
Proceedings of the Sixth International Conference on Conceptions of Library and Information Science—"Featuring the Future"

Resolvability of references in users' personal collections

[Nishikant Kapoor](#)¹, [John T Butler](#)², [Gary C Fouty](#)², [James A Stemper](#)², [Joseph A Konstan](#)¹

¹GroupLens Research, Department of Computer Science and Engineering, 200 Union Street SE, University of Minnesota, Minneapolis, MN 55455, USA

²University Libraries, 117 Pleasant St SE, University of Minnesota, Minneapolis, MN 55455, USA

Abstract

Introduction. Digital library users collect, enhance and manage their online reference collections to facilitate their research tasks. These personal collections, therefore, are likely to reflect users' interests, and are representative of their profile. Understanding these collections offers great opportunities for developing personalized digital library services, such as reference recommender systems.

Method. We recruited subjects by individual e-mails to the users of RefWorks - a web-based personal reference management tool installed for use at the University of Minnesota. To participate, subjects needed to give their consent and share their references with us. 96 subjects participated, majority (65) of who were graduate students, resulting into 30,336 references. Based on the type of the reference, these were stratified into one of the three valid identifying IDs - DOI, ISBN, or URL. Multiple reference resolvers (CrossRef, WorldCat) were used to enhance the overall resolvability of these collections.

Analysis. Descriptive statistics and simple graphics analysis were used to describe the dataset.

Results. Over 90% of the total references in users' personal collections could possibly have a valid ID (DOI, ISBN, URL), and therefore, are potentially resolvable. However, only about 17% of the references in these collections had a valid ID, and fewer than 11% actually resolved successfully. Using a combination of reference resolvers, the total resolvability of the references in these collections was enhanced from under 11% to over 41%.

Conclusions. Users' personal reference collections have a tremendous potential of building, supporting, and enhancing personalized digital library services, such as reference recommender systems.

Introduction

Digital libraries continue to grow enormously, and rapidly as more and more people access the networked digital environment. Additionally, people gather, build and manage their very own personal collections, and have come to expect those to be integrated with online digital libraries. While this ongoing growth offers an immense power of information to its users, it poses several challenges as well. Finding useful information, effectively and efficiently, continues to be the primary challenge.

Recommender Systems ([McNee et al. 2002](#), [Torres et al. 2004](#), [McNee et al. 2006](#)) offer a viable solution, but rely heavily on personalization ([Rashid et al. 2002](#)) i.e. they need to know the user before they can help her find something useful. Recommender Systems compare users with similar interests, and predict new items of interest to the user, given some information about her profile. These systems build users' profile by collecting information using a combination of explicit and implicit methods, and draw on the similarities (or dissimilarities) of these profiles to generate recommendations for users ([Maltz et al. 1995](#)).

Users' personal reference collections are an implicit means to learn about their interests, and are a close representative of their profiles. These collections, therefore, offer a great potential for systems such as a reference recommender system, to offer personalized tools and services in digital libraries. However, recommender systems first need to be able to harvest these collections effectively before they can use these to build users' profile. And, that can be made possible if these references had a valid ID that uniquely identified them. Unique identification of identical references in different users' collections is the key to building similarities between the users. It opens the door to matching references between collections, and to obtaining additional metadata from which to generate recommendations. In order to realize the potential of references in users' personal collections, it is important that we understand their nature and assess them in terms of their resolvability.

Most recommenders in digital libraries have focused on mining the implicitly rated references in the reference section of a paper (i.e., public collections of rated references). The cited references in a technical paper are indicative of their support for the paper, and are considered by the recommenders as implicit 'ratings' for the paper. Quickstep and Foxtrot recommender systems explored recommending on-line research papers to academic researchers ([Middleton et al. 2004](#)). TechLens used references in the published paper to build correlations between papers ([McNee et al. 2002](#)).

Our work focuses on references in users' personal collections - to understand how these collections can be uniquely identified, and to assess their potential for developing digital library services using recommenders. We are not the first to explore the quality and nature of references. Numerous reference analysis tools and methods have been developed to support bibliometric research, including citation counts and frequencies, impact factors ([Garfield 1955](#)), clustering ([Carpenter et al. 1973](#)), bibliographic coupling ([Kessler 1963](#)), and co-citation analysis ([Small 1973](#)). However, all these analyses have been performed on published citations, and hinge on the degree of citation accuracy and completeness.

In this study, we explore resolvability of references in users' personal collections to unique identifiers, and to their online sources. We analyzed the contents of 96 reference collections from university students, faculty and staff, maintained in the RefWorks (<http://www.refworks.com/>) references management system, to address two primary research questions:

- How resolvable are references in users' personal collections? And, how many of those do actually resolve to their online sources?
- How can we enhance resolvability of references in these collections?

Resolvability

A reference is resolvable if it has (a) a valid unique ID that leads to its online source, or (b) enough information such as title, author, etc., that could be used to resolve it to its online source. An online source of a reference establishes definitive online presence of the reference, either by leading to its full-text, or to its description. For example, a valid

unique ID for a reference to a Journal article is a DOI (Digital Object Identifier), which leads to its full-text, and a valid unique ID for a reference to a book is its ISBN (International Standard Book Number) number, which leads to its description (in some cases, to its full-text). A reference to an online entity (URL - Universal Resource Locator) may not be its unique identifier, but it reaffirms its existence. A valid DOI, ISBN or a URL of a reference is expected to lead us to its online source, provided the online source is available. If the online source is not available (it might not be in digital form yet), we conclude that the ID for the reference is not valid. Our dataset consisted of references to various types such as references to Journal articles, whole books, book chapters, web pages, newspaper articles, maps, reports, etc. Of all the different reference types, the ones that could possibly have a unique ID are

- DOI based i.e. references to articles in Journals, Journal-Electronic and Conference Proceedings - 25,267 references (83.29%)
- ISBN based i.e. references to book Whole, Book Edited, Monograph - 2,216 references (7.30%)
- URL based i.e. references with URL (includes references that have DOI and/or ISBN) - 622 references - (2.05%)

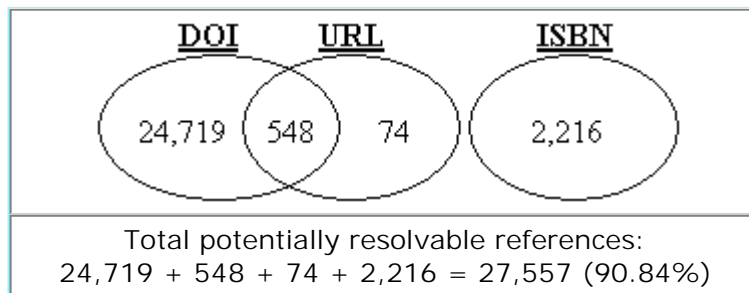


Figure 1: Total potentially resolvable references

Figure 1 shows the total number of potentially resolvable references, i.e., references that can be represented using a valid unique ID. Since DOI and ISBN numbers are unique IDs, combining these (minus the ones that are common in both) gives the total number of references that could possibly have a unique ID. In addition, there are 74 references that contain a URL. Assuming that all these references have a valid ID and are available online, we get the total number of references that are potentially resolvable as 27,557 (90.84%).

Reference Types	DOI	ISBN	URL
Journal	3,765		545
Conference Proceedings			3
Book Whole		777	
Book Edited		15	
Monograph		1	
Report			16
Generic			36
Web Page			14
Newspaper Article			5
Book Section			1
Map			1
Hearing			1
Totals	3,765	793	622

Table 1: Distribution of references with IDs

Table 1 shows the distribution of references that in actual have at least one of the three possible IDs. For example,

number of references having an ISBN ID is the sum of references to [Book Whole (777) + Book Edited (15) + Monograph (1)], which totals to 793. There are other reference types that have ISBN numbers, but they are not included in ISBN count (i.e. towards references that are resolved via ISBN query) because those references are defined under reference type other than Book Whole, Book Edited, and Monograph. Similarly, there are 3,765 references having DOIs. However, all references having a URL, regardless of their reference type, are counted towards the references that can be resolved via URL query, because a URL in any reference can lead us to its online source, in which case it would be considered resolvable.

Not all the IDs that were available in the reference collections resolved to their respective online sources. Reasons varied from reference not available online to mal-formatted IDs to actually incorrect IDs. Figure 2 shows the original resolvability of users' personal reference collections. Out of the total of 30,336 references, there were 5,180 references that had an ID, out of which 3,247 actually resolved, giving a total original resolvability of 10.70%.

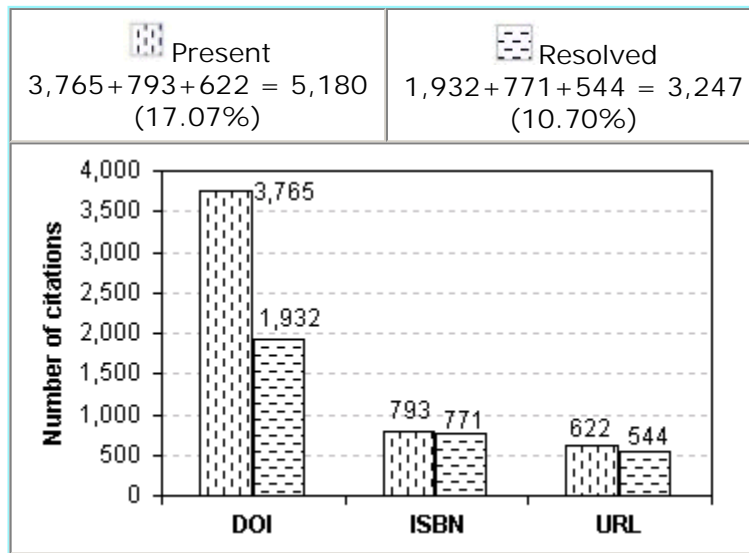


Figure 2: Original resolvability

Resolving references

We used a combination of reference resolving queries to assess the resolvability of references in our dataset.

- **DOI Query Interface** - For all the references that were to articles in Journal, in Conference Proceedings, and in Journal-Electronic, we created a query using the available information in the reference, and submitted it to CrossRef.org for fetching its DOI. The only required values being a journal title and either author or first page to help identify the article. The remaining fields are optional, but recommended.
- **OpenURL Query Interface** - Similar to DOI Query Interface, CrossRef provides an OpenURL interface that also fetches the DOIs. The key information for these queries is author last name, journal title, publication year, volume, issue and date. We used this query to augment the resolvability in the DOI Query.
- **ISBN Query Interface** - WorldCat Libraries (<http://www.worldcat.org/>) provides an interface for querying a specific title, based on ISSN or ISBN. Since ISBN uniquely identifies a reference, whereas an ISBN is a unique identifier for a collection, we limited this query interface to references of only the types that could possibly have an ISBN number i.e. to references to Book Whole, Book Edited, and to Monograph. A successful query returned the ISBN ID for the reference.
- **URL Validation** - References having a URL were validated against their online sources i.e. if the URL existed, the reference was concluded to have a valid ID.



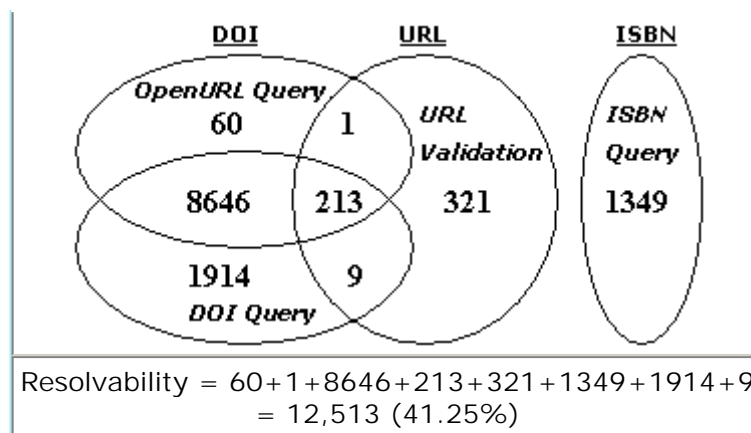


Figure 3: Resolvability computation

We executed four different queries to assess the resolvability of our dataset. Recall that when a query fetched a valid unique ID for a reference, the reference was concluded to be resolved. Figure 3 shows the breakdown of the four queries, and the contribution of each towards the cumulative resolvability. The cumulative resolvability is expressed in terms of the cumulative resolvability of the three valid IDs i.e. DOI, ISBN and URL. ISBN query resolved 1,349 references with ISBN IDs, whereas URL validation resolved 544 references. However, out of the 544 references with URL, there were 223 references that had a valid DOI as well. These DOIs were fetched using one of the two queries - OpenURL query, or DOI query. Together, the two queries resolved 11,164 references with valid DOI IDs, with an overlap of 8859 references between them. ISBN query did not have any overlap with any other query.

Almost all but 61 DOIs that were resolved by OpenURL query were also resolved by the DOI query. Additionally, DOI query resolved another 1,923 valid unique DOIs that OpenURL query could not. These numbers should be interpreted mindful of the fact that all these resolved references include the references that resolved in the original dataset as well i.e. 3,247 (figure 2). The total resolvability from figure 3 can thus be computed by summing up all the individual components, which comes out to be 12,513 references, or approximately 41.25% of the total references.

Discussion

Since a large number of references with unique IDs had DOIs (figure 1), we wanted to ensure their validity i.e. if the reference type had been correctly assigned to the references. We created a random sample of 100 references to articles in Journals and in Conference Proceedings, and manually verified them online. Only 3 of the 100 references could not be traced back to their respective online sources, giving us an assurance that at least 97% of the total references with DOIs were indeed references to articles in Journals and in Conference Proceedings.

Our resolvability computation (figure 3) found that over 41% of references in our dataset are reliable and accurate. This is significantly higher than the original resolvability (10.70%, figure 2). We expect the computed resolvability of the references to be even higher, because not all the intellectual content has been digitized yet. So, if there were any references that did not have any DOIs, we assumed them to be not resolvable programmatically. However, as the visual inspection results revealed, the very same references could actually be resolved.

We noticed that some references had ISBN numbers even when they did not belong to any of the reference types - Book Whole, Book Edited, and Monograph. Similar was the case with DOIs. References in such cases were ignored for this study, and were not attempted to resolve. However, we believe that a closer look at these might uncover references that either have an incorrect reference type, or are additional reference types that can possibly have a valid unique ID.

We resolved the references via querying them at some of the authoritative resolvers in the digital libraries arena, i.e., CrossRef and WorldCat Libraries, which provide services, as well as the APIs to accomplish the reference lookup. However, we believe that adding additional services for lookup (for example, Citation Matcher from PubMed) will considerably enhance the resolvability.

The original dataset from RefWorks did not differentiate between references with ISBN numbers, and the ones with ISSN numbers. An ISBN is a unique ten or a thirteen digit ID of an intellectual entity called a book. The ISSN is a unique ID of a much larger package called a journal which may contain thousands of intellectual entities called articles, many, though not all, are now digital. We needed to separate these two types of references to accurately compute resolvability because a single ISSN applies to the entire population of references within a journal, and is not a unique ID of a reference. We used the checksum verification, and the format checking to ensure that we take into account only the references with ISBN numbers, and not the ones with ISSN numbers.

Results and contributions

To understand the nature of users' personal reference collections in terms of their resolvability, and how it could be used to enhance library services such as building a reference recommender system, we examined 30,336 references maintained by 96 users. These users were selected from faculty, researchers, graduates, and undergraduates who used RefWorks reference management web-based tool at the University of Minnesota. The references in users' collections were of many different types, including but not limited to references to Journals, to Books, to Reports, and to Web pages.

To the best of our knowledge, there has not been any study that evaluates resolvability of users' personal references collections. In this study, we establish an in-depth understanding of uniquely identifying such collections, and assessing their quality in terms of their resolvability. Using the composite approach to resolvability, we were able to enhance the original resolvability of these collections from under 11% to over 41%. We should state that this is a complementary reference resolution technique that should not attempt to replace formal reference resolution methods, but rather augment them.

With over 90% of potentially resolvable references, we believe, users' personal collections have a tremendous potential of building, and supporting digital library services. Vastly enhanced resolvability of these collections allows more perspective to the designers of systems like reference recommender systems, and offers them endless opportunities to build tools, and personalized digital library services that are built on top of personal collections.

Acknowledgements

We would like to thank RefWorks for allowing us to use a reference-sharing feature to harvest references from consenting participants. This research is funded by a grant from the University of Minnesota Libraries, and grant IIS-0534939 from the National Science Foundation.

References

- Carpenter, M.P.; Narin, F. 1973. Clustering of scientific journals. *Journal of the American Society for Information Science*, **24**(6), 425-436
- Garfield, E. 1955. Citation indexes for science: A new dimension in documentation through association of ideas. *Science* **122**(July 1955), 108-111, 473-476
- Kessler, M.M. 1963. Bibliographic coupling between scientific papers. *American Documentation*, **14** (Jan. 1963), 10-25
- Maltz, D. and Ehrlich, K. 1995. Pointing the way: active collaborative filtering. In I. R. Katz, R. Mack, L. Marks, M. B. Rosson, and J. Nielsen (Eds.) *Conference on Human Factors in Computing Systems* (pp 202-209). New York, NY: ACM Press/Addison-Wesley Publishing Co.
- McNee, S., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., Konstan, J.A., & Riedl, J. (2002). On the Recommending of Citations for Research Papers. *Proceedings of ACM 2002 Conference on Computer Supported Cooperative Work (CSCW2002)*, 116-125
- McNee, S. M., Kapoor, N., and Konstan, J. A. 2006. [Don't look stupid: avoiding pitfalls when recommending research papers](#). *Proceedings of the 2006 Anniversary Conference on Computer Supported Cooperative Work*, 20, 171-180. Retrieved 3 October, 2007 from <http://doi.acm.org/10.1145/1180875.1180903>

- Middleton, S. E., Shadbolt, N. R., and De Roure, D. C. 2004. [Ontological user profiling in recommender systems](#). *ACM Transactions on Information Systems*, **22**(1) (Jan. 2004), 54-88. Retrieved 3 October, 2007 from <http://doi.acm.org/10.1145/963770.963773>
- Rashid, A. M., Albert, I., Cosley, D., Lam, S. K., McNee, S. M., Konstan, J. A., and Riedl, J. 2002. Getting to know you: learning new user preferences in recommender systems. *Proceedings of the international Conference on intelligent User interfaces*, 7, 127-134.
- Small, H. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, **24**(4), 265-269
- R. Torres, S.M. McNee, M. Abel, J.A. Konstan, and J. Riedl. Enhancing Digital Libraries with TechLens+. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 228-237

How to cite this paper

Kapoor, N., Butler, J.T., Fouty, G.C., Stemper, J.A. & Konstan, J.A. (2007). "Resolvability of references in users' personal collections" *Information Research*, **12**(4) paper colis13. [Available at <http://InformationR.net/ir/12-4/colis/colis13.html>]

Find other papers on this subject

Scholar Search

Google Search

Windows Academic

000009
[Web Counter](#)

© the authors, 2007.
Last updated: 18 August, 2007




[Contents](#) | [Author index](#) | [Subject index](#) | [Search](#) | [Home](#)