Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Predicting the long tail of book sales: Unearthing the power-law exponent

Trevor Fenner, Mark Levene *, George Loizou

Department of Computer Science and Information Systems, Birkbeck, University of London, London WC1E 7HX, UK

ARTICLE INFO

Article history:
Received 6 May 2009
Received in revised form 24 September 2009
Available online 24 February 2010

Keywords:
Long tail
Power-law distribution
Stochastic model

ABSTRACT

The concept of the long tail has recently been used to explain the phenomenon in e-commerce where the total volume of sales of the items in the tail is comparable to that of the most popular items. In the case of online book sales, the proportion of tail sales has been estimated using regression techniques on the assumption that the data obeys a power-law distribution. Here we propose a different technique for estimation based on a generative model of book sales that results in an asymptotic power-law distribution of sales, but which does not suffer from the problems related to power-law regression techniques. We show that the proportion of tail sales predicted is very sensitive to the estimated power-law exponent. In particular, if we assume that the power-law exponent of the cumulative distribution is closer to 1.1 rather than to 1.2 (estimates published in 2003, calculated using regression by two groups of researchers), then our computations suggest that the tail sales of Amazon.com, rather than being 40% as estimated by Brynjolfsson, Hu and Smith in 2003, are actually closer to 20%, the proportion estimated by its CEO.

1. Introduction

The long tail is the phenomenon that allows an e-commerce business to make significant profit from small sales volumes of a large number of less popular items. It is well known that in aggregate the tail sales of many online retailers, offering products such as books, music and films, are comparable to the sales of the most popular items, i.e. the “blockbusters”. The proportion of tail sales may be estimated using regression techniques on the assumption that the data obeys a power-law distribution. Newman, in Ref. [1], provides evidence that books sales do indeed follow a power law. In this paper we present a different method for estimating the tail sales based on a generative model that simulates a simplified sales process, and results in an asymptotic power-law distribution. Our generative model also has the advantage that the proportion of tail sales can be estimated from the number of available products and the total volume of sales, when these are known. The methodology we present may be useful in providing sales analytics for an e-commerce business in relation to prediction and validation of sales volumes from the tail.

A power-law distribution taking the mathematical form

$$g(i) = \frac{C}{i^{\tau}}, \tag{1}$$

where C and τ are positive constants, represents the proportion of observations having the value i. The constant τ is called the *exponent* of the distribution [1]. There are many well-known examples of power-law distributions [2]; for example, *Lotka’s law* states that the number of authors publishing a prescribed number of papers is inversely proportional to the

[*] Corresponding author.
E-mail addresses: trevor@dcs.bbk.ac.uk (T. Fenner), mark@dcs.bbk.ac.uk (M. Levene), george@dcs.bbk.ac.uk (G. Loizou).

0378-4371/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2010.02.021

Table 1
Product variety comparison for large Online and Brick-and-mortar retailers.

Product category	Online	Brick-and-mortar
Books	2,300,000	40,000–100,000
CDs	250,000	5,000–15,000
DVDs	18,000	500–1,500
Digital cameras	213	36
Portable MP3 players	128	16
Flatbed scanners	171	13

square of the number of publications. *Pareto's law*, which is a cumulative version of (1), states that the number of people whose personal income is above a certain level follows a power law with an exponent between 1.5 and 2. *Zipf's law*, which states that the relative frequency of a word in a text is inversely proportional to its rank, is the inverse of the cumulative power-law distribution for the proportion of words whose frequency is above a certain level. (We note that for $\tau > 1$ the cumulative distribution corresponding to the distribution (1), i.e. the proportion of observations greater than i, also follows a power law, but with exponent $\tau - 1$. Its inverse is a *Zipfian* distribution of frequency against rank, which also follows a power law, now with exponent $1/(\tau - 1)$.)

The tail of a power-law distribution decays polynomially, in contrast to the exponential decay characteristic of distributions such as the Normal and geometric. Power-law distributions are notoriously hard to fit [3], and often there is an exponential cutoff present in the power-law scaling, although this cutoff may only be observable in the tail of the distribution for extremely large data sets [4]. A power-law distribution with exponential cutoff [5] is of the mathematical form

$$g(i) = \frac{C\, q^i}{i^\tau}, \tag{2}$$

where $0 < q < 1$, and frequently $q \approx 1$.

The concept of the *long tail* has been recently popularised by Anderson [6] (see also www.longtail.com) and is currently used to explain the phenomenon in e-commerce where the total volume of sales of the items in the tail of a Zipfian distribution of sales volume against sales rank is comparable to that of the most popular items. One category of sales to which long tail analysis has been applied is online book sales. In Refs. [7,8] it was argued that the considerable increase of product variety in online book stores has a significant positive impact on consumer welfare. Their analysis also applies to other products such as CDs and DVDs. Table 1, taken from Ref. [7], shows the numbers of products available from Online and Brick-and-mortar stores.

In Ref. [9] an analysis of online book sales data was carried out to compare the demand and price competition between Amazon.com and BarnesandNoble.com. In order to analyse the long tail, the exponent of the assumed cumulative power-law distribution relating the sales rank of a book to the number of copies sold needs to be estimated. In Ref. [9] the estimate of $\tau - 1$ used was 1.2, while in Ref. [7] the slightly lower value of 1.1481 was used. (The sales rank of a book is one greater than the number of books that have sold more copies.) Based on the latter estimate of the power-law exponent and assuming, as shown in Table 1, that the most popular 100,000 titles are stocked in Brick-and-mortar stores, Brynjolfsson et al. [7] concluded that about 40% of Amazon.com's sales are represented by titles that would not normally be found in these stores. It is interesting to note that Jeff Bezos, the CEO of Amazon.com, thought that the 40% figure was too high and the real figure was closer to 20% [10]. So, assuming that Bezos is correct, how can we explain this discrepancy?

There is some inconsistency in the estimation of the power-law exponent for Amazon.com's sales data, and different researchers have reported values for $\tau - 1$ in the range from just below 1.0 to approximately 1.3 [7,9]. This is not surprising, since there are inherent difficulties in fitting power-law distributions [3,4] and it is often unclear whether or not the distribution is indeed a pure power law. We will show that the generative model we describe in the next section supports the exponent $\tau - 1$ being in the region of 1.1 for Amazon.com's sales data, assuming that Jeff Bezos's estimate of 20% tail sales is closer to reality than 40%.

A recent approach, which to a certain extent circumvents the above problems, is to assume a generative model that results in a distribution that is asymptotically either a pure power-law distribution or a power-law distribution with exponential cutoff, where q in (2) is close to 1.0 [1]. (The latter covers a wider range of real-world scenarios than a pure power law.) The details of such a model are given in Section 2. We use this model to investigate the possible range of power-law exponents that are consistent with the book data given in Table 1 corresponding to 20%, 30% or 40% of the sales being in the tail of the distribution. The methodology we use and our results are presented in Section 3, and analysis of a sparse data set from Ref. [6] is presented in Section 4. Finally, in Section 5 we give our concluding remarks.

2. A stochastic model exhibiting power-law behaviour with an exponential cutoff

The stochastic model presented in Ref. [5] can be described, in the context of the sales of products, and in particular books, as follows. We have at our disposal a countable number of urns, say $urn(i)$, $i = 1, 2, \ldots,$ where each urn contains a number of products, for example, books or CDs. A product is in $urn(i)$ if i copies of it have been sold since it entered the system. Initially all the urns are empty except $urn(1)$, which has one product in it (of which one copy has been sold). At time