Developing an Assessment Index for Collection-User Suitability: Application of Information Entropy in Library Science

(Le Yang, Fuyi Wei *, Enci Chen)

Abstract

What is lacking from the literature and practice of library print collection assessment is a mathematical function that can integrate variables and quantify the overall measurement of the collection as a final score for the library science practitioners and administrators. The study uses the information entropy to explain a mathematical procedure on how to define variables from the regular circulation statistics. Based on the assessment concepts of circulation statistics and user behavior, the study explores to develop mathematical functions and proposes a Collection-User Suitability index to quantify the suitability measurement of the overall collection and the user intention. The study then considers the extreme conditions to validate the functions and verifies the application of the function via empirical and simulated data. Finally, the authors developed a computing tool and uploaded it to GitHub for free access based on the study's results. It is hoped that library science practitioners and administrators can utilize the mathematical function.

Keywords: Collection Assessment; User Behavior; Usage Evaluation; Information Entropy; Library Science; Information Science

1. Introduction

Resource acquisition is one of the academic library's roles that have remained the most valued for over a decade, according to a faculty survey (Long & Shonfeld, 2014). If collection-building is a role valued the most by faculty, libraries should assess how well they fulfill that role (Johnson, 2016). Henry et al. (2008) hold a similar point of view, claiming that if libraries do not critically analyze collections, then the purpose of the library's existence can be in question. Indeed, collection evaluation and assessment have always been an essential part of libraries, many scholars and librarians consider it a core mission of the library to ensure the collection meets the needs of users and to demonstrate the value to the institutions (Borin & Yi, 2011; Duncan & O'Gara, 2015). Collection assessment is also necessary for better management of resources and provides the library administration with documented evidence on the effective stewardship of the library (Henry et al., 2008).

In the recent decade, the library's traditional role as a repository for print materials has been replaced by a learning center that facilitates communication and interaction (Kaplan, Steinberg, & Doucette, 2006). Library collection management is also transitioning from a print model to a hybrid model that includes electronic and digital resources. According to Tabacaru and Pickett (2013), librarians must invest in building and maintaining hybrid collections, although libraries have embraced the digital environment. Cook and Maciel (2010) revealed that library patrons who indicated a preference for online reading vs. print copies are split equally, with fifty percent of survey respondents for each. The changing landscape of library collections makes it more important for librarians to assess and understand the library print collection's value in fulfilling

the needs of library patrons. However, traditional evaluation methodology may not address the emerging needs of collection assessment (Borin & Yi, 2008).

Establishing a formal procedure to ensure collection evaluation in libraries becomes critically crucial as resource cost has been inflating without comparable increasing funds in collection development (Lantzy, Matlin, & Opdahl, 2020). Wilde and Level (2011) reported that only 11% of survey respondents had developed a formal process for collection evaluation. An ARL spec kit also reported that 58% of ARL member institutions had developed a formal, entirely, and partially process of collection evaluation in the libraries (Harker & Klein, 2016). ACRL Research Planning & Review Committee (2016) called for the need to establish formal review protocols of collections rather than ad-hoc project-based models. Lantzy, Matlin, and Opdahl (2020) summarized that this dramatic increase in the development of formalized collection assessment reflects how libraries respond to the pressures of demonstrating more value in collections with less funding.

Reviewing the current studies, what is lacking from the literature is a mathematical function or model that can integrate all different types of collection and circulation statistics so that a final and single score can quantify how well the library collection is matching the user needs. The study explains a mathematical procedure of how to define and quantify variables for collection and circulation statistics and how to develop, test, and optimize mathematical functions. The study firstly introduces probability for individual items that have historical circulation statistics. The study further argues that linear probability cannot ideally nor objectively describe the distribution of circulation data when dealing with a random variable. The linear probability also cannot reflect the suitability of circulated items and potential user needs. On the other hand, during the research procedure, the authors identify the need to find a similar normal distribution that can synthetically characterize its average variability (Shannon, 1948) and measure absolute randomness (Matricciani, 1994).

Hence, the study proposes to use Shannon's mathematical function of information entropy to quantify each circulated item’s uncertainty of being successfully borrowed again in the future first. The study then advances to calculate the average entropy value of aggregated individual entropy values for the assessment index of the collection-user suitability, which is defined to evaluate library collection’s suitability for its patron. The study cites the developed mathematical analysis to support the equation construction and tests five extreme conditions to validate the mathematical functions. At the final step, both empirical and simulated data were used to examine and further verify the practicability of the functions. Besides, the authors developed an application and uploaded it to GitHub for free access.

2. Related Works

2.1 Library Collection Assessment

According to Borin and Yi (2008), the literature in the library collection assessment and evaluation shows that methodologies are generally grouped into two categories, the traditional category, which is usually criteria-based, and the new category, usage-based. A similar classification is also claimed by Johnson, Hille, and Reed (2005), that library assessment topologies can be categorized into collection-based and user-based. Borin and Yi (2008) insisted that the traditional methods, including indicators such as collection size, acquisition rates, annual

expenditures, circulation data, in-house usage, and more, cannot be solely applied to the current environment for collection evaluation. When the paradigm-shifting of collection development occurs, so does the focus of library collection evaluations and the relevant indicators. Although there is no single correct way to evaluate collections, indicators primarily used or weighted nowadays include general capacity, usage, users, subject-specific standards, scholarly publishing, and environmental factors (Borin & Yi, 2008; Dinkins, 2003; Murphy, 2013).

Library circulation data has been used as a collection evaluation indicator (Day & Revill, 1995). Dee et al. (1998) analyzed collection usage statistics and concluded that these studies are helpful in evaluating collection performance relative to library use. Gaining a comprehensive overview of the current state of a library's collections requires accurate and accessible data. Therefore, a collection assessment project usually starts with recent inventory statistics, ensuring the assessors have an accurate image of their current holdings (Intner, 2003; Johnson, 2016). Wilde and Level (2011) encouraged libraries to maintain a pool of collections statistics that can always be accessible to library staff. Moreover, analyzing and interpreting the data would become the next task for assessors in this process (Duncan & O'Gara, 2015).

Practical collection assessment provides quantitative and qualitative data for evaluating a library's holdings (Henry et al., 2008). It is essential to use a measurement that is not solely based on volumes added or the number of titles (Kyrillidou, 2006). Thus, there is an increasing focus on collection evaluation from users' perspectives, which helps determine how well the library collection meets the needs of its patrons (Agee, 2005). Kelly (2014) carried out an assessment program on their library collection, involving multiple methods of measurement and including inputs from all stakeholders. Borin and Yi (2011) proposed that capacity and usage are the two good indicators of assessing the general collection and applied the indicators by looking at system and circulation statistics, given definite evidence of the overall usage of the collection. However, overall usage and circulation statistics usually provide only a longitudinal comparison, instead of demonstrating accurate utility or suitability of the said collection and the patrons.

The realization of lack of suitability measurement signifies the importance of including liaison librarians in the process of collection assessment, who can provide insightful interpretations on both collect data and usage statistics (Gregory, 2011; Johnson, 2016; Luther & Guimaraes, 2016). Duncan and O'Gara (2015) developed such a collection assessment model to evaluate subject collections at their libraries; while this model included several collection-related projects, the focus is an assessment rubric that is used to evaluate disciplinary collections, combining student enrollment data, survey results, and collection statistics. Another approach that has been used to assess the relevance of a collection is citation analysis (Broadus, 1997); though it addresses the need of suitability measurement on a collection to some extent, it does not provide a thorough and comprehensive solution to it. Finally, with the analytic and visualization tools being rapidly developed, scholars started to use bibliometric techniques on collection assessment, specifically on electronic subscriptions (Ho, 2018).

Library performance measurement is another aspect that libraries use to identify budgets and resources' management objectiveness and implementation. According to Prathap and Mittal (2010), the review of the literature showed that relevant studies had attracted the attention of library and information scholars, who have started to utilize the approach of performance measurement for quality improvement and organizational effectiveness (Day & Revill, 1995; Henderson, 2000; Lu, 2005). Based on the idea of h-index and the p-index developed by Liu and Rousseau (2009), Prathap and Mittal (2010) demonstrated two separate quantitative and

qualitative indices on a coordinate grid that aims to reflect an evaluation of their library circulation data directly. Unfortunately, no further quantitative studies or mathematics-oriented models are found in library collection assessment or performance measurement studies.

Reviewing the literature, library practitioners are aware of the difficulty in building and evaluating library collection, as it involves a few factors that are hard to express or quantify. Historical circulation statistics can only explicitly outline the usage data without further analysis, and survey results are challenging to apply to any extensive collection, while collection profiling is subject to selection preference and personalized interpretation. The study hopes to explore the possibility of quantifying both individual and collective suitability of library collection and the patron group it serves by investigating the applications of information entropy in related literature.

2.2 Application of Information Entropy

The influence and application of the entropy concept have been found in many disciplines. For example, a recent bibliometric study has presented a significant impact of Shannon's information entropy in STEM fields, whereas the disciplines of Social Sciences are the least influential categories (Basurto-Flores et al., 2018). In bibliometrics and information science, Shannon entropy has been used in quantitative analysis of subject fields and journal publications. An early study firstly adopted the concept and used Shannon entropy to measure the obsolescence of the literature by studying IEEE journals (Matricciani, 1994) and concluded that 40-50 years is roughly the age for a reference to be historical and fundamental.

In 2002, Leydesdorff used probabilistic entropy as indicators to study structural changes in journal clusters in subject fields and suggested that although general patterns of change cannot be indicated, probabilistic entropy can often indicate structural erasure caused by the process of codification. In studying interdisciplinarity vector-based indicators, Leydesdorff and Rafols (2011) explored using entropy to measure unevenness of citations and suggested that Shannon entropy, though affected by the size of the population, qualifies for an indicator of interdisciplinarity. Similarly, Silva et al. (2013) used the entropy equation to measure the interdisciplinarity index of subject categories for citation analysis and found that impact factors of journals positively correlate with high entropies related to the wideness of readership of journals.

Rare literature on entropy is found in the field of library and information science in recent years. However, a recent study of books' depth and breadth entropy method was found to be used by Zhang and Zhou (2020) in the research methodology steps to estimate the regional distribution of books for further data selection in the study. Thus, the application of information entropy in library collection assessment and suitability evaluation for user intention will fill the gap in the profession and contribute to library and information science literature.

3. Research Questions

The relevant studies show that circulation statistics such as times and length of checked-out items are merely collected, aggregated, and presented in library collection assessment. What is lacking from the literature about library collection assessment or suitability evaluation of collection and users is a mathematical function that can provide a direct, accurate, and numerical reflection on

either a disciplinary collection or the general collection. How to make the best use of the circulation statistics in a mathematical model and generate a numerical value from the modeling to evaluate each circulated item, a disciplinary collection, or the suitability for the patrons remain unresolved.

Before developing a mathematical function, measurable variables need to be identified during the circulation cycle. As usual, quantitative values in granular levels of a whole population firstly need to be clearly defined and calculated before a function can be developed. In the study, the times and length of each circulated physical item within a set period, e.g., a year, can be tallied. Without considering the item's intrinsic value, the times and length of each checked-out physical item can be considered as "random" variables that can be calculated for the probability of this item being successfully circulated within the set period in the future. The probability here is another presented form of the usage statistics, converted from the historical circulation data. The probability itself can certainly be used to evaluate the circulated books based on the frequency of the physical items being needed and circulated. For example, the book circulated more often and checked-out longer is more probably than the one circulated less and shorter.

However, the probability here in the study is not as simple as it indicates. The authors argue that the larger value of probability does not necessarily represent a better evaluation. For example, if the circulated book is circulated so frequently or so long that other patrons cannot successfully borrow the item within a set period, the probability here will be minimal. In this case, a small probability, instead, means many users are eager for the item, which should also represent the higher suitability of the collection and the user. That is to say, the linear probability cannot ideally nor objectively represent what collection-user suitability entails. For instance, when $p = 0.99$, it indicates that the patron can 99% successfully borrow the item and it indicates a positive suitability of the item and the patron; however, when $p = 0.01$, it indicates the patron will probably not borrow the item because it is not available in the collection, whereas it still indicates a positive suitability for the item.

Therefore, the study proposes to use Shannon's mathematical function of information entropy to synthetically characterize the average variability (Shannon, 1948) and measure absolute randomness (Matricciani, 1994). When considering a discrete random variable, as Matricciani (1994) added his explanation to the entropy, that the fundamental concept of the information entropy is "entirely based on the probabilities of the symbols, not on the meaning they carry" (p.131). Moreover, because the discrete entropy is an absolute measure of randomness, it is instrumental in information application relative to the more common discrete random variables (Matricciani, 1994).

In this case, the probability of each item can be used to calculate "entropy," a fundamental quantity first introduced by Shannon (1948) that quantifies the uncertainty of the information outcome. In the mathematical model of entropy, both sides of the distribution can measure absolute certainty, while the middle of the distribution represents absolute uncertainty. That is to say, the probability will be converted to the entropy so that in the middle when the probability is 0.5, which means it is the most uncertain if the physical item is being needed and borrowed. Further, with a set of probabilities of occurrence, following Shannon's entropy algorithm, measuring the total entropy (uncertainty of overall outcome) can be calculated, which can be used to assess and measure if the overall collection meets the need of user intention. Therefore, the study hereby defines the total entropy as the assessment index for the collection-user suitability.

In the procedure for calculating the entropy, the study aims to answer two research questions step by step:

> *RQ1: How to measure the probability of each print item being successfully borrowed within a set period?*

The probability defined in RQ1 is meant for the entropy calculation. Thus, historical probabilities should not be used for assessing what has happened but for estimating the numerical value of entropy.

> *RQ2: How to utilize Shannon's entropy function to measure the collection-user suitability with the calculated probabilities within a set period?*

In Shannon's theorem (p.633-634), the accumulated value of probabilities plays a central role in information theory as information, choice, and uncertainty measures. That is, a set of possible events whose probabilities of occurrence once set, a measure of "how much choice is involved in the selection" and of "how uncertain we are of the outcome" can be found. Therefore, RQ2 aims to turn the entropy value into a measurement of collection-user suitability.

4. Construction of Functions

In order to develop a function to measure the probability of each item being circulated, the study breaks down the inferring procedures and bases on a precondition that individual items being studied have circulation statistics (being circulated at least once). In general, any physical item being circulated needs to fulfill two prerequisites: the item being needed and available in the library. The study takes Book A and Book B as examples, assuming Book A was circulated twice within a year, one for ten days and the other for 20 days, and Book B was circulated four times with ten days for each time. The circulation statistics about Book A and Book B are summarized in Table 1.

Table 1: Circulation Statistics about Book A & Book B

Book	Times of Circulation	Length of Circulation	Total Days of Being Checked-out
A	2	10 days & 20 days	30
B	4	10 days each time	40

With these provided numbers, two questions can be directly answered: 1) which book has a higher chance to be available in the library (the answer is Book A), and 2) which book is more frequently needed by patrons (the answer is Book B). However, a question such as "which book has a higher probability of being successfully borrowed" cannot be quickly or directly answered by looking at the statistics provided. For example, is it true that Book A has a higher probability of being borrowed in the future because its availability in the library is more extended than Book B? Is it true that Book B has a higher probability of being borrowed because it is more frequently needed than Book A? Questions like these cannot be easily answered without considering all variables.

4.1 Preliminary Construction of the Probability Function

In order to answer this question, one needs a function that includes the currently associated statistics for calculation. The study sets the variables as follow:

P_x = the probability of a book being available in the library

P_y = the probability of a book being needed by patrons

P = probability of a book being successfully borrowed (needed and available); $P = P_x \times P_y$

c = times of the book being circulated within the set period T_0; c should be an integer; $c > 0$

T_0 = the set period for the calculation (e.g. one year = 365 days);

T_j = the days of the book being checked out from the library; $T_j \in (0, T_0)$

T_i = the days of the book being available in the library; $T_i \in (0, T_0)$; $T_i = T_0 - T_j$

From these variables, the study proposes the function of P_x as follow:

$$P_x = \frac{T_i}{T_0} \quad (1)$$

As for P_y, how and why an un-circulated book is needed depends on many unpredictive reasons. However, if a book has been circulated a certain number of times for a duration of period, it is feasible to develop a function by using these two variables c and T_i. The more a book being circulated indicates that the book is more likely borrowed again in the future. Therefore, variables c and P_y should be positively correlated. Conversely, the longer a book being available in the library indicates that the book is less likely borrowed again in the future. That is, variables T_i and P_y should be negatively correlated. In the meanwhile, the probability P_y should be ≤ 1 and infinitely close to 1. Based on this reasoning procedure, the study preliminary proposes the function of P_y as follow:

$$P_y = 1 - \frac{T_i}{T_0 \times c} = 1 - \frac{P_x}{c} (c > 0) \quad (2)$$

Based on *Function (1)* and *(2)*, P can be transformed into *Function (3)*:

$$P = P_x \times P_y = P_x - \frac{{P_x}^2}{c} \ (c > 0) \quad (3)$$

Putting the numbers (Table 1) from the example of Book A and Book B into *Function (3)*, the results are presented in Table 2, which is meant to answer the question "which book has a higher probability of being successfully borrowed." Concluding from Table 2, although Book A is longer available in the library than Book B, it is less likely to be successfully borrowed.

Table 2: Probabilities of Book A & Book B

Book	Probability of Being Available (P_x)	Probability of Being Needed (P_y)	Probability of Being Borrowed (P)
A	0.92	0.54	0.50
B	0.89	0.78	0.69

To test *Function (3)*, the study needs to test the extreme conditions in the function, which can be transformed into *Function (4)*:

$$P = \frac{c^2}{4} - \frac{1}{c}\left(P_x - \frac{c}{2}\right)^2 \ (c > 0) \ (4)$$

Variable *c* represents the number of times a book being circulated. According to *Function (4)*, when $c = 1$, P will have the maximum value 0.25 when $P_x = 0.5$ ($= T_i / T_0$). That is, when an item had been checked out half of the set period, e.g., half a year within one year, this item has the highest probability (0.25) of being borrowed again. When $c \geq 2$ and $P_x \in (0, 1)$ get closer to 1, P will also become larger and larger and being closer to the value of $(1 - 1/c)$.

Because P_x is determined by variables T_i and T_0, if the set period of T_0 is one year, then P_x will have 364 different probabilities as T_i can range from 1 to 365. Thus, when $c = 1$ and $T_i = 364$, P_x will reach its maximum value $Max\ (P_x) = 364/365 = 0.997$; P_y will reach its minimum value $Min\ (P_y) = 1 - [\ Max\ (P_x) / 1] = 0.003$; $P = Max\ (P_x) \times Min\ (P_y) = 0.0029$. This value (0.0029) indicates that the book is nearly impossible of being successfully borrowed. However, when c = 2 and $T_i = 363$, $P_y = 0.503$ and $P = 0.5$. This value (0.5) indicates the most uncertainty for the book being borrowed or not.

4.2 Re-construction of the Probability Function

Theoretically, when $c = 1$ and $c = 2$ (meaning the item is circulated once and twice, one day for each time, within a set period), the difference of the final value of P should not be as this large (0.0029 vs. 0.5). The inference from the calculation of extreme conditions necessitates the adjustment of *Function (3)* by adding fixed variable and exponential. It is noteworthy that, even adding the fixed variable in the denominator, c still represents the circulation time, which means the transformed equation still holds the premise that c>0. The study proposes to adjust the function of P_y to be as follow, which is meant to replace *Function (2)*:

$$P_y = 1 - \frac{T_i}{T_0 \times (c+1)^z} = 1 - \frac{P_x}{(c+1)^z} \ (c > 0) \ (5)$$

Based on *Function (1)* and *(5)*, P can be formulated as *Function (6)*, which is meant to replace *Function (3)*:

$$P = P_x \times P_y = P_x - \frac{{P_x}^2}{(c+1)^z} \ (c > 0) \ (6)$$

In doing so, the value of logarithm Z needs to be calculated by reasonably considering extreme conditions to define the function. The reason why introducing a logarithmic measure is explained by Shannon (1948, p.379) that in the monotonic function, the logarithmic measure being more practically useful, intuitively proper, and mathematically suitable. Consequently, the logarithmic function of Z is as follow:

$$Z = log_{c+1}^{\left(\frac{{P_x}^2}{P_x - P}\right)} \quad (7)$$

To ease the uncertainty in the information and to calculate the probability, *Function (7)* is designed to quantify P_y by using c and T_i. In general, when the circulation gets more often (value of c gets larger), this item is more needed (value of P_y gets larger). When c and T_i have the minimum values $c = 1$ and $T_i = 1$ (checked out once for one day within a year), it is the most uncertain that this book would be borrowed again or not, which in this case $P = 0.5$. Putting these values into *Function (7)*, the logarithmic value of Z will be closest to 1 ($Z \approx 1$). Thus, regarding Z =1, the adjusted *Function (6)* can be confirmed as follow:

$$P = P_x \cdot P_y = P_x - \frac{{P_x}^2}{c+1} \quad (c > 0) \quad (8)$$

Testing the extreme conditions with $c = 1$ and $T_i = 364$; $c = 2$ and $T_i = 363$; $c = 1$ and $T_i = 1$; and $c = 364$ and $T_i = 1$; *Function (8)* produces reasonable probabilities that can be deemed as mathematically suitable and proper. Further validation will be provided in the following sections as well. Putting the numbers (Table 1) from the example of Book A and Book B into *Function (8)*, the adjusted results are presented in Table 3.

Table 3: Adjusted Probabilities of Book A & Book B

Book	Probability of Being Available (P_x)	Probability of Being Needed (P_y)	Probability of Being Borrowed (P)
A	0.92	0.69	0.63
B	0.89	0.82	0.73

4.3 Entropy & Final Assessment Function

Now that the function can estimate the probability of each item (at least circulated once) being borrowed, the entropy of each circulated item can be formulated to Function (9) following Shannon's entropy information theory (p.389):

$$H = -[PlogP + (1-P)log(1-P)] \quad (9)$$

The entropy of each item is converted from the probability of each item being successfully needed and borrowed. As indicated in the Research Questions section, the linear probability alone cannot be used to measure each's suitability for patrons because it fails to represent the situation when the physical item is being needed but unavailable. Instead, the individual entropy can be used to quantify the uncertainty of information outcome, which is the discrete randomness of the physical items being borrowed. The quantified uncertainty then reversely reflects the suitability of the physical item and the patrons' needs. Putting the application in Shannon's position, the measurement of uncertainty can be used to reduce the uncertainty of the bit value.

For those items that do not have circulation statistics (never circulated within a set period), because no previous statistics can be found for a valid mathematical function, the H value must be manually defined. When physical items do not have historical circulation statistics, they cannot be mathematically predicted if the uncirculated items are successfully borrowed in the future. In other words, it is in the most uncertain situation when c=0, where the Shannon's entropy value should be manually defined as H = 1. Consequently, all uncirculated items' entropy values shall be 1 when doing the calculation.

Now that the entropy of individual circulated items is calculated and defined in all situations, from the mathematical standpoint, the collection can also be estimated for user suitability by calculating the average entropy value. Similar to the measurement of the individual uncertainty, the average entropy is used to quantify the collective entropy that reflects how uncertain, or reversely how well, the collection is being used. When adding all entropy of each item together, the entropy of total circulated items can be formulated as follow:

$$H_{(n)} = -\sum_{i=1}^{n}[P_i log P_i \ + \ (1-P_i) log(1-P_i)] \quad (10)$$

Based on *Function (10)*, the average entropy value of the whole collection is $H_{(n)}/n$. From the practical experience, the circulated items usually form a small, if not tiny, fraction of the total collection in many libraries, so the average entropy value $H_{(n)}/n$ of the whole collection will be very close to 1. Considering that entropy is used to measure the uncertainty of information outcome, a larger value of the final score indicates more uncertainty of the occurrence (e.g., when c=0 then H=1). Therefore, following the common readability standard that in a grading scale the larger numerical value usually indicates more positivity to the readers, the study proposes to make a small adjustment on reversing the entropy value to finalize the assessment function as follow:

$$f(P) = (1 - \frac{H_{(n)}}{n}) = \{1 + \frac{\sum_{i=1}^{n}[P_i log P_i \ + \ (1-P_i) log(1-P_i)]}{n}\} \quad (11)$$

Where, *Equation (1)* and *(8)* applied: $P = P_x \times P_y = P_x - \frac{{P_x}^2}{c+1} = \frac{T_i}{T_0} \times [1 - \frac{T_i}{T_0 \times (c+1)}]$

Where,

c = times of the book being circulated within the set period T_0; c is an integer; $c > 0$

(when $c=0$, then $H=1$)

T_0 = the set period for the calculation (e.g. one year = 365 days);

T_i = the days of the book being available in the library; $T_i \in (0, T_0)$.

In *Function (11)*, a coefficient of 10, 100, 1000, or 10000 can be added to the function to increase the readability of the final value, because from the practical experience only a small portion of the overall collection is circulated, which will result in value (with four or five decimal places) being very small that is close to zero. Thus, adding coefficients can ensure the value of the assessment function between 0 and 10 for better comprehension of the final value. The final score, representing collection-user suitability, can be used to assess the suitability of how library collection matches users' intention within a set period, such as a year or a specified period after some specific collection promotion programs.

4.4 Validation of Functions

Firstly, the study examines how the change of checked-out days impact the functions when the circulation time c is unchanged. That is, during the setting period of 365 days (one year), when $c = 1$, $c \to 0$, and $c \to +\infty$, how the value of P_y, P, and $f(P)$ is distributed.

When c = 1, the possible checked-out days range from 1 to 364 days (when 365 days then P = 0). According to Figure 1, value distributions of three functions all meet the expectation within a mathematically proper and reasonable scale.

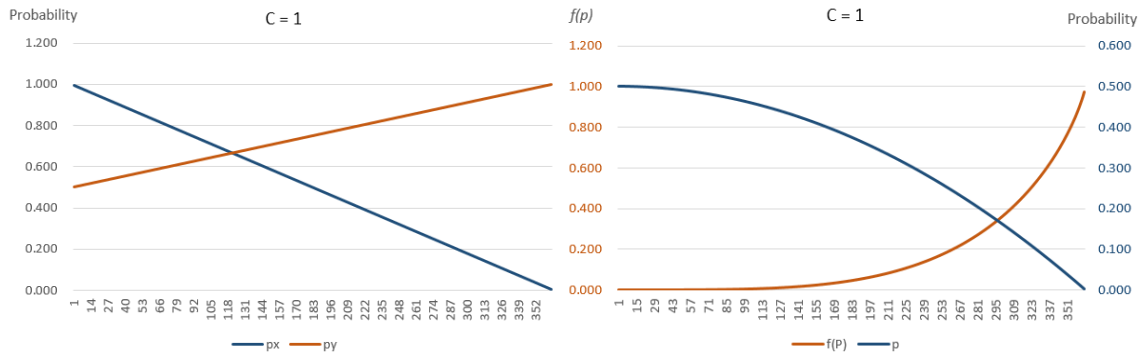

Figure 1: When c = 1, value distribution of *Py*, *P*, and *f(P)*

Figure 2 and Figure 3 further illustrate the distributions of extreme conditions when $c \to 0$ and $c \to +\infty$. Though in the practical situation, these two extreme conditions do not exist, the distributions demonstrate that all possibilities are within a mathematically reasonable scale.

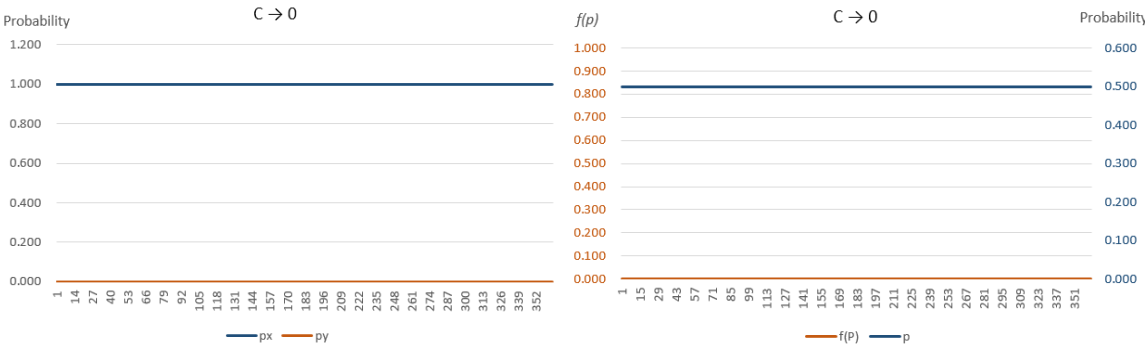

Figure 2: When $c \to 0$, value distribution of *Py*, *P*, and *f(P)*

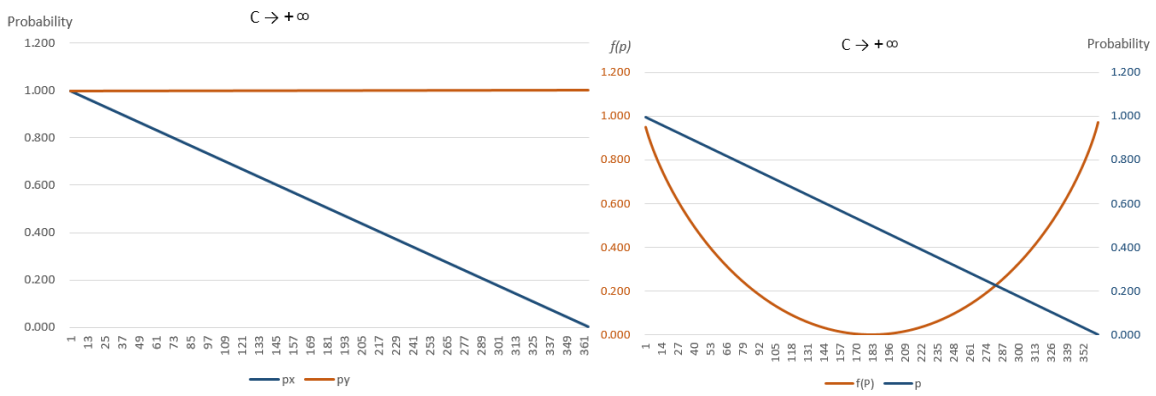

Figure 3: When $c \longrightarrow +\infty$, value distribution of *Py*, *P*, and *f(P)*

Secondly, the study continues to examine how circulation time *c* impacts the functions when the circulation day T_i is unchanged. That is, during the setting period of 365 days (one year), when $T_i \longrightarrow 0$ and $T_i = 364$, how the value of P_y, *P*, and *f(P)* is distributed. Figure 4 and Figure 5 illustrate the distribution in the two extreme conditions and demonstrate that the values are within mathematically reasonable scales.

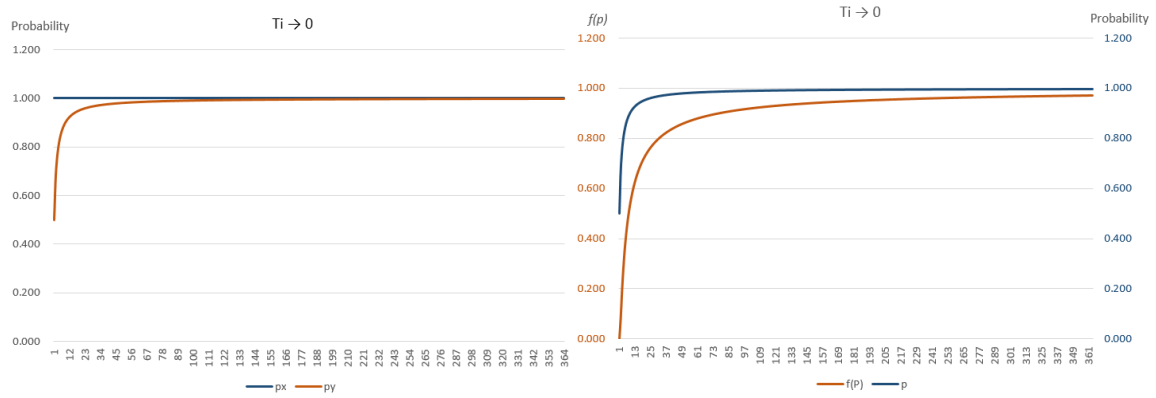

Figure 4: When $T_i \longrightarrow 0$, value of distribution of *Py*, *P*, and *f(P)*

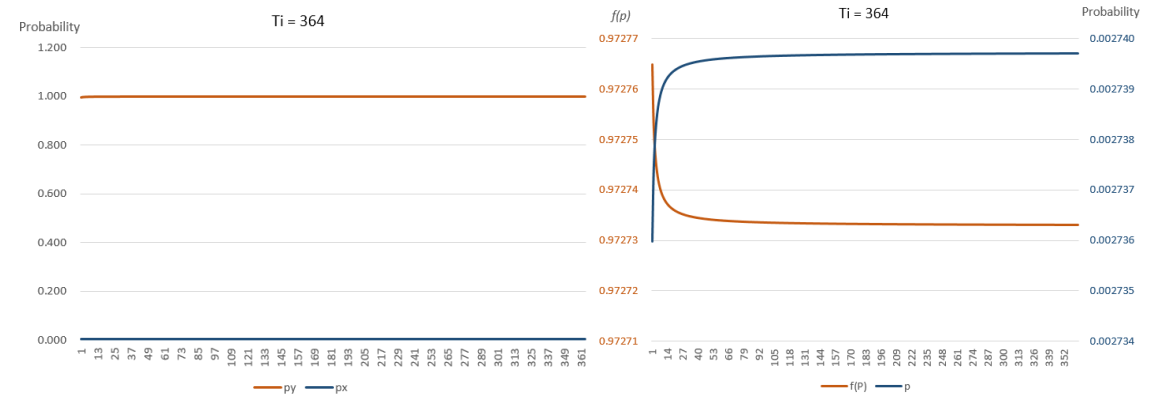

Figure 5: When $T_i = 364$, value of distribution of *Py*, *P*, and *f(P)*

5. Discussion

5.1 Practical Case Study

For the study, the authors collected detailed circulation statistics from a disciplinary collection that was developed for two years of Wenzhou-Kean University (WKU) Library. As shown in Table 4, the first three columns outline the aggregated statistics for the collection, showing that between 2018 and 2019, the collection was added another 3,005 items, 351 more items were circulated in 2019, and the total days of overall circulation are increased by 42,221. There are many reasons as to why the circulation of items may have increased between the two years. However, it is somewhat dispersed to comprehend these statistics because it is uncertain whether the increased number of circulated items and circulation days were resulting from the grown collection. It is challenging to interpret statistics that measure different things. It is also difficult to compare the two years' data by looking at all different statistics scattered in different categories. Therefore, putting all statistical data into the collection-user suitability equation to yield a single and final index score for an overarching evaluation will help understand the measure of collection quality and users' borrowing behaviors.

Table 4: Assessment of A Disciplinary Collection at WKU Library

Year	Total Collection	Circulated Items	Total Circulated Days	Total Checked-out No. of Times (c)	Score [$f(P)$]	Score [$f(P) \times 10000$]
2018	23814	771	38226	1286	0.000796	7.96
2019	26819 (+3005)	1122 (+351)	80447 (+42221)	2072 (+786)	0.002603	26.03 (+227.01%)

In order to apply the retrieved statistics to the final function, the authors process all numerical data for each circulated item and use *Function (8)* to calculate the probability of each circulated item within each year. The authors then applied all *P* values into *f(P) Function (11)* to get the final scores for 2018 and 2019, as presented in the last two columns of Table 4. The result shows that the collection-user suitability score in 2018 is 0.000796, and in 2019 is 0.002603, offering a direct quantitative reflection on the suitability comparison. The coefficient 10000 is applied to help increase comprehension of the score. The assessment shows that the collection-user suitability score is increased by 227% from 2018 to 2019. The authors completed the batch processing of data and calculations through MS Excel.

The collection-user suitability math model is derived from the entropy application. According to Frenken and Leydesdorff (2000), the expected information is contained in a message built by posteriori upon the historical time series. In other words, the entropy application tries to answer, "what does the past mean for the present." For example, in the study, the probability calculation of *Function (8)* is only meant for the entropy calculation that quantifies the uncertainty of existing information outcomes. Historical probabilities are not meant for assessing what has happened but for estimating the numerical value of entropy that can mean something for the present. Similarly, the probability of the un-circulated books in the historical time frame is zero, reflecting that these books were not circulated in the past and might not be needed at present. In this regard, these uncirculated items represent the most uncertain cases, in which it is impossible to calculate via

mathematical equation. Thus, the entropy values for the uncirculated items are defined as 1, indicating the most uncertain situation.

It shall be reiterated here that the fundamental concept of the information entropy is "entirely based on the probabilities of the symbols, not on the meaning they carry" (Matricciani, 1994, p.131). For instance, *Function (5)* aims to calculate the probability of P_y of an item being needed based on circulation and statistics. The function offers a mathematical method to develop the correlation between circulation statistics and availability, trying to quantify users' needs. However, the mathematical equation is not meant to quantify the books' intrinsic value that might or might not cause the need in the future, especially for those items never circulated.

Moreover, there is no global benchmark for the final score because libraries and collections have distinct situations, including the size of the collection, circulation policy, target readers, and more. Instead, libraries can establish their benchmark score for suitability comparison based on their historical data. Peer institutions can also establish a benchmark score or conduct comparative studies among libraries.

5.2 Practice of Simulated Data

Because there is no benchmark of the final score for comparison, the study simulated another three sets of data based on the 2018 data of WKU Library to examine how the change of each variable affects the final score and verify the practicability of the function. The simulated data and results are presented in Table 5, where, in 2018a, the circulated days (*Ti*) are doubled, in 2018b checked-out times (*c*) are doubled, and in 2018c, the number of circulated items is increased to 1286 (the maximum number of the circulated items should be 1286 because it cannot exceed the total checked-out times).

Table 5: Three Sets of Simulated Data based on 2018 Practical Data

Year	Total Collection	Circulated Items	Total Circulated Days (T_i)	Total Checked-out No. of Times (c)	Score [$f(P)$]	Score [$f(P) \times 10000$]
2018	23814	771	38226	1286	0.000796	7.96
2018a	23814	771	76452 (+100%)	1286	0.001300	13 (+63.32%)
2018b	23814	771	38226	2572 (+100%)	0.002897	28.97 (+263.94%)
2018c	23814	1286 (+67%)	38226	1286	0.003214	32.14 (+303.77%)

According to Table 5, the independent increase of each variable results in a positive effect of the final collection-user suitability score, meeting the study's expectation and verifying the practicability of the final function. Furthermore, the result indicates that the increase of circulated items in 2018c positively impacts the final score. It implies that diversifying items to satisfy patrons' different needs can be one of the proper means to maintain a healthy collection. Meanwhile, the increase of the checked-out number of times in 2018b suggests a stronger and more frequent needs by different patrons, resulting in a notable growth of the collection score.

When the circulated items and checked-out times do not change, the sole increase of circulated days in 2018a brings some positive impacts to the final score, with a 63.32% addition,

not as much as the other two sets 2018b and 2018c. The limited suitability score in 2018a denotes that the mere growth of circulated days of the same items cannot significantly influence the suitability measurement since the change does not signify a diversified collection or expanded needs.

5.3 Computing Tool

For application purposes and the practitioners' convenience, the authors developed a computing tool using Function (8) and Function (11) and uploaded the tool to GitHub for free access via (https://github.com/yanglegd/collectionentropy) with illustrated instructions. Once deployed on the local server, the application allows users to input variables and calculate the probability result of individual items. The application also allows users to load a pre-processed Excel sheet, select a coefficient, and calculate a final assessment score for the collection.

6. Conclusion

The study provides explanations of an inferring procedure of defining variables and constructing functions for collection usage statistics. The study then proposes a final mathematical function, integrating variables from regular circulation statistics, to yield a final score for collection-user suitability assessment and performance comparison. The final score provides a direct and quantitative reflection on how the library collection meets the needs of patrons so that assessment librarians and library administrators can use the final scores to obtain a scope picture by comparing different years' collection quality. It aids decision-making on budget allocation and performance measurement of the organization. In addition to the final function, the probability of item circulation can also be used to evaluate individual print items and be referenced by subject librarians and collection development librarians when making acquisition decisions.

The Data process and calculation of functions can all be performed in the MS Excel sheet. A global benchmark score can also be established for different types and sizes of libraries if more library science practitioners adopt the concept of a collection-user suitability assessment and share the assessment scores. The function can also be applied to in-library usage and eBook usage when the usage statistics are transformed and expressed in terms of statistical circulation format following uniform standards, or a similar function can be developed to reflect the measurement and assessment of electronic resources. It is also hoped that the study can be used to aid decision-making of collection management.

Although the mathematical functions have been validated and testified via simulated data, the authors welcome further questioning, testing, and validating with different data pools from other libraries. Later studies can also consider using Bayes' theorem to optimize the function by considering the probability of items' being successfully borrowed (P) as the needed prior probability.

Reference

ACRL Research Planning & Review Committee (2016). *2016 top trends in academic libraries A review of the trends and issues affecting academic libraries in higher education*. Retrieved March 16, 2020, from http://hdl.handle.net/1805/11853.

Agee, J. (2005). Collection evaluation: A foundation for collection development. *Collection Building, 24*(3), 92-95.

Basurto-Flores, R., Guzmán-Vargas, L., Velasco, S., Medina, A., & Hernandez, A. C. (2018). On entropy research analysis: Cross-disciplinary knowledge transfer. *Scientometrics, 117*(1), 123-139.

Borin, J., & Yi, H. (2008). Indicators for collection evaluation: a new dimensional framework. *Collection Building, 27*(4), 136-143.

Borin, J., & Yi, H. (2011). Assessing an academic library collection through capacity and usage indicators: testing a multi‐dimensional model. *Collection Building, 30*(3), 120-125.

Broadus, R. N. (1997). The applications of citation analysis to library collection building. In J. V. Melvin (Eds.), *Advances in librarianship* (pp. 299-335). New York: Academic Press.

Cook, C., & Maciel, M. (2010). A decade of assessment at a research-extensive university library using LibQUAL+. *Research Library Issues, 271*(8), 4-12.

Day, M., & Revill, D. (1995). Towards the active collection: the use of circulation analyses in collection evaluation. *Journal of Librarianship and Information Science, 27*(3), 149-157.

Dee, C. R., Rankin, J. A., & Burns, C. A. (1998). Using scientific evidence to improve hospital library services: Southern Chapter/Medical Library Association journal usage study. *Bulletin of the Medical Library Association, 86*(3), 301-306.

Dinkins, D. (2003). Circulation as assessment: Collection development policies evaluated in terms of circulation at a small academic library. *College & Research Libraries, 64*(1), 46-53.

Duncan, C. J., & O'Gara, G. M. (2015). Building holistic and agile collection development and assessment. *Performance Measurement and Metrics, 16*(1), 62-85.

Frenken, K., & Leydesdorff, L. (2000). Scaling trajectories in civil aircraft (1913–1997). *Research Policy, 29*(3), 331-348.

Gregory, V. L. (2011). *Collection development and management for 21st century library collections: An introduction.* New York: Neal-Schuman Publishers.

Harker, K. R., & Klein, J. (2016). SPEC Kit 352: Collection assessment. Retrieved March 16, 2020, from https://doi.org/10.29242/spec.352.

Henderson, A. (2000). The library collection failure quotient: The ratio of interlibrary borrowing to collection size. *Journal of Academic Librarianship, 2*6(3), 159-170.

Henry, E., Longstaff, R., & Van Kampen, D. (2008). Collection analysis outcomes in an academic library. *Collection Building, 27*(3), 113-117.

Ho, L. H. S. (2018). Collection assessment for a Middle Eastern, English curriculum university library. *Collection and Curation, 37*(3), 128-133.

Intner, S. (2003). Making your collections work for you: collection evaluation myths & realities. *Library Collections, Acquisitions, and Technical Services, 27*(3), 339-350.

Johnson, P., Hille, J., & Reed, J. A. (2005). Collection analysis: Evaluation and assessment. In P. Johnson (Eds.), *Fundamentals of collection development and management* (pp. 225-263). Chicago: ALA Editions.

Johnson, Q. (2016). Moving from analysis to assessment: Strategic assessment of library collections. *Journal of Library Administration, 56*(4), 488-498.

Kaplan, R., Steinberg, M., & Doucette, J. (2006). Retention of retrospective print journals in the digital age: Trends and analysis. *Journal of the Medical Library Association, 94*(4), 387-393.

Kelly, M. (2014). Applying the tiers of assessment: A holistic and systematic approach to assessing library collections. *Journal of Academic Librarianship, 40*(6), 585-591.

Kyrillidou, M. (2006). The impact of electronic publishing on tracking research library investments in serials. *ARL: A Bimonthly Report on Research Library Issues and Actions from ARL, CNI, and SPARC, 249*, 6-7.

Lantzy, T., Matlin, T., & Opdahl, J. (2020). Creating a library-wide collection management cycle: One academic library's approach to continuous collection assessment. *Journal of Library Administration, 60*(2), 155-166.

Leydesdorff, L. (2002). Indicators of structural change in the dynamics of science: Entropy statistics of the SCI Journal Citation Reports. *Scientometrics, 53*(1), 131-159.

Leydesdorff, L., & Rafols, I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics, 5*(1), 87-100.

Liu, Y., & Rousseau, R. (2009). Properties of Hirsch-type indices: The case of library classification categories. *Scientometrics, 79*(2), 235-248.

Long, M. P., & Schonfeld, R. C. (2014). Ithaka S + R US library survey 2013. New York: Ithaka S + R.

Lu, S. (2005). Performance measurement and measure indicators for public libraries. *Bulletin of Library and Information Service, 53*, 23-42.

Luther, M., & Guimarães, A. (2016). *Applying the principles of total library assessment to inform sustainable collection development*. Retrieved March 16, 2020, from http://works.bepress.com/ana_luther_guimaraes/24/.

Matricciani, E. (1994). Shannon's entropy as a measure of the "life" of the literature of a discipline. *Scientometrics, 30*(1), 129-145.

Murphy, E. (2013). Assessing university library print book collections and deselection: A case study at the National University of Ireland Maynooth. *New Review of Academic Librarianship, 19*(3), 256-273.

Prathap, G., & Mittal, R. (2010). A performance index approach to library collection. *Performance Measurement and Metrics, 11*(3), 259-265.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3), 379-423.

Silva, F. N., Rodrigues, F. A., Oliveira Jr, O. N., & Costa, L. D. F. (2013). Quantifying the interdisciplinarity of scientific journals and fields. *Journal of Informetrics, 7*(2), 469-477.

Tabacaru, S., & Pickett, C. (2013). Damned if you do, damned if you don't: Texas A&M University Libraries' collection assessment for off‐site storage. *Collection Building, 32*(3), 111-115.

Wilde, M., & Level, A. (2011). How to drink from a fire hose without drowning: Collection assessment in a numbers-driven environment. *Collection Management, 36*(4), 217-236.

Zhang, C., & Zhou, Q. (2020). Assessing books' depth and breadth via multi-level mining on tables of contents. *Journal of Informetrics, 14*(2), 101032.