Exploring the Stability of IDF Term Weighting

Xin Fu* and Miao Chen*

University of North Carolina, Chapel Hill, NC 27599 USA
{xfu,mchen}@unc.edu

Abstract. TF·IDF has been widely used as a term weighting schemes in today's information retrieval systems. However, computation time and cost have become major concerns for its application. This study investigated the similarities and differences between IDF distributions based on the global collection and on different samples and tested the stability of the IDF measure across collections. A more efficient algorithm based on random samples generated a good approximation to the IDF computed over the entire collection, but with less computation overhead. This practice may be particularly informative and helpful for analysis on large database or dynamic environment like the Web.

Keywords: term weighting, term frequency, inverse document frequency, stability, feature oriented samples, random samples.

1 Introduction

Automatic information retrieval has long been modeled as the match between document collection and user's information needs. In any implementation based on this model, the representation of document collection and users' information need is a crucial consideration. Two main questions are involved in the representation: decision on what terms to include in the representations and determination of term weights [1].

TF·IDF is one of the most commonly used term weighting schemes in today's information retrieval systems. Two parts of the weighting were proposed by Gerard Salton [2] and Karen Spärck Jones [3] respectively. TF, the term frequency, is defined as the number of times a term in question occurs in a document. IDF, the inverse document frequency, is based on counting the number of documents in the collection being searched that are indexed by the term. The intuition was that a term that occurs in many documents is not a good discriminator and should be given lower weight than one that occurs in few documents [4]. The product of TF and IDF, known as TF·IDF, is used as an indicator of the importance of a term in representing a document.

The justification for and implementation of IDF has been an open research issue in the past three decades. One thread of research focuses on IDF calculation itself and proposes alternative IDF computation algorithms [5]. The other thread of research seeks theoretical justifications for IDF and attempts to understand why TF·IDF works so well although TF and IDF exist in different spaces [6].

* These authors contribute equally to this paper.

H. Li et al. (Eds.): AIRS 2008, LNCS 4993, pp. 10–21, 2008.

There have been a vast number of studies on distribution of word frequencies and other man made or naturally occurring phenomena. Studies found that these phenomena often follow a power-law probability density function and a Zipf or a Poisson mixture frequency rank distribution [7, 8]. However, there are different opinions on values of the parameters in the distribution function. It was also noted that the parameters might change across genre, author, topic, etc. [7]. Finally, many early experiments were conducted on abstracts rather than full text collections. The language patterns in the full text can be very different from the abstracts.

This study does not intend to test the term frequency distributions, or to derive the estimators for distribution parameters. Instead, it aims to investigate the similarities and differences between IDF distributions based on the global collection and on different samples and to test the stability of the IDF measure across collections. The study examines how IDF varies when it is computed over different samples of a document collection, including feature-oriented samples and random samples. As Oard and Marchionini [9] pointed out, estimates of IDF based on sampling earlier documents can produce useful IDF values for domains in which term usage patterns are relatively stable.

The motivation of this study comes from the observation that advance knowledge of IDF is either impossible for a real world collection or too expensive to obtain. The practical aim of the study is to develop a more efficient algorithm that requires less computational time/cost, but at the same time, generates a good approximation to the IDF computed over the entire collection. In a dynamic world where new information is added accretionally to a collection, it would be informative to understand how collection based weights will evolve when new information is added. With this understanding, we may make recommendations such as if the collection size increases by more than x percent, then the IDF weights should be updated. A recommendation like this will be particularly useful in a dynamic environment like the Web.

2 Methods

The experiments were conducted on a 1.16GB collection of full text journal articles published mostly between 1997 and 2006. The articles came from 160 journals and all of them were available in XML format. In order to allow for comparison of IDF computed from different components of articles (title, abstract, reference and table or figure caption), articles which missed any of the above four components were discarded. After the pruning, 15,132 articles were left and used for experiments. The rest of this section describes the pre-processing and the statistical methods.

2.1 Preprocessing

The pre-processing of the collection consists of four steps: extracting, tokenizing, removing stop words, and stemming. The first step was to extract from each document the information that would be used in analysis. First, information from the following XML tags was extracted: journal ID, journal title, article identifier, article title, abstract, titles of cited works and table/figure captions tags. If a cited work was a journal article, the title of the article was used. If a cited work was a book, the book

title was used. Then, for the <body> section which contained the text of the articles, all information was extracted unless it was under headings which were non-topic in nature (e.g., those for structural or administrative information). Since there was variation across articles and across journals in the use of headings, we manually examined all the headings that appeared five times or more in the collection and identified non-topic headings for removal. We also sorted the headings alphabetically and identified variants of these headings and occasional misspellings.

The second step of pre-processing was to stripe off XML tags and tokenize the extracted parts into terms. We used space as the delimiter; therefore, only single words (instead of phrases) were considered. The tokenizer removed all the punctuation. In specific, this means that all hyphenated terms were broken into two. Next, we applied a basic stop word list called Libbow Stop Word List [10] which we expanded by adding all-digit terms, such as "100" and "1000", and octal control characters, such as "&#x0002b". We also removed all the DNA sequences (strings of 8 characters or longer which were formed solely by letters a, c, g, and t) and information in the <inline-formula> tags, which contained noise.

Finally, we used the popular Porter stemming algorithm [11] to stem the words. All the stemmed words were stored into a table in the Oracle database. For each stemmed term, the article ID and source ('T' for title, 'A' for abstract, 'C' for caption, 'R' for reference, and 'B' for body) were stored. Body is the part of an article other than the title, abstract, table/figure and reference. Headings are regarded as part of the body.

In sum, the experiment collection consisted of 15,132 full text articles of 3,000 words long on average. Each article contained a title, an abstract, a body, some table/figures and a list of references. For tables and figures, we were only interested in their captions, so we discarded table/figure contents. For references, we were only interested in titles of cited works (either book titles or titles of journal articles) and did not consider other information, such as authors and journal names. The vocabulary size was 277,905, which was the number of unique stemmed terms in the collection.

2.2 Statistical Methods

Various ways have been proposed to calculate IDF. A basic formula was given by Robertson [4]. A later discussion between Spärck Jones [12] and Robertson resulted in the following formula of IDF:

$$idf(t_i) = \log_2(\frac{N}{n_i}) + 1 = \log_2(N) - \log_2(n_i) + 1 \quad (1)$$

where N is the total number of documents in the collection and n_i is the number of documents that contain at least one occurrence of the term t_i.

For a particular collection (fixed N), the only variable in the formula is n_i, the number of documents in which the term t_i appears. Let us call it document frequency (DF). IDF is then a monotone transformation to the inverse of the document frequency. When a different collection is used, the IDF will differ only by a constant, $\log_2(N/N')$.Therefore, we can approach the IDF comparison problem by comparing the document frequency distribution in the global collection and in each of the subsampling collections.

There were 14 collections involved in the analyses: the global collection, four feature oriented sub-sampling collection (title, abstract, caption and reference) and nine random sample collection (from 10% to 90% at 10% interval). Each collection can be represented as N feature-value pairs with the feature being a term in the vocabulary and value being the document frequency of the term. Formally, the document frequency representation of a collection is: DF_xx {($term_1$, df_1), ($term_2$, df_2), …, ($term_{277905}$, df_{277905})}, with xx being G (global), T (title), A (abstract), C (caption), R (reference), 10 (10% random sample), …, 90 (90% random sample). For example, the global collection document frequency is represented as: DF_G {(a, 10), (a0, 17), …, (zzw, 1), (zzz, 2)}. When a term was missing in a sub-sampling collection, its document frequency was defined as 0.

With the above representation, the problem of comparing document frequency features of two collections (the global collection and a sub-sampling collection) was abstracted as comparing two data sets, DF_G and DF_sub, and asking if they follow the same distribution.

For each data set, we first summarized the data by plotting document frequencies against their ranks and showing a histogram of document frequency distribution. Some summary statistics were also reported to characterize the distribution.

To compare the two data sets, we first looked at their histograms to visually compare the shapes of the distributions. Then we generated scatter plots and computed correlation coefficients to estimate the strength of the linear relationship between the two data sets. As the data were by far not normally distributed, we used Spearman Rank Order Correlation Coefficient, instead of Pearson's Correlation Coefficient.

Finally, we computed the Kullback-Leibler distance between the two distributions as a numeric characterization of how close the distribution of the sub-sampling data set was from the distribution of the global data set. The Kullback-Leibler distance, also called the relative entropy [13], between the probability mass function based on a sub-sampling collection p(x) and that based on the global collection q(x) was defined as:

$$D(p \parallel q) = \sum_{x \in \mathrm{X}} p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(X)}{q(X)} \tag{2}$$

in which X was the vocabulary space, x was a term in the vocabulary space, p(x) and q(x) were the document-wise probability that the term occurred in the sub-sampling collection and the global collection, respectively. Specifically, p(x) was computed as $n_i/\sum n_i$, so all p(x) added up to 1. q(x) was computed in a similar way for the global collection. Note that in this definition, the convention (based on continuity arguments) that 0log(0/q)=0 was used. As Cover and Thomas [13] pointed out, the Kullback-Leibler distance is not a true distance between distributions since it is not symmetric and does not satisfy the triangle inequality; nonetheless, it is a useful and widely used measure of the distance between distributions.

Although there are many studies on the distribution of word frequency and frequency rank, it remains unclear which distribution and which parameters are the most appropriate for what type of collection. In this study, we used non-parametric methods without assuming any particular form of distribution.

3 Results

3.1 Feature Oriented Samples

We carried out the first set of experiments on feature based sub-sampling collections. We started with calculating document frequency, n_i in formula (1), for each stemmed term. The vocabulary size (i.e., the number of rows in the df_xx table) for the global collection and each of the special feature collections are summarized in Table 1.

Table 1. Vocabulary size for each collection

Data Set Code	Global	T	A	C	R
Collection Name	Global	Title	Abstract	Caption	Reference
Size	277,905	16,655	44,871	71,842	94,007

Figure 1 displays the scatter plots between the global DF and sample DFs. When a term was missing from a sub-sampling collection, its sample DF was defined as 0. The plots showed that the linear relationship between the global DF and any of the sample DFs was not very strong. The Spearman Correlation Coefficients were 0.5691 (global vs. reference) > 0.5373 (abstract) > 0.4669 (caption) > 0.3845 (title).

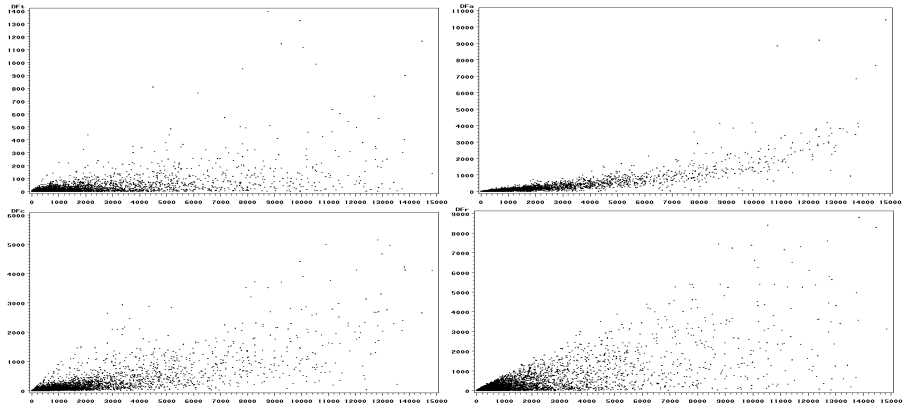

Fig. 1. Scatter plots between global DF and DF of Title (upper left), Abstract (upper right), Caption (lower left), Reference (lower right) collections

To help understand why the linear relationship was not very strong, we replotted the four pairs of DFs, but on a log-log scale. The plots are displayed in Figure 2. Since the logarithm function has a singularity at 0, these plots excluded all the terms that only appeared in the global collection, but not in a sub-sampling collection.

From these plots, we could easily see that the deviation from linear relationship happened more with terms that had low document frequencies in the sub-sampling collections. For example, in the upper left plot, y value of 0 (i.e., DF_T=1)

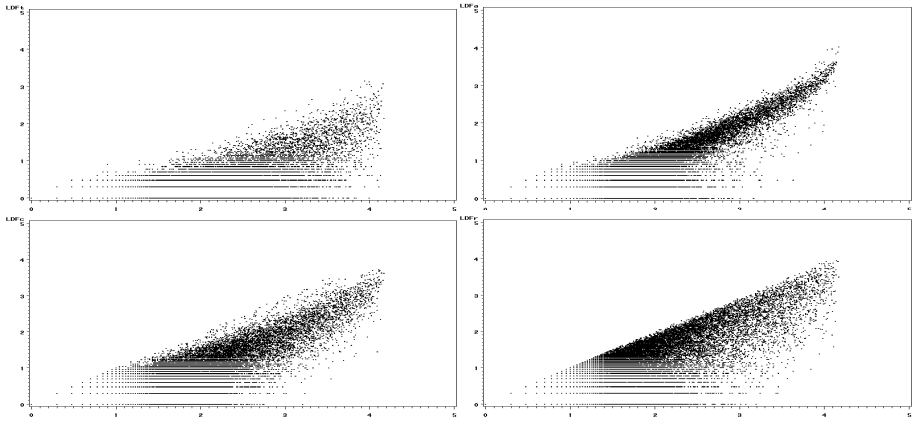

Fig. 2. Scatter plots between global DF and DF of Title (upper left), Abstract (upper right), Caption (lower left), Reference (lower right) collections on a log-log scale (base 10)

corresponded to a wide range of x values from 0 up to about 4 (i.e., DF_Global from 1 to 9,950). This means that a term that only appeared once in document titles appeared somewhere else in 9,950 documents.

Next, we computed the relative entropy for each of the feature-oriented samples compared to the global collection. Recall that the relative entropy was defined as:

$$D(p \parallel q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(X)}{q(X)} \tag{2}$$

in which X was the vocabulary space and x was a term in the vocabulary space. For each term t_i, p(x) and q(x) were computed as $n_i/\sum n_i$ with n_i being the document frequency for term t_i in the sub-sampling collection and in the global collection, respectively. The results were: 0.2565 (abstract) < 0.3838 (caption) < 0.3913 (reference) < 0.7836 (title). This means that the abstract data set distributed most similarly to the global data set, followed by the caption data set and the reference data set. Note that this order was different from the order for Spearman Correlation Coefficients.

In summary, the above analyses compared the four sample distributions with the distribution of the global data set from different approaches. All five sets of data were heavily skewed. The scatter plots and the correlation coefficients showed that the linear relationships between the sample data sets and the global data set were not very strong. So, IDFs calculated based on those samples may not be very good estimates for those of the global collection. In particular, the title was probably too crude to be a good compression of the full text. Besides, qualitative analyses suggested that there were some systematic missing of terms in the title and reference collections. In other words, their language patterns were not the same as the rest of the collection. Furthermore, distribution distance measures led to different results than the correlation coefficients. The disagreement might be an indication that there were more complex relationships in the data.

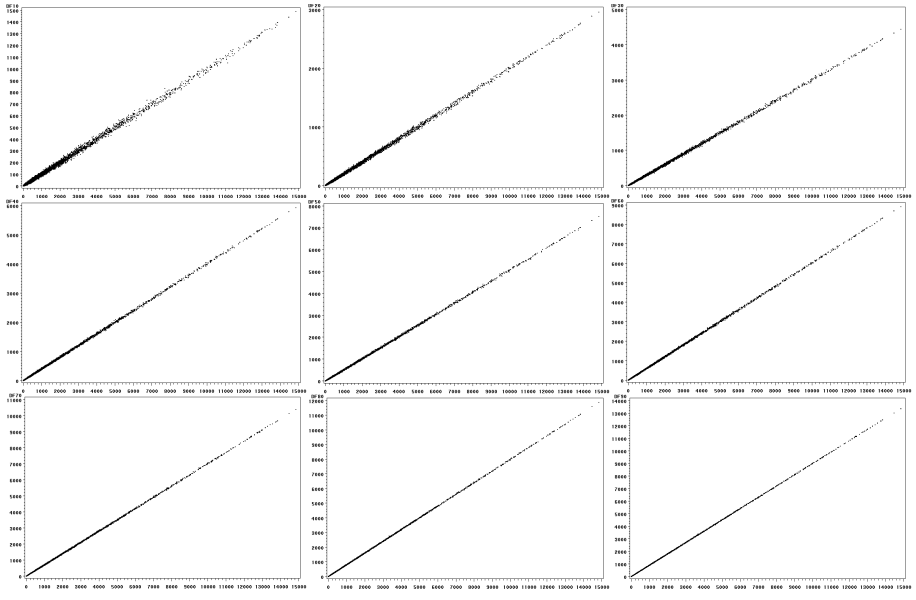

Fig. 3. Scatter plots between global DF on x-axis and DF of random 10% to 90% collections on y-axis (from left to right, then top to bottom)

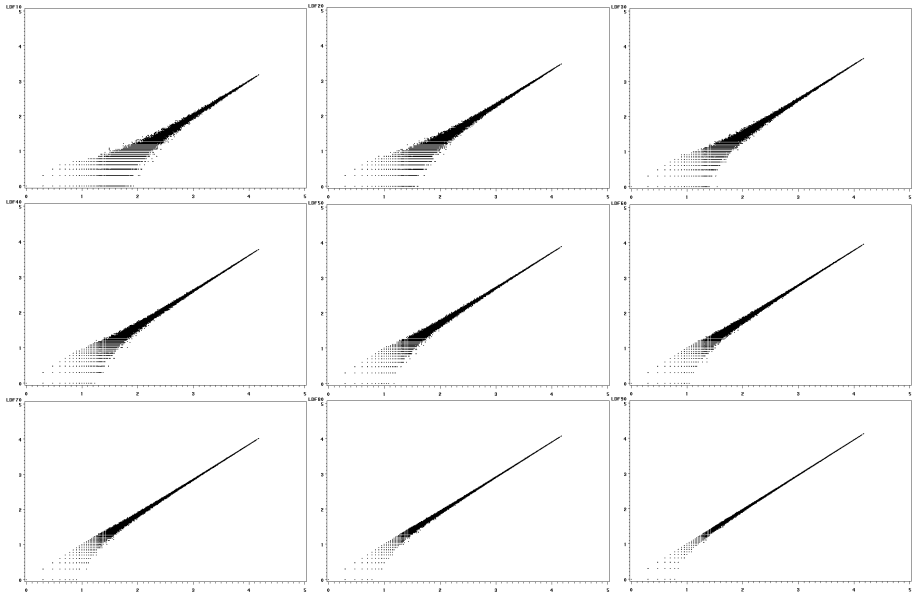

Fig. 4. Scatter plots between global DF and DF_10 to DF_90 (from left to right, then top to bottom) on a log-log scale (base 10)

3.2 Random Samples

In this set of experiments, we used random samples of the collection. For a 10% sample, we randomly picked 1,513 articles out of the 15,132 articles and built the data set with terms that came from this subset of articles. The vocabulary size for the global collection and each of the random sample collections are summarized below.

Table 2. Vocabulary size for the global collection and each of the random sample collections

Data Set	10%	20%	30%	40%	50%
# articles	1,513	3,026	4,540	6,053	7,566
Vocabulary	68,724	102,208	130,001	158,318	181,041
Data Set	60%	70%	80%	90%	Global
# articles	9,079	10,592	12,106	13,619	15,132
Vocabulary	202,236	222,170	241,675	260,225	277,905

In Figure 3, we display the scatter plots between the global DF and random sample DFs. When a term was missing from a random sample collection, its sample DF was defined as 0. The plots showed that the linear relationship between the global DF and any of the random sample DFs was stronger than those between the global DF and DFs of feature oriented samples. As the sample size increased, the Spearman Correlation Coefficient kept increasing from 0.5591 (10%) to 0.9299 (90%). It should be noted that although all these scatter plots look much closer to a straight line than the previous set, the smallest Spearman Correlation Coefficient between the random data set and the global data set [Corr(DF_10, DF_Global)=0.5591] was actually smaller than the largest one between the feature oriented data set and the global data set [Corr(DF_R, DF_Global)=0.5691]. We attributed this to the fact that Spearman Correlation Coefficient was rank based statistic while our data sets were seriously skewed and had a significant portion of "ties" in DF rank. Therefore, the scatter plots might be better descriptions of the relationship between data sets than simply looking at the correlation coefficients.

We also plotted the nine pairs of data sets on a log-log scale (base 10). The plots are displayed in Figure 4. Again, these plots excluded all the terms that only appeared in the global collection, but not in a sub-sampling collection.

The deviation from the linear relationship happened again with terms that had low document frequencies in the sub-sampling collections. However, unlike the title and reference collection in the first experiment, we did not notice any systematic missing of terms in the random samples. For example, the terms which were missing from the 10% collection while appearing in the largest numbers of documents in the global collection (number in parenthesis) were: array (64), darwin (62), bisulfit (59), subgraph (57), repercuss (56), Shannon (56), notochord (53), peyer (52), ced (51), puc (50). These numbers were also much smaller than those associated with the feature oriented samples.

Next, we computed the relative entropy for each of the random sample data set distribution compared to the global data set distribution. The values went down consistently as the sample size increased, from 0.0883 (10%) to 0.0025 (90%). Also, all of these relative entropies were much smaller than those we obtained in the previous section.

Overall, the results suggested that compared to distributions of feature oriented sample data sets, distributions of random sample data sets were much more similar to that of the global data set. IDFs calculated based on those samples should be better estimates for the global IDF. It is not exactly clear how large a sample will be "sufficient" to represent the whole; however, it is clear that terms that have low document frequencies in the sub-sampling collections are less stable then those that appear in a lot of the documents in the samples. In other words, we can safely use high-DF terms in a sample collection to predict the DF or IDF in the global collection, while we are much less confident in using low-DF terms.

Finally, back to the ultimate task of using sample data to predict global IDF, we ran a simple linear regression of IDF Global on IDF_10 to IDF_90. The regression equations are:

$$IDF\_G = -2.4096 + 1.3661 \bullet IDF\_10$$
$$IDF\_G = -1.9296 + 1.2615 \bullet IDF\_20$$
$$IDF\_G = -1.5332 + 1.1957 \bullet IDF\_30$$
$$IDF\_G = -1.2108 + 1.1507 \bullet IDF\_40$$
$$IDF\_G = -0.9306 + 1.1134 \bullet IDF\_50$$
$$IDF\_G = -0.7199 + 1.0850 \bullet IDF\_60$$
$$IDF\_G = -0.5164 + 1.0595 \bullet IDF\_70$$
$$IDF\_G = -0.3171 + 1.0366 \bullet IDF\_80$$
$$IDF\_G = -0.1614 + 1.0179 \bullet IDF\_90$$

R-Squares for the regressions ranged from .8076 (10%), .9211 (40%), .9555 (60%), to .9907 (90%). This means that on average IDFs calculated with 10% of the sample can predict 80 percent of the variation in IDFs calculated with the entire collection. We did not use the same model for feature based sampling data sets because we found that the error distributions were seriously non-normal for those data sets.

4 Conclusions and Discussions

This study explored the relationship between IDF distributions (via DF distributions) in sub-sampling collections and in the global collection with the intention to derive an optimal sampling method that used a minimal portion of the entire collection, but generated a satisfactory approximation to the IDF generated over the whole collection. We looked at two different sampling methods, feature based sampling and random sampling. Feature based sampling resulted in four sample data sets: title, abstract, caption, and reference. Random sampling resulted in nine sample data sets: from 10% to 90% at 10% intervals. Several strategies were used to compare the distribution of each of the sample data sets with the distribution of the global data set. Each data set was first summarized using two graphs: a plot of ranked document frequency against the rank and a histogram of the document frequency. The relationship

between the two data sets in question was then characterized in three ways: a scatter plot, a Spearman Correlation Coefficient, and a Kullback-Leibler distance between their distributions. Finally, for random sampling data sets, we performed simple linear regression models to predict global IDFs from sample IDFs.

The results suggested that IDFs computed on random sample collections had stronger association with the global IDF than IDFs computed on feature based sample collections. On average, the IDF computed on the 10% random sample could explain about 80% of the variations in IDFs calculated with the global collection. We also noted from scatter plots that high DF terms in the samples (i.e., terms which appeared in a lot of articles in the sample collection) were much more reliable than low DF terms in predicting global DFs.

At the beginning of the study, we were interested in finding out an optimal sampling method which would begin with a collection as small as possible, but generate IDFs close enough to the global IDF that no significant difference can be found in their distributions. This then became a hypothesis testing problem: "can we disprove, to a certain required level of significance, the null hypothesis that two data sets were drawn from the same population distribution function?" It required establishing a difference measure and choosing a test statistic so that we could compare the difference to a critical value of the test statistic distribution.

To achieve this goal, we considered two goodness-of-fit tests: the Chi-square goodness-of-fit test and the Kolmogorov-Smirnov test. The Chi-square test is used to test if a sample of data came from a population with a specific distribution [14]. The Kolmogorov-Smirnov test has a similar purpose, but is based on the maximum distance between two cumulative distribution functions. The test statistics will tell us whether the sample DF follows the same distribution as the global DF.

The Chi-square test could potentially be carried out at three levels. At the first level, we could look at the raw data, i.e., the document frequency of each individual term, and directly compared the distribution of DF_G and DF_Sub. This was like measuring the association between two categorical variables, each having 277,905 categories. The frequency count in each cell was simply the document frequency of the term in that collection.

At the second level, we could look at the summary data of document frequency versus the number of terms with that document frequency. The method put the data into bins, each bin corresponding to a unique value of the document frequency. Then it counted the number of terms that fell into each bin. We can think of this as generating a histogram and forcing each bar to be of unit width. So, the feature space became unique document frequency values (not necessarily consecutive) and the value was the number of terms with that feature.

At the third level, we could form wider bins for either of the first two methods. Binning for the second method was easy to understand and was what statistical software would do by default (choosing an optimal bar width to plot the histogram), but grouping "bag of words" in the raw data, as at the first level, could only be arbitrary.

There were problems with each of the three levels. For the first level, there were two serious problems. Firstly, all the data sets contained significant proportion of low DF terms, which translated into cells with low frequency in the bivariate table. An appropriate Chi-square test would require that expected frequencies in each cell be at least 5 (possibly allowing a few exceptions if there are a large number of categories).

Obviously, our data do not meet the requirement. Secondly, it is argued that a Chi-square test should only be used when observations are independent, i.e., no category or response is dependent upon or influenced by another [15]. This assumption is seriously in jeopardy in our data. It is well known that some words turn to occur together, e.g., forming a phrase, while other words rarely co-occur. The test results would be very unreliable if we ignored this important linguistic phenomenon and assumed that term frequencies were independent. A direct consequence of working with dependent data was the difficulty in choosing the appropriate number of degrees of freedom, which was determined by the number of independent observations in the data.

The second and the third methods, both looking at distribution of data, instead of the raw data, introduced another problem. To give a simple example, assume there were two terms in the vocabulary. In the global collection, Term 1 appeared in 100 documents ($DF\_Global_1=100$) and Term 2 appeared in 10 documents ($DF\_Global_2=10$). In one sub collection (SubOne), $DF\_SubOne_1=9$ and $DF\_SubOne_2=1$. In another sub collection (SubTwo), $DF\_SubTwo_1=1$ and $DF\_SubTwo_2=9$. If we looked at the raw data, we would probably conclude that the SubOne collection had closer document frequency distribution with the global collection than the SubTwo collection. However, if we only looked at the summary data, the histograms of the two sub collections were identical. What was missing in the summary data was the mapping of terms between the global collection and the sub-sampling collection. Even if we found that the summaries of two data sets had same or similar distributions, we could not know for sure if the raw data had similar distributions.

Compared with the first and the second method, the third method should eliminate low count cells, but as in any other statistical methods, binning involves a loss of information. Plus, as we mentioned above, binning based on the raw data was very arbitrary.

Due to these problems, we felt that it was inappropriate to use Chi-square tests on the current data. The Kolmogorov-Smirnov test would be inappropriate either due to the dependency of terms. The Kolmogorov-Smirnov test is based on cumulative density functions. Since the probabilities of some terms are dependent on each other, the cumulative density functions generated from the data are unreliable.

To address the term dependence problem, we will consider in a future study using the principal component analysis (PCA) technique to project the original term space into an orthogonal space and use the principal components (linear combination of terms) to compute correlation coefficients and relative entropies, and perform goodness-of-fit tests. Analyses in that space should be more reliable than what has been done with the current data sets.

We are also aware of other possibilities for a future study. Instead of using simple random sampling to form the random sample collections, we can consider stratified sampling methods such that the 10% sample collection is formed by 10% of articles from Journal 1, 10% of articles from Journal 2, etc. This would account for the possibility that language patterns are different across journals.

This study only considered single terms. Taking phrases into consideration in a future study may better model the language use. In this study, we noted that terms that appeared in a lot of articles in the sample collection were more reliable than terms that

appeared in a few articles in predicting global DFs. It would be interesting to follow up this with a term level analysis and see if there are certain terms that are more or less stable.

Acknowledgments. We thank Drs. Chuanshu Ji, Cathy Blake and Haipeng Shen for their guidance on this work and the anonymous reviewers for their feedback.

References

1. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5), 513–523 (1988)
2. Salton, G.: Automatic information organization and retrieval. McGraw-Hill, New York (1968)
3. Spärck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28, 11–21 (1972)
4. Robertson, S.: Understanding inverse document frequency: on theoretical arguments for IDF. Journal of Documentation 60, 503–520 (2004)
5. Wang, J., Rölleke, T.: Context-specific frequencies and discriminativeness for the retrieval of structured documents. In: Lalmas, M., MacFarlane, A., Rüger, S.M., Tombros, A., Tsikrika, T., Yavlinsky, A. (eds.) ECIR 2006. LNCS, vol. 3936, pp. 579–582. Springer, Heidelberg (2006)
6. Blake, C.: A Comparison of document, sentence and term event spaces. In: Coling & ACL joint conference, Sydney, Australia (2006)
7. Church, K.W., Gale, W.A.: Poisson mixtures. Natural Language Engineering 1(2), 163–190 (1995)
8. Newman, M.E.J.: Power laws, Pareto distributions and Zipf's law (2005), Available at: http://aps.arxiv.org/PS_cache/cond-mat/pdf/0412/0412004.pdf
9. Oard, D., Marchionini, G.: A conceptual framework for text filtering (1996), Available at: http://hcil.cs.umd.edu/trs/96-10/node10.html#SECTION00051000000000000000
10. McCallum, A.K.: Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering (1996), Available at: http://www.cs.cmu.edu/_mccallum/bow
11. Porter, M.F.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)
12. Spärck Jones, K.: IDF term weighting and IR research lessons. Journal of Documentation 60, 521–523 (2004)
13. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley, New York (1991)
14. Snedecor, G.W., Cochran, W.G.: Statistical Methods. Iowa State University Press, Ames (1989)
15. Conner-Linton, J.: Chi square tutorial (2003), Available at: http://www.georgetown.edu/faculty/ballc/webtools/web_chi_tut.html