Automatically Identifying Records from the Extracted Data Fields of Genealogical Microfilm

Kenneth Tubbs
tubbsk@cs.byu.edu

David W. Embley
embley@cs.byu.edu

Department of Computer Science,
Brigham Young University Provo,
Utah 84602, U.S.A.

Introduction

Imagine the ability to search thousands of microfilm documents at once. Yet extracting and organizing this information by hand is nearly impossible. This research proposes an algorithmic process to identify and extract records automatically from tabular tables found in images. The proposed algorithm extracts record patterns by using knowledge of the structure and geometry of tables and a genealogical ontology. The algorithm accepts raw data collected from zoned microfilm images. It receives this data as an XML input file that describes the coordinates of each table cell, the printed text in each cell, if any, and whether or not each table cell is empty. The process produces record patterns that express the geometry and attributes of the records within a table.

Given the extracted values for hand-written fields, the process can use the derived record patterns to extract the records into an XML file. Individuals can then query the XML files for the information in these microfilm documents. Since the process extracts both the values and geometry of each table, it can display query results as localized regions of the original microfilm images.

Method

The process identifies records by iterating over the following three steps: (1) recognize the structure of the table, (2) map the attributes of the table to a genealogical ontology, and (3) verify the correctness of the extracted record patterns using the genealogical ontology. The algorithm generates and ranks multiple candidates at each step. After each step, the algorithm evaluates the candidates and sorts them according to their likelihood. Following an iteration of the three steps, the algorithm automatically creates rules to guide the next iteration. The process converges when the candidates' likelihood evaluations do not change. The algorithm then selects the set of record patterns with the highest likelihood. The table structure, attribute mappings, and records of this candidate are stored with the microfilm document.

1. Identify the Table's Structure

The algorithm derives table structure by identifying and aggregating table primitives. Table primitives are regular expressions that describe the relationships between table labels and table values. The table zones with printed text are labels, and the nonempty zones without printed text are values. The algorithm uses a minimum distance classifier to recognize the primitive patterns within a table. It then uses the types of extracted primitives to assign the table to a table class. Each table class specifies probabilistic rules that are used to aggregate and factor the table primitives. The algorithm aggregates primitives to form a tree where each level of the tree represents a factoring level. This algorithm creates multiple candidate trees. The trees are evaluated based on the confidence of the table primitive matches, the confidence of the table class match, and the probabilistic rules used.

2. Match the Table's Attributes

The algorithm maps the table's printed text to a genealogical ontology. The table zones with printed text are the labels or attributes of the table. The algorithm checks for four possible label types. The first type consists of labels that match the attributes of the ontology exactly or by a narrow edit distance. The second type is a label that matches a synonym of the ontology's attributes. The third label type includes labels that are composites of the ontology's attributes. The last type describes the labels that require a human-user to define the mapping function. The algorithm creates multiple mappings for the attributes of a table. Each set of attribute mappings is evaluated based on the likelihood of the each match.

3. Verify the Correctness of the Record Patterns

For each of the table structures and attribute mappings created, the algorithm identifies the individual record patterns within the microfilm document. The algorithm examines the records patterns against the constraints of ontology to measure whether feasible relationships and cardinalities exist. Each candidate is evaluated based on its variance from the ontology.

Final Remarks

The proposed algorithm describes a process to automatically extract record patterns from the tables of genealogical microfilm. The research proposes methods to exploit knowledge of the structure and geometry of tables and a genealogical ontology to identify record patterns. It derives the structure, geometry, attributes and attribute mappings for each record type with in a table. In addition, it provides a means for assigning a confidence to extracted record patterns.