The Impact of User Corrections On A Crawl-Based Digital Library: A CiteSeerX Perspective

(Invited Paper)

Jian Wu, Kyle Williams
Information Sciences and Technology
The Pennsylvania State Univeristy
University Park, PA, 16802
Email: jxw394@ist.psu.edu,
kwilliams@psu.edu

Madian Khabsa
Computer Science and Engineering
The Pennylvania State University
University Park, PA, 16802
Email: madian@psu.edu

C. Lee Giles
Information Sciences and Technology
Computer Science and Engineering
The Pennylvania State University
University Park, PA, 16802
Email: giles@ist.psu.edu

***Abstract*—CiteSeerX is a crawl-based digital library search engine providing free access to more than 4 million academic papers. Since metadata in the digital library is obtained through automatic extraction, it is inevitable that errors will occur. CiteSeerX offers a feature allowing registered users to correct paper metadata including titles, authors, abstracts, publication years, venues, etc. We claim that user corrections, as a form of crowd-collaboration, provide a useful and efficient way to improve metadata quality and the impact of the digital library. As evidence to support this claim, we investigate user corrections from the last 5 years and analyze: the nature of the corrections; the quality of the corrections; and the impact of the corrections on downloads.**

I. Introduction

Digital libraries generally obtain data from two sources: user submissions and Web crawling. In the first case, the documents and associated metadata are manually entered by users and the data are accurate and in standard formats. However, collecting such a corpus is slow and requires a large amount of manual effort. Some well-known examples of digital libraries that allow for the manual capture of metadata are PubMed, arXiv, IEEE Xplore, ACM Digital Library, Springer, and most other publishers. Some other digital libraries receive metadata that is "pushed" from publishers, as is the case for DBLP and Harvard ADS.

In the second case, documents are first harvested by crawlers. The associated metadata are then extracted in an automated manner, which is much faster. Some famous examples are Google Scholar, Microsoft Academic Search, and CiteSeerX. However, because the documents are downloaded from the public Web, their formats vary significantly. Furthermore, since many documents available on the Web are preprints, the metadata embedded inside the documents themselves may not be the same as the final published version. These complexities make it challenging to extract metadata from these documents and errors in extraction are common. Automatic data correction after extraction is possible, but it is usually biased towards certain types of documents and although it can correct a fraction of metadata, it cannot guarantee to correct all mistakes and it is hard to verify which documents still have incorrect metadata. Therefore, human interactions have to be involved to improve the metadata quality in a collaborative manner.

CiteSeerX is a digital library search engine providing over four million publicly available academic documents. It has a focused crawler, which actively crawls the Web and downloads documents in PDF formats. These documents are processed by a filter, which removes non-academic documents. The academic documents are then sent to a metadata extraction module, which identifies, extracts, and parses header, text body, and citation metadata. The header information is extracted by SVMHeaderParser [1], a SVM-based extractor. The header includes 15 fields: titles, authors, affiliation, address, note, email, date, abstract, introduction, phone, keywords, web, degree, publication number, and page information. Citations generally contain the same fields as the headers, but are retrieved by ParsCit [2], which is a citation string parser that uses conditional random fields for citation parsing.

Due to the complexities mentioned above, it is inevitable that there will be some errors in the metadata that is automatically extracted from papers crawled from the public Web. To address this, CiteSeerX has a feature allowing registered users to make corrections to header metadata. This feature has been available in CiteSeerX since 2008 and, thus far, we have collected 400,000 corrections. We view the process of users correcting metadata as a from of crowd-collaboration. In this paper, we study the impact of the user corrections by investigating the correction history. We analyze the quality of the metadata produced by user corrections and we examine the impact that user corrections have on the number of times papers are downloaded.

In describing the above, this paper is structured as follows. In Section II, we introduce the user correction feature in CiteSeerX, and describe how user corrections changes the paper cluster and citation graph. In Section III, we investigate the CiteSeerX user correction history to find out who corrected our metadata and the types of corrections that were made. In Section IV, we evaluate the automatically extracted metadata quality. Lastly, in Section V, we evaluate the impact of user corrections by comparing the download rates before and after a user correction takes place.

II. User Corrections in CiteSeerX

Many digital libraries allow users to manually correct the metadata. For example, Microsoft Academic Search allows users to change header information including titles, authors,

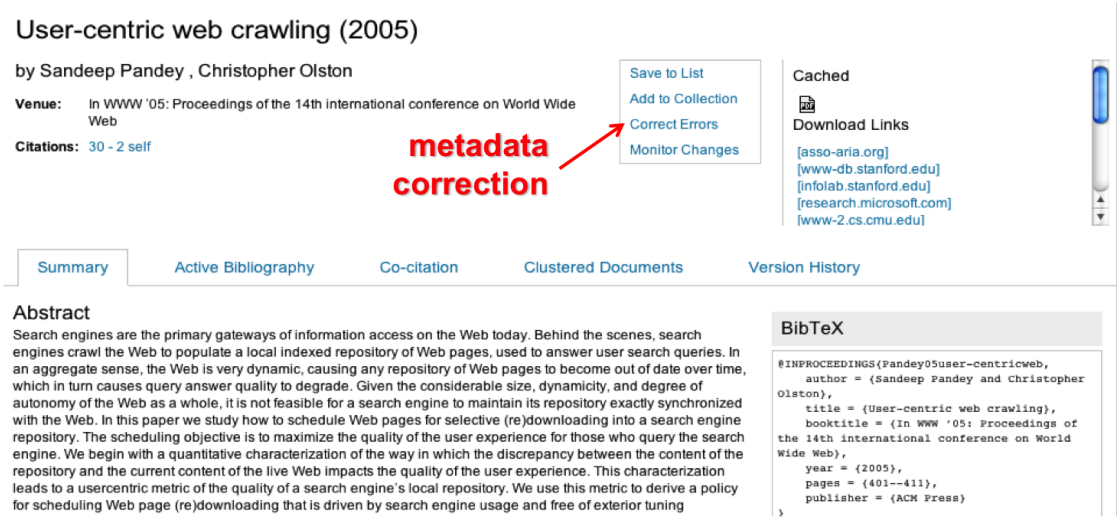

Fig. 1. The user correction web interface.

year, DOI, conference, journal, PDF URL and abstract. In Google Scholar, users are only allowed to edit metadata of their own papers. CiteSeerX allows users to correct metadata of all papers as long as they log in to their CiteSeerX account. Figure 1 shows a summary page of a paper in CiteSeerX. After clicking the "Correct Errors" link, users are directed to a page where they are allowed to change all header metadata assuming they have logged in to their accounts. The changes are effective right after they click the "submit" button.

In addition to changing the metadata of a paper, a user correction action also might alter the citation graph structure. The citation graph represents the citation relationships of all in-collection papers, i.e., papers with actual PDF files in our repository, as well as papers listed in the "reference" or "bibliography" sections of papers, which we call *citations*. Ideally, each citation in the reference section should have a corresponding PDF file in the repository. In practice, some citations do not have in-collection counterparts since we have not found them when crawling the public Web. Furthermore, some papers for which we have the PDFs might not be cited by any other papers.

CiteSeerX uses title and author information to group an in-collection paper and its citations into a *paper cluster*. If two papers have the same title and authors, but different content, we refer to them as near duplicates. We combine a paper, its citation mentions and any near duplicates into a *paper cluster* and each cluster represents a node in the citation graph.

After a user correction is submitted, the cluster where the corrected paper was in is deleted. Papers are *re-clustered* based on the updated metadata. The corrected paper could be grouped into another paper cluster if its title and authors are changed or it could also stay in the same cluster if other fields are changed but the title and authors remain the same. Figure 2 illustrates the changes of paper cluster after a user correction is submitted.

Besides updating metadata values and citation graphs in the main database, CiteSeerX preserves all paper version history. First, the main database contains a table to record information about each user correction action, including the paper ID, the user ID, and the version number. The repository server also writes a new XML file containing the updated metadata as well as the source of each field, i.e., from user correction or from an automated parser. In the next section, we study these version history files to understand details about user correction, including who corrected the metadata, what metadata were corrected, and which papers were corrected.

III. User Correction History

CiteSeerX contains more than four million documents. The challenge to evaluate the metadata is to obtain the ground truth, e.g., the true values of fields, from a big sample. Another challenge is that the CiteSeerX repository does not only contain papers, but also other academic documents, such as books, and theses. There are also as many as 15 metadata fields, so assigning equal weights to all fields is biased because certain fields can be more important than the others.

In this work, we randomly draw 1000 documents from the entire CiteSeerX collection between 2008 and 2013. We manually inspect these documents, label them, and extract metadata from of them. Although the sample size is only $1/4000$ of the entire repository, it covers a relatively large time range and the manual inspection of these documents is still the best way to generate a golden standard and gives the most meaningful evaluation.

We assign these documents to several pre-defined categories including "paper", "report", "book", "thesis", "slide", "resume", "poster", "abstract", "non-en", and "others" [3]. The general coverage of these categories is tabulated in Table I. Except "resume", "non-en", and "others", all categories are academic. The classification results indicate that more than 92% of the sampled documents are academic. In this work, we focus on "papers", because the metadata extractor was originally designed for this type of document. They are also what most people are interested in and make up the majority of the papers in the CiteSeerX collection.

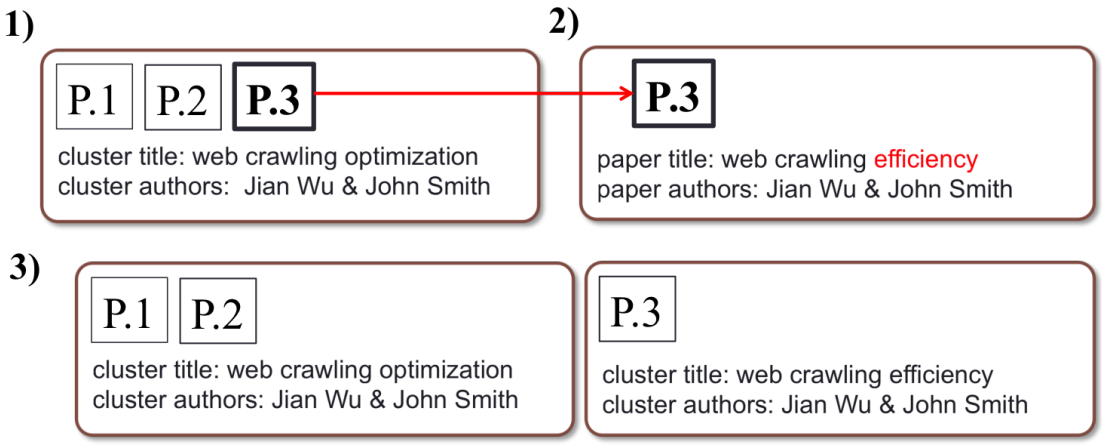

Fig. 2. Changes of paper cluster after a metadata correction is submitted. 1) P.1, P.2, and P.3 are grouped into the same cluster; 2) The title of P.3 is corrected by a user; 3) The initial cluster is deleted. All papers are re-clustered.

TABLE I. DOCUMENT CATEGORIES AND THEIR COVERAGE.

#	Category	Coverage[1]
831	paper	All journal articles, conference proceedings, and their manuscripts or pre-print versions. Research-oriented magazine articles.
45	report	Similar to "paper" except that the front page contains "tech report" with a number.
7	book	All published documents with ISBNs, or multi-chapter documents with similar structures.
26	thesis	A multi-chapter document with "thesis" on the front page.
8	slide	Lecture, conference, and product demonstration presentations.
0	resume	All commonly defined resumes, and CVs.
2	poster	A single-page document used for exhibition, usually containing faceted text, tables, and plots.
3	abstract	A paper without text-body.
3	non-en	Documents whose titles, and text-bodies are written in non-English languages.
75	others	All documents other than the categories above

[1] The descriptions here are just guidelines. The detailed category definitions are beyond the scope of this paper.

Next, we visually inspect all papers in our sample and manually extract titles and authors, which are the most important fields for to identify a document in a crawl-based digital library. This is different from submission-based digital libraries, in which the primary key for a document is a combination of a list of publication information, such as venue name (conference or journal), year, issue number, and page number. For a crawl-based digital library, these information fields are difficult to extract using the current techniques and often do not exist in the document. On the other hand, the combination of titles and authors can be used to uniquely identify the majority of papers published[1], and these are the most commonly used terms in search activities. In addition, titles and authors can be used to match a collection of paper metadata against another collection with more complete and accurate metadata. The results can be used for data cleaning tasks and consolidating metadata collections, e.g., [4]. Therefore, we argue that titles and authors are vital metadata of papers for a crawl-based digital library. Thus, when we present the evaluation of metadata quality, we choose to only focus on these fields.

A. Categories of Documents

To understand what categories of documents were corrected, we evaluate the categories of the random sample. The results are tabulated in Table II. This table shows that approximately 8% of the sampled documents were corrected by individual users. Among the corrections, 17% were *reports* and 8% were *papers*. The metadata of some *theses* were also corrected. Only one *other* document was corrected in this sample, which looks like a manuscript of a research note except that the letters printed are overlapped and unreadable. None of *books*, *slides*, *poster*, and *abstract* have been corrected.

TABLE II. COUNTS OF USER CORRECTED DOCUMENTS IN EACH CATEGORY.

Category	Documents	Corrected	Percentage
paper	831	73	8%
thesis	26	2	7%
report	45	8	17%
book	7	0	0%
slide	8	0	0%
poster	2	0	0%
abstract	3	0	0%
non-en	3	0	0%
others	75	1	1%
total	1000	84	8%

Percentages are rounded up to integers.

It should be noted that because the sample is small compared with the entire repository, the zero values of uncorrected categories only put lower limits to the real fractions.

B. Who Did User Corrections?

In order to make a correction, the person must have an active account in CiteSeerX. Most people use their real names when creating their accounts, which allows us to track their

[1]Sometimes, a conference paper can be followed with a journal or magazine article with the same title and authors.

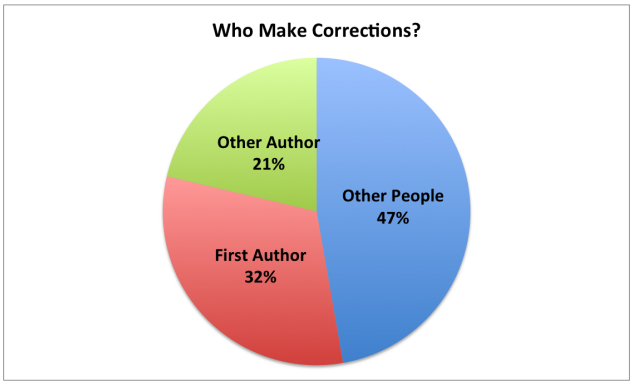

Fig. 3. Categories of correctors.

relationships to the authors of the paper they are correcting. In this section, we attempt to investigate whether the person who corrected the metadata (hereafter the corrector) is one of the authors of this paper.

The random selected sample contains the most accurate data in terms of author names. However, the sample is too small to form a statistically important sample. Therefore, we use all the papers in the user correction history. We match the corrector's last name against the last names of all authors extracted by CiteSeerX. Here, we assume that the correctors registered with their real last names. We also assume that the authors of these papers in the CiteSeerX database are accurate and complete. We admit that it is difficult to directly test and verify these two assumptions.

The correction history we studied covers a time frame between October 2008 and September 2014. The set contains 297,355 user correction records. Before doing statistical analysis, we filter out correctioned caused by two "outlier correctors". One of them is our former software engineer, who corrected over 1800 times. The other one is a user who made about 285,000 corrections. It is interesting to probe the motivation of this massive amount of corrections, but this is beyond the scope of this paper. After excluding these "outlier correctors", we end up with a sample of 10,561 random user corrections. We then match the corrector's last name with the last names of paper authors. The result, shown in Figure 3, indicates that more than half of the corrections were performed by authors of papers that were corrected. This fraction can be just a lower limit, because in many of the "non-author" correctors, the author list was written as "John Smith et al.", so only the first author was parsed out and used for matching. As a result, the actual fraction of "author" correctors is higher than 55%.

Figure 3 has at least two implications. First, it is important to get a *complete* list of authors since both the first author and the other authors seems to care about the quality of the metadata. The author strings ending with "et al." narrow the population of *potential* correctors, thus reducing the possibility of being corrected. Second, it is reasonable to allow users to correct metadata of *any* paper since a significant fraction of correctors do not only correct their own papers, but also papers written by other people.

IV. Assessing User Corrections

In the previous section we attempted to characterize the types of corrections made and the users who make the corrections. In this section, we analyze the quality of the collections. We consider the random sample of papers described in the previous section for which we have manually captured the titles and authors and which forms the ground truth for metadata. We evaluate the quality of the automatic extraction of metadata for a subset of these records; evaluate the quality of user corrected metadata; and assess the extent to which the manually corrected metadata improved on the automatically extracted metadata.

A. Methodology

Corrections to CiteSeerX metadata can come from many sources, such as system corrections (as new information becomes available), inference-based corrections or user corrections via the interface described in Section II. For this reason, a check was made to each of the papers in the ground truth collection of 1000 papers to see if it had been corrected by a user and all other sources of corrections were discarded. Furthermore, when more than 1 correction has been made to a paper, we consider only the latest correction for the sake of this evaluation. Through this filter method, a total of 73 user corrections were evaluated.

A user study was conducted for the 73 papers to evaluate the quality of the automatically extracted and user corrected metadata. For each record, the users were provided with the ground truth metadata and asked to answer the following questions about the automatically extracted and user corrected metadata for both the title and author fields:

1) Is the automatically extracted metadata correct?
2) Is the manually corrected metadata correct?
3) Does the manually corrected metadata improve on the automatically extracted metadata?

Three users completed the survey and the results reported are based on the majority answer for each question.

B. Results

Figure 4 shows the percentage of *yes* responses for each of the questions stated above. The y-axis is the percentage of papers for which the majority response was *yes* and the x-axis represents the question.

As can be seen from Figure 4, users marked the automatically extracted titles and authors as being correct about 50% of the time. For titles, this number is similar to that found in [5]; however, for authors, the accuracy of automatically extracted authors was found to be higher in [5]. One possible reason for this is that in [5] the authors make use of a sliding scale for judging metadata quality whereas for our evaluation it is a binary judgment.

Figure 4 also shows that the quality of the user supplied corrections is very high. For authors it is 100% correct and for titles it is 94.55%. This shows that user corrections are a reliable source for high quality metadata.

Lastly, Figure 4 shows that user corrections lead to an improvement in the titles in about 67% of cases and an

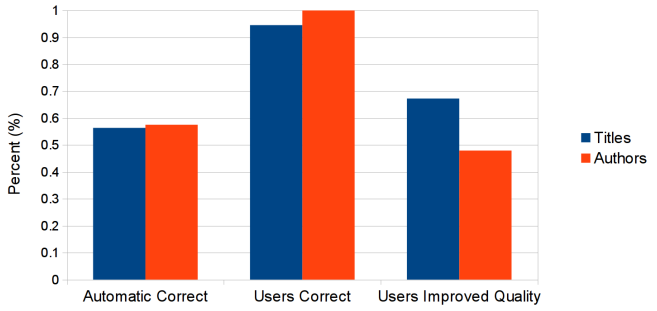

Fig. 4. Results of metadata quality evaluation

improvement in authors in about 48% of cases. The reason that the improvement is so much higher for titles is that, in many cases, the automatic extractor will mistakenly include additional words in the automatically extracted titles. For instance, in many cases the word *Abstract* was incorrectly tagged as belonging to the title and users correct many of these cases.

It is clear from this analysis that user corrections are beneficial and lead to higher quality metadata. This higher quality metadata is important for several reasons. First, document linking and clustering is improved since it is based on metadata. Secondly, the titles and authors are commonly used as search queries and improved metadata leads to a higher probability of returning relevant results in response to a user query. Lastly, given the high quality of user supplied metadata, it essentially provides a ground truth that can be used for training the classifiers that are used for metadata extraction. In this sense, it provides a feedback mechanism that can be used to continuously improve the quality of metadata in CiteSeerX.

V. Impact of User Correction

In this section we aim to study the impact of user corrections on the papers receiving corrections, and on the overall system. There are multiple ways through which the impact of user corrections can be studied and analyzed. Since the goal of user correction is to enhance the quality of the metadata, it should therefore make the document whose metadata was corrected more discovarable by end users. One way through which discoverability might be estimated is the number of downloads a given paper receives. We conjecture that the number of downloads a paper recieves should increase after a user correction as the new metadata has increased discoverability of the paper.

It turns out that this conjecture can not be answered simply for many reasons. First, the majority of the downloads (more than 70%) are referred from Google, Google Scholar, and Bing. These search engines use their own metadata extraction to identify titles and authors of academic papers. Thus, it can be the case that users discovered the paper with correct metadata from search engines and the manual correction did not affect the discovarability of the paper. Second, the average download behavior tends to change from month to month and, more importantly, year to year. Downloads in July are usually much less than March for example as many researchers are on vacation in July.

Measure	Year Before	Year After
Average number of downloads	21.07	16.79

TABLE III. The average number of downloads in the year before and the year after a paper has its metadata corrected in 2011

Year	Average # Downloads Per Paper
2010	42.03
2011	58.39
2012	21.40

TABLE IV. The average number of downloads per paper in each year

Nevertheless, using the historical access logs of CiteSeerX we proceed to analyze the download patterns in the previous and following 365 days for papers that were corrected in 2011. We identified 41,000 papers whose metadata was corrected in 2011, and used historical access logs of the three years (2010, 2011 and 2012) to identify download requests. Logs were pre-processed to remove all requests from search engine bots. Furthrmore, all download requests that were referred by Google and Bing were filtered out to mitigate the effect of external metadata. We report the average number of downlaods in the year before and the year after the manual correction in Table III.

The initial take away from Table III is that the average number of downloads decreased after the metadata was corrected. This is counter-intuitive and is not expected. However, to put this number into persepctive we need to compare it against the average number of downloads papers in CiteSeerX receive in general. Table IV shows the average number of downloads per paper in the years 2010, 2011, and 2012. It shows that in 2012, there was decline in the average number of downlaods by more than 50%. However, manaully corrected papers had their average number of downloads decline by 20% only, which is much lower than the average.

Although these numbers can not be used in a statistical hypothesis tests as is, they however give an indication about the download behavior before and after a user correction. It is evident that there is a decrease in the number of downloads after the correction, however, there is an overall decrease in the number of downloads during that period and the papers whose metadata was corrected had a smaller decrease in the number of downloads compared to papers in general. From this we might conclude that metadata correction is increasing the discoverability of papers in the digital library.

VI. Conclusion

This paper described the user feedback within a digital library in the form of user corrections and showed how CiteSeerX, an existing popular digital library, utilizes that to enhance its metadata. We first studied the types of corrections made by users. The quality of the corrections was then analyzed in a user study where it was found that manual corrections indeed enhance the quality of the metadata significantly. In the end, we compared the download behavior before and after the corrections were made and showed that, on average, this led to an improvement in the discoverability of documents.

ACKNOWLEDGMENTS

We acknowledge partially support from National Science Foundation.

REFERENCES

[1] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox, "Automatic document metadata extraction using support vector machines," ser. JCDL '03, 2003, pp. 37–48. [Online]. Available: http://dl.acm.org/citation.cfm?id=827140.827146

[2] I. G. Councill, C. L. Giles, and M.-Y. Kan, "Parscit: an open-source crf reference string parsing package," ser. LREC '08, 2008.

[3] C. Caragea, J. Wu, K. Williams, S. D. G., M. Khabsa, and C. L. Giles, "Automatic Identification of Research Articles from Crawled Documents," ser. WSDM-WSCBD '14, 2014.

[4] C. Caragea, J. Wu, A. Ciobanu, K. Williams, J. Fernandez-Ramirez, H.-H. Chen, Z. Wu, and C. L. Giles, "Citeseerx: A scholarly big dataset," ser. ECIR '14, 2014, pp. 311–322.

[5] M. Lipinski, K. Yao, C. Breitinger, J. Beel, and B. Gipp, "Evaluation of header metadata extraction approaches and tools for scientific PDF documents," *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries - JCDL '13*, pp. 385–386, 2013.