Chapter 21
A Grid and Cloud Based System for Data Grouping Computation and Online Service

Wing-Ning Li
University of Arkansas, USA

Jonathan Baran
University of Arkansas, USA

Donald Hayes
University of Arkansas, USA

Cameron Porter
Acxiom Corporation, USA

Tom Schweiger
Acxiom Corporation, USA

ABSTRACT

Record linkage deals with finding records that identify the same real world entity, such as an individual or a business, from a given file or set of files. Record linkage problem is also referred to as the entity resolution or record recognition problem. To locate those records identifying the same real world entity, in principle, pairwise record analyses have to be performed among all records. Analytical operations between two records vary from comparing corresponding fields to enhancing records through large knowledge bases and querying large databases. Hence, these operations are complex and take time. To reduce the number of pairwise record comparisons, blocking techniques are introduced to partition the records into blocks. After that records in each block are analyzed against one and another. One of the effective blocking methods is the closure approach, where a "related" equivalence relation is used to partition the records into equivalence classes. This paper introduces the closure problem and describes the design and implementation of a parallel and distributed closure prototype system running in an enterprise grid.

DOI: 10.4018/978-1-4666-2065-0.ch021

1. INTRODUCTION

A record may be viewed conceptually as consisting of a set of fields. When unique identifiers are unavailable or do not exist in records, determining records that represent the same real world entity is an important and challenging problem, which has many applications. For instance, it addresses data quality issues such as "data accuracy, redundancy, consistency, currency and completeness" (Li, Zhang, & Bheemavaram, 2006). Ensuring data quality is becoming a critical issue that impacts organizational performance(Ballou, Wang, & Pazer, 1998; Ballou, 1999; Delone & Mclean, 1992; Redman, 1998) This problem is also referred to in the literature as record linkage problem (Fellegi & Sunter, 1969; Newcombe, 1988), data cleaning problem (Do & Rahm, 2002), object identification problem (Tejada, Knoblock, & Minton, 2001; Tejada, Knoblock, & Minton, 2002), or entity resolution problem (Benjelloun, Garcia-Molina, Su, & Widom, 2005). All these research efforts deal with the fundamental question of how to effectively identify record "duplicates" when unique identifiers are unavailable or do not exist in records. The main idea is to rely on matching of other fields in records such as name, address, and so on. It is not uncommon for a record having over hundred fields in real data files. Therefore only a relatively small subset of fields is used to carry out the matching. The set of fields selected is application dependent and is often referred to as keys.

1.1. Motivation for Closure Computation

The most basic application is to identify duplicates within a single file or between two files. In the single file situation, in principle, each record must be checked against every other record in the same file in order to find its duplicates. Similarly, in the two files scenario, each record in one file must be compared against every record in the other file. Both schemes amount to carrying out all pairwise analyses among records and have a time complexity that is quadratic to the number of records (the input size of an algorithm). The methods used by analytical tools to decide if two records is a match vary from comparing corresponding keys (fields selected for matching) to enhancing records (correcting certain fields, appending additional fields, etc.) through large knowledge bases and querying large databases. Since analytical tools are complex and time consuming, each pairwise analysis takes much more time than that of a simple instruction. For large files having hundreds of millions to billions of records, the performance of such a scheme is unacceptable.

To overcome the poor performance, the total number of pairwise record analyses must be reduced. To understand how this could be done, let us consider the case where records are in a single file. Conceptually each record may be viewed as being associated with a potential set of records from which to find its duplicates. Records not in the potential set are guaranteed not to be duplicates, and therefore need not be compared with. Hence, pairwise analyses are needed only between records in the same potential set. For a record, a straightforward way of defining its potential set is to let all other records in the file to be its potential set. This leads to the quadratic pairwise comparisons. Now imagine that a scheme exists that reduces the potential set from the whole file to a small fraction of that. Furthermore, this scheme can be carried out efficiently. What the scheme does is that it effectively partitions the records in the input file into many relatively small groups within which all pairwise record analyses are needed. The scheme that reduces the record pairs to be compared is called blocking in the literature (Baxter & Christen, 2003). Closure operation, of which the definition is given in the preliminary section, is one of the blocking schemes. This paper presents a parallel and distributed, grid based prototype that carries out the closure operation.