Journal of Computer Science 4 (10): 826-833, 2008
ISSN 1549-3636

Performance Analysis of Data Mining Tools Cumulating with a Proposed Data Mining Middleware

Lai Ee Hen and Sai Peck Lee
Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

Abstract: Data mining has becoming increasingly popular in helping to reveal important knowledge from the organization's databases and has led to the emergence of a variety of data mining tools to help in decision making. Present study described a test bed to investigate five major data mining tools, namely IBM intelligent miner, SPSS Clementine, SAS enterprise miner, oracle data miner and Microsoft business intelligence development studio. Present studies focus on the performance of these tools. Results provide a review of these tools and propose a data mining middleware adopting the strengths of the tools.

Key words: knowledge discovery, performance metrics, test bed

INTRODUCTION

In today's information age, in order to stay competitive in the market, there is a need for a powerful analytic solution to help in the extraction of useful information from the large amount of data collected and stored in an organization's databases or repositories. This has led to the emergence of Knowledge Discovery in Databases (KDD) which is responsible to transform low-level data into high-level knowledge for decision support. According to Fayyad *et al.*[12] "Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"[13]. Knowledge discovery process consists of the list of iterative sequence steps of processes and data mining is one of the KDD processes. Data mining is the application of algorithms for extracting patterns from data[13]. These extracted patterns will provide useful knowledge to decision makers. As such, there has been an increasing demand for data mining tools to help organizations uncover knowledge that can guide in decision making. Each tool offers a wide range of functionalities to users. The functionalities provided are important in a data mining tool, however, the performance of the tool is also a necessary feature. Therefore, in this research, we have selected five widely used data mining tools to study their performance.

MATERIALS AND METHODS

Motivation: Successful implementation of the data mining effort requires a careful assessment of the various tools and algorithms available. Frequent pattern mining forms a core component in mining associations, correlations, sequential patterns, partial periodicity, and so forth, which are of great potential value in applications. Many methods have been proposed and developed for efficient frequent pattern mining in various kinds of databases, including transaction databases, time-series databases, and so forth. As we are flooded with the wide range of features, we should not omit data mining tools' performance. In today's information age, companies are no longer dealing with megabytes of database but instead gigabytes and terabytes. Based on the polls from KDNuggets, in the year 2006, the median value for the largest database size used to perform data mining is between 1.1 and 10 Gigabytes, and 12% report mining terabyte size of databases[27]. Therefore, the scalability of a data mining tool should be leveraged to enable massive data sets to be analyzed. As data sets increase in size, data mining tools become less and less efficient. This is an important feature for a data mining tool which motivates us to study the performance of the data mining tools.

Analysis of data mining tools: In present study, we identified five major data mining tools which we perceived as market leaders. We study the performance

Corresponding Author: Lai Ee Hen, Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

criteria of each tool based on six attributes which we have introduced. Each of the attributes is associated with metrics that we perceived as important measurement to the attribute. A known limitation to our study is that we only study a subset of the data mining tools that are currently available. However, they are either market leaders or major data mining tools that are deemed adequate to represent the current state of technologies in data mining.

Selected data mining tools: There are numerous data mining tools available in the market. Based on the META Group's META spectrum report for data mining, a few data mining tools have been cited as leaders in the data mining market, namely SPSS Clementine, Oracle Data Mining, and SAS Enterprise Miner 0. Therefore we limit our scope of study to these few data mining market leaders. In addition, we have also chosen IBM Intelligent Miner and Microsoft Business Intelligence Development Studio that we considered as the major data mining tools in our analysis. The selected data mining tools are shown in Table 1.

Data mining performance measures: We identified several recommended metrics that we perceived as significant to determine the performance attributes of data mining tools such as IBM Intelligent Miner, SPSS Clementine, SAS Institute Enterprise Miner, Oracle Data Miner, and Microsoft Business Intelligence Development Studio. These metrics are measures on the application tier and thus also referred to as application tier metrics. Table 2 shows the attributes measured and the associated metrics as included in[31]. Specific thresholds are used as a baseline to measure the performance attributes of IBM Intelligent Miner, SPSS Clementine, SAS Institute Enterprise Miner, Oracle Data Miner, and Microsoft Business Intelligence Development Studio. We identified several metrics from Table 2 to measure the performance of these tools. Table 3 shows these metrics together with the description of the metrics. These metrics were chosen due to our focus on increasing performance of data mining by maximizing memory usage and reducing I/O.

Analysis strategy: Test bed: To facilitate an analytical comparison of different frequent mining methods based on the listed performance metrics in Table 3, an open test bed has been constructed to study the performance of the various data mining tools based on the adventure works database[25] with different types of algorithms that the data mining tools supported as depicted in Fig. 1.

The test bed consists of a synthetic data generator. It is used to generate large sets of synthetic data in various kinds of distributions. These data will act as the data source for the data mining tools to generate mined results.

The data will be mined by various data mining methods ranging from different types of algorithms that are supported by IBM Intelligent Miner[22], SPSS Clementine[26], SAS Institute Enterprise Miner[24], Oracle Data Miner[23] and Microsoft Business Intelligence Development Studio[25]. We will test the performance of the tools based on a few data mining algorithms which are Classification algorithms,

Table 1: Selected data mining tools

No	Product Name	Vendor
1	IBM Intelligent Miner	IBM corporation
2	SPSS Clementine	SPSS Inc
3	SAS Institute, Enterprise Miner	SAS institute inc
4	Oracle, Data Miner	Oracle corporation
5	Microsoft, business intelligence development studio	Microsoft corporation

Table 2: Attributes and metrics extracted from[31]

Attribute	Metric
To measure memory shortages	Memory\Available Bytes
	Process(All_processes)\Working Set
	Memory\Pages/sec
	Memory\Cache Bytes
To measure excess paging with a disk bottleneck	Memory\Page Reads/sec
	Physical Disk\Avg. Disk Bytes/Read
	Logical Disk\% Free Space
	Physical Disk\% Disk Time
	Logical Disk\% Disk Time
	Physical Disk\Disk Reads/sec
	Physical Disk\Disk Writes/sec
To measure paging file fragmentation	Physical Disk\Split IO/sec
	Physical Disk\% Disk Read Time
	Physical Disk\Current Disk Queue
Length	
	Process(All processes)\Handle Count
	Paging File\% Usage
	Processor\ % Processor Time
	Processor\ Interrupts/ sec
	Server\ Bytes Total/ sec
	Server\ Pool Paged Peak
To measure memory leaks	Memory\Available Bytes
	Memory\Committed Bytes
	Memory\Pool Nonpaged Bytes
	Memory\Pool Nonpaged Allocs
	Process(process_name)\Private Bytes
	Process(process_name)\Working Set
	Process(process_name)\Page Faults/Sec
	Process(process_name)\Page File Bytes
	Process(process_name)\Handle Count
To measure cache manager efficiency	Cache\Copy Read Hits %
	Cache\Copy Reads/sec
	Cache\Data Map Hits %
	Cache\Data Maps/sec
	Cache\MDL Read Hits %
	Cache\MDL Reads/sec
	Cache\Pin Read Hits %
	Cache\Pin Reads/sec
	Memory\Pages Input/sec

Table 3: Types, metrics and descriptions

Type	Subtype	Metric	Measured As	Description
Memory	-	Available bytes (MAB)	Megabytes	Available Bytes (measured in bytes) is the amount of free physical memory available for processes to allocate
Memory	-	Pages/sec (MP)	Percentage	Pages/sec is the number or pages read from or written to disk in seconds to resolve unreferenced pages in memory
Paging File	-	% Usage (PFU)	Percentage	The measure of usage of page file.
Disk	Logical disk	% Free space (LDFS)	Percentage	% Free Space is the percentage of total usable space on the selected logical disk drive that was free
Disk	Physical disk Logical disk	% Disk time (PDDT)% Disk Time (LDDT)	Percentage	% Disk Time is the elapsed time used by the disk drive during the service of read or write request

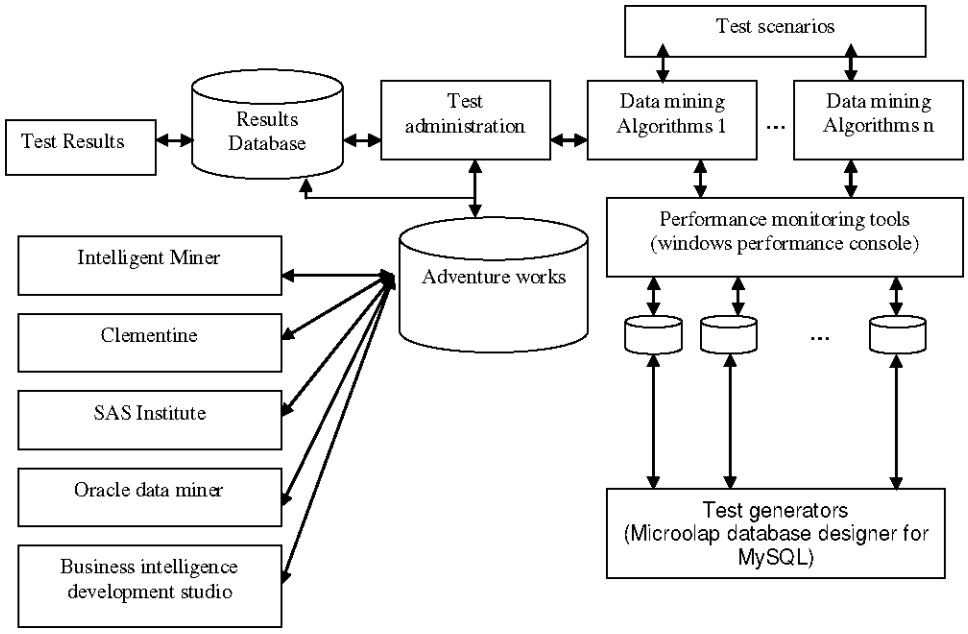

Fig. 1: Architecture of the open test bed

Regression algorithms, Segmentation algorithms, Association algorithms and Sequence analysis algorithms.

Configuration and selection of data mining algorithms will be done during Test Administration. Data mining process is done by the selected data mining algorithms and the result will be either stored in the Results Database or in file based on the nature of the data mining tools.

Adventure Works Cycles, a fictitious company on which the Adventure Works database is based[25], is a large multinational manufacturing company. The company manufactures and sells metal and composite bicycles to North America, European and Asian commercial markets. 5 schemas and 14 tables are used in the Adventure Works database. These schemas and tables provided are used thoroughly to test the different types of algorithms supported by IBM Intelligent Miner, SPSS Clementine, SAS Institute Enterprise Miner, Oracle Data Miner, and Microsoft Business Intelligence Development Studio.

Throughout the process of the analysis in the test bed, we used Windows Performance Monitor to monitor the performance of each data mining tool based on the performance measures defined in Table 3.

Analysis strategy:

Test environment: Each tool may perform differently in behaviour. In order to achieve standardization during testing we have proposed a fix test environment shown in Table 4 and 5.

Table 4: Server 1 Specification

Processor	Pentium IV 3.0-GHz processor
Operating system	Windows 2003 server enterprise Edition Service Pack 1
File format	Disk partitions are formatted with the NTFS file system
Memory	512 MB of RAM
Available hard disk space	60 GB of available hard-disk space
Optical Drive	52X CD-ROM drive
Other devices	Microsoft windows-compatible network interface card
Other software requirements	Sun java runtime environment
Microsoft Internet Explorer 6.0 SP1	version 1.4

Table 5: Server 2 specification

Processor	Pentium IV 2.8-GHz processor
Operating system	Windows 2003 Server Enterprise Edition Service Pack 1
File format	Disk partitions are formatted with the NTFS file system
Memory	512 MB of RAM
Available hard disk space	40 GB of available hard-disk space
Optical Drive	52X CD-ROM drive
Other devices	Microsoft Windows-compatible network interface card
Other software requirements	Sun Java Runtime Environment
Microsoft Internet Explorer 6.0 SP1	version 1.4

RESULTS AND DISCUSSION

Table 6, Fig. 2-6 show the test results of each data mining tool mentioned above. We measured the performance of five algorithms, namely classification algorithms, regression algorithms, segmentation algorithms, association algorithms, and sequential analysis algorithms based on a subset of the metrics in Table 3. Throughout testing, we encountered several exceptions which cause data mining computations to terminate unexpectedly. Such exceptions are denoted by an "E" notation on the table.

Test results comparison: Memory access times are measured in Nanoseconds. Disk access times are measured in Milliseconds. The difference factor is a million which means that disk access times are about a million times slower than memory access. As such, we believe that data mining should be performed in-memory. Based on our study, tools like Microsoft Business Intelligence Development Studio and IBM Intelligent Miner consume a large amount of both memory usage and disk usage on the application tier. These tools are distributing part of the data mining load to the application tier. Such a strategy delegates computing cycles from the backend systems right to the application systems. However, this strategy potentially might lead to problems such as disk issue and memory issue. With reference to Table 6 and the given charts, there might be possibly of slight variations on the results collected on the metrics on different algorithms such as sequence analysis algorithms, association algorithms and segmentation algorithms. A possible hypothesis is that some of the tools might produce better performance on different algorithms and as such the tools are able to compute better performance attributes. For example, Microsoft Business Intelligence Studio (90% of physical disk\ disk time and 80% logical disk\disk time) consumes a large amount of disk usage on classification algorithms as compared to tools like SPSS Clementine (30% of physical disk\disk time and 30% logical disk\disk time) but only consumes a small amount of disk usage on segmentation algorithms (10% of physical disk\disk time and 10% logical disk\disk time) as compared to SPSS Clementine (40% of physical disk\disk time and 50% logical disk\disk time). Such a result might be likely caused by the implementation strategy on the algorithms by each vendor. For example, vendors like SPSS might be more efficient in implementing classifications as compared to Microsoft. On the other hand, Microsoft might be more efficient in implementing segmentation algorithms as compared to IBM. In short, the implementation strategy of algorithms indirectly affects the performance of data mining tools. A recommended solution would be to allocate, if possible, a major percentage of data mining computations at the memory level to minimize disk activity and maximize memory activity.

Our study also reveals that data mining at the memory level will lead to better performance. For example, IBM Intelligent Miner consumes only 15% of physical disk\disk time and 100mb of memory\available bytes. This explains that there is a trade-off between memory and disk. If we spend more time at the memory level, then we should spend less time on disk activity (also referred to as Disk I/O). Disk I/O is often a major bottleneck to data mining performance.

We discovered that inefficient memory management might lead to potential memory issues. For example, if a memory access results in a hard page fault, the access will be a million times slower as compared to disk access. Excessive hard page faults can result in a system that thrashes (i.e., almost no useful work gets

Table 6: Data Mining Tools Performance Metrics

Data mining tools	Memory\available bytes (MAB)	Memory\pages/ sec (%) (MP)	Paging file\% Usage (%) (PFU)	Physical disk\ Free Space (%) (PDFS)	Logical disk \Free Space (%) (LDFS)	Physical Disk\ Disk Time (%) (PDDT)	Logical Disk\ Disk Time (%) (LDDT)
Classification algorithms							
IBM	100	20	50	15	15	15	15
SPSS	110	20	30	15	15	30	30
SAS	180	80	80	1	1	1	1
Oracle	250	90	90	0	0	0	0
Microsoft	100	20	50	15	15	90	80
Regression algorithms							
IBM	220	80	70	10	10	10	10
SPSS	E	E	E	E	E	E	E
SAS	140	80	60	1	1	1	1
Oracle	180	80	80	0	0	0	0
Microsoft	120	70	50	15	15	20	20
Segmentation algorithms							
IBM	90	40	30	10	10	30	30
SPSS	70	40	40	6	6	40	50
SAS	100	50	60	1	1	1	1
Oracle	120	60	60	0	0	0	0
Microsoft	90	50	30	4	4	10	10
Association algorithms							
IBM	210	70	50	20	20	20	20
SPSS	240	90	60	40	40	70	60
SAS	190	50	40	1	1	1	1
Oracle	E	E	E	E	E	E	E
Microsoft	250	70	50	10	10	20	20
Sequence analysis algorithms							
IBM	100	30	30	10	10	30	30
SPSS	180	50	80	20	20	40	40
SAS	170	40	40	1	1	1	1
Oracle	190	60	40	0	0	0	0
Microsoft	190	60	40	10	10	40	40

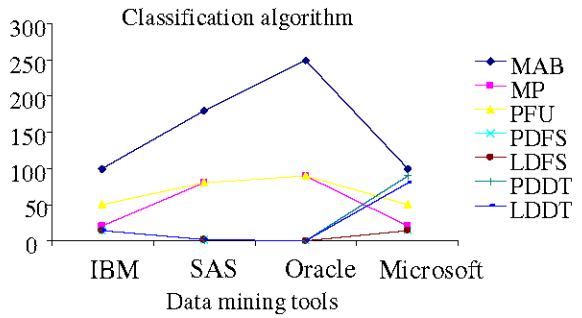

Fig. 2: Performance metrics of data mining tools based on classification algorithms

done). As such, each data mining product, both server-based and client-based, should be carefully designed to minimize page faults. Metrics such as Memory\Pages/sec is the metrics that we used to monitor the behaviour of disk during paging (i.e., whether the disk is busy with other work or with handling page faults).

Proposed architecture of server-based data mining middleware: Based on the test results, the strategies used by data mining tools both show their strength and weaknesses. The proposed data mining middleware will adopt the strengths of dominant data mining tools, coupled with its added features.

With reference to the test results in Table 6, Fig. 2-6, we noticed that there is a correlation between memory and disk. If the memory activity is high, then the disk activity will be low and vice versa. Such a relationship explains that computations performed at the memory level are faster than computations performed at the disk level. This leads to the proposed data mining middleware, namely Java-Based Data Mining Middleware (also referred to as JDMM). JDMM will use memory extensively on data mining computations as we believe that the memory is more

efficient on handling computations as compared to disk which requires input/output. The objective is to minimize disk usage and to maximize memory usage.

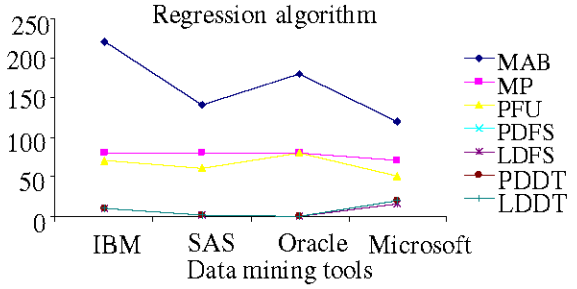

Fig. 3: Performance metrics of data mining tools based on regression algorithms

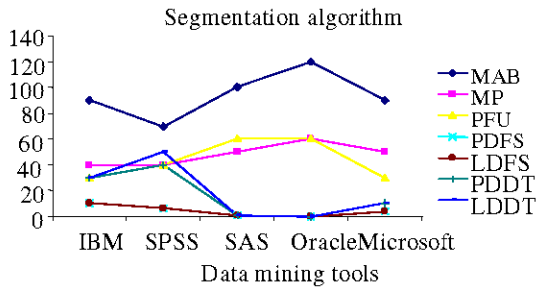

Fig. 4: Performance metrics of data mining tools based on segmentation algorithms

With 64-computing, we believe that memory limitation is no longer a bottleneck in data mining. In addressing efficiency of data mining algorithms,

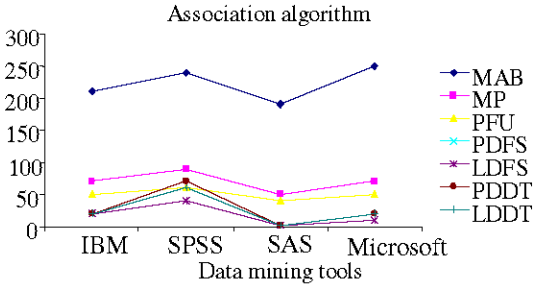

Fig. 5: Performance metrics of data mining tools based on association algorithms

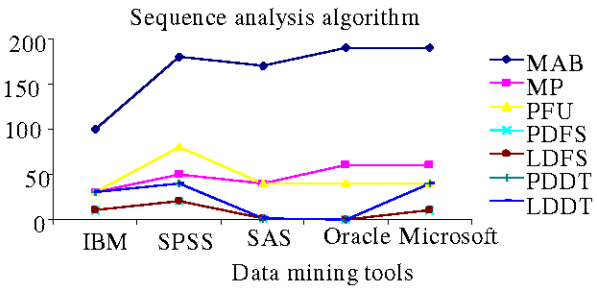

Fig. 6: Performance metrics of data mining tools based on sequence analysis algorithms

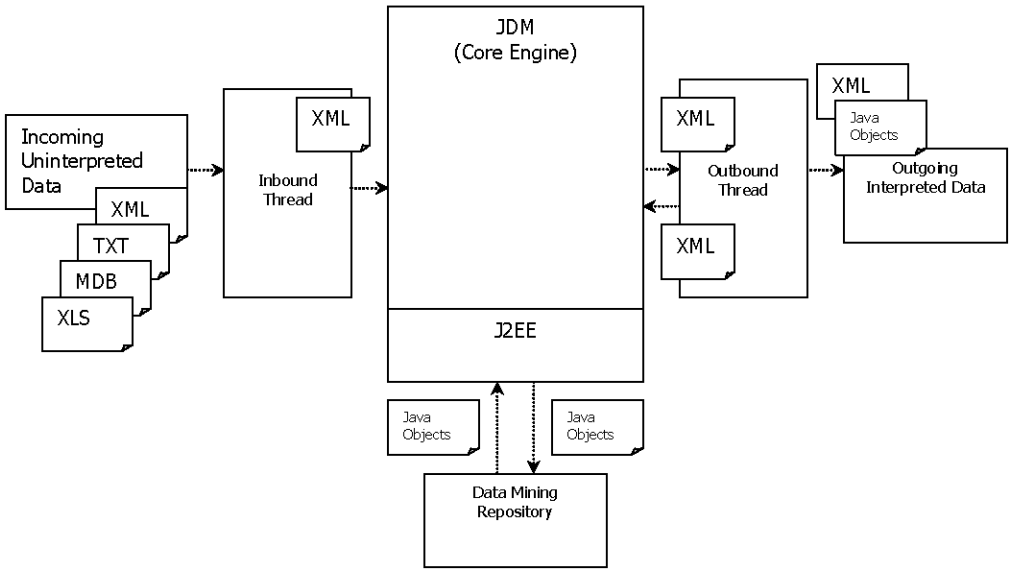

Fig. 7: JDMM High level Architecture

JDMM proposes a plugin-based data mining algorithms. As such, the number of algorithms supported by JDMM is scalable. If a specific algorithm is identified to be the potential bottleneck to JDMM, we can easily remove the plugin of this algorithm. Alternative, JDMM can also tackle such inefficiency through its data mining repository which is memory-based. We ultimately transfer the computations of these inefficient algorithms to the memory repository and let the repository tackle the issue of performance. In addition, the proposed JDMM will be able to handle multiple data sources.

JDMM is a server centric middleware. It is platform-data source and data mining technique-independent which are accessible from front-, back- and web-office environments. The high-level architecture of the JDMM is depicted in Fig. 7. JDMM will allow users to mine data from multiple data sources ranging from relational databases, object-oriented databases, flat files and so forth. This is possible through the in bound thread where different adapters are allowed to plugin into the JDMM. These adapters allow users to connect to different data sources. The Java-based Data Miner (JDM) will be the core engine to perform data mining operations on the incoming uninterpreted data from the data sources. The Data Mining Repository is used to persist mined objects. The sole objective of the memory repository is to reduce any I/O during the process of mining the data sources. Majority of the data mining processes within JDMM are performed using the cached data from the memory repository. Lastly the results are delivered to users through the Outbound thread where the mined data are formats into either PDF file, XLS file, XML file, text file, html/htm file or any proprietary formats that are incorporated into the JDMM.

CONCLUSION

We believe, in the near future, most data mining products will effectively utilize the memory extensively during data mining. With 64-bit computing, we believe that memory limitation is no longer an issue. Alternatively, the continual reduction in the cost of memory will benefit 64-bit computing. 64-bit computing has gradually shifted down to the personal computer desktop which means, in the near future, data mining products no longer need a powerful dedicated server to compute small to medium size data mining computations. Apart from 64-computing, future research on data mining products will need to be more comprehensive covering attributes such as maintainability, adaptability, scalability, reliability and portability.

ACKNOWLEDGEMENT

This study was undertaken as part of the dissertation project undertaken by Lai Ee Hen for her Master of Software Engineering programme at University of Malaya, Malaysia.

REFERENCES

1. Zanasi, N.F.F. Ebecken and C.A. Brebbia, 2004. Data Mining V: Data Mining, Text Mining and Their Business Applications. 5th International Conference on Data Mining, Wit Press.
2. Michael Goebel and Le Gruenwald, 1999. A Survey of Data Mining and Knowledge Discovery Software Tools. SIGKDD Explorat., 1: 20-33. http://doi.acm.org/10.1145/846170.846172
3. John, F. Elder IV and Dean W. Abbott Elder, 1998. A Comparison of Leading Data Mining Tools, 4th International Conference on Knowledge Discovery and Data Mining, pp.1-31 http://www.abbottanalytics.com/assets/pdf/Abbott-Analytics-Comparison-High-End-DM-Tools-No-Pics-1998.pdf
4. Charles Berger, 2001. Oracle9i data mining: An oracle white paper. Oracle Corporat., 1-17
5. David J. Hand, 1999. Statistics and Data Mining: Intersecting Disciplines, SIGKDD Explorations, 1: 16-19. http://doi.acm.org/10.1145/846170.846171
6. Dr. Diego Kuonen, 2004. A Statistical Perspective of Data Mining, Statoo Consulting.
7. Jeffrey W. Seifert, 2004. Data Mining: An Overview, Congressional Research Service. The Library of Congress, pp: 1-19
8. Charlie Berger, 2004. Oracle Data Mining: Know More, Do More, Spend Less, Oracle Corporation. pp: 3-5
9. Grossman, R., S. Kasif, R. Moore, D. Rocke and J. Ullman, 1999. Data mining research: opportunities and challenges. A Report of three NSF Workshops on Mining Large, Massive and Distributed Data, Jan. 1999, pp: 1-11. http://www.rgrossman.com/dl/misc-001.pdf
10. Pavel Berkhin, 2002. Survey of clustering data mining techniques. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.3739
11. Dean W. Abbott, I. Philip Matkovsky and John F. Elder IV, 1998. An evaluation of high-end data mining tools for fraud detection. Proceeding of the IEEE International Conference on Systems, Man and Cybernetics. Oct. 11-14 IEEE Explore, San Diego, CA, USA., pp: 2839-2841. DOI: 10.1109/ICSMC.1998.725092

12. Michel A. King, John F. Elder IV, Brian Gomolka, Eric Schmidt, Marguerite Summers and Kevin Toop, 1998. Evaluation of Fourteen Desktop Data Mining Tools, pp: 1-6. http://www.dataminginglab.com/Portals/0/tool%20eval%20articles/smc98_king_elder.pdf
13. Fayyad, U., G. Piatetsky-Shapiro and P. Smyth, 1996. Advances in Knowledge Discovery and Data Mining, MIT Press, ISBN-10: 0262560976 pp: 560.
14. Peter M. Chen and David A. Patterson, 1993. Storage performance-metrics and benchmarks. Proceeding of the IEEE., 81: 1-33. DOI: 10.1109/5.236192.
15. Ian Foster, Carl Kesselman, and Steven Tuecke, 2001. The Anatomy of the Grid: Enabling Scalable Virtual Organizations, int. J. High Perform. Comput. Appli., 15: 200-222. DOI: 10.1177/109434200101500302
16. Floyd Marinescu, 2005. Announcing terracotta-Java clustering and caching without API's. http://www.theserverside.com/news/thread.tss?thread_id=34294
17. Zanasi, A., N.F.F. Ebecken and C.A. Brebbia, 2004. Data mining V: Text mining and their business applications. Proceeding of the 5th International Conference on Data Mining, (ICDM'04), WIT Press, pp: 5-7
18. Kirk Pepperdine, 2006. Clustering POJOs in the Java Runtime Environment, Terracotta White Paper. Terracotta, Inc., pp. 1-7
19. TPC-W Terracotta-Sessions v/s Popular Commercial Application Server Test Results, 2006. Terracotta Inc.
20. Terracotta Inc., 2006. Apache tomcat and terracotta session clustering comparison and real-world benchmarks. http://www.terracotta.org/confluence/download/attachments/1509755/TerracottaSessions_Tomcat_Comparison.pdf
21. Predictive Model Markup Language (PMML), 2005. Technology reports. http://xml.coverpages.org/pmml.html
22. DB2 Intelligent Miner Library, 2002. Using the Intelligent Miner for Data, IBM, Version8 Relase 1. http://www-306.ibm.com/software/data/iminer/library-v81.html
23. Oracle Corporation, 2006. Oracle Data Miner, http://www.oracle.com/technology/products/bi/odm/odminer.html
24. SAS Enterprise Miner Documentation, What's New in SAS Enterprise Miner 5.1, SAS Institute Inc, http://support.sas.com/documentation/onlinedoc/miner/
25. Microsoft Corporation, 2005. SQL server 2005 documentation and samples. http://blogs.x2line.com/al/archive/2004/07/23/490.aspx
26. SPSS Inc., 2005. Maximize Your Returns with Data Mining and Predictive Analysis, Clementine, http://www.spss.com/clementine/index.htm,
27. KDnuggets, 2005. Largest Database Data-Mined, http://www.kdnuggets.com/polls/2006/largest_database_mined.htm
28. KDnuggets, 2006. Data mining/analytic tools you used in 2006. http://www.kdnuggets.com/polls/2006/data_mining_analytic_tools.htm
29. Microsoft Corporation, 2005. Windows XP Professional Documentation.
30. Oracle Corporation, 2004. Data Mining Tools, METAspectrum Evaluation. http://www.oracle.com/technology/products/bi/odm/pdf/odm_metaspectrum_1004.pdf
31. Microsoft Corporation, 2003. Windows server 2003 documentation. http://www.microsoft.com/windowsserver2003/proddoc/default.mspx