Chapter 6.5
Management of Data Streams for Large-Scale Data Mining

Jon R. Wright
AT&T Labs – Research, USA

Gregg T. Vesonder
AT&T Labs – Research, USA

Tamraparni Dasu
AT&T Labs – Research, USA

ABSTRACT

In an enterprise setting, a major challenge for any data-mining operation is managing data streams or feeds, both data and metadata, to ensure a stable and certifiably accurate flow of data. Data feeds in this environment can be complex, numerous and opaque. The management of frequently changing data and metadata presents a considerable challenge. In this chapter, we articulate the technical issues involved in the task of managing enterprise data and propose a multi-disciplinary solution, derived from fields such as knowledge engineering and statistics, to understand, standardize, and automate information acquisition and quality management in preparation for enterprise mining.

INTRODUCTION

Witten & Frank (2000), Piatetsky-Shapiro et al. (1996), and others have noted that industrial *data mining* applications frequently require substantial effort in the activities that occur prior to the application of data-mining algorithms. These observations match very closely with our experience in working with real-world applications. We have been able to work with a telephony data-mining project over the past several years, and it has served as a test bed for our ideas and experiments on data quality. The project has visibility at all levels of the enterprise, and it provides us with access to complex data on a large scale. The monitoring methodology we have developed has largely been driven by our experiences in working with this application and therefore represents, in part, a

response to practical needs. Fundamentally, the amount of data flowing into that project is at a scale that demands automated monitoring tools, or at least, machine-assisted monitoring of some sort. Our ideas, therefore, receive serious testing in the crucible of a real-world application, a facet occasionally missing from research projects. Certain aspects of the application are covered in more detail in a later section.

We find that perhaps 90% of the effort in a real-world data-mining project is spent in acquisition, preparation, management of data, and other related activities, with about 10% spent on analysis. Notwithstanding this fact, much of the research on data mining focuses on providing increasingly sophisticated analysis and discovery algorithms, typically relying on pre-existing databases, such as the World Wide Telescope (Gray, 2003). Consequently, the core practical issues involved in acquiring and managing data on a large scale have not received the attention they deserve.

Successfully managing data at scale, including preparation prior to analysis, is probably the key determiner of success on any practical data mining application. Typically, the most interesting data originates within the computer applications that support the core business processes of an enterprise. This data takes a variety of forms — transaction records, application logs, Web scrapings, database dumps, and so forth. The supporting computer systems, because of their critical role in the business, are nearly always off-limits to CPU and space-hungry data-mining activities, and there is no choice but to transfer the data to a computing environment that is more data-mining friendly. Because of the increased computer support and automation of the core business processes in the modern enterprise, the scale of data available for mining is both massive and growing. In addition, data-mining is sometimes the only window on important business processes within an enterprise. The *metadata* so essential in interpreting and understanding enterprise data can be of significant size in its own right. By necessity, both data and *metadata* evolve continuously with the changing business environment of the enterprise. Originally, it was thought that most problems could be solved through data warehousing but the scale and flexibility required by modern data-mining applications make this a less than efficacious and flexible solution.

Many of the key issues in data quality and information management have been described previously (Pipino, Lee, and Wang, 2002; Huang, Lee, and Wang, 1999). In particular, maintaining a high level of data quality is closely tied with data-feed management. In this chapter, we focus on tools and methodologies needed to implement data quality management techniques such as those discussed in these publications within the context of enterprise data quality management, from data gathering to data mining. We view the whole process as somewhat organic in nature, emphasizing the growth and evolution of data in a real setting.

The *mise en place* Problem in Industrial Data Mining Applications

The term we use to designate the process associated with assembling, preparing, managing and interpreting data for the data mining process is *mise en place*, a cooking term, meaning "put in place," which describes the activities of measuring, chopping, peeling, and so forth, in preparation for the actual cooking process. In a similar fashion, enterprise data must be prepared for data mining. Such activities include: (a) assembling time-dependent metadata, (b) providing and managing a long-term repository for ephemeral data and metadata, so that longitudinal trends may be deduced, and (c) monitoring and improving the quality of the data. In many industrial settings the bulk of the effort for a data mining project focuses on these *mise en place* activities. Without these activities, meaningful data mining would be impossible (see Witten & Frank, 2000).