Identifying potentially attractive publications in a digital library which remain invisible to users

John Catlow, Mirosław Górny, Rafał Lewandowski
Adam Mickiewicz University, Poznań, Poland

Abstact: Digital library collections contain publications of different degrees of attractiveness to users. However, some publications belonging to categories that are generally sought by users appear to be "invisible". The Digital Library of Wielkopolska has implemented a method enabling such items to be identified. This involves monitoring the average daily numbers of views of individual publications available in the library. Among the 100 most frequently accessed items, certain categories are seen to be well represented; these indicate the types of publications which are most popular among users. The lowest daily rate of views among these 100 items is taken as a threshold for visibility. If a publication belongs to one of the popular categories, but its average daily number of views is below this threshold, it may be classified as a "lost" item. It is then necessary to determine the reason for this. The cause is usually the inadequacy of the information contained in the item's description.

Keywords: digital libraries; users; collections**.**

1. Introduction

The creation of digital libraries has brought changes in the methods used to track the popularity of books and journals. It is true that traditional lending libraries are able to provide statistical data concerning numbers of borrowings of particular items. In digital libraries, however, data are collected concerning the number of times publications are accessed. It must be remembered, of course, that certain publications may be printed out by users or downloaded onto disk, in which case the situation is similar as with traditional libraries. Nonetheless, the great majority of digital library users make use of publications in online mode.

Whether the data concern borrowings from traditional libraries or views in digital libraries, it is not revealed how the publication is in fact used – whether it was read as a whole or merely browsed. The digital library data also fail to provide information on who accessed the item; hence it is possible that very

Received: 31.3.2017 Accepted: 21.5.2017 ISSN 2241-1925

intense use of a given publication may be generated by a relatively small number of users. It is possible, of course, to analyse the IP addresses of those accessing publications, but this would require a more advanced system for analysing the frequency of views.

The main use of information of this type is to identify items which belong to intensively used categories, but which are accessed a relatively small number of times. Such an item can be assumed to have low visibility to information users, and may be referred to as a “lost” item.

In every digital library there are categories of items which are used with very high (above-average) frequency. An item belonging to such a category is usually found to have been accessed a large number of times after the elapse of a certain period since its being placed in the library. If the level of use of such an item is found to be low, then it is desirable to determine the reasons for this.
The existence of such items in the library, even if only in small number, can be considered to represent “intellectual loss”. The value of a publication is hard to measure in itself, and an appropriate evaluation can be made solely in the context of the environment of the information user.It is often the case that a single publication can turn out to be of key importance to particular readers.

2. Analysis and results

An analysis was carried out on data from the Digital Library of Wielkopolska, one of Poland’s largest regional digital libraries. It contains scans of many items from the collections of the University Library in Poznań and the library of the Polish Academy of Sciences (PAN) in Kórnik, as well as resources supplied by the National Archives in Poznań and the library of the Poznań Society of Friends of Science. Other items come from the Provincial Public Library, the Raczyński Library, the public libraries in Ostrów Wielkopolski and Krotoszyn, and the libraries of Poznań’s College of Physical Education and the Universities of Medicine, Technology, Life Sciences and Fine Arts.

At the time when the analysis was carried out, the library contained a total of 288,909 publications. Approximately 73% of these were newspapers and magazines from the 19thcentury and from the interwar period. This collection consists of 632 titles, chiefly with a connection to Poznań and the Wielkopolska region. The remaining publications include books, manuscripts (approximately 8000), other old printed matter (approximately 4000 items) and archive materials (approximately 6000 items). Polish-language publications dominate, although 21% of items are in German, and approximately 3% in Latin.

Based on analysis of the frequency with which library items were accessed, it was found overall that the most frequently used 20%of items (57,782) accounted for 75.25% of the total number of views. The “80/20 rule «was thus borne out in relation to the whole of the library’s collection.

The distribution of cumulative numbersof views for individual items revealed that the most popular item had attracted approximately one million views, the next two items had one million views in total, and the next three again had a total of one million views (to an accuracy of around 10,000 views). The continuation of this sequence (successive numbers of items totalling approximately one million views) was as follows: 5, 7, 11, 23, 45, 98, 187, 280, 427,670, 1,045, 1,493, 2,098, 3,024, 4,028, 6,136, 8,318, 10,637, 13,594, 16,217, 22,384, 30,755, 35,118, 60,119. The final group, made up of the remaining 71,821 publications, attracted only around 300,000 views in total.

Thus the sequence of average numbers of views for successive groups of items, each accounting for one million views, was as follows:

Number of publications in the group	Average cumulative number of views (approx.)
1	**1,000,000**
2	**500,000**
3	**330,000**
5	**200,000**
7	**140,000**
11	**91,000**
23	**43,000**
45	**22,000**
98	**10,000**
187	**5,300**
280	**3,500**
427	**2,300**
670	**1,500**
1,045	**960**
1,493	**670**
2,098	**480**
3,024	**330**
4,028	**250**
6,136	**160**

8,318	**120**
10,637	**94**
13,594	**74**
16,217	**62**
22,384	**45**
30,755	**33**
35,118	**28**
60,119	**17**

Table 1. Average cumulative numbers of views within groups of publications attracting one million views in total.

The number of times a given item is accessed can be assumed to depend on its subject area, its standing within its field, the technique typically used to access it (the number of views depends on whether the item is used online or copied locally), and the length of time for which it resides in the library's collections. External factors naturally also play a role, such as the size of the population interested in a given topic, and the characteristics of that population (intensity of work, preference for online working, etc.).

It is not possible, of course, to measure these factors precisely, or to construct mathematical formulae which could be used for exact prediction of levels of use of library items. It may be supposed that the cumulative number of views will be approximately proportional to the length of time for which an item has been available in the library. Analysis of the 100 most popular items, however, showed a lack of correlation between these two parameters (the Pearson correlation coefficient was 0.172). Analysis for the first 50 items revealed only a weak correlation, at a level of 0.21. It must be concluded, then, that the frequency of use of library items is highly variable under the influence of the other factors mentioned. An analysis was thus made of the frequencies with which individual items had been accessed over the period for which they had been available in the library, expressed in average numbers of views per day.

For the 100 most popular items, the sequence of these daily access frequencies (rounded to the nearest integer) was as follows: 243, 218, 203, 128, 120, 113, 100, 89, 88, 65, 63, 62, 61, 59, 50, 49(x2), 44(x3), 43, 42(x2), 38(x2), 36(x2), 31, 30, 29(x2), 28(x3), 27, 24, 23, 22, 21, 20, 18(x2), 15(x4), 14(x2), 13, 12(x8), 11(x4), 10(x2), 9(x2), 8(x6), 7(x7), 6(x8), 5(x11), 4(x2). The arithmetic mean was 30.5.

This list of the most popular 100 library items was selected as a basis for comparison, among which the lowest value for the daily number of views was 4.

Certain categories of items in the collections are found to be well-represented in this set. In turn, items belonging to those same categories for which the daily number of views is well below the value of 4 might be considered to be items whose potential for use is possibly not being fully realised. These may be referred to as “lost” items.

The first hundred items included:

- 13 textbooks for chemistry students (e.g. *Laboratory Exercises in Inorganic Chemistry*);
- 4 dissertations in economics (e.g. *An evaluation of the implementation and effectiveness of methods of quality management according to suppliers in the motor industry*);
- 28 publications relating to genealogy (e.g. *Genealogies of Living Titled Polish Families*);
- 4 items relating to history (e.g.*Diary of Stefan Batory’s Expedition to Pskov*);
- 18 dissertations in medicine (e.g. *British policy on the promotion of health after the Second World War: a comparative study*);
- 10 works on sports science (e.g. *Analysis of selected determinants of the formation of healthy attitudes in students*);
- 26 other works (e.g. *Theatrum chemicum, praecipuos selectorum auctorum tractatus de chemiae et lapidis philosophici antiquitate, verit*).

3. Interpretation of the results

3.1. Medical dissertations

Medical dissertations are supplied by the University of Medicine in Poznań; the library currently has 1105 of them. Why are 18 of them found among the library’s 100 most popular items? They are mostly works which were added to the Digital Library of Wielkopolska in 2011–2013. This was a time in which medics were “discovering” the digital library and beginning to use it in large numbers. They use it for two main purposes: to search for data, and to search for works considered as models for their own doctoral and habilitation dissertations. Hence certain items, regarded as exemplary, enjoy high levels of popularity. Because these works have long, descriptive titles, it is not difficult to reach the sought item.

3.2. Textbooks for chemistry students

The Digital Library of Wielkopolska has a relatively small number of such books (around one hundred), and these are hugely popular among students. There are 13 of them among the 100 most popular items in the library. Their titles are precise, and it is easy to find them within the library’s collection.

3.3. Works on sports science

The exceptional popularity of these works results from the fact that they are intensively used as materials for exercises in using the digital library by students of Poznań's College of Physical Education. Moreover, the first item on the list of the 100 most frequently used appears on the College's website at the head of the list of most popular works, and is therefore very frequently clicked on as an example. These works carry detailed descriptions, and it is relatively easy to reach them.

3.4. Genealogy

Genealogical works enjoy a very high level of interest, since 60% of digital library users are people with an interest in genealogy or amateur historians, often simply seeking information about their ancestors. It is this category that contains the largest number of "lost" items. Why is this?

Firstly, such items often have titles that do not indicate that their content relates to genealogy. For example, the work titled *Kalisz Calendar for the Year 1914*contains the names of many government employees, landowners and entrepreneurs. However, an amateur reader is likely to expect the work to contain information of the kind typical for a calendar, and not to identify it as a potential source of information about specific people. Similarly, the monthly *Forestry Review* is viewed as a specialist journal; few users will suspect that it contains lists of employees of the national forestry service.

A second reason, in some cases, is the language and font used. This applies especially to German directories of addresses printed in Gothic type, and also to the handwritten records of craftsmen's guilds, particularly those dating from the time of the Prussian partition and thus written in German.

3.5. Others

In the case of other items appearing on the list of the most frequently accessed publications, it was not possible to determine the reason for their exceptionally high level of interest. They also concern very diverse subject matter. For these reasons they will not be considered further here.

4. Solving the problem

What should be done, therefore, to minimise the number of "lost" items? First of all, a significant improvement needs to be made in the available information concerning them. It is therefore necessary to do the following:

1. Provide them with additional keywords (lists of names, addresses, surnames, etc.);
2. Group them into collections, which are de facto lists of publications appearing on the lists visible on the digital library's homepage;
3. Supplement the titles with detailed information about the content of the publications;

4. Where possible, apply OCR processing;
5. Enable Google search bots to access the library catalogues.

5. Conclusions

Works on genealogy are currently the only group which is monitored in terms of the retrieval of lost items. If the daily number of hits on an item is below 4.0, then it may be considered to require intervention.

Nonetheless, it is becoming more common for the library's cataloguers to attempt to predict such cases based on features of publications, and to take appropriate action at the cataloguing stage. Contact with readers, to whom assistance is given particularly when making genealogical searches, is extremely helpful in this.

Unfortunately, a significant problem at the present time is the use of the DjVu format. Since March of this year, Mozilla 52 has not supported plugins with the NPAPI infrastructure, and Chrome ceased doing so several months previously. Users are forced to use HTML 5DjVu conversion, which has many faults. For example, it runs slowly, and does not enable text searching.

The only solution to this problem is to copy items to a local disk and to browse them using a DjVu viewer. This is a further obstacle which has a very significant effect on the frequency of access for publications belonging to the most intensively used categories, and this is a problem that will be very difficult to solve