Evaluation of INCISO: A system for automatic elaboration of a Citation Index in Social Science Spanish Journal

José Manuel Barrueco Cruz
University of Valencia
barrueco@uv.es

Pedro Blesa
Polytechnic University of Valencia
pblesa@dsic.upv.es

Thomas Krichel
Long Island University, Novosibirsk State University
krichel@openlib.org

Julia Osca-Lluch
University of Valencia-CSIC
m.julia.osca@uv.es

Elena Velasco
Polytechnic University of Valencia
elvear@aaa.upv.es

Abstract

We have developed a system that can elaborate a citation index in an automated way. It has been tested with Spanish journals. We need evaluate our system, mainly in effectiveness of the retrieval of citations. Criteria for evaluation of the system is presented and discussed, and the results of the application to our system are showed and analyzed.

1 Introduction

Citation indexes are key tools for two reasons. Firstly, there are tools for searching the scientific literature. Second, they allow the evaluation of the production scientific papers, because it is commonly accepted that good papers receive more citations than lousy ones. Therefore citations counts are a common approach to evaluating the quality of a research paper.

In practice, compiling a citation index is a complicated exercise. Traditionally there has been one monopoly supplier of such information in the Institute for Scientific Information. Their products are very expensive. They are mainly geared to Anglo-Saxon publications. However, in many countries national publication traditions persist, mainly in the social sciences and the humanities. These can not be evaluated using the indices produced by ISI.

Spain is no exception. There is no tool for the evaluation of research (mainly in Social Sciences and Humanities) published in Spain [6]. The resources required for building general citation indexes by traditional methods are too expensive to be carried out at national level. However, with the generalization of the Internet as a new communication channel, with electronic journals that are proliferating, new avenues become available.

If articles are available in digital formats, there is a possibility for a computer system of extracting the references automatically. In this paper we present research project we have created a citation index by automatic means. We developed a computer system which is able to automatically create citation indexes for Spanish publications. The proposal was funded from the Spanish Ministry of Science and Technology with a research grant for three years starting in July 2005. We named the project INCISO (*in*dice de *ci*encias *so*ciales). [2] is an early account of our efforts.

More formally, INCISO has two main objectives:

1) Design a computer system for the elaboration of a citation index in an automated way. It has been tested with a selection of Spanish journals in Social Sciences.

2) Elaborate and disseminate a citation index for Social Sciences based on a selection of Spanish journals. This index will be available for all the scientific community and it will be freely accessible at the project web site at http://inciso.openlib.org/.

Obviously, the next step is the evaluation of the system. Information systems can be evaluated in term of many criteria, including execution efficiency, storage efficiency, effectiveness of the retrieval, and the features they offer a user. Our interest in INCISO at this moment is not related to system efficiency, our aims focused in effectiveness of the retrieval of citations, and specially recall and precision.

The remaining of this paper is organized as follows. Section two discusses background material. Section 3 reports on methodology and work plan of our project, and the INCISO architecture. Section 4 describes criteria for system evaluation, and its application to our project. Section 5 shows and discusses results obtained. Section 6 concludes the paper.

2 The background to the INCISO project

INCISO is not the first project to try to build an autonomous citation index [7]. In this section we first survey earlier work. We distinguish three types of projects, commercial projects, academic projects, and hybrid projects.

Commercial projects usually sell the data that they produce. The classic commercial projects are the

Proceedings of the 9th Russian Conference on Digital Libraries RCDL'2007, Pereslavl, Russia, 2007

citation indices produced by the Institute for Scientific Information, now owned by Thomson Scientific and integrated into their Web of Science produced. Recently a competitor to this product has emerged in the Scopus product of Elsevier. Both products use a rather narrow list of quality controlled sources. The inclusion of a source in the citation index is an indication of prestige. We are not sure about the actual production methods, but it is generally believed that these commercial product undergo extensive human control and there
One project, Google Scholar, is a hybrid project. It is based like an academic project on a broad collection of documents. It is only loosely quality controlled, and its elaboration is mostly computer-based. It is freely available. An illuminating account of the differences between Google Scholar, Web of Science and Scopus for a test set of documents is found in [4].

Academic project works tend to work with a broad range of documents that are available on the Internet. They analyze the documents' full text in order to extract the references. Then citations can be linked to the documents they represent if they are also available. The data is extracted automatically by computer programmes. The quality of data gathered by such project varies. In general it is not as good as the commercial projects. The improvement of the technical processes in order to extract better metadata from documents is the key technical challenge to these projects. The collection of a good source dataset of citing papers is they management issue.
The CiteSeer project at http://citeseer.ist.psu.edu is the pioneering work is this area. It mostly avoids the managerial issue. It extracts documents from the Internet that look like academic documents. Then it looks through the reference section. When it finds a reference that contains a hyperlink, it extracts that document and the process is starting anew. In this way, the project can gather a large database of documents without human and administration of source document.

Jose Manuel Barrueco Cruz is the creator of CitEc at http://citec.repec.org, see [3]. This is a citation index for Economics based on documents available in the RePEc digital library as described at http://repec.org. CitEc uses a modified version of the CiteSeer software to reference linking of documents which are available in open access (mainly working papers). For each paper in RePEc it provides the features: “cited by”, when the paper has been cited by others also available in RePEc and “get references” when the references of the citing paper have been successfully linked to the cited documents. The CitEc system is lays the technical groundwork for the INCISO project.

3 The INCISO project

The methodology we are going to use in order to extract and link the information about references can be described in the following seven steps:

1. We need to define a set of journals we are interested in. We test our system with a sample of Spanish journals in social sciences.
2. We need to obtain bibliographic information about the articles published on the selected journals.
3. The bibliographic information for each article with the electronic address pointing to the documents full text is stored in a mySQL database.
4. For each citing document, the file containing the document full text is downloaded. At the moment INCISO only deals with files in PDF format. The file is converted to ASCII so that the text could be easily extracted and manipulated.
5. Once the file has been successfully converted starts the parsing of the whole text in order to identify and delimit the references section.
6. All the data extracted in the previous steps is stored in a database of references. This database is used for bibliometric studies of the results.

The project offers two types of deliverables. On one hand, there is the citation index itself, that will be useful for evaluation of research carried out in Spain in social sciences, and on the other hand, a set of technical documents about the system that will be of mayor interest for the researchers community on the digital libraries area. All results are freely published on the web.

The design of the system is based in the following basic characteristics.

1. It is as much as possible multi-discipline,
2. The system is based on open source software as much as possible.
3. It runs autonomously and continuously
4. The project is open. Data generated will be accessible freely accessible for any use.

There are three basic software modules that INCISO use, 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 [5], to identify each reference and to extract their elements (authors, title, etc.).
3. Linking of references with the document they represent if available on INCISO.

4 A Retrieval Performance Evaluation for INCISO

The first type of evaluation to be done with a software piece is a functional analysis. Given that the system has passed this phase, one should proceed to evaluate the performance of the system.

Traditional performance evaluation deals with time and resources. The shorter the response time, and the smaller the resource consumption, the space used, the better is the system. At this moment, we are not interested in these metrics.

In a system such as INCISO, designed for providing retrieval of information, specifically references, other metrics are of interest. Information retrieval systems require the evaluation of how precise is the answer set. This type of evaluation is referred to as Retrieval Performance Evaluation, see [1]. Such an evaluation is usually based on a test reference collection and on an evaluation measure. The test reference collection consists of a collection of documents (academic papers in our case), an example of information requests, and a set of relevant documents for each information request. For each information request, the valuation measure

quantifies the similarity between the set of references retrieved by the system, and the set of relevant references found in the documents. This provides an estimate of the performance of the system.

Let us apply cover the two most used retrieval evaluation measures, recall and precision, to INCISO. We know that our system for creating a citation index by automatic means will not retrieve exactly every reference that is in the documents. We are also afraid that it may retrieve as references information that, in fact, is not a correct reference.

Let *d* be document. Let there be *C* strings in it that we call references. We assume that a human can find these references without any error. Now assume that INCISO processes *d* and generates an answer with *R* retrieved references. INCISO may also produce a number *F* of references it thinks are in the document, but they are strings that are not references, as the human would have identified them. In addition, there are *M* references in the document that INCISO has missed. Ideally $F=M=0$, but usually at least one of them is positive.

Recall is the fraction of all correct references that have been found

$$\text{Recall} = (R-F) / C$$

Precision is the fraction of the correct retrieved references over the total number of retrieved references.

$$\text{Precision} = (R-F) / R$$

Recall is the most important measure in our case. Except for small collections, the denominator in recall is unknown and must be estimated. We will use samples of papers of different Spanish journals. Given that we will usually have bigger collections, we must be careful for the generalization of results.

5 Results obtained in evaluation of INCISO

In this phase of our work we are focused in the detailed examination of results obtained, in order of evaluating its quality. Moreover, we intend to obtain a diagnostic of possible mistakes, and from them, try to look for changes that make INCISO more efficient.

In order to test INCISO we have built a bibliographic database with metadata of articles published in Spanish social sciences journals in the last ten years (1994–). This is the population we are working with. This collection has 133798 metadata records about articles. We call this the metadata base. We test INCISO with a sample of six open access journals available in electronic format. The time span of publication varies from journal to journal. For the largest journal it ranged form to 2004. The total number of articles published in this set of journals was 1493. Every article is available in a PDF file. For 465 of the full-text files the conversion from PDF to text failed. In that case, the quality of the text version was not enough to carry on the detection of references. For 219 the remaining 1028 articles, INCISO was not able to find a reference section. In about the 40% of the cases there was no reference section in the article. But otherwise the section was there but the system was unable to identify it correctly. Therefore 809 articles remain.

In order to evaluate the quality of the process of finding of references we have defined a small database with two tables. The first table is

Document (
DocId,
Ref_real,
Ref_found,
Ref_false,
Note)

where DocId is the primary key. Ref_real is the real number of references in an article as counted by a human, Ref_found is the number of references found by INCISO, and Ref_false is the number of references identified by INCISO that after checking by a human are not references, but something else.

The other table is

Reference (
RefId,
DocId,
Author,
Title,
Year,
Text,
Citation),

where RefId is the primary key, and DocId is a foreign key to a record in the table **Document**. In the fields Author, Title, Year, Text and Citations our evaluator notes if it is correct or wrong. Text is the full text of the citation.

We have started with a representative sample of papers. At the time of writing, we have worked with 236 documents of 809 analyzed. This is 29% of the total available. In the set of 236 selected documents INCISO has found $R=13279$ references. Thus we are looking at an average of 16 references found per article. For evaluating the equality of the references, we classify them as follows

1. Real references (detected manually in the original document)
2. References found by INCISO
3. Junk strings found by INCISO

From there we can calculate precision and recall. Precision is good at 93%. That means that most of the references suggested are actual references, rather than junk strings. But recall is 52%, which is really low. We are going to study with some detail the problems we have found. This completes the first stage of our evaluation.

In the second stage, we will look at the structure of the references found by INCISO. For each reference found the system will aim to locate each of the fields Author, Title, Year. For each field and each value of the field as suggested by INCISO, our evaluator finds,

- if it is correct
- if it is not correct
- if it is too long, probably because it contains other elements of the references

A manual check of all correctly found references finds that INCISO has correctly found

- 96,3 % of years
- 62,5 % of titles
- 60,7 % of texts
- 58,0 % of authors

When parsing the references, we have found four main sources of mistakes, most of them related with the use of Spanish language:

- Use of hyphen when the author is repeated
- Papers without a year of edition, and then you can find "en prensa" (in press), given that the program do not find the four digits corresponding to the year.
- The system identifies "et al.", but not its Spanish version "y otros"
- The title is formed for more than a sentence, o contains punctuation marks.

We think that these four mistakes can explain more than a half of all the mistakes of INCISO. The two first have an easy solution when we modify the corresponding programs. This completes the second step of our evaluation.

In the final stage we analyze the quality of reference linking. Once each the reference has been identified the system try to find if it cites a document to other social sciences Spanish journals, i.e. any article in our metadata collection. If so, we consider there is a citation from the reference to the document.

To do that, our evaluator notes, in the column Citation of the table Reference, one of the following codes.

- Code "I" is used if the citation is and internal correct citation. An internal citation is a citation to a document that is in our metadata database that has correctly identified.
- Code "E" is used for an external s An external citation is a citation to a document that is not in our database.
- Code "P" is used for a lost citation. This is the case where the document contains a citation, that can, by human checking, found to go to a document that is in our database, but the system has not detected it.

Related to citations, in the test dataset INCISO has found and linked 28 of a total of 141 citations our database. It means 20 % of recall.

6 Conclusions

Our study has revealed that there are some simple steps that we can take to improve performance. In the next version we will try to correct mistakes, and compare the new results with the ones obtained in the current version.

References

[1] Baeza-Yates, Ricardo and Berthier Ribeiro-Neto, (1999) "Modern Information Retrieval" *Addison-Wesley*

[2] Barrueco Cruz, José Manuel, Blesa, P., Osca-Lluch, Julia., Krichel, Thomas Velasco, Elena and Leonardo Salom. (2006). "INICSO: Elaboration automatique d'un index de citations des revues espagnoles en sciences sociales". In: *Ametist*, vol. 0, no. 1, pp.113–129.

[3] Barrueco Cruz, José Manuel, and Thomas Krichel (2005) "Building an autonomous citation index for grey Literature: RePEc, the economics working papers case" In: *The Grey Journal*, *An International Journal on Grey Literature*, vol. 1, no. 2, pp. 91–97

[4] Bauer, Kathleen, and Bakkalbasi, Nisa (2005) "An Examination of Citation Counts in a New Scholarly Communication Environment", In: *D-Lib Magazine*, September, vol. 11, no. 9

[5] Lawrence, Steve, Kurt Bollacker, and C. Lee. Giles (1999) "Indexing and retrieval of scientific literature", In: *Proceedings of eighth International Conference on Information and Knowledge Management*, CIKM99, pp. 139–146.

[6] Osca-Lluch Julia and Julia Haba (2005) "Dissemination of Spanish Social Sciences and Humanities Journals." *Journal of Information S*cience, vol. 31, no. 3, pp. 229–236.

[7] Roth, Dana L. (2005) "The emergence of competitors to the Science Citation Index and the Web of Science", *Current Science*, 2005, vol. 89, no. 9, pp. 1531–1536.

♣ This research is supported by grant HUM2004-05532 from the Spanish Science and Education Ministry.