

Rule based Autonomous Citation Mining with TIERL

Muhammad Tanvir Afzal¹, Hermann Maurer¹, Wolf-Tilo Balke², Narayanan Kulathuramaiyer³

¹Institute for Information Systems and Computer Media
Graz University of Technology, Austria
{mafzal, hmaurer}@iicm.edu

²Institute for Informationssysteme
Technische Universität Braunschweig, Germany
balke@ifis.cs.tu-bs.de

³Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak, Malaysia
nara@fit.unimas.my



Journal of Digital
Information Management

ABSTRACT: Citations management is an important task in managing digital libraries. Citations provide valuable information e.g., used in evaluating an author's influences or scholarly quality (the impact factor of research journals). But although a reliable and effective autonomous citation management is essential, manual citation management can be extremely costly. Automatic citation mining on the other hand is a non-trivial task mainly due to non-conforming citation styles, spelling errors and the difficulty of reliably extracting text from PDF documents. In this paper we propose a novel rule-based autonomous citation mining technique, to address this important task. We define a set of common heuristics that together allow to improve the state of the art in automatic citation mining. Moreover, by first disambiguating citations based on venues, our technique significantly enhances the correct discovery of citations. Our experiments show that the proposed approach is indeed able to overcome limitations of current leading citation indexes such as ISI Web of Knowledge, Citeseer and Google Scholar.

Categories and Subject Descriptors

H.3.7 [Digital Libraries]; System issues: H.3.3 [Information Search and Retrieval]; Retrieval models: I.1.2 [Algorithms]; Analysis of algorithms: I.2.8 [Problem Solving, Control Methods, and Search]; Heuristic methods

General Terms: Autonomous citation management, Citation disambiguation, Heuristics

Keywords: Citation mining, Digital libraries, Citation index, Information extraction

Received: 17 September 2009; Revised 1 December 2009; Accepted 8 December 2009

1. Introduction

Digital libraries (DL) collect, organize, and provide access to large collections of diverse knowledge resources. A well-managed digital collection of electronic published works and artefacts is of great importance in providing a strong impact for forthcoming 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 (Garfield, 1955), 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 (Hu and Jaffe, 2003), network studies (Dorogovtsev and Mendes, 2002) and in finding relationships between documents (Small, 1973). 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 (PLoS Medicine Editors, 2006). 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 (Giles et al., 1998). Fully autonomous citation mining as such has to rely on community effort for the verification and regular updating of citation records for example: Citeseer (Giles et al., 1998). The remainder of this paper is organized as follows. Section 2 describes the chain method techniques. Section 3 presents the Personal Information Ontology. Section 4 discusses the database design and implementation of the Chain-ontology method.

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 to current state of the art by applying our methods to the dataset of the Journal of Universal Computer Science (J.UCS)¹ as case study.

2. Citation Mining and Discovery

ISI citation index is the premier service provided by the ISI Web of Knowledge². 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

¹<http://www.jucs.org>

²http://apps.isiknowledge.com/UA_GeneralSearch.do

a number of factors³. This citation index is further used for the ranking of journals (Garfield, 1972). It is a manually created index making it extremely expensive. Some thoughts and issues on this manual approach are discussed in (Garfield, 1964). 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.

CiteSeer⁴ on the other hand provides an autonomous citation indexing service automating the entire process from crawling to extraction of citations from the Web (Giles et al., 1998). 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%) (Giles et al., 1998).

Google Scholar⁵, an open source multi disciplinary citation indexing service, was established in fall 2004 as a beta release. 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"⁶. 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 (Price, 2008). It has gradually improved its algorithm and has been able to overcome previously encountered problems of finding citations backward in time (Jacsó, 2008). Its algorithm, however, has not been made known publicly.

Apart from the aforementioned citation indexes, there have been some other systems developed for a local dataset to extract references. For example Day (Day, 2007) 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 (Ding et al., 1999). 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 (Agichtein and Ganti, 2004)(Borkar et al., 2001).

³<http://scientific.thomsonreuters.com/free/essays/selectionofmaterial/journalselection/>

⁴<http://citeseer.ist.psu.edu/>

⁵www.scholar.google.com

⁶<http://scholar.google.at/intl/en/scholar/about.html>

Such approaches as used for CiteSeer (Giles et al., 1998) 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 (Ding et al., 1999) 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 citation entry.
3. Linking 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#0Cations of realistic programming 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, authors from a citation entry is not at all a trivial issue as:

- 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.

Apart from spelling mistakes were made by authors re-wording of titles also occurs e.g in the 3th 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 Rule using based Learning (TIERL)

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. The system architecture for TIERL is shown in Fig. 3. TIERL is a layered

1. Aha, D. and Kibler, D. (1989) Noise-tolerant instance-based learning algorithms. Proceeding of the Eleventh International Joint Conference on Artificial Intelligence (pp. 794-799). Detroit, MI: Morgan Kaufmann.
2. Ortega, J., and Fisher, D. 1995. "Flexibly exploiting prior knowledge 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., Ledecz A., Szilapanovits J.: "Towards Two-Level Formal Modeling of Computer-Based Systems", *Journal of Universal Computer Science*, Vol. 6, No. 11, pp. 1131-1144, November, 2000.

Figure 1. Badly formatted references by authors

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.

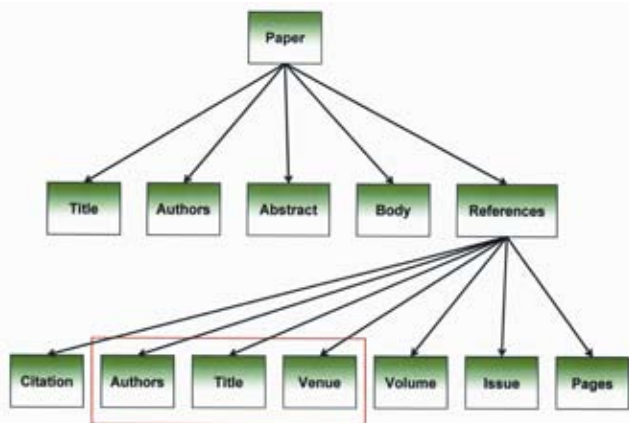


Figure 2. Template based Information Extraction

4.1 TIERL Algorithm

The generic rules to identify a citation entry are depicted below:

Step 1. Extract references from each document using template based information extraction technique.

Step 2. Tokenize each citation string and extract citation components (title, authors, and venue) using FLUX-CIM (Cortez et al., 2007).

For each citation string repeat step 3 to step 8.

Step 3. Disambiguate extracted venue in step 2 from DBLP for the focused citation string using rule based approach as given in section 4.2.

Step 4. Select all papers (their titles and authors) from DBLP which are published in the disambiguated venue in step 3.

Step 5. Apply direct match between the extracted title in step 2 and the titles of the papers selected in step 4.

If (exact match is found) **then** link the citation, focus the next citation entry and go to step 3.

Else if (direct match fails) **then** continue to step 6.

Step 6. Remove stop-words from extracted title in step 2 and focused titles in step 4.

Step 7. Apply approx. matching on the paper's titles returned by step 6.

Approximate matching is calculated as:

(No. of words found in the compared title in a sequence * 100)

(max. number of words of a paper's title in extracted or compared title)

If (match) > threshold **then** link the citation, focus the next citation entry and go to step 3.

Else if (match of more than one records) > threshold **then** select all matched papers as candidates and go to step 8.

Else if (match) < threshold **then** select max. matched paper as candidate and go to step 8.

Step 8. Match author's list of both extracted and candidate papers.

If (all authors matched) **then** link the citation, focus the next citation entry and go to step 3.

Else show to user/community for verification, focus the next citation entry and go to step 3.

Different techniques for the extraction of citation components have been proposed and used in the past. For our experiments, we used technique proposed quite recently (Cortez et al., 2007). This technique gives precision and recall of more than 94% on a generic dataset. This technique uses a knowledge base (KB) which contains pairs of (m_i, o_i) where m_i is metadata field like author, title, and venue and o_i is different occurrences of this field. This KB is used to calculate the field frequency. A citation string is split into blocks on the occurrence of any character other than the characters A, ..., Z, a, ..., z, 0, ..., 9. For each block field frequency is calculated as shown in (1).

$$FF(b, m_i) = \frac{\sum_{t \in T(m_i) \cap T(b)} fitness(t, m_i)}{|T(b)|} \quad (1)$$

Where fitness (t, mi) is defined as follows:

$$fitness(t, m_i) = \frac{f(t, m_i)}{N(t)} \times \frac{f(t, m_i)}{f_{max}(m_i)} \quad (2)$$

The block b is associated with the field which gives the maximum value of FF. More details about the technique can be found in (Cortez et al., 2007).

4.2 Searching Articles by Venue

Venue disambiguation is an important task for citation indexes like Thomson ISI, Google Scholar, and CiteSeer. Accurately

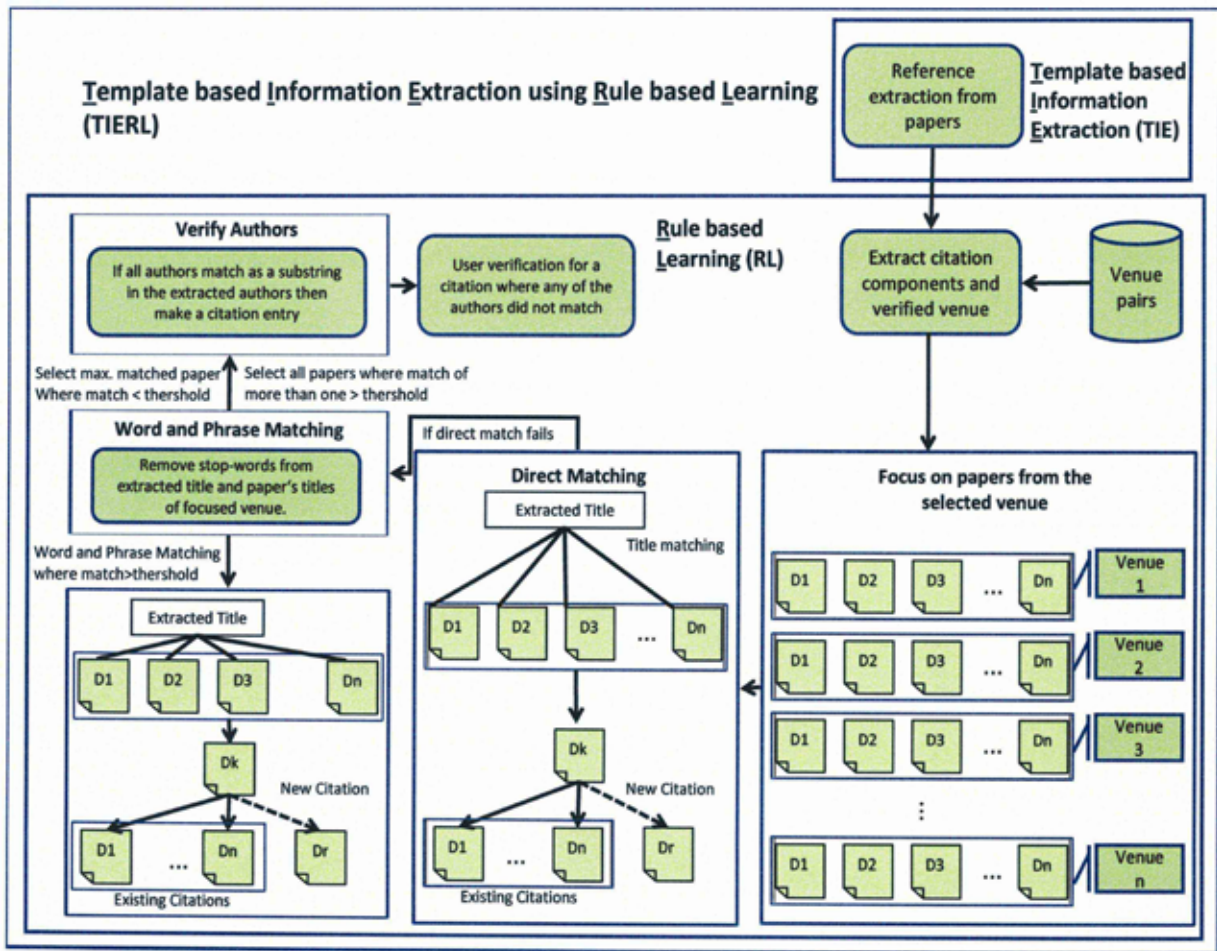


Figure 3. System architecture for TIERL

disambiguated venues are further used for user interfaces and for performing data mining of research literature. We try to cleverly use this venue information to accurately link the “cited” and “cited by” paper. Hall (Hall et al., 2008) have recently suggested an unsupervised method for venue disambiguation. They assume that venues tend to focus on particular research areas and these areas are reflected in the titles of the published papers in a venue. Consequently, they made a venue over title model and disambiguate venues based on Dirichlet process mixture. This model works fine when the venue is focused. They also applied this model to two venues which share the same “acronym” like ISWC (International Semantic Web Conference and International Symposium on Wearable Computing). The venues were accurately disambiguated because the focus of both venues was quite different. But if the venues share the same acronym and the focus of the venue is also the same, then it becomes difficult to disambiguate. These types of venues are listed in Table 1. Also, venues which are not focused are also difficult to disambiguate like the venues “Communications of the ACM”, “IEEE Computer”, “Journal of Universal Computer Science” etc.

In DBLP⁷, the venues are indexed by acronym along with the full venue title. There are more than 5000 unique venues listed in DBLP. A knowledge base was built which comprises of a set of pairs of the form $KB = \{(a_1, f_1), \dots, (a_n, f_n)\}$ in which each a_i is an acronym and f_i is a full name of the venue where a_i and f_i both are pointing to the same venue. A typical example of this pair is a venue pair where a_i is “AAAI” and f_i is “National Conference on Artificial Intelligence”.

ID	Venue acronym	Venue Full Name
1	ICIS	International Conference on Information Systems
		IEEE/ACIS International Conference on Computer and Information Science
2	ICDM	Industrial Conference on Data Mining
		IEEE International Conference on data Mining
3	AIPR	Applied Imagery Pattern Recognition Workshop
		Artificial Intelligence and Pattern Recognition
4	PDCS	Parallel and Distributed Computing Systems (IASTED)
		Parallel and Distributed Computing Systems (ISCA)

Table 1. Venues sharing same acronym with almost same focus

The rules to disambiguate venue is illustrated here:

Step 1. Make venue pairs from DBLP as (a_i, f_i) where a_i (acronym) and f_i (full name) are pointing to the same venue.

Step 2. Remove stopwords from the extracted venue (in step 2 of section 4.1).

⁷<http://www.informatik.uni-trier.de/~ley/db/>

Step 3. Apply direct match between the cleaned venue string from step 2 with the pairs (a_i, f_i) .

If (one match is found) **then** note the corresponding DBLP venue and exit.

Else if (more than one venues in (a_i, f_i) share the same a_i) **then** go to step 4.

Else if $LD(\text{substring}(\text{venue in step 2}), \text{any value in pairs } (a_i, f_i)) = 1$ OR $LD(\text{substring}(\text{any value in pairs } (a_i, f_i)), \text{venue in step 2}) = 1$ (where LD is Levenshtein distance) **then** note the corresponding DBLP venue and exit.

In step 3, treat the words (Journal, International, National, European, Asian, publishers like (IEEE, ACM, WSEAS, Springer, and Elsevier etc) as general words, if they match in a sequence then okay, otherwise they will be ignored while matching.

Else if all patterns of a venue in step 2 match in a sequence as a substring with any pair of (a_i, f_i) **then** note the corresponding DBLP venue and exit.

Step 4. select all papers from the venues which share the same acronym. Disambiguate venue and citation based on matched titles of the paper as described in section 4.1.

The matching of patterns in the extracted venue string means that it should match as a substring in a sequence with any of the venue pair (a_i, f_i) . For example in the case of venue "Journal of Universal Computer Science", all of the following extracted venues will find their match: "Jour. Univers. Comp. sci.", "J. Uni. Comp. Science" and "J. Uni. Computer Sci." etc.

4.3 Dataset

For our initial experiments, we collected texts of citations already hand-clustered into groups referring to the same paper from Cora⁸. For this dataset, we collected the extracted citation components. Our main task was to disambiguate venue and link the citation accurately. Within this dataset, we further focused on the venues listed in DBLP. In this dataset, there were 7 unique venues with different strings mentioning to the same venue. These venues belonged to a focused area where venue over title model may work fine (Hall et al., 2008). This dataset was enhanced with three further venues. One of the venues is "Journal of Universal Computer Science" which belongs to a list of venues that publish papers in broad categories. Two remaining venues belong to the similar focus area and share the same acronym, i.e. ICIS (International Conference on Information Systems, IEEE/ACIS International Conference on Computer and Information Science). In this way, we have approximately 400 citation strings which were used to disambiguate venues and then accurately linked with cited papers.

From the citation strings, we first need to extract the part of the venue string which actually referring to some venue. Stop-words like ('proc', 'proceedings', years, months, 'in', ':', ':', 'published', numeric values, corresponding alphabets for numeric values like eleventh, twelfth etc., 'of', 'the', '(', ')', '{', '}', '[', ']', '-', 'to appear', 'accepted', 'vol', 'issue', 'no', leading and trailing spaces) are removed. By means of this process, we clean the venue string. However, it may still contain some discrepancies along with typographical errors.

⁸<http://www.cs.umass.edu/~mccallum/code-data.html>

In the first run of matching a cleaned venue string with the venue pair (a_i, f_i) , 89% of the venues were matched. The remaining 9% venues were found during step 3 and 4 of section 4.2. 8.5% of the venues found their match in step 3 resulting in $LD(s, t) = 1$ while comparing individual strings.

For a citation entry, we focused only on the paper's titles published in the extracted and verified venue. The results are shown in Table 2. This algorithm achieved an overall accuracy of 99.23%. A small fraction (0.77%) of the citations were unidentifiable as authors wrongly recorded venue information in their citations e.g. The paper "Learning subgoal sequences for planning" was actually published in venue 'IJCAI' but was wrongly cited as being published in 'AAAI'.

Matching Steps	Accuracy
Direct matching	89.05%
Approximate matching > threshold	7.38%
Author's verification where approximate matching < threshold	2.80%
Overall accuracy	99.23%

Table 2. TIERL algorithm results

4.4 Added Value

The extraction of venues and focusing on the papers published in particular venues was significant in linking the citations properly. For example, the same team of authors has written the following two papers in two different venues with a slight change in title.

"Instance-Based Learning Algorithms", published in "Machine Learning".

"Noise-tolerant instance-based learning algorithms", Published in "IJCAI".

Although the authors are the same and title of the paper is also similar, it was successfully disambiguated because of the focused dataset (searching for articles within the articles published in the verified venue). For another citation string, the cited title was "Instance-Based Learning." instead of "Instance-Based Learning Algorithms", published in 'Machine Learning'. Without focusing by venue resulted in 62 unique records from DBLP dataset where this title was matched 100% as a substring. Focusing by venue then significantly helped by reducing the choices to only three candidate papers to select. As a result CiteSeer which selects citation strings of similar lengths from its huge index (Giles et al., 1998) gets too many similar records. This makes it very difficult to disambiguate.

Some times, while making a citation, authors write some additional words or omit or change some words from the title e.g. paper "Instance-Based Learning Algorithms" was cited as "Instance-Based Learning Methods". During an approximate matching process, 67% was matched and then citation was derived based on the matched authors' list. It was noted that there was not a single false positive citation. This is predictable as the same team of authors normally do not submit a paper with almost same title to the same venue.

5. Experimental Case Study

The Journal of Universal Computer Science (J.UCS) 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 articles which makes it an interesting dataset for this case study. J.UCS has published more than 1200 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 PDFBox⁹, Ghostview¹⁰ and PDFTextStream¹¹. 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: '[author's years]', '[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 INTRAPDF¹². 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 is 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 (Giles et al., 1998). (2) 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 (Cortez et al., 2007). 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. The details of venue disambiguation can be found in section 4.2. In this way we extracted citation

components from 133 J.UCS to J.UCS citations. This technique when applied on a generic dataset (Cortez et al., 2007) 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. But of course the accuracy of results for a venue for which one does not have complete bibliographies to compare with the extracted token would not be 100%. The results of our TIERL algorithm (as described in section 4.1) on J.UCS dataset gives the results as shown in Table 3. 3% citations were unidentified. On manual inspection, it was found that 2.25% were referring to papers which were not indexed by DBLP but in fact were published by J.UCS. This is however not the fault of our algorithm. While the match for 0.75% (only one record) was less than threshold. Subsequently, list of extracted authors for the maximum matched paper was compared to DBLP. 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 revised the same pattern that we did not find any 'False Positive'.

Matching Steps	Accuracy
Direct matching	69.17%
Approximate matching > threshold	24.06%
Author's verification where approximate matching < threshold	3.76%
Overall accuracy	97%

Table 3. TIERL algorithm results on J.UCS dataset.

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 ("Web of Science", "Current Contents Connect", and "ISI Proceedings"). To compare with CiteSeer and Google Scholar, we used their standard websites. 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. The reasons for these missed citations by TIERL are explained in section 6.5. So 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%.

⁹<http://www.pdfbox.org/>

¹⁰<http://pages.cs.wisc.edu/~ghost/index.html>

¹¹<http://snowtide.com/>

¹²<http://www.intrapdf.com/>

TIERL indexes 98% because overall 2% J.UCS papers were not indexed by DBLP. 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.

6.3 Missed Citations within the Indexed Papers

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 4 and the missed percentages are shown in Fig. 6.

Citation Index	Indexed papers	All Citations within J.UCS	Found by Citation Index
ISI	42	58	41
GS	93	146	113
CiteSeer	53	78	44
TIERL	91	143	133

Table 4. Found citations within the index by Citation Indexes

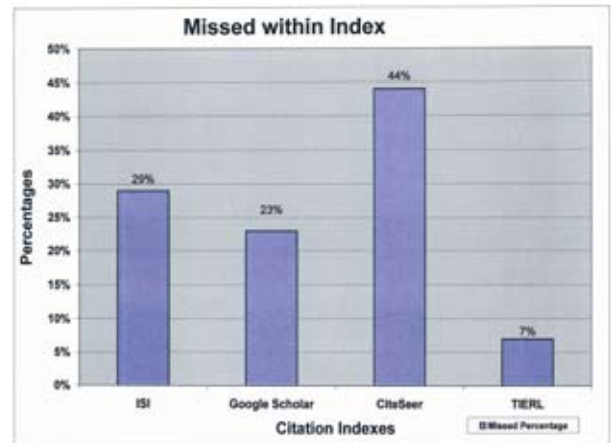


Figure 6. Missed citations within citation index and their overall impact

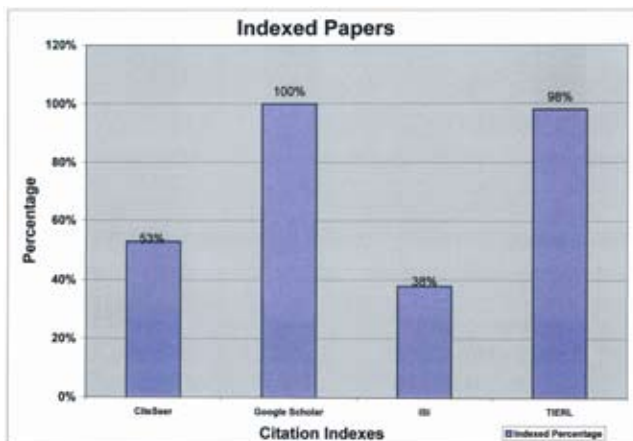


Figure 4. Indexed papers

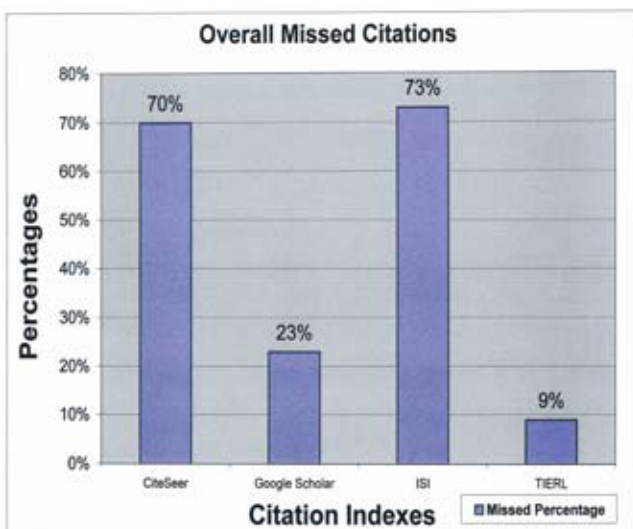


Figure 5. Missed citations

6.4 Misleading Impact Factor

Being an authority in measuring impact factors of journals, Thomson ISI publishes a Journal Citation Report every year. Thomson ISI calculates an impact factor for a particular venue in a given year based on the citations for the papers published in the last two years. For example the impact factor of J.UCS in 2005 would be the number of citations made by the papers in 2005 (which are published in ISI indexed venues) to papers published in J.UCS during the years 2003 and 2004 divided by the total number of papers published in J.UCS during 2003 and 2004. The impact factor of J.UCS in 2005 by ISI is as follows:

Cites in 2005 to articles published in: 2004 = 26
 Cites in 2005 to articles published in: 2003 = 33
 Sum = 59
 Number of articles published in: 2004 = 89
 Number of articles published in: 2003 = 86
 Sum = 175
 Impact factor = $59/175 = 0.337$

But within our small focused dataset of citations from J.UCS to J.UCS articles, it has been observed that there were extra 4 citations in J.UCS papers published in 2005 to J.UCS articles published in 2004. With this small information the actual impact factor of J.UCS for the year 2005 becomes 0.36 instead of 0.337. But it has been shown that the impact of missed J.UCS citations by ISI within their index was 29.3%. And if ISI is missing citations to J.UCS papers by the same ratio for other sources then the impact factor of J.UCS should be 0.48 instead of 0.336 i.e. almost equivalent to the J.UCS impact factor in 2003.

6.5 Missed Citation Snippets

This section first describes the reasons for missed citations from TIERL and then by other systems. As discussed in

section 5 that TIERL had missed 13 citations which is 9% of the total. There are the following reasons for: 1.5% was due to unspecified venue information or citing a venue wrongly. 7.5% were due to bad conversion from PDF to text as discussed in section 5. The reason for this failed conversion was due to PDF files encoding that prevented editing. But Google Scholar was able to find these citations as it had indexed HTML versions of these documents. For future experiments we will consider PS and HTML versions to overcome this limitation.

Typical missed citations by TIERL are shown below:

In the following entry, the authors have specified each component correctly but the venue is cited wrongly. The article was published in Journal of Universal Computer Science but while citing authors have written J. Universal Computations and Systems.

M. Margenstern, K. Morita, A polynomial solution for 3-SAT in the space of cellular automata in the hyperbolic plane, *J. Universal Computations and Systems*, 5-9, (1999), 563-573.

In the following case authors have not provided the venue information and that is why the citation was not found by TIERL.

[Borghoff and Pareschi, 1998] Borghoff, U. M. and Pareschi, R. (1998). *Information Technology for Knowledge Management*. Springer.

If we carefully look at the missed citations by major citation indexes then we will find some interesting patterns. For example, in the following reference entry:

227. K. Kwon. [A Structured Presentation of a Closure-Based Compilation Method for a Scoping Notion in Logic Programming]. *Journal of Universal Computer Science*, 3(4):341-376, 1997.
An extension to logic programming which permits scoping of procedure definitions is described at a high level of abstraction (using ASMs) and refined (in a provably-correct manner) to a lower level, building upon the method developed in [100]. The PhD thesis upon which this paper is based was submitted to Duke University on December 12, 1994, under the title "Towards a Verified Abstract Machine for a Logic Programming Language with a Notion of Scope", number CS 1994-36, pp.189.
228. L. Lamport. A new solution of Dijkstra's concurrent programming problem. *Comm. ACM*, 17(8):453-455, 1974.

The authors have written an explanation after the reference entry 227. Usually, it is not expected that authors would write some explanation within the references. But in this case the reference entry 227 would be considered until the next entry 228 starts although the actual reference entry is only the first three lines. But in this case the 227 reference entry is assumed to comprise 10 lines. When this reference entry would be compared in the citation index, it will not find a match with any reference entry.

Let us consider the following case:

[SN01] R. F. Stärk and S. Nanchen. A Complete Logic for Abstract State Machines. *Journal of Universal Computer Science (JUCS), Abstract State Machines 2001: Theory and Applications*, 2001. (this volume).

The authors have made a mistake while writing the title. The word "complete" was added additionally which means that the citation may not be found.

In the following reference entry, the authors have made two errors while writing a title. "." is replaced with "-". However, this is not a big problem. But the other mistake is crucial: "Computer-supported" is replaced by "Computer-based". Thus it becomes difficult to identify the corrected cited article when the comparison is made within the huge index. Our word and phrase matching algorithm working on a focused subset of the huge index has discovered the correctly cited article.

[3] Maurer, H., Stubenrauch, R., Camhy, D.: Foundations of MIRACLE - Multimedia Information Repository: A Computer-based Language Effort, *J.UCS* 9, 4 (2003), 309-348

In the following reference entry, the title of the paper seems correct but still it did not find a match within the existing citation index. The reasons for this are that after the venue name, there is no volume and issue number. It is written as "This Volume" which did not find its match. But our technique first identified the venue and then checked for the title as a substring in this entry and found it correctly.

[Problem Description, 2000] [The Light Control Case Study: Problem Description]. *Journal of Universal Computer Science, Special Issue on Requirements Engineering (This Volume)*.

The results of citation mining are also questionable as the citation indexes have difficulties in distinguishing individuals precisely. For example, Ann Arbor, Walton Hall and Milton Keynes (the name of cities) were wrongly classified as actively cited authors (Postellon, 2008).

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 case study and 0.8% for generic experiments). These figures have indicated that although authors make many mistakes when citing references, mistakes in writing venue strings are not as significant. 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/online_material.rar

8. Acknowledgements

This contribution is partly funded by the Institute for Information Systems and Computer Media (IICM), Graz University of Technology, Austria and the Higher Education Commission (HEC) of Pakistan.

References

- [1] Agichtein, E., Ganti, V. (2004). Mining reference tables for automatic text segmentation. *In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 20-29.
- [2] Borkar, V., Deshmukh, K., Sarawagi, S. (2001). Automatic segmentation of text into structured records. *In: Proceedings of the ACM SIGMOD Conference*, p. 175-186.
- [3] Cortez, E., da Silva, A. S., Goncalves, M. A., Mesquita, F., de Moura, E. S. (2007). FLUX-CIM: Flexible Unsupervised Extraction of Citation Metadata. *In: Joint Conference on Digital Libraries*. p. 215-224, Vancouver, British Columbia, Canada.
- [4] Ding, Y., Chowdhury, G., Foo, S. (1999). Template mining for the extraction of citation from digital documents. *In: Proceedings of the Second Asian Digital Library Conference*, Taiwan, p. 47-62.
- [5] Day, M., Tsai, R. T., Sung, C., Hsieh, C., Lee, C., Wu, S., Wu, K., Ong, C., Hsu, W. (2007). Reference Metadata Extraction using a Hierarchical Knowledge Representation Framework. *Decision Support Systems*. 43, 152-167.

- [6] Dorogovtsev, S. N., Mendes, J. F. F. (2002). Evolution of Networks, *Advances in Physics*, 51, 1079-1187.
- [7] Garfield, E. (1955). Citation Indexes for Science. *Science*. 122, 108-111.
- [8] Garfield, E. (1964). Can Citation Indexing be Automated. In: Symposium proceedings of Statistical Association Methods for Mechanized Documentation. 189-192, Dec.15.
- [9] Garfield, E. (1972). Citation analysis as a tool in Journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science* 178, 471-479.
- [10] Giles, C. L., Bollacker, K. D., Lawrence, S. (1998). CiteSeer: An Automatic Citation Indexing System. In: Proc. of Third ACM Conference on Digital Libraries, p. 89-98, Pittsburgh, Pennsylvania, United States. Jun. 23-26.
- [11] Hall, R., Sutton, C., McCallum, A. (2008). Unsupervised Deduplication using Cross-field Dependencies. In: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. p. 310-317. Las Vegas, Nevada, USA.
- [12] Hu, A. G. Z., Jaffe, A. B. (2003). Patent citations and international knowledge flow: the cases of Korea and Taiwan. *International Journal of Industrial Organization*. 21, 849-880.
- [13] Jacsó, P. (2008). Reference Reviews, <http://www.gale.cengage.com/reference/peter/200708/SpringerLink.htm> (accessed 23, May).
- [14] PLoS Medicine Editors. (2006). The impact factor game. It is time to find a better way to assess the scientific literature. *PLoS Medicine*. 3, 707-708.
- [15] Postellon, D. C. (2008). Hall and Keynes join Arbor in the citation indices. *Nature*, 452, 282.
- [16] Price, G. (2008). Google Scholar Documentation and Large PDF Files, <http://blog.searchenginewatch.com/blog/041201-105511> (accessed 23, May).
- [17] 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, 265-269.

Authors Biographies



Muhammad Tanvir Afzal studied Computer Science at Graz University of Technology, Austria and was awarded Ph.D. with distinction in 2010. He received his master's degree in Computer Science from Quaid-i-Azam University, Islamabad, Pakistan and secured Gold Medal in 2004. During his Ph.D., he spent one month each in Technical

University Braunschweig and Universiti Malaysia Sarawak for research activities. He worked in software houses, R&D institutes, and universities at various levels. He worked on Context-aware systems for Journal of Universal Computer Science (J. UCS). He authored more than 20 publications in international journals and conferences. He is/was serving as editor/reviewer/session-chair for various reputable international journal and conferences. His research areas include personalized services, Semantic Web and web/text mining.



Hermann Maurer got his Ph.D. from the University of Vienna in 1966. He was professor for Computer Science at the Universities Calgary, Canada; Karlsruhe, Germany; Auckland, New Zealand and for many years at Graz University of Technology. He authored of some 20 books and 650 publications, has graduated some 500 M.Sc. and 60 Ph.D. students, and was co-founder of a number of companies. He has obtained a

number of distinctions (including three honorary doctorates) and was elected chairperson of the Informatics Section of the

Academia Europaea in spring 2009. His main research areas today include digital libraries and on-line encyclopedias.



Wolf-Tilo Balke currently is a full professor at Technische Universität Braunschweig and a director of the L3S Research Center, Hannover, Germany. Before that he was a research fellow at the University of California at Berkeley. His research is in the area of information systems and service provisioning,

including preference-based database retrieval algorithms and ontology-based discovery and selection of Web services. Wolf-Tilo Balke is the recipient of two Emmy-Noether-Grants of the German Research Foundation (DFG) and the Scientific Award of the University Foundation Augsburg. He has received his B.A. and M.Sc. degree in mathematics and a PhD in computer science from University of Augsburg, Germany.



Narayanan Kulathuramaiyer is the Dean and Professor of Computer Science at the Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak (UNIMAS). He received his Ph.D. in Computer Science from Graz University of Technology, Austria. He serves as the Director of the Web Intelligence Consortium (WIC),

Malaysia Research Centre and as an Editor in Chief for the Journal of Universal Computer Science. His research interests include Semantics-Aware Systems, Knowledge Management, E-Learning and Web Science.