Library and Information Services in Astronomy IV
July 2-5, 2002, Prague, Czech Republic
B. Corbin, E. Bryson, and M. Wolf (eds)

Identifying the Important Papers for One Astronomical Object

Soizick Lesteven & Pascal Dubois

Observatoire Astronomique de Strasbourg, CDS, Strasbourg, France
lesteven@astro.u-strasbg.fr & dubois@simbad.u-strasbg.fr

Abstract. In astronomical databases, such as SIMBAD, the number of bibliographical references attached to one astronomical object is continuously growing with the accumulation of published literature over the years. Some objects are cited in many papers, and it may be difficult to identify the most relevant papers, i.e. those which contain extensive results about that object.

A first aid for the evaluation of the relevance of a paper consists in reporting additional information about the importance of the object in the article. The number of citations of the object in the paper gives one measure. The relevance can also be evaluated from the location of the citations of the object name in the paper. As an example, when the object name appears in the title of the paper or/and as an "Individual" keyword, that object is generally extensively studied. The variable nomenclature used to designate the astronomical objects is one of the obstacles to be taken into account to derive a significant measure of the impact of the various papers dealing with one object. An evaluation of the classification criteria will be presented.

1. Introduction

When you start a literature search for one astronomical object the main problem is not to retrieve bibliographical references but how to select the most relevant papers, those which give important information about that object.

We have investigated that selection problem by using the location in the paper of the citation of the object name. In this paper, we present the method we have used to evaluate these criteria in a selected paper (section 2), the tool we have developed (section 3), the first results we have obtained (section 4), and the validation of our application (section 5).

2. Method

We have selected 17 astronomical objects with different object types (stars, planetary nebulae, galaxies, quasars, radio sources and X-ray sources). All these objects are linked to about one hundred references into SIMBAD. For the 17 objects, the total number of references is 1739 with extracted titles, abstracts

and keywords from SIMBAD (Wenger et al. 2000).

Our selection is based on the locations of the object name citations in the article. We have used three criteria :

- The object is cited in the title ;
- The object is cited in the abstract ;
- The object is cited in the keywords.

These criteria are objective and independent of any a priori judgement on the paper.

3. Tool

The detection of the names has been made automatically with specific software. To retrieve all the different known acronyms for one astronomical object, we have used SIMBAD. SIMBAD is, in the first place, a database of identifications, aliases, and names of astronomical objects. This implies a continuous careful cross-identification of objects from catalogues, lists, and journal articles. Furthermore, to take into account the variable nomenclature used to designate astronomical objects, we have used the "Dictionary of Nomenclature of Celestial Objects". This dictionary (Lortet et al. 1994) is a reference work which tracks all the designations quoted in the literature. It provides full references and synonyms about the different acronyms. Presently, it contains more than 10,000 acronyms and is updated on a regular basis. Each designation, coming from SIMBAD and the dictionary, is automatically translated into a regular expression, and all these regular expressions are searched in the text. The identifiers are retrieved, and some of the inaccurate recognitions can be detected by filtering the results (for example, space mission names, spectral types, atomic or molecular species, can easily be confused with object names). This automatic recognition tool already has been validated by another application (automatic recognition of object name in the abstracts, which is systematically checked by an expert, Lesteven et al. 1998).

4. Results

With this automatic recognition tool, we have selected 146 articles from the 1739 retrieved references, in other words 8 percent of the papers. This results are shown in the Figure 1. The tool is able to recognize all the known acronyms found in the literature. Figure 2(a) shows the total number of quotations by object, and for each object the number of different acronyms is given. For each object, the number of different acronyms found in the texts is small compared to the number of known names in SIMBAD.

- The number of acronyms varied from 1 to 8 even though the number of acronyms in SIMBAD varied from 8 to 46 for the same objects;

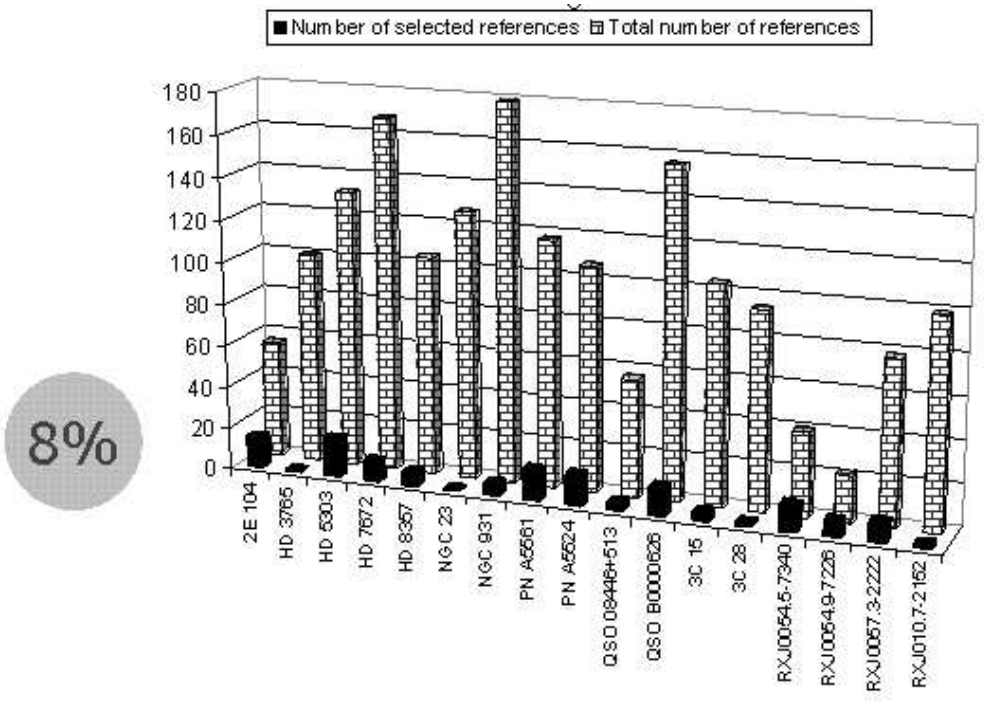

Figure 1. Number of selected references

- 3 objects are detected in the texts with always the same acronym;
- half of the objects were never detected with the starting acronym.

Another interesting result is the location of the object name citation. In the 146 selected papers, we found 228 citations of objects names.

- 101 citations are detected in the titles, which means that 68% of the selected papers have the object name in the title;
- 97 citations are in the abstracts (but 55 are common with the title)
- 30 objects names are cited in the keywords (only 3 are just in the keywords).

These results are shown in the Figure 2(b).

5. Validation

5.1. Validation of the selection

To validate the selection, we have read the 146 selected papers and attributed each of them a weight relating to its importance. To evaluate the importance of one paper for a specific object, we have considered three levels:

- The whole paper or at least a paragraph with a subtitle, is devoted to the object;

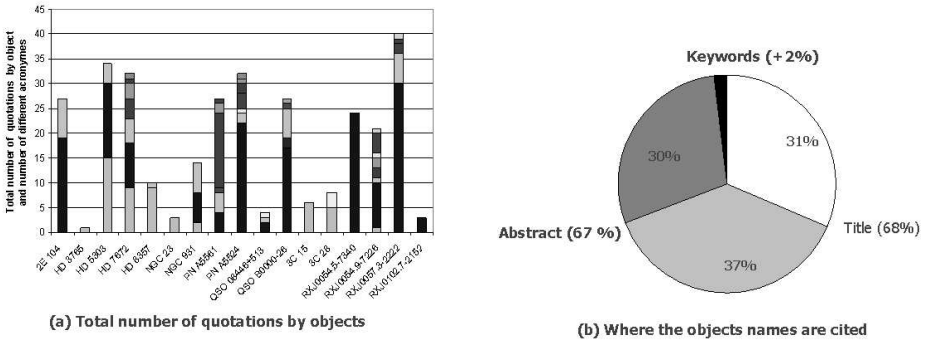

Figure 2. (a) Total number of quotations by object (the different colours correspond to different acronyms) and (b) Where the objects' names are cited

- the paper gives data for this object (magnitude, period, redshift, ...), the object is often quoted in a table;
- the object is quoted in the text as an example (e.g. the same study has been made on ...). This level is useful to detect similar object or theoretical papers concerning this object.

We applied these criteria to the 146 articles, and 139 of them received the first level of recognition. That means that 96% of the selected references have been considered important for the object by the reader with these criteria. The question is now: How many other papers were missed?

5.2. Improvement of the selection

Making a histogram of the number of objects present in each of these 146 references, we see that these references usually quote only a small number of objects (this number is provided by SIMBAD). There is a drop-off after 9 objects quoted by reference as shown in Figure 3. We decided to use this new criterion (references with less than 10 indexed objects) to select other papers and detect some other relevant ones. We read 171 other papers and found 31 relevant references in addition to the 146. We found then the following percentages for the detection of important papers for an object:

- 60% are selected using the detection in the title;
- 25% are added using the detection in the abstract. In other words, 60% of important papers are also detected in the abstracts but with 35% in common with the titles;
- 1% is added using the detection in the keywords (in total 17%);
- 14% could be added by inspection of papers quoting less than 10 objects.

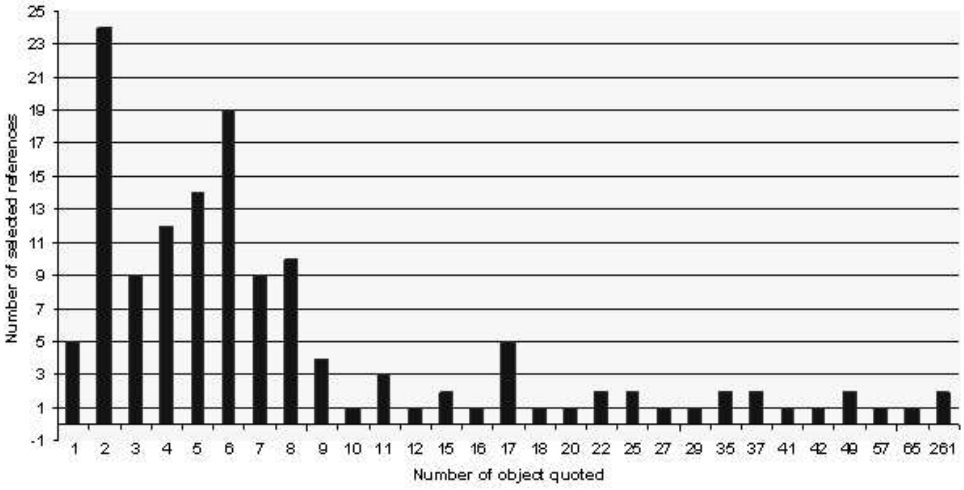

Figure 3. Where the objects names are cited

This last criterion, independent of the object names, allows detection of other relevant papers, but it is less efficient. Only 18% of the read papers are relevant. It has to be considered as a secondary criterion because it retrieves relevant papers (14% more) that are not found by the preceding indexations, but the selection can be a cumbersome process. We tried to explain these 14% additional references:

- The object's name is not detected by the program (the form used is not admitted by the Dictionary of Nomenclature, or it is found in a list of objects not easily recognizable, ...);
- a paragraph is devoted to the object, but there is no mention of it in the title, abstract or keywords ;
- the paper is a theoretical one, the object is cited in the text as a possible application;
- the object appears in the printed version but not the electronic one. There exist some differences between the two versions, in general in "old" papers.

6. Conclusion

The location of the object name in a paper is a simple but interesting criterion to identify relevant papers for one specific astronomical object. It permits making a first selection of about 10% of the papers which permits starting a bibliography for this object.

If the name of the object is in the title of a paper, it is not very surprising that this paper is important for this object. The presence of the name in the abstract has the same importance. The presence of the object name in the

keyword concerns only a small number of papers, but it is a relevant selection criterion. Authors must be encouraged to use this type of keywords more often.

In the near future, we wish to go further into this work to improve it by using other criteria (as the references citations) and offering this new capability to the SIMBAD users.

References

Lortet M.-C., Borde S., Ochsenbein F.,
"Second Reference Dictionary of the Nomenclature of Celestial Objects." Astron. Astrophys. Suppl. Ser., 107, 193-218 (1994) (http://vizier.u-strasbg.fr/cgi-bin/Dic).

Lesteven Soizick, Bonnarel F., Dubois P., Egret D., Fernique P., Genova F., Murtagh F., Ochsenbein F., Wenger M. "Information Extraction: New Developments in Astronomical Information Retrieval for Electronic Publications."
LISA III: ASP Conference Series, Vol. 153, 1998, p. 61.

Wenger M., Ochsenbein F., Egret D., Dubois P., Bonnarel F., Borde S., Genova F., Jasniewicz G., Laloe S., Lesteven S., Monier R.
"The SIMBAD astronomical database. The CDS reference database for astronomical objects." Astron. Astrophys. Suppl. Ser., 143, 9-22 (2000)