The Librarian: Entity Matching Across Media Types

Joshua Blum, Nolan Eastin
{joshblum, neastin}@mit.edu

December 9, 2013

1 INTRODUCTION AND MOTIVATION

Entity matching is nontrivial and especially difficult when each entity is large. The Librarian aims to perform automatic entity resolution across various media types. Given sets of text, audio, image, and video files, The Librarian can merge the sets of entities by de-Âŋduplicating the files, gathering metadata from trusted sources, (i.e. IMDB, Rotten Tomato, and Spotify APIs), and categorizing the media based on the metadata that is found.

The motivation behind the system is to be able to handle large dumps of data as well as incremental updates over a large corpus of files with many contributers. The Librarian serves as a background system upon which clients can be built which handler user interaction to add and modify data within the system.

The initial system has been tested with approximately 6 TB of seed data that has been collected from several sources. This data was used as training data for testing and developing our matching algorithms. Only movie entities were categorized from this corpus, although the matching algorithms can be extended to work with other media types such as audio, image, or text files.
In addition to trying to algorithmically resolve two entities, crowd sourcing is used to establish ground truth where necessary. If there are multiple matches for a single entity a user can select the correct metadata for the entity or input their own.

We begin by discussing related work (2), providing a system design overview (3), showing an analysis of the system's performance (4), list goals for future work (5), and conclusions of the project (6).

2 Related Work

3 System Overview

3.1 Architecture

3.2 Implementation

3.2.1 Handlers

3.2.2 Identifiers

4 Performance

5 Future Work

6 Conclusions