More Than Just Black and White: A Case for Grey Literature References in Scientific Paper Information Retrieval Systems

Sesagiri Raamkumar, Aravind; Foo, Schubert; Pang, Natalie

2015

Sesagiri Raamkumar, A., Foo, S., & Pang, N. (2015). More Than Just Black and White: A Case for Grey Literature References in Scientific Paper Information Retrieval Systems. Lecture Notes in Computer Science, 9469, 252-257.

<https://hdl.handle.net/10356/81219>

<https://doi.org/10.1007/978-3-319-27974-9>

© 2015 Springer International Publishing Switzerland. This is the author created version of a work that has been peer reviewed and accepted for publication by 17th International Conference on Asia-Pacific Digital Libraries (ICADL 2015), Lecture Notes in Computer Science, Springer. It incorporates referee's comments but changes resulting from the publishing process, such as copyediting, structural formatting, may not be reflected in this document. The published version is available at:
[<http://dx.doi.org/10.1007/978-3-319-27974-9>].

Downloaded on 26 Apr 2023 12:33:19 SGT

More Than Just Black and White: A Case for Grey Literature References in Scientific Paper Information Retrieval Systems

Aravind Sesagiri Raamkumar, Schubert Foo, Natalie Pang

Wee Kim Wee School of Communication and Information,
Nanyang Technological University, Singapore
{aravind002, sfoo, nlspang}@ntu.edu.sg

Abstract. In this paper, we start by analyzing the presence of grey literature (GL) references in the bibliographies of different article-types using an extract of 122,406 articles from the ACM Digital Library. GL articles accounted for about 16% of the overall references with highest presence in proceedings (17.61%). Boosting techniques for promoting GL references are proposed for IR-based systems for two specific scenarios. A simple IR experiment was conducted with 103,739 articles to validate the proposed boosting techniques with 10 research topics. Results show that GL references were consistently pushed to the top of the search results along with an increased visibility in top 20 results.

Keywords: grey literature, citation analysis, boosting techniques, scientific paper information retrieval, digital libraries

1 Introduction

Scientific papers are the key information resources for literature review and information seeking purposes of researchers. The scientific nature of these papers is asserted through the rigorous review process enforced by commercial publication houses. The findings of studies in such papers are accepted as part of the body of knowledge for corresponding research topics. Grey Literature (GL) materials are alternate resources produced by researchers, practitioners and the common public. These resources are not part of the traditional publishing lifecycles. Grey Literature is defined as “*information produced on all levels of government, academia, business and industry in electronic and print formats not controlled by commercial publishing*” [1]. Websites, technical reports and dissertations are considered as GL across all disciplines. Conference proceedings are considered as GL in medical, social sciences and humanities disciplines while computer science and engineering disciplines consider them as scientific materials. In this paper, we regard proceedings as scientific papers while the article-types websites, technical reports, dissertations, patent and unpublished are considered as GL materials.

Even though, there has been debate on whether GL materials should be included in systematic literature reviews and meta-analyses [2], inclusion of GL in scientific papers has been largely left to authors' discretion. There have been efforts in the past to accumulate and index GL materials at both institutional and national levels for facilitating easier access [3]. There have also been arguments made for utilizing open access repositories in indexing both GL and non-GL articles [4]. Academic databases and search systems that provide an integrated access to scientific and GL papers are few since the scope of collecting and indexing papers is immense. In the systems where both types of papers are available, lesser preference is given to GL materials while ranking the search results, thereby reducing their visibility to searchers. Nevertheless, past studies have shown sustained referencing of GL materials in scientific papers over the years [5].

Research interest in GL has ranged from recognition of the availability of GL to development of GL initiatives such as conferences, journals and repositories [2]. The usability of GL materials has been acknowledged in the context of conducting literature reviews [6]. However, the inclusion of GL articles in meta-analysis has been subject to debate due to the difficulty in assessing the methodological quality of the studies [7]. Yasin et al. [8] reported the presence of GL references in about 9.22% of 183 Systematic Literature Reviews (SLR) articles with conference proceedings and technical report as the most cited GL types (68%). In a citation analysis study [5] conducted on a sample set of articles in Google Scholar, data indicated that around 34.3% of the retrieved articles were of the GL type with technical reports (26%) being the most used type. These studies indicate the notable presence of GL materials in scientific papers.

There are two main objectives in the work proposed in the paper. The first objective is to analyze the referencing extent of GL materials in an extract of papers from ACM Digital Library (ACM DL). The analysis aims to examine the extent of GL cited by authors, and hence its importance if so, to warrant the motivation to push GL materials higher in the results list in specific scenarios. The second objective is to propose boosting techniques for pushing GL materials in Information Retrieval (IR) system settings due to their applicability in digital libraries. Digital libraries which are repositories of various resource types can benefit from having such functionality especially if such GL resources are found to be particularly relevant for specific domains and contexts. Our scope was restricted to the GL materials that are cited as bibliographic references in the articles of ACM DL extract used in the current study. We conducted a simple experiment to validate the IR boosting technique with 10 research topics selected from the ACM DL. Our experiment showed that the boosting technique helped in pushing GL references to the top 20 search results.

2 Analysis of Bibliographic References in ACM DL Extract

As a part of ongoing studies towards building a task-based scientific paper recommender system, we have been using an extract of the ACM Digital Library (ACM DL). The extract contains full text and metadata of articles from the ACM DL pub-

lished between the years 1951 and 2010. A sample set was built from the extract, containing the articles that have keywords, abstract and references entirely present in the extract. The total count of valid articles in the sample set is 122,406 with 103,739 proceedings articles and 18,667 periodicals articles. The bibliography parsing service AnyStyle was used for parsing the bibliographies. AnyStyle assigns the references type based on the BibTeX entry types. A total of 2,320,345 references were parsed and the required fields were extracted and stored in a MySQL database.

Based on this information, the percentages of GL and non-GL references by different article types are shown in Figure 1. The GL Percentage (GLP) on the whole for proceedings and periodicals was 17.61% and 14.48%, indicating the nature of the publications. In ACM DL, journals, magazines and transactions constitute periodicals. Under periodicals, articles are mainly classified as research articles while some journals also allow review/survey articles. Survey articles have the lowest GLP (12.86%). Contrastingly, referencing behavior in conferences, symposiums and workshops is observed to be comparatively less stringent. There are different article-types under proceedings. The types demos (18.71%) and tutorials (17.98%) have the highest GLP as the references in articles of such types are of miscellaneous nature.

The current examination is limited to quantifying the referencing of GL articles in the extract. In-depth qualitative studies need to be conducted to identify the citation motivations of authors based on the citation contexts of GL references in articles. As stated earlier, the argument being made in this paper is to better utilize the available data within digital libraries in order to help researchers in finding sufficient and useful articles as a part of their literature review. Since digital libraries parse the full text of articles to extract structured data, the bibliographic references could be utilized for enhancing user search.

Fig. 1. GL vs Non-GL References Percentage in ACM DL Articles

3 Experimentation with GL Boosting Technique

In academic search systems and digital libraries, the article metadata along with full text is indexed. Papers are retrieved based on the matching of search keywords and

the indexed text. The final results are sorted using citation count, search keyword relevance, recency, and other factors. Bibliographic references from articles are also indexed in some of these systems although these items are rarely displayed in the search results. For promoting GL references in IR systems, the technique of document boosting can be adopted. A separate field known as ‘boosting weight’ can be added to each document while the indexing process is performed. The boosting weights can be used during the retrieval process to compute new similarity score on top of the basic similarity score. By using this technique, GL materials that get additional weightage due the boosting rules will have a higher probability of getting a better rank in the search results, thereby increasing their visibility to users.

We briefly discuss the results of an experiment carried out to validate the GL references boosting technique applicable for scientific paper IR systems. For this experiment, the proceedings articles of the ACM DL extract were used. A total of 103,739 articles and corresponding 2,320,345 references were indexed in a Lucene index file. The fields indexed were *article id*, *article title*, *article abstract*, *article-type* and the *boosting weight* (gl_w). For all the proper full-text articles, article-type was set as ‘inproceedings’ and for the references, the type was set based on the reference-type identified using the AnyStyle reference parsing service. Okapi BM25[9] ($k=2$, $f=0.75$) was used for the initial similarity score calculation. The boosting weight rules used are provided as follows: *Rule 1: If the article or the reference is of non-GL type, the boosting weight is 1.0, Rule 2: If the reference is of GL type and its reference count is more than 2, the boosting weight is 1.5, Rule 3: If the reference type is a thesis, the boosting weight is 1.25.* Rules 2 and 3 are the two scenarios used for boosting GL materials in this study. The boosting weights for these rules are samples for this study since the aim is elevate certain GL articles. A separate experiment is required to arrive at ideal weights for boosting rules.

Ten research topics were used as the search keywords for the experiment. These topics are part of the top 50 author specified keywords in the ACM DL. For each research topic, a search was performed and the top 100 results were retrieved based on the BM25 similarity score. For the BM25 similarity score calculation, the title and abstract fields of the articles were used. From the retrieved 100 results, the top 20 results were ranked using four techniques. The proposed GL Boosting Technique (GBT) was benchmarked against two traditional ranking techniques and one combined ranking technique. The three benchmarking ranking techniques were Citation Count Technique (CCT) in which the results were ranked based on article’s citation count, BM25 Similarity Score Technique (BST) in which the ranking was based on the computed BM25 similarity score and thirdly, the Combined Score Technique (CST) where the values of citation count and BM25 similarity score were added to form a combined score. In the proposed GBT, the ranking was based on the values which were computed by multiplying the BM25 similarity score with the boosting weight (gl_w).

In order to perform a proper evaluation, a novel evaluation metric Grey Literature Availability Measure (GLAM) had to be conceptualized that accounted for both the retrieved GL articles count as well as the corresponding ranks. GLAM is based on two base metrics GL Count (GLC) which is the count of GL materials that are re-

trieved in the query and in-query MRR (iMRR), a modified version of Mean Reciprocal Rank [10]. Unlike MRR which is calculated across a set of queries, iMRR is calculated within a single query. First, the reciprocal ranks of GL articles in the query are identified. Second, the sum of the reciprocal ranks is divided by the GL articles count (GLC) to form the iMRR value. The main evaluation metric Grey Literature Availability Measure (GLAM) is calculated by adding up GLC and iMRR for each query. Higher GLAM values indicate higher presence of GL articles in the query results, along with better ranks.

The evaluation was done with two ranked lists (N@10 and N@20) for each research topic. Figures 2 and 3 provide the GLAM values computed using the four techniques for the 10 research topics at N@10 and N@20 respectively. For N@10, GL boosted technique (GBT) has higher GLAM than the benchmarking techniques as the GL count (GLC) is higher for all research topics. For N@20, except for two research topics ‘interaction design’ and ‘wireless networks’ where CCT technique has a higher GLAM, the GBT technique produces the best results for all other research topics. Thus, the experiment shows the GL boosting technique produces the expected results for majority of the input research topics.

Fig.2. Comparison of GLAM Values for the Four Techniques (N@10)

Fig.3. Comparison of GLAM Values for the Four Techniques (N@20)

4 Conclusion and Future Work

In this paper, we have analyzed the bibliographic references of articles from an extract of the ACM Digital Library for ascertaining the level of GL referencing. We found that GL materials were found in nearly 16% of the references at an overall level. GL referencing was higher in proceedings article-types particularly tutorial, demo and workshop papers. The referencing was comparatively lower in periodicals particularly in the literature survey article-type. Two scenarios were identified for boosting

GL materials in the context of digital libraries where bibliographic references are parsed and indexed along with article metadata and full text. Boosting techniques for pushing GL references have been proposed for IR settings. An IR experiment was conducted to validate the effectiveness of the boosting technique using a novel evaluation metric. Future user evaluation techniques can also be carried out to evaluate the relevance of the GL found in the result lists. These proposals have practical implications for implementation in academic digital libraries since current systems don't properly make use of all types of bibliographic references of articles indexed in the corpus. GL articles which satisfy certain pre-set rules such as candidates for a higher presence in the search results. As a part of our current work, the proposed IR boosting techniques will be implemented in a task-based scientific paper recommender system for reading list generation and also for shortlisting papers for inclusion in research manuscripts.

Acknowledgements. This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative and administered by the Interactive Digital Media Programme Office.

References

1. Schöpfel, J., Stock, C., Farace, D.J., Frantzen, J.: Citation analysis and grey literature: Stakeholders in the grey circuit. In: GL6: Sixth International Conference on Grey Literature (2005).
2. Farace, D., Schöpfel, J. eds: Grey literature in library and information studies. Walter de Gruyter (2010).
3. Stock, C., Henrot, N.: From OpenSIGLE to OpenGrey: Changes and Continuity. *Grey J.* 7, 93–97 (2011).
4. Banks, M.A.: Towards a continuum of scholarship: The eventual collapse of the distinction between grey and non-grey literature. *Publ. Res. Q.* 22, 4–11 (2006).
5. Di Cesare, R., Luzi, D., National, I.: The impact of Grey Literature in the web environment: A citation analysis using Google Scholar. *Grey J.* 4, 83–96 (2008).
6. Booth, A., Papaioannou, D., Sutton, A.: Systematic approaches to a successful literature review. Sage (2011).
7. Cook, D.J., Guyatt, G.H., Ryan, G., Clifton, J., Buckingham, L., Willan, A., McIlroy, W., Oxman, A.D.: Should unpublished data be included in meta-analyses?: Current convictions and controversies. *Jama.* 269, 2749–2753 (1993).
8. Yasin, A., Hasnain, M.I.: On the quality of grey literature and its use in information synthesis during systematic literature reviews, (2012).
9. Jones, K.S., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Inf. Process. Manag.* 36, 809–840 (2000).
10. Voorhees, E.M.: The TREC-8 Question Answering Track Report. *TREC.* 99, 77–82 (1999).