Nonparametric Estimation of Search Query Patterns

Soohyung Joo sjoo@uwm.edu
Dietmar Wolfram dwolfram@uwm.edu
Suyong Song songs@uwm.edu

University of Wisconsin-Milwaukee

Abstract

In this poster, we adopted nonparametric regression as a method to identify the unique distribution of query log data collected from the Excite search service in May 2001. In Informetrics, parametric modeling has been widely used in tracing term frequency data, such as Zipf's law, Lotka's law, or Bradford's law. However, these traditional parametric methods have had limited application when detecting distributions for large datasets with a nonlinear pattern and a long tail. This study tested kernel regression as an alternative tool to model nonlinearity of term frequency patterns. The results indicated that the kernel regression produced an improved model fit compared to previous parametric approaches in modeling query patterns.

Keywords: power law, non-parametric estimation, kernel regression, query log analysis

Introduction

In informetrics, many researchers have attempted to generalize the patterns of different types of information production. Based on observed regularities in the distribution of data, mathematical models are developed to match the observed pattern (Wolfram, 2003). Query logs have been a focus of research to informetricians for several decades, and it is widely known that many size-frequency patterns show a reverse-J-shape with a long tail. This reverse J-shaped distribution has been modeled using parametric methods, in particular power law curves (Newman, 2005). Parametric models identify a specific equation, and attempt to estimate corresponding parameters from the observed data. For example, Zipf's law, which posits two parameters in its power law equation, has been most widely applied to explain a relation between term frequency and its rank. However, modeling based on power law sometimes misspecifies nonlinear $_{\text{patterns}}$ due to the noise mainly caused from long tails. Also, as the size of a dataset grows, it becomes more difficult to adequately fit the observed frequency distributions to mathematical models (Ajiferuke, Wolfram, & Famoye, 2006).

Nonparametric regression can be a compelling alternative to predominant power low method. Nonparametric regression is more flexible and accurate to articulate the nonlinearity of any function, as it determines the local shape of the conditional mean relationship (Blundell & Duncan, 1998). This study intends to adopt nonparametric estimation, to be more specific--kernel regression--as a way to model search query patterns. To the best of our knowledge, this study is the first attempt to apply kernel regression to estimate term frequency patterns.

Kernel Regression

The kernel regression is one of nonparametric methods in statistics that estimate a non-linear relation between random variables. The kernel regression can be represented with a regression function, g(x):

$$y_i = g(x_i) + \varepsilon_i,\ i = 1, 2, \ldots, n \quad (1)$$

Joo, S., Wolfram, D., & Song, S. (2013). Nonparametric estimation of search query patterns. *iConference 2013 Proceedings* (pp.919-924). doi:10.9776/13479

where *g(x)* is the unknown regression function and ε_i's are the independent and identically distributed zero mean errors (Eubank, 1999).

Unlike parametric regression, *g(x)* is a form of unknown smooth function. In this case, *g(x)* can be estimated based on the following nonparametric regression function:

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} Y_i K(\frac{x - X_i}{h})}{\sum_{i=1}^{n} K(\frac{x - X_i}{h})} \qquad (2)$$

where *K(·)* is the kernel function (weight function) that penalizes distance from the local position where the approximation is centered. *h* is the bandwidth that controls the width of the weight function (Takeda et al., 2007).

The selection of the kernel and bandwidth determines the shape and the smoothness of the estimation in kernel regression. The selection of the kernel function is open. *K(·)* can be selected from a Gaussian, Epanechinkov, or other forms that satisfy the requirements of the kernel function:

$$\int K(z)dz = 1\,;\ \int zK(z)dz = 0\,;\ \int z^2K(z)dz = c \qquad (3)$$

where c is a constant value (Takeda et al., 2007). In this study, an Epanechnikov kernel was selected as a weight function, which is commonly used in kernel regression:

$$K(u) = \frac{3}{4}(1 - u^2)\mathbf{1}_{\{|u| \le 1\}} \qquad (4)$$

Silverman's (1986) rule was employed for the bandwidth calculation (where σ is standard deviation of *x* and *n* is sample size):

$$h = 1.06 * \sigma * n^{-1/5} \quad (5)$$

Data Collection and Analysis

Search term frequency data extracted from the 2001 Excite query transaction log data set (Spink et al., 2002). The dataset consists of 587,145 non-repeating queries submitted to the Excite search engine in May 2001. Individual terms were identified within queries by parsing for standard delimiters, with exceptions for entries such as URLs and email addresses, which were treated as single terms. The term frequency distribution represents the number of terms that occur one time, two times, etc., up to a maximum of 129,170 times. The dataset yielded 1,538,120 tokens for 182,012 term types. The tokens represent the individual occurrences of specific terms. The term types represent the distinct terms entered by searchers. Even though the dataset is not very recent, it was appropriate to test the new method as it has shown typical query patterns represented by power law.

As is usually done in power law function fits, both the size of the query term (x-axis) and the number of query terms that occur with a given frequency (y-axis) were transformed using logarithmic scaling. This study fitted the transformed data using both parametric and nonparametric methods. In parametric estimation, both linear modeling and polynomial modeling, in particular quadratic equation modeling, were examined. In nonparametric analysis, kernel regression was conducted using MATLAB. *Root Mean Squared Error* (*RMSE*) and R^2 values were used to compare the model fits across three different estimations.

Results

First, we modeled the size-frequency function of query terms using both parametric and nonparametric methods. The observed data were fitted to test the power law identified in Equation 6, essentially a Lotka function:

$$Y = A/X^b \quad (6)$$

where *A* and *b* are parameters to be estimated. To easily visualize the patterns, we applied a logarithmic transformation to both axes:

$$ln(Y) = ln(A) - b*ln(X) \quad (7)$$
$$y_i = c + b_1 x_i \quad (8)$$

where $ln(Y) = y$, $ln(A) = c$, and $ln(X) = x$. In addition, quadratic regression was employed to achieve a better model fit:

$$y_i = c + b_1 x_i + b_2 x_i^2 \quad (9)$$

To the same function, we applied kernel regression based on an Epanechinkov kernel. As shown in Table 1, the kernel regression resulted in the lowest *RMSE* (0.154, R^2=.956), while linear and quadratic models resulted in values of .312 and .195, respectively.

Table 1 *Estimation of tf-count (log transformed)*

Model	*RMSE*	R^2	Estimates		
			c	*b1*	*b2*
Linear	0.312	.821	3.763	-1.270	
Quadratic	0.195	.930	5.994	-3.275	.426
Kernel	0.154	.956	(non-parametric)		

Using the parameters estimated from ordinary least squares (OLS), we plotted the obtained linear and second-order polynomial curves. As shown in Figure 1, the linear equation had some defects in correctly articulating the pattern of actual observations. In particular, there were evident limitations in modeling low frequency terms and the long tail. The curve was slanted toward the x-axis due to the relatively long tail. The y-intercepts of the linear and quadratic (polynomial) models were 3.763 and 5.994 respectively, far from the observed value (5.050).

On the other hand, the result of the kernel regression showed an improved estimation result by achieving a better *RMSE* value. The kernel regression result not only specified the unique feature of the long tails but also precisely followed the nonlinear pattern in the region of highly ranked terms.

Second, we modeled the token distribution by term frequency using the same methods. As the pattern was shaped as “V”, a linear model was not appropriate in this case (*RMSE*=6.263; R^2=0.172). Quadratic model showed a better model fit (*RMSE*=0.195; R^2=0.676) as it depicts “U” shape. Kernel regression exhibited a better model fit by achieving an *RMSE* value of 0.154 (R^2 = 0.798).

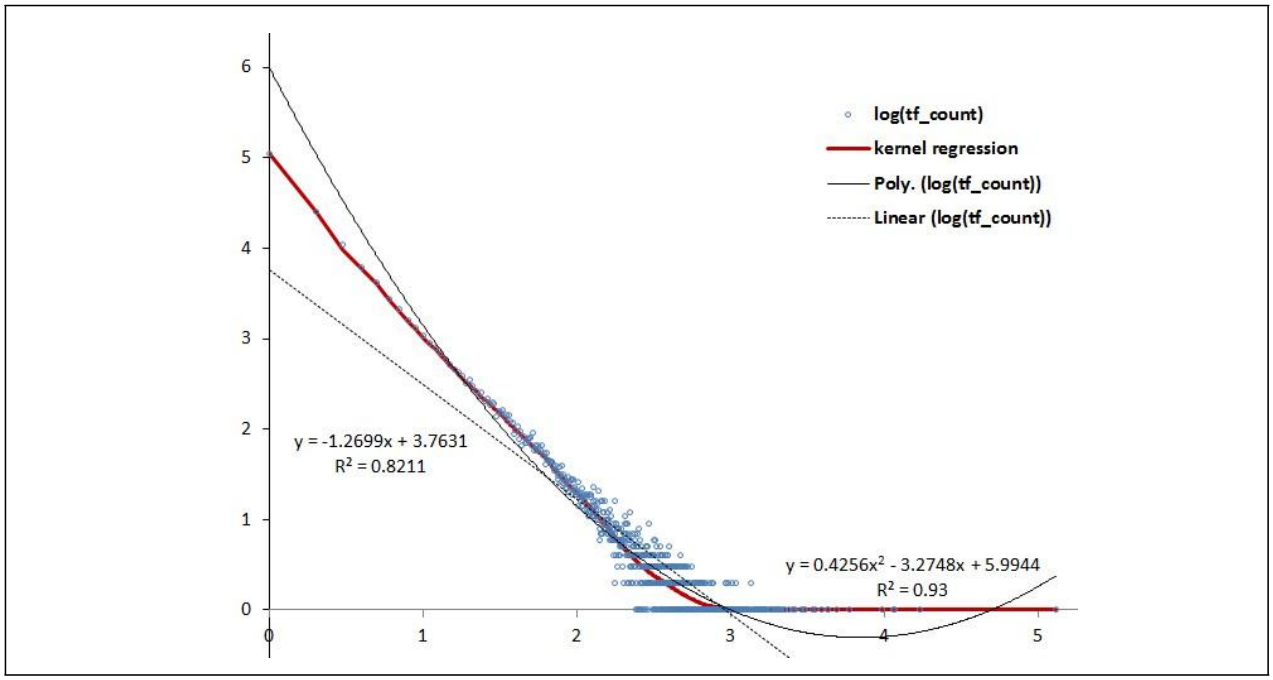

Figure 1. Size-frequency distribution model fits (log transformed)

Table 2 *Estimation of token frequency distribution (log transformed)*

Model	*RMSE*	R^2	Estimates		
			c	*b1*	*b2*
Linear	6.263	.172	3.763	-.270	
Quadratic	0.195	.676	5.994	-2.275	.426
Kernel	0.154	.798	(non-parametric)		

Figure 2 compares three lines of different estimation methods. As shown in Figure 2, the kernel regression presented more elaborated estimation than the other two parametric regressions.

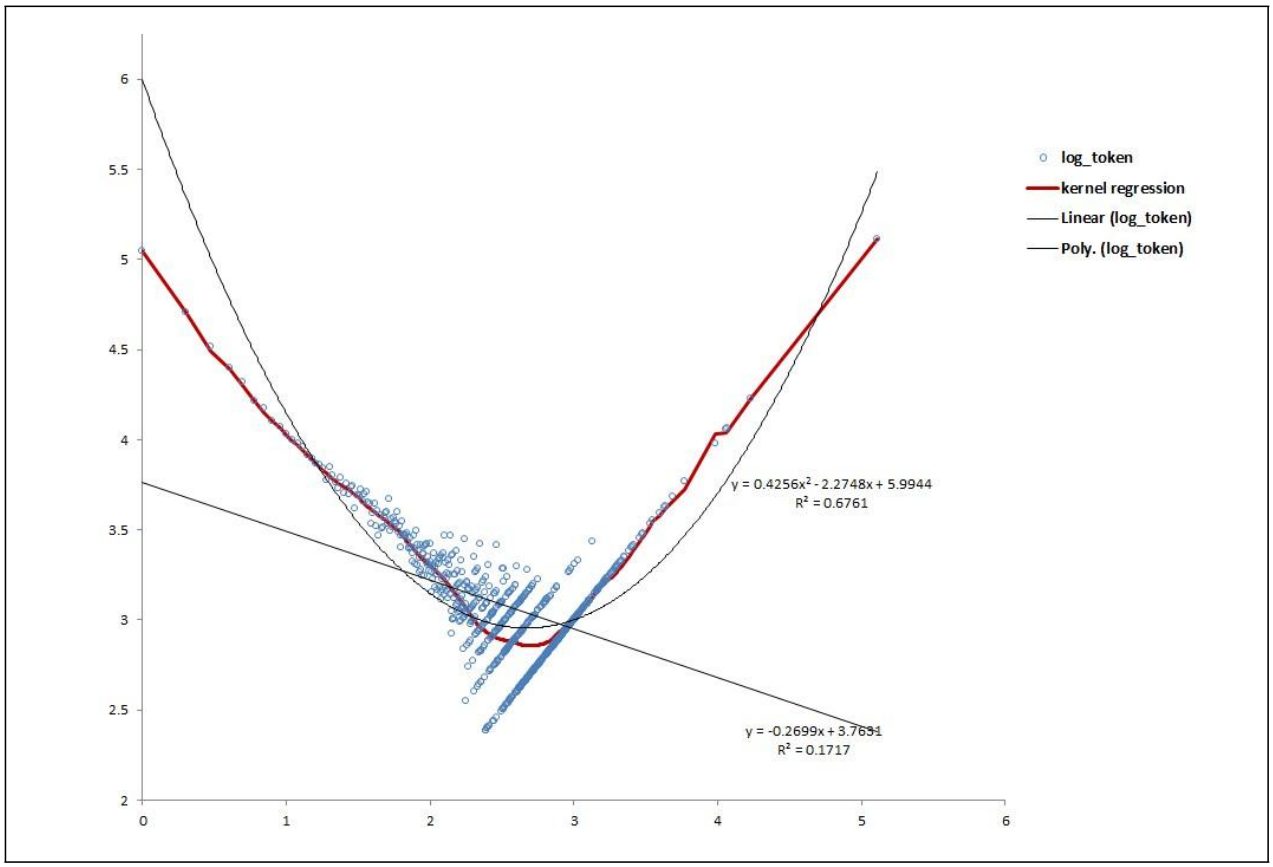

Figure 2. Token frequency distribution model fits (log transformed)

Conclusion

This study introduced nonparametric estimation to model query term frequency distribution in Informetrics. We compared first- and second-order regression models and a nonparametric method. This study supports that kernel regression could be useful in exploring query distributions, or potentially other frequency distribution data for large, nonlinear datasets with long tails. Effective modeling is particularly important for accurately predicting outcomes for informetric data. This has applications for more accurate estimation of informetric model outcomes not only in information retrieval environments, but also more generally for information production and use.

Classical parametric methods, represented by a power law in this example, rely on a specific model and seek to compute associated parameters in the presence of noise mostly coming from the long tail end and nonlinearity. In contrast to parametric methods, nonparametric estimation relies on the observed data itself to dictate the structure of the model (Wand & Jones, 1995).

In this poster, we described and proposed the use of kernel regression as an alternative tool for tracing query term frequency distribution patterns. This study yields some methodological contributions for the field of informetrics. Nonparametric methods can be applied to different types of data such as size-frequency form distributions used in journal productivity, author productivity, citation distributions and indexing exhaustivity distributions. As many data distributions found in metrics studies exhibit non-linear patterns and long tails, nonparametric regression can be useful to model observed data, particularly for larger, more difficult to model datasets which are now common.

References

Ajiferuke, I., Wolfram, D., & Famoye, F. (2006). Sample size and informetric model goodness-of-fit outcomes: A search engine log case study. *Journal of Information Science*, 32(3), 212-222.

Blundell, R. & Duncan, A. (1998). Kernel Regression in Empirical Microeconomics. *The Journal of Human Resources*, 33(1), 62-87.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing. Second edition.* New York: Marcel Dekker.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5), 323-351.

Silverman, B.W (1986*). Density Estimation for Statistics and Data Analysis, ser. Monographs on Statistics and Applied Probability*. New York: Chapman & Hall.

Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *IEEE Computer Magazine*, 35(3), 107-109.

Takeda, H., Farsiu, S., & Milanfar, P. (2007). Kernel regression for image processing and reconstruction. *IEEE Transactions on Image Processing,* 16(2), 349–366.

Wand, M. P. & Jones, M. C. (1995). *Kernel Smoothing, ser. Monographs on Statistics and Applied Probability.* New York: Chapman & Hall

Wolfram, D. (2003). *Applied informetrics for information retrieval research.* Westport, CT: Libraries Unlimited.