A Comparison of Document, Sentence, and Term Event Spaces

Catherine Blake
School of Information and Library Science
University of North Carolina at Chapel Hill
North Carolina, NC 27599-3360
cablake@email.unc.edu

Abstract

The trend in information retrieval systems is from document to sub-document retrieval, such as sentences in a summarization system and words or phrases in question-answering system. Despite this trend, systems continue to model language at a document level using the inverse document frequency (IDF). In this paper, we compare and contrast IDF with inverse sentence frequency (ISF) and inverse term frequency (ITF). A direct comparison reveals that all language models are highly correlated; however, the average ISF and ITF values are 5.5 and 10.4 higher than IDF. All language models appeared to follow a power law distribution with a slope coefficient of 1.6 for documents and 1.7 for sentences and terms. We conclude with an analysis of IDF stability with respect to random, journal, and section partitions of the 100,830 full-text scientific articles in our experimental corpus.

1 Introduction

The vector based information retrieval model identifies relevant documents by comparing query terms with terms from a document corpus. The most common corpus weighting scheme is the term frequency (TF) x inverse document frequency (IDF), where TF is the number of times a term appears in a document, and IDF reflects the distribution of terms within the corpus (Salton and Buckley, 1988). Ideally, the system should assign the highest weights to terms with the most discriminative power.

One component of the corpus weight is the language model used. The most common language model is the **Inverse Document Frequency (IDF)**, which considers the distribution of terms between documents (see equation (1)). IDF has played a central role in retrieval systems since it was first introduced more than thirty years ago (Sparck Jones, 1972).

$$IDF(t_i) = \log_2(N) - \log_2(n_i) + 1 \quad (1)$$

N is the total number of corpus documents; n_i is the number of documents that contain at least one occurrence of the term t_i; and t_i is a term, which is typically stemmed.

Although information retrieval systems are trending from document to sub-document retrieval, such as sentences for summarization and words, or phrases for question answering, systems continue to calculate corpus weights on a language model of documents. Logic suggests that if a system identifies sentences rather than documents, it should use a corpus weighting scheme based on the number of sentences rather than the number documents. That is, the system should replace IDF with the **Inverse Sentence Frequency (ISF)**, where N in (1) is the total number of sentences and n_i is the number of sentences with term i. Similarly, if the system retrieves terms or phrases then IDF should be replaced with the **Inverse Term Frequency (ITF)**, where N in (1) is the vocabulary size, and n_i is the number of times a term or phrases appears in the corpus. The challenge is that although document language models have had unprecedented empirical success, language models based on a sentence or term do not appear to work well (Robertson, 2004).

Our goal is to explore the transition from the document to sentence and term spaces, such that we may uncover where the language models start

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 601–608,
Sydney, July 2006. ©2006 Association for Computational Linguistics

to break down. In this paper, we explore this goal by answering the following questions: How correlated are the raw document, sentence, and term spaces? How correlated are the IDF, ISF, and ITF values? How well does each language models conform to Zipf's Law and what are the slope coefficients? How sensitive is IDF with respect to sub-sets of a corpus selected at random, from journals, or from document sections including the abstract and body of an article?

This paper is organized as follows: Section 2 provides the theoretical and practical implications of this study; Section 3 describes the experimental design we used to study document, sentence, and term, spaces in our corpora of more than one-hundred thousand full-text documents; Section 4 discusses the results; and Section 5 draws conclusions from this study.

2 Background and Motivation

The transition from document to sentence to term spaces has both theoretical and practical ramifications. From a theoretical standpoint, the success of TFxIDF is problematic because the model combines two different event spaces – the space of *terms* in TF and of *documents* in IDF. In addition to resolving the discrepancy between event spaces, the foundational theories in information science, such as Zipf's Law (Zipf, 1949) and Shannon's Theory (Shannon, 1948) consider only a term event space. Thus, establishing a direct connection between the empirically successful IDF and the theoretically based ITF may enable a connection to previously adopted information theories.

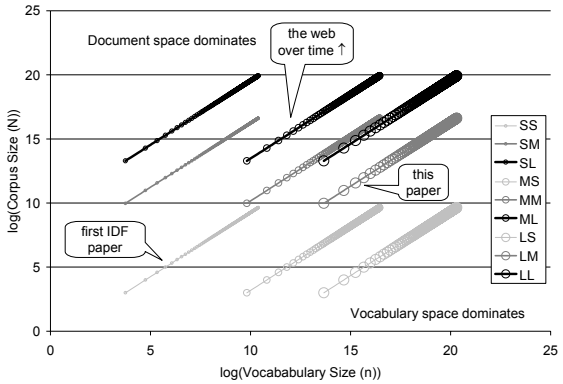

Figure 1. Synthetic data showing IDF trends for different sized corpora and vocabulary.

Understanding the relationship among document, sentence and term spaces also has practical importance. The size and nature of text corpora has changed dramatically since the first IDF experiments. Consider the synthetic data shown in Figure 1, which reflects the increase in both vocabulary and corpora size from small (S), to medium (M), to large (L). The small vocabulary size is from the Cranfield corpus used in Sparck Jones (1972), medium is from the 0.9 million terms in the Heritage Dictionary (Pickett 2000) and large is the 1.3 million terms in our corpus. The small number of documents is from the Cranfield corpus in Sparck Jones (1972), medium is 100,000 from our corpus, and large is 1 million

As a document corpus becomes sufficiently large, the rate of new terms in the vocabulary decreases. Thus, in practice the rate of growth on the x-axis of Figure 1 will slow as the corpus size increases. In contrast, the number of documents (shown on the y-axis in Figure 1) remains unbounded. It is not clear which of the two components in equation (1), the $\log_2(N)$, which reflects the number of documents, or the $\log_2(n_i)$, which reflects the distribution of terms between documents within the corpus will dominate the equation. Our strategy is to explore these differences empirically.

In addition to changes in the vocabulary size and the number of documents, the average number of terms per document has increased from 7.9, 12.2 and 32 in Sparck Jones (1972), to 20 and 32 in Salton and Buckley (1988), to 4,981 in our corpus. The transition from abstracts to full-text documents explains the dramatic difference in document length; however, the impact with respect to the distribution of terms and motivates us to explore differences between the language used in an abstract, and that used in the body of a document.

One last change from the initial experiments is a trend towards an on-line environment, where calculating IDF is prohibitively expensive. This suggests a need to explore the stability of IDF so that system designers can make an informed decision regarding how many documents should be included in the IDF calculations. We explore the stability of IDF in random, journal, and document section sub-sets of the corpus.

3 Experimental Design

Our goal in this paper is to compare and contrast language models based on a document with those based on a sentence and term event spaces. We considered several of the corpora from the Text Retrieval Conferences (TREC, trec.nist.gov); however, those collections were primarily news

articles. One exception was the recently added genomics track, which considered full-text scientific articles, but did not provide relevance judgments at a sentence or term level. We also considered the sentence level judgments from the novelty track and the phrase level judgments from the question-answering track, but those were news and web documents respectively and we had wanted to explore the event spaces in the context of scientific literature.

Table 1 shows the corpus that we developed for these experiments. The American Chemistry Society provided 103,262 full-text documents, which were published in 27 journals from 2000-2004[1]. We processed the headings, text, and tables using Java BreakIterator class to identify sentences and a Java implementation of the Porter Stemming algorithm (Porter, 1980) to identify terms. The inverted index was stored in an Oracle 10i database.

	Docs		Avg	Tokens	
Journal	**#**	**%**	**Length**	**Million**	**%**
ACHRE4	548	0.5	4923	2.7	1
ANCHAM	4012	4.0	4860	19.5	4
BICHAW	8799	8.7	6674	58.7	11
BIPRET	1067	1.1	4552	4.9	1
BOMAF6	1068	1.1	4847	5.2	1
CGDEFU	566	0.5	3741	2.1	<1
CMATEX	3598	3.6	4807	17.3	3
ESTHAG	4120	4.1	5248	21.6	4
IECRED	3975	3.9	5329	21.2	4
INOCAJ	5422	5.4	6292	34.1	6
JACSAT	14400	14.3	4349	62.6	12
JAFCAU	5884	5.8	4185	24.6	5
JCCHFF	500	0.5	5526	2.8	1
JCISD8	1092	1.1	4931	5.4	1
JMCMAR	3202	3.2	8809	28.2	5
JNPRDF	2291	2.2	4144	9.5	2
JOCEAH	7307	7.2	6605	48.3	9
JPCAFH	7654	7.6	6181	47.3	9
JPCBFK	9990	9.9	5750	57.4	11
JPROBS	268	0.3	4917	1.3	<1
MAMOBX	6887	6.8	5283	36.4	7
MPOHBP	58	0.1	4868	0.3	<1
NALEFD	1272	1.3	2609	3.3	1
OPRDFK	858	0.8	3616	3.1	1
ORLEF7	5992	5.9	1477	8.8	2
Total	100,830			526.6	
Average	4,033	4.0	4,981	21.1	
Std Dev	3,659	3.6	1,411	20.3	

Table 1. Corpus summary.

[1] Formatting inconsistencies precluded two journals and reduced the number of documents by 2,432.

We made the following comparisons between the document, sentence, and term event spaces.

(1) Raw term comparison

A set of well-correlated spaces would enable an accurate prediction from one space to the next. We will plot pair-wise correlations between each space to reveal similarities and differences.

This comparison reflects a previous analysis comprising a random sample of 193 words from a 50 million word corpus of 85,432 news articles (Church and Gale 1999). Church and Gale's analysis of term and document spaces resulted in a p value of -0.994. Our work complements their approach by considering full-text scientific articles rather than news documents, and we consider the entire stemmed term vocabulary in a 526 million-term corpus.

(2) Zipf Law comparison

Information theory tells us that the frequency of terms in a corpus conforms to the power law distribution K/j^{θ} (Baeza-Yates and Ribeiro-Neto 1999). Zipf's Law is a special case of the power law, where θ is close to 1 (Zipf, 1949). To provide another perspective of the alternative spaces, we calculated the parameters of Zipf's Law, K and θ for each event space and journal using the binning method proposed in (Adamic 2000). By accounting for K, the slope as defined by θ will provide another way to characterize differences between the document, sentence and term spaces. We expect that all event spaces will conform to Zipf's Law.

(3) Direct IDF, ISF, and ITF comparison

The $\log_2(N)$ and $\log_2(n_i)$ should allow a direct comparison between IDF, ISF and ITF. Our third experiment was to provide pair-wise comparisons among these the event spaces.

(4) Abstract versus full-text comparison

Language models of scientific articles often consider only abstracts because they are easier to obtain than full-text documents. Although historically difficult to obtain, the increased availability of full-text articles motivates us to understand the nature of language within the body of a document. For example, one study found that full-text articles require weighting schemes that consider document length (Kamps, et al, 2005). However, controlling the weights for document lengths may hide a systematic difference between the language used in abstracts and the language used in the body of a document. For example, authors may use general language in an

abstract and technical language within a document.

Transitioning from abstracts to full-text documents presents several challenges including how to weigh terms within the headings, figures, captions, and tables. Our forth experiment was to compare IDF between the abstract and full text of the document. We did not consider text from headings, figures, captions, or tables.

(5) IDF Sensitivity

In a dynamic environment such as the Web, it would be desirable to have a corpus-based weight that did not change dramatically with the addition of new documents. An increased understanding of IDF stability may enable us to make specific system recommendations such as if the collection increases by more than n% then update the IDF values.

To explore the sensitivity we compared the amount of change in IDF values for various subsets of the corpus. IDF values were calculated using samples of 10%, 20%, …, 90% and compared with the global IDF. We stratified sampling such that the 10% sample used term frequencies in 10% of the ACHRE4 articles, 10% of the BICHAW articles, etc. To control for variations in the corpus, we repeated each sample 10 times and took the average from the 10 runs. To explore the sensitivity we compared the global IDF in Equation 1 with the local sample, where N was the average number of documents in the sample and n_i was the average term frequency for each stemmed term in the sample.

In addition to exploring sensitivity with respect to a random subset, we were interested in learning more about the relationship between the global IDF and the IDF calculated on a journal sub-set. To explore these differences, we compared the global IDF with local IDF where N was the number of documents in each journal and n_i was the number of times the stemmed term appears in the text of that journal.

4 Results and Discussion

The 100830 full text documents comprised 2,001,730 distinct unstemmed terms, and 1,391,763 stemmed terms. All experiments reported in this paper consider stemmed terms.

4.1 Raw frequency comparison

The dimensionality of the document, sentence, and terms spaces varied greatly, with 100830 documents, 16.5 million sentences, and 2.0 million distinct unstemmed terms (526.0 million in total), and 1.39 million distinct stemmed terms. Figure 2A shows the correlation between the frequency of a term in the document space (x) and the average frequency of the same set of terms in the sentence space (y). For example, the average number of sentences for the set of terms that appear in 30 documents is 74.6. Figure 2B compares the document (x) and average term freq-

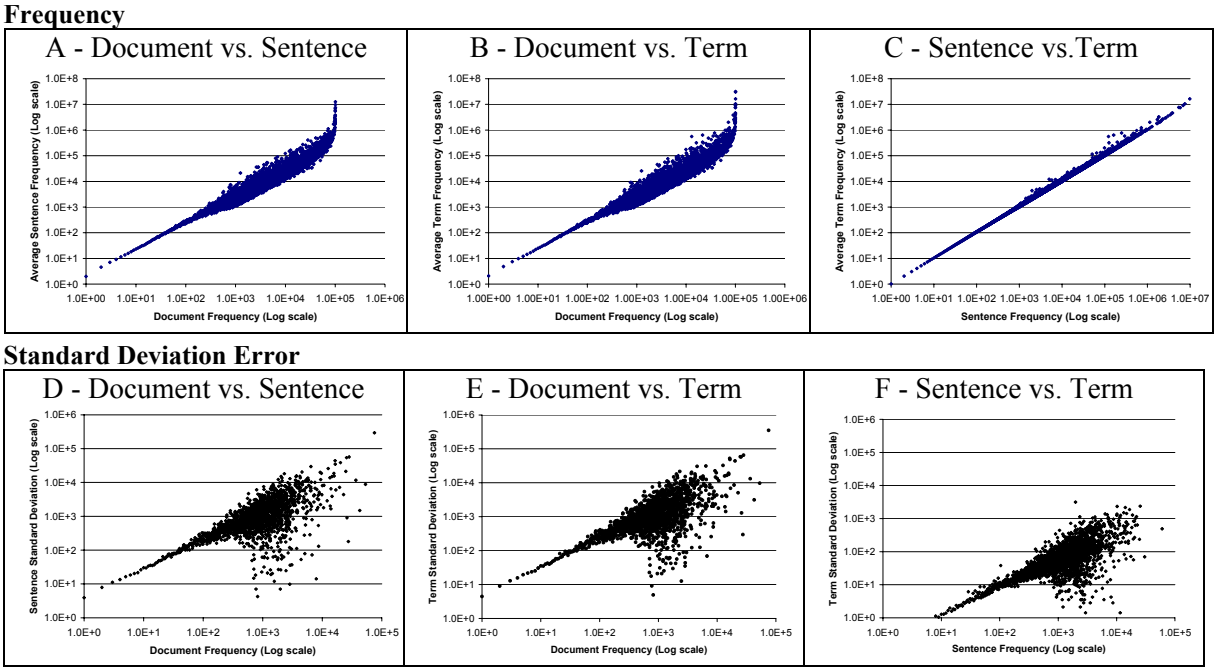

Figure 2. Raw frequency correlation between document, sentence, and term spaces.

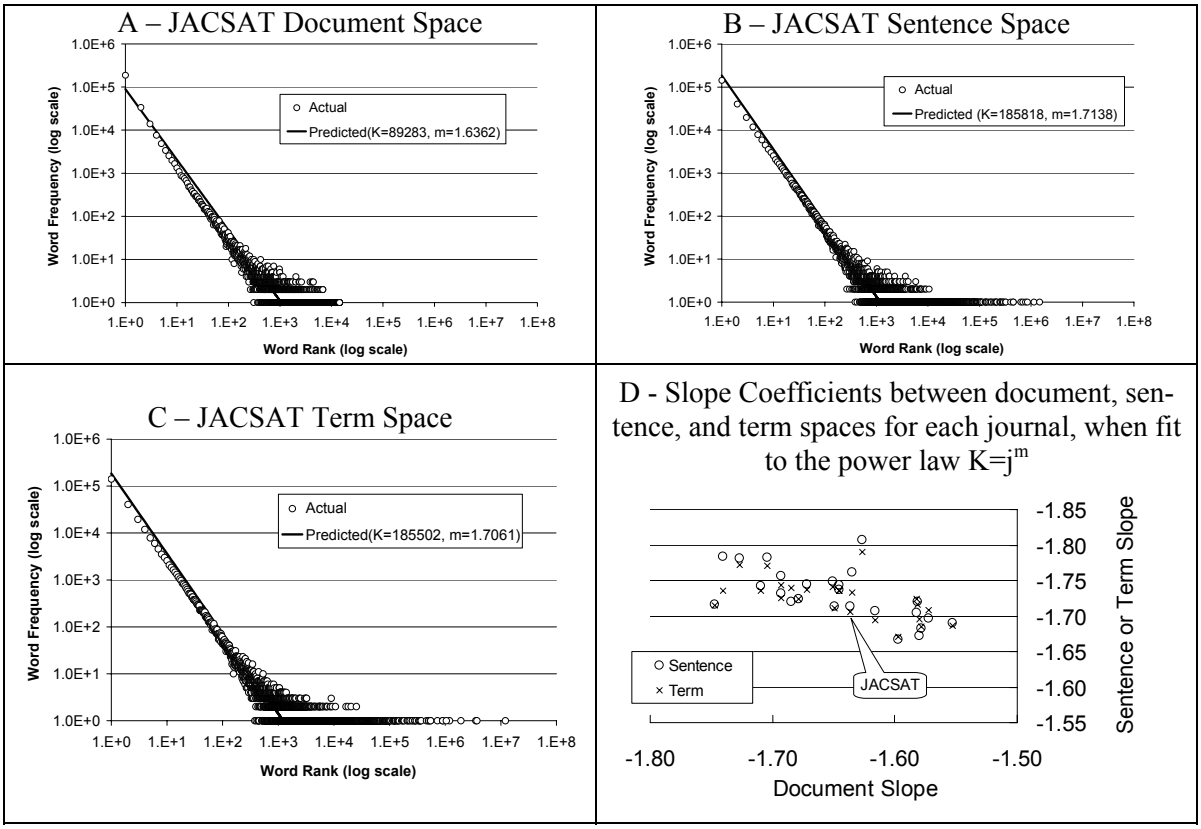

Figure 3. Zipf's Law comparison. A through C show the power law distribution for the journal JACSAT in the document (A), sentence (B), and term (C) event spaces. Note the predicted slope coefficients of 1.6362, 1.7138 and 1.7061 respectively). D shows the document, sentence, and term slope coefficients for each of the 25 journals when fit to the power law $K=j^m$, where j is the rank.

quency (y) These figures suggest that the document space differs substantially from the sentence and term spaces. Figure 2C shows the sentence frequency (x) and average term frequency (y), demonstrating that the sentence and term spaces are highly correlated.

Luhn proposed that if terms were ranked by the number of times they occurred in a corpus, then the terms of interest would lie within the center of the ranked list (Luhn 1958). Figures 2D, E and F show the standard deviation between the document and sentence space, the document and term space and the sentence and term space respectively. These figures suggest that the greatest variation occurs for important terms.

4.2 Zipf's Law comparison

Zipf's Law states that the frequency of terms in a corpus conforms to a power law distribution K/j^{θ} where θ is close to 1 (Zipf, 1949). We calculated the K and θ coefficients for each journal and language model combination using the binning method proposed in (Adamic, 2000). Figures 3A-C show the actual frequencies, and the power law fit for the each language model in just one of the 25 journals (jacsat). These and the remaining 72 figures (not shown) suggest that Zipf's Law holds in all event spaces.

Zipf Law states that θ should be close to -1. In our corpus, the average θ in the document space was -1.65, while the average θ in both the sentence and term spaces was -1.73.

Figure 3D compares the document slope (x) coefficient for each of the 25 journals with the sentence and term spaces coefficients (y). These findings are consistent with a recent study that suggested θ should be closer to 2 (Cancho 2005). Another study found that term frequency rank distribution was a better fit Zipf's Law when the term space comprised both words and phrases (Ha et al, 2002). We considered only stemmed terms. Other studies suggest that a Poisson mixture model would better capture the frequency rank distribution than the power model (Church and Gale, 1995). A comprehensive overview of using Zipf's Law to model language can be found in (Guiter and Arapov, 1982).

4.3 Direct IDF, ISF, and ITF comparison

Our third experiment was to compare the three language models directly. Figure 4A shows the average, minimum and maximum ISF value for each rounded IDF value. After fitting a regression line, we found that ISF correlates well with IDF, but that the average ISF values are 5.57 greater than the corresponding IDF. Similarly, ITF correlates well with IDF, but the ITF values are 10.45 greater than the corresponding IDF.

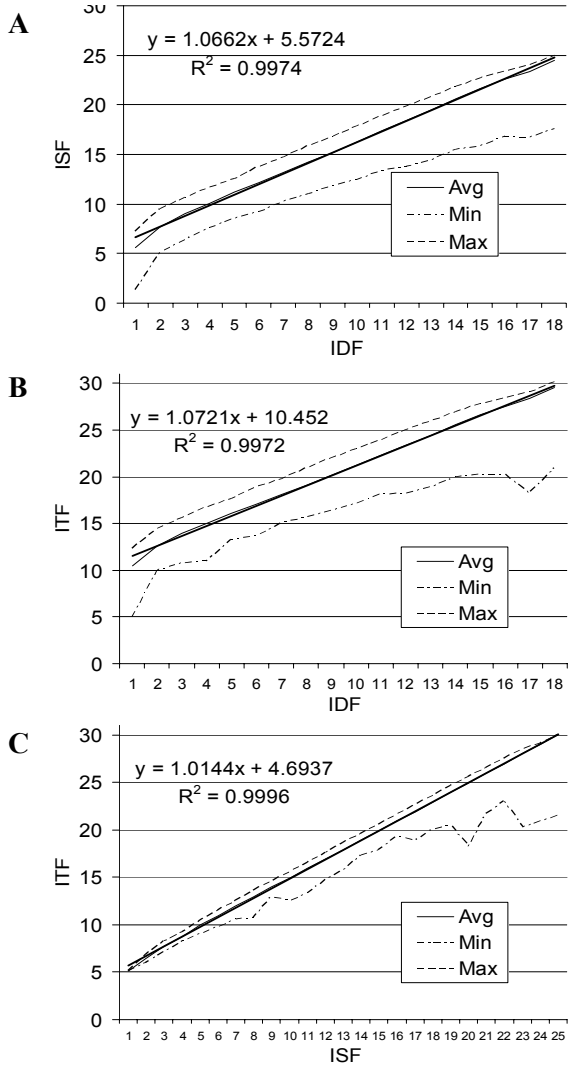

Figure 4. Pair-wise IDF, ISF, and ITF comparisons.

It is little surprise that Figure 4C reveals a strong correlation between ITF and ISF, given the correlation between raw frequencies reported in section 4.1. Again, we see a high correlation between the ISF and ITF spaces but that the ITF values are on average 4.69 greater than the equivalent ISF value. These findings suggests that simply substituting ISF or ITF for IDF would result in a weighting scheme where the corpus weights would dominate the weights assigned to query in the vector based retrieval model. The variation appears to increase at higher IDF values.

Table 2 (see over) provides example stemmed terms with varying frequencies, and their corresponding IDF, ISF and ITF weights. The most frequent term "*the"*, appears in 100717 documents, 12,771,805 sentences and 31,920,853 times. In contrast, the stemmed term "*electrochem"* appeared in only six times in the corpus, in six different documents, and six different sentences. Note also the differences between abstracts, and the full-text IDF (see section 4.4).

4.4 Abstract vs full text comparison

Although abstracts are often easier to obtain, the availability of full-text documents continues to increase. In our fourth experiment, we compared the language used in abstracts with the language used in the full-text of a document. We compared the abstract and non-abstract terms in each of the three language models.

Not all of the documents distinguished the abstract from the body. Of the 100,830 documents, 92,723 had abstracts and 97,455 had sections other than an abstract. We considered only those documents that differentiated between sections. Although the number of documents did not differ greatly, the vocabulary size did. There were 214,994 terms in the abstract vocabulary and 1,337,897 terms in the document body, suggesting a possible difference in the distribution of terms, the $\log(n_i)$ component of IDF.

Figure 5 suggests that language used in an abstract differs from the language used in the body of a document. On average, the weights assigned to stemmed terms in the abstract were higher than the weights assigned to terms in the body of a document (space limitations preclude the inclusion of the ISF and ITF figures).

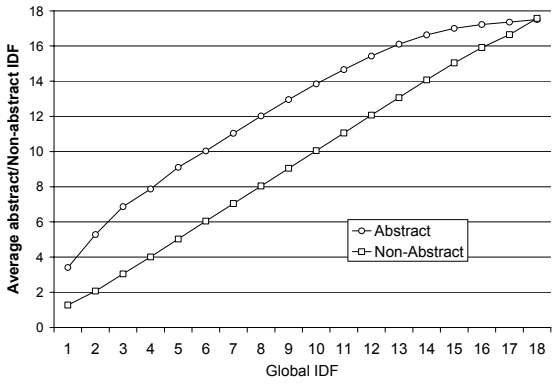

Figure 5. Abstract and full-text IDF compared with global IDF.

	Document (IDF)			Sentence (ISF)			Term (ITF)		
Word	**Abs**	**NonAbs**	**All**	**Abs**	**NonAbs**	**All**	**Abs**	**NonAbs**	**All**
the	1.014	1.004	1.001	1.342	1.364	1.373	4.604	9.404	5.164
chemist	11.074	5.957	5.734	13.635	12.820	12.553	22.838	17.592	17.615
synthesis	14.331	11.197	10.827	17.123	18.000	17.604	26.382	22.632	22.545
eletrochem	17.501	15.251	15.036	20.293	22.561	22.394	29.552	26.965	27.507

Table 2. Examples of IDF, ISF and ITF for terms with increasing IDF.

4.5 IDF sensitivity

The stability of the corpus weighting scheme is particularly important in a dynamic environment such as the web. Without an understanding of how IDF behaves, we are unable to make a principled decision regarding how often a system should update the corpus-weights.

To measure the sensitivity of IDF we sampled at 10% intervals from the global corpus as outlined in section 3. Figure 6 compares the global IDF with the IDF from each of the 10% samples. The 10% samples are almost indiscernible from the global IDF, which suggests that IDF values are very stable with respect to a random subset of articles. Only the 10% sample shows any visible difference from the global IDF values, and even then, the difference is only noticeable at higher global IDF values (greater than 17 in our corpus).

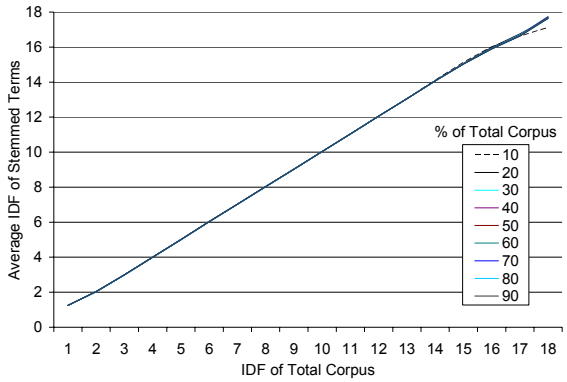

Figure 6 – Global IDF vs random sample IDF.

In addition to a random sample, we compared the global based IDF with IDF values generated from each journal (in an on-line environment, it may be pertinent to partition pages into academic or corporate URLs or to calculate term frequencies for web pages separately from blog and wikis). In this case, N in equation (1) was the number of documents in the journal and n_i was the distribution of terms within a journal.

If the journal vocabularies were independent, the vocabulary size would be 4.1 million for unstemmed terms and 2.6 million for stemmed terms. Thus, the journals shared 48% and 52% of their vocabulary for unstemmed and stemmed terms respectively.

Figure 7 shows the result of this comparison and suggests that the average IDF within a journal differed greatly from the global IDF value, particularly when the global IDF value exceeds five. This contrasts sharply with the random samples shown in Figure 6.

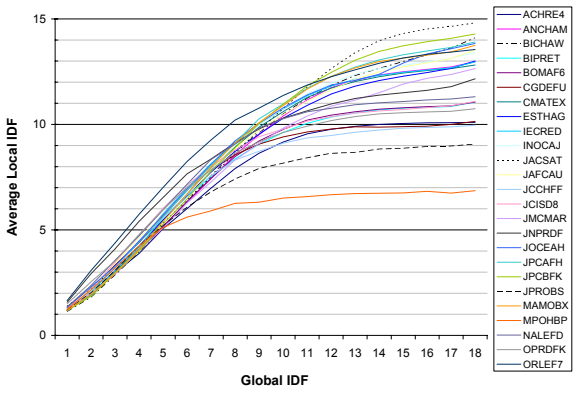

Figure 7 – Global IDF vs local journal IDF.

At first glance, the journals with more articles appear to correlated more with the global IDF than journals with fewer articles. For example, JACSAT has 14,400 documents and is most correlated, while MPOHBP with 58 documents is least correlated. We plotted the number of articles in each journal with the mean squared error (figure not shown) and found that journals with fewer than 2,000 articles behave differently to journals with more than 2,000 articles; however, the relationship between the number of articles in the journal and the degree to which the language in that journal reflects the language used in the entire collection was not clear.

5 Conclusions

We have compared the document, sentence, and term spaces along several dimensions. Results from our corpus of 100,830 full-text scientific articles suggest that the difference between these alternative spaces is both theoretical and practi-

cal in nature. As users continue to demand information systems that provide sub-document retrieval, the need to model language at the sub-document level becomes increasingly important. The key findings from this study are:

(1) The raw document frequencies are considerably different to the sentence and term frequencies. The lack of a direct correlation between the document and sub-document raw spaces, in particular around the areas of important terms, suggest that it would be difficult to perform a linear transformation from the document to a sub-document space. In contrast, the raw term frequencies correlate well with the sentence frequencies.

(2) IDF, ISF and ITF are highly correlated; however, simply replacing IDF with the ISF or ITF would result in a weighting scheme where the corpus weight dominated the weights assigned to query and document terms.

(3) IDF was surprisingly stable with respect to random samples at 10% of the total corpus. The average IDF values based on only a 20% random stratified sample correlated almost perfectly to IDF values that considered frequencies in the entire corpus. This finding suggests that systems in a dynamic environment, such as the Web, need not update the global IDF values regularly (see (4)).

(4) In contrast to the random sample, the journal based IDF samples did not correlate well to the global IDF. Further research is required to understand these factors that influence language usage.

(5) All three models (IDF, ISF and ITF) suggest that the language used in abstracts is systematically different from the language used in the body of a full-text scientific document. Further research is required to understand how well the abstract tested corpus-weighting schemes will perform in a full-text environment.

References

Lada A. Adamic 2000 Zipf, Power-laws, and Pareto - a ranking tutorial. [Available from http://www.parc.xerox.com/istl/groups/iea/papers/ranking/ranking.html]

Ricardo Baeza-Yates, and Berthier Ribeiro-Neto 1999 *Modern Information Retrieval*: Addison Wesley.

Cancho, R. Ferrer 2005 The variation of Zipfs Law in human language. The European Physical Journal B 44 (2):249-57.

Kenneth W Church and William A. Gale 1999 Inverse document frequency: a measure of deviations from Poisson. *NLP using very large corpora*, Kluwer Academic Publishers.

Kenneth W Church.and William A. Gale 1995 Poisson mixtures. *Natural Language Engineering,* 1 (2):163-90.

H. Guiter and M Arapov 1982. Editors *Studies on Zipf's Law*. Brochmeyer, Bochum.

Jaap Kamps, Maarten De Rijke, and Borkur Sigurbjornsson 2005 The Importance of lenght normalization for XML retrieval. *Information Retrieval* 8:631-54.

Le Quan Ha, E.I. Sicilia-Garcia, Ji Ming, and F.J. Smith 2002 Extension of Zipf's Law to words and phrases. *19th International Conference on Computational linguistics*.

Hans P. Luhn 1958 The automatic creation of literature abstracts *IBM Journal of Research and Development* 2 (1):155-64.

Joseph P Pickett et al. 2000 *The American Heritage® Dictionary of the English Language*. Fourth edition. Edited by H. Mifflin.

Martin F. Porter 1980 An Algorithm for Suffix Stripping. *Program*, *14* (3). 130-137.

Stephen Robertson 2004 Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation* 60 (5):503-520.

Gerard Salton and Christopher Buckley 1988 Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24 (5):513-23.

Claude E. Shannon 1948 A Mathematical Theory of Communication *Bell System Technical Journal*. 27 379–423 & 623–656.

Karen Sparck Jones, Steve Walker, and Stephen Robertson 2000 A probabilistic model of information retrieval: development and comparative experiments Part 1. *Information Processing & Management*, 36:779-808.

Karen Sparck Jones 1972 A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11-21.

George Kingsley Zipf 1949 *Human behaviour and the principle of least effort. An introduction to human ecology,* 1st edn. Edited by Addison-Wesley. Cambridge, MA.