Accepted for publication, *IEEE Computer*, 1998. Copyright © IEEE.

Digital Libraries and Autonomous Citation Indexing

Steve Lawrence, C. Lee Giles, Kurt Bollacker
NEC Research Institute, 4 Independence Way, Princeton, NJ 08540
Phone: 609 951 2676 (lawrence) 2642 (giles) Fax: 609 951 2482
{lawrence,giles,kurt}@research.nj.nec.com

Abstract

The World Wide Web is revolutionizing the way that researchers access scientific information. Articles are increasingly being made available on the homepages of authors or institutions, at journal Web sites, or in online archives. However, scientific information on the Web is largely disorganized. This article introduces the creation of digital libraries incorporating *Autonomous Citation Indexing* (ACI). ACI autonomously creates citation indices similar to the *Science Citation Index* ®. An ACI system autonomously locates articles, extracts citations, identifies identical citations that occur in different formats, and identifies the context of citations in the body of articles. ACI can organize the literature and provide most of the advantages of traditional citation indices, such as literature search using citation links, and the evaluation of articles based on citation statistics. Furthermore, ACI can provide significant advantages over traditional citation indices. No manual effort is required for indexing, which should result in a reduction in cost and an increase in the availability of citation indices. An ACI system can also provide more comprehensive and up-to-date indices of the literature by indexing articles on the Web, technical reports, conference papers, etc. Furthermore, ACI makes it easy to browse the context of citations to given articles, allowing researchers to quickly and easily see what subsequent researchers have said about a given article. Digital libraries incorporating ACI may significantly improve scientific dissemination and feedback.

Short Abstract: *Autonomous Citation Indexing* (ACI) automates the construction of citation indices similar to the *Science Citation Index* ®. ACI aims to improve the dissemination and retrieval of scientific literature, and provides improvements in cost, availability, comprehensiveness, efficiency, and timeliness.

Keywords: citation indexing, digital libraries, bibliometrics, information retrieval, scientific literature, literature search.

1 Introduction

The rapid increase in the volume of scientific literature has led to researchers constantly fighting information overload in their pursuit of knowledge. Staying up to date with the literature is becoming increasingly difficult, if not impossible. A fundamental problem that researchers face is that of obtaining relevant articles. Experience varies significantly, but the time when every essential journal was held in all major academic libraries has passed [8]. The most common reason cited is that the price inflation of journals has outstripped library (and personal subscriber) budgets [8]. The Web promises to make more scientific articles easily available to more scientists. Many print journals now provide access to the full-text of articles on the Web, and the number of electronic journals continues to increase (about 1,000 journals in 1996). An increasing number of authors, journals, institutions, and archives (e.g. the physics e-Print archive at `http://xxx.lanl.gov` [7]) make research articles available for almost immediate access. However, scientific literature on the Web remains remarkably disorganized. A scientist may post a relevant preprint on his or her Web site, but finding the article in a timely fashion may be difficult – the Web search engines have difficulty keeping up to date [9] and do not index the contents of Postscript and PDF files, and no researcher has the time to find and evaluate all of the articles on the Web that might be of interest. Autonomous Citation Indexing (ACI) addresses these problems.

2 Citation Indexing

A citation index [5] indexes the citations that an article makes, linking the articles with the cited works. Citation indices were originally designed mainly for information retrieval [5], and allow navigating the literature in unique ways. Papers can be located independent of the particular words or language used. A citation index allows navigation backward in time (the list of cited articles) and forward in time (which subsequent articles cite the current article?) Citation indexing is a valuable tool for literature search, and allows a researcher to find out what subsequent papers cite a given paper. Citation indexing can improve scientific communication by helping to limit the wasteful duplication of prior research, revealing relationships between articles, drawing attention to important corrections or retractions of published work, and identifying significant improvements or criticisms of earlier work. Citation indices can also be used to find out where and how often a particular article is cited (thus providing an indication of the importance of an article), to analyze research trends, and to identify emerging areas of science [5].

3 Digital Libraries and Citation Indexing

The Institute for Scientific Information (ISI) (`http://www.isinet.com`) produces multidisciplinary citation indices including the *Science Citation Index* ® (SCI), which indexes the significant scientific journals. The ISI databases are constructed using manual assistance from human indexers. ISI provides the *Web of Science* ®, which is a Web version of the ISI citation indices. The ISI databases have a number of limitations. For example, they are expensive (which limits availability), have low coverage of the literature, and do not show the context of citations. The limitations of traditional citation indexing systems are expanded on in the following section.

Citation links, linking the bibliography of an article to the abstract or full-text of cited articles, are starting to be used in limited domains on the Web. HighWire Press (`http://highwire.stanford.edu`), a leading Web scientific publisher, includes citation links within the journals that they publish. The Open Journal project [8] has been investigating the provision of a *link service*, which is intended to provide citation linking across journal sites from different publishers. However, most scientific papers available on the Web currently provide no form of citation linking.

Cameron proposed a universal bibliographic and citation database that would link every scholarly work ever written [3]. He described a system in which all published research would be available to and searchable by any scholar with Internet access. The database would include citation links and would be comprehensive and up to date. Cameron's proposed system would avoid the requirement for manual assistance in citation indexing by transferring the manual effort to the authors or institutions – authors or institutions would be required to provide citation information in a specific format. This requirement of significant work placed on the authors or institutions is probably a major factor preventing the realization of Cameron's proposal. *Autonomous Citation Indexing* (ACI), described in detail below, overcomes this problem by completely automating the citation indexing process without requiring any extra effort from authors or institutions. Additionally, ACI goes beyond previous proposals by extracting and making the context of citations quickly and easily available, thereby improving literature search and evaluation.

4 Limitations of Current Tools

The TULIP project from Elsevier [1] tested systems for networked delivery and use of journals, and involved one of the most extensive user evaluations to date [8]. The final report summarized user requirements: ease of use, access to all information from one source, timeliness of information, effective search capabilities, fast downloading and printing, high image text quality, sufficient journal and time coverage, and links between documents [1]. Current access to scientific information on the Web is limited, and does not meet most of these requirements: the sources of scientific information are many and varied, ease of use and query interfaces vary dramatically, and no service provides comprehensive coverage. Furthermore, it seems unlikely that current services will meet all of the requirements above, for example the citation indexing performed by the *Web of Science* is limited in coverage because of the requirement for manual assistance during indexing.

Citation indexing is a valuable tool that greatly enhances digital libraries of scientific articles. However, citation indexing has received much criticism. For economic reasons (and for quality concentration) ISI primarily indexes only top-ranking journals in SCI [5]. A recurrent criticism against SCI is that the index is biased because of the journal selection process. The selection of journals tends to change slowly, often failing to keep up with the development of new journals [3]. For literature search, the restriction to journal articles can be very significant. In many research fields such as computer science and mathematics, significant results can appear in conference proceedings that either do not appear in journal form for months or years, or do not appear in a journal at all. More details on the limitations of traditional citation indices are contained in the discussion section.

Autonomous Citation Indexing (ACI) addresses many of these and other limitations.

5 Autonomous Citation Indexing

An Autonomous Citation Index autonomously creates a citation index from literature in electronic format. An ACI system autonomously locates new articles, extracts citations, identifies citations to the same article that occur in different formats, and identifies the context of citations in the body of articles. The viability of autonomous citation indexing depends on the ability to perform these functions accurately. We have built a prototype digital library with ACI called *CiteSeer* [6], that performs these tasks sufficiently accurately. CiteSeer works by downloading papers that are made available on the World Wide Web, converting the papers to text, parsing them to extract the citations and the context in which the citations are made in the body of the paper, and storing the information in a database. CiteSeer includes full-text indexing of the articles and citations. CiteSeer allows the location of papers by keyword search or by citation links. Papers related to a given paper can be located using common citation information or word vector similarity. Given a particular paper of interest, CiteSeer can display the context of how the paper is cited in subsequent publications.

CiteSeer operates completely autonomously. The following sections describe the document acquisition, document processing and parsing, and query and browsing aspects of CiteSeer.

5.1 Locating Documents

Finding articles can be accomplished by searching the Web, monitoring mailing lists or newsgroups for announcements of new articles, or by direct links with publishers. Once familiar with ACI systems, researchers may send notification of new papers directly, allowing these papers to be indexed almost immediately. Journal papers are increasingly being made available online at journal Web sites. Journals typically charge for access to online papers, and as such one way to index these papers would be to make agreements with the publishers. An ACI system is likely to be beneficial to publishers, because users can be directed to the journal's Web site, increasing subscriptions. Currently, CiteSeer uses Web search engines (e.g. AltaVista, HotBot, Excite) and heuristics to locate good starting points for crawling the Web (e.g. CiteSeer can search for pages that contain the words "publications", "papers", "Postscript", etc.). CiteSeer locates and downloads Postscript or PDF files. Duplicate URLs and files are avoided.

5.2 Document Processing

The downloaded Postscript or PDF files are converted into text using PreScript from the New Zealand Digital Library project (`http://www.nzdl.org/technology/prescript.html`). The text file is checked to verify that the document is a valid research document, for example by testing for the existence of a references or bibliography section. Many Postscript files print the pages in reverse order – CiteSeer detects and reorders these documents.

Autonomously finding the citations in an article can be done by locating the section containing the reference list, either by identifying the section header or the citation list. It is necessary to detect inserted material, such as figures and page numbers of the document being indexed, which can appear within the reference list. Once the set of references has been identified, individual citations are extracted. The list of citations is typically formatted such that citation identifiers, vertical spacing, or indentation can be used to delineate individual citations. Each citation is parsed using heuristics to extract fields including: the title, author, year of publication, page numbers, and the citation identifier. The citation identifiers (e.g. "[6]", "[Giles97]", "Marr 1982") are used to find the locations in the document body where the citations are actually made, allowing CiteSeer to extract the context of the citations. Variations in the citation identifier, such as listing all authors or only the first author, or varying use of initials, are handled using regular expressions.

The heuristics used to parse the citations were constructed using an "invariants first" philosophy. That is, subfields of a citation that had relatively uniform syntactic indicators as to their position and composition given all previous parsing, are always parsed next. For example, the label of a citation to mark it in context always exists at the beginning of a citation and the format is uniform across all citations in an article. Once the more regular features of a citation are identified, trends in syntactic relationships between subfields to be identified and those already identified are used to predict where the desired subfield exists (if at all).

For example, author information almost always precedes title information. CiteSeer also uses databases of author names, journal names, etc. to help identify the subfields of citations.

Citations to a given article can be made in significantly different ways. For example, Figure 1 shows a sample of citations to the same article extracted from machine learning publications on the Web. Much of the significance of ACI and CiteSeer derives from the ability to recognize that all of these citations refer to the same article. For example, this ability allows a detailed listing of a cited article to show all instances of the citation across multiple articles. Identifying variant forms of citations to the same article also enables statistics on citation frequency to be generated, allowing the estimation of the importance of articles.

```
Aha, D. W. (1991), Instance-based learning algorithms, Machine Learning 6(1), 37-66.
D. W. Aha, D. Kibler and M. K. Albert, Instance-Based Learning Algorithms. Machine
        Learning 6 37-66, Kluwer Academic Publishers, 1991.
Aha, D. W., Kibler, D. & Albert, M. K. (1990). Instance-based learning
        algorithms. Draft submission to Machine Learning.
```

Figure 1. A sample of citations to the same paper showing typical variations in citation format. These citations were extracted from machine learning publications on the Web.

As suggested by the example citations in Figure 1, the problem is not completely trivial. All fields, including the title, author names, and even the year of publication routinely contain errors. Autonomously determining the subfields of a citation is not always easy. For example, commas are often used to separate fields, but they are also used to separate lists of authors, and are embedded in some titles. Full stops are used to separate fields, but are also used to denote abbreviations. Sometimes there is no punctuation at all between fields. We have considered four broad classes of methods for identifying and grouping citations to identical articles:

1. *String distance* or *edit distance* measures, which consider distance as the amount of difference between strings of symbols. The *Levenshtein distance* is a well known edit distance where the difference between two strings is simply the number of insertions, deletions, or substitutions required to transform one string into another. A more recent and sophisticated example is *LikeIt*, an intelligent string comparison algorithm introduced by Yianilos [12].

2. *Word frequency* or *word occurrence* measures are based on the statistics of words that are common between strings. Word frequency measures such as *Term Frequency* $\times$ *Inverse Document Frequency* (TFIDF) are common in information retrieval.

3. Knowledge about *subfields* or the *structure* of the data can also be used. In the case of citations, subfields such as author name(s), title, year of publication, etc. can be used.

4. *Probabilistic models* can be trained using known bibliographic information in order to identify subfields from the words and/or structure of citations. These subfields could be used with the previous methods.

We have investigated algorithms from each of these classes, and have performed quantitative tests by extracting a number of sets of citations from online papers, manually grouping the identical citations, tuning the algorithms on a training set, and comparing the correct groupings with the automated groupings on test sets. CiteSeer currently uses an algorithm based on normalization of the citations, sorting according to length, and matching words and phrases (within subfields). On tests covering 1158 citations, this algorithm resulted in about 5% of the automated groupings containing an error with respect to the correct grouping of citations (this does not mean 5% of citations are incorrectly grouped – just one incorrect citation in a group marks the entire group as incorrect).

While the algorithm currently used by CiteSeer is sufficient for practical use, there are many avenues for improvement. For example, the use of learning techniques and probabilistic estimation based on training sets of known bibliographic data should be able to boost performance substantially. Large quantities of bibliographic information is freely available on the Web (e.g. the collection of computer science bibliographies at `http://liinwww.ira.uka.de/bibliography/index.html`), and this information provides labeled training data that learning techniques can use in order to associate the words and/or the structure of citations with the corresponding subfields. We chose not to use models trained on specific words initially because the sole use of such models will bias the errors made by the system – errors are more likely to be made for new authors, new journals, new areas, etc. that are not contained in the training data (which could potentially have a negative impact on scientific progress). Preliminary investigations suggest that probabilistic information from specific words and learning techniques can provide very good performance, and future research will consider adding these techniques to the methods above. Another method of improving citation matching performance would be to allow (certain) users to correct errors.

An ACI system also benefits from the ability to identify the bibliographic details of the papers that are indexed. Without this ability, it is not possible to properly analyze the graph formed by citation links; for example it is not possible to detect whether or not citations are self-citations. For specific publication venues, it should be possible to take advantage of the regularity across individual articles in order to autonomously extract items such as the title, authors, etc. with high accuracy. However, for papers that are autonomously located on the Web, on a researcher's homepage for example, the problem is more difficult. We have investigated the problem of determining author and title information from Postscript or PDF files that have been autonomously found on the Web. By analyzing font information, it is relatively easy to extract the title and the first author. Accurately identifying all authors and author addresses is somewhat harder, however it is relatively simple to detect the header of an article, and thus to detect self-citation. CiteSeer currently marks a citation as a self-citation if one of the author's names is located in the header of the indexed article. There are avenues for accurately extracting all author names and addresses from arbitrary articles. For example, recognizing and parsing the various forms of article headers, using databases of author names, cross-correlating with bibliographic records, or learning techniques may enable accurate author identification for arbitrary articles. As before, sole use of learning techniques based on specific word information is not preferred because of the potential for biasing the errors.

5.3 Query and Browsing

The first query to CiteSeer is a keyword search, which can be used to return a list of citations matching the query, or a list of indexed articles. The literature can then be browsed by following the links between the articles made by citations. Figure 2 shows a sample response for the query `quinlan` in a CiteSeer digital library of the machine learning literature. The number of citations to each article is given in the leftmost column. The "hosts" column indicates the number of unique hosts (Web servers) that the articles containing the citations originated from, and the "self" column indicates the number of citations to the given paper that are predicted to be self-citations. At the end of the response is a graph showing the number of citations versus the year of publication of the cited articles. The number of self-citations is not included in the main number of citations or the graph.

CiteSeer indexes the full text of citations and articles, providing full Boolean search with phrase and proximity support. When searching for citations, the default mode of operation is to retrieve all citations matching the given query, group the citations to identical papers, and order the results by the number of citations to each paper (options include ordering by date and field restrictions). CiteSeer does not currently perform any special processing in order to account for the variant ways of referencing proper names (e.g. first name or one or more initials), however the Boolean and proximity support can be used to cover variant forms of author names (if an author's last name is unique then it is sufficient to just search for the last name). CiteSeer does not use any "stop" words (common words like "the" typically excluded from indexing), so it is possible to search for phrases containing initials. When searching in the full text of articles, CiteSeer returns the header for matching documents along with the context of the article where the keywords occur. Documents can be ordered according to the number of citations to them, citations to important articles (e.g. review articles), or date. Details of particular documents, including the abstract, full text, list of citations, and an "active bibliography" of related documents can be obtained.

Once an initial keyword search is made, the user can browse the digital library using citation links. The user can find which papers are cited by a particular publication and which papers cite a particular publication, including the context of those citations. Figure 3 lists the papers that cite a particular article in Figure 2, along with the context of the citations (obtained by clicking on the appropriate [Context] link in Figure 2). The context of citations can be very helpful for literature search, for author feedback, and for evaluation. The context may contain a brief summary of the paper, another author's response to the article, limitations or criticism of the original work, or subsequent work that builds upon the original article. The context of citations can help a researcher to determine whether or not the citing or cited article should be read in full.

CiteSeer can find articles related to a given article. A number of algorithms are used [6]: a) *word vectors* – a TFIDF scheme is used to locate articles with similar words, b) *LikeIt* edit distance comparison of the article headers is used to find similar headers, c) a new algorithm called *Common Citation* $\times$ *Inverse Document Frequency* (CCIDF) is used to find articles with similar citations, and d) a combination of the previous three algorithms is used. The CCIDF scheme is particularly interesting. Citations are hand picked by the authors as being related documents, and it seems intuitive to use citation information to judge the relatedness of documents. CCIDF is analogous to the word oriented TFIDF, and considers the common citations between any pair of documents weighted by the inverse frequency of citation (so that, for example, common citations to highly-cited methodological papers are weighted lower). When viewing the details of an article, CiteSeer displays an "active bibliography" of related documents.

6 Discussion

NEC Research plans to make CiteSeer freely available to researchers. For current information, please contact `citeseer@research.nj.nec.com` (see `http://www.neci.nj.nec.com/homepages/lawrence/citeseer.html`). To subscribe to the CiteSeer announcements mailing list, send a message to `majordomo@research.nj.nec.com` with the text `subscribe citeseer-announce` in the body of the message.

While CiteSeer is already in practical use, there are many avenues for further improving the dissemination and access of scientific information on the Web. For example, printed literature may be processed with OCR and stored efficiently using technology such as the DjVu image compression technology (`http://djvu.research.att.com/`). Digital libraries with ACI can provide many additional services such as current awareness and community features. For example, papers or research topics may be linked to a

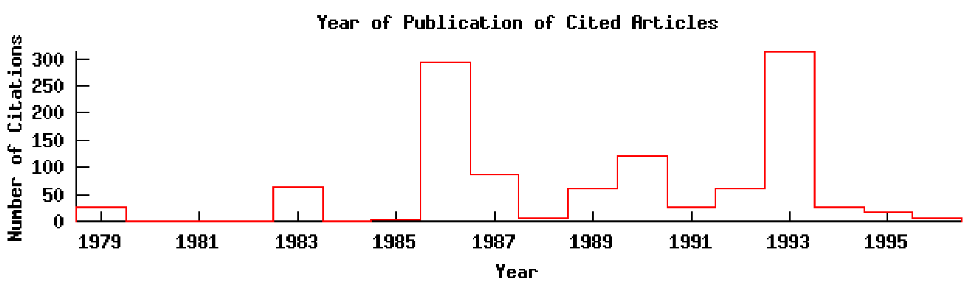
CiteSeer
Autonomous Citation Indexing

New Search Options Help Add Documents Feedback

Searching for **quinlan** in **Machine Learning (small test database)** (15260 documents 273478 citations total).
Retrieving citations... 1162 citations found
Click on the [Context] links to see the citing documents and the context of the citations. Track All Documents

Citations [hosts] (self)	Article
266 [70] (6)	J. R. **Quinlan**. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Mateo, California, 1993. Context Bib Track Check
243 [73] (1)	**Quinlan**, J. (1986). *Induction of Decision Trees*. Machine Learning, 1:81-106. Context Bib Track Check
96 [28] (2)	**Quinlan** J. R., "*Learning Logical Definitions from Relations*", Machine Learning 5 (1990) 239-266 Context Bib Track Check
49 [22]	J. R. **Quinlan**.*"Learning efficient classification procedures and their application to chess end games"*, In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, eds, Machine Learning: An Artificial Intelligence Approach, Palo Alto: Tioga, 1983: 463-482. Context Bib Track Check
42 [22] (1)	**Quinlan**, J.R. (1987). *Simplifying decision trees*. International Journal of Man-Machine Studies, 27(1):221-234. Context Bib Track Check
27 [18] (3)	J. R. **Quinlan** and R. L. Rivest. *Inferring decision trees using the minimum description length principle*. Information and Computation, 80:227-248, 1989. Context Bib Track Check

[... section deleted ...]



Self-citations are not included in the graph or the main number of citations.

Figure 2. An example of how a digital library with ACI can group citations to the same paper and rank papers based on the number of citations. This example shows the results of a search for the author `quinlan` in a small test digital library of the machine learning literature (we have not attempted to index all of the machine learning literature available on the Web). The `Context` links show the context of the individual citations, a sample of which can be seen in Figure 3, the `Bib` links provide the bibliographic details of the respective articles, the `Track` links activate tracking (the context of new citations will be emailed to the user), and the `Check` links show the variant forms of the citations, which can be used to check for errors in the grouping algorithm. The `hosts` and `self` numbers indicate the number of distinct hosts the citing articles were found on, and the number of citations predicted to be self-citations. The graph at the bottom shows the number of citations versus the year of publication of the cited articles. The titles of the articles are italicized automatically.

J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Mateo, California, 1993. Summary Details

This paper is cited in the following contexts:

Towards a Framework for Memory-Based Reasoning - Simon Kasif kasif@cs.jhu.edu - Steven Salzberg salzberg@cs.jhu.edu - David Waltz waltz@research.nj.nec.com - John Rachlin rachlin@cs.jhu.edu - David Aha aha@aic.nrl.navy.mil Details

......when using symbolic-valued features. The VDM is an adaptive distance metric that adjusts itself to a database of examples, and can then be used for retrieval (see Section 4). **Tree-based methods for partitioning data into regions (e.g., [Omo89, Omo87]) such as k-d trees or decision trees [Qui93] also can be used to define a relevant local neighborhood.** Thus, instead of seeing a decision tree as a classification device in the MBR context, a decision tree defines a static partitioning of space into regions. In other words, the distance between data instances that are grouped in the same......

[Qui93] J. R. **Quinlan**. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

Learning Symbolic Rules Using Artificial Neural Networks - Mark W. Craven and Jude W. Shavlik - Computer Sciences Department - University of Wisconsin - 1210 West Dayton St. - Madison, Wi 53706 - email:craven, shavlik@cs.wisc.edu, Appears in Machine Learning: Proceedings of the Tenth International Conference, - P. E. Utgoff (editor), Morgan Kaufmann, San Mateo, Ca, 1993 Details

......was designed as a technique for improving generalization in neural networks, we explore it here as a means for facilitating rule extraction. **We present experiments that demonstrate, for two difficult learning tasks, our method learns rules that are more accurate than rules induced by Quinlan's (1993) C4.5 system.** Furthermore, the rules that are extracted from our trained networks are comparable to rules induced by C4.5 in terms of complexity and understandability. Towell and Shavlik (1991) demonstrated that concise and accurate symbolic rules can be extracted in the restricted case of......

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Design and Evaluation of the Rise 1.0 Learning System - Pedro Domingos - pedrod@ics.uci.edu - Technical Report 94-34 - August 30, Department of Information and Computer Science - University of California, Irvine - Irvine, California 92717, U.S.A. Details

......as the domain difficulty grows, without sacrificing speed. Introduction and motivation. Current machine learning approaches to the induction of concept definitions from examples fall mainly into two categories: "divide and conquer" and "separate and conquer. **"Divide and conquer" methods [11, 14] recursively partition the instance space until regions of roughly constant class membership are obtained.** This approach has often worked well in practice, but is plagued by the splintering of the sample that it causes, resulting in decisions being made with less and less statistical support as......

[14] J. R. **Quinlan**. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[... section deleted ...]

Figure 3. An example of how an ACI system can show the context of citations to a given article. A list of citing articles is shown. For each article, the header, the context of the citation, and the specific form of the citation is shown. The sentence containing the citation is automatically highlighted. The `Details` links allow viewing the full details of the articles (header, abstract, citations, source location, related documents, etc.). The `Summary` link shows a summary of citing documents without citation context.

discussion area where scientists may post formal or informal comments, reviews, responses, new results, etc. Researchers may sign up to receive email when new citations are found to given papers, when new articles are indexed that match a personal profile, or when a new response is posted regarding a paper of interest. In addition to autonomous citation indexing, the CiteSeer project encompasses areas including: the location of articles, efficient full-text indexing, information extraction, the display of query-sensitive summaries and citation context, browsing, clustering, hubs and authorities, related document detection, detection of overlapping documents, current awareness, user modeling, error correction, and graph analysis.

CiteSeer should be considered complementary to commercial citation indices such as SCI. Although CiteSeer is sufficiently accurate to be very useful, SCI can provide greater accuracy, especially in areas where they index informal citations (e.g. referencing a work of art within the body of an article). There are many avenues for improvement to the algorithms that enable ACI. For example, combinations of the methods for identifying citations to the same article may substantially improve performance, and learning techniques may be used to improve many aspects of the system.

Citation indices such as SCI are limited in the number of journals that they can index, because the indexing process requires manual effort, and corresponding economic reward. The justification for such selective indexing is that a relatively small number of journals accounts for the bulk of significant scientific results [11]. However, the fact that a small set of journals accounts for a large percentage of citations is probably partially due to information overload – researchers may only read a small set of journals and miss significant results published elsewhere (it has been argued that the use of citation indices tends to emphasize the particular journals indexed and devalue other forms of scientific communication [3]). Widespread use of digital libraries with ACI should promote the dissemination of more literature.

There are definite disadvantages to limited journal selection. Journal selection typically follows a review process, which necessarily means that articles making the journal worthy of indexing have already been published, and were previously unavailable in the citation index. The limitation to journals excludes conferences, monographs, technical reports, and preprints (SCI indexes some non-journal items). Important new findings may be conveyed with short conference papers, while not justifying journal publication, and many ideas or feedback on previous work do not require lengthy papers to present. These ideas or feedback can be presented in technical reports, working papers, or conference papers far in advance of the corresponding journal publication [4]. In areas such as computer science, significant work is often presented in conferences and the journal publication, if at all, comes much later. These areas can be rapidly moving, and the time between conference and journal publication can be significant. The broader coverage that can be provided by ACI can clearly be helpful for literature search, allowing a scientist to find work that cites his or her work, or that is relevant to the scientist's research. Features other than "appears in journal x" may be indicative of importance, e.g. preprints from specific authors, or feedback on a researcher's own work. For work that reaches journal publication, broader coverage of preprints, technical reports, and conference proceedings can provide very significant advantages of timeliness (which can increase productivity, point out limitations before starting research of limited value, help reduce the duplication of research, etc.). Even work that does not reach journal publication may contain important and/or useful feedback, or connections within the scientific community.

Citation statistics are widely used for evaluation. However, evaluation based on citation statistics can lead to erroneous conclusions. Authors that receive a similar number of citations may do so for different reasons, e.g. a small number of researchers may cite each other but have little influence outside of their clique, and researchers can produce significantly different quantities of citations to their own papers. The underlying assumption that citations imply scholarly impact is not always true (e.g. [2]). What is actually said about a cited document can be very important, but is typically not considered when evaluating citation statistics. Evaluation based on journal citation statistics is also delayed by the journal reviewing and publication processes, meaning that statistics on recent work may not be available, although citations may exist in conference proceedings and preprints. By making the context of citations easily and quickly browsable, ACI can help to more accurately evaluate the importance of individual contributions. By indexing a broader selection of the literature including conference papers and preprints, ACI can help to provide more timely evaluation of articles.

ACI facilitates the possibility of autonomously classifying the context of citations using learning techniques, which may be useful when there are a large number of citations to a given paper. We are investigating the classification of citation context into classes such as "Supporting Reference", "Criticism", "Followup Work", "Review", and "Praise". Although we do not expect accuracy comparable to humans, we do expect that learning techniques can provide useful predictions by analyzing the manner of citation and the usage of words in the sentences surrounding the citation. For example, if a citation occurs in a group of citations (e.g. "see [1,3,4,7,11,14]") then the classification "Supporting Reference" may be more probable. If a paper is cited several times in the same paper, then the citing paper may be more likely to be "Followup Work".

7 Summary

The World Wide Web has revolutionized the way that people access information. Researchers and institutions are increasingly making articles available on the homepages of authors or institutions, or in online archives. However, scientific information on the Web is largely disorganized. This article introduced the creation of digital libraries incorporating *Autonomous Citation Indexing* (ACI). ACI systems autonomously locate articles, extract citations, identify identical citations that occur in different formats, and identify the context of citations in the body of articles.

We have built a working, practical prototype system: CiteSeer. CiteSeer operates on papers available in electronic form. Specifically the system locates and indexes the rapidly increasing number of articles that are made available on the Word Wide Web. However, ACI is not limited to indexing literature on the Web. Any literature can be indexed, e.g. printed articles can be converted to electronic form using optical character recognition (OCR). Compared to the current commercial citation indices, CiteSeer does not cover the significant journals as comprehensively, and provides lower accuracy. These disadvantages are expected to decrease over time. However, CiteSeer provides significant advantages compared to current citation indices. Because CiteSeer can index articles as soon as they are available on the Web, it should be of greater use to researchers for finding recent relevant literature, and for keeping up to date. CiteSeer is autonomous, requiring no manual effort during indexing, which should lead to wider availability, more comprehensive indices, and lower cost. CiteSeer makes the context of citations easily browsable, facilitating more efficient literature search, and more informed estimation of the impact of given articles. By covering more recent work and operating autonomously, CiteSeer facilitates an increased rate of dissemination and feedback.

The revolution that the Web has brought to information dissemination is not so much due to the availability of information (huge amounts of information has long been available in libraries), but rather the improved efficiency of accessing information. Digital libraries incorporating ACI can help to organize the literature, and may significantly improve the efficiency of scientific dissemination and feedback. The transition to scholarly electronic publishing has been slow [10]. ACI may help to speed up this transition. For example, having a widely/freely available linked network of the literature may encourage scientists to pursue publication avenues that make their work available in the online network as quickly as possible (and therefore be more likely to be seen, cited, and create impact earlier).

Acknowledgments

We would like to thank Haym Hirsh, Bob Krovetz, Michael Lesk, Michael Nelson, and Craig Nevill-Manning for useful comments and suggestions.

References

[1] Marthyn Borghuis, Hans Brinckman, Albert Fischer, Karen Hunter, Eleonore van der Loo, Rob ter Mors, Paul Mostert, and Jaco Zijlstra. TULIP final report, 1996, `http://www.elsevier.nl/inca/homepage/about/resproj/tulip.shtml`.

[2] T. A. Brooks. Evidence of complex citer motivations. *Journal of the American Society for Information Science*, 37:34–36, 1986.

[3] Robert D. Cameron. A universal citation database as a catalyst for reform in scholarly communication. *First Monday*, 2(4), 1997, `http://www.firstmonday.dk/issues/issue2_4/cameron/index.html`.

[4] S.L. Esler and M.L. Nelson. Evolution of scientific and technical information distribution. *Journal of the American Society for Information Science*, 49(1):82–91, 1998.

[5] Eugene Garfield. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. Wiley, New York, 1979.

[6] C. Lee Giles, Kurt Bollacker, and Steve Lawrence. CiteSeer: An automatic citation indexing system. In Ian Witten, Rob Akscyn, and Frank M. Shipman III, editors, *Digital Libraries 98 - The Third ACM Conference on Digital Libraries*, pages 89–98, Pittsburgh, PA, June 23–26 1998. ACM Press.

[7] P. Ginsparg. First steps towards electronic research communication. *Computers in Physics*, 8:390–396, 1994.

[8] S. Hitchcock, L. Carr, S. Harris, J.M.N. Hey, and W. Hall. Citation linking: Improving access to online journals. In Robert B. Allen and Edie Rasmussen, editors, *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pages 115–122, New York, NY, 1997. ACM.

[9] Steve Lawrence and C. Lee Giles. Searching the World Wide Web. *Science*, 280(5360):98–100, 1998.

[10] A.M. Odlyzko. Tragic loss or good riddance? The impending demise of traditional scholarly journals. *International Journal of Human-Computer Studies*, 42:71–122, 1995.

[11] James Testa. The ISI database: The journal selection process, `http://www.isinet.com/whatshot/essays/esay9701.html`, 1997.

[12] Peter Yianilos. The LikeIt intelligent string comparison facility. Technical Report 97-093, NEC Research Institute, 1997, `http://www.neci.nj.nec.com/homepages/pny/papers/likeit/main.html`.