

Grid Enabled Distributed Data Mining and Conversion of Unstructured Data

Paul Donachy
Terrence J harmer
Ron H Perrott
Belfast e-Science Centre
www.qub.ac.uk/escience

Jens Rasch
Sarah Bearder
Martin Beckett
Dataactics Ltd
www.dataactics.co.uk

Abstract

With the explosion in size of data warehouses and the proliferation of databases, handling large volumes of unstructured data is still the most critical element currently affecting companies attempting to control their data assets. This presents, even experienced data managers, with a host of potential problems, including matching, transformation, and integration of various disparate data sources.

At present, it can not be stressed enough how poorly developed many of the current practices are and how as the size of datasets is only going to increase massively in the near future, it is of significant commercial importance to develop scalable affordable techniques for handling these issues. In terms of both crime prevention and national security, it is imperative to streamline and provide a mechanism for seamless access and mining to such volumes of unstructured data.

A grid enabled environment has the potential to solve this problem by providing the core data mining engine with secure, reliable and scaleable high bandwidth access to the various distributed data sources and formats across various administrative domains. The architecture and a roadmap for a Grid-enabled distributed data matching solution based on such technology will be presented in this paper.

1 Introduction

The explosion in the size of data warehouses and the proliferation of databases, handling large volumes of unstructured data is still the most critical element currently affecting companies attempting to control their data assets. This presents, even experienced data managers, with a host of potential problems, including matching, transformation, and integration of various disparate data sources. The emergence of Grid technology and the ability to provide secure, reliable and

scaleable high bandwidth access to distributed data sources across various administrative domains is set to play an important role in the area of data mining.

This paper will present the background and motivation for the industrial e-Science project GEDDM. The industrial partner in this project, Dataactics Ltd, has developed the world's first fully fuzzy parallelised data matching algorithm. We describe the effect fuzzy matching has on data mining and the computational consequences. This paper will present the nature of this industry with typical business examples to explain the types of

errors encountered in real world data and the fuzzy approach to data matching. The CPU intensive demands and consequences of this process will be described and a solution using message passing will be outlined. Results are presented for a wide range of hardware scenarios to implement parallel solutions from SMP to Beowulf clusters. The architecture for a fully service oriented Grid enabled distributed data matching solution is presented.

2 Current Practice

Volume and structure of data are still the most critical elements currently affecting companies attempting to control their data assets. At present, it can not be stressed enough how poorly developed many of the current practices are and how as the size of datasets is increasing massively in the near future, it is of significant commercial importance to develop scalable affordable techniques for handling these issues. In terms of both crime prevention and national security, it is imperative to streamline and provide a mechanism for seamless access and mining to such volumes of unstructured data.

Bringing together data from different sources poses a number of difficulties. There are straightforward problems of format and standards in representing addresses and dates and employing different systems, which were intended for different uses and emphasis different parts of the data.

In addition to the rich source of natural errors there are also deliberate errors introduced for fraud.

We are particularly concerned with errors, which lead to duplication - that is multiple physical records in a database, which actually refer to the same entity. We have classified typical errors in databases, which lead to duplicate entries into a number of classes:

- **Deletion, insertion, and replacement:** Single character errors

in a field with no obvious reason. Often either “typos” or extra characters inserted when translating between different data formats.

- **Phonetic errors:** Single errors due to mis-hearing or “mis-thinking”. Especially common with numbers eg. 15 / 50. The specific errors are often dependant on the language being used.
- **Visual errors:** Similar to typing errors but it is possible to predict which character was intended. These are becoming more common as legacy paper based data is computerised or systems are used to automate entering of manual forms. Typical errors are confusing “m” and “n” and OCR system confusing 1(one) and l (letter L).
- **Equivalent words:** An interesting class of errors occurs when words are regarded as equivalent by human operators such as the Ms/Mrs or road/avenue/street/terrace. This can even be deliberate in some areas to claim a more desirable address.
- Names are a rich source of these errors. To everyone except a computer “Richard” is equal to “Dick”, this problem is even more prevalent in cultures where names contain honorifics or have different forms depending on the situation.
- **Inconsistency errors:** These are errors, which can be identified automatically. Mismatches between the gender of a title and first name or between a postcode and post town.

Of late, the Industrial partner has had requests from major customers in the US (banking and legal sectors) to interrogate data sources in numerous structures, formats and disparate administrative locations. These business opportunities all present a similar technical

problem, in that interfacing to such vast amounts of information in a common structured parallel approach across such disparate structure and sources was a bottle neck and problematic. (E.g. one legal customer held over 45Tb of data in various formats including email, PDF, web logs, various RDBMS).

3 Deduplication

The Datactics “DataTrawler” product consists of a graphical user interface running on Windows and most Unix like systems. This shows the current view of the data and a tree view of all the operations performed. Operations range from simple import and conversion of data to complex fuzzy logic deduplication of large datasets but all can be performed with no programming or specialized database knowledge. Each operation is actually carried out by a separate engine, which may be running locally or on a different platform or even in a cluster. The results of the engine are reflected in the GUI. In addition a batch script can be generated allowing the sequence of operations to be performed automatically. An audit report of all operations is generated for all runs.

The range of errors described above need a variety of approaches which are all contained. The mixture of mostly random point errors and structural errors lead to our unique approach combining user supplied domain IP and fuzzy matching technology together with HPC computing power.

Match files: All the operations in the tool are first subject to a set of user editable templates. These allow specific patterns of letters and numbers to be matched, or equivalent words to be compared or certain characters to be removed. This has the advantage of allowing non-programming users to contribute domain specific business intelligence and allowing easy specialization and customization for a specific country.

Fuzzy: All the matching engines within Duplitrix can employ Fuzzy matching. This allows a defined number of allowable errors to be specified for each match. An error is a single missing character or a swapped pair of characters. The level of Fuzziness can be tuned depending on the data to be matched. For example in first names we might specify an overall level of 1 error but for long, complicated, foreign or often mis-spelt names a higher error level would be permitted.

Since many of the errors are just as likely to occur in the first character allowing a fuzzy match prevents many of the indexing techniques commonly used to search data. Ultimately this becomes a brute force approach of matching every character of every record against every other record.

Two recent advances have made possible this level of error matching on large datasets. Cheap commodity high performance PCs with large memory/storage coupled with fast networking and cheap (free) available operating systems, which allow clusters of machines to be built easily.

Messaging: The Datactics matching engines use MPI (specifically mpich) to operate seamlessly on SMP and clusters under Windows, Linux and a wide range of Unix-like operating systems.

4 Towards a Grid Architecture

The Grid based Distributed Data Mining architecture presented here is based on the Open Grid Services Architecture (OGSA) model [1] derived from the Open Grid Services Infrastructure specification defined by the OGS Working Group within the GGF [2].

The Open Grid Services Architecture (OGSA) represents an evolution towards a Grid architecture based on Web services concepts and technologies. It describes and defines a service oriented architecture composed of a set of interfaces and their

corresponding behaviors to facilitate distributed resource sharing and accessing in heterogeneous dynamic environments [3].

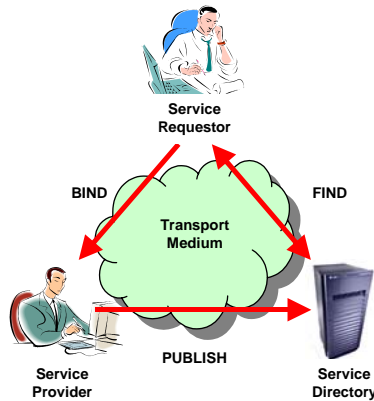


Figure 1

Figure 1 shows the individual components of the service-oriented architecture (SOA). The service directory is the location where all information about all available grid services is maintained. A service provider that wants to offer services publishes its services by putting appropriate entries into the service directory. A service requestor uses the service directory to find an appropriate service that matches its requirements.

An example of such a requirement is the maximum price a service requestor is willing to accept for a specific data mining service. When a service requestor locates a suitable service, it binds to the service provider, using binding information maintained in the service directory. The binding information contains the specification of the protocol that the service requestor must use as well as the structure of the request messages and the resulting responses. The communication between the various agents all occurs via an appropriate transport mechanism.

5 Summary

Grid technology presents a framework that aims to provide access to heterogeneous resources in a secure, reliable and scalable

manner across various administrative boundaries. The data-mining sector is an ideal candidate to exploit the benefits of such a framework.

However before widespread adoption happens within this sector a number of fundamental areas will need to be addressed:

Heterogeneous Resource: The challenge in grid enabling a commercial product comes from the requirement to use a mixture of different resources and platforms, efficiently use mixtures of nodes with very different performance and capabilities and manage external resources with unknown availability. All this must be provided to a non-technical user in a simple and straightforward manner.

Management: As grids evolve as a heterogeneous array of heterogeneous nodes the monitoring and management of such grids becomes more and more critical. To date, little or no work has been undertaken to investigate a cohesive strategy to managing such arrays of heterogeneous grid elements and how such monitoring and management strategies will be integrated into applications and existing enterprise management solutions.

References

- [1] OGSA
<http://www.globus.org/ogsa/>
- [2] OGSI
<http://www.gridforum.org/ogsi-wg/>
- [3] S. Burbeck, "The Tao of e-Business Services," IBM Corporation (2000); see <http://www-4.ibm.com/software/developer/library/ws-tao/index.html>.