Merge Document Sources

Description

To ensure data integrity, by default, the ISI database loader only identifies (in the sense of recognizing many as one) journal entities based on the most stringent of exact string matching. This means that it is generally possible for two documents, or two references, or a document and a reference to specify equal journal identifiers but to have those two journal entities treated as if they are distinct.

This algorithm makes a best-effort attempt to merge (that is, identify) journal entities.

This algorithm does not exist on the menu, but rather is run automatically when the ISI database is loaded into the tool.

Outputs

- A database where the identified identical journals have been merged.
- The merging table used to merge the identical journals . This can be used to rerun the merge manually, likely to correct for errors, with Merge Entities.
- A Merge Report as a text file. It will give a simple description all the journals who were merged, identified by their value for the *TWENTY_NINE_CHARACTER_SOURCE_TITLE_ABBREVIATION* and *FULL_TITLE* columns.

The Basic Problem

A paper from the *Proceedings of the National Academy of Sciences of the United States of America* might indicate that it is from *PROC NAT ACAD SCI USA*, but a reference to that journal might identify it by *P NATL ACAD SCI USA*. On the other hand, did you know that *MOL CELL* and *MOL CELLS* are different journals (*Molecular Cell* and *Molecules and Cells*, respectively)? Linking up journals is a subtle problem.

A Conservative Solution

ISI publishes an official association between journal names and canonical journal identifiers. This, combined with that ISI records typically indicate the "full title" of each document's source journal, generally allows for the identification of document-source journals in your database. In some cases, we can even unify journals which occur only as a reference source (and our ability will increase in time). These methods, combined with identification based on exact string matching, provide nearly as strong an automatic ISI journal merging operation as is possible while minimizing false positives.

Usage Hints

Load an ISI file into the tool, then create a database from it using the ISI database loader.

This algorithm should be suitable for many journal merging needs. Nevertheless, it is strongly recommended that you check the table of journal entities in the database both before and after merging to confirm desired behavior.

For even greater merging power, consider using this algorithm in conjunction with manual journal merging. Just create a journal merging table for your database and apply it.

Implementation Details

The merging is performed as indicated in Merge Entities. This algorithm uses the authoritative journal merging list (AJML). It will first *normalize* the *FULL_TITLE* column from the Sources Table table. If the AJML contains this normalized value, all journals with the AJML's matching value will be merged together. If the normalized *FULL_TITLE* column is not found in the AJML, the *TWENTY_NINE_CHARACTER_SOURCE_TITLE_ABBREVIATION* column from the Sources Table table will be *normalize*. If the AJML contains this normalized value, all journals with the AJML's matching value will be merged together; if it does not contain the normalized value, then the normalized value for the *TWENTY_NINE_CHARACTER_SOURCE_TITLE_ABBREVIATION* column will be used and all entries with the same value will be merged. In the case that an entry does not contain a value for the *TWENTY_NINE_CHARACTER_SOURCE_TITLE_ABBREVIATION* column, a random, unique value will be used for the merging.

Definitions

1. Normalize
   1. For this article, "normalize" means to convert the value of the column to lower case and remove all leading and trailing spaces.
2. Authoritative Journal Merging List (AJML)
   1. A mapping from an Authoritative Journal Name to other common names for the journal. It can be found in the Sci2 tool's configuration folder as *JournalGroups.txt*.

See Also