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

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.

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''. 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 two describes the RePEc data set which has been used as test bed for the citation index that we have developed. Section three describes the CitEc architecture. Section four is devoted to the a alysis 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 five concludes the paper.