



TECHNOLOGY

D-Dupe: A Visual Interface for Relational DeDuplication

OVERVIEW

It is no surprise that the amount of data being collected today is growing almost exponentially. This has led to multiple references to the same underlying entity leading to what is known as deduplication. Discovering, visualizing, analyzing and resolving duplicate records within the social networks are the challenges faced by the database communities.

Researchers at University of Maryland College Park have developed D-Dupe: an interactive tool that combines data mining algorithms for entity resolution with task-specific network visualization. The two novel features of D-Dupe are:

1. Stable Visual Layout Optimized for Entity Resolution: The stable and meaningful layout presents small sub-networks from large databases in a task-appropriate, simple, and surprisingly effective design for visually presenting information about potential duplicates.
2. Fine-grained Control for Combining Entity Resolution Algorithm: D-Dupe allows users the flexibility to apply and interleave different entity resolution algorithms. This feature when integrated with visualization of the common social context proves extremely efficient in resolving duplicates. The flexible combination of similarity measure provides a potent environment for decision making and recording of user actions for latter review.

Inventors have also explained the performance of D-Dupe on bibliographic datasets and walked through the procedure for removing duplicate entities. Challenge of data representation is solved effectively by combining visual and analytic information of data cleaning in an interactive tool. Powerful filtering and search techniques are also integrated into the tool to make it versatile.

Researchers have also investigated and demonstrated the application of D-Dupe for name resolution in email collections, place resolution in geospatial databases, and name resolution in academic genealogy datasets and found the tool to be highly effective. The video illustrating the tool is available at <http://www.cs.umd.edu/lings/ddupe>.

For additional information please contact the Office of Technology Commercialization, University of Maryland. Phone: 301-405-2924.

APPLICATIONS

eliminating data redundancy, social networking

ADVANTAGES

visual way to correct errors in a network graph

CONTACT INFO

UM Ventures
0134 Lee Building
7809 Regents Drive
College Park, MD 20742
Email: umdtechtransfer@umd.edu
Phone: (301) 405-3947 | Fax: (301) 314-9502

Additional Information

INSTITUTION

University of Maryland, College Park

PATENT STATUS

Patent(s) pending

LICENSE STATUS

Contact OTC for licensing information

CATEGORIES

- Information Technology

EXTERNAL RESOURCES

IS-2006-047