SIMBAD Quality-control

Soizick Lesteven
Centre de données astronomiques de Strasbourg
Observatoire astronomique
F-67000 Strasbourg
lesteven@simbad.u-strasbg.fr

The astronomical database SIMBAD developed at the Centre de données astronomiques de Strasbourg presently contains 760 000 objects (stellar and non stellar). It has the unique characteristic of being structured specifically for astronomical objects. All types of heterogenous data (bibliographic references, measurements and sets of identification) are connected with each object. The attributes that define quality of the database include :
- Reliability : cross-identification should not rely upon just exact values object coordinates. It also means that information attached to one simple object should be consistent. We have to control the existing data in order to start with a reliable base and to cross-identify new data assuring the quality as data grows.
- Exhaustivity : delays between publication of new informations and their inclusion in the database should be as short as possible. We have to maintain the integrity of the database as data accumulates.
Taking the amount of data into consideration and the rate of new data production, it is necessary to use automatic methods.

One of the possibilities is to use multivariate data analysis. The Factor-space is a n-dimensional relevancy space. Which is described by the n-axes representing a set of n subject matter headings, the words and phrases can be used to scale the axes and the documents are then a vector average of the terms within them.

Our application is based on the NASA-STI bibliographical database. The selected data concern astronomy, astrophysic and spacl radiation (102,963 references from 1975 to 1991 included 8070 keywords).
The F-space is built from this bibliographical data. By comparing the F-space position obtained from the NASA-STI keywords with the F-space position obtained from the SIMBAD references, we will be able to show whether it is possible to retrieve information with a restricted set of words only. If the comparison is valid, this will be a way to enter bibliographic information in the SIMBAD quality control process.
Futhermore, it is possible to connect the physical measurements of stars from SIMBAD to literature concerning these stars from the NASA-STI abstracts. The physical properties of stars (e.g. UBV colors) are not randomly distributed. Stars are distributed among different clusters in a physical parameter space. We will show that there are some relations between this classification and the literature concerning these objects clusters in a factor space. We will investigate the nature of the relationship between the SIMBAD measurements and the bibliography. These would be new relationships that are not pre-established by an atronomer. In addition, the bibliography could be neutral information that can be used in combination with the measured parameters.