A formal derivation of Heaps’ Law

D.C. van Leijenhorst, Th.P. van der Weide *

Department of Computer Science, Faculty of Mathematics and Computing Science, Radboud University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, Netherlands

Received 25 September 2003; received in revised form 26 February 2004; accepted 2 March 2004

Abstract

Word frequencies in text documents can be reasonably described by the Mandelbrot distribution, which has Zipf’s Law as a special case. Furthermore, the growth of vocabulary size as a function of the text size (its number of words) has been described in Heaps’ Law. It has been shown that these two experimental laws are related.

In this paper we go a step further, and provide a (formal) derivation of Heaps’ Law from the Mandelbrot distribution. We also provide a specification of the validity area for applying Heaps’ Law.

© 2004 Elsevier Inc. All rights reserved.

1. Introduction

In many practical situations, a connection has been shown between the order of probability of events, and the probability itself. The most well-known models for such connections are Zipf’s Law [12] and the Mandelbrot distribution [8].

Let the $r$th most probable event have probability $p$, then Zipf’s Law states that $p \cdot r$ is (almost) equal for all events, while the Mandelbrot distribution claims this for the expression $p \cdot (c + r)^\theta$ for some parameters $c$ and $\theta$. In case of $c = 0$, the distribution is also referred to as the generalized Zipf’s Law. Some authors motivate the validity of these laws from physical phenomena, see for...
example [4] for Zipf’s Law in the context of cities. But it is also possible to derive Zipf/Mandelbrot’s Law from a simple statistical model [7]. For example, Zipf’s Law can be derived for word occurrences in artificial language, when it is assumed that letters that compose a word are drawn randomly from some distribution. In practice, however, words are thoughtfully selected by the author; yet on the long run this selection process may adjust to such a statistical description.

Another experimental law of nature is Heaps’ Law [6], which describes the average growth in the number of unique elements (also referred as the number of records), when elements are drawn randomly without replacement from some statistical distribution. For example, in the case of word occurrences in natural language, Heaps’ Law predicts the vocabulary size of a document from its text size, i.e., the number of words it contains. Heaps’ Law states that this number of unique elements will grow according to \( a k^b \) for some application dependent constants \( a \) and \( b \), \( 0 < b < 1 \), where \( k \) is the number of drawings. See Table 1 for an overview of used symbols.

In this paper we focus on the relation between Zipf’s Law and the Mandelbrot distribution on the one hand, and Heaps’ Law on the other hand. This relation has been recognized, for example in [3], but this relation has not been formally motivated. In this paper we assume that elements are drawn according to the Mandelbrot distribution, and derive Heaps’ Law for the number of unique elements drawn. As a consequence, Heaps’ Law can also be regarded in a natural way as a complexity estimate.

Unfortunately, this analysis leads to a rather untractable recurrence relation that has no analytical solution. By applying techniques from complexity theory, restricting ourselves to first-order terms, Heaps’ Law is obtained. Note that by involving second order terms, a more advanced formulation of Heaps’ Law may be obtained.

Table 1
Table of most important symbols used in this paper

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>Vocabulary size</td>
</tr>
<tr>
<td>( c )</td>
<td>Constant in Mandelbrot distribution</td>
</tr>
<tr>
<td>( \theta )</td>
<td>Constant in Mandelbrot distribution</td>
</tr>
<tr>
<td>( a_N )</td>
<td>Normalization constant of Mandelbrot distribution</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>Constant in Heaps’ Law</td>
</tr>
<tr>
<td>( \beta )</td>
<td>constant in Heaps’ Law</td>
</tr>
<tr>
<td>( S_k )</td>
<td>Probability of new word in ( k )th drawing</td>
</tr>
<tr>
<td>( M_k )</td>
<td>( k )th inverse moment of probability distribution</td>
</tr>
<tr>
<td>( N_k )</td>
<td>Expected vocabulary size after ( k ) drawings</td>
</tr>
</tbody>
</table>
In Fig. 1 we see how nicely average growth can be fitted by a power function of the form \( nk^h \) in the case of a set of 100 elements (denoted as \( N = 100 \) in this figure; \( \theta \) and \( c \) refer to the parameters of the Mandelbrot distribution, and \( a_N \) is a normalization constant for this distribution that will be introduced in a later section).

However, this approximation expressed by Heaps’ Law is not valid everywhere. For one reason, the number of records is bounded by the total number of events, while a power function will exceed this number eventually. In order to express the limited validity of Heaps’ Law, we also focus on the validity area of the approximations in our analysis. The validity area is described rather defensively, in practice the area will be larger.

The structure of this paper is as follows. In Section 2 we discuss related work. In Section 3 we present a statistical model for the vocabulary size in a text, i.e. the average number of unique occurrences after a series of drawings. In Section 4 we solve the resulting equation, leading to Heaps’ Law. We also give bounds for the validity area of the approximations. In Section 5 we draw some conclusions and discuss further research.

2. Related work

In [3], Heaps’ Law and the generalized Zipf’s Law are related. It is shown that under a plausible assumption, it can be derived that if both Heaps’ Law and the generalized Zipf’ Law hold, the relation \( \beta = 1/\theta \) is a consequence. The argumentation is as follows. Heaps’ Law predicts the vocabulary size in a text consisting of a given number of words (the text size). It is assumed that the
frequency of the least frequent word in this text is $\Theta(1)$. As a result, the prediction of this frequency as obtained by the generalized Zipf’s is also $\Theta(1)$. Working out the details leads to $\beta = 1/\theta$. In our paper we go a step further, and derive Heaps’ Law directly from the Mandelbrot distribution, which has the generalized Zipf’ Law as a special case.

In [9], it is shown that lognormal distributions are comparable in their generative power to power law distributions such as Zipf Law and the Mandelbrot distribution. This suggest the validity of Heaps’ Law also for these distributions.

Related fundamental research may also be found in the context of the behavior of population frequencies of animals of various species which for example is described by Turing’s formula. See for example [5]. In [10], by techniques somewhat similar to ours, Turing’s Formula is related to Zipf’s Law.

3. A probabilistic model for Heaps’ law

Let $W$ be a set of $N$ words numbered $1, \ldots, N$, and let $p_i$ the probability that word $i$ is chosen. The underlying text model is that words are taken independently at random with replacement from the set $W$ according to this probability distribution. We will be interested in the asymptotic behavior (for $N \to \infty$) of the expected resulting number of different words taken.

After taking $k$ words $w_1, \ldots, w_k$ from $W$, let $D_k = \{w_1, \ldots, w_k\}$ be the set of different words, and let $n_k$ be the number of such words: $n_k = \#D_k$. Then obviously $n_k \leq k$. We analyze the drawing of the $k$th word for $k > 0$ in detail. There are two possibilities depending on whether this $k$th word has been drawn before. Let $a < k$, then

\[
\text{Prob}(n_k = a) = \text{Prob}(n_{k-1} = a - 1 \land w_k \notin D_{k-1}) + \text{Prob}(n_{k-1} = a \land w_k \in D_{k-1}) = \text{Prob}(n_{k-1} = a - 1) \ast \text{Prob}(w_k \notin D_{k-1}) + \text{Prob}(n_{k-1} = a) \ast \text{Prob}(w_k \in D_{k-1})
\]

As $a > k$ is not possible, we have in that case $\text{Prob}(n_k = a) = 0$. For the remaining case we have $\text{Prob}(n_1 = 1) = 1$.

Note that $\text{Prob}(w_k \in D_{k-1}) = 1 - \text{Prob}(w_k \notin D_{k-1})$. This latter probability is further elaborated by

\[
\text{Prob}(w_k \notin D_{k-1}) = \sum_{i \in W} \text{Prob}(w_k = i \land i \notin D_{k-1}) = \sum_{i \in W} p_i (1 - p_i)^{k-1}
\]

For notational convenience, let $S_k = \sum_{i \in W} p_i (1 - p_i)^{k-1}$, and $M_k = \sum_{i \in W} (1 - p_i)^k$. 

We will refer to $M_k$ as the $k$th reverse moment of the probability distribution. Then the following relation is an immediate consequence: $S_k = M_{k-1} - M_k$.

If we further simplify notation by introducing $N(k, a) = \text{Prob}(n_k = a)$, then we get the following recurrence relation:

$$
N(1, 1) = 1 \\
N(k, a) = 0 \quad \text{if } k < a \\
N(k, a) = N(k - 1, a - 1) * S_k + N(k - 1, a) * (1 - S_k) \quad \text{if } k \geq a
$$

We will be interested in the expected number of different words $N_k$ after taking $k$ words randomly from the set $W$ of words. Using this recurrence relation, we get for $N_k$ ($k > 1$) the following recurrence relation:

$$
N_k = N_{k-1} + S_k
$$

As a consequence:

**Lemma 1.** The expected number of different words in a random selection of $k$ words is $N_k = N - M_k$.

In order to estimate the asymptotic behavior of $N_k$ we will estimate $M_k$ in the next section.

### 4. Approximating reverse moments for Mandelbrot distribution

The Mandelbrot distribution provides a reasonable approximation of frequency of word usage in natural language. The Mandelbrot distribution assumes words to be ranked according to their frequency of usage. The probability of the word ranked at position $i$ then corresponds to

$$
p_i = a_N(c + i)^{-\theta}; \quad a_N = \left(\sum_{i=1}^{N} (c + i)^{-\theta}\right)^{-1}
$$

for some constants $c \geq 0$ and $\theta$, where $a_N$ is such that $\sum_{i=1}^{N} p_i = 1$.

Usually the constant $\theta$ ranges over $[1, 2]$. A special case is Zipf's Law, which results from the Mandelbrot distribution by choosing $\theta = 1$ and $c = 0$. In this paper we shall assume $c > 0$ and $1 < \theta \leq 2$. In this context the significant variables are $N$ and $k$.

We shall use some conventional $o, \sim$, and $O$-symbolism [2]. If both $N$ and $k$ are involved, $f = O(g)$ will mean that for some constants $C$ and $C' > 0$, $|f(N, k)| \leq C|g(N, k)|$ for all $N$ and $k$ that are $> C'$.

$f = o(g)$ means that for any $\epsilon > 0$ there exists a $C'(\epsilon) > 0$ such that $|f(N, k)| \leq \epsilon|g(N, k)|$, for all $N$ and $k > C'(\epsilon)$.  


Our main result will be a rigorous proof of Theorem 1 below, giving asymptotics for a vocabulary of large size \( N \) and a random sample of size \( k \), also large.

However, \( k \) must be reasonably small with respect to \( N \)-namely, such that the error term of Theorem 1 can be neglected. This is what occurs in practice: one observes that Heaps’ power Law does not hold any more for large \( k \). Theorem 1 displays this behavior very well.

**Theorem 1.** The expected number \( N_k \) of different words in a random selection of \( k \) words from \( N \) is

\[
N_k = \alpha k^\beta (1 + o(1)) + O\left( \frac{k}{N^{\gamma - 1}} \right) \left( N, k \to \infty, \frac{k}{N^{\gamma - 1}} \to 0 \right)
\]

where \( \beta = \theta^{-1} \) and \( \alpha \) is the constant \( a_\infty^\beta \Gamma(1 - \beta) \) with \( a_\infty = \lim_{N \to \infty} a_N \). (\( \Gamma \) is the well-known gamma function \([11]\).)

The error term \( O\left( \frac{k}{N^{\gamma - 1}} \right) \) is of smaller order than the main term \( a_\infty^\beta \Gamma(1 - \beta) k^\beta (1 + o(1)) \) if \( k \ll N^{\theta} \), and close to zero if \( k \ll N^{\theta - 1} \). The proof furthermore shows that the constants implied by the \( O \) and \( o \) symbols are not “large” indeed; the first is easily seen to be at most \( 2 a_N \). Thus we have rather precise information about the validity interval. As an immediate consequence we have:

**Theorem 2** (Heaps’ Law). \( N_k = \alpha k^\beta \); \( k, N \to \infty \), with validity region \( k \ll N^{\theta - 1} \).

The proof, though at places not totally straightforward, is conceptually simple. However, it involves some calculations the details of which we have conveniently placed in Appendix A. Here we shall describe the general argument.

I. We work with \( M_k = N - N - k \). Let \( t(x) = \alpha_N (c + x)^{-\theta} \) and \( \phi_k(x) = (1 - t(x))^k \), so that \( M_k = \sum_{i=1}^{N} \phi_k(i) \). First the summation is replaced by an integral:

\[
M_k = \int_{1}^{N} \phi_k(x) \, dx + O(1)
\]

II. An appropriate substitution transforms this integral into \( A \int_{t(1)}^{t(N)} (1 - t)^k t^\mu \, dt \) where \( \mu = -1 - \frac{1}{\beta} \), \( A = \frac{1}{\beta} a_N^{1/\beta} \).

III. The latter integral resembles an Euler Beta function \([11]\), but the integration interval is \([t(N), t(1)] \) instead of \([0, 1] \). By manipulations such as partial integration, the difference with the Beta function can be shown to equal \( N \) (needed since \( N_k = N - M_k \)) plus the right error term of Theorem 1.
IV. Finally, the Beta function can be expressed in terms of the Gamma function. Stirling’s approximation for the Gamma function now yields the main term \( a \cdot k^b \) of Heaps’ Law.

To our satisfaction, the exponent obtained from the Beta function calculation appeared to be exactly the heuristically plausible value \( \theta^{-1} \) (cf. [3]). Moreover, our approach yields a validity interval.

5. Conclusions and further research

In this paper we have derived the Heaps’ Law from the Mandelbrot distribution, and provided a validity area for Heaps’ Law. As a next step, a second order approximation may be employed, providing a sharper formulation for Heaps’ Law, and a larger validity area. Furthermore, other distributions may be examined, leading to Heaps’ Criterion as a sufficient condition for a distribution to imply Heaps’ Law.

Acknowledgements

The authors want to express their gratitude to the (anonymous) referees. Their comments in many improvements of this paper. The authors also want to thank Franc Grootjen, who was involved in an earlier version of this paper.

Appendix A

Here, we present the details of the estimate of the \( k \)th reverse Mandelbrot moment \( M_k \), and the expected number \( N_k = N - M_k \) of different words in a random selection of \( k \) words.

Notations and conventions. There are numbers \( c \) and \( \theta \) that belong to our probabilistic model and are fixed and satisfy \( c > 0 \) and \( 1 < \theta \leq 2 \). \( a_N \) is the Mandelbrot parameter \( (\sum_{i=1}^{N}(c+i)^{-\theta})^{-1} \) and \( a_\infty = \lim_{N\to\infty} a_N \).

It is convenient to have the further notations

\[
\beta = \frac{1}{\theta}; \quad A = \beta d_N; \quad \mu = -1 - \beta; \quad \text{and} \quad \alpha = a_\infty \Gamma(1 - \beta) \tag{A.1}
\]

with \( \Gamma \) the ordinary Gamma-function [11], and

\[
t(x) = a_N(c + x)^{-\theta}; \quad \phi_k(x) = (1 - t(x))^k; \quad \text{and} \quad \Psi(t) = (1 - t)^{k-1} t^{\mu+1} \tag{A.2}
\]
Note that $-2 < \mu \leq -\frac{1}{2}$ and $0 < t(x) < 1$ for all $x$. We shall frequently quote

**Remark 1.** $a_N = O(1)$; i.e. $a_N$ and $A = \beta a_N^\theta$ are bounded, uniformly in $N$ and $k$.

(Indeed: $a_{\infty} \leq a_N \leq a_1 - (c + 1)^\theta$.)

**Proof of Theorem 1.** We shall follow the lines of Section 4.

*Ad I.* For any $k > 0$ and $N$, $\phi_k(x) = (1 - t(x))^k$ is a monotonically increasing function: $[1, \infty) \to (0, 1)$; so the integral method (see [1]) applies: $\sum_{i=1}^{N-1} \phi_k(i) \leq \int_1^N \phi(x) \, dx \leq \sum_{i=2}^N \phi_k(i)$. Hence

$$M_k = \sum_{i=1}^N \phi_k(i) = \int_1^N \phi(x) \, dx + \epsilon$$

with error $|\epsilon| \leq \phi_k(1) + \phi_k(N) = (1 - a_N(c + 1)^{-\theta})^k + (1 - a_N(c + N)^{-\theta})^k$. By Remark 1, $a_N$ and thus $\epsilon$ are uniformly bounded in $N$ and $k$. In this way, the reverse moment $M_k$ is approximated by an integral.

*Ad II.* By substitution of $t(x) = a_N(c + x)^{-\theta}$, one has $dx = -At^\mu \, dt$ (A and $\mu$ as in as in (A.2)) and

$$\int_1^N \phi(x) \, dx = A \int_{t(N)}^{t(1)} (1 - t)^k t^\mu \, dt$$

*Ad III.* Integrating by parts, (A.4) can be written as

$$A \frac{(1 - t)^k t^{\mu+1}}{\mu + 1} \bigg|_{t(N)}^{t(1)} - A0k \int_{t(N)}^{t(1)} \Psi(t) \, dt,$$

with $\Psi(t) = (1 - t)^{k-1} t^{\mu+1}$

In terms of the original parameters, the first part of (A.5) equals

$$(c + N)\phi_k(N) - (c + 1)\phi_k(1).$$

Now $\phi_k(1) \leq 1$. By Taylor expansion (and Remark 1), $\phi_k(N) = (1 - a_N(c + N)^{-\theta})^k = 1 + O\left(\frac{k}{(c+N)^{\theta}}\right) \left(\frac{k}{(c+N)^{\theta}} \to 0\right)$. In this way, the first part of (A.5) has been estimated as $O(1) + N + O\left(\frac{k}{(c+N)^{\theta}}\right)$. Thus, we obtain approximately the number $N$ of words in the set $W$, which is also the main term of the reverse moment $M_k$.

The second part of (A.5) is the integral $-A0k \int_{t(N)}^{t(1)} \Psi(t) \, dt$ that can be split into three terms as $-A0k \int_0^1 \Psi(t) \, dt + A0k \int_0^{t(N)} \Psi(t) \, dt + A0k \int_{t(1)}^1 \Psi(t) \, dt$.

- The second term equals $A0k \int_0^{t(N)} \Psi(t) \, dt = A0k \int_0^{t(N)} (1 - t)^{k-1} t^{\mu+1} \, dt$. Since $t(N) \to 0$ as $N \to \infty$, this term has order $A0k \cdot O\left(\int_0^{t(N)} 1 \cdot t^{\mu+1} \, dt\right) = O\left(kt(N)^{\mu+2}\right) = O\left(\frac{k}{(c+N)^{\theta}}\right)$ (the same error as came from the first part of (A.5)).
By partial integration (as with (A.4)), the third term is found to equal
\[ A\theta k (1 - t)^{k-1} \frac{t^{\mu+2}}{\mu + 2} \bigg|_{t(1)}^1 + \frac{A\theta k (k - 1)}{\mu + 2} \int_{t(1)}^1 (1 - t)^{k-2} \mu^{\mu+2} dt \]

By Remark 1, \( A \) is bounded. Also, \( \mu + 2 > 0 \) so
\[ A\theta k (1 - t)^{k-1} \frac{t^{\mu+2}}{\mu + 2} \bigg|_{t(1)}^1 = O(k(1 - t(1))^{k-1}) \quad (k \to \infty; \; \text{all } N). \]
Similarly,
\[ \frac{A\theta k (k - 1)}{\mu + 2} \int_{t(1)}^1 (1 - t)^{k-2} \mu^{\mu+2} dt \leq \frac{A\theta k (k - 1)}{\mu + 2} \int_{t(1)}^1 (1 - t)^{k-2} \mu^{\mu+2} dt = O((k^2(1 - t(1))^{k-1}) \quad (k \to \infty; \; \text{all } N). \]

Thus one observes that the third term decreases exponentially with \( k \).

The first term will be treated below.

Summarizing: up to now we have proved that
\[ M_k = -A\theta k \int_0^1 \Psi(t) dt + N + O(1) + O \left( \frac{k}{N^{\theta-1}} \right) \quad (A.6) \]
where \( \Psi(t) = (1 - t)^{k-1} \mu^{\mu+1} \) and \( N, k \to \infty, \frac{k}{N^{\theta-1}} \to 0 \).

Ad IV. Cf. [11], the integral in (A.6) can be recognized as a Beta function
\[ -A\theta k B(k, \mu + 2), \]
\[ \frac{\Gamma(k)}{\Gamma(k + \mu + 2)} \]
in terms of the \( \Gamma \) function. This expression is valid only for \( \mu + 2 \neq 0 \), equivalent with \( \theta \neq 1 \) (which in this paper we have assumed).

Cf. [11], one has Stirling’s approximation of the \( \Gamma \) function: \( \Gamma(x + 1) \sim \sqrt{2\pi x^x e^{-x}} \).

Applying this for the terms \( \Gamma(k) \) and \( \Gamma(k + \mu + 2), k \to \infty \), yields after some straightforward regrouping of factors
\[ -A\theta k \frac{\Gamma(k)\Gamma(\mu + 2)}{\Gamma(k + \mu + 2)} \]
\[ \sim -A\theta \Gamma(\mu + 2) \left( \frac{k - 1}{k + \mu + 1} \right)^{1/2} \frac{e^{\mu+2}}{(1 + \frac{\mu+2}{k-1})^{k-1}} \frac{k}{(k + \mu + 1)^{\mu+2}} \]
\[ \sim -a_N^\beta \Gamma(\mu + 2) \cdot 1 \cdot \frac{e^{\mu+2}}{e^{\mu+2}} \cdot (k - \mu - 1)^{-\mu-1} \left( \frac{1 + \frac{\mu + 2}{k - 1}}{k - 1} \right)^{k-1} \to e^{\mu+2}, k \to \infty \]
\[ = -a_N^\beta (1 - \beta) k^\beta \]

Finally, \( a_N^\beta = a_\infty^\beta (1 + o(1)), N \to \infty \). We now have obtained
\[ -A\theta k \int_0^1 \Psi(t) dt = -a_N^\beta \Gamma(1 - \beta) k^\beta (1 + o(1)) \]

Substituting this into (A.6) (absorbing the \( O(1) \)), and using \( N_k = N - M_k \) completes the proof of Theorem 1. \( \square \)
References

[9] M. Mitzenmacher, A brief history of generative models for power law and lognormal distributions, TR-08-01 Computer Science Group, Havard University, Cambridge, MA.