Viability of Implementing Data Mining Algorithms as a Web Service

Mohamad Saraee, Nick Howard, Carl Stent, Edward Thompson
School of Computing, Science and Engineering
University of Salford, UK

ABSTRACT

This paper describes an experiment into the viability of implementing data mining algorithms within a W3C standards compliant web service. The experiment shows that it can be done by the successful deployment of a prototype based on an implementation of the K-means clustering algorithm. The prototype produced demonstrates how the concept of a data-mining web-service can be a reliable and effective data-mining tool especially in environments where raw processing power is a valuable commodity. The slim-client to fat-server model is demonstrated effectively showing how a user armed with a simple web browser can potentially harness super computing power. In addition the foundation for the development of an advanced data-mining framework is presented which can include the implementation of any number of data mining techniques. The paper also seeks to highlight some ideas for future research and development of more sophisticate web services that are more scalable to suit both very specific tasks and very large datasets

Keywords

Data Mining, Web Service, K-means, XML, DMSP (Data mining service provider)

1. INTRODUCTION

Web services are a standardized, platform independent way to making an API publicly available over the Internet. Developers can use standard languages to write web service that are translated into and XML document known as a WSDL(Web Services Description Language). Since all the interactions with the web-service are completed in plain text they are very platform independent making them much more useable by a wider range of people.

This paper is a discussion of how a data mining web service could be designed and implemented. It describes the successful implementation details of a prototype system designed as an experiment.

Technical contributions

- Implementing clustering by K-means in c#
- Delivering data to be mined to the Web Service
- Reporting results in a universal manner.

Data Mining technology is a proven and valuable way for large commercial organizations to identify previously unforeseen and unrecognizable trends in the data they collect. These trends and patterns can be interpreted and used to make commercial decisions that both cut costs for an organization and provide an improved experience for their customers.

Data mining requires very intensive computer processing so until now has been the domain of large commercial or scientific organizations. These organizations have the necessary resources, in both hardware and software, to perform data mining for themselves. Smaller organizations on a more restrictive budget simply don't have the money to invest in data mining infrastructure of their own but would find data mining commercially beneficial. A service-based approach is therefore necessary to fill this gap.

Such an approach would ideally be as low cost as possible. Since data Mining lends itself to the client server model a web service was an obvious approach. The Web Service provides the flexibility of a data mining API while removing the most expensive parts of the data mining process; the actual data processing on to hardware best equipped for the task.

The most significant cost advantage of this approach is due to web services requiring the minimum of human interaction between service and service provider thus the service provider requires less staff and can pass this cost saving onto the customer.

For a client to implement a web service still carries a certain cost so a browser based interface is also described in this paper. This interface is designed to demonstrate the ease of implementing the web service API.

In summary the problem being investigate here is to whether a Web Service based approach is sensible and usable for data mining. Also to determine the viability of using data mining in a "self service" context useful to small and medium sized businesses.

2. METHODOLOGY

The alteration to the data mining process is highlighted below in figure 1. The highlighted are the part of the process that are implemented in the data mining web service.

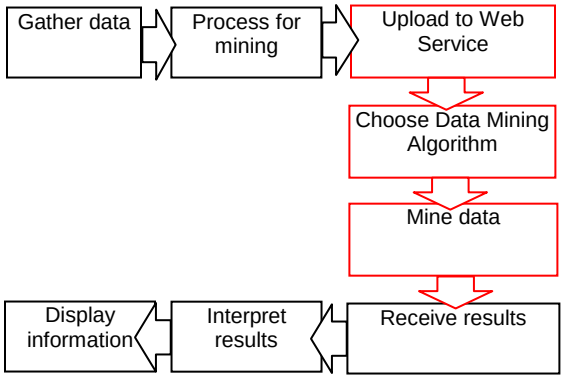


Figure 1: The alteration to the data mining process

Each of these stages will be dealt with below.

2.1 Design

The key points for the design of a data mining web service are

- Input data must be in a simple, extensible, textual data format that can be interpreted on any platform.
- All the complex processing must completed by the server.
- Output must be in a predictable format that can be processed by the client and used to create something that can be easily interpreted by a computer. IE the output can be ***data*** but must be able to be easily converted to *information* without complex processing.

2.2 Uploading to the web service.

Most data mining applications can be tailored to a specific purpose, however we do not have that luxury. Because the exact use of our data mining service is unknown a very open ended and modular approach is required. This means expressing the data in a simple yet extensible manner, for this we propose the use of XML, therefore our data sets must be encoded in XML for upload.

A good data mining web service will allow clients to make use of a variety of data mining algorithms, however, each data mining algorithm requires a slightly different format for the input data and will produce differing output. Expressing input and output data as XML provides the perfect technological solution for data transfer because DTD (Data Type Declaration) technology can be used to ensure the XML data format corresponds to the requirements of the algorithm. Here is how the system would work:-

- Data Mining Service Provider (DMSP) can implement a data mining algorithm to be included in the service
- The DMSP will then publish 2 DTD's, one for the *input* to the service and one for the *output* this is in addition to the WSDL (discussed later).
- Data passed to the web service by a client will have to conform to the input DTD and the client should expect data that conforms to the output DTD encoded by the web service.
- When the data is uploaded it will be checked for validity against the input DTD and then sent to the correct algorithm based on the chosen DTD. Data that does not conform to the catalogue of DTDs will result in an error.

The data mining algorithms also require parameters. These could be included in the sent data or they could be uploaded separately. It would be better practice to separate the data to be mined from the parameters that describe how to mine it.

In the example created for this paper the data upload and parameterization is completed in a single transaction called as if it were a method. This is shown below

```
kmeansService.kmeansProcess2D
(data,colIndex1,colIndex2,numberOfClusters);
```

As you can see this sends the data as a long string in the first parameter thus keeping the system as simple and transparent and possible. The other parameters are specific to the k-means clustering algorithm, since k-means clustering is most useful at processing 2d data such as location information we pass in 2 column numbers that are to be processed and then the number of clusters that we would like to receive.

The method returns the data directly meaning that the client must wait for the server to complete the mining process before the clustered output is received.

2.3 Implementing a client

Since the only algorithm to be implemented in the web service is K-means clustering there is actually very little complexity in terms of code. The web services main role in the system is to abstract the processor intensive data mining figure 2 shows the role of the client.

In order to use the web services functionality it is necessary to implement a client interface. The client interface must handle the following:-

- The upload of the data to be processed
- The ability for the user to specify parameters
- A human readable interpretation of mined data

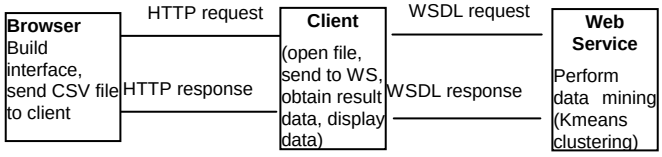


Figure 2: The role of the client.

Theoretically the client can be implemented on almost any computing platform, for the purposes of this experiment a

browser-based solution was written. This relied on Microsoft ASP (Active Sever Page) technology to upload the data and interpret the results.

Data is submitted using a standard browser upload control. The uploaded data must be in CSV(Comma Separated Variable) format. The file stream is broken down and sent to the Web Service as an XML request together with the parameters.

The client waits for the web service to complete the clustering and then receives the output in the form of an XML response In this case the outputted points are plotted with different colors representing the separated clusters. A human can easily interpret this output image.

The client does all processing of data for presentation; this can include all manner of HCI principals as demonstrated by the experiment. These include:-

- File upload is a single stage process provided by the generic file upload control
- Clear presentation of the processed data in the form of a chart.
- Chart colour coded intuitively representing clusters and centroids effectively.
- The ability to identify the location of a centroid and see the table of data corresponding to that cluster.

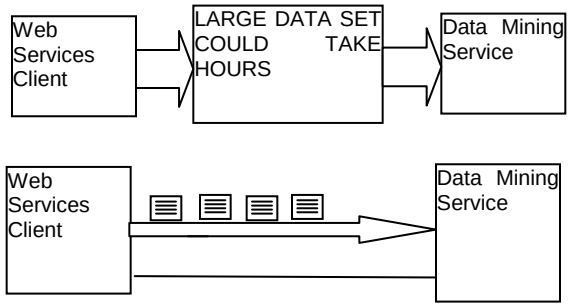
Please refer to appendices for more information on our protoype client website.

3. EVALUATION

While this solution does enable users to data mine in a remote manner it doesn't remove the need for prior knowledge of data mining techniques. The proposed system will only be useful to people who are already competent data miners and are aware of the appropriate algorithms for use in different situations. Users also have to be able to warehouse their own data before it can be send to the web service. This will also probably require a high level of skill.

3.1 Uploading data

One of the problems we encountered HTTP transactions are not ideal for sending and receiving large data large data sets. Since HTTP has a single threaded challenge response framework a large database could take hours to upload in a single go without any indication of progress or success. To avoid this an implementation could be provided where the client was responsible for splitting and sending the data set in smaller chunks and then once data mining was complete the results could be returned to the client in the same manner. New multi threaded Web Technologies such as XMLHTTPRequest could be used to implement progress indicators.



This solution would also allow for the implementation of features such as resuming of a dataset, addition of a dataset, and the ability to save a data set on the server.

The web service could start performing data mining as soon as a chunk of data is uploaded thus giving more instant results which are then refined over time as the amount of data increases.

Thus the web service calls could be :-

uploadDataChunk(data) – *to upload a sensibly sized chunk of data*

setMiningParameters(parameters) - *to upload the mining parameters*

startMining() – *to launch the miner()*

CheckMiningStatus() – *to check the progress of the mining algorithm thus allowing the client to give feedback to the user*

3.2 Redistribution of the data mining task

We have discussed at length how the web service can be can utilize a powerful central server however this idea could be reversed. The web service could be modified to redistribute data to be mined to a distributed cluster of personal computers as was implemented by the SETI (Search for Extra Terrestrial Intelligence) project.

3.3 Efficiency

The web service could be made more efficient by using compression for the transfer of data. This would simplify the upload and reduce wastage. This is especially important because XML is a very verbose and wasteful way of expressing data however it does compress very efficiently.

4. CONCLUSION

Data mining via a web service is an unexploited opportunity in a growing area of computer science. Admittedly it is not something revolutionary simply and evolution made possible by the maturity of web service technologies, XML and different data mining techniques. It is to be expected that many more commercial organizations such as Scientio [5] will start making data mining web services for interested customers to use.

Oracle's leadership in the implementation of the JSR 73 standard outlining an API for data mining proves that the technologies are stabilizing and maturing.

Since the technological solutions are now a proved concept the next stage is to educate the public about the usefulness of data mining that data they collect every day. This will probably be a difficult and slow process initially because the benefits are not immediately obvious indeed when testing out web service users found the output to be "pretty" but didn't seem to tell them much. Large companies and organizations such as supermarkets and governments have leaded the way in this respect and it is to be expected that their progress will filter down to smaller

organizations. This is when data mining web services will be taken very seriously as a concept, a standard and a business.

5. REFERENCES

[1] MSDN Web Service tutorials – *http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnwebsrv/html/webservbasics.asp*

[2] W3C Web Service Spec 2002 - *http://www.w3.org/2002/ws/#drafts*

[3] Web Services Standards for Data Mining http://www.ncdm.uic.edu/workshops/dm-ssp04/web_services_standards.pdf

[4] Java Specification Request 73: Java **Data Mining** (JDM), Version 1.0, Final Review *http://www.jcp.org/en/jsr/detail?id=73*

[5] http://www.scientio.com/XmlMinerWebService.aspx