Using Sketches to Estimate Associations

Ping Li
Department of Statistics
Stanford University
Stanford, California 94305
pingli@stat.stanford.edu

Kenneth W. Church
Microsoft Research
One Microsoft Way
Redmond, Washington 98052
church@microsoft.com

Abstract

We should not have to look at the entire corpus (e.g., the Web) to know if two words are associated or not.[1] A powerful sampling technique called *Sketches* was originally introduced to remove duplicate Web pages. We generalize sketches to estimate contingency tables and associations, using a maximum likelihood estimator to find the most likely contingency table given the sample, the margins (document frequencies) and the size of the collection. Not unsurprisingly, computational work and statistical accuracy (variance or errors) depend on sampling rate, as will be shown both theoretically and empirically. Sampling methods become more and more important with larger and larger collections. At Web scale, sampling rates as low as 10^{-4} may suffice.

1 Introduction

Word associations (co-occurrences) have a wide range of applications including: Speech Recognition, Optical Character Recognition and Information Retrieval (IR) (Church and Hanks, 1991; Dunning, 1993; Manning and Schutze, 1999). It is easy to compute association scores for a small corpus, but more challenging to compute lots of scores for lots of data (e.g. the Web), with billions of web pages (D) and millions of word types (V). For a small corpus, one could compute pair-wise associations by multiplying the (0/1) term-by-document matrix with its transpose (Deerwester et al., 1999). But this is probably infeasible at Web scale.

Approximations are often good enough. We should not have to look at every document to determine that two words are strongly associated. A number of sampling-based randomized algorithms have been implemented at Web scale (Broder, 1997; Charikar, 2002; Ravichandran et al., 2005).[2]

A conventional random sample is constructed by selecting D_s documents from a corpus of D documents. The (corpus) sampling rate is $\frac{D_s}{D}$. Of course, word distributions have long tails. There are a few high frequency words and many low frequency words. It would be convenient if the sampling rate could vary from word to word, unlike conventional sampling where the sampling rate is fixed across the vocabulary. In particular, in our experiments, we will impose a floor to make sure that the sample contains at least 20 documents for each term. (When working at Web scale, one might raise the floor somewhat to perhaps 10^4.)

Sampling is obviously helpful at the top of the frequency range, but not necessarily at the bottom (especially if frequencies fall below the floor). The question is: how about "ordinary" words? To answer this question, we randomly picked 15 pages from a Learners' dictionary (Hornby, 1989), and selected the first entry on each page. According to Google, there are 10 million pages/word (median value, aggregated over the 15 words), no where near the floor.

Sampling can make it possible to work in memory, avoiding disk. At Web scale ($D \approx 10$ billion pages), inverted indexes are large (1500 GBs/billion pages)[3], probably too large for memory. But a sample is more manageable; the inverted index for a 10^{-4} sample of the entire web could fit in memory on a single PC (1.5 GB).

[1] This work was conducted at Microsoft while the first author was an intern. The authors thank Chris Meek, David Heckerman, Robert Moore, Jonathan Goldstein, Trevor Hastie, David Siegmund, Art Own, Robert Tibshirani and Andrew Ng.

[2] *http://labs.google.com/sets* produces fascinating sets, although we don't know how it works. Given the seeds, "America" and "China," *http://labs.google.com/sets* returns: "America, China, Japan, India, Italy, Spain, Brazil, Persia, Europe, Australia, France, Asia, Canada."

[3] This estimate is extrapolated from Brin and Page (1998), who report an inverted index of 37.2 GBs for 24 million pages.

Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 708–715, Vancouver, October 2005.

Table 1: The number of intermediate results after the first join can be reduced from 504,000 to 120,000, by starting with "Schwarzenegger & Austria" rather than the baseline ("Schwarzenegger & Terminator"). The standard practice of starting with the two least frequent terms is a good rule of thumb, but one can do better, given (estimates of) joint frequencies.

Query	Hits (Google)
Austria	88,200,000
Governor	37,300,000
Schwarzenegger	4,030,000
Terminator	3,480,000
Governor & Schwarzenegger	1,220,000
Governor & Austria	708,000
Schwarzenegger & Terminator	504,000
Terminator & Austria	171,000
Governor & Terminator	132,000
Schwarzenegger & Austria	120,000

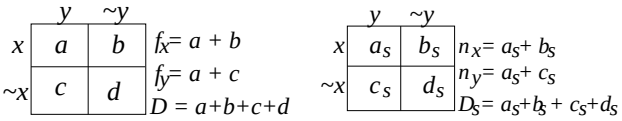
1.1 An Application: The Governator

Google returns the top k hits, plus an estimate of how many hits there are. Table 1 shows the number of hits for four words and their pair-wise combinations. Accurate estimates of associations would have applications in Database query planning (Garcia-Molina et al., 2002). Query optimizers construct a plan to minimize a cost function (e.g., intermediate writes). The optimizer could do better if it could estimate a table like Table 1. But efficiency is important. We certainly don't want to spend more time optimizing the plan than executing it.

Suppose the optimizer wanted to construct a plan for the query: "Governor Schwarzenegger Terminator Austria." The standard solution starts with the two least frequent terms: "Schwarzenegger" and "Terminator." That plan generates 504,000 intermediate writes after the first join. An improvement starts with "Schwarzenegger" with "Austria," reducing the 504,000 down to 120,000.

In addition to counting hits, Table 1 could also help find the top k pages. When joining the first pair of terms, we'd like to know how far down the ranking we should go. Accurate estimates of associations would help the optimizer make such decisions.

It is desirable that estimates be consistent, as well as accurate. Google, for example, reports 6 million hits for "America, China, Britain," and 23 million for "America, China, Britain, Japan." Joint frequencies decrease monotonically: $s \subset S \implies \text{hits}(s) \geq \text{hits}(S)$.



(a) (b)

Figure 1: (a): A contingency table for word x and word y. Cell a is the number of documents that contain both x and y, b is the number that contain x but not y, c is the number that contain y but not x, and d is the number that contain neither x nor y. The margins, $f_x = a + b$ and $f_y = a + c$ are known as document frequencies in IR. D is the total number of documents in the collection. (b): A sample contingency table, with "s" indicating the *sample space*.

1.2 Sampling and Estimation

Two-way associations are often represented as two-way contingency tables (Figure 1(a)). Our task is to construct a sample contingency table (Figure 1(b)), and estimate 1(a) from 1(b). We will use a maximum likelihood estimator (MLE) to find the most likely contingency table, given the sample and various other constraints. We will propose a sampling procedure that bridges two popular choices: (A) sampling over documents and (B) sampling over postings. The estimation task is straightforward and well-understood for (A). As we consider more flexible sampling procedures such as (B), the estimation task becomes more challenging.

Flexible sampling procedures are desirable. Many studies focus on rare words (Dunning, 1993; Moore, 2004); butterflies are more interesting than moths. The sampling rate can be adjusted on a word-by-word basis with (B), but not with (A). The sampling rate determines the trade-off between computational work and statistical accuracy.

We assume a standard inverted index. For each word x, there are a set of postings, X. X contains a set of document IDs, one for each document containing x. The size of postings, $f_x = |X|$, corresponds to the margins of the contingency tables in Figure 1(a), also known as document frequencies in IR.

The postings lists are approximated by *sketches*, skX, first introduced by Broder (1997) for removing duplicate web pages. Assuming that document IDs are random (e.g., achieved by a random permutation), we can compute skX, a random sample of

X, by simply selecting the first few elements of X.

In Section 3, we will propose using sketches to construct sample contingency tables. With this novel construction, the contingency table (and summary statistics based on the table) can be estimated using conventional statistical methods such as MLE.

2 Broder's Sketch Algorithm

One could randomly sample two postings and intersect the samples to estimate associations. The sketch technique introduced by Broder (1997) is a significant improvement, as demonstrated in Figure 2.

Assume that each document in the corpus of size D is assigned a unique random ID between 1 and D. The postings for word x is a sorted list of f_x doc IDs. The sketch, skX, is the first (smallest) s_x doc IDs in X. Broder used $\text{MIN}_s(Z)$ to denote the s smallest elements in the set, Z. Thus, $skX = \text{MIN}_{s_x}(X)$. Similarly, Y denotes the postings for word y, and skY denotes its sketch, $\text{MIN}_{s_y}(Y)$. Broder assumed $s_x = s_y = s$.

Broder defined resemblance (R) and sample resemblance (R_s) to be:

$$R = \frac{a}{a+b+c}, \quad R_s = \frac{|\text{MIN}_s(skX \cup skY) \cap skX \cap skY|}{|\text{MIN}_s(skX \cup skY)|}.$$

Broder (1997) proved that R_s is an unbiased estimator of R. One could use R_s to estimate a but he didn't do that, and it is not recommended.[4]

Sketches were designed to improve the coverage of a, as illustrated by Monte Carlo simulation in Figure 2. The figure plots, $\text{E}\left(\frac{a_s}{a}\right)$, percentage of intersections, as a function of (postings) sampling rate, $\frac{s}{f}$, where $f_x = f_y = f$, $s_x = s_y = s$. The solid lines (sketches), $\text{E}\left(\frac{a_s}{a}\right) \approx \frac{s}{f}$, are above the dashed curve (random sampling), $\text{E}\left(\frac{a_s}{a}\right) = \frac{s^2}{f^2}$. The difference is particularly important at low sampling rates.

3 Generalizing Sketches: $R \rightarrow$ Tables

Sketches were first proposed for estimating resemblance (R). This section generalizes the method to construct sample contingency tables, from which we can estimate associations: R, LLR, cosine, etc.

[4]There are at least three problems with estimating a from R_s. First, the estimate is biased. Secondly, this estimate uses just s of the $2 \times s$ samples; larger samples $\rightarrow$ smaller errors. Thirdly, we would rather not impose the restriction: $s_x = s_y$.

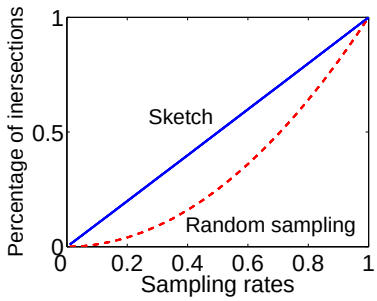


Figure 2: Sketches (solid curves) dominate random sampling (dashed curve). a=0.22, 0.38, 0.65, 0.80, 0.85f, f=0.2D, D=10^5. There is only one dashed curve across all values of a. There are different but indistinguishable solid curves depending on a.

Recall that the doc IDs span the integers from 1 to D with no gaps. When we compare two sketches, skX and skY, we have effectively looked at $D_s = \min\{skX_{(s_x)}, skY_{(s_y)}\}$ documents, where $skX_{(j)}$ is the jth smallest element in skX. The following construction generates the sample contingency table, a_s, b_s, c_s, d_s (as in Figure 1(b)). The example shown in Figure 3 may help explain the procedure.

$$D_s = \min\{skX_{(s_x)}, skY_{(s_y)}\}, \quad a_s = |skX \cap skY|,$$
$$n_x = s_x - |\{j : skX_{(j)} > D_s\}|,$$
$$n_y = s_y - |\{j : skY_{(j)} > D_s\}|,$$
$$b_s = n_x - a_s, \quad c_s = n_y - a_s, \quad d_s = D_s - a_s - b_s - c_s.$$

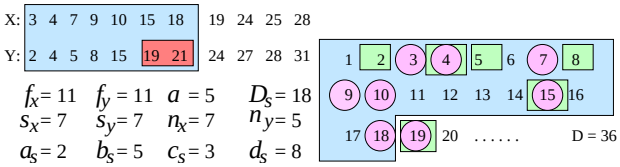
Given the sample contingency table, we are now ready to estimate the contingency table. It is sufficient to estimate a, since the rest of the table can be determined from f_x, f_y and D. For practical applications, we recommend the convenient closed-form approximation (8) in Section 5.1.

4 Margin-Free (MF) Baseline

Before considering the proposed MLE method, we introduce a baseline estimator that will not work as well because it does not take advantage of the margins. The baseline is the *multivariate hypergeometric* model, usually simplified as a *multinomial* by assuming "sample-with-replacement."

The sample expectations are (Siegrist, 1997),

$$\text{E}(a_s) = \frac{D_s}{D}a, \qquad \text{E}(b_s) = \frac{D_s}{D}b,$$
$$\text{E}(c_s) = \frac{D_s}{D}c, \qquad \text{E}(d_s) = \frac{D_s}{D}d. \tag{1}$$



(a) (b)

Figure 3: (a): The two sketches, skX and skY (larger shaded box), are used to construct a sample contingency table: a_s, b_s, c_s, d_s. skX consists of the first $s_x = 7$ doc IDs in X, the postings for word x. Similarly, skY consists of the first $s_y = 7$ doc IDs in Y, the postings for word y. There are 11 doc IDs in both X and Y, and $a = 5$ doc IDs in the intersection: {4, 15, 19, 24, 28}. (a) shows that $D_s = \min(18, 21) = 18$. Doc IDs 19 and 21 are excluded because we cannot determine if they are in the intersection or not, without looking outside the box. As it turns out, 19 is in the intersection and 21 is not. (b) enumerates the $D_s = 18$ documents, showing which documents contain x (small circles) and which contain y (small squares). Both procedures, (a) and (b), produce the same sample contingency table: $a_s = 2$, $b_s = 5$, $c_s = 3$ and $d_s = 8$.

The margin-free estimator and its variance are

$$\hat{a}_{MF} = \frac{D}{D_s} a_s, \quad \text{Var}(\hat{a}_{MF}) = \frac{D}{D_s} \frac{1}{\frac{1}{a} + \frac{1}{D-a}} \frac{D - D_s}{D - 1}. \quad (2)$$

For the multinomial simplification, we have

$$\hat{a}_{MF,r} = \frac{D}{D_s} a_s, \quad \text{Var}(\hat{a}_{MF,r}) = \frac{D}{D_s} \frac{1}{\frac{1}{a} + \frac{1}{D-a}}. \quad (3)$$

where "r" indicates "sample-with-replacement."

The term $\frac{D-D_s}{D-1} \approx \frac{D-D_s}{D}$ is often called the "finite-sample correction factor" (Siegrist, 1997).

5 The Proposed MLE Method

The task is to estimate the contingency table from the samples, the margins and D. We would like to use a maximum likelihood estimator for the most probable a, which maximizes the (full) likelihood (probability mass function, PMF) $P(a_s, b_s, c_s, d_s; a)$. Unfortunately, we do not know the exact expression for $P(a_s, b_s, c_s, d_s; a)$, but we do know the conditional probability $P(a_s, b_s, c_s, d_s | D_s; a)$. Since the doc IDs are uniformly random, sampling the first D_s contiguous documents is statistically equivalent to randomly sampling D_s documents from the corpus. Based on this key observation and Figure 3, *conditional* on D_s, $P(a_s, b_s, c_s, d_s | D_s; a)$ is the PMF of a two-way sample contingency table.

We factor the full likelihood into:

$$P(a_s, b_s, c_s, d_s; a) = P(a_s, b_s, c_s, d_s | D_s; a) \times P(D_s; a).$$

$P(D_s; a)$ is difficult. However, since we do not expect a strong dependency of D_s on a, we maximize the partial likelihood instead, and assume that is good enough. An example of partial likelihood is the Cox proportional hazards model in survival analysis (Venables and Ripley, 2002, Section 13.3) .

Our partial likelihood is

$$P(a_s, b_s, c_s, d_s | D_s; a) = \frac{\binom{a}{a_s}\binom{f_x - a}{b_s}\binom{f_y - a}{c_s}\binom{D - f_x - f_y + a}{d_s}}{\binom{D}{D_s}}$$
$$\propto \prod_{i=0}^{a_s-1} (a - i) \times \prod_{i=0}^{b_s-1} (f_x - a - i) \times \prod_{i=0}^{c_s-1} (f_y - a - i)$$
$$\times \prod_{i=0}^{d_s-1} (D - f_x - f_y + a - i), \quad (4)$$

where $\binom{n}{m} = \frac{n!}{m!(n-m)!}$. "$\propto$" is "proportional to."

We now derive an MLE for (4), a result that was not previously known, to the best of our knowledge. Let $\hat{a}_{MLE}$ maximizes $\log P(a_s, b_s, c_s, d_s | D_s; a)$:

$$\sum_{i=0}^{a_s-1} \log(a - i) + \sum_{i=0}^{b_s-1} \log(f_x - a - i)$$
$$+ \sum_{i=0}^{c_s-1} \log(f_y - a - i) + \sum_{i=0}^{d_s-1} \log(D - f_x - f_y + a - i),$$

whose first derivative, $\frac{\partial \log P(a_s, b_s, c_s, d_s | D_s; a)}{\partial a}$, is

$$\sum_{i=0}^{a_s-1} \frac{1}{a - i} - \sum_{i=0}^{b_s-1} \frac{1}{f_x - a - i} - \sum_{i=0}^{c_s-1} \frac{1}{f_y - a - i}$$
$$+ \sum_{i=0}^{d_s-1} \frac{1}{D - f_x - f_y + a - i}. \quad (5)$$

Since the second derivative, $\frac{\partial^2 \log P(a_s, b_s, c_s, d_s | D_s; a)}{\partial a^2}$, is negative, the log likelihood function is concave, hence has a unique maximum. One could numerically solve (5) for $\frac{\partial \log P(a_s, b_s, c_s, d_s | D_s; a)}{\partial a} = 0$. However, we derive the exact solution using the following updating formula from (4):

$$P(a_s, b_s, c_s, d_s | D_s; a) = P(a_s, b_s, c_s, d_s | D_s; a-1) \times \frac{f_x - a + 1 - b_s}{f_x - a + 1} \frac{f_y - a + 1 - c_s}{f_y - a + 1} \frac{D - f_x - f_y + a}{D - f_x - f_y + a - d_s} \frac{a}{a - a_s} = P(a_s, b_s, c_s, d_s | D_s; a-1) \times g(a). \quad (6)$$

Since our MLE is unique, it suffices to find a from $g(a) = 1$, which is a cubic function in a.

5.1 A Convenient Practical Approximation

Rather than solving the cubic equation for the exact MLE, the following approximation may be more convenient. Assume we sample $n_x = a_s + b_s$ from X and obtain a_s co-occurrences without knowledge of the samples from Y. Further assuming "sample-with-replacement," a_s is then binomially distributed, $a_s \sim \text{Binom}(n_x, \frac{a}{f_x})$. Similarly, assume $a_s \sim \text{Binom}(n_y, \frac{a}{f_y})$. Under these assumptions, the PMF of a_s is a product of two binomial PMFs:

$$\binom{f_x}{n_x} \left(\frac{a}{f_x}\right)^{a_s} \left(\frac{f_x - a}{f_x}\right)^{b_s} \binom{f_y}{n_y} \left(\frac{a}{f_y}\right)^{a_s} \left(\frac{f_y - a}{f_y}\right)^{c_s} \propto a^{2a_s} (f_x - a)^{b_s} (f_y - a)^{c_s}. \quad (7)$$

Setting the first derivative of the logarithm of (7) to be zero, we obtain $\frac{2a_s}{a} - \frac{b_s}{f_x - a} - \frac{c_s}{f_y - a} = 0$, which is quadratic in a and has a solution:

$$\hat{a}_{MLE,a} = \frac{f_x(2a_s + c_s) + f_y(2a_s + b_s)}{2(2a_s + b_s + c_s)} - \frac{\sqrt{(f_x(2a_s + c_s) - f_y(2a_s + b_s))^2 + 4 f_x f_y b_s c_s}}{2(2a_s + b_s + c_s)}. \quad (8)$$

Section 6 shows that $\hat{a}_{MLE,a}$ is very close to $\hat{a}_{MLE}$.

5.2 Theoretical Evaluation: Bias and Variance

How good are the estimates? A popular metric is mean square error (MSE): $\text{MSE}(\hat{a}) = \text{E}(\hat{a} - a)^2 = \text{Var}(\hat{a}) + \text{Bias}^2(\hat{a})$. If $\hat{a}$ is unbiased, $\text{MSE}(\hat{a}) = \text{Var}(\hat{a}) = \text{SE}^2(\hat{a})$, where SE is the standard error. Here all expectations are conditional on D_s.

Large sample theory (Lehmann and Casella, 1998, Chapter 6) says that, under "sample-with-replacement," $\hat{a}_{MLE}$ is asymptotically unbiased and converges to Normal with mean a and variance $\frac{1}{\text{I}(a)}$, where $\text{I}(a)$, the Fisher Information, is

$$\text{I}(a) = -\text{E}\left(\frac{\partial^2}{\partial a^2} \log P(a_s, b_s, c_s, d_s | D_s; a, r)\right). \quad (9)$$

Under "sample-with-replacement," we have

$$P(a_s, b_s, c_s, d_s | D_s; a, r) \propto \left(\frac{a}{D}\right)^{a_s} \times \left(\frac{f_x - a}{D}\right)^{b_s} \times \left(\frac{f_y - a}{D}\right)^{c_s} \times \left(\frac{D - f_x - f_y + a}{D}\right)^{d_s}, \quad (10)$$

Therefore, the Fisher Information, $\text{I}(a)$, is

$$\frac{\text{E}(a_s)}{a^2} + \frac{\text{E}(b_s)}{(f_x - a)^2} + \frac{\text{E}(c_s)}{(f_y - a)^2} + \frac{\text{E}(d_s)}{(D - f_x - f_y + a)^2}. \quad (11)$$

We plug (1) from the margin-free model into (11) as an approximation, to obtain

$$\text{Var}(\hat{a}_{MLE}) \approx \frac{\frac{D}{D_s} - 1}{\frac{1}{a} + \frac{1}{f_x - a} + \frac{1}{f_y - a} + \frac{1}{D - f_x - f_y + a}}, \quad (12)$$

which is $\frac{1}{\text{I}(a)}$ multiplied by $\frac{D - D_s}{D}$, the "finite-sample correction factor," to consider "sample-without-replacement."

We can see that $\text{Var}(\hat{a}_{MLE})$ is less than $\text{Var}(\hat{a}_{MF})$ in (2). In addition, $\hat{a}_{MLE}$ is asymptotically unbiased while $\hat{a}_{MF}$ is no longer unbiased under margin constraints. Therefore, we expect $\hat{a}_{MLE}$ has smaller MSE than $\hat{a}_{MF}$. In other words, the proposed MLE method is more accurate than the MF baseline, in terms of variance, bias and mean square error. If we know the margins, we ought to use them.

5.3 Unconditional Bias and Variance

$\hat{a}_{MLE}$ is also unconditionally unbiased:

$$\text{E}(\hat{a}_{MLE} - a) = \text{E}(\text{E}(\hat{a}_{MLE} - a | D_s)) \approx \text{E}(0) = 0. \quad (13)$$

The unconditional variance is useful because often we would like to estimate the errors before knowing D_s (e.g., for choosing sample sizes).

To compute the unconditional variance of $\hat{a}_{MLE}$, we should replace $\frac{D}{D_s}$ with $\text{E}\left(\frac{D}{D_s}\right)$ in (12). We resort to an approximation for $\text{E}\left(\frac{D}{D_s}\right)$. Note that $skX_{(s_x)}$ is the order statistics of a discrete random variable (Siegrist, 1997) with expectation

$$\text{E}\left(skX_{(s_x)}\right) = \frac{s_x(D+1)}{f_x + 1} \approx \frac{s_x}{f_x} D. \quad (14)$$

By Jensen's inequality, we know that

$$\text{E}\left(\frac{D_s}{D}\right) \leq \min\left(\frac{\text{E}\left(skX_{(s_x)}\right)}{D}, \frac{\text{E}\left(skY_{(s_y)}\right)}{D}\right) = \min\left(\frac{s_x}{f_x}, \frac{s_y}{f_y}\right) \quad (15)$$

$$\text{E}\left(\frac{D}{D_s}\right) \geq \frac{1}{\text{E}\left(\frac{D_s}{D}\right)} \geq \max\left(\frac{f_x}{s_x}, \frac{f_y}{s_y}\right). \quad (16)$$

Table 2: Gold standard joint frequencies, a. Document frequencies are shown in parentheses. These words are frequent, suitable for evaluating our algorithms at very low sampling rates.

	THIS	HAVE	HELP	PROGRAM
THIS (27633)	—	13517	7221	3682
HAVE (17396)	13517	—	5781	3029
HELP (10791)	7221	5781	—	1949
PROGRAM (5327)	3682	3029	1949	—

Replacing the inequalities with equalities underestimates the variance, but only slightly.

5.4 Smoothing

Although not a major emphasis here, our evaluations will show that $\hat{a}_{MLE+S}$, a smoothed version of the proposed MLE method, is effective, especially at low sampling rates. $\hat{a}_{MLE+S}$ uses "add-one" smoothing. Given that such a simple method is as effective as it is, it would be worth considering more sophisticated methods such as Good-Turing.

5.5 How Many Samples Are Sufficient?

The answer depends on the trade-off between computation and estimation errors. One simple rule is to sample "2%." (12) implies that the standard error is proportional to $\sqrt{D/D_s - 1}$. Figure 4(a) plots $\sqrt{D/D_s - 1}$ as a function of sampling rate, D_s/D, indicating a "elbow" about 2%. However, 2% is too large for high frequency words.

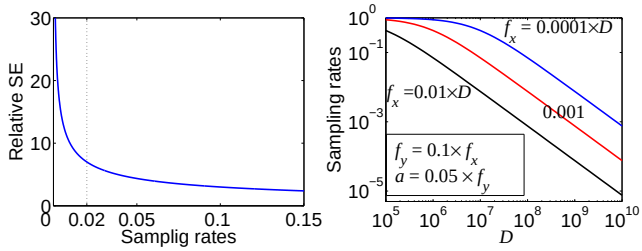
A more reasonable metric is the "coefficient of variation," cv = $\frac{\mathrm{SE}(\hat{a})}{a}$. At Web scale (10 billion pages), we expect that a very small sampling rate such as 10^{-4} or 10^{-5} will suffice to achieve a reasonable cv (e.g., 0.5). See Figure 4(b).

6 Evaluation

Two sets of experiments were run on a collection of $D = 2^{16}$ web pages, provided by MSN. The first experiment considered 4 English words shown in Table 2, and the second experiment considers 968 English words with mean df = 2135 and median df = 1135. They form 468,028 word pairs, with mean co-occurrences = 188 and median = 74.

6.1 Small Dataset Monte Carlo Experiment

Figure 5 evaluates the various estimate methods by MSE over a wide range of sampling rates. Doc IDs were randomly permuted 10^5 times. For each permutation we constructed sketches from the inverted index at a series of sampling rates. The figure shows that the proposed method, $\hat{a}_{MLE}$, is considerably better (by $20\% - 40\%$) than the margin-free baseline, $\hat{a}_{MF}$. Smoothing is effective at low sampling rates. The recommended approximation, $\hat{a}_{MLE,a}$, is remarkably close to the exact solution.



(a) (b)

Figure 4: How large should the sampling rate be? (a): We can sample up to the "elbow point" (2%), but after that there are diminishing returns. (b): An analysis based on cv = $\frac{\mathrm{SE}}{a}$ = 0.5 suggests that we can get away with much lower sampling rates. The three curves plot the critical value for the sampling rate, $\frac{D_s}{D}$, as a function of corpus size, D. At Web scale, $D \approx 10^{10}$, sampling rates above 10^{-3} to 10^{-5} satisfy cv < 0.5, at least for these settings of f_x, f_y and a. The settings were chosen to simulate "ordinary" words. The three curves correspond to three choices of f_x: $D/100$, $D/1000$, and $D/10,000$. $f_y = f_x/10$, $a = f_y/20$. SE is based on (12).

Figure 6 shows agreement between the theoretical and empirical unconditional variances. Smoothing reduces variances, at low sampling rates. We used the empirical $\mathrm{E}\left(\frac{D}{D_S}\right)$ to compute the theoretical variances. The approximation, $\max\left(\frac{f_x}{s_x}, \frac{f_y}{s_y}\right)$, is $> 0.95\mathrm{E}\left(\frac{D}{D_S}\right)$ at sampling rates > 0.01.

Figure 7 verifies that the proposed MLE is unbiased, unlike the margin-free baselines.

6.2 Large Dataset Experiment

The large experiment considers 968 English words (468,028 pairs) over a range of sampling rates. A floor of 20 was imposed on sample sizes.

As reported in Figure 8, the large experiment confirms once again that proposed method, $\hat{a}_{MLE}$, is considerably better than the margin-free baseline (by

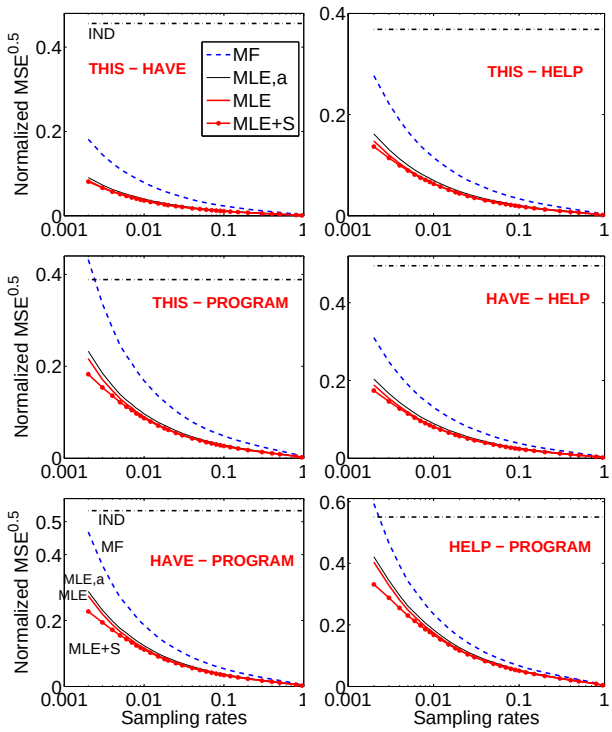


Figure 5: The proposed method, $\hat{a}_{MLE}$ outperforms the margin-free baseline, $\hat{a}_{MF}$, in terms of $\frac{\text{MSE}^{0.5}}{a}$. The recommended approximation, $\hat{a}_{MLE,a}$, is close to $\hat{a}_{MLE}$. Smoothing, $\hat{a}_{MLE+S}$, is effective at low sampling rates. All methods are better than assuming independence (IND).

$15\% - 30\%$). The recommended approximation, $\hat{a}_{MLE,a}$, is close to $\hat{a}_{MLE}$. Smoothing, $\hat{a}_{MLE+S}$ helps at low sampling rates.

6.3 Rank Retrieval: Top k Associated Pairs

We computed a gold standard similarity cosine ranking of the 468,028 pairs using a 100% sample: $\cos = \frac{a}{\sqrt{f_x f_y}}$. We then compared the gold standard to rankings based on smaller samples. Figure 9(a) compares the two lists in terms of agreement in the top k. For $3 \leq k \leq 200$, with a sampling rate of 0.005, the agreement is consistently 70% or higher. Increasing sampling rate, increases agreement.

The same comparisons are evaluated in terms of precision and recall in Figure 9(b), by fixing the top 1% of the gold standard list but varying the top percentages of the sample list. Again, increasing sampling rate, increases agreement.

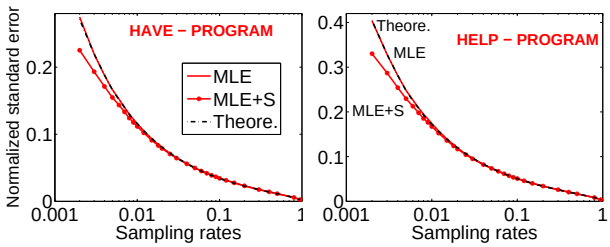


Figure 6: The theoretical and empirical variances show remarkable agreement, in terms of $\frac{\text{SE}(\hat{a})}{a}$. Smoothing reduces variances at low sampling rates.

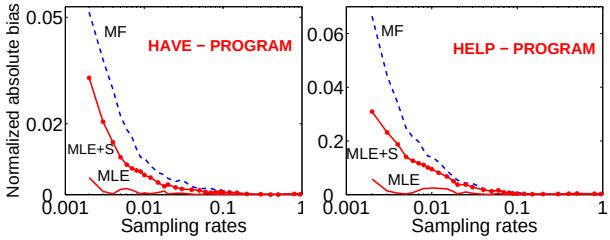


Figure 7: Biases in terms of $\frac{|\text{E}(\hat{a})-a|}{a}$. $\hat{a}_{MLE}$ is practically unbiased, unlike $\hat{a}_{MF}$. Smoothing increases bias slightly.

7 Conclusion

We proposed a novel sketch-based procedure for constructing sample contingency tables. The method bridges two popular choices: (A) sampling over documents and (B) sampling over postings. Well-understood maximum likelihood estimation (MLE) techniques can be applied to sketches (or to traditional samples) to estimate word associations. We derived an exact cubic solution, $\hat{a}_{MLE}$, as well as a quadratic approximation, $\hat{a}_{MLE,a}$. The approximation is recommended because it is close to the exact solution, and easy to compute.

The proposed MLE methods were compared empirically and theoretically to a margin-free (MF) baseline, finding large improvements. When we know the margins, we ought to use them.

Sample-based methods (MLE & MF) are often better than sample-free methods. Associations are often estimated without samples. It is popular to assume independence: (Garcia-Molina et al., 2002, Chapter 16.4), i.e., $\hat{a} \approx \frac{f_x f_y}{D}$. Independence led to large errors in our experiments.

Not unsurprisingly, there is a trade-off between computational work (space and time) and statistical

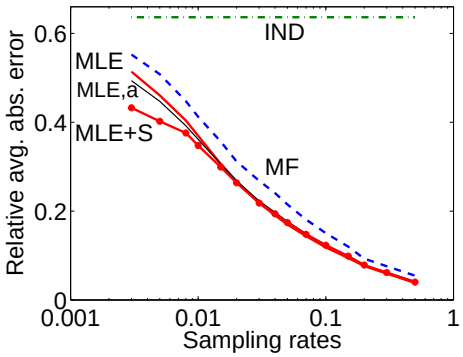
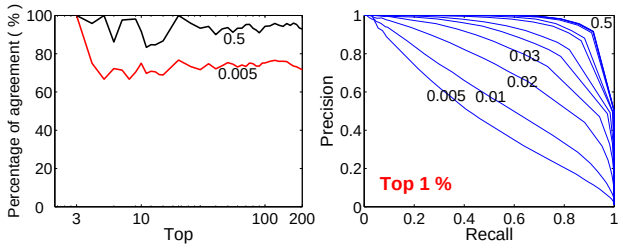


Figure 8: We report the (normalized) mean absolute errors (divided by the mean co-occurrences, 188). All curves are averaged over three permutations. The proposed MLE and the recommended approximation are very close and both are significantly better than the margin-free (MF) baseline. Smoothing, $\hat{a}_{MLE+S}$, helps at low sampling rates. All estimators do better than assuming independence.

accuracy (variance or errors); reducing the sampling rate saves work, but costs accuracy. We derived formulas for variance, showing precisely how accuracy depends on sampling rate. Sampling methods become more and more important with larger and larger collections. At Web scale, sampling rates as low as 10^{-4} may suffice for "ordinary" words.

We have recently generalized the sampling algorithm and estimation method to multi-way associations; see (Li and Church, 2005).



(a) (b)

Figure 9: (a): Percentage of agreements in the gold standard and reconstructed (from samples) top 3 to 200 list. (b):Precision-recall curves in retrieving the top 1% gold standard pairs, at different sampling rates. For example, 60% recall and 70% precision is achieved at sampling rate = 0.02.

References

S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International World Wide Web Conference*, pages 107–117, Brisbane, Australia.

A. Broder. 1997. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences*, pages 21–29, Positano, Italy.

M. S. Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388, Montreal, Quebec, Canada.

K. Church and P. Hanks. 1991. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29.

S. Deerwester, S. T. Dumais, G. W. Furnas, and T. K. Landauer. 1999. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

H. Garcia-Molina, J. D. Ullman, and J. D. Widom. 2002. *Database Systems: the Complete Book*. Prentice Hall, New York, NY.

A. S. Hornby, editor. 1989. *Oxford Advanced Learner's Dictionary*. Oxford University Press, Oxford, UK.

E. L. Lehmann and G. Casella. 1998. *Theory of Point Estimation*. Springer, New York, NY, second edition.

P. Li and K. W. Church. 2005. Using sketches to estimate two-way and multi-way associations. Technical report, Microsoft Research, Redmond, WA.

C. D. Manning and H. Schutze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA.

R. C. Moore. 2004. On log-likelihood-ratios and the significance of rare events. In *Proceedings of EMNLP 2004*, pages 333–340, Barcelona, Spain.

D. Ravichandran, P. Pantel, and E. Hovy. 2005. Randomized algorithms and NLP: Using locality sensitive hash function for high speed noun clustering. In *Proceedings of ACL*, pages 622–629, Ann Arbor.

K. Siegrist. 1997. *Finite Sampling Models*, *http://www.ds.unifi.it/VL/VL_EN/urn/index.html*. Virtual Laboratories in Probability and Statistics.

W. N. Venables and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Springer-Verlag, New York, NY, fourth edition.