Multivariate Data Analysis Applied to Bibliographical Information Retrieval: SIMBAD Quality-Control

Soizick Lesteven

CDS, Observatoire de Strasbourg, Strasbourg, France

1 Introduction

The astronomical database SIMBAD developed at the Centre de Données de Strasbourg (CDS) currently contains about one million objects (stellar and non-stellar), all described by heterogeneous data (bibliographical references, measurements and identifiers). SIMBAD supplies the existing links between the different synonym names of one object and describes astronomical objects with information coming from the bibliography. SIMBAD reflects a state of astronomical knowledge that evolves permanently.

Taking into consideration the amount of data as well as the rate of new data production (data from new catalogues and satellites), it was necessary to develop automatic methods to control the coherence of the existing data in order to check the quality of the database and verify the relevancy of the new data.

We designed two methods for the quality control project at CDS, then we developed tools that we applied to SIMBAD data in order to improve its content.

For the first method, we conceived an expert system based on pre-established knowledge, to model astronomer's reasoning. This method is efficient to analyse the astronomical data, but is unable to control the bibliography and to use the information included in the bibliography. Moreover, the creation and updating of the knowledge database represents a heavy work load.

For the second method, we used multivariate data analysis applied to information that comes from the published articles in the main astronomical journals. The bibliography contains quantifiable information that characterizes astronomical objects. That information is used to single out anomalies in SIMBAD.

In this paper, we describe the second method.

Preprint submitted to Elsevier Science 5 July 1995

2 Parametrization of the Bibliographical Information

2.1 Method

The articles in major astronomical journals are an important source of useful information to check the coherence of the SIMBAD data. We wanted to show that the bibliographical reference information concerning one object is quantifiable and can be compared to other information in the database to detect anomalies.

One of the possibilities to analyse the content of a document is to use multivariate data analysis. Factor-space (Ossorio, 1965) is an n-dimensional relevancy space. This space is described by n axes representing a set of n subject matter headings. The words and phrases can be used to scale the axes, and documents are then a vector average of the terms within them. These relevancy scores may be obtained either directly, using human judgement (Ossorio, 1965) or via automated evaluation of classified collections of documents using statistical analysis (Kurtz, 1992). Presently, we simplify our Factor-space by using keywords instead of subject matter headings (Lesteven, 1994).

Our research is presently based on SIMBAD and on the NASA-STI bibliographic database. In the NASA-STI database, references are described by the title, the authors, the journal, a list of keywords, the publication year, an abstract and other information. For each category of SIMBAD data (object type, spectral type, IRAS flux, ...) we can build a specialized "bibliographical space" where the variables are the keywords associated with the selected references and the individuals are the different values of the checked data. To extract the information, we used a Principal Component Analysis (PCA). PCA is the simplest of the multivariate methods. The object of this analysis is to take n variables and find combinations of these to produce new variables that are uncorrelated. The lack of correlation is a useful property because it means that the new variables measure different "dimensions" in the data.

2.2 Application

The first application concerned the object types associated with SIMBAD objects. The object type defines the astronomical nature of the object. In SIMBAD, it results from a diagnosis based on the membership of an astronomical catalogue (Ochsenbein and Dubois, 1992). We selected a data set consisting of 43,345 different objects listed in 5915 different references published in 1989 and common to the two databases (SIMBAD and NASA-STI). These 5915 references are characterized by 463 keywords which occur more than 8 times. The 43,345 objects refer to 40 different object types.

Each object type corresponds to a list of SIMBAD objects having that object type. Each SIMBAD object corresponds to a list of SIMBAD references where that object is mentioned. Each reference corresponds to a list of NASA-STI keywords. Therefore, each object type corresponds to a list of keywords with their frequency. It is now possible to place the 40 different object types in the "bibliographical space" of 463 dimensions to which the PCA is applied. Contrary to a straightforward application, the PCA applied to that application did not permit the reduction of the number of axes to a small number but it did permit the selection of the most discriminant axes for the object types by reducing the weight of meaningless keywords. In that application, the PCA is used to weight the information and therefore to improve the system.

In that application, we kept 39 dimensions after the PCA. In the new 39 dimensional space, S-Space, each axis is a linear combination of keywords and individuals (object types) that can be placed along it. The position of the keywords along each axis, as well as the position of the object types in the S-Space allows us to immediately detect clusters of characters closely related or completely unrelated to each other. It is possible to give an astronomical meaning to the new axes. The distribution of keywords and objects types brought the "knowledge" that is necessary to cluster the SIMBAD objects according to object type and permitted the detection of anomalies.

3 How to use S-Space to Find the Type of a SIMBAD Object ?

3.1 Principles

We placed each SIMBAD object in S-Space from the list of keywords in the bibliographical references. For each object we computed the distance with all the other 43,345 objects and selected the thirty nearest objects from the target object. We selected the most frequent object type in this set of 30 objects (it became the S-Space object type) and we compared it to the SIMBAD object type (taking into account compatibility between certain types). When the system detected an incompatibility between the two object types, it flagged the object with an error signal.

3.2 Results

We kept 17,934 different objects from the data set (we eliminated the objects that have a null distance with the nearest objects). 470 of them gave an error signal, that is to say 2.6% of possible incoherencies. For each of these objects, we tried to analyse the diagnosis and understand the anomaly. The error analysis goes through at least one of the following steps:
- a SIMBAD interrogation to have the complete list of cross-identifiers (to extract the information coming from acronyms), the fundamental data and the measurements (the coherence of acronyms, data and measurement was checked), and the complete list of associated bibliographical references;
- a search in the bibliography to look for typing errors and authors' errors;
- a check of the Palomar maps to have one observational indication. In the near future this step will be replaced by the use of "ALADIN".

A first detailed study of the objects that gave error signals permits the classification of the anomalies into three main categories :
- complex objects;
- errors in the SIMBAD object type;
- SIMBAD type correct.

3.2.1 The Complex Objects

We give the name "complex object" to any object which includes two physically associated components with different object types.
Example :

PK 133-08 1

Type: V*
Coord 1950.0 = 01 55 32.9 +52 39 15 mb, mv = 13.2 14. :
Coord 2000.0 = 01 58 49.6 +52 53 49 sp type = G2Ib

PK 133-08 1	PN M 1-2	p PN VV 8
PN VV' 11	PN VV 8	PN ARO 116
V* V741 Per	LS V +52 1	CSI+52-01555
IRAS 01555+5239		

Measurements:
IRAS: 1

References: 11

1989A&AS...78..301	*1989ApJ...346..201*	*1989IBVS.3364....1*	*1989IAUS..131..261*
1988AJ.....95.1817	*1988AJ.....96..337*	**1988AJ.....96.1407**	*1988BICDS..35...52*
1988PAZh...14..445	*1988PAZh...14..510*	*1988IUE88...2..179*	

"PK 133-08 1" is flagged with an error signal because its SIMBAD object type is *variable star* even though the S-Space object type is *planetary nebulae.* In this case, the inconsistency comes from the fact that we find a planetary nebula linked to its central variable star. In SIMBAD these two elements are associated. Is this association justified or not?

31% of the detected incoherencies concern complex objects. We found planetary nebulae and their central stars, HII regions and their excited stars, dark nebulae and embedded stars, molecular clouds and stars in formation. These confusions often occur from the bibliography, since the authors designated a component of a complex object (or the complex object itself) by the name of the other component.

3.2.2 Errors Detected by a Wrong S-Space Object Type

Example :

Z 0754.1+5300

Type: Galaxy
Coord 1950.0 = 07 54.1 +53 00 mb, mv = 13.9 :
Coord 2000.0 = 07 58.0 +52 52 ./.

Z 0754.1+5300	NGC 2474	Z 262 -52
UGC 4114	MCG+09-13-096	

Measurements:
RVEL: 1

References: 6

1989ApJ...343..811	*1989ApJ...345..871*	*1989S&T....77..227*	*1989IAUS..131...39*
1988A&A...190..237	**1988A&A...202..203**		

"Z 0754.1+5300" is flagged with an error signal because its SIMBAD object type is *galaxy* even though the S-Space object type is *planetary nebulae.* In this case, the contradictions come from the fact that astronomers have named a nearby planetary nebula after this galaxy. Some articles concerning the planetary nebula "PK 164 +31 1" are associated with the galaxy "Z 0754.1+5300".

36% of the detected incoherences concern SIMBAD objects having a correct object type. We found anomalies that are due to wrong keywords, some errors in the name of the object, galaxies designated by supernovae, objects not mentioned in the references and standard stars. It is interesting to note that this check detects the standard photometric stars which often appear in papers dealing with non-stellar objects.

3.2.3 *Errors Coming from the SIMBAD Object Type*

Example :

ESO 495- 21		
Type: Em*		
Coord 1950.0 = 08 34 07.1 -26 14 04	mb, mv = 12.65	12.11
Coord 2000.0 = 08 36 15.0 -26 24 32	./.	
ESO 495- 21	com not a PN..	PK 248+08 1
He 2-10	WRAY 15-241	PN SaSt 2-4
MCG-04-21-005	IRAS 08341-2614	CGMW 2-3967

"ESO 495- 21" is flagged with an error signal because its SIMBAD object type is *emission star* even though the S-Space object type is *galaxy*. For that object, the problem comes from the object type associated with the acronym WRAY 15-, which is too restrictive. WRAY 15- is a list of objects that can be emission stars but also planetary nebulae or emission objects.

33% of the detected incoherencies concerned errors in the SIMBAD object type. We found others cases that produced error signals: acronym problems, authors' errors, wrong fusions, ...

4 Other Results

The evaluation of the results from the first application based on the object types and the year 1989 was promising. To check the validity of the method (reliability, completeness and evolution in time) we again trained the system with different ranges of time. In the meantime, the object types were modified in SIMBAD, with the inclusion of more precise types. We applied the S-Space with the new list of object types to three different data sets. We selected the object types associated with SIMBAD objects having bibliographical information in 1984, then in 1991 and again in 1989.

4.1 *The Keywords and the Object Types*

The keywords and the object types are the "knowledge" base of this research. It was interesting to follow the evolution of keywords and object types to understand the S-Space evolution. Since astronomical terminology evolves, it is probable that the S-Space will also be modified. We compared the three different lists of keywords and particularly the percentage of their frequencies. We extracted three categories of keywords. The first category contains the new or more and more frequent keywords. We checked that some keywords appear because astronomy is a science in permanent evolution. We found some new object types, and some new techniques.

keywords	frequency (%) 84	frequency (%) 89	frequency (%) 91	variance
HIPPARCOS SATELLITE	0.000000	0.000000	0.125462	2.000000
INFRARED IMAGERY	0.000000	0.041517	0.138669	0.936069
GALACTIC BULGE	0.000000	0.087646	0.171685	0.65752
CHAOS	0.000000	0.041517	0.079239	0.64637
DARK MATTER	0.000000	0.364425	0.627311	0.605399
FRACTALS	0.000000	0.046130	0.079239	0.604621
COMPUTATIONAL ASTROPHYSICS	0.000000	1.692961	1.148970	0.554960
INFRARED RADIATION	0.000000	0.078421	0.105652	0.532830
VERY LARGE ARRAY (VLA)	0.000000	0.147615	0.184892	0.518852
X RAY BINARIES	0.069583	0.553557	0.680137	0.366806
COMPUTERIZED SIMULATION	0.049702	0.110711	0.184892	0.230646

In the second table, we observed keywords having less and less importance. This corresponds to meaningless keywords or objects no longer studied. Many of them decrease simply because more precise keywords about the same subject appear (e.g. X-ray sources).

keywords	frequency (%) 84	frequency (%) 89	frequency (%) 91	variance
OPTICAL EMISSION SPECTROSCOPY	0.059642	0.000000	0.000000	2.000000
PLEIADES CLUSTER	0.054672	0.000000	0.000000	2.000000
ASTROPHYSICS	0.616302	0.184519	0.132066	0.48682
X RAY SOURCES	1.655070	0.572008	0.515055	0.32927
ASTRONOMICAL MAPS	0.313121	0.124550	0.085843	0.32369

In the last table we observed keywords showing a big peak in 1989 corresponding to particular events.

keywords	frequency (%) 84	frequency (%) 89	frequency (%) 91	variance
SUPERNOVA 1987A	0.000000	0.931820	0.231115	1.044567
EXOSAT SATELLITE	0.000000	0.119937	0.059429	0.67069
HALLEY'S COMET	0.198807	0.798044	0.369783	0.306113

For the object types, we checked the same categories as these examples show. In the first table we have object types that considerably increase during between 1984 and 1991 or we have new object types.

object type	frequency (%) 84	frequency (%) 89	frequency (%) 91	variance
AGN	0.102829	0.071357	0.809320	1.080043
Seyfert	0.152586	0.144944	0.257614	0.077175
Seyfert_2	0.384781	0.367934	0.645176	0.074179

In that second table we show object types that are less studied by the astronomers.

object type	frequency (%) 84	frequency (%) 89	frequency (%) 91	variance
PM⋆	8.096991	2.522020	3.216761	0.289296
⋆⋆	3.419909	3.211060	1.732628	0.072571

4.2 Which Keywords for Which Object Types

To construct an object type, the system kept the nearest objects from the target object. To understand the part of the PCA on the new combination of keywords, we listed, for different object types, a list of keywords associated with the different nearest objects.

object type	keywords
IR	infrared astronomy satellite, infrared spectroscopy, near infrared radiation, infrared spectra, infrared astronomy, red giant star, carbon star, ...
star	astrometry, astronomical photometry, reference stars, stellar parallax, stellar spectra, stellar color, stellar radiation, emission spectra, astronomical catalogs, stellar composition, stellar magnitude, B stars, carbon stars, late stars, M stars, O stars, Blue stars, ...
⋆⋆	double stars, orbital elements, stellar orbits, stellar parallax, binary star, visual observation, astrometry, stellar motion, stellar rotation, triple stars, quantum counters
Galaxy	galactic clusters, irregular galaxies, spiral galaxies, intergalactic media, sky survey, AGN, QSO, infrared sources, redshift, galactic nuclei, radio astronomy, accretion disks

These results show how the PCA combined keywords having the same meaning for one object type.

4.3 The Detected Anomalies

For the three different years, we found the same kind of anomalies in the same proportion.

- 33 % of complex objects. We found planetary nebulae and their central stars, HII regions, dark nebulae, molecular clouds ...
- 36 % of the detected incoherences have a correct object type. We found the standard stars, some new author's errors (e.g. cross-identification mentioned for the object GRS 264.29 +01.47 with the object FMC 27 in the catalog GRS) and some objects not mentioned in a reference.
- 31% of the detected incoherences have a wrong object type: We found some bad fusions, acronym problems (e.g.: RCW). Between the two applications, we already corrected some incoherences concerning acronyms in SIMBAD, so this category decreased a little, and these improvements allowed to detect some other anomalies.

5 Conclusions

The S-Space is an efficient tool to detect anomalies in fundamental data, measurements and identifiers. These functions are similar to those of the expert system. Moreover, it is efficient at detecting anomalies in the bibliographical content and presently it is the only such tool that exists. Furthermore, S-Space uses the information that comes from the bibliography. By the way, this system is not based on pre-established knowledge and takes into account the evolution of astronomy. Other applications can be developed using S-Space. It can be used to insure the relevancy of the bibliography, and can be used as an assistant for the bibliographical usage of SIMBAD. For example it allows to mark up the standard stars. Another use of the S-Space is as an automatic information retrieval tool.

References

[1] D. Egret, M. Wenger and P. Dubois, in *Databases and On-line Data in Astronomy*, 79–88, ed. D. Egret and M. Albrecht (1991).

[2] M.J. Kurtz, in Advice from the Oracle: Really Intelligent Information Retrieval. *Intelligent Information Retrieval: The Case of Astronomy and Related Space Sciences*, 21–28. Kluwer Academic Publishers, A. Heck and F. Murtagh (eds.) (1993).

[3] S. Lesteven *Méthodes d'analyse multivariée appliquée à la recherche d'information: le contrôle de la Qualité de SIMBAD.* Thesis, Strasbourg (1994).

[4] F. Ochsenbein and P. Dubois, Object classification in SIMBAD *in Astronomy from Large Database II*, 405–410. Proceedings. A. Heck and F. Murtagh (eds.) (1992).

[5] P.G. Ossorio, Classification space: a multivariate procedure for automatic document indexing and retrieval, 479–524, *Behavioral Research* 2 (1965).