Reference Linking in Economics: the CitEc Project

Abstract: Over the last few years, reference linking has become an active research field in digital libraries. In this paper, we present a software system that gives a solution for solving the automatic allocation reference linking. The system is based on the agent paradigm which has been applied in different approaches, in order to find a solution for the information explosion problem. The agent is called CitEc agent and it has been tested with the documents available on the RePEc data set.

1. Introduction

The relationship between documents established through citations and bibliographic references is a characteristic that differentiates scientific literature from other literary representations. Such a relationship has been investigated from multiple disciplines like scientometrics or information retrieval. The possibility of navigating the scientific literature through such relationships or to automatically access the full text of a cited document by simply clicking on its reference have been dreams for a long time. However, two requisites were needed for such dreams to become reality:

1. documents should be available in electronic format.
2. algorithms should be created to automatically extract the reference data from the document full text.

The first requisite is already achieved because more and more documents are available on the Internet. The second one is an active research field at the moment. Reference linking is the process of automatically linking each work cited in a document with its electronic full text. This is a straightforward process when metadata about references included in a paper is available as a result of the publishing process. Problems arise when it needs to be identified and extracted out of the papers full text. Several software solutions have been created to automatically deal with these problems. The agent described in this paper is one of them.

An increasing number of different applications based on the agent paradigm have appeared over the last few years in order to find a solution for the explosion of information (Julian, 2000). Taking into account other factors, Internet becomes an excellent proving ground for agent and multi-agent development. An agent is defined by its flexibility, which implies that an agent is (Wooldridge, 1995):

- Reactive: an agent must answer to its environment.
- Proactive: an agent has to be able to fulfil his own plans or objectives.
- Social: an agent has to be able to communicate with other agents by means of some kind of language.
The agent based paradigm is a suitable model to support this kind of problems because it allows an automatization of the different involved processes, an autonomy degree in order to act independently of the user, to solve complex and dynamic problems like reference linking in a flexible way and to apply intelligent techniques for locating and retrieving information.

Following this line of research, an approach to a system architecture which is based on the agent paradigm for solving the automatic allocation of reference linking data is presented in this paper. Basically, the reference linking work consists of three parts: reading, parsing and linking of documents. We have developed an agent to accomplish autonomously each one of these steps. So, we present CitEc. CitEc has been developed at the Universidad Politécnica de Valencia with the objective of reference linking documents available on the RePEc (Research Papers in Economics) data set.

RePEc (http://www.repec.org) is a distributed digital library specialised in Economics. Since 1997 is providing the latest research results in Economics. At present it holds bibliographic descriptions of more than 50,000 full text documents (articles published in journals and working papers). Most of them are freely available on the net. RePEc is based in a distributed architecture where institutions worldwide share information about the documents they publish or distribute. For more information (Barrueco, 2000).

The rest of the paper is structured as follows: Section 2 shows the related work, in section 3 we present the architecture of CitEc agent. Section 4 gives the results obtained and, finally, section 5 explains some conclusions.

2. Related work

Reference linking has become over the last few years an active research field in digital libraries. The publishing industry and the academic communities of electronic archives have carried out projects in this area. This section describes some of these initiatives.

OpCit (Open Citation Project) is a three years project funded jointly by the Join Information Systems Committee (JISC, UK) and the National Science Foundation (NSF, USA). One of the OpCit aims is to link the documents distributed through the arXiv preprint archive. This archive is the largest repository of scientific documents available on Internet. At the moment, with more than 150,000 documents freely available, it holds almost half of the literature produced in the field of High Energy Physics.

OpCit has developed software to automatically extract the citation information out of the documents full text and it has created links between cited and citing documents which are available on the archive. A full version of the archive with interconnected documents is running at: http://arabica.ecs.soton.ac.uk/cgi-bin/search_tj. OpCit and RePEc work on completely different environments. While in Economics there is a very large range of documents contributed by different institutions and produced with a variety of software and over different platforms, in the Physics archive almost all files are created using TeX source files submitted by the authors.

In a similar working line the CERN Document Server has reference linked their technical documents. CERN holds more than 170.000 full text documents. This initiative is similar to OpCit in the sense that both are working with
homogeneous documents. The system can be accessed at: http://weblib.cem.ch. They are offering information about the references of each document and about the cites such document has received. By contrast with OpCit, in CERN the new data has been fully integrated in the system that researchers have been using for years. As reported in (Claivaz, 2001) near three million references have been extracted from the processed documents. For 1,937,162 of them the system has been able to create a link between the reference and the document it represents. Links have been established with documents available both in the repository and in journals subscribed by the CERN library.

CrossRef is an initiative carried out by the publishing industry. It is managed by PILA (Publishers International Linking Association) a society made up of more than 80 publishers and abstracting services around the worldwide. It is different of previous initiatives in the sense that it does not need to look into the documents to extract their references. On the contrary, publishers submit to CrossRef the metadata of the documents they publish. In this sense, is a more simple system.

CiteSeer is not just a reference linking software that could be applied to any electronic document in any discipline, it goes further in building true autonomous citation indexes. It has been developed at the NEC Research Department and described in papers like (Lawrence, 1999). A sample CiteSeer database is available at the url: http://csindex.org. It contains more than 200,000 documents with over two millions references. CiteSeer is a free of charge distributed software for non-commercial institutions. It can be used to create repositories of scientific documents and citation indexes following the model of the ISI Citation Indexes. It is able to locate scientific documents on the Web, to download the documents full text in order to extract their list of references, to split each reference in different elements (title, author, etc). It can also determine if two references with different format represent the same document, etc.

CiteSeer is relevant for our purposes because it works in environments similar to RePEc, since it deals with heterogeneous sets of documents that have been created with a variety of software, with different formats and more important of all, with different citation styles. Each discipline has a particular citation style, which is employed by almost all authors. In Physics, for instance, a typical reference is quite simple because the work title is omitted and the source data is abbreviated. It could look like:


Nevertheless in Economics references use to be more complex. Most of them use the Chicago Manual of Style (Chicago, 1993). That makes software developed by CERN or OpCit unusable in our dataset.

3. The CitEc agent
CitEc is an agent designed to autonomously operate on the RePEc universe of data, but it can be applied to other information sources too. It basically has three functions: it reads documents in RePEc, it extracts their citation data and, finally, it links references with the full text of the documents they represent if are available on RePEc. CitEc is structured as different modules with different tasks. Its architecture is presented in the following point.
3.1 Architecture

It can be seen from the chart in figure 1 that CitEc is made up of three components: the processing subsystem, the knowledge base (KB) and the communications interface.

The most important part of CitEc is the processing subsystem. Basically it is in charged of detecting new full text documents, downloading and parsing them in order to find out their lists of references and finally linking the references with the full text of the documents they represent. Each task is executed by one of the following modules: reading, parsing and linking.

The CitEc inputs are templates containing bibliographic information about scientific documents. The CitEc agent processes this data and, as a result, it modifies the RePEc environment by adding a new template type called “citation template” (Barrueco, 2002). There will be a citation template for each document that has been successfully processed. It will include the list of references with the different identified elements of each one. In addition, a link from the cited document will be included if it is available on RePEc data set.

The parsing module is the core part of the agent. It accomplishes three main tasks. Since RePEc holds only bibliographic data with pointers to the full text of the documents, the first operation performed is to download the files containing the full text of the documents. Files in other formats than PDF or PostScript are excluded. This is a little constraint because more than 94% of documents in RePEc are made available in one of these formats. Secondly, it converts the files from the original format to ASCII, so the text can be parsed. CitEc uses pstotext for this operation. By far, this is the most complicated task since there is no a reliable application to convert PDF files to ASCII. See (Robinson, 2001) for a comparison of conversion tools. Finally, the last task of this module is to parse the ASCII file in order to find out if the author has included a list of references. We have used in this task CiteSeer procedures rather than to design new ones. CiteSeer has been probed as a very
successful tool in the citation analysis of heterogeneous sets of documents (Lawrence, 1999). Once the references have been extracted, the module tries to isolate each individual reference and then to parse it in order to find out its elements. The module is able to identify the publication year, authors, title and context in which the citation appears in the body of the document.

Finally, the linking module reads the information about references extracted by the previous module. For each reference found, it tries to identify if the document is available in RePEc. It compares the reference elements against the KB. If it finds a similar document, a link is established between both documents. This module is also in charged of modify the RePEc environment by adding new templates to the data set. Each time a reference is read and processed a new template is created.

The second component of CitEc is the knowledge base. It maintains an internal representation of the RePEc data set. As described in (Barrueco, 2000), RePEc is based on a relational structure that can be easily translated to a SQL database. In this way the KB of CitEc has been developed as a relational database and has been implemented using a MySQL database management system. The KB contains three basic tables: DOCUMENT is a table that holds the basic bibliographic information for each document available in RePEc. REFERENCE is a table that contains a row for each reference extracted of the documents. Finally, CITATION is a table that contains a row for each relation established between cited and citing documents in RePEc. It has two columns, the first one holds the key of the citing document and the second one the key of the cited document.

Finally, the communications interface provides an easy way to interrogate directly the KB from third part applications. This module accepts the methods described by (Bergmark, 2000) in the Cornell reference linking API. It is implemented as a CGI script available at: http://netec.ier.hit-u.ac.jp/adnetec-cgi-bin/get_data.pl. This script requires two arguments. The first argument is a handle of the document which we need to get information of, and the second argument is one action. There are three implemented actions: getMyData returns the document metadata, getReferenceList gets the list of references of the document and getCurrentCitationList returns the list of documents which have cited the current document.

4. Results

CitEc has been on since January 2002. Over this period of time, 28157 documents have been analyzed. 13446 (48%) have been successfully processed. 293069 correct references were extracted out of them. This figure gives us an average of 22.5 references by document. 52352 (18%) references represented a document available in RePEc, therefore a link between both documents was created. Nevertheless, for the purposes of this study is more important to look into the errors.

It can be seen from the chart in figure 2 that the most important cause of error is conversion from PDF/PS to ASCII. Since pstotext has a high rate of errors, more sophisticated conversion tools are required. Nevertheless this point is out of our scope. Secondly, CitEc has been unable to find a list of references in the 12% of documents. This figure is quite high because it is not usual to find scientific documents without a bibliography. We have detected that in some cases the
problem behind this error is that \textit{pstotex} crashes before to finish the process. That produces a correct but uncompleted ASCII file. Since references are usually placed at the end of the paper, CitEc can not find a list of references.

Other problems with lesser incidence are documents with wrong number of references. We consider that the process to isolate references is correct when it finds between 1 and 75 references. If the number is out of these limits the document is discarded. Another cause of errors is documents which have been distributed in other formats than PDF or PS. As we can see this is a minimal part of documents. Moreover, there is a 3\% of documents, which are not in English language. This constraint will

![Figure 2. CitEc processing results](image)

be removed in future versions. Finally, since the public release of CitEc on January, 14th until March, 9th, it has got 4674 queries (77 request by day). At the moment, the results of our work have been implemented in three RePEc services: WoP\textsc{e}c: (http://netec.mcc.ac.uk/WoP\textsc{e}c), Socio\textsc{n}et (http://socionet.ru) and Bib\textsc{s}oc (http://www.uv.es/bibsoc/GM).

5. Conclusions

An approach to a system architecture which is based on the agent paradigm has been presented in this paper. This approach gives a solution for solving the automatic allocation reference linking. CitEc is an agent that automates the process of reference linking between different documents in RePEc. It allows to automatically obtain information about the references of a document in a digital library. CitEc has been tested with a great quantity of documents and it seems to work well to reduce enormously the time needed in this process.

As future works we want to reduce the error percentage of the evaluated documents in the different modules of the CitEc agent. Another direction for future work is to improve the research algorithms of the references.
References


