

Comparison of Two “Document Similarity Search Engines”

Philippe Poinçot

Observatoire de Strasbourg, Université Louis Pasteur, Strasbourg, France

Soizick Lesteven

Observatoire de Strasbourg, Université Louis Pasteur, Strasbourg, France

Fionn Murtagh

Observatoire de Strasbourg, Université Louis Pasteur, Strasbourg, France; Faculty of Informatics, University of Ulster, Londonderry, Northern Ireland

Abstract. We have developed and used the “CDS document map” based on neural networks (Kohonen maps)

<http://simbad.u-strasbg.fr/A+A/map.pl>

In this self-organizing map, documents are gradually clustered by subject themes. The tool is based on keywords associated with the documents. For one selected document, we locate it on the CDS document map and retrieve articles clustered in the same area.

The second search engine, used by the ADS¹, has the capability to find all similar abstracts in the ADS database, with “keyword request”.

We have compared the results of the document similarity search engines, using the same set of documents. One example will be described and results will be discussed.

1. Introduction

Information broadcasting techniques are constantly in development. This is the reason why electronic publications have become more and more important in many domains, and particularly in astronomy. It is possible to access many documents on-line, and it is important to have efficient information retrieval tools at our disposal.

In this article, we will discuss the bibliographical information retrieval system we have developed at the Centre de Données astronomiques de Strasbourg (CDS). In order to validate our system, we compare our results with those from the ADS. The initial results are then discussed.

¹NASA Astrophysics Data System
(<http://cdsads.u-strasbg.fr>, <http://adswww.harvard.edu>, <http://ads.nao.ac.jp>)

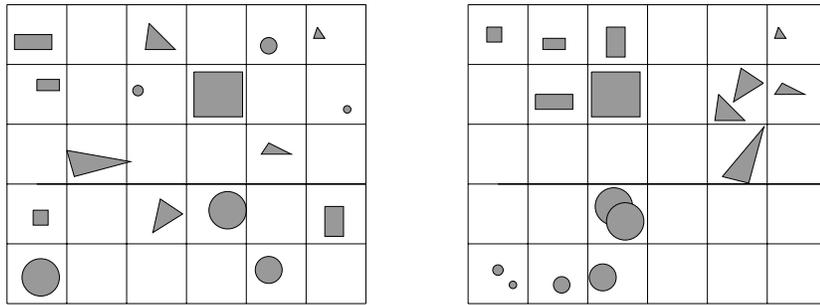


Figure 1. Object locations. Left: before learning. Right: after learning

2. Our Document Retrieval Tool

We are developing, at the CDS, an information retrieval tool, based on Self-Organizing Maps (SOM). These maps are one of the artificial intelligence techniques, and more precisely an unsupervised neural network.

2.1. Self-Organizing Maps (SOM): Basic Principles

A SOM can be considered as a table where objects are classified (Figure 1). In such a table, similar objects are located in the same area. In the example (Figure 1–right), a global classification is shown: the three different shapes are located in three different clusters, furthermore the largest objects are located towards the center of the table, and each cluster is ordered: the largest objects are at one side of a cluster, smaller shapes are at the other side.

Now, let us see how a SOM works:

- each object is described by a vector. In the example, the vector has two components: the first corresponds to the number of angles, and the other to the width of the area.
- initially, a vector is randomly associated with each box (or “node”) of the table.
- each document is located in a box whose descriptive vector is the most similar to the object’s vector.
- during a cyclic learning process, the components of the nodes describing vectors are modified. The learning process produces the classification.

2.2. The Bibliographic Map

Our set of documents. We have used the SOMs in order to classify some bibliographical data. In this study, our set of documents comes from the journal “Astronomy and Astrophysics”, from 1994 to 1998. The descriptive vector is based on the journal keywords associated with each document. We eliminated the rare keywords (keywords found in less than 5 documents). Finally, we used about 4000 documents described by 269 keywords.

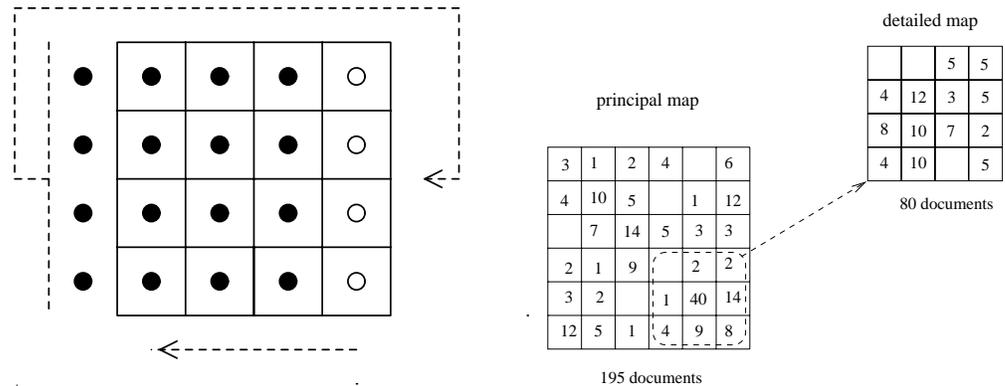


Figure 2. Our use of SOM. Left: shifting the rows or the columns. Right: two classification levels

Our use of SOM. We used the SOMs, adapting them to our own needs:

- Documents located at a map edge have neighbors at the other side of the map. It is then possible to reconfigure the map without losing the similarity of closely clustered documents (Figure 2–left).
- When there are many documents (more than 30) in a node of the map, we create a new map with the documents attached to this node and the 8 nodes around it. Such a map is called *local* map, or *detailed* map (Figure 2–right). The first map is called the *principal* map. We used a 15x15 table for the principal map, and 5x5 for the detailed maps.

The graphical interface. We display this classification with a density map, which represents graphically the areas containing papers of similar content and the number of documents in the areas. Then, the map is labeled to locate on it the themes dealt with (Figure 3–left).

The user can select one node of the map (by clicking on the picture) to obtain some information about the articles located in it (the number of documents and the keywords describing them appear on the right side of the interface) (Figure 3–right). The user can also access the detailed map, and/or the article content (title, authors, abstract) and all the facilities provided by the CDS bibliographical service (including a link to ADS and to the on-line full paper when available).

The user interface allows one to select and display on the map only a part of the database. This is used with keyword queries (only the documents containing selected keywords are shown), or with an external list of documents (bibcode queries).

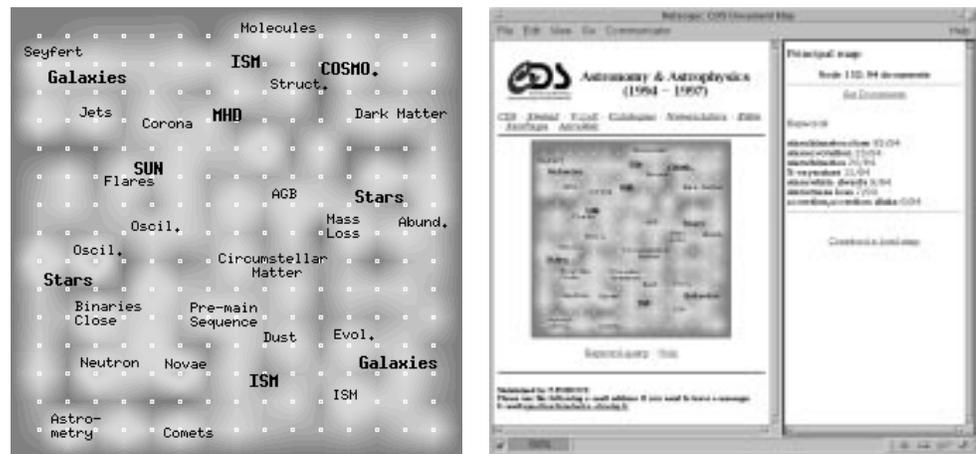


Figure 3. Left: the principal map. Right: the user interface.

3. Validation of the Bibliographic Map

3.1. Method

In order to validate our document retrieval tool, we have begun a study based on the comparison of our system with the ADS system, taken as a reference. The ADS provides wide query possibilities, so we have been able to use exactly the same set of documents with the two systems.

Two different types of comparison have been done. We first use the ADS with keyword query to validate more precisely our system. For the second type of comparison we use the ADS with full-text query to compare the two systems with similar criteria for document retrieval.

What is different in the two systems

- The bibliographic map is based on keywords only; the ADS can process a keyword search² or a full-text analysis.
- The bibliographic map provides a graphical view of classified documents; the ADS returns a sorted list of documents, corresponding to the user's query.

The different steps. The method consists of selecting one document, called “starting document”, and retrieving similar documents (coming from the same bibliographic database) using the two systems. Then, the different sets of documents are compared by an expert of the field. More precisely:

²This is an old version of the ADS Abstract Service, but still accessible.

Results from “keywords only” ADS version. For the keyword query, the ADS retrieved 323 documents. The 315 last ones have the same small score because they are described only by one keyword of the query. We eliminated them for our study. 17 documents remain, only one is scored as less relevant by the specialist.

Comparison

- 11 relevant documents are simultaneously retrieved by both systems.
- The bibliographic map retrieves 3 documents scored as relevant that the ADS does not retrieve, but it misses 5 others. Three of these documents are located in the Interstellar Medium (ISM) zone, and can be retrieved by a keyword query with the bibliographic map, using the keywords describing the starting document. The two others are located in another node of the detailed map.
- The ADS retrieves 5 relevant documents (scored as relevant by the expert) that the CDS does not retrieve, but misses 3 other relevant documents. One of them is only described by one keyword of the starting document, the two others are described by another set of keywords (NASA-STI).

These initial results show that both systems retrieve almost the same documents. They allow our system to be validated.

node	relevant documents	less relevant documents
6*	8	1
5	4	0
11	2	4

*node containing the starting document.

Table 1. Results for the documents coming from the detailed map (list 1).

4.2. Validation with the ADS full text query

Results from the CDS map. We take into consideration all documents of the detailed map, which contains 80 documents (list-2); a check by an expert shows that 22 of them are relevant, 58 are less relevant.

Results for the ADS full text query. To compare the second ADS list (list B) with the map, we selected the first 80 documents (the ADS sorts the documents by decreasing similarity order). As we have seen in section 2.2., we can visualize on the density map an external list of documents when they are already classified on the map. In Figure 4, we can see the location of the documents retrieved by the ADS on the map, corresponding to the given starting document. There is a spreading of the documents while the list becomes longer and longer. This means that the smaller a document’s score is, the farther away it will be located from the “starting document”.

A check by an expert gives 25 relevant documents, and 55 less relevant.

Comparison

- 15 relevant documents are simultaneously retrieved by both systems (Table 2).
- The bibliographic map misses 10 other relevant documents. These documents may be lost because associated descriptive keywords are wrong (typographic error), or not accurate enough: relevant or less relevant articles are found to be described by the same set of keywords. Other lost documents may be retrieved with a keyword query (using the keywords describing the “starting document”), which shows articles in the ISM region for example.
- The ADS retrieves 10 relevant documents that the map does not retrieve, but it misses 7 other relevant documents. There are 7 relevant documents in the 20 first retrieved documents, 14 relevant documents among the 20 subsequent retrieved documents, and 25 in the whole set (list B).

The ADS missing relevant documents are retrieved if we examine the ADS result list beyond the first 80 retrieved documents. These results show that the list of documents retrieved by the ADS (the first documents of the list B) is noisier. But if we examine the complete list, the ADS retrieves more relevant documents than our system.

Full text analysis allows to retrieve more relevant documents, but these are drowned in a longer set of documents.

	relevant	less relevant
ADS (first 80 documents)	25	55
map	22	58
common	15	22

Table 2. Results for all the documents coming from the detailed map (list 2), compared to the ADS (list B).

5. Conclusion

This study shows that the bibliographic map gives results comparable with the ADS request by keywords. These first results allow us to validate our information retrieval system.

When we compare our results to the ADS full text, we show that the ADS retrieves more relevant documents, but it seems that the bibliographic map is more accurate for the first retrieved documents. It seems logical that the full text search is noisier, but more complete if full results are taken into account.

The two systems appear to be complementary, because about 60% of the retrieved documents are identical, and both systems pull in other similar documents. It is recommended to use both systems to get as much relevant information as possible.

Furthermore the bibliographic map, with its graphical interface, is a new approach which is well-designed for the Web. The keyword and bibcode queries give some complementary access to the map. The links existing between two neighboring nodes provide a new mechanism for helping users retrieve similar documents.

But the bibliographic map works only with documents described by a well-defined set of keywords. To extend this tool, we are going to develop a new map based on full text analysis.

Further Reading

- Honkela, T., Pulkki, V., & Kohonen, T. 1995, ICANN-95, 2, 3-7,
Kohonen, T. 1995, *Self-Organizing Maps*, (Berlin: Springer)
Lesteven, S., Poinçot, P., & Murtagh, F. 1996, *Vistas in Astronomy*, 40(3),
395-400
Lin, X., Soergel, D., & Marchionini, G. 1991, In: Proc. 14th. Ann. Int.
ACM/SIGIR Conf. on R & D in Information Retrieval, 262-269
Murtagh, F. & Hernández-Pajarez, M. 1995, *Journal of Classification*, 12, 165-
190
Poinçot, P. 1997, *Actes des Conférences INFORSID*, 625-641,
Poinçot, P., Lesteven, S., & Murtagh, F. 1998, *A&AS*, 130, 183
Salton, G. 1991, *Science*, 253, 974