Enriching a Document Collection by Integrating Information Extraction and PDF Annotation

Brett Powley, Robert Dale, and Ilya Anisimoff

Centre for Language Technology, Macquarie University, Sydney, Australia
{bpowley,rdale,ilya}@ics.mq.edu.au

ABSTRACT

Modern digital libraries offer all the hyperlinking possibilities of the World Wide Web: when a reader finds a citation of interest, in many cases she can now click on a link to be taken to the cited work. This paper presents work aimed at providing the same ease of navigation for legacy PDF document collections that were created before the possibility of integrating hyperlinks into documents was ever considered. To achieve our goal, we need to carry out two tasks: first, we need to identify and link citations and references in the text with high reliability; and second, we need the ability to determine physical PDF page locations for these elements. We demonstrate the use of a high-accuracy citation extraction algorithm which significantly improves on earlier reported techniques, and a technique for integrating PDF processing with a conventional text-stream based information extraction pipeline. We demonstrate these techniques in the context of a particular document collection, this being the ACL Anthology; but the same approach can be applied to other document sets.

1. INTRODUCTION

When a researcher is reading a scholarly article, she encounters citations which correspond to complete references provided in footnotes, in end notes, or in a bibliography at the end of the article. The purpose of these references, of course, is to enable the reader to locate the actual documents referred to. In a pre-web world, locating such a document meant a trip to one's bookcase, or perhaps to the library; today, where an increasing number of documents are to be found floating in cyberspace, it often means the posing of a query to a search engine such as Google.

Given the relative ease with which document production systems such as LaTeX and Microsoft Word now enable the incorporation of hyperlinks, we foresee a future where documents are not cast adrift upon creation, but are born as nodes in a rich network of easy-to-follow citation linkages: if you are curious about another paper referred to in the article you are currently reading, that paper will be retrievable in the blink of an eye by simply clicking on the citation. Indeed, some publishers already produce documents containing hyperlinks that provide this kind of connectivity. However, there exist on the web a vast number of documents that pre-date the availability of these possibilities. Our interest is in exploring how techniques from language technology and document processing can be used to retrofit such documents with links and annotations that integrate them into a true web of science.

Our goal in this paper is to describe a system that automatically enriches documents within a particular online document collection (the ACL Anthology*) by automatically generating pop-up annotations that appear when the reader hovers her mouse over citations. These annotations provide the full details of the reference, thus removing the need to flip forwards to the reference list (at the risk of losing one's place) when one's curiosity is piqued. At the same time the annotations on the citations provide a direct hyperlink to the external document itself, so removing the need to actively search for the referred-to document. The extent to which this is possible depends, of course, on the availability of the target work in online form.

Although we demonstrate our approach on one specific document collection, the techniques are more broadly applicable; we believe they can be ported to other document collections relatively straightforwardly.

The paper proceeds as follows. First, in Section 2, we briefly describe some related work. Then, in Section 3, we introduce the ACL Anthology, an open access repository of research articles and papers in the field

*See `http://aclweb.org/anthology-index/`.

of computational linguistics; this serves as the test bed for our study. In Section 4, we describe our customised PDF text extraction tool, which labels extracted text elements with their physical locations on document pages. Section 5 then describes the process we use for the automatic extraction of citations from the documents and the identification of these citations with their references. Section 6 describes the results of running our text extraction process on PDF files; Section 7 describes how this information is used to enrich the documents within the corpus we worked with. Section 8 points to some directions for future work.

2. RELATED WORK

The idea of a richly linked hyperspace of knowledge is the very foundation of the World Wide Web, and has a long history, with the basic idea often being credited to Vannevar Bush in his 1945 *Atlantic Monthly* article [4]. It took another 40 years for Bush's vision to become a reality. It is natural nowadays to create knowledge resources, regardless of their form, with their incorporation into hyperspace as a concern from the outset; but the integration of previously existing resources requires often laborious manual editing and annotation.

The automation of this integration of legacy material requires a range of technologies; for the kind of material we are concerned with, we are primarily concerned with identifying the formal linkages between documents, and making it easy for readers to follow these links even when they reside in documents that are not purpose-made HTML files.

The potential for automatic extraction of citation links was first identified by Garfield in 1955 [5]. In the last decade, a variety of approaches to extracting citations and references from academic papers have been explored.

Bergmark et al. [2] report on heuristics for extracting citations (which they call 'contexts' and 'reference anchors') from ACM papers, reporting precision of 0.53, based on randomly selected papers. These papers exclusively use numbered citation keys (e.g. *[1]*) rather than the textual keys which we aim to extract. Bergmark [1] reports in more detail on extracting information from digital library papers, including citations in a variety of formats. She does not report results for individual extraction tasks, but reports 86.1% 'average accuracy', for the number of 'elements' correctly extracted from each document; 'elements' include the title, author, year of publication, references, and citations. The widely-used CiteSeer system developed by Giles et al. [6] attempts to extract bibliographic information from references as well as the citation context (i.e. words surrounding the citation). They report being able to extract authors from references 82.1% of the time.

A number of web-based systems have been implemented to provide a user interface to citation linking data. Notable examples are the ISI Web of Science (founded by Garfield himself), and more recently applications such as CiteSeer[†] [6] and Google Scholar.[‡]

3. THE CORPUS

The focus of our work is the ACL Anthology Reference Corpus [3] (ACL ARC), a collection of 10,921 documents in PDF form drawn from the ACL Anthology website as it stood at Februrary 2007.[§] The Anthology itself is a resource provided freely to the community by the Association for Computational Linguistics; it contains the complete collection of papers from the Association's annual conferences since its inception, as well as the papers from a host of associated workshops and other conferences. The collection also includes the past contents of the *Computational Linguistics* journal. Consequently, the Anthology represents a significant and comprehensive resource for anyone working in the field of computational linguistics.

Material dated prior to 2000 has generally been incorporated into the Anthology by means of scanning and optical character recognition being applied to original hard copies. Inevitably, this process introduces errors into the text representations of the papers, and results in problems in searching the repository. Since 2000, the Association's conference proceedings have been 'born digital', and so do not suffer from this problem; the nature of PDFs, however, means that there are other issues that arise when trying to find specific text strings

[†] `http://citeseer.ist.psu.edu`

[‡] `http://scholar.google.com`

[§] New material is added to the Anthology on an ongoing basis, with the proceedings of conferences often appearing simultaneously with, or even prior to, the occurrence of the conferences themselves.

within these files. For the work described in this paper, approximately 6,300 documents from the Anthology were used, representing those which did not have major text extraction errors due to either OCR errors (for older papers), or custom PDF font subset encodings. This subcorpus contains approximately 115,000 citations and 89,000 references. One particularly useful property of the Anthology is that around 20% of citations are to other documents in the Anthology, thus making it possible to envisage the provision of hyperlinks between a significant number of documents.

4. INTEGRATING PDF AND TEXT STREAM PROCESSING

The first step in any information extraction task involving a PDF document is the extraction of a text stream from the document. However, extracting a usable text stream from a PDF file is non-trivial. The PDF format was designed for accurate rendering of a document on screen or on a printer, and not for recovery of the original text. Therefore, features such as the order of text in the PDF file, the division of text into chunks for rendering, and the presence or absence of explicit spaces between words, are dependent not on any meaningful features of the original document such as word breaks or text flow, but rather on the rendering path used to produce the PDF document. There are, of course, a large number of PDF text extraction tools available both commercially and in the public domain; however, each of these has its idiosyncracies, and no tool that we have tested is able to reliably produce with high reliability a clean stream of text that corresponds to the actual text flow in the document. Faced with this problem, we decided that we had to develop a more robust mechanism that carried out significant processing of the elements in a PDF file, with the specific aim of producing output that would integrate readily with an information extraction pipeline.

For this project in particular, we wanted to provide the infrastructure for adding annotations to the PDF files from which bibliographic data had been extracted. This implied that the physical positions of extracted elements in the PDF documents were required. More precisely, we required the page number, $\langle x, y \rangle$ coordinates, and the width and height of each element. While it would be theoretically possible to construct an information extraction system based on raw PDF data, in practice such an approach is problematic: most existing information extraction tools and pipelines are designed to work on text streams.

Our approach therefore was to develop a dual-stream processing pipeline, an overview of which is depicted in Figure 1. The idea behind this pipeline is to produce simultaneously, for each document, output suitable for information extraction and for PDF manipulation.

We developed a custom PDF text stream extractor, using the open-source tool PDFBox [¶] as our starting point. Our version of this tool produces two outputs for each processed PDF file. The first contains the text stream alone; this is the output file that would be used as the input to a standard information extraction pipeline. The second file contains, in XML format, a stream of tokens from the source PDF file, along with both each token's page location in the PDF file and also the location in the text stream (i.e., the other output file) of each token. This approach allows the information extraction pipeline to operate on the raw text stream, and at the end of that pipeline, text stream positions to be trivially mapped back into PDF page positions.

Another aspect of our information-extraction tuned PDF extraction is the provision of semantically useful chunks of text to later stages in the pipeline. Generally, the division of raw text drawing primitives in a PDF file have no relationship to semantically useful divisions such as word breaks; instead, they tend to be somewhat arbitrary clusters of characters whose generation depends on the rendering pipeline used to produce the PDF file. To produce a stream of words from these rendering commands, we processed each rendered chunk of text from the PDF files character-by-character. The position, width, and spacing of each character were analysed to detect whether it was a continuation of an existing word, or whether it represented a word or line break. At each detected break, a word was emitted together with its page position, width, and height, calculated using its font metrics.

Using this mechanism, we are able to deliver a richer representation of the text that can be used for a variety of other information extraction purposes; for example, it makes it easier to locate and skip over figures and other floating material, and it provides a basis for more reliable table extraction.

[¶] available at http://www.pdfbox.org/

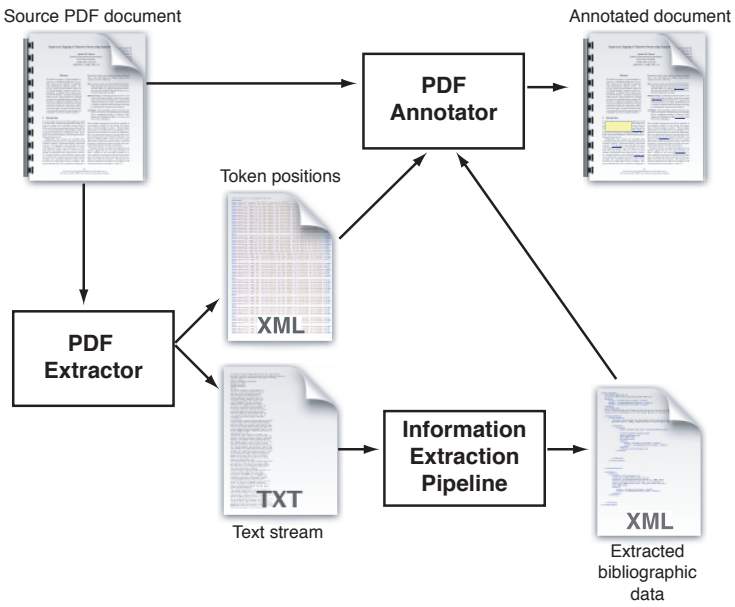


Figure 1. The dual-stream PDF processing pipeline

5. EXTRACTING CITATIONS AND REFERENCES

Once we have a clean, reliable and correctly ordered text stream, the next step is the extraction of bibliographic data from the documents in the corpus. This section describes the motivation and algorithms employed for extracting this data.

5.1 Some Terminology

First, to avoid any confusion, we make explicit our use of terminology:

- a **reference** appears, in the data considered here, in a list of works at the end of a document, and provides full bibliographic information about a cited work;
- a **citation** is a mention of a work in the body of the text, and includes enough information (typically, an author–year pair or an alphanumeric key) to uniquely identify the work in the list of references.

Powley and Dale [7] introduced terminology to describe the variety of citation styles encountered in academic literature; Table 1 summarises the major styles. For the present work, we are interested in *textual citations* only, although the work described here can be extended to other forms of citations.

As an output of the extraction process, we aim to provide the following metadata for each document in the collection:

- a bibliographic record for the document, comprising at least author name, title, year, and containing publication;
- the citations from the body of the document;
- the sentences containing those citations;
- the references from the reference section, segmented into individual references;
- each citation linked to its corresponding reference; and

Textual — Syntactic **Levin (1993)** provides a classification of over 3000 verbs according to their participation in alternations involving NP and PP constituents.
Textual — Parenthetical Two current approaches to English verb classifications are WordNet (Miller et al., 1990) and Levin classes **(Levin, 1993)**.
Prosaic **Levin** groups verbs based on an analysis of their syntactic properties especially their ability to be expressed in diathesis alternations.
Pronominal **Her** approach reflects the assumption that the syntactic behavior of a verb is determined in large part by its meaning.
Numbered There are patterns of diathesis behaviour among verb groups **[1]**.

Table 1. Citation styles [7]

- the target document for each reference, where it exists in the corpus.

There are three phases to the extraction process for producing this data. The *document mining* phase processes individual documents to extract citing sentences, citations, and references; the *web mining* phase uses web resources to obtain bibliographic information not readily available from the documents themselves; and the *interlinking* phase integrates data from the first two phases to produce interdocument links based on citations. These phases are described in detail in the following sections.

5.2 Document mining

In earlier work [8], we presented techniques for high accuracy extraction of citations from academic papers, designed for applicability across a broad range of disciplines and document styles. We integrated citation extraction, reference parsing, and author named entity recognition to significantly improve performance in citation extraction, and demonstrated this performance on a cross-disciplinary heterogeneous corpus. We showed that the retrieval performance of our algorithm significantly improved on earlier work, with f-measure results of 0.98 for the ACL Anthology corpus and 0.97 for the heterogeneous corpus, when applied to a sample of previously unseen documents. We leverage the results of that work in the present paper. The algorithm is described in detail in Powley and Dale [7], and a graphical representation of the processing associated with a sample citing sentence is shown in Figure 2. We briefly recap that algorithm here.

The citation extraction algorithm works at the sentence level to isolate and tag citations. We begin with the observation that textual citations are anchored around years; earlier work showed that we could identify candidate sentences containing citations with a recall of better than 0.99 simply by using the presence of a year as a cue [7].

Our first step is therefore to search each sentence for a candidate year token (a 'year' for this purpose being a 4-digit number between 1900 and the current year, potentially with a single character appended to it). If we find such a token, our task is then to determine whether it forms part of a citation, and if it does, to extract the author names that accompany it.

In general, we may say that a textual citation comprises one or more authors followed by one or more years; in practice, the variety of constructions which a writer might use to format a citation is somewhat more complicated.

Writers often use a list of years as shorthand for citing multiple papers by the same author: consider *Smith (1999; 2000)*, which represents citations of two separate works. Given the candidate year, we therefore first search backwards and forwards to isolate a list of years.

Our task is then to find the list of authors. While this often immediately precedes the list of years, this is not always the case; consider, for example, *Knuth's prime number algorithm (1982)*. We therefore search backwards from the year list, skipping words until we find an author name; currently, we choose to stop searching after

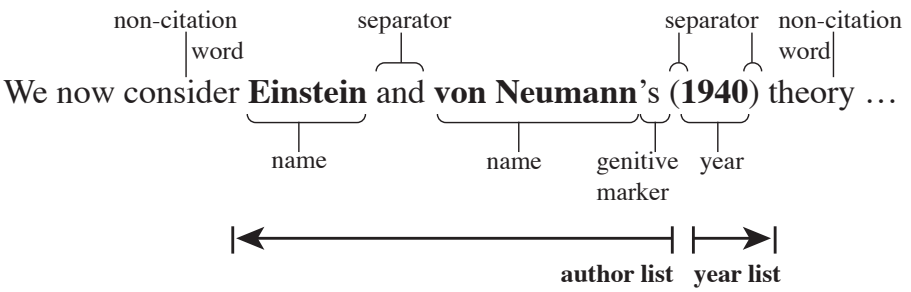


Figure 2. Extracting citation information [8]

10 words, as we have found that this choice gives good performance. Having found a single author name, we continue searching backwards for additional names, skipping over punctuation and separators, and stopping when we encounter a non-surname word.

If no author names are found, we conclude that the candidate year was a number unrelated to a citation. We also treat a small number of temporal prepositions which commonly appear before years as stopwords, concluding that the candidate year is not a citation if preceded by one of these (*in*, *since*, *during*, *until*, and *before*). Otherwise, having found a list of authors and list of years, we normalise the citation instance into a list of citations each consisting of a list of authors and a single year.

Distinguishing author names from other words requires an accurate named entity recogniser. We employ a named entity recognition algorithm based on the observation that any author name in the body of the document ought also to appear in the references section. Given a candidate name (generally, a capitalised word preceding a year), we attempt to locate the same name in the references section of the document in a context which also includes the candidate year. Additionally, we use alignment between words in the body of the document and the references section to detect multi-word author names; for example, *van den Bosch*, *Al Shalabi*, *Gaustad van Zaanen*, *Tjong Kim Sang*. This algorithm gives very good performance for generalised author name detection and also for the specific case of compound author names; for a detailed presentation of the algorithm and evaluation of its performance, the reader is referred to our earlier work [8].

A fortunate side-effect of using the references section to perform named entity recognition is that each citation is at the same time matched to its corresponding reference, and also author names in references can be tagged with high reliability. This provides the basis for segmenting the references section into individual references.

5.3 Web mining

Acquiring document metadata — effectively, a BibTEX record for each document in the Anthology — requires a different approach, primarily because some key metadata such as the journal or conference name and year of publication is frequently not present in the document itself. Our approach to generating these records was therefore to scrape the ACL Anthology web site. We searched for the HTML anchors and associated text that pointed to the PDF documents. The text surrounding these anchors was then segmented into BibTEX fields, and a unique identifier for the document generated based on the linked filename. This identifier was used to associate the record with the citation and reference metadata extracted in Phase 1, and also with the corresponding PDF file.

5.4 Document interlinking

Given the citation data from Phase 1 and the document metadata from Phase 2, we now had all the necessary data to link citations, via references, to their target documents. For each reference, candidate matches from the corpus were found based on the author list and year. These initial matches, however, may not represent the cited document: an author may publish more than one paper in a given year, and even when there is only a single match from the Anthology, the actual citation may be to another paper not present in the corpus. To perform reliable matching, document titles were matched between the reference from the document and the metadata

record from the candidate target. A fuzzy matching algorithm was employed, checking for content words (and excluding common words such as determiners and prepositions), and assigning a match if the proportion of content words from the reference and BibTEX title exceeds a threshold of 0.75. This accounted for variations in article titles, typographical errors, and extraction errors, while still allowing us to match with high accuracy.

6. PRODUCING A REFERENCE CORPUS

The initial aim of this project was to produce citation and reference data that would form the basis of our research into the analysis of citing sentences. However, a broader aim was to produce a corpus that could be readily used by other researchers in natural language processing, and that also could be used to produce new and richer digital library representations of the ACL Anthology that would provide enhanced navigation and representation of the document links.

The corpus consists of, for each document in the Anthology:

- the source PDF file;
- the extracted text stream, as a UTF-8 encoded text file;
- the extracted stream of tokens with physical page information (as described in Section 4), as an XML file; and
- an XML file containing the document metadata, extracted citations and references, and interlinking data (as described in the previous section).

The major parts of the XML data format used to represent the extracted data are illustrated in Figure 4.

The parallel text stream approach described in Section 4 considerably enhances the utility of this resource. As was the case with our citation and reference extraction system, any information extraction pipeline can operate on the raw text stream without requiring any knowledge of PDF representation. The tokenized text stream containing PDF location information can then be employed at any stage to determine a page location corresponding to an extracted element, even on extraction pipelines designed to work only with raw text.

The next section describes a prototype application which we developed to demonstrate the use of this resource.

7. A PROTOTYPE APPLICATION: DOCUMENT ANNOTATION

Our goal was to validate the utility of the corpus we had produced by generating an enriched document collection based on the intra- and inter- document links. Our goal was to develop a system which would automatically enrich the original PDF documents in the library with in-document annotations representing the citations, references, and inter-document links. When reading a paper, a researcher could mouse over a citation to see the corresponding reference, removing the need to flip forwards to the reference list (at the risk of losing one's place) to see details of a cited work. At the same time, clicking on a citation (or reference) would provide a direct hyperlink to the external document, removing the need to actively search for a cited document. Where the linked-to document existed in the Anthology, that document in turn would be enriched with citation and reference links.

Production of this enriched collection of annotated original documents was facilitated by the parallel text stream approach. The position in the raw text stream of each citation was used to obtain a corresponding series of tokens from the tokenized representation of the file; these tokens were then used to calculate the geometric bounds on the page of the citation. This gave an anchor to which the pop-up annotation containing the reference could be attached, using the API provided by PDFBox. The reference text to display in the popup was itself extracted from the text stream using the supplied positions.

Interdocument links were constructed using one level of indirection. Rather than pointing directly to the target document, a script was deployed at a known URL which translates a document ID into a document location. This allows us to direct a link to an enriched document from our library where it exists, or to an external source where it does not. It also allows us to expand the collection of documents to which enriched documents can point without needing to re-annotate the source documents.

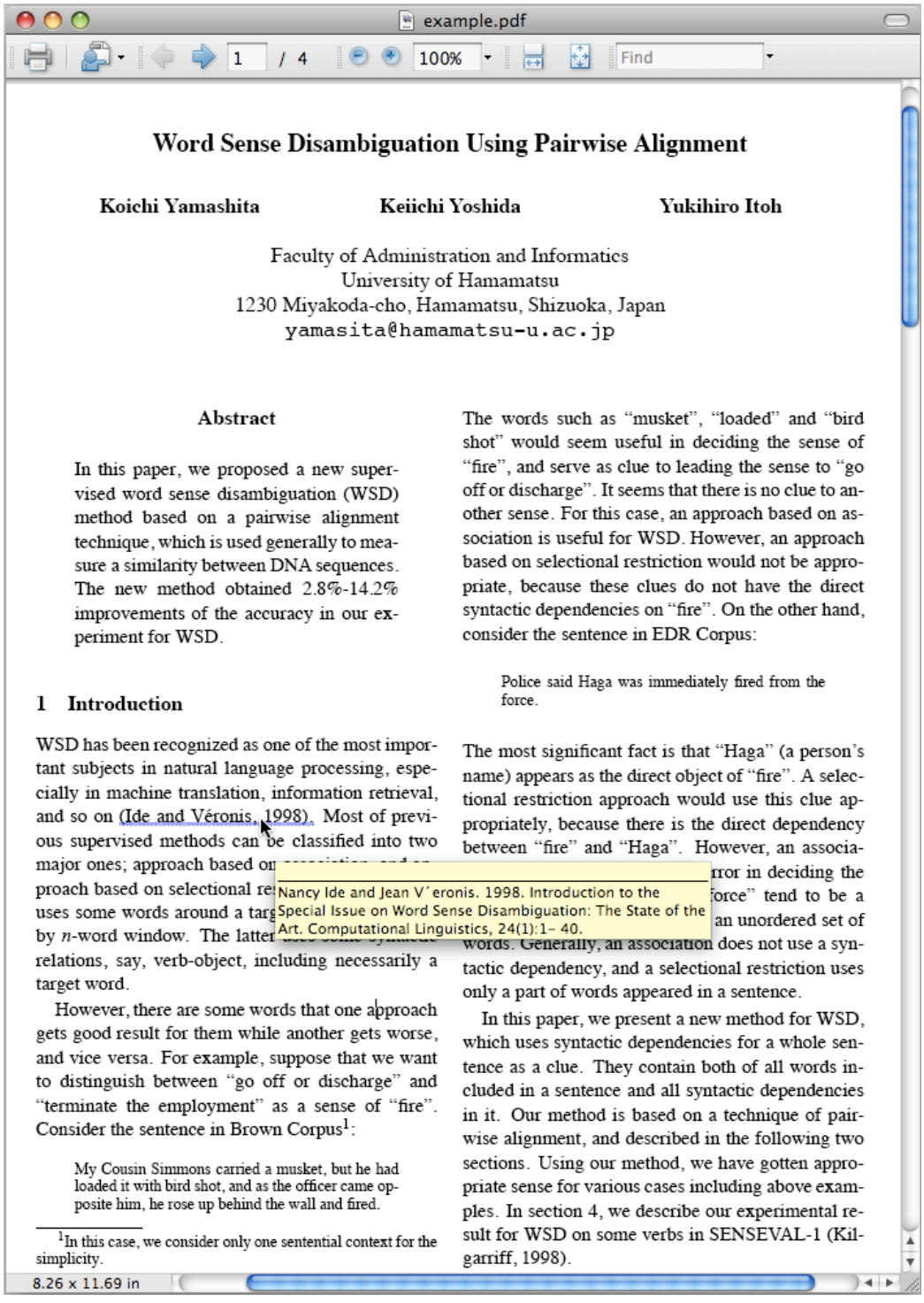
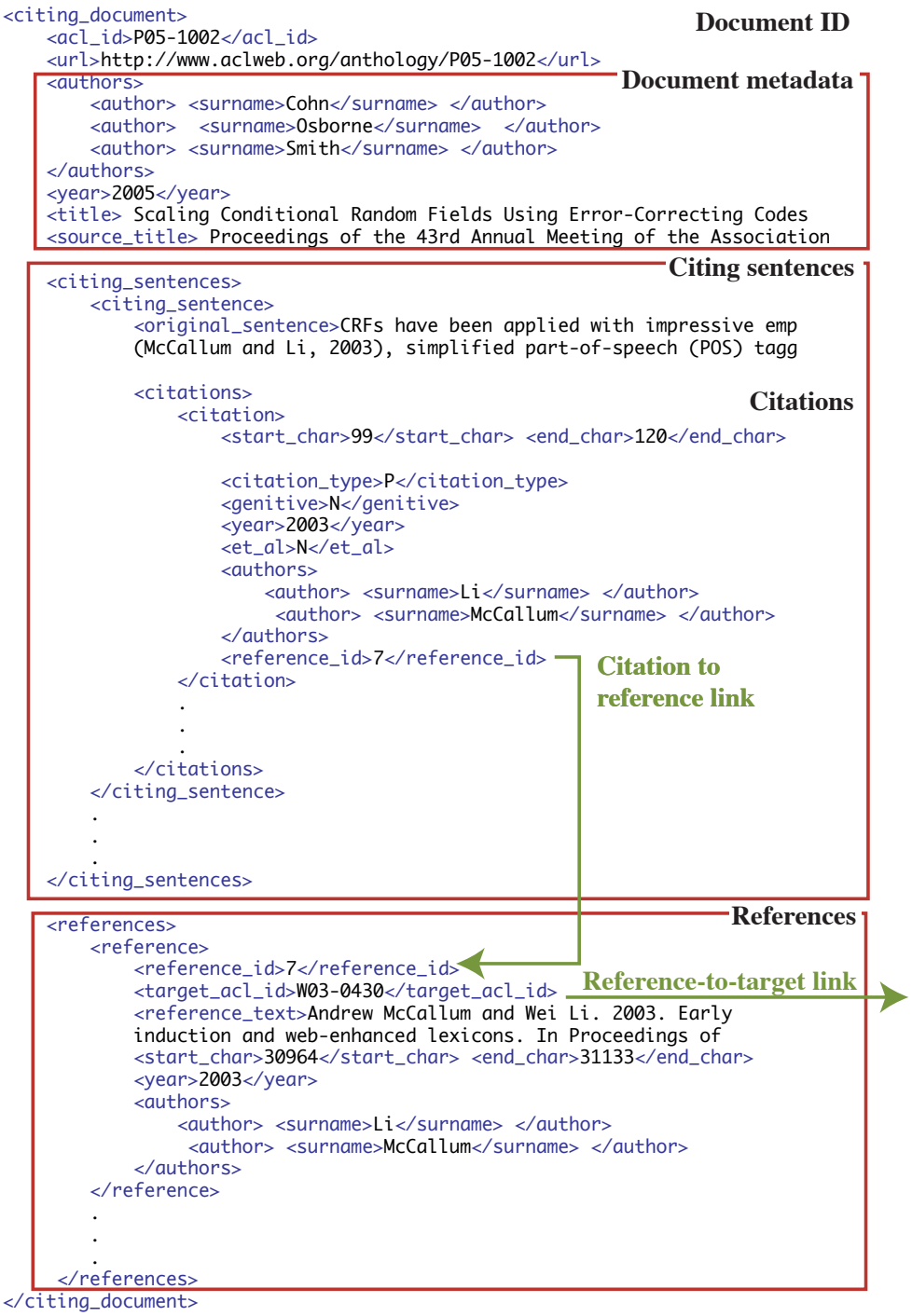


Figure 3. An annnotated paper, showing a citation and its corresponding reference.




```
<citing_document>
    <acl_id>P05-1002</acl_id>
    <url>http://www.aclweb.org/anthology/P05-1002</url>
    <authors>
        <author> <surname>Cohn</surname> </author>
        <author>  <surname>Osborne</surname>  </author>
        <author> <surname>Smith</surname> </author>
    </authors>
    <year>2005</year>
    <title> Scaling Conditional Random Fields Using Error-Correcting Codes
    <source_title> Proceedings of the 43rd Annual Meeting of the Association

    <citing_sentences>
        <citing_sentence>
            <original_sentence>CRFs have been applied with impressive emp
            (McCallum and Li, 2003), simplified part-of-speech (POS) tagg

            <citations>
                <citation>
                    <start_char>99</start_char> <end_char>120</end_char>

                    <citation_type>P</citation_type>
                    <genitive>N</genitive>
                    <year>2003</year>
                    <et_al>N</et_al>
                    <authors>
                        <author> <surname>Li</surname> </author>
                         <author> <surname>McCallum</surname> </author>
                    </authors>
                    <reference_id>7</reference_id>
                </citation>
                .
                .
                .
            </citations>
        </citing_sentence>
        .
        .
        .
    </citing_sentences>

    <references>
        <reference>
            <reference_id>7</reference_id>
            <target_acl_id>W03-0430</target_acl_id>
            <reference_text>Andrew McCallum and Wei Li. 2003. Early
            induction and web-enhanced lexicons. In Proceedings of
            <start_char>30964</start_char> <end_char>31133</end_char>
            <year>2003</year>
            <authors>
                <author> <surname>Li</surname> </author>
                 <author> <surname>McCallum</surname> </author>
            </authors>
        </reference>
        .
        .
        .
     </references>
</citing_document>
```

Figure 4. XML data format depicting intra- and inter- document links.

An example of an annotated PDF paper from the Anthology is shown in Figure 3. A significant advantage of this approach over the web-based systems mentioned in Section 2 is that the annotations are available in the document itself, with the consequence that they can be accessed in the normal course of reading an article. It also means that the enriched data does not rely on a web connection, but is available for offline reading.

8. CONCLUSIONS AND FUTURE WORK

Our approach to enriching a document collection using citation and reference links has proven particularly successful. The parallel text stream approach shows that sophisticated annotation of PDF documents is possible even when employing information extraction pipelines designed to work only on text streams.

The prototype application provides rich annotations within documents, delivering a valuable aid to navigating the literature, and allowing enriched citation information to be accessed naturally while reading a paper. The application of our technique to a large collection of legacy documents (for which page location information was not always directly available) demonstrates that enrichment even of older documents in large collections is feasible. The application of the approach to the ACL Anthology, as described here, is in the final stages of tidying up and will be released to the public by the time of the conference.

Our current work has focussed on a self-contained collection of documents, providing links between documents in the collection. A natural extension of this work is linking to documents outside the corpus, automatically locating cited documents on the web and in other online repositories. One can imagine a range of possibilities that would make the in-document information even more useful; for example, a straightforward extension would be to populate the popups that appear on citations with the abstracts of the cited works. It is also possible to conceive of in-document user interfaces to even richer data about cited documents, such as automatically extracted abstracts or even automatically generated summaries, with the information that appears on a citation being derived from the target document on a citation-by-citation basis by taking into account the context surrounding the citations.

References

[1] Donna Bergmark. Automatic extraction of reference linking information from online documents. Technical Report CSTR2000-1821, Cornell Digital Library Research Group, 2000.

[2] Donna Bergmark, Paradee Phempoonpanich, and Shumin Zhao. Scraping the ACM digital library. *SIGIR Forum*, 35(2):1–7, 2001.

[3] S Bird, R Dale, B J Dorr, B Gibson, M T Joseph, M-Y Kan, D Lee, B Powley, D Radev, and Y F Tan. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08)*, Marrakech, Morocco, May 2008.

[4] Vannevar Bush. As we may think. *The Atlantic Monthly*, pages 101–108, July 1945.

[5] Eugene Garfield. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159):108–111, July 1955.

[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, 1998. ACM Press. ISBN 0-89791-9653.

[7] Brett Powley and Robert Dale. Evidence-based information extraction for high-accuracy citation extraction and author name recognition. In *Proceedings of the 8th RIAO International Conference on Large-Scale Semantic Access to Content*, Pittsburgh, PA, 2007.

[8] Brett Powley and Robert Dale. High accuracy citation extraction and named entity recognition for a heterogeneous corpus of academic papers. In *Proceedings of the 2007 IEEE International Conference on Natural Language Processing and Knowledge Engineering*, Beijing, China, 2007.