Building an autonomous citation index for grey literature: the Economics working papers case*

José Manuel Barrueco
Universitat de València
Spain

Thomas Krichel
Long Island University
New York, USA

December 15, 2004

Abstract

This paper describes an autonomous citation index named CitEc that has been developed by the authors. The system has been tested using a particular type of grey literature: working papers available in the RePEc (Research Papers in Economics) digital library. Both its architecture and performance are analysed in order to determine if the system has the quality required to be used for information retrieval and for the extraction of bibliometric indicators.

1 Introduction

The main characteristic that differentiates the scientific literature from other literary representations is the relationship between documents established trough citations and bibliographic references. The scholarly work can't exist on its own. It must always be related to documents in the same subject area that have been published earlier on. In this way we can see the literary corpus as a complex semantic network. In that network, the vertices are documents and the edges are citations and references.

It is important to differentiate between citations and references. Citations are referrals that a scientific work receives from other documents published later on. References are referrals that one document makes to other works published before.

*We are grateful to S. Lawrence for given us access to the CiteSeer code. Some parts of the system described here are based in its algorithms.

In the 1960s Eugene Garfield developed the first tool devoted to the representation of relationships between scientific documents: the Science Citation Index. Since then, citation indexes have become an important study tools in some areas. In Scientometrics, citation indexes have become an essential tool for the evaluation of scientific activity. In Information Science researchers have studied the possibility of browsing the scientific literature using references and citations. In this way, once an interesting document has been found, it would be possible to use its references to find similar ones.

Compiling large scale citation indexes for printed literature, using human labour, has been an expensive task. In the past only the ISI (Institute for Scientific Information) has carried out this type of work. However, nowadays all scientific documents are generated in electronic form. If they are available on the Internet this allows the possibility of extracting the references automatically. The references of a scientific paper identify the cited documents and create the appropriate links if they are available in electronic format. With such system the costs would be dramatically reduced and new indexes covering new document types (i. e. grey literature) could arise.

The pioneers in this research area were Steven Lawrence and C. Lee Giles with the CiteSeer autonomous citation index (ACI) for Computer Science. They define an ACI as a system which "can automatically create a citation index from literature in electronic format. Such a system can autonomously locate articles, extract citations, identify citations to the same article that occur in different formats, and identify the context of citations in the body of articles. The viability of ACI depends on the ability to perform these functions accurately" (Lawrence, Bollacker, and Giles 1999). In this paper we describe a similar system called Citations in Economics (CitEc). This system uses CiteSeer technology to automatically build a citation index for documents contained in the RePEc (Research Papers in Economics) digital library.

The remainder of this paper is organised as follows. Section 2 describes the RePEc data set which has been used as test bed for the citation index that we have developed. Section 3 describes the CitEc architecture. Section 4 is devoted to the analysis of the system performance in order to determine whether it could be used to extract bibliometric indicators. Otherwise it would be limited to information retrieval. Section 5 concludes the paper.

2 RePEc: a digital library for Economics

RePEc (Research Papers in Economics) is the largest decentralised non-commercial academic digital library in the world. Its home page lives at `http://repec.org`. A gateway that is compatible with the OAI-PMH (Open Archives Initiative, Protocol for Metadata Harvesting) [1] is available at `http://oai.repec.openlib.org`. RePEc describes two types of documents: grey literature, namely working papers, and articles published in peer-reviewed journals. In November 2004 there were 140.000 working papers and 144.000 articles. RePEc is based on a distributed architecture where research institutions worldwide share information about the documents they publish. The details are contained in two documents: the Guilford Protocol (Krichel a) and ReDIF (Research Documents Information Format) (Krichel b).

The Guilford Protocol (GP), named after the town where it was created, is a set of technical requirements that an institution should accomplish to become a member of the RePEc community. It covers only institutional collaboration, i.e. individuals can not join RePEc. There are two ways for an institution to participate in RePEc: archives or services. Archives collaborate by providing bibliographic information about the documents their institution publishes. They also may provide the full-text of these papers. Technically an archive is a space in the hard disk reachable by an HTTP or FTP server. There, files containing bibliographic information are stored. The structure of this space is defined in the GP.

The second pillar of RePEc is ReDIF. All data transmitted between archives and services is encoded in a specific bibliographic format named ReDIF. ReDIF was created to meet the RePEc needs and therefore it is not aimed to become a widely used format for interchange of bibliographic data between libraries. Its main characteristic is that it is simple enough to be used by non-technical people outside of the library world. This is because the most archives are maintained by administrative staff or academics without knowledge of library procedures.

ReDIF allows to describe working papers, articles in journals, physical persons and software compo-

[1] `http://www.openarchives.org`

nents. Each object is described by a template made up of several fields like a traditional database record. They use an *attribute: value* syntax. Fields can be optional or mandatory. The main mandatory field is the Handle which holds a code that identifies the object within the RePEc dataset.

RePEc was created in 1997 from a collaboration of several projects working on electronic document dissemination in the discipline. Since then the number of institutions collaborating as archives has been increasing. At the time of writing, in November 2004, there are 413 archives. The number of documents by archive depends on the kind of institution it belongs to. For example, the prestigious NBER (National Bureau of Economic Research, USA) provides a very large archive describing all the 10897 papers it has published.

Metadata as it appears in the archives is of little utility for researchers. Further processing is needed to take the information and present it in a user-friendly way. This is the objective of the user services. User services are the main way a user works with RePEc. A complete list of user services can be found at the RePEc home page. They add value to the data provided by the archives in several ways:

- Scanning for new data and creating an awareness system to announce new additions.
- Creating a searchable database of papers.
- Creating some type of filtering service to help the user in the selection of the most relevant documents.

Building a citation index allows to create an additional user service that has citations as its prime focus. This is done in the CitEc project. Its home page can be found at: `http://netec.ier.hit-ac.jp/CitEc`.

3 Citations in Economics architecture

The CitEc architecture is based in two main elements as it is shown in Figure 1. First, we have a knowledge base [2] where all authoritative metadata about RePEc documents is stored. This base

[2] The precise detail of this base is beyond the scope of this paper.

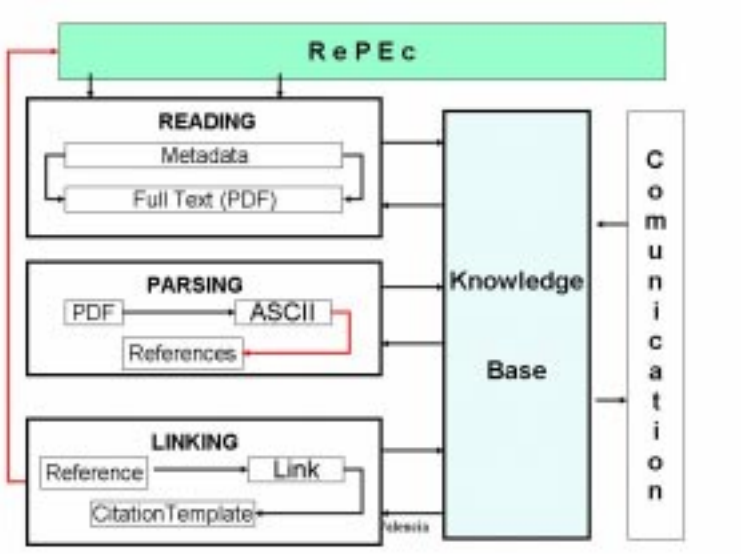

Figure 1: CitEc Architecture

represents the main improvement we have implemented in the CiteSeer software. The quality of the bibliographic references provided in the papers is variable. For instance, it is usual to find different forms for the same author names, journal titles, etc. We use the knowledge base to complete and improve the references quality with metadata provided by the publishing institutions. Secondly, we have a series of three software modules, one for each step in the reference linking process:

1. **collecting** metadata and documents' full text.
2. **parsing** of documents in order to find the references section, to identify each reference and to extract their elements (authors, title, etc.).
3. **linking** of references with the document full text they represent if available on RePEc. [3]

It is important to note that each module is based on the output of the previous one. In this way, the successful processing of each document implies to successfully surpass the sequence of three levels.

[3] Linking of documents out of our data set, using technologies based in DOI identifiers, in the same way it is done in CrossRef at the moment, is in our to-do list.

3.1 Collecting

Collecting involves three different steps: (1) to collect the documents' metadata, (2) to download the documents' full text and (3) to convert them to a parseable form.

The metadata quality varies from archives. There are archives that provide very complete records for each paper, including abstracts, JEL (Journal of Economics Literature Classification) codes, etc. On the other hand, other archives may only provide titles and authors. There are three main problems with the metadata that seriously affects the processing of papers:

1. The absence of publication dates. This field is optional in ReDIF and some archive maintainers don't use it. In our research this data is fundamental because the publication year is one of the attributes we use to check whether a citation goes to a RePEc document or not. Fortunately the publication year usually forms part of the working paper number. Most series are numbered like: 9901, 9902 .... Taking advantage of this convention, we have developed procedures that guess the year from the paper numbers.

2. The format in which the author names are written. ReDIF requires that each name be placed in a different field but some archive maintainers write all authors in a single field, separated by some punctuation. We have also developed procedures to cope with such problems and to correct them as far as possible.

3. Wrong URLs. The URLs provided by the archives to retrieve the documents full text are incorrect. This is a rare but serious problem. If we can not access the paper is not possible to circumvent the problem as we have done with the other cases.

We are working with a distributed library of metadata. There is no a single place where all full text documents live. They are dispersed in multiple servers from multiple institutions. Therefore the second step is to download to our hard disk those documents that are available in full text. This is done by going trough each archive, reading the bibliographic information and, if a File-URL field is found, retrieving the resource contained in the URL. Usually such resource will be the document

itself, but in some cases archive maintainers could point the URLs to abstract pages. In these cases the paper will be discarded.

Once the document is saved in our hard disk, we start the conversion process. First, we check if the full text file is compressed. If that is the case, a decompression algorithm is used. Second, we check the file format. Only PDF and PostScript documents are accepted at the moment. Fortunately both are quite popular formats in Economics. More than 95% of the RePEc documents are in either PostScript or PDF.

The last step is to convert the document from PDF or PostScript to ASCII. For this purpose, we use the software **pstotext** developed by Andrew Birrell and Paul McJones as part of the Virtual Paper[4] project.

3.2 Parsing

Parsing is the most complicated process. Authors usually construct references in a variety of formats, even within the same paper. In addition disciplines vary with respect to the traditions in the way citations are marked in the documents.

Due to the importance of the parsing process we decided to start with a software that has been already tested rather than develop new software from scratch. Our choice has been **CiteSeer** by S. Lawrence, Kurt Bollacker and C. Lee Giles, which has been described in papers like (Lawrence, Bollacker, and Giles 1999).

CiteSeer is able to identify the part of the document containing the list of references. Then it can split the list into different references. Finally it parses each reference to find the elements. At the moment it only identifies the publication year, the title and the authors. However, as we will see, these four elements are enough for our purposes.

[4] `http://www.research.compaq.com/SRC/virtualpaper/home.html`

3.3 Linking

Once we have parsed the documents, the next stage is to look if some of the references successfully found go to documents identified in RePEc. In such cases, some type of link between both documents should be established. We are doing that by comparing each reference successfully parsed, with the authoritative metadata stored in the CitEc knowledge base. At the moment we consider that a reference represents a RePEc document when:

- The parsed reference title and the title in our metadata collection are close enough.
- The publication year of both items is the same.

In this process we take each reference, extract the parsed title and convert it to a normalised version called *key title*. Here all multiple spaces and articles are removed and upper case letters are converted to lower case. Then we select from our knowledge base of metadata all documents that contain in their title all the words of the reference key title. All selected papers are suspect of being the cited document. In a second step we compute the Levenshtein distance of each suspect title with the reference title. If this distance is greather than 8% of the title length, the suspect document is rejected. Finally, we check if the publication year of the suspect papers and the reference is the same. If this is the case we assume that the reference is to the document we have.

4 Internal Evaluation

In this section we provide a detailed description of errors detected in the processing of RePEc documents in order to determine if our autonomous citation index could be used to provide bibliometric indicators or to assist in information retrieval.

In order to evaluate the system behaviour we define a series of stages in the reference extraction and linking process that every paper should pass. In this way, the initial stage for all documents is "notprocessed". It will be changed to the final stage of " linked" for papers which have successful passed all stages in the reference linking process. If the process fails, an error status describing the

Error	*Documents*	*%*
restricted	51418	28%
not found	4221	2%
bad document	7702	4%
available	121111	66%

Table 1: Documents' availability

problem detected is associated with the document. All information about document status and errors is recorded in the knowledge base.

The current version of the system is dated August 2004. It contains 175452 metadata records with information about electronic documents. Such documents are distributed in the 1591 series coming from 378 institutions worldwide contribute to RePEc. The number of documents per institution ranges from those that only provide one or two documents to those with a national scope which provide documents coming from several institutions.

We face three problems when downloading documents. Firstly, there are institutions that charge for access to the full text of their publications. In such cases the documents are simply ruled out. We found 51418 documents with restricted access. That cut down the number of documents to be processed to 124034. The lion share of the restricted documents comes from commercial publishers and the JSTOR project. Secondly, we found in the metadata wrong URLs to the documents' full text. That means the documents are not found at the specified location due to an error in the metadata provided by the archives. Finally, with the error "baddocument" there is a variety of problems. For instance, documents digitalized as images, even using the PDF format, and URLs that instead of pointing to the document's full text go to an abstract page. This practice is not allowed in RePEc but some institutions work in this way to make sure their web sites get as much hits as possible. Table 1 shows the ratio of not available documents.

Once the 121111 available documents have been downloaded the system starts the conversion from the original PDF or PostScript formats to ASCII. In this processing step we found three possible problems: "incompatible format" when the format of the file containing the paper is not PDF or PS, "conversion error" when the program that convert the formats fails or "no references" when even having

Error	*Documents*	*%*
conversion error	10304	9%
no English	2062	2%
no references	24663	22%
incompatible format	13708	12%
converted	61474	55%

Table 2: Errors found in the conversion process

a text version of the paper, the system has been unable to find a references section in the corpus of the document.

61474 documents out of 112111 have been successfully converted to text format. Table 2 shows the error distribution for this particular stage.

At this step it is important to note the large number of documents in which the process of conversion has failed. An initial conclusion to be taked into account in future system updates would be the need of testing new conversion programs.

61474 documents were parsed in order to locate and identify their bibliographic references. Documents in which the number of references identified by the system is greater than seventy are discarded. In such cases is quite probably the process has failed since such a large bibliography is unusual. As it is shown in table 3 almost the 90% of documents stay below the limits.

In total, 1165075 references have been identified in 53201 documents. That represents an average of 22 references by document.

The linking module is the last one in the process. It is in charge of creating a link between each reference correctly parsed and the full text of the document it represents if such document is available in RePEc. 307094 out of the 1165075 references identified are representation of RePEc documents.

In conclusion, 44% of documents available in RePEc were successfully processed. That is, the system was able to extract and link their references. More than half of the documents could not been linked for different reasons. The most important cause of problems is the conversion from PDF to text formats. The second most important is that the system was unable to find the references list in the

Error	*Frequency*	*%*
wrong number of references	2081	13%
correctly parsed	53224	87%

Table 3: Excluded documents due to problems with references

12% of the documents. Since it is unusual to find an scientific document without bibliography, we could conclude that the algorithm of analysis needs to be considerable improved in order to extract bibliometric indicator with enough quality to be used in bibliometric studies.

5 Conclusion

To sum up, along this paper we have described a system that makes possible to automatically extract citation data of documents from a distributed digital library. We have designed a procedure to automatically retrieve the documents' full text from the servers and extract the citation data. Whereas this procedure has been proved successful, a few remarks should be taken into account for future work:

- The collaboration of archives maintainers is a key point to allow a correct administration of the system. Good metadata is essential to obtain relevant results in the citation linking process. The main problem we face is wrong URLs to the documents' full text. This will make impossible to analyse the documents. To solve that we are planing to automatically inform the maintainers about each document that could not be downloaded from their archives.

- Better conversion programs from PDF to text are needed. We work with files generated by a wide range of applications. It is possible even to find scanned documents saved as simple images. As a result we need to use a powerful tool that would allow us to obtain an usable text representation of the document.

- The parsing algorithm could be clearly improved. In our example it was able to parse correctly only the 75% of the references. While this can be an acceptable rate of errors when working with reference linking, it is not enough to create more complicated applications like the bibliometric analysis of a discipline.

References

Krichel, T. Guildford protocol. ftp://netec.mcc.ac.uk/pub/NetEc/RePEc/all/root/docu/guilp.html.

Krichel, T. Redif (research documents information format). ftp://netec.mcc.ac.uk/pub/NetEc/RePEc/all/root/docu/redif_1.html.

Lawrence, S., K. Bollacker, and C. L. Giles (1999). Indexing and retrieval of scientific literature. In *Eighth International Conference on Information and Knowledge Management, CIKM99*, pp. 139–146.