

Spotfire Decision Site

David Kellogg

October 11, 2004

Abstract

Spotfire Decision Site is a flexible visualization tool for XML analysis. Because Spotfire analyzes mostly numerical data, a programmer must act as a conduit between the XML and Spotfire. After this transformation, Spotfire loads the data. Quickly, the user can find missing and duplicate fields. Numerical data can be viewed more easily in Spotfire than in Excel, and with greater speed. The recommendation is to use Spotfire for visual XML validation.

1 Introduction

Spotfire Decision Site is a product of Spotfire, Inc.¹ The strength of this application is visualization of numerical data from a database. Though our data of interest is text in XML form, not numbers in a database, data transformations can integrate XML data with this application. A prototype transformation from XML (extensible markup language) to CSV (comma separated value) has been written in Perl. Spotfire can load CSV data, then help the user find missing data among many records. Quickly, the user can switch among fields to verify the data are complete.

2 Books Example

One example of XML data analysis finds empty subject and author fields. The transformation software counts the number of completed fields for each data type.

¹<http://www.spotfire.com>

2.1 Missing Fields

An author is missing in Figure 1. The Perl transform counted the number of authors for each book. The books are assigned node numbers.

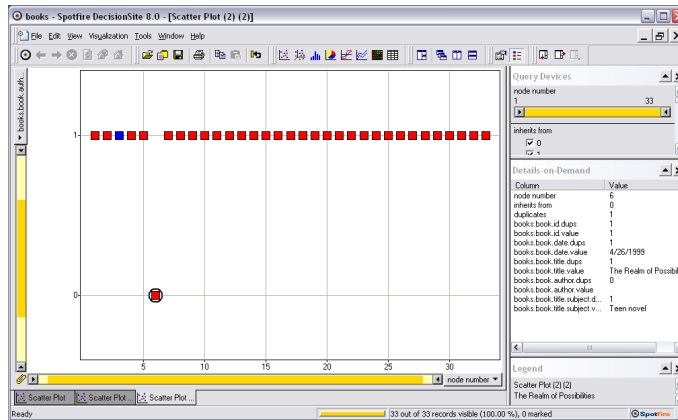


Figure 1: The author of *The Realm of Possibility* is missing, as is shown on the right.

2.2 Binning

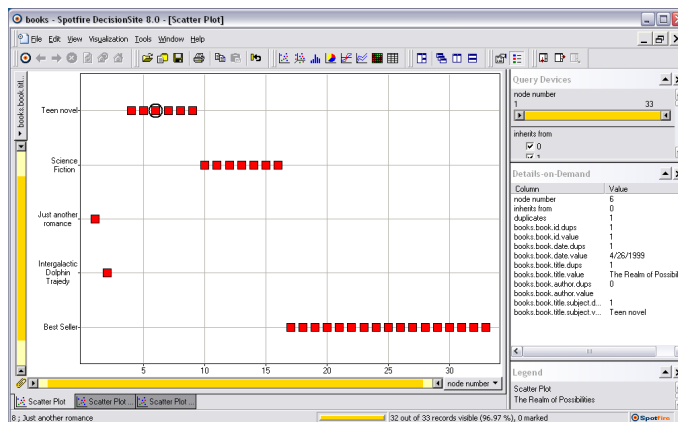


Figure 2: Data are placed in bins by subject, quickly showing the portion of each.

Binning of genres is a good example of visual inspection of fields. In Figure 2, there are 6 teen novels. The graphic quickly reveals this. Notice too, that the subject for node 3 is missing. A graph similar to Figure 1, but for subjects would reveal this missing subject more easily, still.

2.3 Duplicate Fields

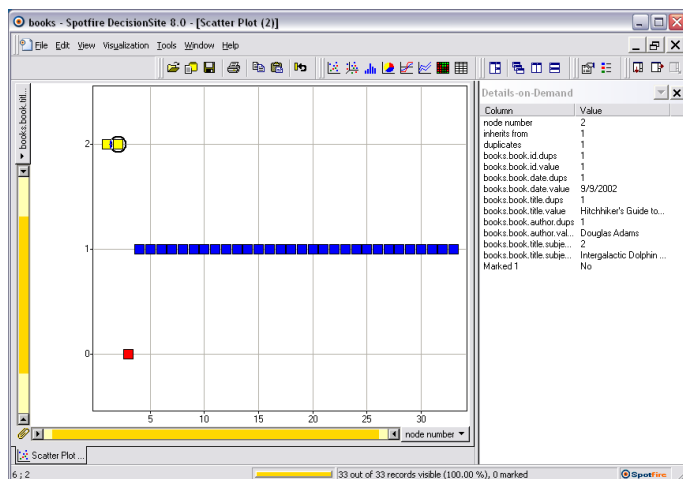


Figure 3: The subject is missing for one book, and another (shown twice) has two subjects.

The Hitchhiker’s Guide to the Galaxy has two subjects, “just another romance” and “intergalactic dolphin trajedy.” This might be a problem, depending on the specification for the XML. Errors buried deep in the XML tree are apparent immediately with this tool.

2.4 Out of Range Data

The book, The Color Purple, has a date that appears out of range, as in Figure 4. This is found by plotting the dates for all books.

3 Data Integration

Spotfire can export data into a database. A database integrates easily with most programming languages. The visual inspector might choose to delete

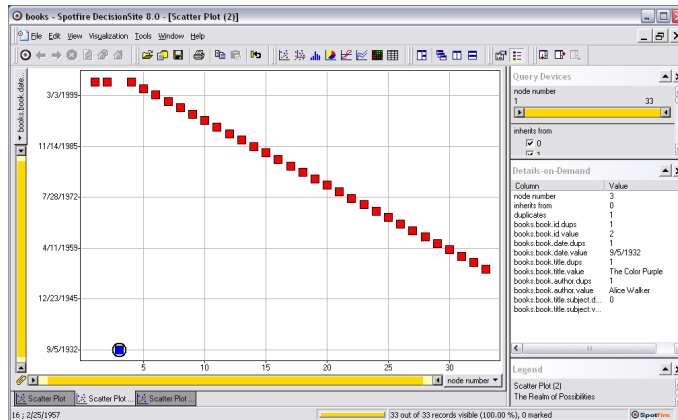


Figure 4: *The Color Purple* stands out as a date out of the range of the others.

some records or fields. This is accomplished by clicking on the record in the graph and marking it with a “No” label. This is noted in the database and the XML after a simple transformation. Beginning with Spotfire, a seamless visual tool can be built for XML analysis and editing.

4 Conclusion

Spotfire Decision Site is an interesting and flexible tool. Only part of the usefulness of Decision Site is documented here. The ease of use compares very well to Excel and other data analysis tools. Despite the orientation of this program to numerical data, using a subset of its features makes it worthwhile to use on non-numerical data.

This document is written in \LaTeX .