Predicting clicks of PubMed articles

Yuqing Mao, PhD[1], Zhiyong Lu, PhD[1]
[1] National Center for Biotechnology Information (NCBI), National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA

Abstract

Predicting the popularity or access usage of an article has the potential to improve the quality of PubMed searches. We can model the click trend of each article as its access changes over time by mining the PubMed query logs, which contain the previous access history for all articles. In this article, we examine the access patterns produced by PubMed users in two years (July 2009 to July 2011). We explore the time series of accesses for each article in the query logs, model the trends with regression approaches, and subsequently use the models for prediction. We show that the click trends of PubMed articles are best fitted with a log-normal regression model. This model allows the number of accesses an article receives and the time since it first becomes available in PubMed to be related via quadratic and logistic functions, with the model parameters to be estimated via maximum likelihood. Our experiments predicting the number of accesses for an article based on its past usage demonstrate that the mean absolute error and mean absolute percentage error of our model are 4.0% and 8.1% lower than the power-law regression model, respectively. The log-normal distribution is also shown to perform significantly better than a previous prediction method based on a human memory theory in cognitive science. This work warrants further investigation on the utility of such a log-normal regression approach towards improving information access in PubMed.

Introduction

Millions of users access biomedical articles in PubMed each day (1, 2). However, the total number of article access is not evenly distributed. Rather, less than half of the PubMed articles account for almost 90% of the access traffic during a given time period. Because the popularity of an article is a result of collective user behavior, from a sociological point of view, it may be correlated with the quality of the article. Previous studies (3) have confirmed there is a correlation between an article's accesses and its citations, which is commonly used as an indicator of its impact. Hence, if the access pattern of an article can be predicted, then relevant search results may be supplemented by document popularity for improved ranking (4).

Overall, PubMed articles exhibit very different and diverse access patterns (5). Some articles receive a large number of clicks in the few days after they are first indexed in PubMed but the number of clicks decay rapidly shortly after, while others have a much slower decrease in clicks over time. There are also articles that consistently receive very few clicks over time. In this paper, we present a study of utilizing the previous access history of an article to forecast their access patterns in the future, as part of our long-term investigation of PubMed logs (6-11). More specifically, we propose using the log-normal regression to model the click trend of each article, then estimating the parameters of the model based on the past article access data. There have been few previous attempts to address this problem. In particular, Goodwin and colleagues (12) adopted an approach based on the Anderson and Schooler model (13). In this study we use different modeling methods and conduct benchmarking experiments on a much larger PubMed document set (vs. a local medical library's proxy logs in (12)) with multiple evaluation metrics.

Predicting usage is a common approach to estimate values at some specific future times. Past research has shown that the future demands can be estimated as a function of past data using formal statistical methods, such as Bayesian inference (14), trend estimation (15) and time series analysis (16). These methods have applications in economics, medical research and many other areas. Regression analysis is one of the most widely used statistical techniques in forecasting and prediction, and a large number of techniques for carrying out regression analysis have been developed (17). For example, proportional hazards models, also known as Cox models (18), are well-recognized techniques for exploring the relationship between the survival of a patient and several explanatory variables.

In library sciences, Burrell proposed a stochastic model to predict the borrowing of books from a library collection based on the circulation statistics for a fixed period of time (19). This model was adapted by Anderson and Schooler to explain the usage of human memory (13). They elaborated the model by changing the underlying function to a

power-law distribution, which was derived from their assumption in cognitive science that the human memory retention functions tend to satisfy a power relationship. Recently, as social media has become more and more popular, there has been some research focusing on predicting the popularity of online social media content, such as video on YouTube, or news on Digg (20). For example, Szabo and Huberman proposed a methodology to predict the popularity of online content based on a finding of the correlation of popularity between early and later times (21). Their models were based on an observation that the logarithmically transformed long time popularity of an online content is highly correlated with its early measured popularity. Lee *et al.* used a Cox model to predict the number of comments for the threads of two online discussion forums (22).

Methods

PubMed access data

For this research, we obtained PubMed query logs for all articles that were accessed from July 1, 2009 to June 30, 2011. As such, we have daily click data for each paper that was accessed within this 2-year time duration. In order to compare with the previous study, a subset of 1,840,558 articles was selected and that each is associated with access data for more than 30 days during the 2-year time period.

Log-normal distribution

Previous work (12) suggested that the document access pattern follows a power-law distribution. When plotted on a log-log scale, the power-law curve appears to be linear, which means the access pattern of each article is monotonically decreasing. However, our query log data shows that the log-log plot of the access pattern is a parabola rather than a straight line.

Figure 1 shows the power-law plot of the clicking numbers for articles in different ages. A straight line is assumed to be able to fit the data if the clicking numbers are exponentially distributed, which can be expressed as:

$$\ln C(d) = \ln C(d_0) + \alpha \ln(\frac{d}{d_0}) \quad (1)$$

where $C(d)$ is the clicking numbers in day d, d_0 is the first day when the article becomes available in PubMed, and α is the decay which indicates the decreasing rate with d_0. As can be seen in Figure 1, despite a high correlation, the computed clicking numbers based on the equation 1 (the solid straight line) do not fit perfectly with the real data (in blue diamond dots).

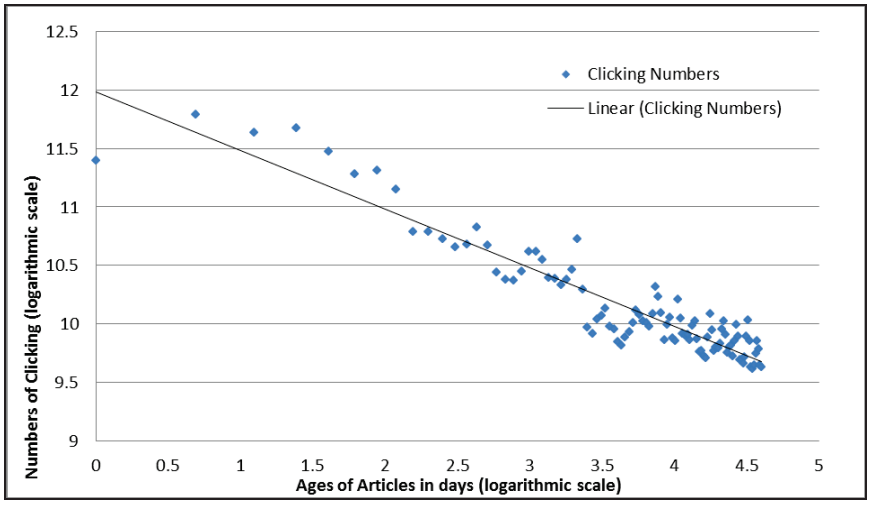

Figure 1. Scatter plot with the fitted line of clicking numbers versus article ages, both in natural log scales.

Further, when we plot the actual clicking numbers C(d) over the logarithm of article ages ln(*d*), it appears that the random variable follows a normal distribution with standard deviation σ (red curve in Figure 2). This suggests that the clicking number *C(d)* is distributed normally around an logarithmic-age-dependent variable, and the probability of being clicked can be derived based on the probability density function (PDF) of the normal distribution:

$$\Pr[y = x] = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

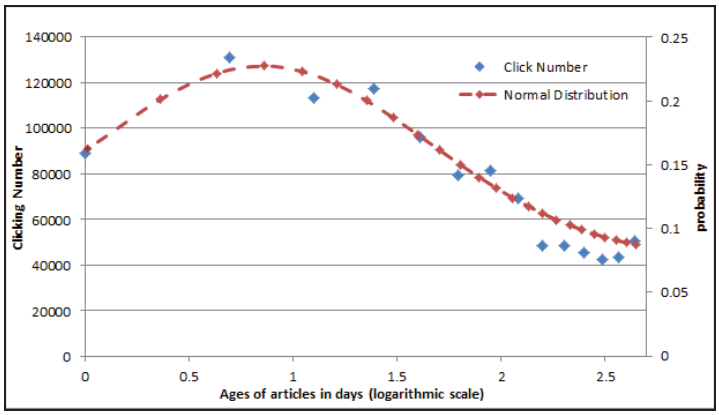

Figure 2. Scatter plot of actual clicking numbers versus article ages in natural log scale.

Let $x = \ln x'$; and because we need to keep the integration of density over all possible values as 1 ($\int \tilde{f}(x')dx' = \int f(\ln x')d(\ln x') = 1$), we have the distribution of clicking number over the age:

$$\tilde{f}(x') = f(\ln x')\frac{d(\ln x')}{dx'} = \frac{1}{x'\sigma\sqrt{2\pi}} e^{-\frac{(\ln x'-\mu)^2}{2\sigma^2}} \qquad (2)$$

Eq. 2 is a log-normal distribution. For estimating the two parameters μ and σ in log-normal distributions, we can use maximum likelihood because it is efficient in empirical settings:

$$\hat{u} = \frac{\sum_k \ln(x_k)}{n}, \quad \sigma^2 = \frac{\sum_k (\ln x_k - \hat{u})^2}{n} \qquad (3)$$

where (x_1, x_2, ..., x_n) is a sample of observations, *n* is the total observed number of clicks an article, and x_k is the day when the k^{th} click was observed. As such, we can predict the clicking number of an article at day *d* as:

$$C(d) = C_t \cdot \frac{1}{d\sigma\sqrt{2\pi}} e^{-\frac{(\ln d-\hat{\mu})^2}{2\sigma^2}} \qquad (4)$$

where C_t is the total clicking number of an article in its lifetime, which can be estimated based on the observation and Eq. 2:

$$C_t = \frac{n}{\sum_m \tilde{f}(m)}$$

where *n* is the total observed number of clicks, and *m* is the day when the clicks were observed.

From Eq. 4, we can have:

$$\ln C(d) = \ln C_t - \frac{1}{2\sigma^2}(\ln d)^2 + (\frac{\mu}{\sigma^2} - 1)\ln d - \ln\sqrt{2\pi}\sigma - \frac{\mu^2}{2\sigma^2} \qquad (5)$$

Let $\alpha = -\frac{1}{2\sigma^2}, \beta = \frac{\mu}{\sigma^2} - 1, \gamma = \ln C_t - \ln\sqrt{2\pi}\sigma - \frac{\mu^2}{2\sigma^2}$ and we can estimate the parameters of Eq. 5 using least squares, which appeared to be more accurate in our prediction than the maximum likelihood method in Eq 3. It should be

noted that, as pointed by Mitzenmacher (23), if σ is sufficiently large, then the quadratic term of Equation 6 will be small enough so that the result will appear almost linear for a large range of values. That explains why the log-normal distribution was often approximated by power-law distributions in previous studies (24).

Anderson and Schooler Model

This Anderson and Schooler model (13) predicts the odds of a memory being retrieved based on the recency and frequency of past retrieval (referred as Recency of Access (ROA) and Frequency of Access (FOA) respectively). Although such a model was for explaining the human memory phenomena in psychological area when first proposed, it is actually applicable to many other prediction tasks.

When both ROA and FOA models were applied for predicting document access (12), Equation 6 was used where the parameter *d* is the decay that controls how quickly the access of an article decreases. The parameter *n* is the total number of accesses for a given document. The parameter t_k is the time (in days) since the *k*th access of the article.

$$\ln\left(\sum_{k=1}^{n} t_k^{-d}\right) \qquad (6)$$

By such calculation, a higher score will be assigned to a document if it has been accessed within a shorter and more recent period. The basic assumption is that the "strengths" of individual accesses decay as a power function of the time elapsed between two consecutive article accesses. Thus if the document is accessed recently (t_n is small) and it is a short time for all accesses (t_1 is small), the total "strength" will be high.

The authors in (12) also proposed to use FOA along to predict document access as in Equation 7, which essentially assumes that the *n* accesses are evenly spaced over a given period of time T.

$$\ln\left(\frac{n}{1-d}\right) - d\ln(T) \qquad (7)$$

In (12), *d* was empirically determined to be 0.1 and 0.05 in (6) and (7) respectively based on the data for all articles that were accessed fewer than 100 times, which in fact ignores the fact that the decreasing rate of access number vary between articles.

Power-law distribution

Instead of using the Anderson and Schooler Model as in (12), we also explored power-law distribution directly to model the correlations between past and later times of the logarithmically transformed document accesses. Such a correlation can be described by a linear model:

$$\ln C(d) = \ln\left[r(d,d_0)C(d_0)\right] = \ln C(d_0) + \ln r(d,d_0) \qquad (8)$$

where *C(d)* is the clicking numbers of an article in day *d*; d_0 is the day where the article is available; $r(d, d_0)$ accounts for the linear relationship between the log-transformed accesses at different times.

The straight-line on the log-log plot is often regarded as the signature of a power-law distribution. The main attribute of power-law is its scale invariance. In general, a power-law distribution has the form:

$$f(x) = ax^k$$

which produces the linear relationship when logarithms are taken of both $f(x)$ and x. The linearity of accesses on the logarithmic scale makes it possible to predict the number of clicking at any given time in the future as they are approximated to be a constant product of the accesses measured at an earlier time. Similar to Eq. 4, the clicking number of an article at day *m* can be calculated as:

$$am^k \qquad (9)$$

The parameter values for the predictions *a*, similar to $\hat{u}$ and σ in Eq. 4, can be obtained with maximum likelihood fitting from the training data as in the case of the log-normal distribution.

Evaluation metrics

We first evaluated the results with the correlation coefficients between the predicted and actual number of access. Suppose $\mathbf{X}=[x_1, x_2,\ldots x_n]$ and $\mathbf{Y}=[y_1, y_2,\ldots y_n]$ are series of predicted and actual clicking numbers of *n* articles respectively, the sample correlation coefficient is used to estimate the Pearson correlation *r* between **X** and **Y**:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2(y_i-\bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of **X** and **Y**, and s_x and s_y are the sample standard deviations of **X** and **Y**.

In addition, we used two commonly used metrics for measuring prediction errors (differences between predicted vs. actual number of access): Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i \quad y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i| \qquad MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|x_i \quad y_i|}{x_i+1.0}$$

As the name suggests, MAE is an average of the absolute errors $|e_i| = |x_i \quad y_i|$. Note that the MAPE formulation is includes an adjustment factor *1.0* as the actual value of access may be 0.

Results

Table 1 shows the correlation between the predicted data and actual data using different methods and parameters. As in (12), the results were based on the prediction for a 30-day period using access data from past 364 days. Using the exact prediction models and parameter values in (12), we obtained the correlation scores of 0.634 for the model based on FOA and ROA, and 0.656 for the model based on FOA along. Compared to the results previously reported (See Figure 5 and Figure 6 in (12)), FOA and ROA results on our log data were lower, especially for the model based only on FOA (0.656 vs. 0.932).

To optimize performance using the Anderson and Schooler model, we modified the value of parameter d based on our dataset, and by doing so we were able to achieve higher performance (setting *d* to be 0.5 for the ROA and FOA model and 0.9 for the FOA model respectively). However, despite this effort, our correlation scores (0.638 and 0.879) are still lower than those results reported in (12). We believe such a discrepancy in performance is likely due to the discrepancies between the two datasets. As pointed out in (12), their results were specific to the Houston Academy of Medicine-Texas Medical Center (HAM-TMC) users, while PubMed users include both professionals and the general public (i.e. not just the academic users). Our data is also different from theirs in that we include an article's accesses since it first becomes available in PubMed. No such arrangements were made to the dataset in (12).

Table 1. Correlation results of different methods

Methods	Pearson Correlation
ROA and FOA (d=0.1)	0.631
FOA (d=0.05)	0.656
ROA and FOA (d=0.5)	0.638
FOA (d=0.9)	0.879
Power-law distribution	0.885
Log-normal distribution	**0.891**

By contrast, the power-law and log-normal distributions both yield higher correction scores (0.885 and 0.891 respectively) than the Anderson and Schooler model. One plausible explanation for this is that the parameters of these two models were derived for each article individually while the parameters of the Anderson and Schooler model were considered identical for all articles. As such, the two higher-performing models are able to take into account the fact that different articles may have different decay rate in terms of the number of accesses after they become available in PubMed.

Table 2 shows the prediction results with a moving window using both the power-law and log-normal models. That is, unlike the previous experiments in Table 1 where the same 364 days were used for predicting future access for each of the next 30 days, we chose to use the last 364 days when predicting for the very next day. So even though the number of days for each prediction did not change (fixed to be 364), the actual 364 days for each of predicted 30 days were different. As can be seen in Table 2, the log-normal model yielded smaller errors and higher correlation scores than the power-law model. Although the overall trend in Table 2 is consistent with that in Table 1, both correlations in Table 2 are greater than those in Table 1, suggesting that the prediction with a moving window is more effective than using a fixed time period.

Table 2. Prediction with a moving window and with access data from past 364 days.

Methods	Correlation	MAE	MAPE
Power-law distribution	0.893	0.7490	32.78%
Log-normal distribution	0.907	0.7204	30.97%

Finally, in addition to using a moving window, we evaluated results by tuning another parameter: the different number of days to be used in prediction. To find a time period that can achieve the most accurate prediction, we tested our two models with historical access from the 2 most recent days before the prediction day to the entire lifetime of the article. Table 3 shows that the power-law model achieved the smallest error and highest correlation if the previous 76 days was used for prediction, while the log-normal model achieved the best performance with data from the previous 204 days. Once again, the table shows that the log-normal model predicts the article's accesses more accurately than the power-law model. Note that in this case, the MAPE score of the log-normal model is less than 30%, which is typically considered as good performance for forecasting (25).

Table 3. Prediction with a moving window and with optimized length of past access data.

Methods	Correlation	MAE	MAPE
Power-law distribution (document access data from past 76 days)	0.915	0.7301	31.70%
Log-normal distribution (document access data from past 204 days)	0.939	0.7012	29.13%

Discussion

To understand in what typical circumstances the log-normal regression model performs better than the power-law model, we show two real-world cases of article access, along with the prediction patterns of the two models. As shown in Figure 2, the power-law regression model (the red line) has difficulty in making accurate predictions based on a short period past usage. Such a situation applies to those new PubMed articles that have only limited past access data and before the prediction and the escalating trend of its access had just ended (Figure 2), the power-law regression model would forecast a reverse trend for such articles, while the log-normal regression model (the green line) could model such situations more correctly.

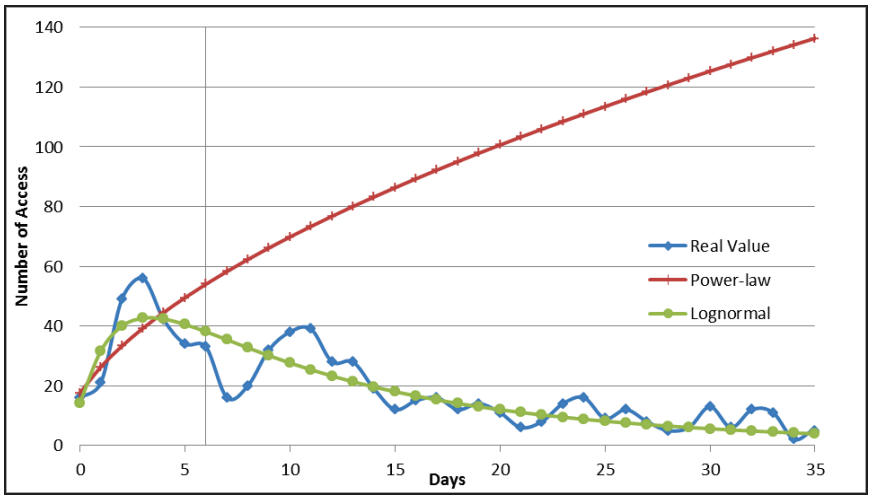

Figure 2. Actual and predicted access trends of a PubMed article. Both prediction methods were executed at day 6 as illustrated by the gray vertical line.

The log-normal regression model performs better also in cases like Figure 3. Despite the power-law regression model captures the correct trend in general, the variations in the first 20 days of the article made the predicted value by power-law regression model shape like a flat curve.

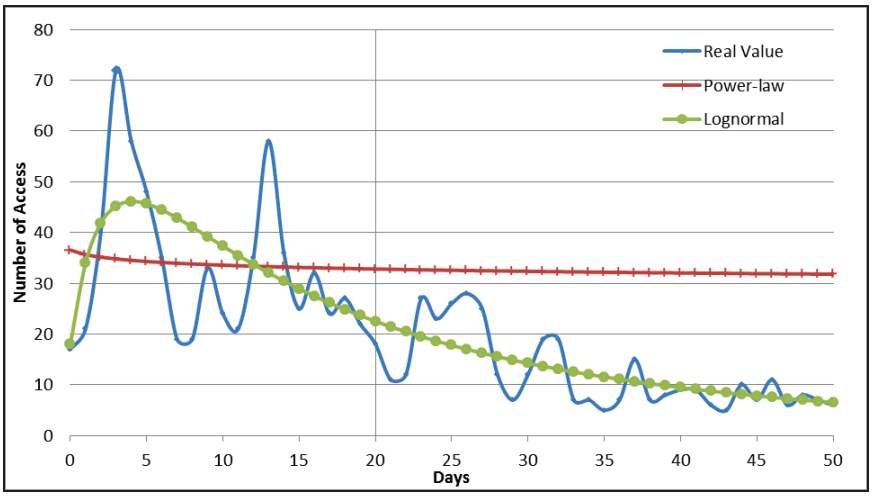

Figure 3. Actual and predicted access trends of a PubMed article. Both prediction methods were executed at the day 20 as illustrated by the gray vertical line.

The log-log plots of the actual and predicted access for the above articles are shown in Figure 4 and Figure 5 respectively. These figures further illustrate that fitting the specific historical data with a straight line in these cases results in a wrong trend (Figure 4), or could not accurately capture the gradient of the data (Figure 5). As log-normal distributions sometimes are very similar to power-law distributions and are often mistaken for power-law distributions (26), these two examples demonstrate that the log-normal distribution is more appropriate in describing the access trends in PubMed.

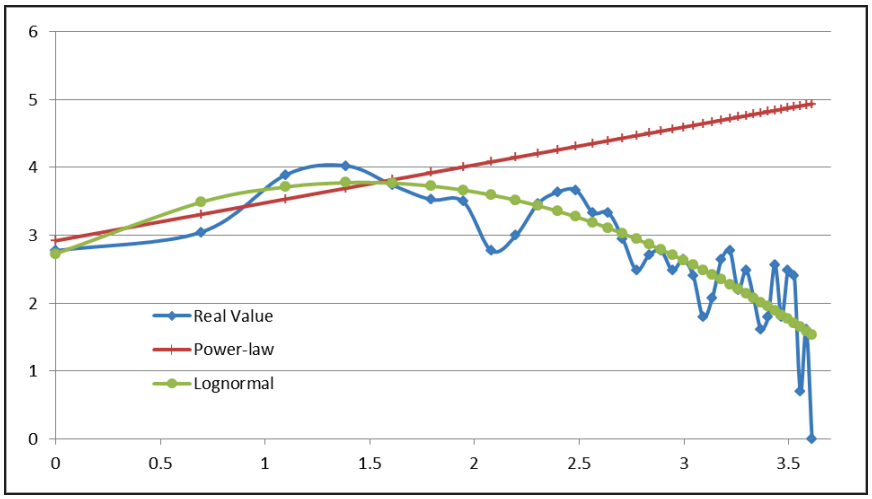

Figure 4. Log-log plot of same data in Figure 2.

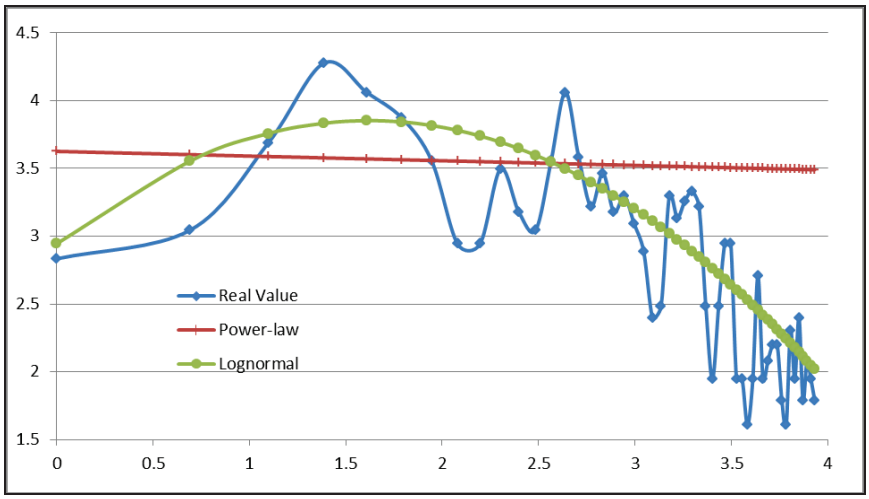

Figure 5. Log-log plot of same data in Figure 3.

However, it should be pointed out that the log-normal regression model also needs sufficient historical data to forecast the correct access trend. The historical data should contain at least one extra day of the access after reaching peak access (for most articles in our dataset, the peak of access generally happens within the first 7 days). Otherwise the log-normal regression model is unable to locate the vertex of the parabola and is subject to making erroreous predictions. Moreover, the log-normal regression model takes advantage over the power-law regression model in forecasting articles with a relatively short period of historical data. For articles with a long period of historical data, the escalating trend of the access in the early days would have less effect on the parameter estimation, and the access of the article falls into a monotonically decreasing interval. Thus the advantage of the log-normal regression model based on the log-normal distribution is weakened.

Finally, as we mentioned before, for some articles their access patterns do not always follow the power-law or log-normal distribution. Figure 6 shows such an example where the article became available on October 2010, and its usage gradually decreased since then. However from May to June 2011, the article received a burst of access. In cases like this, neither of the methods was able to make accurate predictions accordingly.

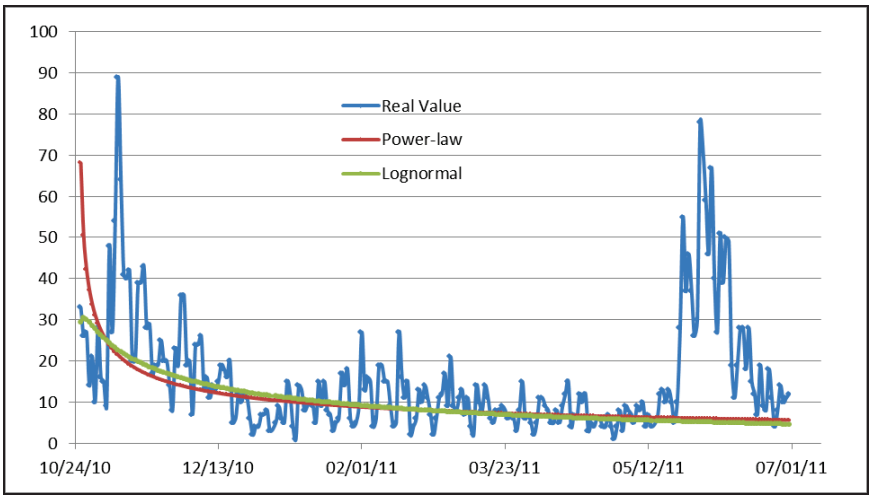

Figure 6. Actual and predicted access trends of a PubMed article

Conclusions

Our results suggest that the access pattern of PubMed articles can be forecasted based on the previous usages of the articles. The performance of the proposed log-normal regression and power-law regression models are better than previously reported the FOA and ROA models (12) because our two models allow parameter estimation for each article individually. More importantly, the fact that predictions by the log-normal model fit best with the actual data indicates that the article access pattern might follow a log-normal distribution, rather than a power-law distribution as previously reported.

Several research issues remain for the future work. First as Cha and colleagues suggested, the exact popularity distribution might depend on the content (27). Hence, it would be useful to know in what situations the article access distribution can be suitably approximated by a power-law distribution. We would also like to improve the prediction accuracy through combining other techniques with the proposed model. Finally, as PubMed articles currently are not ranked by relevance or popularity (28, 29), we would like to enable efficient ranking of the articles, based on probabilistic IR models integrated with the predicted access probabilities as the prior distribution.

Acknowledgements

We would like to thank Dr. John Wilbur for his help discussion on this project and Dr. Eric Sayers for providing us the PubMed document access data. We also thank Dr. Robert Leaman for proofreading this manuscript.

Funding: This research is supported by NIH Intramural Research Program, National Library of Medicine.

References

1. Herskovic JR, Tanaka LY, Hersh W, Bernstam EV. A day in the life of PubMed: Analysis of a typical day's query log. Journal of the American Medical Informatics Association. 2007;14(2):212-20.
2. Dogan RI, Murray GC, Névéol A, Lu Z. Understanding PubMed® user search behavior through log analysis. Database: the journal of biological databases and curation. 2009;2009.
3. Lawrence S. Free online availability substantially increases a paper's impact. Nature. 2001 May;411(6837):521-.
4. Chapelle O, Zhang Y. A dynamic bayesian network click model for web search ranking. Proceedings of the 18th international conference on World wide web; 2009: ACM; 2009. p. 1-10.
5. Smith L, Wilbur W. The Popularity of Articles in PubMed. Open Information Systems Journal. 2011;5:1-7.
6. Doğan RI, Lu Z. Click-words: learning to predict document keywords from a user perspective. Bioinformatics. 2010;26(21):2767-75.

7. Lu Z, Wilbur WJ, McEntyre JR, Iskhakov A, Szilagyi L. Finding query suggestions for PubMed. AMIA Annual Symposium Proceedings; 2009: American Medical Informatics Association; 2009. p. 396.
8. Lu Z, Xie N, Wilbur WJ. Identifying related journals through log analysis. Bioinformatics. 2009;25(22):3038-9.
9. Lu Z, Wilbur WJ. Improving accuracy for identifying related PubMed queries by an integrated approach. Journal of biomedical informatics. 2009;42(5):831-8.
10. Névéol A, Islamaj Doğan R, Lu Z. Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. Journal of biomedical informatics. 2011;44(2):310-8.
11. Li J, Lu Z. Developing Topic-specific Search Filters for PubMed with Click-through Data. Methods of information in medicine. 2013;52(4).
12. Goodwin JC, Johnson TR, Cohen T, Herskovic JR, Bernstam EV. Predicting biomedical document access as a function of past use. Journal of the American Medical Informatics Association. 2012;19(3):473-8.
13. Anderson JR, Schooler LJ. Reflections of the environment in memory. Psychological science. 1991 Nov;2(6):396-408.
14. Box GEP, Tiao GC. Bayesian inference in statistical analysis: DTIC Document; 1973.
15. Greenland S, Longnecker MP. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. American Journal of Epidemiology. 1992;135(11):1301-9.
16. Brockwell PJ. Time Series Analysis: Wiley Online Library; 2005.
17. Denison DGT. Nonparametric Bayesian regression methods. Proceedings of the Section on Bayesian Statistical Science; 1998: Citeseer; 1998.
18. Cox DR. Regression Models and Life-Tables. J Roy Stat Soc B. 1972;34(2):187-220.
19. Burrell Q. A simple stochastic model for library loans. Journal of Documentation. 1980;36(2):115-32.
20. Lerman K, Hogg T. Using stochastic models to describe and predict social dynamics of web users. ACM Transactions on Intelligent Systems and Technology (TIST). 2012;3(4):62.
21. Szabo G, Huberman BA. Predicting the popularity of online content. Communications of the ACM. 2010;53(8):80-8.
22. Lee JG, Moon S, Salamatian K. An approach to model and predict the popularity of online contents with explanatory factors. Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI) 2010; 2010; 2010.
23. Mitzenmacher M. A brief history of generative models for power law and lognormal distributions. Internet mathematics. 2004;1(2):226-51.
24. Clauset A, Shalizi CR, Newman ME. Power-law distributions in empirical data. SIAM review. 2009;51(4):661-703.
25. Lewis C. Demand forecasting and inventory control: Routledge; 2012.
26. Newman MEJ. Power laws, Pareto distributions and Zipf's law. Contemporary physics. 2005;46(5):323-51.
27. Cha M, Kwak H, Rodriguez P, Ahn YY, Moon S. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. Proceedings of the 7th ACM SIGCOMM conference on Internet measurement; 2007: ACM; 2007. p. 1-14.
28. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. Database: the journal of biological databases and curation. 2011;2011.
29. Lu Z, Kim W, Wilbur WJ. Evaluating relevance ranking strategies for MEDLINE retrieval. Journal of the American Medical Informatics Association. 2009;16(1):32-6.