Astronomical Data Analysis Software and Systems VIII
ASP Conference Series, Vol. 172, 1999
D. M. Mehringer, R. L. Plante, and D. A. Roberts, eds.

An Application of XML and XLink Using a Graph-Partitioning Method and a Density Map for Information Retrieval and Knowledge Discovery

Damien Guillaume, Fionn Murtagh

Université Louis-Pasteur, University of Ulster

Abstract. We have defined an XML language for astronomy, called AML (Astronomical Markup Language), able to represent meta-information for astronomical objects, tables, articles and authors. The various AML documents created have links between them, and an innovative tool can cluster the documents with a graph-partitioning algorithm using the links. The result is displayed on a density map similar to Kohonen Self-Organising Maps. AML and its advantages will be briefly described, as well as the clustering program, which is one of the many possible applications of AML.

1. Introduction

XML (Extensible Markup Language; Cover 1998; Bray, Paoli, & Sperberg-McQueen 1998) is a new markup language similar to SGML (Standard Generalised Markup Language), but simpler and having other advantages over SGML. It is becoming a new standard for publishing information on Internet and will in many cases replace HTML. XLink (a part of XLL, Extensible Linking Language; Maler & DeRose 1998) is the XML language used in XML documents to create links between the documents, and is also better than the simple link used in HTML.

We have defined an XML language for astronomy, called AML (Astronomical Markup Language), able to represent meta-information for astronomical objects, tables, articles, authors and images. The different AML documents created have links between them, and we have created an innovative tool to cluster the documents with the use of the links, and display the result with a density map similar to Kohonen Self-Organising Maps.

After briefly introducing XLink, we present here the specific AML features, the issue of the display of clustered documents, the problem of partitioning the graph of the links, and we finally give an example of the maps we can now create.

2. XLink

XLink, a standard mechanism to link XML documents, is still under development. However, its objectives and features are clearly defined, and a W3C working draft proposes a syntax for it. The main feature of XLink is that it allows the definition of links between more than two documents. The links are also labeled so they have semantics, and minimum information can be specified about the desired browser behaviour when a link is used. While in HTML different tags are used as links (simple `<A href="...">` links, or links to images, for instance), XLink will provide a unifying linking method for XML documents.

3. AML Features

AML is defined by a single file describing the XML tags and their hierarchy, called a "DTD". But, as AML describes different kinds of documents, it is composed of different parts: the first part is a meta-data block, which is used for all AML documents. This gives information about the AML file itself: the AML version, the date of creation, the author and so on. The other parts describe the different "AML objects": astronomical object, table, set of tables, article, person and image.

We have developed a Java applet to be used as an AML browser, able to display AML documents with a nice interactive user interface, and allowing the use of the links between the documents as easily as hypertext links.

There are currently two types of links between AML documents: links by URL, and links by identifier. A link by URL can link together two AML documents such as an article and an author, and are usually bidirectional and included in the two documents. Links by identifier use a resource name instead of a URL, and specify the AML object of the linked document. A list of servers serving AML documents for the given AML object can then be used to let the user choose a server. A new URL is created and from that the link can be used as a link by URL. This is like using URNs (Uniform Resource Names), but we are using a system called GLU[1] (Fernique, Ochsenbein, & Wenger 1998) to resolve the URNs, and this system is currently only used in astronomy and has some specific features.

The AML DTD, together with some examples, interfaces to get AML documents from different servers, and the AML browser, can be found at `http://strule.cs.qub.ac.uk/~damieng/these/`.

4. Discovering Knowledge by Using the Links: a Problem of Visualisation

Different existing servers place astronomical documents at the disposal of Internet users, e.g., ADS, NED or Simbad, and one can use the links between the astronomical objects and the articles to surf from one server to another. However, the documents have more and more links, and when a document has a hundred links, it becomes impossible for the user, unless (s)he is very patient, to look at all the links and to summarise the information.

The most common displays for displaying linked documents, trees and graph representations, are not ideal here: a tree would be too big and wouldn't show well the links that are not in the tree structure, and usual graph representations become unreadable with more than 30 documents.

[1] `http://simbad.u-strasbg.fr/glu/glu.html`

5. Self-Organising Maps

Self-Organising Maps (SOM) displayed as density maps are more and more used as a convenient display solution for information retrieval. In the astronomy field in particular, Kohonen SOM have been used to display summarised information on the content of astronomical bibliographic databases and allow the user to search for documents by subject and get similar documents at the same time (Poinçot, Lesteven, & Murtaugh 1998).

The density map shows clusters of similar documents, with a colour corresponding to the density of documents in each square of the map. For each square, a title (difficult to compute automatically) indicates the subject of the documents in the space of the map.

In the case of the Kohonen SOM built with astronomical articles, the clustering is done with the keywords associated with the articles. In the case we are working on, the relationship between the documents is expressed with links between them, which is very different, but the same user interface could be used to display the clusters of documents.

6. An Efficient Graph-Partitioning Algorithm

To cluster the documents using their links, the problem is not an agglomerative question, because we already have a tree (plus some links between the nodes that cannot be displayed on a classic tree). It is instead a graph-partitioning problem: we have a graph of connected nodes (the documents), with weighted edges (links) between them, and we want to partition this graph so that it can be displayed on a density map. This problem of clustering documents with their links is more interesting with XML and XLink than with HTML, because we have weighted links, the weights corresponding to the semantics of the link: the semantics are specified with XLink but not HTML.

Many proposals have been made related to this NP-hard (Garey & Johnson 1979) problem, such as simulated annealing algorithms (solving the local minima problem by accepting non-improving solutions with a probability based on the *temperature*; Sheild 1987, Johnson et al. 1989), and noising algorithms (solving the local minima problem by adding noise to the data, decreasing with the number of iterations; Sharon & Hudry 1993). A recent proposal for a modified noising algorithm by Sudhakar & Murthy (1997) seems to be one of the best solutions known to us. We used this algorithm, which is rather fast, and gives good results. The two important problems this algorithm is solving, are to obtain balanced clusters, and to minimise the number of cut links between the clusters. A constant (α) is used to provide a trade-off between these two constraints, so that the cost function can be expressed as: $T = I + \alpha * C$, I being the total imbalance between the clusters, and C the weighted number of cut links.

7. Examples

Any AML objects could be used for clustering (currently: astronomical objects, articles, people, images, tables and set of tables), but we need to have access to a database giving AML documents as a result of a query. Currently, the

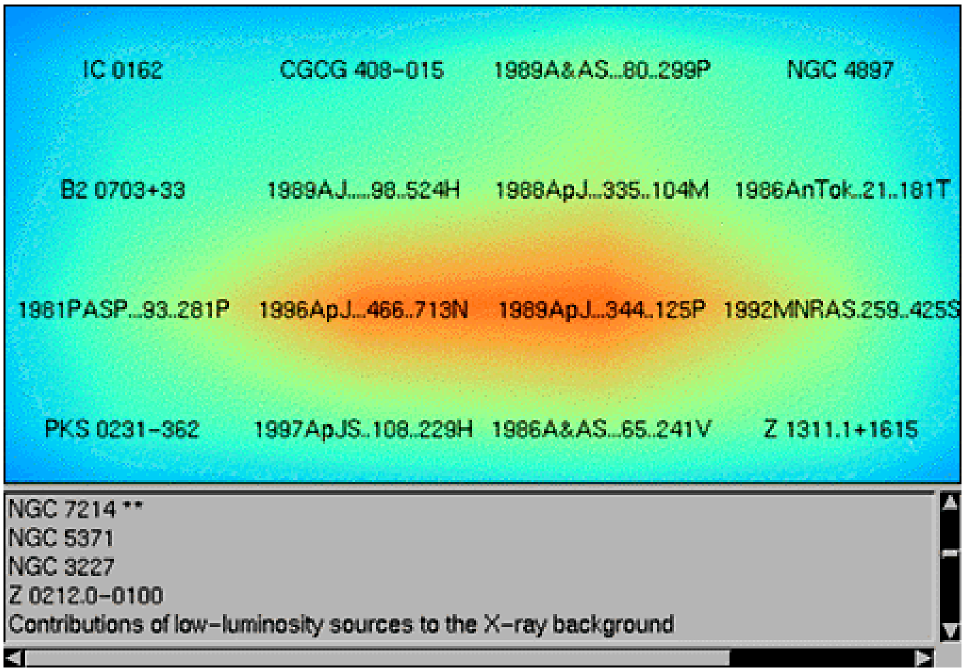

Figure 1. Map for MRK334.

only servers supplying AML documents are: Simbad, NED and ADS. We can also use the database of astronomers created by Chris Benn (La Palma) and Ralph Martin (Royal Greenwich Observatory, Cambridge) to get people-related information. We use 3 different AML objects: astronomical object, article and person, and we have a 3-dimensional space of relations between these objects.

Here is a map example, for MRK334 (an astronomical object). The map is composed of 16 sets of documents, and each set is labeled with the identifier of the document having the most links with other documents in the same set (this may not be the most representative). In the case of an article, it's a 19-characters code called a "bibcode", which is a common identifier for the astronomical bibliography. If the user clicks on a space of the map, the corresponding list of documents appears below the map, with the full title of the article in case of an article. When the user clicks on a document in the list, the content is shown in the AML browser.

Another interesting feature of this program is the selection of two documents, for the whole set of the documents retrieved, being most similar to the first document. The comparison is done with the content of the files, using DTD-specific functions thanks to the XML formatting. For instance, the most similar document for MRK334 is the one for NGC7214, because they are both Seyfert galaxies and many articles have been written about the two objects together. This is a typical "knowledge discovery" application, and it goes much further than what was previously possible with HTML: instead of comparing single words using their frequencies in the texts, it becomes possible to do complex numerical comparisons.

8. Conclusion

Using the ideal features of XML for information retrieval, and its associated language for the links, XLink, we have implemented a new tool for knowledge discovery and used it with astronomical documents. This tool seems very useful, and the only drawback is that, as it is using distributed resources dynamically, it cannot be used in real-time because of the time required to download the documents. Its computational requirements are, however, not far from real-time, with access time dominating processing time.

References

Bray, T., Paoli, J., & Sperberg-McQueen, C. M. 1998, "Extensible Markup Language (XML) 1.0: W3C Recommendation 10-February-1998", `http://www.w3.org/TR/REC-xml`

Cover, R. 1998, "Extensible Markup Language (XML)", `http://www.oasis-open.org/cover/xml.html`

Fernique, P., Ochsenbein, F., Wenger, M., 1998, in ASP Conf. Ser., Vol. 145, Astronomical Data Analysis Software and Systems VII, ed. R. Albrecht, R. N. Hook, & H. A. Bushouse (San Francisco: ASP), 466

Garey, M. R. & Johnson, D. S. 1979, Computers and Intractability, a Guide to the Theory of NP-Completeness, (New York: Freeman)

Johnson, D. S., Aragon, C. R., McGeoch, L. A., & Sheron, C. 1989, Oper. Res., 37(6), 865

Maler, E. & DeRose, S. 1998, "XML Linking Language (XLink): W3C Working Draft 3-March-1998", `http://www.w3.org/TR/WD-xlink`

Poinçot, P., Lesteven, S., & Murtagh, F. 1998, A&AS, 130, 183

Sharon, I. & Hudry, O. 1993, Oper. Res. Lett., 14, 133

Sudhakar, V. & Siva Ram Murthy, C. 1997, Integration, the VLSI Journal, 22, 101