The Feasibility of Investing in Manual Correction of Metadata for a Large-Scale Digital Library

Hung-Hsuan Chen
Computational Intelligence Technology Center,
Industrial Technology Research Institute
Hsinchu, Taiwan
hh.chen@itri.org.tw

Madian Khabsa[†], C. Lee Giles[†‡]
[†]Computer Science and Engineering,
[‡]Information Sciences and Technology,
The Pennsylvania State University,
University Park, PA, US
madian@psu.edu, giles@ist.psu.edu

ABSTRACT

Given a large-scale digital library that automatically crawls and parses PDF files to generate metadata for documents and authors, we estimate the number of person-hours required to correct a small portion of the metadata, in the hope that a large portion of users can benefit from these corrections. We obtain users requests by analyzing CiteSeerX's log files from September 2009 to March 2013. We found that the distribution of users requests for search is highly imbalanced: most document search queries and author search queries concentrate on a small set of terms. As a result, even for a large-scale digital library, we estimate it is affordable to invest a few person-hours to check the correctness of a few metadata, and thus provide benefits to a good portion of document search and author search requests.

Categories and Subject Descriptors

H.1.2 [**MODELS AND PRINCIPLES**]: User/Machine Systems—*Human information processing*; I.7.1 [**DOCUMENT AND TEXT PROCESSING**]: Document and Text Editing—*Document management*; H.3.7 [**INFORMATION STORAGE AND RETRIEVAL**]: Digital Library—*Collections, System issues, User issues*

General Terms

Languages, Algorithm, Experimentation

Keywords

Digital Library, User Satisfaction, User Experience, Metadata Correction, Human-Aided Metadata Generation, Practicability

1. INTRODUCTION

In the last few decades, the number of scientific publications has grown rapidly. These publications include not only traditional peer-reviewed papers, but also papers published in other media, such as e-print services (e.g., arXiv[1]) and personal web pages. However, traditional academic literature databases, such as Science Citation Index (SCI) and Social Science Citation Index (SSCI), cover only a small portion of these publication platforms [10]. In addition, conference publications, which are important and peer-reviewed papers in certain fields, are usually ignored by these databases. As a result, several researchers have criticized that these traditional databases may be biased toward a limited number of journals due to source selection [2].

To include a large number of academic papers in the databases, researchers have proposed autonomous platforms, such as Google Scholar[2] and CiteSeerX[3], that crawl and digest scientific papers from the Internet with little manual effort. Unfortunately, these papers, which are usually in a PDF format, are challenging to parse, because the PDF format contains limited structured metadata. As a result, these autonomous platforms sometimes extract incorrect document metadata, such as incorrect titles, author names and affiliations, references list, etc. This incorrect information effects user experience and scientific research statistics, such as the h-index of authors and the impact factor of venues.

Several modern digital libraries still collect bibliography manually because of the high quality of the collected data. One typical example is DBLP[4], which includes a large number of conference and journal publications in the Computer Science domain. However, such a process is laborious. Although it is possible to rely on crowdsourcing platforms, such as Amazon Mechanical Turk[5], to collect the metadata more efficiently, certain tasks, such as author name disambiguation, require domain knowledge, and may not be appropriate by such an approach.

Instead of completely manual editing or completely automatic parsing, we show that automatic parsing with a small amount of manual checking can provide high quality content to a good portion of query requests. Given that manual resources are expensive, it is impractical to manually check every piece of metadata for a large-scale digital library. Therefore, a natural question to raise is which metadata should be examined? One may choose to examine the fields in which the parsers frequently make mistakes.

[1] http://arxiv.org/
[2] http://scholar.google.com/
[3] http://citeseerx.ist.psu.edu/
[4] http://dblp.uni-trier.de/
[5] https://www.mturk.com/

978-1-4799-5569-5/14/$31.00 ©2014 IEEE.

Table 1: The number and the percentage of each search type based on the logs.

Search Type	Number	Percentage
Document	46,770,997	62.07%
Author	28,502,533	37.82%
Table	35,073	0.04%

Therefore, human resources are allocated to review the most error-prone fields. Such a choice requires some understanding about the capability of parsers. In contrast, we show that most users submit similar requests. Hence, we may allocate the human resources to check the metadata related to these highly demanded requests. Using the CiteSeerX digital library as the experimental target, we estimate that investing less than 160 person-hours of manual correction on selected documents can benefit 10% of the document searches, and investing less than 7 person-hours of manual checking on selected authors can benefit 20% of the author queries. This seems to be a worthwhile and affordable investment for most digital library providers.

2. DISCOVERING THE MOST DEMANDED SEARCH REQUESTS

CiteSeerX currently collects over 4 million academic documents in the Computer Science and the Information Science domain, with a recent addition in Physics and Medicine. To discover the most demanded search requests from the CiteSeerX users, we rely on the logs collected between 2009/09 and 2013/03. These logs contain over 3 billion entries.

Processing such a large number of entries by a traditional computing environment is infeasible. To investigate the huge log files within affordable time, we import the logs into Apache Hive warehouse[6], a framework supporting distributed storage and Map-Reduce computing. We utilize HiveQL[6] and Apache Pig Latin[7] to communicate with Hive for data processing.

2.1 Search Queries: Understanding Users' Interests

Among the 3 billion log entries, over 70 million of them are search requests. We utilize these requests to figure out users' search interests.

CiteSeerX currently supports document search, author search, and table search. In this paper, we focus on document search and author search, since the two types together account for over 99% of the search requests. Table 1 lists the percentage of each type of searches based on the logs.

2.2 The Measurement of Inequality

The Pareto principle describes a highly skewed distribution in which a large portion of effects is contributed by a small portion of causes. This phenomenon is observed in many situations, including citation distribution, network degree distribution, word frequency distribution, etc. We surmise that users' search queries follow a similar condition: most users' queries concentrate on a small number of terms. By manually correcting the metadata related to these queries, we can better serve most of the users.

[6]`http://hive.apache.org/`

[7]`https://pig.apache.org/`

Table 2: The most popular document queries and author queries (punctuation marks are removed, queries with more than 5 words are pruned because they are regarded as "titles").

Rank	Document Query	Author Query
1	imbalanced	Chen
2	DNS	Lee
3	control system security	Smith
4	scada	Wang
5	pattern matching	Li
6	workload characterization cloud computing	Zhang
7	data mining	Johnson
8	van	Liu
9	brain computer interface	Anderson
10	personal learning environment	Miller

To measure the inequality, a popular method is fitting the distribution to the power-law distribution, and inferring the level of inequality by the fitted exponent: a larger exponent usually indicates a larger degree of inequality. However, even if a small set of terms include most queries, their distribution may not follow power-law. In addition, the estimation of exponent and other parameters in the power-law distribution is expensive [5]. Alternatively, we quantify the inequality by showing the Lorenz Curve, the Gini coefficient, and the balanced inequality ratio [7,9]. These scores are easy to compute. More importantly, they do not assume the underlying distribution of the effects and the causes. Therefore, these indexes can be applied on a wider range of applications.

The Lorenz Curve is widely used by Economists to visualize the wealth distribution. However, it did not receive much attention in the Computer Science field until recently [9]. Traditionally, the Lorenz Curve shows the cumulative percentage of the income earned by the bottom percentage of the individuals.

By the Lorenz Curve, one can easily figure out "the bottom x% of individuals earn y% of the total income". In a perfectly balanced situation, the value of x always equals y. The area between the perfectly balanced situation and the Lorenz Curve is a measurement of income inequality. The Gini coefficient is defined as the twice of the area, and therefore the value is always between 0 and 1, which represents fully equal (everyone earns the same income) and completely unequal (the richest person earns all the income) respectively.

The balanced inequality ratio states that p% of the richest individuals earn the $(1-p)$% of the income [9]. The balanced inequality ratio is commonly known as the 80-20 rule: 80% of the wealth is owned by 20% of the rich. The value of p is obtained by finding the intersection between the Lorenz Curve and the line $x + y = 1$ [9].

2.3 The Most Demanded Queries in Document Search

In this section, we study the most demanded queries in document search, which is the default search of CiteSeerX. From the logs, we found that many CiteSeerX users submit

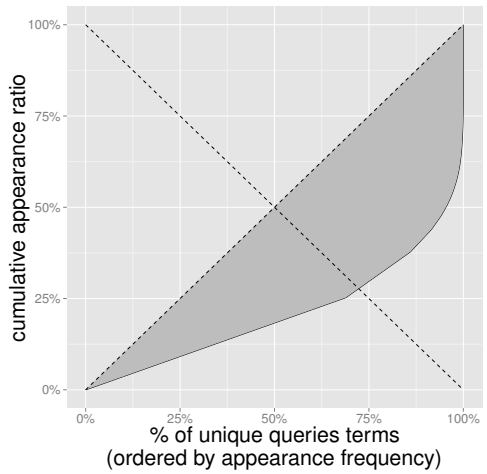

Figure 1: The Lorenz Curve of the document queries and their frequencies. Gini Coefficient: 0.5837**, balanced inequality ratio: the most frequent** 27.75% **document queries account for** 72.25% **of the document query traffic.**

author names in the document search form. This is probably because document search is the default search form, and users simply submit queries without further checking. To prune these "author name queries" from document search, we apply a simple rule to identify these name queries: a query is regarded as a name if 1) the query contains less or equal to 3 words, and 2) every word starts with an uppercase character and all the remaining characters are lowercase.

After pruning author names, we rank all the remaining terms by the number of submitting frequencies and plot the Lorenz Curve, as shown by the solid line in Figure 1. The diagonal dash line connecting $(0, 0)$ and $(1, 1)$ represents the perfect equality case. The area between this diagonal line and the Lorenz Curve represents the level of inequality. The Gini coefficient, as discussed earlier, is defined as twice of the area, and therefore the value is between 0 (perfect equality) and 1 (completely inequality). The Gini coefficient of the query frequency is 0.5836, which suggests that the inequality is extremely manifest. The balanced inequality ratio is obtained by the intersection of the Lorenz Curve and the other diagonal line that connects $(0, 1)$ and $(1, 0)$. It shows that the top 27.75% of query terms accounts for 72.25% of the document query traffic.

2.3.1 *Estimated Person-Hours to Clean the Metadata for Documents*

Now assuming we want to improve 5% of the document search requests, we only need to correct documents related to the 62 most frequent queries (the top 0.0012% of the distinct queries), because these queries account for 5% of the document search traffic. Assuming the metadata of the top 10 returned documents of the 62 queries are manually examined and each document takes 5 minutes for correcting, we will need about 51.67 person-hours, which should be affordable for most digital library service providers. We select to clean the top 10 returns because the logs show that more than 92% of users visit only the first page (the top 10 results) after searching.

Table 3: The estimated person-hours needed to correct portions of the metadata of documents.

Number of Queries to Examine	Percentage of Document Search Benefited	Estimated Person-Hours needed
62 (top 0.0012%)	5%	51.67
190 (top 0.0038%)	10%	158.33
540 (top 0.0108%)	15%	450.00
1477 (top 0.0294%)	20%	1230.83

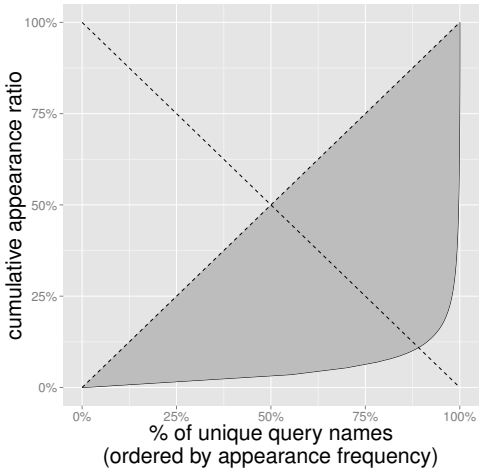

Figure 2: The Lorenz Curve of the name queries and their frequencies. Gini Coefficient: 0.8944**, balanced inequality ratio: the most frequent** 10.96% **name queries account for** 89.04% **of the name query traffic.**

Table 3 lists some more person-hours estimations. By recruiting 10 workers to correct the data for three week (i.e., about 1200 person-hours), we can benefit about 20% of the document search requests.

2.4 The Most Demanded Authors

We include both the query terms in author search and the terms that are identified as "author names" in document search in this section. Similar to Section 2.3, we plot the Lorenz Curve of the author name queries, as shown in Figure 2. We found that the distribution of name queries is even more unequal, compared to the distribution of document queries. The Gini coefficient is 0.8944, and the balanced inequality ratio suggests that the most frequent 10.96% name queries accounts for 89.04% of the name query traffic.

2.4.1 *Estimated Person-Hours to Clean the Metadata for Authors*

Since the name queries distribute more imbalanced than the document queries, we can examine fewer queries to benefit a good portion of the name query requests. Table 4 shows the estimated person-hours needed to benefit 5%, 10%, 15%, and 20% of the name query requests. By only examining the returned names of the most frequently queried name, we can benefit 5% of the author queries. Again, assuming we manu-

Table 4: The estimated person-hours needed to correct portions of the metadata of authors.

Number of Name Queries to Examine	Percentage of Author Search Benefited	Estimated Person-Hours needed
1 (top 0.00006%)	5%	1.40
2 (top 0.00012%)	10%	2.77
4 (top 0.00025%)	15%	5.56
5 (top 0.00031%)	20%	6.93

ally check the metadata of the top 10 returned authors and the correction takes about 5 minutes for each author, we need only 1.40 person-hours. By only examining the most frequent 5 name queries and check the top 10 returned authors for each of these queries (estimated cost is only 6.93 person-hours), we can benefit 20% of the name query requests.

3. RELATED WORK

For various reasons automatic metadata extraction for academic documents continues to be an active topic [1, 8, 11, 12].

Most academic documents present their content in PDF format, because this format encapsulates details for rendering information, such as font type, text size, text color, figure location, line width, etc. This enables PDF to represent a document independently of the computing environment, i.e., a PDF document is highly portable. However, PDF is not designed for presenting structured metadata. Therefore, the PDF parsers mostly infer metadata based on various heuristics, such as the styles [1] (e.g., font size and position, gap between lines), keywords or keyphrases matching [6,8] (e.g., "Section", "Figure", "Reference", "Abstract"), or referring to affiliated knowledge resources [3,4,13] (e.g., looking for similar records from DBLP or CiteSeerX). Since papers published in different venues apply different templates, a parser that performs well for a certain conference or a certain field may be error-prone in others [11]. As a result, manually compiled databases are still desirable in many cases.

4. DISCUSSION AND FUTURE WORK

It is very challenging to build a metadata parser that performs well for all types of academic documents. Until a better retrieving technique is created, we may need a certain degree of manual involvement in metadata editing. In this paper, we found that most users submit a similar set of document queries and author queries. As a result, it is possible to invest a small number of person-hours for metadata correction such that a reasonable portion of user requests are better served.

In addition to person-hours, we may also assign other resources based on the request distribution. For example, assuming we have an accurate but time-consuming parser and a less-accurate but fast parser, it may be worthwhile to apply the first parser to extract the metadata of the few highly requested items.

5. ACKNOWLEDGEMENT

We gratefully acknowledge partial support from the NSF.

6. REFERENCES

[1] J. Beel, S. Langer, M. Genzmehr, and C. Müller. Docear's pdf inspector: title extraction from pdf files. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 443–444. ACM, 2013.

[2] L. Butler. ICT assessment: moving beyond journal outputs. *Scientometrics*, 74(1):39–55, 2008.

[3] H.-H. Chen, L. Gou, X. Zhang, and C. L. Giles. CollabSeer: a search engine for collaboration discovery. In *Proceedings of the 11th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 231–240. ACM, 2011.

[4] H.-H. Chen, P. Treeratpituk, P. Mitra, and C. L. Giles. CSSeer: an expert recommendation system based on CiteseerX. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 381–382. ACM, 2013.

[5] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[6] I. G. Councill, C. L. Giles, and M.-Y. Kan. ParsCit: an open-source CRF reference string parsing package. In *LREC*, 2008.

[7] J. L. Gastwirth. The estimation of the lorenz curve and gini index. *The Review of Economics and Statistics*, pages 306–316, 1972.

[8] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic document metadata extraction using support vector machines. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 37–48. IEEE, 2003.

[9] J. Kunegis and J. Preusse. Fairness on the web: Alternatives to the power law. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 175–184. ACM, 2012.

[10] P. O. Larsen and M. von Ins. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, 84(3):575–603, 2010.

[11] M. Lipinski, K. Yao, C. Breitinger, J. Beel, and B. Gipp. Evaluation of header metadata extraction approaches and tools for scientific pdf documents. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 385–386. ACM, 2013.

[12] F. Peng and A. McCallum. Information extraction from research papers using conditional random fields. *Information Processing & Management*, 42(4):963–979, 2006.

[13] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998. ACM, 2008.