Proceedings of the 2007 IEEE
International Conference on Integration Technology
March 20 - 24, 2007, Shenzhen, China

# Integrating Data Mining Processes within the Web Environment for the Sports Community

Di Dong and Rafael A. Calvo
*Web Engineering Group*
*School of Electrical and Information Engineering*
*The University of Sydney*
*Sydney, NSW 2006, Australia*
{didong727 & rafa}@ee.usyd.edu.au

***Abstract* - Even though a data mining approach has been successfully adopted to accomplish a number of organizations' marketing goals and objectives in business, it is still in the infancy stage in the domain of sport. However, more and more organizations in sport community have already recognized the power of this modern technology. In this paper, we are going to present a powerful web-based analysis platform, which includes modern Data Mining and Statistics mechanisms, for the sport data analysis. Besides it brings a set of new data analysis methods to this particular field, this system makes it is possible that coaches and athletes are able to access the discovery process and results at "any time" and "any where", because of the web-based property. J2EE, the most popular web application platform, and XML, the standard format of content presentation and sharing on the internet, are adopted as the main technologies to implement our design. Our case study shows that the system works well in a real-world sport organization environment as it is supposed to. Some future challenges and research respective about this particular field are also given.**

***Index Terms - Data mining, sports performance analysis, web application, J2EE, XML***

## I. INTRODUCTION

Nowadays, modern electronic and sensor technologies are used more and more widely in the training and competition of all the different sports. As a result of this progress, huge amount of record data has been collected to support coaches for right strategies and decisions to direct athletes for future competitions. However so far, the major use of these records is limited to basic statistics analyses and only from some aspects of sport sciences, such as biomechanics. So some important patterns of these datasets themselves and relationships among the data may still retain hidden. And there is so much data, it is virtually impossible to effectively analyze it all manually. Even worse, the traditional local-based applications limit the flexibility and accessibility of analysis processes and their results. In this paper, a web-based analysis system, which is equipped with the knowledge discovery and statistics mechanisms, is proposed to help the coaches and researchers in the sports communities to gain more useful support from these sets of historical data.

Data Mining, also known as Knowledge Discovery, which is a process of nontrivial extraction of implicit, previously unknown and potentially useful information [1], can be employed to find out the valuable information that we mentioned above. At the stage so far, we are focusing on individual athlete sports, rather than the team ones. More specific in one sport, our process is seeking for factors that can affect individual's performance critically, quantitative associations between them and a mathematic pattern that can summary these hiding relationships. Different from the traditional sports analysis, this is purely "data - oriented" as no interpretation from the sports discipline will be added.

On the other hand, Web services approach is a distributed computing paradigm and allows applications written in diverse languages, and running on multiple platforms to interoperate and integrate more easily and less expensively than other traditional methods [2]. So comparing to the traditional approaches, adopting a web-based platform brings two strong benefits, "anytime" and "anywhere", to the sports community when they are trying to access the analysis results. And considering from the perspective of software project management, web-based systems require lower implementation cost, less programming skills [3] and less time for development and deployment [4]. Moreover, the maintenance complexity of a web system is much less than the traditional local-based systems. According to these reasons, a web system that is built on the combination of J2EE and XML technologies has been designed as an implementation platform for the knowledge discovery process mentioned above in our research.

The rest of the paper is arranged as follow: the second section gives a brief picture of the data mining in the current sports community. Then our proposed web platform is introduced in the section 3, while the fourth part provides a real-world case study. Finally, the last section concludes the paper with a discussion regarding the present research work and the research respective.

## II. DATA MINING IN SPORTS

Data mining is a process of extracting previously unknown, valid, actionable, and ultimately comprehensible information from large databases and then using the information to make crucial business decisions [5]. From a different perspective, Kotler [6] described data mining as "involving the use of sophisticated statistical and mathematical techniques such as cluster analysis, automatic interaction detection, predictive modeling, and neural

networking". Most of the definitions of data mining fall into these two aforementioned categories. In modern competition sport, from the combination of these two definitions, data mining is considered as the process of using sophisticated mathematical or statistical models to extract valuable, valid, and actionable information from a database to accomplish the sport community's goal that helping and directing athletes to achieve better performance in various competition events. In the recent years, both advances in information technology and organizations' needs have facilitated the upsurge of data mining. Even though a data mining approach has been successfully adopted to accomplish a number of organizations' marketing goals and objectives in business, it is still in the infancy stage in the domain of sport [7].

Although data mining has not been as widely applied in sport as it has in other fields, various successful applications still exist, as more and more sport organizations have already recognized the power of modern data mining technologies. In the world-top basketball league, NBA, it has already been applied by coaches to identify player patterns that box scores do not reveal, which helps win games by extracting relevant information from the database [8]. And the practical results of these applications showed utilizing data mining in this way makes it easier for coaches to make decisions about when and how to position their players for maximum effect [9]. Francett [10] and Hudgins-Bonafield [11] stated that data mining applications help analyze a huge amount of data to reveal winning player combinations for coaches. Moreover, the data mining approach to after-game analysis and improvement takes much less time than the traditional approach—forever rewinding the videotape. Among these applications, the most reputable one is Advanced Scout [12], which was created by IBM T.J. Watson Research Centre. This software seeks out and discovers interesting patterns in game data. With this information, a coach can assess the effectiveness of certain coaching decisions and formulate game strategies for subsequent games. Advanced Scout has been distributed to most of NBA teams and quickly integrated into their game preparation and analytical processes. The positive feedback received from coaching staffs indicates that it's been a valuable tool.

The more interesting thing is that data mining now is not only used on the analysis of athletes, but also on coaches' behavior. In Fast and Jensen's [13] research, they applied knowledge discovery technologies on a complex social network that was formed by the interactions of professional coaches and teams in the National Football League (NFL). According to the analysis on this network, they identified notable coaches and characterized championship coaches. Moreover they utilized the coaching network to learn a model of which teams will make the playoffs in a given year.

The data mining technologies cannot only be used to process the pure data based information, but also multimedia materials. Chen, Shyu and Zhao [14] have used data mining techniques to successfully identify different events with the footage of a soccer match. For instance, they could use their system to find every corner-kick in female soccer videos that resulted in a goal within two minutes. This multimedia data mining tool is extremely powerful and reduces much of the manual labor that coaches would have to go through to find a particular piece of footage.

Although we cannot give more insight to such kind of applications in modern sports fields, the examples above have already illustrated that data mining is recognized as a powerful analysis tool by the sport community now. As we mentioned before, the application of data mining in this particular field is still at the beginning phase, lots of work and research need to be performed to solve some critical issues, such as algorithm adjustment and process definition. Fortunately, researchers and experts from both of these two fields, sport and data mining, have already done more and deeper cooperation on these problems.

## III. System Overview

As the analysis tasks are different in the different sports and organizations, the system should be constructed in an easy-deployment and dynamic manner. In this way, a combination of J2EE and XML technologies is adopted in order to make this system dynamic enough to handle different Data Mining tasks.

As a result of strong industry collaboration, J2EE [15] is based on the experience and expertise from leading designers of industry systems that are scalable, well integrated and provides a cross-platform environment for application development, deployment and management. It standardizes the operating environment for server-side Java applications and provides solid baseline standards on various functional components or containers for presentation and business logic with communication links to client-side presentation, as well as back-end database and legacy systems [16]. In one word, J2EE provides the following benefits: platform independence, multiple vendor support, the popularity of Java in both industries and universities, and a larger number of tools and resources from which to choose. In our particular context, this makes J2EE technology allowing the data mining applications that are written with Java language to be deployed in a wide range of configurations with varying performance and scalability according to the different requirements of different sports and organizations.

The language XML that is designed and developed by the World Wide Web Consortium [17] now is becoming the standard format of content presentation and sharing on the internet. Besides working as deployment properties files of J2EE components, XML is mainly used in the following three aspects in our system:

- Presenting the raw record data that are used as input for the discovery processes.
- Describing complex mining processes, mainly including necessary components, implementation sequence and output format.
- Configuring the function components in each mining process.

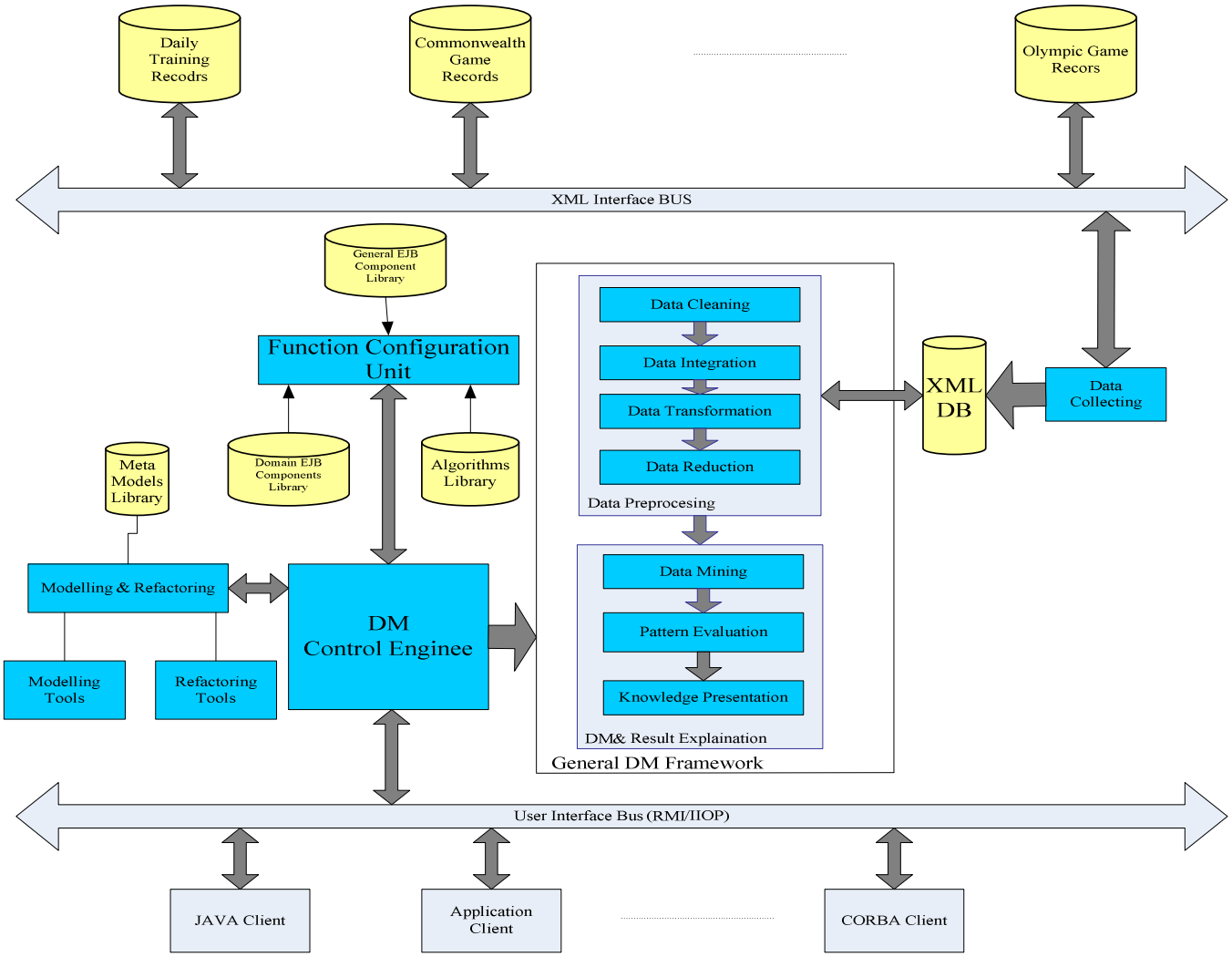


Fig.1 System Structure Overview

The architecture overview of this data mining analysis platform is listed above, and it consists of the following 4 main components:

- **DM Control Engine**: work as a console, main responsibilities include interacting with user input, controlling and coordinating the implementation of various individual components in the system.
- **Modelling & Refactoring**: support the constructing and changing the mining procedures based on the dynamic of mining tasks. The results of these operations are called "Meta-Models", which are represented with XML. These models describe the relationships among mining process, steps, components used and various parameters.
- **Function Configuration**: handle the deployment and configuration of various EJB components according to different "Meta-Models".
- **DM Framework**: provide a general information finding process. The elements in this part are high-level logical virtual components that have a very high granularity. These elements drive one or more low-level functional EJB components to accomplish a specific process step. The relationships between these two sets of components are obtained from the results of Function Configuration.

The four libraries that contain basic knowledge discovery algorithms and function-configuration properties are also critical parts of this online platform.

- **Algorithms Library**: an algorithms package that contains popular and useful data mining algorithms that are widely adopted recently, such as
- **Domain EJB Components Library**: contains the property files of the EJB components that are used to implement specific domain logic in different sport programs.
- **General EJB Components Library**: contains the property files of the EJB components that are used to implement some basic functions, such as message communication.
- **Meta-Models Library**: contains XML files that record the basic and frequently used data mining procedures, such as dimension reduction, clustering and classification. These basal mining procedures are used to construct more complex knowledge discovery processes according to the

specified goals and requirements in the different sports programs and organizations.

## IV. CASE STUDY: ROWING TRAINING IN NSWIS

In this section we are going to give a real-world example of adopting our proposed web analysis system to improve individual rowing performance at the New South Wales Institute of Sports (NSWIS), Australia. The institute, located at Sydney Olympic Park, was established as a statutory body under the Institute of Sport Act, 1995 following a review recommending central coordination and monitoring of high performance sports programs. And today it has almost 700 high performance athletes on squad or individual scholarships and offers 28 sport programs. The organization goal of the institute is ensuring that athletes that are in the New South Wales state-wide have access to leading edge coaching and sports technology while also receiving tailored support to help balance their elite sporting commitments with personal development and a career.

In the rowing sport program, NSWIS has already collect vast amounts of data on each of their athletes during training and on various race-days. Most of the data was produced by a biomechanics rowing performance-monitoring system, named Rowsys2, and fully detailed description of the system and its usage can be found in Richard M. Smith's paper [18]. However, as we mentioned before, it is virtually impossible for a sport institute to do these analysis effectively, because there is so much data and lack of professional data processing experts. Before applying our proposed analysis platform, they only stored the data and ran very simple analyses on it, which typically include basic Microsoft Excel™ graphs that relate two or more variables according to some biomechanics knowledge. Even worse, there was not an effective and efficient way that can deliver such kind of results to the athletes and their coaches, as they were located in several rowing clubs all round the state-wide. As a result, these sets of collected data were not able to aid coaches and trainers to better prepare their athletes for future.

In order to maximize the effect of such a type of data, NSWIS cooperated with us to introduce our online data mining platform into the rowing training field. According the properties of such data sets, we design the following knowledge discovery process to work on them. This process consists of two phases: "**Model Construction**" and "**Quantitative Rules Mining**". In the first phase, the main task is to construct a linear model that can best show the linear relationships between the various record attributes and the performance indicator, boat's average speed. Two knowledge discovery algorithms, Principal Component Analysis (PCA) [19] and Multiple Linear Regression (MLR) [20], are used sequentially here. In order to access the rowing activity comprehensively, 28 attributes have been recorded, which makes the database to have the property of high dimensionality. So PCA, the widely used dimension reduction method, is applied firstly to express the original database with the several most significant principal components that are uncorrelated to each other. Such a reduction has important benefits. First, the computational overhead of the subsequent processing stages is reduced. Second, noise may be reduced, as the data not contained in the first $d$ components may be mostly due to noise. Third, a projection into a subspace of a very low dimension, for example two, is useful for analyzing and visualizing the data, which make these data more clear to the sports community. The number of components, $d$, is determined according to the ratio of variance that can be explained by these new linear combinations of original attributes. After this, it is the right time to introduce MLR to build the linear regression model between these principal components and boat's velocity. As a result, the linear relationships that are contained in the rowing record database will be concluded into a formula as follow:

$$\begin{aligned}\mathbf{AverageSpeed} &= b_0 + b_1 PC_1 + b_2 PC_2 + \ldots + \ldots + b_d PC_d \\ &= b_0 + b_1 (W_{11}A_1 + W_{12}A_2 + \ldots + W_{1n}A_n) \\ &\quad + b_2 (W_{21}A_1 + W_{22}A_2 + \ldots + W_{2n}A_n) + \ldots \\ &\quad + b_d (W_{d1}A_1 + W_{d2}A_2 + \ldots + W_{dn}A_n) \\ &= b_0 + \Sigma^{d}_{i=1} (b_i * \Sigma^{n}_{j=1} W_{ij}A_j) \qquad (1)\end{aligned}$$

And the attributes that have the substantially large loading, $W_{ij}$, in the first few principal components are considered as the critical factors that can affect average boat speed significantly. This is because that the attribute with large loading has a notable influence on the principal component, and if the loading were substantially larger than other attributes', the component would be essentially equal to the attribute with this large loading [21]. The next phase is to look for the quantitative relationships that are contained in the rowing record database. More specifically, we aim to find the Quantitative Association Rules [22] between the critical attributes we found (or their combinations) and the boat's speed. This set of rules also can give evidence to prove the linear model we build in phase 1.

As a result of implementing the above discovery process in our web system, some valuable and unexpected results have been obtained. The new results brought us the new viewpoints of these historical rowing performance data, which are different from the ones gained through the traditional biomechanics analysis. For example, one result indicated the average value of Work made by a rower in one single stroke should be considered as the significant factor that affects the average speed of a boat, which means rowers should optimize their skills to maximize the value of average work per stoke, rather than only focus on increasing some basic biomechanics measurements, such as stroke rate. Another unexpected result showed that max foot force and max force on pin have the great effect on the performance indicator. From the perspective of biomechanics, these two just measure the value at a single time, which cannot be used to explain the change of the boat's average speed. But we know that if the max pin force and foot force of a rower are high, it means the rower may have a good physical status or sophisticated rowing skills. Such kind of rowers certainly can give a good performance,

which means he can make a boat move in a high speed. So these two face attributes can be used as the benchmark by coaches to access rowers' status or select the potential rowers.

As we mentioned before, this web-based system also let the coaches and trainers to access analysis mechanisms and results at any time or any location. In this way, the meetings that were arranged to distribute any analysis results could be canceled, and time and funds that were assigned for such a kind of meeting had been saved. Moreover, as all the information is stored in a sever located in the headquarter of NSWIS, the integrity of analysis processes and results can be maintained easily.

## V. CONCLUSION

In this paper, we present a web-based data mining platform that is used to deal with the historical data in the sport community, and a real world example shows that such a system can bring the benefits of modern knowledge discovery techniques to this particular field. However like the whole picture of applying data mining in sport community, this online system is also in the infancy stage. In order to make it so sophisticated that can accomplish more analysis tasks and give better performance, some research and development work still need to be done.

From the perspective of software system development, these three tasks should be done in order to extend the functionality of the proposed platform:

- Extend the types of presentation devices, such as cell phone, by adding more communication protocol interfaces to the User Interface Bus.
- Extend the existing algorithms library to equip the system with the most recent Data Mining and Statistics algorithms. An "algorithm-adding" interface can be designed to accomplish this target.
- Extend the Meta-Model library with more knowledge discovery meta-processes, in order to make the system has the ability to implement complex mining processes in different sport programs.

Another category of issues falls into the area of data mining models or algorithms. Various models have been commonly and successfully employed to solve other real world problems, and tasks that are performed vary from model to model. Consequently, no rule of thumb exists that explains which model is the best model in solving a practical problem. This situation may be even worse in sports community, as it is still a very short time since we started to apply these technologies. In other words, the selection of the model depends heavily on the type of problems, the data structure of raw records, and the objective of a discovery process. Therefore, it is critical to develop the methodology of constructing a thorough examination of organizational goals and data structure in different sports programs and organizations before choosing data mining techniques.

## ACKNOWLEDGMENT

Di Dong is supported by a Norman I Price scholarship. And the authors gratefully acknowledge the solid support from Kenneth Graham (manager of sport science services) and Margy Galloway (coordinator of the Technical Analysis department) in the New South Wales Institute of Sports, Australia.

## REFERENCES

[1] W. J. Frawley, G. Piatetsky-Shapiro and C. J. Matheus, "Knowledge Discovery in Databases: An Overview", in G. Piatetsky-Shapiro and W. J. Frawley (eds.), *Knowledge Discovery in Databases,* AAAI/MIT Press, pp.1̃27, 1991.

[2] M. Haines, "Levels of Web Services Adoption: From Technical Solution to Business Opportunity," presented at *Ninth Americas Conference on Information Systems*, Tampa, Florida, USA, 2003.

[3] Y. Huang and J. Chung, "A Web Services-based Framework for Business Integration Solutions," *Electronic Commerce Research and Application*, vol. 2, pp. 15-26, 2003.

[4] P. Ratnasingam and P. Pavlou, "The Role of Web Services In Business To Business Electronics Commerce," presented at *Eighth Americas Conference on Information Systems*, Dallas Texas, 2002.

[5] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, & A. Zanasi. *Discovering data mining: From concept to implementation*. NJ: Prentice Hall, 1998, pp. 189.

[6] P. Kotler, *Marketing management*. (11th ed.). New Jersey: Upper Saddle River Pearson, Education, Inc. 2003, pp.54.

[7] L. Fielitz, & D. Scott, "Prediction of physical performance using data mining". *Research Quarterly for Exercise and Sport*, vol. 74, no. 1, pp. 25, 2003.

[8] C.Y. Chen, Y.H. Lin, "A New Market Research Approach in Sport – Data Mining", *The Sport Journal*, vol. 9, no. 3, pp. 46-51, July 2006.

[9] H. Baltazar, "NBA coaches' latest weapon: Data mining". *PC Week*, vol. 17, no. 10, pp. 69, 2000.

[10] B. Francett, "The NBA gets a jump on data mining". *Software Magazine*, vol. 17, no. 9, pp. 24-25, 1997.

[11] Hudgins-Bonafield, C., "Data mining software scores high with the NBA". *Network Computing*, vol. 8, no. 11, pp. 50, 1997.

[12] I. Bhabdari, E. Colet, J. Parker, Z. Pines, R. Pratap, & K. Ramanujam, "Advance Scout: Data Mining and Knowledge Discovery in NBA", *Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 121-125, March 1997.

[13] A. Fast, and D. Jensen, "The NFL coaching network: analysis of the social network among professional football coaches". *AAAI Fall Symposium on Capturing and Using Patterns for Evidence Detection*, 2006

[14] S.C. Chen, M.L. Shyu, and N. Zhao, "An enhanced query model for soccer video retrieval using temporal relationships", presented at *21st International Conference on Data Engineering,* 2005.

[15] Jiang Guo; Yuehong Liao; B. Parviz. "A performance validation tool for J2EE applications". *13th Annual IEEE International Symposium and Workshop on Engineering of Computer Based Systems*, 2006.

[16] Borland, *Performance Management for J2EE*, 2003.

[17] "Extensible Markup Language (XML) 1.0." http://www.w3.org/TR/REC-xml, 1998.

[18] R.M. Smith, C. Loschner, "Biomechanics Feedback for Rowing", *Journal of Sports Sciences,* vol. 20, no. 1, pp. 783-791, 2002

[19] I. T. Jolliffe, *Principal Component Analysis*. Springer-Verlag, New York, pp. 2, 1986.

[20] B. G. Tabachnick, L. S Fiedell, *Using Multivariate Statistics*. Allyn& Bacon, US, pp. 139, 2000.

[21] A. C. Rencher. *Methods of multivariate analysis*. Wiley, New York, 1995, pp. 423.

[22] R. Srikant, R.Agrawal, "Mining Quantitative Association Rules in Large Relational Tables", *Proceedings of the 1996 ACM SIGMOD international conference on Management of data. Montreal*, Quebec, Canada, vol. 2, no. 1, pp. 1 – 12 , 1996.