all possible choices of this second axis. Thus, the first two principal components together can be used to define two orthogonal dimensions, and hence can be used to visualize the mutual performance relationships between considered retrieval systems in a two dimensional representation, by means of a scatter plot, which accounts for the major part of the total performance variations observed on the original topics. The scatter plot of the component scores of IRRA runs and the loadings of 49 topics on the first and the second principal axes is given in Figure 2. There are three extra points in the plot labeled as “APWorst”, “APMedian” and “APBest”, in addition to the IRRA runs. The point “APWorst” can, for instance, be thought of as a run that performs each topic with a statAP score equal to the statAP score of the worst run.

Interpretation of a PCA biplot is simple. Runs, which are cumulated at the same location in the plot, are the runs that show relatively close average performances over the same set (subset) of topics. Groups of runs, which are compatible in performance, may scatter to different locations in the plot. Subsets

[5] Interested readers may obtain the MATLAB M file for PCA by e-mail.

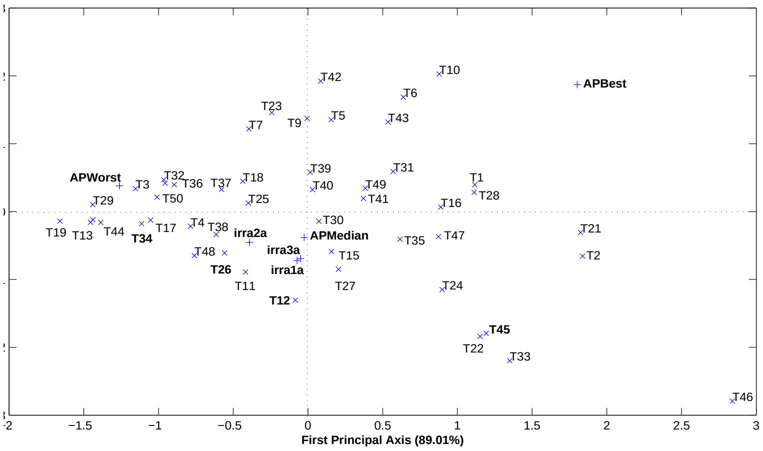


Figure 2: PCA biplot of IRRA adhoc runs and 49 TREC topics based on the statAP scores. Component scores of runs and the loadings of topics are standardized so that the component scores plot of runs and the component loadings plot of topics can be superimposed on one another.

of topics associated with those locations vary and can be determined by the loadings of topics on the associated principal components. Origin of the given PCA plot represents the average performance over all topics, and the gradient levels of observed performance scores are depicted towards the corners of the plot. The first principal component accounts for the 89.01% and the second principal component accounts for the 8.92% of the total performance variation observed on all topics; thus, the given PCA plot explains about 99% of the total performance variation among the considered runs across the 49 adhoc topics. In the given PCA plot, first principal component is positively related to almost all topics; hence, the first principal axis acts as an index variable, and agrees with the *statMAP* measure. On the other hand, second principal component contrasts the performance of runs on two subsets of topics.

The first principal component is dominated by the topics, including 46, 2, 21, 33, and 45; thus, the runs that are effective on these topics tend to have high scores on the first principle axis, and the ineffective runs tend to have low scores. On the other hand, the second principal component contrasts the performance scores observed on topics, including 10, 42, 6, 23, and 9, with the performance scores observed on topics, including 46, 33, 22, 45, and 12. This means that the high component scores along the second principal axis indicate high values of effectiveness on the former topics, and low values of effectiveness on the later topic. Conversely, the low component scores indicate high values of effectiveness on the later topics, and low values of effectiveness on the former topics. However in particular to the case at hand, interpretation should slightly be different, because there can be no run that could perform a topic better than the "APBest" run. That is to say, on the former topics, the observed differences in performance between the "APBest" run and the IRRA runs are higher than the corresponding differences observed on the former topics. For example, on topic 10, the observed statAP score of "APBest" is 0.8994, and the best score amongst others is of "irra2a", with 0.1604. But on topic 33, the observed score of "APBest" is 0.5224, and the observed score of "irra3a" is 0.4736. In this respect, it can be said that the contrast between the former topics and the later topics is related to the difference between the "APBest" run and the IRRA runs. On the former topics (i.e., the topics located on the top half panel of the plot), magnitude of difference is high, while on the later topics (i.e., the topics located on the bottom half panel of the plot), it is low. That's why all topics on which the IRRA runs have the best score are located in the bottom half panel of the plot. But notice that, except for the topic 45, all of those topics are located on the left-bottom panel of the plot. With respect to the first principle component, this means that the magnitudes of the scores observed on that topics are low in general, compared to the magnitudes of the scores observed the topics located on the right-bottom panel of the plot.

In PCA biplots, the proximities between runs and topics can only be interpreted through the principal components. That is to say, the closeness of two objects in a PCA biplot is meaningful only if the objects are of the same kind. Cross-proximities between a run and a topic has no direct interpretation, topics and runs can only be related through principal components. In the given PCA biplot, the topics that dominate the first principle component tend to have high scores on the first principal axis, proportional to the weights assigned by the first eigenvector, obtained by the spectral decomposition of the data matrix in use. For example, topic 19 is the topic that has the lowest score along the first principal axis, and so, it is the topic that is farthest from the "APBest" run. But "APBest" is still the run that has the highest score on this topic. Position of topic 19 indicates that the magnitude of the performance scores observed on topic 19 is lower than the magnitude of the performance scores observed on the topics that have high scores on the first principal axis, such as topic 46. This means that the effect of topic 19 on the average performance of the considered runs is relatively less than the effect of topic 46. It can, therefore, be said that the average performance scores of the considered runs are mostly determined by the topics that have high scores on the first principle axis. This suggests that the performance ranking of the considered runs could remain constant, even if the most of the topics located on the left-half panel of the plot are ignored (starting from the left most topic).

Recall that "irra1a" and "irra3a" are the median runs, on the average. This is clearly depicted by the given PCA plot: they are located close to the "APMedian". But, as seen, "irra2a" is also located close to the "APMedian". This suggests that, for the influencing topics, the performance differences between the IRRA runs are negligible, when they are compared to the performance differences observed among "APWorst", "APMedian" and "APBest" runs.

Comparison of the Performance Profiles of IRRA runs

Retrieval effectiveness has many dimensions, and the average performance is just one of those dimensions. *Performance profile* of a retrieval strategy is also an important dimension of retrieval effectiveness, and refers to the distribution of its total (average) performance on topics. It is a fact that, given a set of topics, a particular level of total performance, say A, can be obtained by infinitely may sets of topic scores, satisfying the equation $A = x_1 + x_2 + \ldots + x_c$, where x_j is the performance score observed on topic j ($j = 1, 2, \ldots, c$). This suggests that two particular runs may have the same average performance over a set of topics, but their observed performance scores on individual topics can be different, as long as the scores of each run sum up to the same total over all topics. In contrast, two runs may have different average performance scores, while they show the same performance profile across the given set of topics. In this case, per topic performance scores of runs must be the multiple of each other on every topic. For instance, suppose that the total performance of a run (r_1) is A_1. Then the total performance of the other run (r_2) must be $A_2 = \epsilon A_1$ (for $\epsilon \geq 0$):

$$\begin{aligned} A_1 &= x_1 + x_2 + \ldots + x_c, \\ A_2 = \epsilon \cdot A_1 &= \epsilon \cdot x_1 + \epsilon \cdot x_2 + \ldots + \epsilon \cdot x_c. \end{aligned}$$

This suggest that, whatever the functional difference between r_1 and r_2 is, it evenly contributes to the performance of r_2 on every topic. Moreover, in relation to the performance profile of retrieval system, we can consider a research question, which may arises as "Is it possible to define an ideal performance profile for a given retrieval system?". To define the ideal performance profiles of retrieval systems and to compare their observed performance profiles, a statistical, multivariate, exploratory data analysis technique, called the *correspondence analysis* (CA), can be used. How CA is applied to the results of retrieval experiments can be found in the work of Dinçer[6].

The independence notion/model, which is used in the DFI based index term weighting methods, is also the notion behind the CA. According to the independence model, a retrieval system is said to be independent from topics, if the observed performance of that system on each topic is equal to the performance suggested by the independence model. For any retrieval strategy, to be independent from topics is important, because, given a test collection with a particular set of topics, if the observed performance of a particular retrieval strategy is independent from the topics, then it is reasonable to expect that a similar performance would also be observed on different sets of topics[7]. The performance suggested by the independence model vary for each pair of system and topic, according to both the total

[6] To the interested readers, I can send MATLAB M file for CA by e-mail.

[7] Of course, provided that the topic set in use is a true representative of the topic population, i.e., only if the topic set is large enough and a proper random sample from the population.

performance of the given system (i.e., the system sum score over all topics) and the total performance score of the given topic (i.e., the topic sum score over all systems). Let $x_{i.} = \sum_{j=1}^{c} x_{ij}$ and $x_{.j} = \sum_{i=1}^{r} x_{ij}$ denote, respectively, the total performance score of system i ($i = 1, 2, \ldots, r$) and the total performance score of topic j ($j = 1, 2, \ldots, c$), where x_{ij} is the performance score of system i observed on topic j. Then the performance expected under independence for system i on topic j is given by

$$e_{ij} = x_{i.} \cdot \frac{x_{j.}}{x_{..}} \qquad \left[= \frac{x_{.i}}{x_{..}} \cdot x_{.j} \right]$$

where $x_{..} = \sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij}$ is the grand total. In consequence, system i is said to be independent from topic j, if the performance score of system i on topic j, x_{ij} is equal to e_{ij}, the performance suggested by the independence model. It immediately follows that, under independence, $x_{i.}$, the sum of observed scores of system i over all topics is expected to distribute on topics, according to the proportion of the total performance scores of topics, $x_{.j}/x_{..}$. This distribution is referred to as the *performance profile expected under independence*, and can be thought of as the ideal performance profile. In here, the proportion of the sum of performance scores observed on a particular topic to the sum of performance scores observed on all topics is the estimate of the population performance proportion of that topic, and hence, defines a conditional performance standard for every system. The independence notion can also be interpreted for topics, symmetrically. Under independence, $x_{.j}$, the sum of observed scores on topic j over all systems is expected to distribute on systems, according to the proportion of the total performance scores of systems, $x_{i.}/x_{..}$ (i.e., the equation given in brackets).

To a certain extent, CA is similar to the PCA. But in fact, they are completely different techniques that can be used to visualize different aspects of the same set of data. Two systems that have different average performance scores can show the same performance profile across a given set of topics; so, opposed to the PCA plots, they can be located at the same position in the CA plots. Conversely, two systems that have the same average performance can show different performance profiles; so, they can be located apart from each other in the CA plots. Moreover, CA accounts for only the *gains* and the *losses* in a retrieval system's performance, relative to the performance expected from that system under independence (i.e., the difference of the observed performance profiles of retrieval systems from their performance profiles expected under independence); it does not account for the magnitudes of the performance scores. In this respect, a particular retrieval system that has a low average performance can gain (or lose) performance on some topics, more than the gains (or the losses) of other systems that have high average performance in general. As a result, CA is unique in that, it enables us to explore the interdependencies between the retrieval systems and the topics, by identifying the systems that response to some topics more than the response expected under independence, and simultaneously, the topics that are responded by some systems more than the response expected under independence.

The CA biplot of the IRRA runs and the adhoc topics is given in Figure 3. Origin of the given CA biplot represents the independence: it is the point of no divergence from the performance profile expected under independence. In other words, the retrieval systems and the topics that are located close to the origin are independent from topics and retrieval systems, respectively. In consequence, the systems that depend on some topics, and the topics that depend on some systems are located far from the origin. For the systems and the topics that are located far from the origin, systems with similar observed performance profiles are located close to each other in the plot. In the same way, topics with similar observed performance profiles are located close to each other, too. On the other hand, the direction of divergence from independence vary system from system, topic from topic. On a subset of topics, the observed performance scores of a particular system may be below the performance scores suggested by the independence model, while on the rest, they may be equal to or above the performance scores expected under independence. In CA biplots, a retrieval system, which departs from independence, is located close to the topics, on which its observed performance scores are above the performance scores expected under independence, while in contrast, it is located far from the topics, on which its performance scores are below the performance scores expected under independence.

It appears that each IRRA run has a different performance profile across the adhoc topics, and gains (or loses) performance on different topics. Along the first principle axis, "irra2a" and the other IRRA runs are contrasted, and this contrast seems to be related mostly to the contrast between two subsets of topics, which can be characterized by the topic 34 and the topic 8. Along the second dimension, it appears that only the "irra1a" and "irra3a" are contrasted, i.e., "irra2a" is located on the horizontal

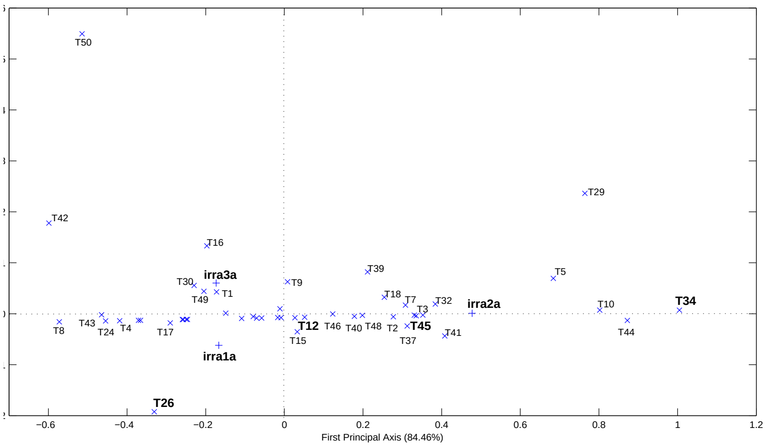


Figure 3: CA biplot of IRRA runs and adhoc topics based on the statAP scores.

(dashed) line crossing the origin. The first principal axis explains about the 84%, and the second principal axis explains about the 16% of the total performance deviations of IRRA runs from independence. This suggests that the difference between the performance profiles of “irra2a” and the others is higher than the difference between the performance profiles of “irra1a” and “irra3a”. Recall that “irra2a” has low average performance, relative to the other IRRA runs. However, when the topics that are located close to the “irra2a” are examined in detail, it is apparent that the performance scores observed on that topics are relatively lower than the performance scores observed on the contrasted topics, in magnitude. For example, on topic 44, the statAP scores of “irra1a”, “irra2a” and “irra3a” are 0.0189, 0.0662, and 0.0173, respectively; similarly, on topic 10, they are 0.0497, 0.1604, and 0.0504. However, on topic 24, they are 0.4753, 0.0635, and 0.4753. In other words, the number of topics, on which “irra2a” gains performance, is relatively more than the number of topics, on which the other IRRA runs gains, but the effects of that topics on the average performance is relatively lower than the effects of others, because of the differences in magnitudes.

On the other hand, the given CA plot also reveals that the performance of “irra2a” depends on the topics in use, more than the performance of other IRRA runs: along the first principal axis, “irra2a” is more distant from the origin than “irra1a” and “irra3a”. In this respect, “irra1a” seems to be the best run amongst the IRRA runs. Its performance shows a considerable divergence from the performance expected under independence, only on the topic 26, where the scores of “irra1a”, “irra2a” and “irra3a” are 0.2352, 0.0527, and 0.1652, respectively. Although “irra1a” has also the best score at topic 12, it appears that this scores is an expected one for it, because topic 12 is located close to the origin. But this case is valid not only for the “irra1a”, but also for the “irra2a” and “irra3a”. The observed performance scores of “irra2a” and “irra3a” on topic 12 are also the expected performance scores from them under independence. On topic 12, the scores of “irra1a”, “irra2a” and “irra3a” are 0.2729, 0.2180, and 0.2729, respectively.

As a result, the considered DFI formulae are good at different subsets (possibly, different types) of topics, or in other words, effectiveness of each DFI formula depends on a subset of topics, which is different from the subsets of topics that the effectiveness of other formulae depend on. The DFI formula used in “irra1a” is actually a derivation of the DFI formula used in “irra3a”. If topic 26 is ignored, it can be said that the derivation used in “irra1a” is beneficial, because this derivation approaches the performance profile of “irra3a” to the performance profile expected under independence (i.e., the ideal performance profile). Performance of “irra1a” is, in this respect, more predictable than the performance of “irra3a”. As a result, if a DFI formula has to be singled out, the DFI formula used in “irra1a” should be preferred to the other DFI formulae.

5.2 Diversity Task

The estimated performance scores of IRRA diversity runs are given in the Table 3.

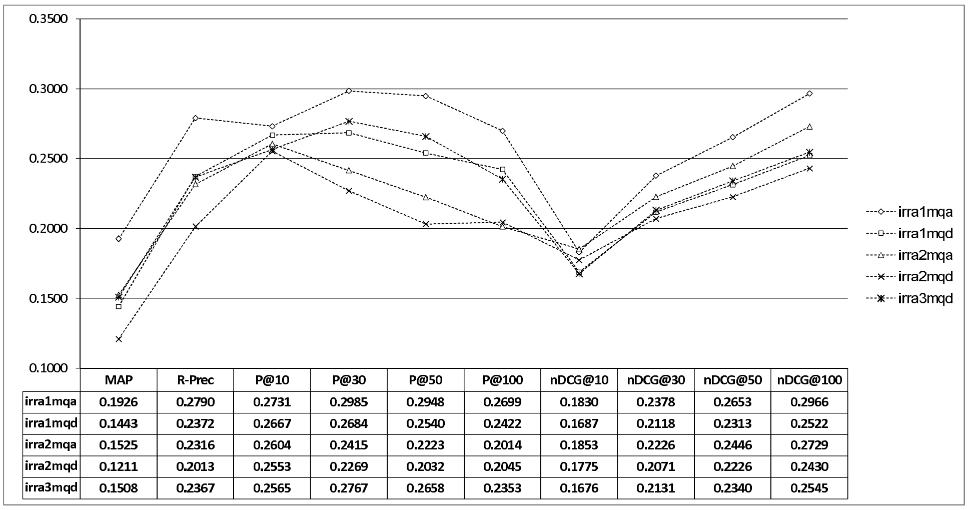
	α-nDCG@10	IA-P@10
irra1d	0.1310	0.0630
irra2d	0.1610	0.0600
irra3d	0.1300	0.0610
best	0.5267	0.2447
median	0.1758	0.0733
worst	0.0000	0.0000

Table 3: Estimated performance scores of IRRA diversity runs.

Based on the *α-nDCG@10* measure, it appears that "irra2d" outperforms other IRRA runs, whereas the results based on *IA-P@10* indicates that they are compatible in performance. These results do not agree with the adhoc results in general. In particular, they differ about the performance of "irra2d". Recall that DFI formula used in "irra2d" is the DFI formula used in "irra2a". Therefore, these results actually indicate that "irra2a" is better than other IRRA runs in returning diverse results to the considered topics. This means that the DFI formula given in Equation 2 is qualitatively different from the other DFI formulae, in that it can satisfy more diverse information needs of users, than the other IRRA runs. Together with the adhoc results, diversity results suggest that, first, a better index term weighting formula could be possible, if the DFI formula used in "irra1a" is somehow fused with the DFI formula used in "irra2a". Second, the average performance of "irra2a" seems to be relatively unstable/uncertain; so, the performance ranking of IRRA runs could change. In this respect, MQ track results may provide more information.

5.3 Million Query Track

Performance summaries of the IRRA runs that were submitted to the MQ track are given in Figure 4. The given performance summaries are of the *base* type of evaluations over 310 topics, where judgment pools include the documents sampled from the IRRA runs.



	MAP	R-Prec	P@10	P@30	P@50	P@100	nDCG@10	nDCG@30	nDCG@50	nDCG@100
irra1mqa	0.1926	0.2790	0.2731	0.2985	0.2948	0.2699	0.1830	0.2378	0.2653	0.2966
irra1mqd	0.1443	0.2372	0.2667	0.2684	0.2540	0.2422	0.1687	0.2118	0.2313	0.2522
irra2mqa	0.1525	0.2316	0.2604	0.2415	0.2223	0.2014	0.1853	0.2226	0.2446	0.2729
irra2mqd	0.1211	0.2013	0.2553	0.2269	0.2032	0.2045	0.1775	0.2071	0.2226	0.2430
irra3mqd	0.1508	0.2367	0.2565	0.2767	0.2658	0.2353	0.1676	0.2131	0.2340	0.2545

Figure 4: The statAP estimates for IRRA MQ runs over valid 310 topics.

MQ results obtained over 310 (valid) topics suggest that the average performances of IRRA runs are higher than the corresponding average performances observed on 49 adhoc topics; for example, the average performance of "irra1mqa" over 310 topics is 0.1926, and approximately 25% higher than its average performance (0.1530) over 49 adhoc topics. The performance scores measured by the *R-Prec* measure agree with the MAP measure about the ranking of IRRA runs. On the other hand, average precision scores measured at 10 documents (i.e., P@10) indicates that all IRRA runs are almost compatible in performance, and centered around the statAP score of 0.2600. However, as the measurement depth increases, they spread towards different directions. The performances of “irra1a” and “irra3mqd” peak at 30 document as 0.2985 and 0.2767, while the performances of others consistently decrease. Up to the 100 document, the ranking of IRRA runs nearly remains unchanged. In general, these results agree with the adhoc results with respect to the relative performance ranking of IRRA runs. As seen, “irra3a” was not submitted to the MQ track. The major reason is that the MQ track is limited to 5 runs for each participating group. But the performance of “irra3mqa” can be estimated from the performance of “irra1mqa”. “irra1a” and “irra3a” are known to be compatible in average performance; they only differ in performance profiles. If it was submitted, its performance should have been close to (or slightly higher at 30 and 50 documents than) the observed performance of “irra1mqa”, as also suggested by the observed similarity in performance between the “irra1mqd” and “irra3mqd”.

In detail, “irra1mqa” is above the median scores of 210 topics out of 310 (68%), while “irra2a” is above median for 147 topic (47%). “irra1mqa” has the best scores on 14 topics, while “irra2a” has the best scores on 16 topics; they coincide on only one topic (topic 20556). The statAP score of “irra1mqa” is zero for 18 topics, and the statAP score of “irra1mqa” is zero for 23 topics. On the other hand, for the IRRA runs that use the diversity strategy, “irra1mqd”, “irra2mqd” and “irra3mqd” are above the median scores of 159 topics (51%), 123 topics (40%), and 176 topics (57%), respectively. They have the best statAP scores, respectively, for 9, 14, and 12 topics. The PCA biplot of the components scores of IRRA runs and the loadings of 310 topics on the first and the second principal components are given in Figure 5, which provides more information about the mutual performance relationships between them across 310 MQ topics.

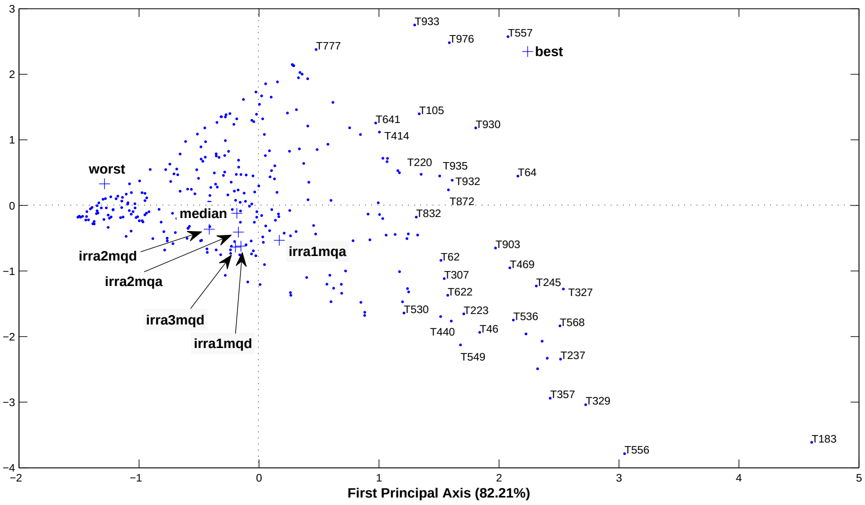


Figure 5: The PCA biplot of component scores of the IRRA MQ base runs, based on the statAP estimates.

In general, this PCA plot agree with the PCA plot given in Figure 2: IRRA runs are scattered close to the median. In particular, the average performance of “irra1mqa” seems to be better than the other IRRA runs: it is the right most run along the first principal axis. The labeled topics are the topics that dominate the first principal component, and so, they are the topics that have the highest influence on the final performance ranking of IRRA runs; for example, topic 557 is the topic where the highest difference in performance between the “APBest” and the IRRA runs is observed, while in contrast, topic 183 is the topic where the lowest difference in performance between the “APBest” and the IRRA runs is observed. The CA biplot of the IRRA runs is given in Figure 6.

As seen in the given CA biplot, along the first principal axis, “irra1mqX” (“irra3mqX”) and “irra2mqX” are contrasted. This contrast indicates that the DFI formulae used in “irra1” and “irra2” produce two different retrieval strategies that are sensitive to different subsets (types) of topics. Along the second principal axis, the strategies used in adhoc runs and diversity runs are contrasted, and again, this contrast suggests that these two strategies are also sensitive to different subsets of topics.

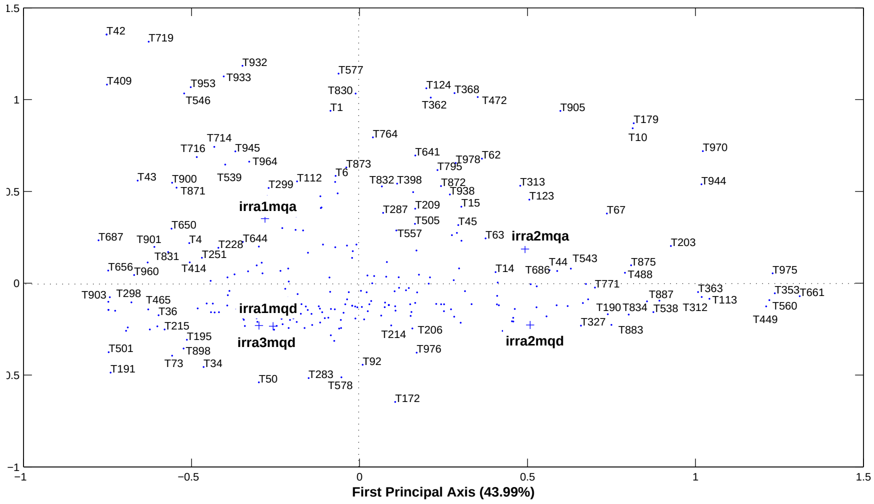


Figure 6: The CA biplot of the IRRA MQ base runs, based on the statAP estimates.

6 Conclusions

IRRA group participated in the 2009 Web track adhoc task and the diversity task, as well as the Million Query (MQ) track. In this year, the major concern is to examine the effectiveness of a novel, nonparametric index term weighting method based on the *divergence from independence* (DFI). All submitted IRRA runs are of the *content-only* type: none of the information in Web pages, such as Meta information, Link information, etc., was distinguished and used. In addition, during searching, none of the common supplementary methods, such as parsing rules, phrase processing, query processing, query expansion, relevance feedback, thesaurus, WordNets, etc., was used. IRRA runs are pure, out-of-the-box DFI runs. In brief, three possible DFI formulae are considered, and used in IRRA runs, as the TF component of the well-known TF×IDF weighting scheme. "irra1" uses a derivation of the DFI formula given in Equation 1 and a new IDF formula developed on the basis of the DFI, as an alternative to the existing IDF formulae. "irra2" and "irra3" use the same IDF formula with "irra1", and they employ the DFI formulae given in Equation 2 and Equation 1, respectively. In summary, empirical results show that "irra1" and "irra3" are compatible in performance on the average, and better than "irra2".

In the *adhoc task*, "irra1" has the best *statAP* scores on topic 12 and 26. Similarly, "irra2" has the best scores on topic 34 and 45; "irra3" has the best score on topic 12. "irra1" are above the median on 24 topics; "irra2a" are above the median on 14 topics; "irra3a" are above the median on 26 topics. The results of the analysis of the performance profiles of IRRA runs reveal that, to a certain degree, performances of considered DFI formulae depend on topics. In other words, effectiveness of each DFI formula depends on the types topics in use. On the other hand, it can also be argued that the derivation used in "irra1" can be a remedy, in this respect. It can correct the performance profile of "irra3" and approach it to the *ideal performance profile*; thereby, the performance of "irra1" might be more predictable (or stable) than the performance of other IRRA runs.

The runs submitted to the *diversity task* are the runs that are submitted to the adhoc task, with only one difference. For the diversity task, the result sets are filtered, such that they consist of at most two Web pages from the same host (URL). Based on the *α-nDCG@10* measure, "irra2" outperforms other IRRA runs, whereas the results based on *IA-P@10* indicates that they are compatible in performance. The diversity task results suggest that "irra2a" is better than other IRRA runs in returning diverse results to the considered topics. The DFI formula given in Equation 2 is qualitatively different from the other DFI formulae. It can, therefore, satisfy diverse information needs of users. Together with the adhoc task results, diversity results actually suggest that a better index term weighting could be possible, if the two DFI formulae can somehow be fused in the same host system.

MQ track results agree with the adhoc results about the relative ranking of IRRA runs, but they differ in that, the average performance of IRRA runs might be underestimated in magnitude with insufficient number of topics. In MQ track, “irra1” is above the median for 210 topics out of 310 (68%), and “irra2” is above median for 147 topic (47%). “irra1” has the best scores on 14 topics (4.5%), and “irra2” has the best scores on 16 topics (5%): they coincide on only one topic (i.e., topic 20556). The number of topics used in the MQ track is approximately the 6 times of the number of topics used in the Web adhoc task. It is worth noting that, for the topics on which the IRRA runs have the best scores, the number of MQ topics are about the 7 times of the number of adhoc topics, e.g., “irra1” has the best scores on 2 adhoc topics and has the best scores on 14 MQ topics. The MQ results of the analysis of the performance profiles of IRRA runs agree with the adhoc results. Both indicate that “irra1” and “irra2” are two different retrieval strategies that are sensitive to different subsets (probably, different types) of topics.

In summary, the results of the TREC 2009 experiments verify that the *independence notion* can provide a simple, but powerful basis for weighting index terms, so that it promises a new direction in the IR research.

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 Trans. Inf. Syst.*, 20(4):357–389, 2002. ISSN 1046-8188. doi: http://doi.acm.org/10.1145/582415.582416.

W. J. Conover. *Practical Nonparametric Statistics.* Wiley, New York, $3^r d$ edition, 1999. ISBN: 0-471-16068-7.

B. T. Dinçer. Correspondence analysis for the evaluation of the results of information retrieval experiments: Comparison of the performance profiles of IR systems. *Journal of Information Retrieval.* (to appear).

B. T. Dinçer. Statistical principal components analysis for retrieval experiments. *Journal of the American Society for Information Science and Technology*, 58(4):560–574, 2007. ISSN 1532-2882. doi: http://dx.doi.org/10.1002/asi.v58:4.

M. Greenacre. *Theory and Applications of Correspondence Analysis.* Academic Press, New York, 1984. ISBN 0122990501.

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.

M. Jambu. *Exploratory and Multivariate Data Analysis.* New York: Academic Press, 1991.

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 *Proceedings of the Seventeenth Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, Dublin, 1994. Springer-Verlag, New York.

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.

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, pages 513–523, 1988.