Predicting biomedical document access as a function of past use

J Caleb Goodwin,1 Todd R Johnson,2 Trevor Cohen,1 Jorge R Herskovic,1
Elmer V Bernstam1,3

ABSTRACT

Objective To determine whether past access to biomedical documents can predict future document access.

Materials and methods The authors used 394 days of query log (August 1, 2009 to August 29, 2010) from PubMed users in the Texas Medical Center, which is the largest medical center in the world. The authors evaluated two document access models based on the work of Anderson and Schooler. The first is based on how frequently a document was accessed. The second is based on both frequency and recency.

Results The model based only on frequency of past access was highly correlated with the empirical data ($R^2=0.932$), whereas the model based on frequency and recency had a much lower correlation ($R^2=0.668$).

Discussion The frequency-only model accurately predicted whether a document will be accessed based on past use. Modeling accesses as a function of frequency requires storing only the number of accesses and the creation date for the document. This model requires low storage overheads and is computationally efficient, making it scalable to large corpora such as MEDLINE.

Conclusion It is feasible to accurately model the probability of a document being accessed in the future based on past accesses.

BACKGROUND AND SIGNIFICANCE

Introduction

Thanks to science and technology, access to factual knowledge of all kinds is rising exponentially while dropping in unit cost...We are drowning in information, while starving for wisdom.—E.O. Wilson, Consilience

Information overload in the biomedical domain is well documented.2–4 In fact, some have argued that expertise in medical subspecialties, given the amount of published literature, can only be obtained just as it is time to retire.5 Because of information overload, significant research effort has focused on information retrieval (IR) systems.

An important goal of IR systems is to prioritize documents likely to be needed from an expansive corpus. In investigating this optimization problem, we looked at similar domains where models exist for estimating the probability of items being retrieved from a large set. The first domain is library science where the problem is predicting the book most likely to be checked out based on past use.6 The second domain is cognitive science where the problem is modeling how human memory selects the memory with the highest probability of being needed based on past use.7 Burrell6 showed that library book circulation can be predicted based on past use. Anderson and Schooler7 linked the optimization problems faced in predicting library book circulation and human memory by adapting Burrell’s model to predict the probability of a memory being accessed as a function of past use.

In this paper, we show that the Anderson and Schooler model, when applied to biomedical documents, successfully predicts the odds of a document being accessed (abstract view or full-text download) based on past use.

There are several practical applications of this work. Previous studies found that relevance judgments can be extracted with high precision from document access patterns,8 9 and document accesses correlate well with citation counts.10 These findings indicate that document accesses contain meaningful information that can be leveraged to improve ranking. Ranking is vital because result sets can be large, and the majority of users look only at the first page of results.11 12 A document access can be viewed as a vote, and in aggregate, these votes can predict whether a document will be accessed in the future. We can then prioritize documents that are most likely to be accessed and thereby provide improved ranking for IR systems.

Another application is enhancing Bayesian IR models, which are a particular type of probabilistic IR model based on Bayes’ theorem. Bayesian IR models require the calculation of the prior probability of a document being relevant. The most common assumption is asserting that documents have an equal probability of access (uniform prior).13–15 Our analysis shows that this assumption is invalid and that the Anderson and Schooler model provides a theoretically motivated method for estimating the prior probability of a document being relevant. In the remainder of this paper, we evaluate the extent to which these findings apply to biomedical document retrieval.

Power-law distribution

Power-law distributions have a large number of small events and a few very large events. For example, if the heights of humans followed a power-law distribution, the majority of people would be a foot tall, and a few rare people would be hundreds or thousands of feet tall.15 Data that follow a power-law distribution can be fitted with a straight line on a log–log scale.

Modeling desirability

Burrell6 modeled library book access as a Poisson process, which predicted the odds that a book
would be checked out as a function of the frequency of past circulation. Burrell defined desirability as a quantifiable property that reflects the probability of an item being accessed. Anderson and Schooler extended Burrell’s model by changing the underlying function to a power-law distribution and extended the model to make predictions based on recency as well as frequency of access.

Anderson and Schooler investigated the statistical regularities of information in different environments to determine how frequency and recency of use affect the odds that an item will be needed in the future. Specifically, they looked at the appearance of words in *New York Times* headlines, utterances spoken by children as a function of words heard, and email correspondences. In all of these situations, the odds of an item appearing as a function of frequency and recency of its past appearance followed a power-law distribution.

Based on the results of the analysis, Anderson and Schooler developed a desirability model which predicts the odds of a memory being retrieved based on the recency and frequency of past retrieval (henceforth Recency of Access (ROA) and Frequency of Access (FOA)). The model is applicable to any domain where the data follow a power-law distribution, the odds of needing an item based on ROA follow a power-law distribution, and the odds of needing an item based on FOA follow a power-law distribution.

**OBJECTIVE**

Our objective was to determine if a model of human memory can predict document accesses based on a quantifiable property known as desirability (probability of an item being accessed).

Many factors impact the desirability of a given document including the publication type (review article, meta analysis, etc), publication date, authors, journal, and citation count. It is important to emphasize that our goal was not to determine which of these factors influence desirability, but to model desirability, which is the aggregate influence of these factors over time. In support of this aim, we show that the statistical regularities of document accesses meet the assumptions of the Anderson and Schooler model. Specifically, we show that (1) document accesses follow a power-law distribution; (2) odds of needing a document based on frequency of past use follow a power-law distribution; and (3) odds of needing a document based on how recently the document was last accessed follows a power-law distribution. Finally, after verifying that document accesses satisfy the assumptions of the Anderson and Schooler model, we evaluated two versions for modeling biomedical document accesses.

**METHODS AND MATERIALS**

This work was approved by the UTHSC-H Committee for the Protection of Human Subjects (Institutional Review Board). We used data from the Houston Academy of Medicine-Texas Medical Center (HAM-TMC) library, which is located in the largest medical center in the world and provides access to information resources for 49 institutions including numerous hospitals, medical schools, nursing schools, public health, and dentistry among others. We used logs from the HAM-TMC server that records queries and document accesses for authenticated users. HAM-TMC users were approximately 80% faculty, with the remaining 20% being administrative employees and students (HAM-TMC library staff, personal communication). We analyzed approximately 2.9 million PubMed queries and approximately 5 million document accesses from August 1, 2009 to August 29, 2010 (394 days). The logs contained search transactions for 14,109 users who accessed 959,491 unique documents representing 4.27% of the articles searchable via PubMed as of August 29, 2010.

The first three experiments investigated whether documents accessed through PubMed satisfied the three assumptions of the Anderson and Schooler model. The fourth experiment applied and evaluated two versions of the Anderson and Schooler model to the PubMed access data: one version that used only frequency, and a second that used both frequency and recency. In all of the experiments, a document access was defined as an abstract view or full-text download, and no distinction was made unless explicitly noted.

**Document accesses**

For this analysis only, we distinguished informational from navigational queries. Informational queries are queries where the underlying information need is to gain information about a topic—for example, ‘link between fish oil and blood pressure.’ Navigational queries are queries where the information need is for a specific item, such as a specific paper or papers published by a particular author.

The queries in the HAM-TMC data set were tagged by PubMed’s automatic term mapping where titles are denoted as (title), PubMed IDs are denoted as (pmid), journal names are denoted as (journal name), and volume is denoted as (vol). Queries that returned only one document and contained only (title) tags were considered navigational queries for specific titles. Queries containing only the (pmid) tag were considered navigational queries for documents using the PMID. Queries containing both the (vol) and (journal) tag were considered navigational queries for a specific volume of a journal. Finally, queries containing none of the aforementioned tags were considered informational queries.

**Desirability as a function of FOA**

We investigated whether document desirability as a function of FOA follows a power-law distribution using the methods described in two previous studies. In this experiment, we used 364 days of data to predict the next 30 days. We grouped documents in the 364-day period based on FOA and calculated the desirability based on the number of documents in a given group that appeared in the 30-day period. In the example in table 1, two documents have an FOA of 3, and documents with an FOA of 3 were present only once in the 30-day period yielding a probability of 50%. Similarly, documents with an FOA of 4 were present three times in the 364-day period, and all were present in the 30-day period, which yields a probability of 100%. We use the formula to convert probabilities into odds.

**Desirability as a function of ROA**

We used previously described methods to investigate whether document desirability as a function of ROA follows a power-law distribution.

<table>
<thead>
<tr>
<th>Document ID</th>
<th>Frequency of Access in 364-day period</th>
<th>Present in 30-day period</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>102</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>103</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>104</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>105</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>
distribution. In this experiment, we used a 7-day period to predict the eighth day. We grouped the documents based on the last day of access (ROA) in the 7-day period. We chose a small period to minimize the effect of PubMed’s reverse chronological ranking, which can bias results toward recently published articles. We repeated the process over the 394-day period by sliding the 7-day testing period. For example, desirability was calculated based on documents accessed on days 1–7 and their access on day 8. Desirability was computed in the next iteration based on documents accessed on days 2–8 and the access of those documents on day 9. We obtained the aggregate desirability by computing the mean of the desirability value for each 7-day period. Table 2 contains example data. Documents 101, 102, and 103 had an ROA of 7 (i.e., it was accessed 7 days ago). Of the three documents that had an ROA of 7 in the 7-day period, one was accessed on the eighth day yielding a probability of 33.3%. Of the two documents with an ROA of 6, one was accessed on the eighth day yielding a probability of 50%.

Evaluation of Anderson and Schooler model

We evaluated two versions of the Anderson and Schooler model. The first version requires storing a time stamp for each access, and the second requires only the aggregate number of accesses and the creation date or the start of the observation period. For example, suppose that two documents, document A and document B, were accessed 120 times over a 300-day period. Document B was accessed 120 times from days 240 to 300, and document A’s accesses are evenly distributed throughout the 300-day period. The calculation based on both ROA and FOA would account for the fact that the accesses for document B occur during a shorter and more recent period compared to document A and assign a higher desirability for document B. In contrast, the method using only FOA would assign an equal desirability to both documents.

Equation 1 presents the model based on ROA and FOA. The parameter d is the decay, which controls how quickly the desirability of a document decreases between accesses. The parameter n is the total number of times that a given document has been accessed. The parameter tk is the time since the last access. Equation 2 presents the model based on FOA, which assumes that the n accesses are evenly spaced over a given period of time T. The parameter d is the decay and is the same as in equation 1.

Desirability equation based on ROA and FOA:

\[
\ln \left( \frac{n}{\sum_{k=1}^{8} t_k^{-d}} \right)
\]  

(1)

Desirability equation based only on FOA:

\[
\ln \left( \frac{n}{1-d} \right) - d \ln(T)
\]  

(2)

We evaluated equation 1 by using 364 days of accesses as input and compared the accuracy of the model’s desirability prediction for the 30-day period with the actual desirability. We grouped the documents according to the number of accesses and applied equation 1 to the access patterns for each document within a given group. For this experiment, the decay parameter d has an empirically derived value of 0.1.

The desirability function in equation 2 assumes that the number of accesses is evenly distributed over the 364-day observation period. In this experiment, we used the number of accesses as input to the desirability function defined in equation 2 and compared the model’s desirability predictions with the actual desirability. For this experiment, the decay parameter d has an empirically derived value of 0.05.

RESULTS

Document accesses follow a power-law distribution

Figures 1, 2 show log–log plots of histograms for abstract views and full-text downloads resulting from both navigational and information queries. The linear regression performed on the data in figure 1 has an R² fit of 0.957. The R² value measures the variation of the data from the fitted line. The data in figure 2 are

<table>
<thead>
<tr>
<th>Document ID</th>
<th>Recency of Access in 7-day period</th>
<th>Present on eighth day</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>Day 7</td>
<td>0</td>
</tr>
<tr>
<td>102</td>
<td>Day 7</td>
<td>0</td>
</tr>
<tr>
<td>103</td>
<td>Day 7</td>
<td>1</td>
</tr>
<tr>
<td>104</td>
<td>Day 6</td>
<td>1</td>
</tr>
<tr>
<td>105</td>
<td>Day 6</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 1 Log–log abstract views versus number of documents (all views).

Figure 2 Log–log number of document downloads versus number of documents.
fitted with a straight line with an $R^2$ value of 0.976. Based on the $R^2$ value, we conclude that the distributions of abstract views and full-text downloads from navigational and informational queries both follow a power-law distribution.

The motivation behind the next analysis was that the previous results could be explained by PubMed’s reverse chronological ranking. PubMed, by default, ranks documents in reverse chronological order, and most users look only at the first page of results. Therefore, the observed power-law distribution could be an artifact of the PubMed ranking where a majority of the recently published articles receive many accesses (because they appear on the first page of results) and older articles receive fewer accesses (because they appear on later pages). To minimize the influence of the PubMed ranking, we looked at a variety of different navigational queries. As shown in table 3, the navigational queries have, on average, smaller result sizes compared to the informational queries, and thus accesses should be less affected by PubMed ranking. If the distribution of accesses for navigational queries followed a power-law distribution, this would support our hypothesis that the distribution of all document accesses follows a power-law distribution independent of the reverse chronological ranking.

Documents accessed via navigational queries in table 3 had an average $R^2$ value of 0.940; providing strong evidence of a power-law distribution. Information queries were defined as documents accessed from queries that do not contain titles, journal, PMID, volume, and dates. The results show that the distribution of accesses from informational queries had an average $R^2$ value of 0.949. Based on these results, we concluded that the distribution of document accesses followed a power-law distribution.

<table>
<thead>
<tr>
<th>Query type</th>
<th>Sample size</th>
<th>Mean of no of documents returned</th>
<th>SD of no of documents returned</th>
<th>Range of no of documents returned</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract views resulting from informational queries</td>
<td>1,324,910</td>
<td>4520.69</td>
<td>95470.62</td>
<td>1–18,859,689</td>
<td>0.936</td>
</tr>
<tr>
<td>Full-text downloads resulting from informational queries</td>
<td>165,469</td>
<td>2658.17</td>
<td>44,837.65</td>
<td>1–6,629,057</td>
<td>0.962</td>
</tr>
<tr>
<td>Abstract views resulting from PubMed ID queries</td>
<td>3991</td>
<td>1.34</td>
<td>15.09</td>
<td>1–953</td>
<td>0.926</td>
</tr>
<tr>
<td>Full-text downloads resulting from PubMed ID queries</td>
<td>7235</td>
<td>1.03</td>
<td>0.43</td>
<td>1–19</td>
<td>0.931</td>
</tr>
<tr>
<td>Abstract views resulting from title searches</td>
<td>17,403</td>
<td>10.03</td>
<td>0.0</td>
<td>1–1</td>
<td>0.969</td>
</tr>
<tr>
<td>Full-text downloads resulting from title searches</td>
<td>16,979</td>
<td>1</td>
<td>0.0</td>
<td>1–1</td>
<td>0.934</td>
</tr>
<tr>
<td>Abstract views resulting from journal and volume search</td>
<td>1,233</td>
<td>3.29</td>
<td>21.61</td>
<td>1–326</td>
<td>0.928</td>
</tr>
<tr>
<td>Full-text downloads resulting from journal and volume search</td>
<td>537</td>
<td>8.41</td>
<td>54.6</td>
<td>1–796</td>
<td>0.954</td>
</tr>
</tbody>
</table>

Desirability as a function of FOA follows a power-law distribution

Figure 3 presents a log–log plot of the desirability based on frequency of past access for articles that were accessed fewer than 100 times (in keeping with). The data are fitted by a straight line ($R^2=0.895$). These results provide strong evidence that document desirability as a function of FOA follows a power-law distribution.

Desirability as a function of ROA follows a power-law distribution

Figure 4 is a log–log plot of the result of the analysis. The linear regression of the data in figure 4 shows that a straight line can fit the data ($R^2=0.938$) and, thus, that the desirability of a document based on ROA follows a power-law distribution.

Evaluating the Anderson and Schooler model

Figure 5 shows the actual desirability of needing a document versus the predicted based on FOA and ROA. The correlation between the predicted desirability and the actual desirability is 0.668. Figure 5 presents the average desirability for the documents within each group.

Figure 6 presents the result of the model using the desirability equation shown in equation 2 (FOA only). The correlation between the predicted desirability and the actual desirability is 0.932.

**DISCUSSION**

We found that the Anderson and Schooler model predicts biomedical document accesses. Specifically, we showed that document accesses followed a power-law distribution,
desirability as a function of FOA followed a power-law distribution, and desirability as a function of ROA and FOA followed a power-law distribution. Finally, we evaluated two versions of the Anderson and Schoeler model for predicting document access. The first model calculated desirability based on ROA and FOA, and attained a 0.668 correlation with the test data. The second model calculated desirability based only on FOA and had a correlation of 0.952 with the test data.

This study is the first to evaluate the Anderson and Schoeler model for predicting biomedical document access, the first to show that the model is generalizable for bibliographic databases, and the first to investigate modeling desirability in the biomedical domain. The most similar study is Recker and Pittkow,23 which validated that the assumptions of the Anderson and Schoeler model were valid for WWW retrieval. The results of Recker and Pittkow23 and the results presented here are mutually reinforcing and show that it is possible to model the desirability of documents in a variety of domains.

A weakness of our study is that the results could be specific to HAM-TMC users. Approximately one-third of PubMed users are from the general public.25 In contrast, the general public for the HAM-TMC comprises approximately 20% of the transactions (HAM-TMC library, personal communication). In addition, the logs captured sessions of authenticated users, and some may use PubMed without logging into the HAM-TMC library. Further, the fact that a document was accessed does not mean that the document met the user’s information need or was necessarily useful to the user.

An interesting finding is that the FOA desirability model outperformed the ROA and FOA model. We hypothesize that poor performance was caused by very few documents in some groups, particularly in the highly accessed groups. Since the number of accesses follows a power-law distribution, there are many documents with few accesses and a small number of documents with many accesses. For example, there were 448,061 documents that were accessed once, but only six documents with 100 accesses.

To test this hypothesis, we created document groups where each group was ensured at least 100 documents. With 100 documents in each group, the FOA and ROA desirability model had a correlation of 0.973, and the FOA desirability model had a correlation of 0.984. Since the FOA and ROA performance improved, this indicates that the low performance was caused by the small number of documents that were accessed many times. The FOA desirability model is insensitive to small sample size since the same desirability is assigned to any document with a given FOA. In contrast, the desirability calculation based on FOA and ROA is sensitive to the small sample size given that desirability is calculated based on individual access patterns and then averaged for the group.

We found that desirability can be modeled, but our results do not explain why documents are desirable. There are many possible correlates including publication type (review article, meta analysis, etc), publication date, journal impact factor (JIF), citation count, and authors that were not investigated in this study. Further, the actual document content (ie, full text) may contain useful information not reflected in meta-data or information about the article such as frequency of past access. One possible effect of the observed power-law distribution could be the correlation between document accesses and citation counts given that citation counts have been shown to follow a power-law distribution.26 In addition, previous studies have shown that full-text downloads are correlated with citation count10 27 and can be used as proxies effectively overcoming some of the weaknesses of citation data such as citation lag.10 27

In future work we will investigate the type of information need best satisfied by desirability given that different metrics are optimal for different information needs. We hypothesize that desirability, given the known correlation between citation counts and document accesses10 27 and the utility of using citation data to find important documents,28 can be used to find important documents while avoiding some of the aforementioned weaknesses of using citation data. Finally, replicating this study using several months of PubMed query logs from the United States of America National Library of Medicine would eliminate bias created by constraining the analysis to a research medical center.

CONCLUSION

We have shown that: (1) document accesses follow a power-law distribution; (2) desirability of documents as a function of FOA follows a power-law distribution; and (5) desirability of documents as a function of ROA follows a power-law distribution. Based on these findings, we tested a desirability model based on both FOA and ROA, and a model based only on FOA. We found that the desirability model based on FOA outperformed the model based on ROA and FOA. We conclude that document desirability for a large corpus such as MEDLINE can be efficiently and accurately predicted based on past use. The desirability metric provides a means to improve document ranking.

Figure 5 Predicted access using Frequency of Access and Recency of Access.

Figure 6 Predicted access using Frequency of Access.
for IR systems and a basis for advancing Bayesian IR models by supplying a theoretically motivated prior probability estimate.

Acknowledgments We also thank the HAM-TMC library, especially C Young, for their support of this research.

Funding This work was supported in part by a training fellowship from the AHRQ and NLM Training Programs of the WM Keck Center for Interdisciplinary Bioscience Training of the Gulf Coast Consortia (AHRQ Grant T32HS017586; NLM Grant 5T15LM007093), NCRR Grant 3ULRR024148, NCRR Grant 1RC1RR028254, NSF IIS-0964613, HRSA Grant D1BRH6261, and the Brown Foundation.

Competition interests None.

Ethics approval This study was approved by The Committee for the Protection of Human Subjects at the University of Texas Health Science Center at Houston under study number HSC-SBNI-11-0071.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES


Downloaded from https://academic.oup.com/jamia/article-abstract/19/3/473/717917 by guest on 26 December 2018