CitEc: an Autonomous Citation Index for Economics

Thomas Krichel and Steve Lawrence

29 October 1999

1 Introduction

This proposal is submitted to the Library and Information Commission as a response to its Digital Libraries Research Call. The proposal is available online at http://openlib.org/home/krichel/CitEc.html. We are grateful to Nir Dagan for comments on an earlier draft. We welcome more feedback.

The broad purpose of the proposal is to build an autonomous citation index for a large collection of scientific documents. This collection is the RePEc dataset of Economics research papers coordinated by the first applicant. In Section 2, we introduce that collection. The autonomous citation index will be prepared by the CiteSeer software, written by the second applicant, as introduced in Section 3. Section 4 details the aims of the applicants. In Section 5 deals with the management of the project. Section 6 describes the deliverables and sets out the timetable to achieve them. Section 7 concludes the application.

2 RePEc

RePEc is a decentralized collection of metadata about research in Economics. At the time of writing, these metadata mainly refer to working papers, i.e., to accounts of recent research results prior to formal publication. These bibliographic data are held on digital archives based on public access computer systems. They are produced by academics for academics; since academic economists do not usually conduct metadata collection, RePEc relies on two principles to lighten the burden of effort and obviate the need for staff dedicated to maintenance of the metadata. First, the decentralization of RePEc's archival materials implies that the collaboration of many academics and many support staff in university departments and research institutions allows the workload to be spread widely, and minimized for each collaborator. Second, a heavy reliance on intelligent technology transforms bibliographic information produced at a wide number of sites ("RePEc archives") into a single, searchable, virtual collection of metadata, automatically performing the syntactical analysis, merging, indexing, and validation needed to update the collection on a daily basis.

In October 1999, there are over 100 RePEc archives, which provide over 16,000 electronic research documents as well as bibliographic data for an additional 60,000 printed documents. After the arXiv.org collection based at Los Alamos National Laboratory RePEc is the second largest collection of free electronic research documents in the world. The crucial difference with arXiv.org is that the collection is sustainable without external funding, because RePEc archives update the data themselves. The total cost of the collection is sufficiently well spread as to ensure that it can be absorbed within each institution.

RePEc has been adopted by four of the five most influential providers of working papers. These are the National Bureau of Economic Research (NBER) and the Federal Reserve System in the US, the Centre for Economic Policy Research (CEPR) in the UK, and the OECD. The last of the five major providers, the International Monetary Fund, are working on an archive. Since the leading providers have adopted RePEc, many small departments have. They all view RePEc as an important tool for the dissemination of information about their work.

The data held in the RePEc archive are simple ASCII templates following a format called "ReDIF". The end user does not access the data in that form. RePEc relies on outside contributors to use the data for user services. User services operate on computer systems that maintain regularly updated copies of remote RePEc archives. This process is known as "mirroring" in the Internet jargon. A central archive provides free mirroring software that makes this process technically quite straightforward. The RePEc data are therefore readily available to any third party who wishes to implement and offer new user services.

Shortly after the foundation of RePEc in May 1997, several user services appeared. "IDEAS" is a set of web pages for all documents and software components in the RePEc dataset, updated daily and searchable with eXcite. "NEP: New Economics Papers" is a current awareness service. It is operated by volunteers who filter data on new additions to RePEc into over 40 subject-based reports that are distributed via electronic mail. The "DECOMATE Working Papers & Research Memoranda" provide a Z39.50 service for the electronic papers in the collection.

The oldest user services using RePEc data are

BibEc and WoPEc. They offer web sites with static pages of the printed papers and the electronic papers contained in RePEc, respectively.[1] BibEc and WoPEc offer a WAIS full text index, ROADS whois++ servers and an mySQL database. Both projects were founding fathers of RePEc. They still contribute to RePEc by running RePEc archives, and they use the data that is provided by other archives.

BibEc and WoPEc are parts of NetEc. NetEc, founded by Thomas Krichel in 1993, is a collection of free online services for economists, offered simultaneously on sites in the United States, the United Kingdom, and Japan. These NetEc web sites are already well established and widely used.

The latest contribution that NetEc has made to the development of RePEc is the HoPEc project. This project allows authors of papers in RePEc to register. Once an author is registered, (s)he can declare which papers (s)he has written. It is then possible—for example—to link from a web page that describes a paper to the homepage of its author, even though the location of that homepage may change over time.

HoPEc opened in October 1999 and for the moment only a few hundred authors are registered. We expect substantial growth in that area. An older service, the "Economics Departments, Institutes and Research Centers in the World" (EDIRC) project has registered over 4000 institutions. The hope is that eventually there will be a relational database that will comprise all researchers, all research institutions they are based at and all papers in the discipline. Clearly reaching this objective will require a lot of time and dedication from the community. But we are clearly on the right path.

3 Autonomous citation indexing

References contained in academic papers are used to give credit to previous work in the literature and provide a link between the "citing" and "cited" papers. A citation index (Garfield 1979) indexes the citations that an paper makes. It links the papers with the cited works. Citation indexes were originally designed mainly for information retrieval (Garfield 1994a). A citation index allows to navigate the literature in unique ways. Papers can be located independent of language, words in the title, or keywords. The index allows navigation backward in time (the list of cited papers) and forward in time (which subsequent papers cite the current paper?). Citation indexes can be used in many ways, e.g.

- citations can help to find other publications which may be of interest,
- the context of citations in citing publications may be helpful in judging the important contributions of a cited paper and the usefulness of a paper for a given query (Garfield 1994a) and (Salton 1971)
- a citation index allows finding out where and how often a particular paper is cited in the literature, thus providing an indication of the importance of the paper,
- a citation index can provide detailed analyses of research trends and identify emerging areas of science. (Garfield 1994b).

Cameron proposed a universal bibliographic and citation database which would link every scholarly work ever written (Cameron 1997). He describes 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. However his proposal requires authors or institutions to provide citation information in a specific format. This is costly to do in the future and very difficult to do on past material. It is not surprising that his proposal has not yet been implemented.

Since 1997, Dr. Steve Lawrence and his team at the NEC Research Institute in Princeton (New Jersey, USA) have been working on CiteSeer. CiteSeer is an autonomous citation indexing software. That means that it can automatically deal with citations as they appear in electronic documents. It is not 100% reliable, but neither are human-created indexes. CiteSeer can deal with PostSript and PDF files. These are the most commonly used formats for Economics papers because these papers tend to make heavy use of mathematical formulae that are difficult to encode in a HTML document.

CiteSeer can identify citations within papers. That means it can very reliably detect that two citations, of different formats, refer to the same paper. This allows a detailed listing of a cited paper to show all instances of the citation across multiple papers. From those listings we can create statistics on citation frequency that allow for a rough estimate of the importance of a paper.

CiteSeer also looks at the citation tag. The citation tag is the information in the citation that is used to cite that paper in the body of the document (e.g. "[6]", "[Giles97]", "Marr 1982"). The citation tags are used to find the locations in the document body where the citations are actually made, allowing CiteSeer to extract the context of these citations. This context may contain a brief summary of the cited paper, another author's response to the cited paper, or subsequent work

[1] The division into electronic papers (WoPEc) and printed papers (BibEc) has historical reasons: when these sites were founded, the number of electronic papers was tiny and none would have been found in most searches on the combined dataset.

which builds upon the cited paper. Any of these elements is of crucial importance to evaluate the cited paper.

CiteSeer can identify citations to the same paper and the context of the citation with a very high degree of accuracy. But its work does not stop there. It also tries to identify subfields of the citation, e.g. the author name, the title, the publication outlet etc. The heuristics used depart from an "invariants first" philosophy. That is, subfields of a citation which have 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. 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, and publisher almost always comes after the title. Using these heuristics CiteSeer has been able to achieve reasonably good results for extracting certain subfields (Lawrence, Giles, and Bollacker 1999).

When CiteSeer is launched on an individual document, the bibliographic data of that *citing* paper to be indexed (title, author, author affiliation, addresses, publication details etc.) may not be easily identified. There just too many ways in which that information is spelled out it the paper. Therefore it will be highly desirable to use CiteSeer on a collection of documents for which a large bibliographic data of which is already available. RePEc provides such a collection of documents.

4 Project aims

We have a scenario where we possess three elements

1. a large set of full text research papers in one subject area
2. an authoritative bibliography of these papers
3. a well used web service that allows to access the papers

We think that adding the citation information as a fourth component will allow to critically augment the services that we offer.

4.1 Complete citation analysis

Clearly the first step is to perform a complete analysis of the citations that are contained in the RePEc papers. This will allow to find out which papers are the most cited. The results will be published on a web page the URL of which will be circulated on the most important lists that are populated by economists. We believe that this is an important step. The publication of a ranking list of most cited papers will cause a sensation in the Economics profession. It will be excellent advertisement for the RePEc dataset and CiteSeer.

4.2 "Usual suspects" extensions

There are two ways in which the RePEc bibliography is useful to enhance an autonomous citation index. First it provides authoritative data on the citing paper. Second it provides for a list of candidate values that is useful for the identification of subfields in the data.

The most important subfield for us is the author subfield. In October 1999, the RePEc dataset covers over 120000 author names fields. These contain about 46000 author names. Not all of these author names correspond to physically different persons. Nevertheless this is an important dataset that will be used to refine the identification of the author name subfield in the citation. The project will deliver software that allows to check a "usual suspects" list of authors when the CiteSeer tries to detect authors names within citation fields. The way that such an author list should be constructed for the author subfield detection to be done in an efficient way is one of the research issues of the project.

In the same way, we intend to use a well-know list of Economics journals and of course RePEc's own list of publication outlets—these are mainly working paper series—to identify the publication channels. We can then compute the impact factor of each channels.

When this software is written, we intend to publish summary statistics of the cited authors and cited publication channels. However, we will not release the detailed data unless the author is registered with HoPEc or the publication channel is providing authoritative data to RePEc. This should act as an incentive for suppliers to deliver more data to RePEc.

4.3 Internal citation

In the first stage we have dealt with citations to external papers, i.e. papers that are not described in RePEc. In a second stage we will seek to map citation data internally. That means that we will seek to identify which paper in RePEc cites which other paper in RePEc.

This problem is by no means straightforward. One important research concern arises in the specific context of a preprint collection. Since most of the papers that we hold are later on published in printed journals, citations of these papers are likely to go to the journal version of the paper. We will have to consider that the two papers are the same if there is a fuzzy match

between author and title only, and that they are even more the same if they are in the same publication outlet. This distinction is at the heart of debates in the digital library community about the "sameness" of publications in different channels (Krichel and Laypunov 1999).

The internal citations will be represented in the ReDIF metadata format. The project will be working on extensions of the format to incorporate citations and citation context. Some special software will be written to allow gather, for each paper, the papers that it is citing and the papers that it is cited by. Special care will be taken to distinguish which type of manifestation ("working paper", "conference paper", "published article") the citation uses. How to make that distinction clear to the user will be a matter to the user services.

4.4 Citations of an author

The ReDIF citation data that is introduced in the previous section will then be matched to the HoPEc author database. The way that information will be encoded in the RePEc dataset has yet to be worked out. We will also have to develop a strategy if and how to involve the registered authors in the correction of the citation data. Will registered authors be able to provide us with data that they have been cited by another author? Will they be able to remove citations data that they know/think is incorrect? These and other question will be an interesting challenge from a conceptual as well as from a software engineering point of view.

For each author that is registered in the HoPEc service, we have precise first and family name information. This very important detail will make the identification of authors in the citations data much easier. However, we will need to extend the CiteSeer software so that it takes account of that extra information.

4.5 Citation as a review

One of the main problems that we have is that the RePEc data is only very weakly peer reviewed. The fact that the papers appear on institutional rather than personal or public archives imply a simple form of peer review. However if the papers that are catalogued by RePEc could be subjected to increased peer scrutiny, the incentives for authors and their institutions to participate in the scheme will be much enhanced.

There are two approaches that WoPEc have taken in the past to express some form of peer review. First WoPEc offered users the opportunity to comment on the paper. This was an embarrassing failure. Over two years, we had about 60 comments submitted, only one of them had any editorial value. The others were gripes from people who could not access the full text over the network or who had questions about how to print the file.

Another possibility are download statistics. However these statistics do not measure the scientific importance of a paper as an intellectual contribution. Many of the users of WoPEc are students, and therefore we expect a bias towards applied papers and towards survey papers. In addition, there is a problem with authors potentially downloading their own papers many times to improve their ratings.

With the advent of autonomous citation indexing, we have the possibility to include the citation context as a comment by the citing paper on the cited paper. This is a genuine instance of peer review. For each paper within our system we will investigate if it cites another paper within RePEc. If it does then we will add the citation context to the metadata of the cited paper. This will give the user valuable information about the cited paper. In the metadata of the citing paper we will add a link to the cited paper, but not list the citation context.

4.6 Finding related papers

Two papers that cite the same earlier work must be related. If two papers cite many papers in common, they must be strongly related. One of the aims of this project is to find and calculate a measure of relation that comes from citation. There are theoretical foundations of these measures through the so-called Erdos numbers. However it would be beyond the scope of this application to detail this background here. An important feature for the WoPEc user service that presents RePEc to the user will be to provide links to related papers. We will use the full (internal and external) citations data as a basis for calculating the relationships.

5 Project management

There are three parties to the project. These are the software contractor (to be nominated), the project advisor (Steve Lawrence), and the project leader (Thomas Krichel).

5.1 The contractor

The contractor will write the software that is required by the project. He[2] will most likely be based in Russia, because it is simply impossible to get the level of computing expertise that is required for the project on the British labour market at anywhere near the cost that we have budgeted. The budgeted cost is about the cost of a full time programmer in Russia. We require the

[2] We use the male form only here for simplification.

programmer to spend 50% of his time on the project. That means that we are willing to offer twice the local wage to hire somebody who is really competent.

This scheme may seem adventurous, but it has worked out well in the past. Much of the software that is used by RePEc and its associated projects has been written in Russia and is maintained there. The only problem is that supervision of such remote work is more time-intensive than the supervision of work that is produced by a local contractor simply because all the communications have to be conducted via email. In the specific case of this project this is less of a problem since one party is based in the United States anyway and there is no face to face meeting planned between any of the project partners.

The software contractor will keep a small sample of the data on a local machine but the services will be produced on the machine that the project will own at the University of Surrey. That machine will host the deliverables. The contractor will operate this machine over the Internet from Russia. For a whole host of technical and organisational reasons this is quite easy to do if, but only if, we have a dedicated machine that produces the deliverable. We have therefore budgeted for such a machine.

The software written for the project will be of two types. There will be enhancements and bug fixes on the current CiteSeer software ("analysis software") and there will be software to make the analysis software interoperate with the ReDIF bibliographical dataset ("gateway software"). We will come back to this distinction in the following.

5.2 The advisor

Dr. Steve Lawrence will advise the project. He will not receive any financial reward for this activity. His participation will be rewarded in kind by the enhancements that the contractor makes to the analysis software. The enhancements to the analysis software will be the intellectual property of NEC Research Institute. However, NEC Research Institute will grant a royalty-free licence to Thomas Krichel to use the software for non-commercial purposes.

Clearly, the division of the contractor's labour between the writing of analysis and gateway software is a decision that will have to be reviewed from time to time. The applicants are hopeful that the split will be about 50% on each of the two components.

5.3 The leader

Dr. Thomas Krichel will be leading the project. He will have sole responsibility to the funders for the receipt and distribution of funds and the execution of the project according to the timetable. He will hire the contractor. He will own all copyright on all intellectual property created by the project apart from the analysis software. He will write all project reports.

The Department of Economics at the University of Surrey has been supporting the activities of Thomas Krichel for the electronic dissemination of research results ever since he became involved in that activity in 1993. This application is particularly interesting because it will deliver a public domain citation index. This index will allow for a quantifiable mapping of the progress of the discipline through citation data. This will yield new insights into the way Economics works.

The department will be releasing Thomas Krichel from his normal duties for 15% of his time to carry out the coordination of the project. The cost of this to the department is £4127 in the first year and £4307 in the second year. It not costed in the funding application. The Department will also donate the institutional overheads to the project. These are 42% of total staff cost as central overheads and 34% of the total staff cost as departmental overheads. Therefore the total contribution of the University to the project is £14843.84, which about 46% of the total cost.

6 Deliverables and timetable

There are 24 months in total that are allocated for the project. The project will consist out of 7 stages. At most stages, a short report will be written by the project leader. At the end all these reports will be combined into a larger report. All reports will be circulated on the WWW and be available as a file to print, just like this application document.

In month 1–4 we will calculate the complete initial citation analysis, with the current version of CiteSeer, without any enhancements. We will publish a league table of the most heavily cited papers.

In month 5, we will publish the enhancement to the ReDIF metadata format that takes account of the citations.

In months 6–10 we will work on the enhancements of the citation matching algorithms of CiteSeer through the bibliographic data in the database. We will write a report on the perceived accuracy improvements, in particular for the author name subfield matching. We will publish league tables for authors and for publication channels.

In month 10–13, we will complete the internal citation data in citation templates, and enhance the WoPEc user service through citation linking. There will be no report needed at that stage because the metadata format of citation is already known since month 5, and the main calculations will have been done in the previous months.

In month 14–15, we will publish an intermediate report that will outline the work in the second part of

the project. In particular we will set out how exactly we are going to estimate the closeness of papers. In particular the report will survey the literature that is relevant to the theory of this problem. We will also document the enhancements that we made to CiteSeer.

In month 15–21, we will work on the relationships between papers. By this time the RePEc dataset will have easily accumulated 30000 papers. That means there are potentially $(30000 \times 29999)/2$ relationships to calculate. Even with an intelligent way to reduce these calculations, this will mean a very large computational job. In computing, size matters a lot. The larger the amount of calculations the more likely it is that a software is running into problems. Therefore we are accounting an important chunk of time to this problem. It is likely that we will have to revise our calculation strategy a few times.

The last three months will be spent on including the citation data into the WoPEc user service. They will also be spent on documenting both software and procedures and on compiling the final report.

7 Conclusions

We fully understand that many a reviewer of this proposal, while appreciating the technical competence of the contents, may wonder: Why bother with all this?

First let us note that both applicants have pioneered in their respective area. The CiteSeer software is the only autonomous citation software in the world. It requires document data that uses the title in the citation. The RePEc collection is the largest bibliography of freely downloadable scientific documents that it can use.[3] The meeting of ambitions in this application is truely unique. Currently all citation indexes have to be compiled by hand. They are therefore very expensive. An improved version of CiteSeer would go a long way towards creating usable indexes that at a tiny fraction of the cost.

This proposal makes no assumption about the behaviour of a third party. Its success is only dependent on the applicants because the input data for the project is already available now. Many digital library research projects construct datasets that are small to start with. Many remain small because they do not pay proper attention to the incentives that need to be created for third parties to feed data into the deliverable. They all too often depart from the assumption that as soon as a convenient digital technology is available users will embrace it with enthusiasm. Our experience is that most users are very slow to change by themselves and many never will. Our project therefore runs on very conservative assumptions about human behaviour.

[3] In Physics citations the title data is not usually present. Therefore the arXive.org is not really suitable for such a project.

This does not mean that we are not interested in cultural change in the user and contributor community. This application strikes at the very heart of the academic business: the recognition by peers. Citations play a crucial rôle there. If we can build citation indexes on the bibliographic data we have a chance to even further expand a system that is already very successful. We can expand it from critical mass gathering to substantial coverage. This will make Economics a model for other disciplines.

An old joke informs us about the difference between the mafioso and the economist. The mafioso makes you an offer you can not refuse. The economist makes you an offer that you can not understand. Economics research has the reputation to be pretty esoteric stuff. That is only partly true. Yes there are a lot of papers written by academics for other academics only. But from the experience that we have in dealing with users, our data is also heavily used by financial institutions and consultancy companies. The financial service sector is very important in the UK economy both internationally and with respect to other domestic sector. Therefore doing work for economists will help to enhance the competitive advantage that the UK has in the sector.

References

Cameron, Robert D. (1997). A universal citation database as a catalyst for reform in scholarly communication. *First Monday 2(4)*.

Garfield, Eugene (1979). *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. Wiley.

Garfield, Eugene (1994a, January 3). The concept of citation indexing: A unique and innovative tool for navigating the research literature. *Current Contents*.

Garfield, Eugene (1994b, January). Where was this paper cited? *Current Contents*.

Krichel, Thomas and Victor M. Laypunov (1999). UPS ReDIF Conversion Report. presented at the first UPS meeting in Santa Fe (New Mexico), October 21–22 1999, available at http://openlib.org/acmes/hist/.papers/ups_redif_conversion_report.a4.pdf.

Lawrence, Steve, C. Lee Giles, and Kurt Bollacker (1999). Digital Libraries and Autonomous Citation Indexing. *IEEE Computer 32(6)*, 66–71.

Salton, Gerard (1971). Automatic indexing using bibliographic citations. *Journal of Documentation 27*, 98–110.