On the Signature Weight in "multiple" m Signature Files

Fabio Grandi

C.I.O.C. – C.N.R. and Dipartimento di Elettronica, Informatica e Sistemistica
Università di Bologna, Viale Risorgimento 2, I-40136 Bologna, Italy
Tel: +39 (0)51 644.3555, Fax: +39 (0)51 644.3540, Email: fgrandi@deis.unibo.it

Abstract

This paper deals with the estimation of the signature weight as generated by the superimposed coding technique adopted in "multiple" m signature files. The estimation is needed for performance evaluation of such organizations used for information retrieval applications. In particular, simple formulas for the probability density function, the expected value and the variance of the signature weight are presented.

The presented formulas can be derived following a general methodology we called the γ-transform approach in a previous work, which results much more simple than the method used by other authors for the derivation of the density function. Equivalence of the results is also shown.

Index terms: Information retrieval, Signature files, Superimposed coding, Discrete probability, Estimation, γ-transform

1 Introduction

The inclusive "or" of bit strings (superimposed coding) is the method adopted for coding data in many signature file organizations used for information retrieval applications. In "standard" superimposed coding [15], signatures are formed from the superimposition of T term signatures, each of which contains the same number, say m, of "1" bits. Superimposed coding has extensively been used for text retrieval [3, 4, 7, 8, 9, 15, 19] but also for other applications, working on formatted or unformatted data, based on the Bloom filter [2, 10, 11, 12, 16, 17, 18].

In a recent proposal [1], signatures are obtained from the superimposition of "multiple" m term signatures, that is term signatures with different numbers of "1" bits (say m_i the number of "1" bits set by the i-th term, $i = 1, \ldots, T$). This method makes it possible to assign higher weights to more frequently accessed terms and improve the overall system performance. However, performance evaluation requires an estimation of the weight (say W, namely total number of "1" bits) of the signatures obtained in this way. To this end, W can be considered as random variable and its distribution and moments can be studied.

The method sketched in [1] and developed in [13] for the evaluation of the signature weight distribution is based on the description of the signature construction as a Markov process. Such a process consists of T stages — each of which corresponds to the superimposition of a term signature — and requires heavy and cumbersome matrix manipulations. Since $W \geq m_1$ if $T \geq 1$, the signature weight distribution was evaluated in [1, 13] as:

$$\Pr[W = m_1 + k] \;=\; \sum_{j=0}^{k} \binom{f - m_1}{j} \binom{f - m_1 - j}{k - j} (-1)^{k+j} \prod_{i=2}^{T} \frac{\binom{m_1 + j}{m_i}}{\binom{f}{m_i}} \tag{1}$$

where $0 \leq k \leq f - m_1$. No evaluation of the moments has been done in [1, 13].

2 Using the γ-transform Approach

In a previous work [6], we introduced a discrete transform (called γ-transform) of the probability density function of a discrete and finite random variable. We showed that the γ-transform approach is very useful in analyzing combinatorial problems which can be described by means of the "principle of inclusion and exclusion" [14], as it provides an easy derivation of the density function and of the moments.

For a given discrete random variable with distribution $\Pr[W = w] = p(w)$ in $\{0, 1, \ldots, f\}$, the *γ-transform* of the density function was defined in [6] as follows:

$$\gamma(x) \;=\; \sum_{w=0}^{f} \frac{\binom{x}{w}}{\binom{f}{w}} \, p(w) \tag{2}$$

Hence, we showed that the *inversion* formula for the γ-transform is given by:

$$p(w) \;=\; \binom{f}{w} \sum_{j=0}^{w} (-1)^j \binom{w}{j} \gamma(w - j) \tag{3}$$

In [6], a physical meaning for the γ-transform $\gamma(\cdot)$ with an integer argument in the range $\{0, 1, \ldots, f\}$ was also provided for a wide class of experiments. Let $p(w)$ be the probability that a given experiment gives rise to w succesful outcomes selected from a set F of f possible outcomes. Then $\gamma(k)$ represents the probability that the experiment is effected as if the outcomes could only be selected from a subset of F with cardinality k, fixed before the experiment. In other words, $\gamma(k)$ is the probability that $f - k$ of the possible outcomes have been excluded *a priori* from the result.

As far as the signature generation is concerned, it is easy to show that it can be described as an outcome inclusion experiment that consists of the selection of the "1" bits to be set. In this case, it is easy to evaluate the γ-transform expression from its physical meaning as:

$$\gamma(k) = \frac{\psi(k)}{\psi(f)} = \prod_{i=1}^{T} \frac{\binom{k}{m_i}}{\binom{f}{m_i}} \tag{4}$$

where $\psi(k) = \prod_{i=1}^{T} \binom{k}{m_i}$ represents the number of ways the "1"s to be set by the T terms can only be selected in a subset of F with cardinality k (i.e. the other $f - k$ are excluded *a priori*). Hence, the γ-transform approach allows a straightforward derivation of the probability distribution thanks to the inversion formula (3):

$$\Pr[W = w] = \Pr[W = w | m_1, \dots, m_T] = \binom{f}{w} \sum_{j=0}^{w} (-1)^j \binom{w}{j} \prod_{i=1}^{T} \frac{\binom{w-j}{m_i}}{\binom{f}{m_i}} \tag{5}$$

which is in a slightly more handy form than (1). As a matter of fact, expression (5) is symmetric with respect to all the m_i's and contains a binomial coefficient, $\binom{f}{w}$ which does not depend on the summation index j. Probability distribution (5) can also be directly derived by means of combinatorial arguments following the same steps of [6, Th. 2 and Sec. 5.1].

Furthermore, the γ-transform approach allows the derivation of all the moments of the random variable W. In particular, we proved that the r-th factorial moment of W is in general given by:

$$E[W^{\underline{r}}] = f^{\underline{r}} \sum_{l=0}^{r} (-1)^l \binom{r}{l} \gamma(f - l) \tag{6}$$

where $f^{\underline{r}} = f(f-1)\cdots(f-r+1)$ is the r-th falling factorial power of f. By means of Eq. (6), the expected value and the variance of W can be evaluated as follows (see [6] for details):

$$E[W] = f\left[1 - \prod_{i=1}^{T}\left(1 - \frac{m_i}{f}\right)\right] \tag{7}$$

$$\sigma_W^2 = f^2\left[\prod_{i=1}^{T}\left(1 - \frac{m_i}{f}\right)\left(1 - \frac{m_i}{f-1}\right) - \prod_{i=1}^{T}\left(1 - \frac{m_i}{f}\right)^2\right] + f\prod_{i=1}^{T}\left(1 - \frac{m_i}{f}\right)\left[1 - \prod_{i=1}^{T}\left(1 - \frac{m_i}{f-1}\right)\right] \tag{8}$$

It can be noticed that, whereas the expected value can also be evaluated in an elementary way (provided that m_i/f is the probability that a given bit is set to "1" by the i-th term), even a direct derivation of the variance from the density function (5) is rather difficult.

Letting V be a random variable with binomial distribution and with the same mean value $E[V] = fp$ as W, being $p = 1 - \prod_{i=1}^{T}(1 - m_i/f)$ the probability that a given bit is set to "1" by the superimposition of the T terms, the distributions of W and V can be compared. Although they have very similar shapes, the W distribution is a bit less dispersed around the mean value than the binomial, as it can easily verified from (8) that $\sigma_W^2 < \sigma_V^2 = fp(1-p)$. This is due to the fact that (5) ensures that $\Pr[W = w] = 0$ for all $w < \min\{m_1, \ldots, m_T\}$ (the i-th term sets m_i bits *at a time*), whereas we can also have $V = 0$ with nonnull probability.

A practical application of the closed formula (8) for the variance of W is, for instance, an accurate evaluation of the false drop probability F_d in "multiple" m signature files, which can be different from the minimal value expected from optimal design, due to the simplifications introduced to obtain tractable formulas [1, 5]. To this end, if $S_1, \ldots, S_n$ are the disjoint sets into which terms have been partitioned according to their query frequency, we assume that D_i and q_i be the number of distinct terms of S_i in a document and the probability that a query term is from S_i, respectively. Being

$$\varphi(w) \quad = \quad \sum_{i=1}^{n} q_i \left(\frac{w}{f}\right)^{m_i} \tag{9}$$

the expression of the false drop probability [5], we can obtain an accurate estimate of its expected value as:

$$F_d \quad \approx \quad \varphi(E[W]) + \frac{\sigma_W^2}{2}\varphi''(E[W]) \tag{10}$$

where:

$$\varphi''(w) \quad = \quad \frac{1}{f^2}\sum_{i=1}^{n} q_i m_i(m_i - 1)\left(\frac{w}{f}\right)^{m_i-2} \tag{11}$$

It should be noticed that signature weights of *documents* instead of *queries* are to be considered here, thus,

$$E[W] \quad = \quad f\left[1 - \prod_{i=1}^{n}\left(1 - \frac{m_i}{f}\right)^{D_i}\right] \tag{12}$$

$$\begin{aligned}\sigma_W^2 \quad = \quad & f^2\left[\prod_{i=1}^{n}\left(1-\frac{m_i}{f}\right)^{D_i}\left(1-\frac{m_i}{f-1}\right)^{D_i} - \prod_{i=1}^{n}\left(1-\frac{m_i}{f}\right)^{2D_i}\right] + \\ & f\prod_{i=1}^{n}\left(1-\frac{m_i}{f}\right)^{D_i}\left[1-\prod_{i=1}^{n}\left(1-\frac{m_i}{f-1}\right)^{D_i}\right]\end{aligned} \tag{13}$$

must be substituted, along with (9) and (11), into (10).

3 Equivalence of the results

Finally, we show in the following that the two expressions of the signature weight distribution, (5) and (1), are equivalent. Letting $w = m_1 + k$ and substituting $l = m_1 + j$ for j in (1) we obtain:

$$\Pr[W = w] \quad = \quad \sum_{l=m_1}^{w}\binom{f-m_1}{l-m_1}\binom{f-l}{w-l}(-1)^{w+l-2m_1}\prod_{i=2}^{T}\frac{\binom{l}{m_i}}{\binom{f}{m_i}} \tag{14}$$

Since $w + l - 2m_1 \equiv w - l \pmod 2$, and via multiplication and division by $\binom{l}{m_1}/\binom{f}{m_1}$, Eq. (14) becomes:

$$\Pr[W = w] \quad = \quad \sum_{l=m_1}^{w} \binom{f-m_1}{l-m_1} \binom{f-l}{w-l} \frac{\binom{f}{m_1}}{\binom{l}{m_1}} (-1)^{w-l} \prod_{i=1}^{T} \frac{\binom{l}{m_i}}{\binom{f}{m_i}} \tag{15}$$

Hence, it is easy to verify (by expanding binomial coefficients into factorials, simplifying, multiplying and dividing by $w!$, and rearranging factorials as binomials) that

$$\binom{f-m_1}{l-m_1} \binom{f-l}{w-l} \frac{\binom{f}{m_1}}{\binom{l}{m_1}} \quad = \quad \binom{f}{w} \binom{w}{l} \tag{16}$$

The application of identity (16) to Eq. (15) yields:

$$\Pr[W = w] \quad = \quad \binom{f}{w} \sum_{l=m_1}^{w} \binom{w}{l} (-1)^{w-l} \prod_{i=1}^{T} \frac{\binom{l}{m_i}}{\binom{f}{m_i}} \tag{17}$$

Since $\binom{l}{m_1} = 0$ if $l < m_1$, we can extend the lower summation limit in (17) down to 0; the substitution $j = w - l$ eventually allows us to obtain Eq. (5).

References

[1] D. Aktug and F. Can, "Analysis of Multiterm Queries in a Dynamic Signature File Organization," *Proc. of 16th ACM-SIGIR Intl. Conf.* (Pittsburgh, PA), 1993.

[2] B. H. Bloom, "Space/Time Tradeoffs in Hash Coding with Allowable Errors," *Comm. ACM* **13** (7), 1970, 422–426.

[3] W. B. Croft and P. Savino, "Implementing Ranking Strategies Using Text Signatures," *ACM TOIS* **6** (1), 1988, 42–62.

[4] C. Faloutsos and S. Christodoulakis, "Signature Files: An Access Method for Documents and Its Analytical Performance Evaluation," *ACM TOIS* **2** (4), 1984, 267–288.

[5] C. Faloutsos and S. Christodoulakis, "Design of a Signature File Method that Accounts for Non-Uniform Occurrence and Query Frequencies," *Proc. of 11th VLDB Intl. Conf.* (Stockholm, Sweden), 1985, 165–170.

[6] F. Grandi, "The γ-transform: A New Approach to the Study of a Discrete and Finite Random Variable," submitted for publication, 1994.

[7] F. Grandi, P. Tiberio and P. Zezula, "Frame-sliced Partitioned Parallel Signature Files," *Proc. of 15th ACM SIGIR Intl. Conf.* (Copenhagen, Denmark), 1992, 286–297.

[8] D. L. Lee and C.-W. Leng, "Partitioned Signature Files: Design Issues and Performance Evaluation," *ACM TOIS* **7** (2), 1989, 158–180.

[9] Z. Lin and C. Faloutsos, "Frame-sliced Signature Files," *IEEE TKDE* **4** (3), 1992, 281–289.

[10] L. F. Mackert and G. M. Lohman, "R* Optimizer Validation and Performance Evaluation for Local Queries," *Proc. of 12th VLDB Intl. Conf.* (Kyoto, Japan), 1986, 84–95.

[11] J. K. Mullin, "A Second Look at Bloom Filters, " *Comm. ACM* **26** (8), 1983, 570–571.

[12] J. K. Mullin and D. J. Margoliash, "A Tale of Three Spelling Checkers," *Software - Practice & Experience* **20** (6), 1990, 625–630.

[13] E. S. Murphree and D. Aktug, "Derivation of Probability Distribution of the Weight of the Query Signature," preprint, Department of Mathematics and Statistics, Miami University, Oxford, Ohio, 1992.

[14] J. Riordan, *An Introduction to Combinatorial Analysis*, New York, NY: Wiley & Sons, 1958.

[15] C. S. Roberts, "Partial Match Retrieval via the Method of Superimposed Codes," *Proc. IEEE* **67** (12), 1979, 1624–1642.

[16] R. Sacks-Davis and K. Ramamohanarao, "Multikey Access Methods Based on Superimposed Coding Techniques," *ACM TODS* **12** (4), 1987, 655–696.

[17] D. G. Severance and G. M. Lohman, "Differential Files: Their Applications to the Maintenance of Large Databases," *ACM TODS* **1** (3), 1976, 256–367.

[18] A. L. Tharp, *File Organization and Processing*, New York, NY: Wiley & Sons, 1988.

[19] P. Zezula, F. Rabitti and P. Tiberio, "Dynamic Partitioning of Signature Files," *ACM TOIS* **9** (4), 1991, 336–367.