Improving Citation Mining

Muhammad Tanvir Afzal
Inst. f. Information Systems and Computer Media, Graz University of Technology, Austria
mafzal@iicm.edu

Hermann Maurer
Inst. f. Information Systems and Computer Media,Graz University of Technology, Austria
hmaurer@iicm.edu

Wolf-Tilo Balke
Inst. f. Informationssysteme Technische Universität Braunschweig, Germany
balke@ifis.cs.tu-bs.de

Narayanan Kulathuramaiyer
Department of Computing & Software Engineering, Universiti Malaysia Sarawak, Malaysia
nara@fit.unimas.my

Abstract

In recent years the number of citations a paper is receiving is seen more and more (maybe too much so) as an important indicator for the quality of a paper, the quality of researchers, the quality of journals, etc. Based on the number of citations a scholar has received over his lifetime or over the last few years various measures have been introduced. The number of citations (often without counting self-citations or citations from "minor" sources, in whatever way this may be defined), or some measurement based on the number of citations (like the h- or the g-factor) are being used to evaluate scholars; the citation index of a journal (again with a variety of parameters) is seen as measuring the impact of the journal, and hence the importance one assigns to publications there, etc. The number of measurements based on citation numbers is steadily increasing, and their definition has become a science in itself. However, they all rest on finding all relevant citations. Thus, "citation mining tools" used for the ISI Web of Knowledge, the Citeseer citation index, Google scholar or software such as the "publishorperish.com" software based on Google scholar, etc., are the critical starting points for all measurement efforts. In this paper we show that the current citation mining techniques do not discover all relevant citations. We propose a technique that increases accuracy substantially and show numeric evaluations for one typical journal. It is clear that in the absence of very reliable citation mining tools all current measurements based on citation counting should be considered with a grain of salt.

1. Introduction

Citation management is of great importance by providing important input for new research that may otherwise not be possible without "standing on the shoulders of giants". Citations allow authors to refer to past research in a formal and highly structured way [1], to systematically construct a citation network that then serves as a means of valuation for published works.

The citation count, which refers to the number of citations a particular paper receives, is used in evaluating bibliometrics such as the quality of a paper, the quality of researchers, the quality of journals, etc. It has been used for knowledge diffusion studies [2], network studies [3] and in finding relationships between documents [4]. Impact factor measurements, as derived from citation counts have been applied in making important decisions such as hiring, tenure decisions, promotions and the award of grants [5]. As such the determination of precise citation counts is of utmost importance.

Citation mining refers to the process of discovering citation counts. This task in itself is not trivial as it involves extensive text analysis to determine the exact intended citation of authors to published works. Owing to the large number of publications, this task involves a great amount of human effort if done manually. Alternatively, an approach for autonomous citation discovery can be applied. This approach, however, tends to be prone to omissions and mistakes [6]. Fully autonomous citation mining as such has to rely on community

978-1-4244-4615-5/09/$25.00 ©2009 IEEE

effort for the verification and regular updating of citation records (e.g. Citeseer [6]).

This paper proposes a novel rule-based autonomous citation mining technique, called Template based Information Extraction using Rule based Learning (TIERL) to address this important task. A two-phase approach is used whereby the system first disambiguates citations based on venues. Subsequently, detailed rule-based mining is performed on a much smaller collection of data within the particular venue. The heuristic approach employed is described in the following sections. We illustrate the benefits of this approach by studying the enhancements of the current state of the art by applying our methods to the dataset of the Journal of Universal Computer Science (J.UCS) [29] as case study.

2. Related Work

ISI citation index is the premier service provided by the ISI Web of Knowledge [7]. It indexes about 9,000 international and regional selected journals and book series. The selection of a journal by ISI depends on the impact factor of the journal and on a number of factors that are listed in [8]. This citation index is further used for the ranking of journals [9]. It is a manually created index making it extremely expensive. Some thoughts and issues on this manual approach are discussed in [10]. In searching for a particular paper's citations, ISI offers different databases such as "Web of Science", "Current Contents Connect", and "ISI Proceedings". One can also select all the databases to be searched for all citations for a given paper.

CiteSeer on the other hand provides an autonomous citation indexing service automating the entire process from crawling to extraction of citations from the Web [11]. Although the primary focus area of CiteSeer is limited to computer and information science, it has nevertheless indexed about 1,077,967 documents and 20,328,278 citations. CiteSeer extracts titles and authors information from a citation entry programmatically. References are used to find the identical match within the collection to ascertain a citation. This service claims that 80% of the titles can be extracted correctly from a number of citations. CiteSeer removes standard words and delimiters such as "-&( )[ ], pp, pages, in press, accepted for publication, vol., volume, no, number et al, isbn, conf, conference, proc., proceeding, international society, transactions, technical reports". Word and phrase matching is subsequently performed on the extracted references (with an error margin of 7.7%) [6].

Google Scholar, an open source multi disciplinary citation indexing service, was established in fall 2004 as a beta release [12]. Its citations are indexed and extracted autonomously and cover a wide range of scientific literature. Google Scholar claims that it covers "peer-reviewed papers, theses, books, abstracts and articles, from academic publishers, professional societies, preprint repositories, universities and other scholarly organizations" [13]. As its search is not restricted to pre-defined journals and conferences, Google Scholar can be applied for the tracking of citations across most open access scholarly documents. One major limitation of Google Scholar is that it considers false positives including citations to press releases, resumes, and even links to bibliographic records for cookbooks [14]. It has gradually improved its algorithm and has been able to overcome previously encountered problems of finding citations backward in time [15]. Its algorithm, however, has not been made public.

Apart from the aforementioned citation indexes, there have been some other systems developed for a local dataset to extract references. For example Day [16] briefly described various systems and introduced a new hierarchical representation framework based on the template mining technique. This survey categorized existing systems into two broad categories "Machine learning" approach and "Rule based" approach. The template mining approach involves a Natural Language Processing (NLP) technique to extract data from text when data exists in recognizable patterns [17]. If a text form matches a template pattern then the data is extracted by using instructions associated with that template. In the current work, we extract references from research papers by employing a template mining approach. As research papers fit into a well defined template, we have used a template-based reference extraction of research papers.

Machine learning approaches discover patterns from a dataset as discussed in [18],[19]. Such approaches as used for CiteSeer [6] take advantage of probabilistic estimation, based on training sets of tagged bibliographic data. Although this technique has a good adaptability, it needs a huge set of labelled sample data for training. This requires a great effort in manually tagging substantial amounts of data.

The rule based approach on the other hand is based on rules defined by an expert in the field. Ding [17] has discussed a template-based mining technique applying pattern matching and pattern recognition in natural language to extract information components. We have augmented our template-based technique by employing heuristic rules to extract the information components from extracted references. Rule-based approaches are straight forward to implement but they are not adaptable and it is often difficult to work with a system with many features. A generalised set of common heuristics has been proposed to overcome this limitation.

3. Problem Statement

Citation mining can be viewed as a three tier process:

1. Reference (citation entries) extraction from documents.
2. Metadata extraction from the citation entry.
3. Linking the citation entry to the cited paper.

Most scholarly works reside in digital libraries as PDF documents. For extracting references, these PDF documents are further converted into plain text. This conversion process may result in errors as shown in the following entry.

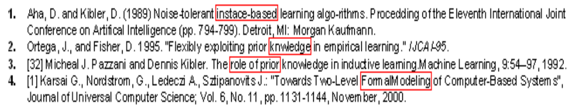
Converted citation entry: 23. P. W. Kutter and A. Pierantonio. Montages: Speci#0Ccations of realistic program-ming languages. Journal of Universal Computer Science, 3#285#29:416#7B442, 1997.

Original citation entry: 23. P. W. Kutter and A. Pierantonio. Montages: Specifications of realistic program-ming languages. Journal of Universal Computer Science, 3(5):416{442, 1997.

The automated extraction of metadata sub field such as title and authors from a citation entry is not at all a trivial issue. Reasons are:

a) All publishers have their own style-guide which needs to be considered while extracting sub fields from a particular reference entry.
b) There are times when authors inadvertently do not follow the style guides properly.

While citing a paper, authors tend to also make mistakes as illustrated in Fig. 1. These mistakes may then lead to improper citation linking.



1. Aha, D. and Kibler, D. (1989) Noise-tolerant instace-based learning algo-rithms. Procedding of the Eleventh International Joint Conference on Artifical Intelligence (pp. 794-799). Detroit, MI: Morgan Kaufmann.
2. Ortega, J., and Fisher, D. 1995. "Flexibly exploiting prior knwledge in empirical learning." *IJCAI-95*.
3. [32] Micheal J. Pazzani and Dennis Kibler. The role of prior knowledge in inductive learning.Machine Learning, 9:54–97, 1992.
4. [1] Karsai G., Nordstrom, G., Ledeczi A., Sztipanovits J.: "Towards Two-Level FormalModeling of Computer-Based Systems", Journal of Universal Computer Science; Vol. 6, No. 11, pp. 1131-1144, November, 2000.

Fig. 1. Badly formatted references by authors

Apart from spelling mistakes made by authors, re-wording of titles also occurs e. g, in the 3rd entry: the word "utility of" was replaced by "role of prior". These types of errors are made mainly because authors simply copy citations from existing references. Mistakes may also arise due to carelessness or negligence.

4. Template based Information Extraction using Rule based Learning

We propose the Template Based Information Extraction using Rule Based Learning (TIERL) technique to increase accuracy of citations obtained. We could make a full text search to link the citations but due to the problems defined in section 3, we have introduced a systematic way of citation linking. TIERL is a layered approach where Template based Information Extraction (TIE) refers to the treatment of a paper as a template from which reference entries are extracted. Rule Based Learning refers to the usage of heuristic rules applied to extract the data and in dealing with uncertainty and the approximate matching of citations.

Research papers are represented as a template structure as shown in Fig. 2. From a given citation string, authors, title and venue information will be used to link citations.

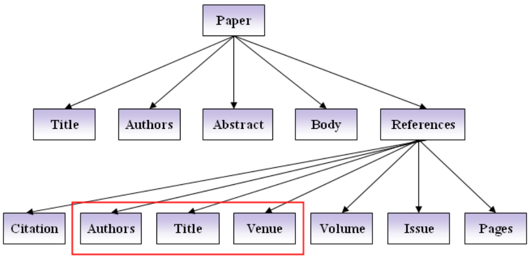


Fig. 2 Template based Information Extraction

The TIERL rules are shown in Figure 3. System works in two steps: 1) extract venue from a citation string. 2) match title from citation string within the papers' title published in the extracted venue.

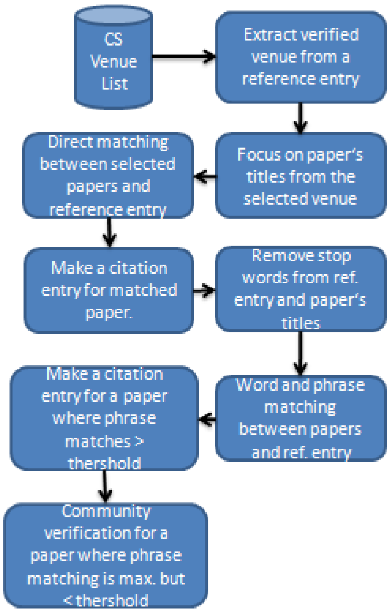

CS Venue List
Extract verified venue from a reference entry
Focus on paper's titles from the selected venue
Direct matching between selected papers and reference entry
Make a citation entry for matched paper.
Remove stop words from ref. entry and paper's titles
Word and phrase matching between papers and ref. entry
Make a citation entry for a paper where phrase matches > thershold
Community verification for a paper where phrase matching is max. but < thershold

Fig. 3 TIERL rules

5. Experimental case study

The Journal of Universal Computer Science (J.UCS) [29] was considered to be a suitable journal to be used for this case study, based on its broad coverage of Computer Science and Information Technology areas. Because of its broad coverage, there is no particular community which is only publishing in J.UCS. Thus, authors from different backgrounds publish their papers. This makes it an interesting dataset for our case study. J.UCS has published more than 1400 peer reviewed papers. J.UCS also provides a large enough document collection to illustrate the workings of the proposed approach.

We applied Template based Information Extraction (TIE) to extract references from PDF versions of J.UCS papers. To perform TIE, we need the full text of all papers in a digital form. The papers are currently available in PDF format and were downloaded automatically from the J.UCS server. Many PDF to text converter tools were tested in terms of accuracy and speed. These include PDFBox [25], Ghostview [26] and PDFTextStream [27]. Based on its performance, PDFBox (open source Java PDF library) was selected for conversion. We then explored the use of layout information of a paper to discover detailed information regarding its structure. For example, a reference starts with the term "references", followed by a delimited list of citation entries. We used three styles of writing a reference entry, which would start from any of the following styles: '[authors year]', '[1]', '1' ". Each citation entry is also expected to have a fixed format. We used intrinsic pattern mining of documents.

13.5% of the papers were editorial columns. Almost 78% out of 86.5% of the papers' references were extracted resulting in over 15 thousands citation entries. 3.5% of the papers have bad references (not complying with any of the templates). 5% of the papers were not compliant with the conversion tool, and were thus not converted correctly into plain text. These 5% papers were not recognized as PDF documents even by the professional converters like INTRAPDF [28]. We propose the use of the postscript and HTML versions of these documents for future experiments.

For the current case study, we focused the citations from J.UCS to J.UCS papers. There were two reasons for the focused dataset (1) J.UCS is indexed by ISI. ISI indexes only a selected number of journals and if we compare the citation out degree for J.UCS then the comparison would not be interesting enough because not all journals and conferences may be indexed by ISI. But if we focus on citations from J.UCS to J.UCS then it is sure that ISI should have all the citations. CiteSeer also claims that it indexes open access journals and tracks when new issues are published. Then the comparison with CiteSeer would be meaningful to know either CiteSeer index all papers of J.UCS if yes then either it is able to find all citations with an error margin of 7.7% as of their claims [6]. (2) The second reason for selecting the dataset was the manual effort required for comparison with the citation indexes because these citation indexes provide free services for community to explore the citations for a focused article most of the time manually. But they (ISI and Google Scholar) do not give their whole data free of charge which could lead to developing an automatic program to compare the results. Consequently, it is a herculean effort to compare each and every paper with ISI, Google Scholar and CiteSeer for checking the citations.

We used the "FLUX-CIM" technique described in [20]. The knowledge base (KB for short) for this was built from all published papers in J.UCS. We extracted the citation components from citation strings where the venue block was represented as J.UCS. In this way we extracted citation components from 133 J.UCS to J.UCS citations. This technique when applied on a generic dataset [20] gives a precision of 95.85% and recall of 96.22% for CS domain. This, however, depends on the complete knowledge base where each and every token represented in the citation string could find its match. In our case, we have focused on the KB built from J.UCS. This is why all tokens found their match in the KB and we were able to extract all the titles and authors of J.UCS citations. The result of our TIERL algorithm on J.UCS dataset gives the results as shown in Table 1. On manual inspection, it was found that the match for 0.75% (only one record) was less than threshold. Subsequently, the list of extracted authors for the maximum matched paper was compared. However, all authors did not find their match and the system was not able to automatically link the citation. This citation was further shown to the user for feedback and on user's response, the citation was linked. Nevertheless, we did not find any 'False Positive'.

Table 1. TIERL algorithm results on J.UCS dataset.

Matching Steps	Accuracy
Direct matching	69.17%
Approximate matching >	24.06%
Author's verification where approximate matching < threshold	6.02%
Overall accuracy	99.25%

After the citation mining for J.UCS articles was completed, we performed comparisons with existing citation indexes. For a comparison with ISI, we selected all of the available databases [7].

To compare with CiteSeer and Google Scholar, we used their standard websites [11] and [12] respectively. We have a total of 133 citations from J.UCS to J.UCS but while comparing, we found 13 more citations which were missed by TIERL. This is 9% of the total. The reasons for these missed citations are that: 1.5% was due to citing a venue wrongly or not including venue information in the citation entry. 7.5% were due to the problem of failure for PDF to text conversion. Now we have total 93 unique J.UCS papers with 146 citations within J.UCS.

6. Experimental results

The measurements selected to compare the citations with other citation indexes were subject to answer three questions. (1) Out of the 146 citations, how many are indexed by each citation index? (2) What was the total missed percentage by each citation index regardless of indexing (the paper or cited by paper). (3) Out of these 146 citations, how many papers and their `cited by` papers were both indexed by each citation index but the citation index has failed to find the citation. The effect of this would be studied by calculating the total number of citations for those papers received within J.UCS. The initial experiment was done during April 2008 and revised in March 2009.

6.1 Indexed papers

The numbers of papers indexed on different citation indexes are listed here. ISI indexes 38% of the papers, CiteSeer indexes about 53% of the papers while Google Scholar indexes 100%. If these citation indexes do not recognize these J.UCS papers then how they can include them for finding citations. The comparison is shown in Fig. 4.

6.2 Overall missed citations

Different citation indexes were compared with the focused citations dataset. The figures represent the percentage of the data missed by citation indexes. These are the overall missed percentages regardless whether the paper is indexed or not. The percentage of missed citations was surprisingly high for the major citation indexes like ISI, Google Scholar and CiteSeer as can be seen in Fig. 5.

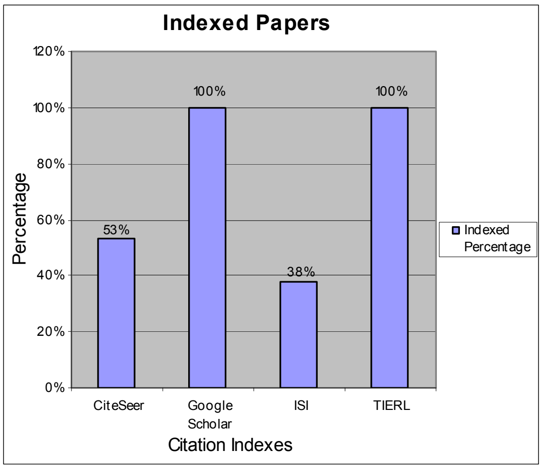


Fig. 4. Indexed papers

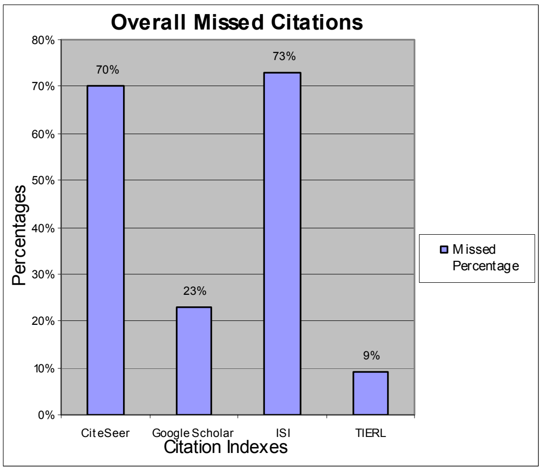


Fig. 5. Missed citations

6.2 Missed citations within the index

Here we focused on missed citations if both the ‘cited’ and ‘cited by’ paper are indexed by the citation index. For example, in the case of ISI, J.UCS was not indexed until 2001. But if we evaluate the missed citations by ISI from 2001, there were a total of 42 articles in J.UCS since 2001 which have been cited by other J.UCS articles. According to our experiments, these 42 articles received 58 citations within J.UCS. All of these ‘cited’ and ‘cited-by’ papers are indexed by ISI. Out of 58 citations, 17 were missed by ISI. This gives an error rate of 29.3%. This is surprisingly high for an established citation index. The comparison with all citation indexes is shown in Table 2.

Table 2. Missed Percentage within the index.

Citation Index	Indexed papers	All Citations within J.UCS	Found by Citation Index	Missed Percentage
ISI	42	58	41	29%
GS	93	146	113	23%
CiteSeer	53	78	44	44%
TIERL	93	146	133	9%

7. Conclusions and future work

As TIERL has focused on venue-specific articles prior to determining citations, it was able to disambiguate papers much more effectively. However, this technique will not work if authors do not specify venues or provide wrong venue information. Our experiments revealed that the error rate in specifying venues was small (1.5% for J.UCS). Our experiments have shown that the proposed approach was able to overcome limitations of current citation mining approaches by providing a layered citation discovery. As the implications of not finding correct citation counts can be serious, this approach should be useful for both autonomous systems such as Citeseer and manual approaches such as ISI. All the experimental and statistical data shown in this paper has been made available at (http://www.jucs.org/jucs_info/downloads/JUCS_Experiments.rar).

8. References

[1] E. Garfield, “Citation Indexes for Science”, Science, 122, pp.108-111 (1955).

[2] A.G.Z. Hu, A.B. Jaffe, “Patent citations and international knowledge flow: the cases of Korea and Taiwan”, International Journal of Industrial Organization. 21, pp. 849-880 (2003).

[3] S.N. Dorogovtsev, J.F.F. Mendes, “Evolution of Networks”, submitted to Advances in Physics.

[4] H.Small, “Co-citation in the scientific literature: A new measure of the relationship between two documents”, Journal of the American Society for Information Science, 24, pp. 265-269 (1973).

[5] Editorial. “The impact factor game. It is time to find a better way to assess the scientific literature”, PLoS Medicine. 3, pp. 707-708 (June 2006).

[6] C.L. Giles, K.D. Bollacker, S. Lawrence, “CiteSeer: An Automatic Citation Indexing System”, proceedings of Third ACM Conference on Digital Libraries, pp.89-98,

[7] ISI Citation Index. http://apps.isiknowledge.com/UA_GeneralSearch.do (accessed 10 March 2009).

[8] ISI journal selection procedure, http://scientific.thomsonreuters.com/free/essays/selectionofmaterial/journalselection/ (accessed 23, May 2008).

[9] E. Garfield. “Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies”, Science 178, pp. 471-479 (1972).

[10] E. Garfield, “Can Citation Indexing be Automated”, symposium proceedings of Statistical Association Methods for Mechanized Documentation, pp. 189-192,

[11] CiteSeer, http://citeseer.ist.psu.edu/ (accessed 08, March 2009).

[12] Google Scholar, www.scholar.google.com (accessed 10, March 2009).

[13] About Google Scholar, http://scholar.google.at/intl/en/scholar/about.html (accessed 23, May 2008).

[14] G. Price, “Google Scholar Documentation and Large PDF Files”, http://blog.searchenginewatch.com/blog/041201-105511 (accessed 23, May 2008).

[15] P. Jacsó, “Reference Reviews”, http://www.gale.cengage.com/reference/peter/200708/SpringerLink.htm (accessed 23, May 2008).

[16] M. Day, R.T. Tsai, C. Sung, C. Hsieh, C. Lee, S. Wu, K. Wu, C. Ong, W.Hsu, “Reference metadata extraction using a hierarchical knowledge representation framework” Decision Support Systems. 43, pp. 152–167 (2007).

[17] Y. Ding, G. Chowdhury, S. Foo, “Template mining for the extraction of citation from digital documents”, Proceedings of the Second Asian Digital Library Conference, Taiwan, pp. 47–62 (1999).

[18] E. Agichtein, V. Ganti, “Mining reference tables for automatic text segmentation”, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 20–29 (2004).

[19] V. Borkar, K. Deshmukh, S. Sarawagi, “Automatic segmentation of text into structured records”, Proceedings of the ACM SIGMOD, pp. 175–186 (2001).

[20] E. Cortez, A.S. da Silva, M.A. Goncalves, F. Mesquita, E.S. de Moura, “FLUX-CIM:Flexible Unsupervised Extraction of Citation Metadata”, proceedings of JCDL, pp. 215-224, (2007).

[22] D.C. Postellon, “Hall and Keynes join Arbor in the citation indices”, Nature, 452, 282 (2008).

[24] Digital Bibliography & Library Project. http://www.informatik.uni-trier.de/~ley/db/.

[25] Apache PDFBox - Java PDF Library, http://www.pdfbox.org/.

[26] Ghostview, a PDF to text convertor. http://pages.cs.wisc.edu/~ghost/index.html.

[27] PDFTextStream 2.0, a PDF to text convertor, http://snowtide.com/.

[28] INTRAPDF - a professional PDF to text convertor, http://www.intrapdf.com/.

[29] Journal of Universal Computer Science, http:www.jucs.org.