South African Journal for Research in Sport, Physical Education and Recreation, 2006, 28(1): 109-125.
Suid-Afrikaanse Tydskrif vir Navorsing in Sport, Liggaamlike Opvoedkunde en Ontspanning, 2006, 28(1): 109-125.
ISSN: 0379-9069

THE VIABILITY OF BUSINESS DATA MINING IN THE SPORTS ENVIRONMENT: CRICKET MATCH ANALYSIS AS APPLICATION

Johan H. SCHOEMAN, Machdel C. MATTHEE & Paul VAN DER MERWE
Department of Informatics, University of Pretoria, Pretoria, Republic of South Africa

ABSTRACT

Data mining can be viewed as the process of extracting previously unknown information from large databases and utilising this information to make crucial business decisions (Simoudis, 1996: 26). This paper considers the viability of using data mining tools and techniques in sports, particularly with regard to mining the sports match itself. An interpretive field study is conducted in which two research questions are answered. Firstly, can proven business data mining techniques be applied to sports games in order to discover hidden knowledge? Secondly, is such an analytical and time-consuming exercise suited to the sports world? An exploratory field study was conducted wherein match data for the South African cricket team was mined. The findings were presented to stakeholders in the South African team to determine whether such a data mining exercise is viable in the sports environment. While many data constraints exist, it was found that traditional data mining tools and techniques could be successful in highlighting unknown patterns in sports match data. However, it is questionable whether this type of data mining is viable in this industry. People in the sports world often do not have the time or the required expertise to acquire, interpret and use the results.

Key words: Business data mining; Cricket match analysis; Knowledge discovery.

INTRODUCTION

The continuous progression of technology has resulted in the ability of businesses, including sports businesses, to gather and store large quantities of data, which consists of a huge amount of potential information. However, our ability to gather and store data has far surpassed our ability to analyse and extract useful information from this data successfully (Fayyad *et al*., 1996). This problem has a direct impact on the ability of businesses to compete successfully, and more efficient techniques with a view to knowledge discovery are therefore required.

Data mining is such a knowledge discovery approach that has been identified as an enabler for businesses to tap the limitless information potential from the vast amounts of data that they collect (Apte *et al*., 2002). This "process of extracting valid, previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decisions" (Simoudis, 1996: 26) can have a direct influence on the success or failure of a business (Chopoorian *et al.*, 2001).

The potential of data mining in sports should be explored with a view to the possibility of applying previously unknown information to influence the success of a team. Information unknown to others has the potential of being employed strategically in order to gain an advantage in this very competitive environment.

The purpose of this paper is to explore the use of data mining on sports data in order to assess whether this practice has any viability in this particular type of business. The following section will describe the research design, while the next section will look into data mining and the data mining process. The *status quo* of data mining in sports is then discussed. After that, the process followed to mine cricket data in order to test its applicability is described. The last section discusses the findings of this study.

RESEARCH DESIGN

The main research question that this study wishes to address is to determine whether proven business data mining techniques can be applied to sports games in order to discover hidden knowledge. In addition, it explores whether such an analytical and time-consuming exercise is suited to the world of sport.

To answer the above questions, an interpretive research approach was followed. Interpretive research, as explained by Klein and Meyers (1999: 69) assumes "…that our knowledge of reality is gained only through social constructions such as language, consciousness, shared meanings, documents, tools, and other artefacts". Unlike positivist research, interpretive research does not attempt to describe the world in terms of dependent and independent variables, but does so by considering the individual, different views of the people involved.

A field study was conducted in which the application of a traditional data mining technique and tool was tested on cricket data for the South African national cricket team. This study is of an exploratory nature since the researcher created a demonstration project without the involvement of the users. Only after completion of this study, were the stakeholders presented with the findings and interviewed to determine the viability of data mining in cricket.

DATA MINING

Data mining is effectively an extension of traditional data analysis approaches, which have been combined with advanced computing technology in order to gain knowledge from large amounts of data more successfully. However, one important difference between traditional techniques and data mining approaches exists. Traditional techniques rely on a hypothesis-driven method, which is used in finding the answers to specific questions. While traditional, verification-driven techniques have a place in data mining, data mining techniques often use a data-driven approach in which patterns in the data are detected without forming a specific hypothesis beforehand (Jackson, 2002). The use of such discovery-driven techniques makes data mining an attractive option, since it enables businesses to gain information from the massive amounts of data that are being collected.

Any data mining effort will be conducted in terms of two dimensions. First of all, the technique that is selected to discover new information should be selected. Secondly, regardless of the selected technique, a particular process or methodology should be followed to mine the data.

Data Analysis and Mining Techniques

Verification-driven analysis techniques (Simoudis, 1996) include q*uery and reporting* (the validation of hypotheses decided on by the user, through the query and reporting function of an information system), m*ultidimensional analysis* (the validation of hypotheses through the use of complex Online Analytical Processing queries) and s*tatistical analysis.*

Discovery-driven data mining techniques include *predictive modelling and classification,* which include techniques that use complex mathematical and statistical algorithms to predict future behaviour based on continuous, historical data (Peacock, 1998a). Another technique is *clustering* which refers to the automatic partitioning of a database into subsets of related records in order to draw conclusions from clusters with similar properties (Hirji, 2001). *Link Analysis*, the technique used in this study, seeks to establish relations (links) between different items in the database (Simoudis, 1996). For example, the use of link analysis could establish that when people buy beer, they usually buy peanuts and chips, too.

The Data mining Process

The data mining process refers to the actual steps that the analyst goes through to apply the selected technique in order to discover new information from data. While various versions of the process do exist (see Brachman *et al.*, 1996; Fayyad, 1998; Peacock, 1998b), one such process methodology, CRISP-DM (Cross-Industry Standard Process for Data Mining), has become the industry standard (Jackson, 2002). This methodology by Chapman *et al.* (2000) consists of several steps and is iterative in nature:

- *Business Understanding:* This phase of the methodology focuses on gaining an understanding of the project objective and requirements from a business point of view. These requirements are then translated into a data mining problem definition, and a preliminary project plan is formulated to achieve the established objectives.
- *Data Understanding:* This phase involves the initial collection of data, becoming familiar with the data, identifying data quality problems, gaining insight into the data and identifying interesting subsets of data that could lead to discovering hidden information.
- *Data Preparation:* In this phase the initial data that was collected is converted into the final dataset that will be used in the modelling phase. This process typically involves the selection of applicable fields and data items, cleaning this data and transforming it into an appropriate format, as required by the analysis and specific data mining tool.
- *Modelling:* In this phase the data mining techniques are applied. Selection of a specific technique would typically depend on factors such as the type of data, as well as the data mining and business objectives. The selected technique would often require a return to the data preparation phase to transform the data into a format that is better suited to the selected technique.
- *Evaluation:* This phase is concerned with reviewing the models built in the previous phase to determine whether they are of a high quality. The results of the models will be evaluated and a decision as to whether these findings should be used in the business has to be made. In addition, it is important to determine whether this whole process has been successful in meeting all the business objectives identified in understanding the business phase.
- *Deployment:* At this part of the process the findings of the data mining effort have to be organised and presented in a format that the business could use. This step could simply

involve drawing up a report with findings, which would be presented to the business users, or it could be the implementation of a repeatable data mining process.

DATA MINING IN SPORT

A great deal of research is available on the analysis of sports data. However, much of this research is hypothesis-driven, thus being aimed at answering a specific question rather than discovering and using unknown information (e.g. Dunn & Syrotuik, 2003; Rahnama *et al.*, 2002; Bartlett, 2003; Field *et al.*, 2003).

Very little data mining has been conducted in sport so far. For example, Delmater and Hancock (2001) identify the use of data mining in customer service, retail, insurance, financial services, health care and medicine, telecommunication, transportation and logistics, energy and government. While the sports industry is not mentioned, the use of data mining techniques tested in some of the areas above could also be extended to sport. Furthermore, data mining on a sports game itself is still a relatively new concept that deserves exploration. As in any business, the sports world is faced with an increasing amount of data and insufficient means with which to analyse it. A few studies have begun to explore data mining techniques on sport games. Bracewell *et al*. (2003) mined data on rugby games to evaluate individual performance. In addition, systems that help with the analysis of sports video footage are being employed in many professional sports for both coaching and strategic purposes (Newell, 2002). One can question whether such systems are simply a more efficient way of analysing video footage, or whether they are geared and used for discovering unknown information. Zhou *et al*. (2002) developed a sport video analysis system that uses several data mining techniques including clustering, in order to help deal with all the available video data. In conclusion, while the data mining of sports data is being touched upon, it is still a new application area and requires a great deal of research to determine its viability and applicability.

DATA MINING APPLICATION

In order to explore the applicability of data mining on sports data, a business data mining application was tested on one-day international cricket data. A proof-of-concept approach was followed as indicated below:

- The authors acquired and mined cricket data in order to test whether the mining of cricket data is applicable and to find initial patterns.
- The findings were presented to the South African cricket team to determine whether such a data mining approach has any viability in the case of cricket.

The business for which the data mining was done is the South African national cricket team, which forms part of Cricket South Africa (Pty) Ltd. An analyst would ideally be intricately involved with the business users to gain a business understanding; but due to the exploratory nature of the research, this was not possible. However, it is not entirely impossible to gain an understanding of this particular business area, since the team's operations are quite well known to the public because of wide media coverage. While it did require certain assumptions to be made in terms of the business and its objectives, this data mining exercise was not primarily intended to add value to the business, but to provide an example of what further

formal data mining projects could do. This section describes the data mining process followed, which was based on the CRISP data mining methodology.

Cricket: A business understanding

The business objective of this data mining exercise was gaining a better understanding of cricket matches in order to discover new information that could assist the South African team in becoming more successful. The discovery of information that could be used strategically by the team particularly led to the following data mining objectives being set:

- To analyse the cricket game (ball-by-ball or over-by-over) itself, rather than review disparate averages of the game.
- To determine strengths and weaknesses of individual teams and players.
- To analyse the performance of a specific player against another player.

Understanding data

One of the main problems with the use of business data mining applications on sports games is that they require data in a traditional database format. However, when considering sports matches, data is mainly collected in video format. While cricket is by nature data intensive, this data is still not in a database format that actually describes the game, but generally in the form of averages or media reports. Therefore, an alternative source of data was required. The closest source that is already in a traditional text format was found to be text transcriptions of matches that have been 'broadcast' over the Internet for the last few years.

End of over 1 (maiden) West Indies 0/0 (RR: 0.00)
SM Pollock 1-1-0-0 - Wynberg End
WW Hinds 0* (0b) CH Gayle 0* (6b)

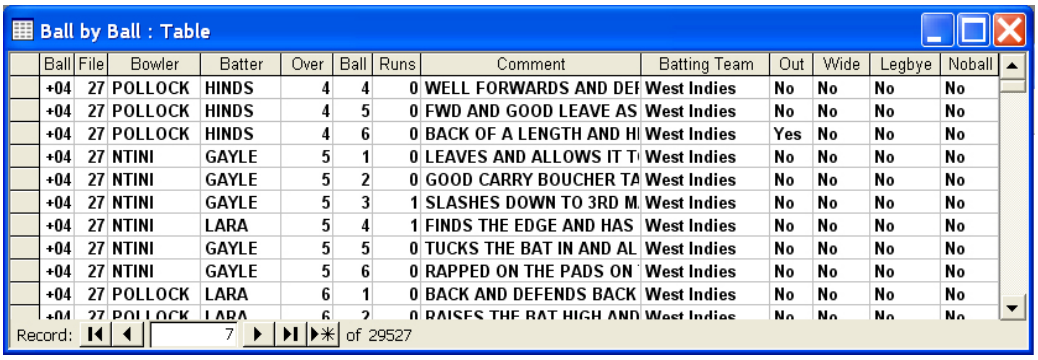
Ntini will have the honour from the Kelvin Grove end
1.1 Ntini to Hinds, no run, oohs and aahs from the crowd as it goes past the outside edge
1.2 Ntini to Hinds, no run, beaten for pace as he played late and Boucher takes
1.3 Ntini to Hinds, no run, stays low and skids through to the keeper
1.4 Ntini to Hinds, no run, pushing through at 140 km/h going across him
1.5 Ntini to Hinds, no run, fuller and Boucher takes as it raced through
1.6 Ntini to Hinds, no run, another Damsel as this was once again in the channel and he leaves

End of over 2 (maiden) West Indies 0/0 (RR: 0.00)
M Ntini 1-1-0-0 - Kelvin Grove End
CH Gayle 0* (6b) WW Hinds 0* (6b)

FIGURE 1. **EXCERPT FROM A CRICKET BALL-BY-BALL REPORT** (CRICINFO, 2003)

Figure 1 displays an excerpt from one such report. While these reports are in text format and provide a 'history' of a game, they are only standardised to a certain degree and contain only a limited amount of information. A program that converts these transcriptions had to be written

to convert it into a database format, as displayed in Figure 2**.** While a sport like rugby is more continuous with no definite beginning and end to events, cricket is based on the recurring event of a ball being bowled to a batsman who attempts to hit it. In this sense, cricket is probably better suited to the use of business data mining applications, since such an event is similar to typical business event, such as a sales transaction.


Ball by Ball : Table

Ball	File	Bowler	Batter	Over	Ball	Runs	Comment	Batting Team	Out	Wide	Legbye	Noball
+04	27	POLLOCK	HINDS	4	4	0	WELL FORWARDS AND DEI	West Indies	No	No	No	No
+04	27	POLLOCK	HINDS	4	5	0	FWD AND GOOD LEAVE AS	West Indies	No	No	No	No
+04	27	POLLOCK	HINDS	4	6	0	BACK OF A LENGTH AND HI	West Indies	Yes	No	No	No
+04	27	NTINI	GAYLE	5	1	0	LEAVES AND ALLOWS IT T	West Indies	No	No	No	No
+04	27	NTINI	GAYLE	5	2	0	GOOD CARRY BOUCHER TA	West Indies	No	No	No	No
+04	27	NTINI	GAYLE	5	3	1	SLASHES DOWN TO 3RD M	West Indies	No	No	No	No
+04	27	NTINI	LARA	5	4	1	FINDS THE EDGE AND HAS	West Indies	No	No	No	No
+04	27	NTINI	GAYLE	5	5	0	TUCKS THE BAT IN AND AL	West Indies	No	No	No	No
+04	27	NTINI	GAYLE	5	6	0	RAPPED ON THE PADS ON	West Indies	No	No	No	No
+04	27	POLLOCK	LARA	6	1	0	BACK AND DEFENDS BACK	West Indies	No	No	No	No
+04	27	POLLOCK	LARA	6	2	0	RAISES THE BAT HIGH AND	West Indies	No	No	No	No

Record: 7 of 29527

FIGURE 2. **EXAMPLE OF BALL-BY-BALL REPORT CONVERTED INTO DATABASE FORMAT**

Data preparation and modelling

The focus of this particular data mining exercise was on the South African (SA) and West Indian (WI) cricket teams. As this was the next series to be played by South Africa, this data was selected to render it meaningful to the stakeholders. The games that were imported into the database are given below:

- The games between SA and WI (Seven One Day International (ODI) Games between SA and WI in the 1998/1999 tournament that took place in South Africa, seven ODI Games between SA and WI in the 2001 tournament that took place in the West Indies). This gives an indication of the two teams performing against each other.
- 11 Games that the two teams played in the 2003 Cricket World Cup which took place in SA. This gives an indication of effects that the venue and weather conditions have on the performance of either country.

The final data consisted of four sets:

- A ball-by-ball analysis that made up 29527 rows against 59 fields in the database;
- An over-by-over analysis that made up 4787 rows against 26 fields;
- A summary of each batsmen's performance per game;
- A summary of each bowler's performance per game.

The data mining technique that was selected for the purpose of this exercise was link analysis. In particular, a visual link analysis application, Netmap, was selected for the following reasons:

- It is well suited to the categorical data of which the database consists predominantly.

- It is not as technical as many of the other data mining techniques and applications and is better suited to users without training in information technology and data mining.
- Whilst link analysis often relies on statistical measures to identify associations, NetMap uses logical, visual representations to enable the users to do so. This visualisation aids the identification, interpretation and understanding of patterns. Visualisation, in effect, reduces the difficulty of understanding by providing perceptual support for cognitive processes (Card *et al*., 1999).

The data preparation and modelling steps that were taken in the mining of the data are as follows (see also Figure 3):

1. A utility was written to extract the ball-by-ball information from the text files downloaded from the Internet. Each file was imported into an Access database. The utility compiled two tables:
    a. One with ball-by-ball data, which resulted in the 29527 rows. This contained information about who bowled each ball, who batted against it, how many runs were scored, what type of runs were scored from the ball, etc.
    b. One with over data, which resulted in the 4787 rows. This contained information on the game status at the end of each over, which included the total runs scored, the number of wickets that had fallen, the batting team and the run rate.
2. This initial data was enhanced to a great extent, to include a great number of additional descriptive and summarised attributes to enhance the filtering capabilities of Netmap. This included adding descriptive attributes to the game that could not be extracted automatically from the files, like the venue, the date, whether it was day or day night game, whether the result was influenced by the Duckworth-Lewis method, etc. In addition, derived figures and statistics such as the total runs scored by a batsman, the runs scored per over and the run rate were compiled. Similarly, figures and statistics were derived for the bowlers which included extras awarded, bowling rates, etc.
3. All this data was categorised into ranges in order to suit the tool's filtering capabilities. For example, instead of importing a "Total Runs Scored" field into Netmap, which would result in a great number of unique values, it is more valuable to band these results into logical bands such as 0-50, 51-100, etc.
4. Finally, various (about 17) studies (data subset or a selection of rows and attributes) were imported into Netmap through the use of another automated tool. This tool translates the selected fields within a traditional DBMS into the proprietary format used by Netmap. A trial-and-error approach to building these studies was followed, in which the combination of attributes selected for the study was changed various times to produce more meaningful results.
5. Once a study is within Netmap, the tool's filtering and display algorithms can be used to search for patterns within the data, which are shown within the paper. A study would often lead to questions that require another view or categorisation, which would require enhancements to the data itself and to the subsequent building of additional Netmap studies.

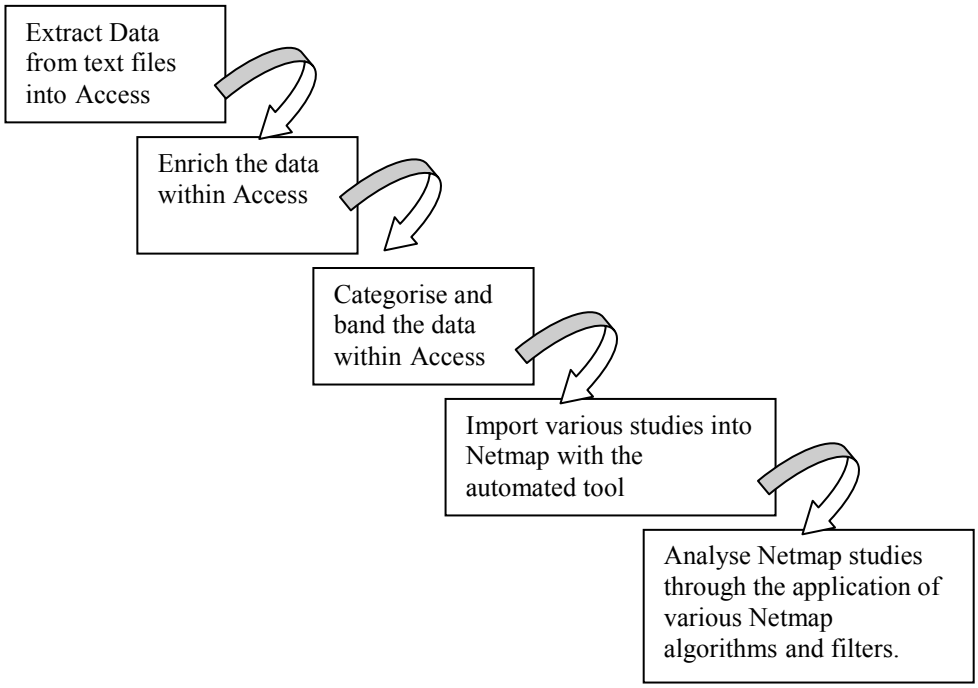


FIGURE 3. **DATA PREPARATION AND MODELLING STEPS**

A number of iterations between the data preparation, modelling and evaluation steps resulted in several models, of which some will be discussed in the next section.

Evaluation

Quite a few models were created from the Netmap studies of which only a few examples are presented below. A top-down analysis approach was followed in this data mining exercise, wherein patterns on a general, game level were first sought out, then became more focused on the over-by-over and consequently ball-by-ball events of the game.

Figure 4 displays a typical *netmap* with which the user can identify links between different data items, or nodes. This model illustrates a summary of the matches between South Africa and West Indies. As one can see, there are distinct relationships (the bold triangles) between SA being the winner, WI being the loser, most of the games being day games and none of the matches between the two countries being influenced by the Duckworth-Lewis system for resetting targets of rain-interrupted games. In addition, a secondary pattern can be detected underneath the extremely bold lines.

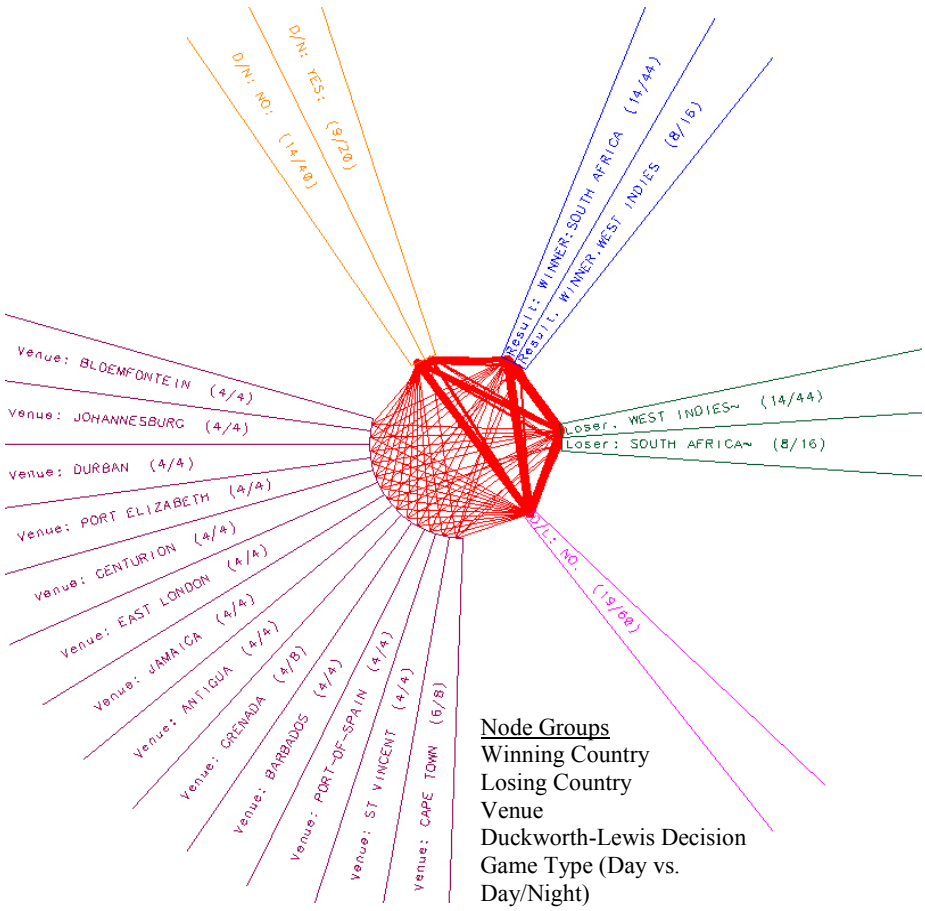


FIGURE 4. **SA VS. WI**

Figure 5 displays this underlying pattern. This model depicts the day/night (D/N) games between the two countries. Here, one can see that there is a strong inclination for SA to win day/night games. This pattern was confirmed by the stakeholders of the cricket team that were interviewed and was said to be a result of WI not being used to playing night games since they do not have suitable lighting facilities at the stadiums in their countries. Although cricket experts probably already know this pattern, this can be seen as a 'new finding' since the tool 'discovered' and highlighted this fact which urge one to further investigation. Such knowledge can then be strategically used to schedule games in one's favour. (In fact, whether this decision was made because of increased ticket sales and coverage, or the knowledge of this pattern, four out of the five one-day games scheduled for the next series between the two countries were day/night games).

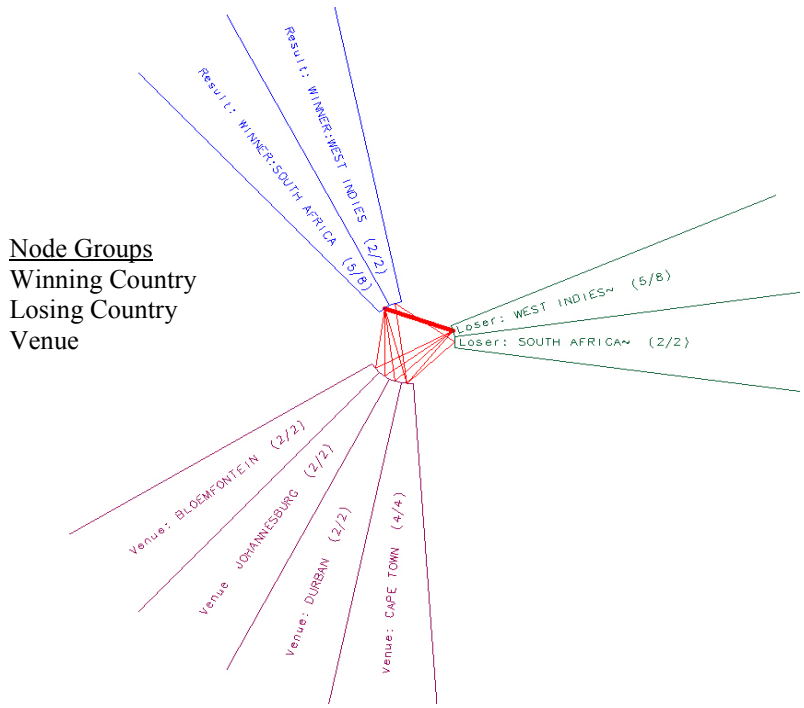
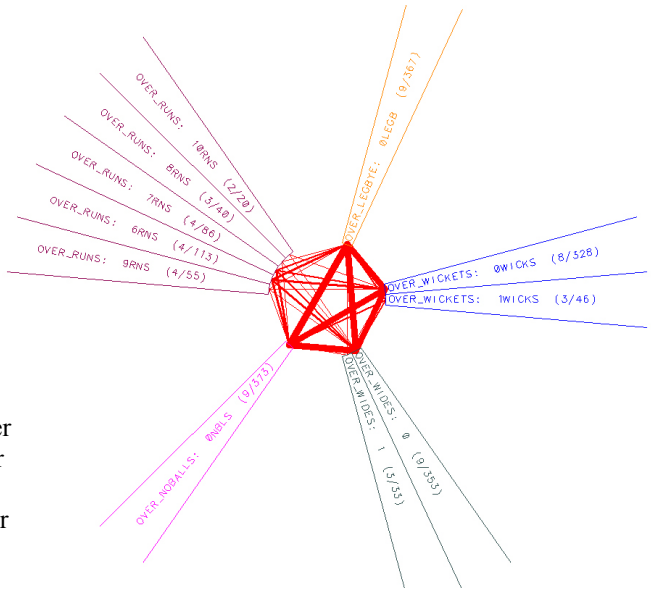


FIGURE 5. **SA VS. WI DAY/NIGHT GAMES**

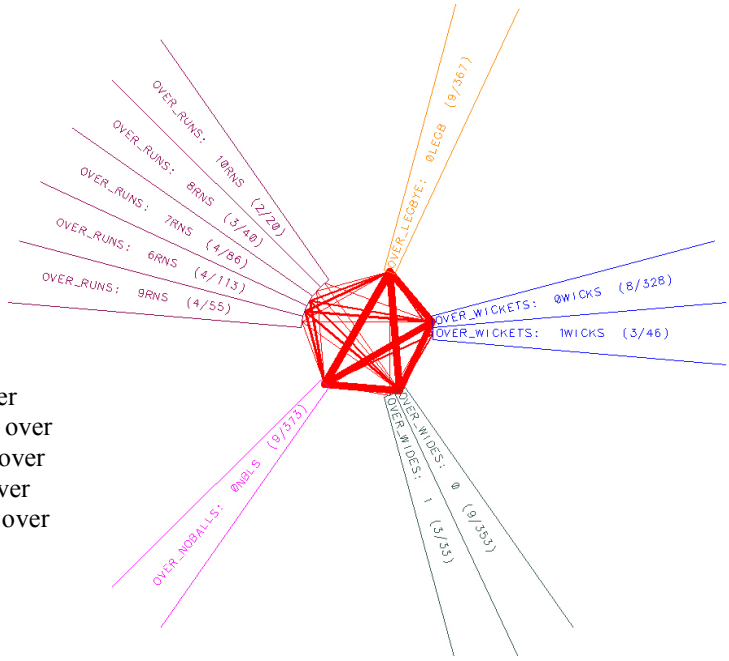
The games are considered in over-by-over detail in Figure 6 and shows the overs in which WI scored more than five runs, when SA bowled. Likewise, Figure 7 shows the overs in which WI scored less than four runs against the bowling of SA. Conventional cricket knowledge would suggest that when a team's bowling is inconsistent in terms of no-balls and wides, the batting team would be able to score more runs, especially since penalty runs are awarded to the batting team. Similarly, when a team's bowling performance is good in terms of no-balls and wides, one would expect the batting team to score less runs. This, in fact was true for SA's batting against WI's bowling. However, it was found that the opposite is true for WI's batting against SA's bowling. SA's bowling was more consistent in the overs that WI scored many runs and quite inconsistent in the overs where WI scored few runs, as indicated by no-balls and wides.

These findings were unknown to the interviewed stakeholders of the cricket team and warrant further exploration. While this pattern could certainly be due to a chance occurrence in the data or other factors not presented in this model, it could also give insight into the mentality of the WI team. WI might not feel the need to score runs when no-balls and wides are bowled, since the bowling team concedes penalty runs. However, this typically results in relatively 'small' overs, which is better for SA in the long run.



Node Groups
Runs in the over
Legbyes in the over
Noballs in the over
Wides in the over
Wickets in the over

FIGURE 6. **SA BOWLING & WI BATTING, > 5 RUNS/OVER**

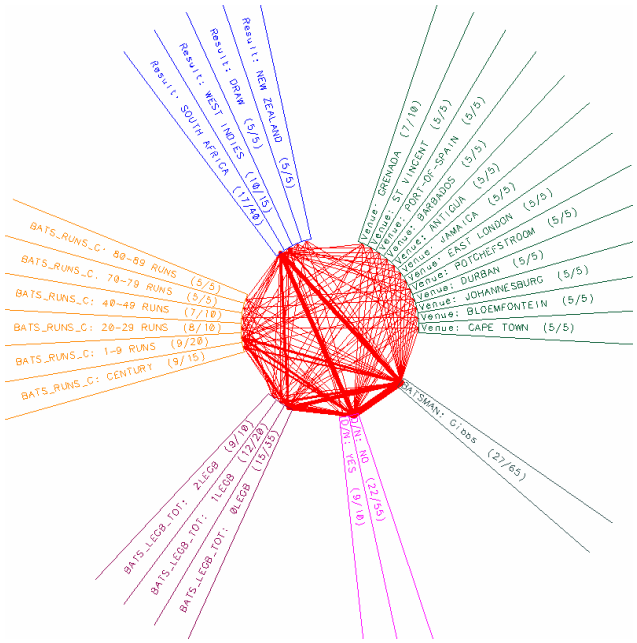


Node Groups
Runs in the over
Legbyes in the over
Noballs in the over
Wides in the over
Wickets in the over

FIGURE 7: **SA BOWLING & WI BATTING, <4 RUNS/OVER**

The batting performance of a specific SA batsman, Herschelle Gibbs, is considered in Figure 8. While it is known that Gibbs tends to either achieve very low (1-9 RUNS) or very high scores (a CENTURY), one can also see that there is no outstanding pattern between his low or high scores and South Africa's tendency to win or lose. Therefore, while he is a valuable player, the SA team is not totally dependent on his good performance in order to win.

While the patterns indicated by Gibbs' batting profile are probably well-known to experts, this profile can be used as a model with which to compare other players who are not as well known. For example, a study was done on the batting profile of a WI batsman, Chris Gayle. It was found that while he also has a tendency to get out before scoring many runs, he does not score centuries as often as Gibbs. In addition, WI did not win many of the games in which he played, especially when he failed to score many runs. Therefore, WI might be more dependent on his success than SA is on Gibbs' performance. Such information could, for example, be used to strategically work out a game plan against a specific batsman.



Node Groups
Winner
Venue
Batsman
Day/Night
Batsman leg bye total
Batsman run total

FIGURE 8: **SA BATTING: GIBBS**

Deployment

Due to the exploratory nature of this project, the deployment phase of this data mining exercise involved the demonstration of models and findings such as the above to three different stakeholders involved with cricket and the national team. The main purpose of these demonstrations and interviews was to determine the viability of NetMap as data mining tool in the analysis of cricket data. The following questions were asked:

- Does such a data mining exercise have any value?
- What changes could be made to the models to improve them?
- Are all of the findings shown already known?
- Are such findings useful at all?
- What will be the difficulties and problems surrounding the adoption of Netmap as data analysis tool?

The findings from the interviews are summarised in Table 1.

DISCUSSION OF FINDINGS

Potential uses for business data mining techniques in the analysis of sports data

The interviewees made it clear that various forms of data analysis are being done in cricket. Firstly, coaches and team members consider various statistics that are compiled and published on the Internet. Secondly, applications that record and file video footage on a ball-by-ball basis are also widely employed in cricket. Finally, applications that use virtual reality to help with training and analysis are also being developed and used.

The interviews resulted in various, and often, conflicting opinions regarding the viability of data mining in cricket. While the director of coaching was extremely positive about the use of data mining tools (specifically Netmap) and techniques in cricket, the national team's coach and analyst seemed more sceptical regarding the actual use of such an application to serve their purposes. However, all three of the participants interviewed in the deployment phase admitted that they were not aware of many of the findings demonstrated to them (see the discussion on the analysis of SA's bowling and WI's batting as well as the performance of WI in day/night games). This demonstrates that data mining techniques and applications that are typically employed in the business world could be successful in discovering unknown information when applied to sports matches. However, to make data mining efforts successful one must not merely uncover hidden information in large quantities of data, but must also be able to use it. In terms of these criteria and incorporating some of the comments of the interviewees, the findings from the data mining process can potentially be used in the following areas of cricket:

- The strategic scheduling of cricket matches
- The profiling of teams or players for purposes of team selection
- Highlighting strengths and weaknesses of players who can be utilised or exploited
- The formulation of a team's playing strategy

While sports analysts would typically focus more on technique or playing strategy, such an analysis could especially be effective in highlighting underlying patterns that would not be noticed from the normal viewing of video footage. In other words, while video footage could

be valuable in discovering the "know-how" of cricket, such an exercise could help to uncover the "know-what" and "know-why" of the game. In addition, it provides a useful way of dealing with the ever-increasing amount of data available. Therefore, this type of analysis could be a complementary addition to the more traditional sports analysis techniques e.g., it was proposed by one of the interviewees to use data mining as complement to video analysis.

TABLE 1. **SUMMARY OF FINDINGS FROM INTERVIEWS**

	Interviewee 1	**Interviewee 2**	**Interviewee 3**
Position	Director of coaching for the South African Cricket Board.	Coach of SA national cricket team.	Data analyst of SA national cricket team.
Perceived value:	Considerable value. It is considered more effective than ordinary statistical analysis due to identification of unknown patterns and visual representation. However, it should be complemented by video analysis.	No real value (e.g. it is of no use telling Gibbs that he often gets out on the fourth ball of the over as this would interfere with individual playing strategies).	Limited value
Proposed changes:	Data should include weather conditions, pitch reports, ball speeds, who won the toss, whether they selected to bat or bowl first.	He needs an analysis method that allows the viewing of player technique, bowling actions, batting shots played, and fielding.	He prefers the current video-based application that allows viewing of elements in game, and which gives descriptive statistics such as counts and averages.
Novelty of findings:	There are unknown findings (e.g. Figure 6 and 7)	There are unknown findings (e.g. Figure 6 and 7)	There are unknown findings (e.g. Figure 6 and 7)
Usefulness of findings:	Can be used in playing strategy, team selection and management strategy.	Will be of limited use.	Useful to him and national coach only (analysis and pattern detection)
Adoption of Netmap:	Netmap is very simple to adopt and use because of visual representation and focus on findings and not on the data mining process itself.	No comment	It will be difficult to use when working with players (not all of them have appropriate training in data analysis and technology).

Obstacles to the adoption of business data mining as sports data analysis tool

One should also question the ability of the sports world to use data mining applications effectively. To be effective, data mining requires people with specific skills that might not be available in the world of sports (this was confirmed by one of the interviewees). The world of business as regards cricket in particular, is typically made up of sportsmen who either play the game or have moved on to administrating or coaching cricket. If the traditional business world is still often struggling to mine their data effectively (Schoeman *et al.*, 2003), can the sports world be expected to do so successfully?

FURTHER RESEARCH

This study was limited in several ways: Only one business data mining tool was applied to the analyses of only cricket match data. Also, due to the exploratory nature of this project, it was not possible to involve the business users, in this case the cricket team and its management in the specification of the business objectives and data mining requirements. A certain lack of direction with regard to the business and data mining objectives therefore existed.

However, despite the exploratory nature of the study it became clear that while the potential for business data mining techniques and applications to discover unknown and usable information from sports data does certainly exist, the viability thereof in the sports industry can be questioned. A reason not mentioned by any of the interviewees is that the transformation of an actual match into the data format required by an application designed for business databases, could be problematic. This is true especially for less data-intensive sports. Data mining applications that use the actual video footage of the game, rather than a textual recreation thereof, are probably more suited to the sports world. However, in order for such an application to be a data mining tool, it would have to be based on some type of data mining technique, as discussed above. In addition, if the analysis is being done on a hypothesis-driven, rather than data-driven basis, the possibility of discovering unknown information is left to chance.

It thus seems as if the adaptation of proven data mining techniques to multimedia data could prove fruitful to the sports industry. However, such applications will probably be limited to the information contained in the actual game itself. An advantage of text-based applications is the ability of the analyst to transform and enrich the data in order to uncover patterns that would normally not be noticed. Therefore, a case for the use of such business applications in sport does still exist.

CONCLUSION

If the data problems particular to the sports environment can be overcome, the possibility of using traditional data mining techniques in order to discover unknown knowledge does indeed exist. Such information could provide a strategic advantage over other teams. However, video-based analysis techniques, while being limited in many ways, are probably more suited to the world of sports, where professional skill is based on physical technique, rather than statistical, technical and academic expertise.

REFERENCES

APTE, C.; LIU, B.; PEDNAULT, E.P.D. & SMYTH, P. (2002). Business applications of data mining. *Communications of the ACM*, 45(8): 49-53.

BARTLETT, R.M. (2003). The science and medicine of cricket: An overview and update. *Journal of Sports Sciences*, 21(9): 733-753.

BRACEWELL, P.J.; MEYER, D. & GANESH, S. (2003). Creating and monitoring Meaningful Individual Rugby Ratings. *Research Letters in the Information and Mathematical Sciences,* 4: 19–22.

BRACHMAN, R.J.; KHABAZA, T.; KLOESGEN, W.; PIATETSKY-SHAPIRO, G. & SIMOUDIS, E. (1996). Mining business databases. *Communications of the ACM*, 39(11): 42-48.

CARD, S.K.; MACKINLAY, J.D. & SHNEIDERMAN, B. (1999). *Readings in Visualisation*. San Francisco, CA: Morgan Kaufmann.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C. & WIRTH, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide,* The CRISP-DM consortium. Hyperlink [www.crisp-dm.org]. Retrieved 21 June 2003.

CHOPOORIAN, J.A.; WITHERELL, R.; KHALIL, O.E.M. & AHMED, M. (2001). Mind your business by mining your data. *SAM Advanced Management Journal,* Spring: 45-51.

CRICINFO, (2003). *Ball-By-Ball Scorecards*. Hyperlink [http://www-usa.cricket.org/link_ to_ database/ ARCHIVE/ 2002-03/]. Retrieved 27 August 2003.

DELMATER, R. & HANCOCK, M. (2001). *Data Mining Explained.* Boston, PA: Digital Press.

DUCKWORTH, F. & LEWIS, T. (2002). Review of the Application of the Duckworth/Lewis Method of Target Resetting in One-Day Cricket, Abstracts from the 6th Australian Conference on Mathematics and Computers in Sport, 1-3 July 2002, Bond University, Queensland, Australia. *Sports Engineering*, 5: 239-244.

DUNN, J.G.H. & SYROTUIK, D.G. (2003). An investigation of multidimensional worry dispositions in a high contact sport. *Psychology of Sport and Exercise*, 4: 265-282.

FAYYAD, U. (1998). Diving into databases. *Database Programming and Design*, 11(3): 24-31.

FAYYAD, U.; PIATETSKY-SHAPIRO, G. & SMYTH, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11): 27-34.

FIELD, M.; COLLINS, M.W.; LOVELL, M.R. & MAROON, J. (2003). Does age play a role in recovery from sports-related concussion? A comparison of high school and collegiate athletes**.** *The Journal of Pediatrics*, 142(5): 546-553.

HIRJI, K.K. (2001). Exploring data mining implementation. *Communications of the ACM*, 44(7): 87-93.

JACKSON, J. (2002). Data mining: A conceptual overview. *Communications of the Association for Information Systems*, 8: 267-296.

KLEIN, H.K. & MEYERS, M.D. (1999). A set of principles for conducting and evaluating interpretive field studies in information systems, *MIS Quarterly*, 23(1): 67-94.

NEWELL, K. (2002). High Tech U. *Coach and Athletic Director*, September: 38-44.

PEACOCK, P.R. (1998a). Data mining in marketing: Part 1. *Marketing Management*, Winter: 9-18.

PEACOCK, P.R. (1998b). Data mining in marketing: Part 2. *Marketing Management*, Spring: 14-25.

RAHNAMA, N.; LEES, A. & REILLY, T. (2002). A novel computerised notation and analysis system for assessment of injury and injury risk in football, *Physical Therapy in Sport*, 3(4): 183-190.

SCHOEMAN, J.H.; GROUND, M. & MATTHEE, M.C. (2003). Getting a clearer picture: A business application of data mining. *Working paper.* Pretoria: Department of Informatics, University of Pretoria.

SIMOUDIS, E. (1996). Reality check for data mining. *IEEE Expert*, October: 26-33.

ZHOU, W.; DAO, S. & JAY KUO, C.C. (2002). On-line knowledge- and rule-based video classification system for video indexing and dissemination. *Information Systems*, 27(8): 559-586.

Dr. Machdel C. Matthee: Department of Informatics, University of Pretoria, Pretoria 0002, Republic of South Africa. Tel.: +27 (0)12 420 3798 (w), +27 (0)82 371 5086 (s), Fax.: +27 (0)12 362 5287, Email: machdel.matthee@up.ac.za

(Subject editor: Prof. J.Z.B. Bloom)

NOTES