RNPS No. 2143
ISSN 2072-6260
Versión Digital

SERIE GRIS

REPORTE TÉCNICO
Minería de Datos

Estado actual de la aplicación de la Computación Paralela a la Minería de Datos.

Dr. José Hernández Palancar

RT_002 **Noviembre 2004**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
Ciudad de La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

RNPS No. 2143
ISSN 2072-6260

REPORTE TÉCNICO
Minería de Datos

SERIE GRIS

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

Estado actual de la aplicación de la Computación Paralela a la Minería de Datos

Dr. José Hernández Palancar

RT_002 **Noviembre 2004**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
Ciudad de La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

Estado actual de la aplicación de la computación paralela a la Minería de Datos

Dr. José Hernández Palancar

Centro de Aplicaciones de Tecnología de Avanzada, 7a #21812 e/ 218 y 222, Siboney, Playa, Habana, Cuba
jpalancar@cenatav.co.cu

RT-MD 002 CENATAV
Noviembre de 2004

Resumen: Como consecuencia de los continuos y acelerados saltos tecnológicos que se vienen produciendo en la Informática, en lo referente al aumento de la capacidad de procesamiento y de almacenamiento, así como en la velocidad de transmisión de datos; se viene desarrollando también un vertiginoso crecimiento de los volúmenes de información científica y comercial en diferentes esferas de la vida humana, que el hombre necesita evaluar con una velocidad de procesamiento directamente proporcional a la importancia de los resultados que se esperan. Aunque ha sido corroborada la utilidad práctica de la aplicación de la Minería de Datos en el procesamiento de gran cantidad de información, se impone cada día más la combinación de dicha disciplina con técnicas y herramientas de procesamiento en paralelo, que permitan enfrentar los grandes volúmenes de datos ante los que los algoritmos actuales de Minería de Datos no funcionan adecuadamente.

Exponer una visión general del estado actual de la aplicación del procesamiento en paralelo a una temática tan importante como la Minería de Datos, es el objetivo esencial de este artículo y constituye un aspecto imprescindible para poder darle solución a diferentes problemáticas asociadas al procesamiento de grandes volúmenes de datos y/o a la implementación de algoritmos de elevada complejidad computacional, que se presentan en esta esfera de investigación científca.

Palabras clave: Minería de datos, algoritmos paralelos, procesamiento en paralelo, reglas de asociación, clasificación, agrupamiento.

Abstract: As a result of the continued and accelerated technological leaps that have been occurring in computer science with regard to increasing the processing power and storage as well as the speed of data transmission; it has been also carried out a dramatic growth of information volumes in different spheres of human life that man needs to assess with a processing speed directly proportional to the importance of the expected outcomes. Although it has been corroborated the practical usefulness of data mining when processing large information amounts, it is not sufficient. Every day, it is imperative to combine this discipline with techniques and tools of parallel processing in order to cope with large data volumes, for which the current data mining algorithms do not work properly.

The main purpose of this paper is to bring an overview of the state of the art of parallel algorithms for data mining as an essential aspect to solve different problems associated with the processing of large data volumes and/ or the implementation of highly-complex computational algorithms presented in this area of scientific research.

Keywords: Data mining, parallel algorithm, parallel processing, association rules, classification, clustering

TABLA DE CONTENIDO

Introducción

La Minería de Datos y el Descubrimiento de Conocimiento abarca un campo interdisciplinario que unifica conceptos de la estadística, del aprendizaje automático, bases de datos y de la computación paralela y distribuida. Esta fusión se debe fundamentalmente al fenomenal crecimiento de los datos en todas las esferas de la actividad humana y a la necesidad económica y científica de extraer información valiosa de los datos recolectados [7], [8], [9], es por ello, que si en un principio el reto principal de la Minería de Datos era la extracción de conocimientos en bases de datos, hoy no es tan sólo eso, sino en general sobre volúmenes de datos muy grandes [6].

Una de las cuestiones más importante de la Minería de Datos y del Descubrimiento de Conocimiento es lograr que los algoritmos, las aplicaciones y los sistemas, sean escalables ante grandes volúmenes de datos, es decir trabajen con la misma eficiencia ante el aumento del tamaño y la dimensionalidad de los datos [1], pero como bien se plantea en [6], alcanzar este resultado constituye un reto. Para lograr que un algoritmo de minería de datos sea escalable ante grandes volúmenes de datos, muchas aplicaciones y sistemas utilizan como tecnología, lo que se conoce por el término de computación de alto rendimiento *(high performance computing)* o computación paralela [2].

Con el abaratamiento relativo de la computación paralela[4], [5] a partir de la introducción de los cluster de *workstations* y otras tecnologías relacionadas, se hace cada vez más frecuente y necesario, contar con una infraestructura de hardware paralela, para la experimentación y aplicación de la computación paralela a la Minería de Datos, de otra manera no sería posible desarrollar nuevos algoritmos, o determinar si muchos de los algoritmos de Minería de Datos frecuentemente utilizados, son escalables ante grandes volúmenes de datos.

En sentido general puede plantearse, que los principales retos[1] para la aplicación del paralelismo a la Minería de Datos son: desarrollar versiones escalables de los algoritmos de Minería de Datos comúnmente utilizados, desarrollar nuevos algoritmos paralelos para minar conjuntos de datos muy grandes y desarrollar software o bibliotecas paralelas orientadas a la Minería de Datos, que faciliten la implementación de los algoritmos y con ello la disminución de los plazos para el desarrollo de las aplicaciones.

En este artículo, pretendemos brindar una visión muy general sobre el estado actual de la aplicación del procesamiento en paralelo a la Minería de Datos, como premisa indispensable para poder abordar futuros proyectos de investigación relacionados con el procesamiento de grandes volúmenes de datos y/o la implementación de algoritmos de elevada complejidad computacional, que se presentan en esta disciplina.

Para el desarrollo de nuestra investigación hemos recopilado y consultado una amplia bibliografía sobre el tema, entre la que destacamos principalmente, el libro de Freitas y Lavington [2], los tutoriales de Kumar et al, así como el *proceeding* [8] del que fueron editores Zaki y Ho.

Como parte de este estudio expondremos algunos de los principales resultados y aspectos de interés relacionados con la aplicación del paralelismo a la Minería de Datos, en función de ello, hemos organizado el artículo en los siguientes tópicos: introducción; conceptos básicos sobre Minería de Datos y procesamiento en paralelo; principales resultados; personalidades más destacadas en la temática; grupos de investigación de mayor renombre; eventos científicos; principales publicaciones

que abordan el tema; tendencias en el desarrollo de algoritmos, software o herramientas; conclusiones y bibliografía.

1. Aspectos introductorios sobre Minería de Datos y Procesamiento en Paralelo

1.1 Aspectos introductorios sobre Minería de Datos

Con el crecimiento explosivo que se viene observando en la disponibilidad de diferentes tipos de datos, se viene produciendo también una oportunidad inaudita de desarrollar de forma automatizada, diferentes técnicas para analizar tales datos y extraer de ellos conocimiento útil. La Minería de Datos es un paso importante en el proceso de descubrir este conocimiento ya que para ello utiliza métodos que descubren patrones, modelos o características interesantes, no-triviales y útiles que se encuentran escondidos en los datos [11].

El descubrimiento de conocimientos en bases de datos (*knowledge discovery in databases - KDD*) como se le denominó inicialmente considerando que el objeto sobre el que se actuaba eran bases de datos, ha pasado a ser el descubrimiento de conocimiento (*knowledge discovery -KD*) a secas, como lo indican ya algunos autores [1,6], dado que en la actualidad los datos pueden ser estructurados o no, de hecho han aparecido otras disciplinas que se desprenden de la Minería de Datos, como lo son la Minería de Textos (*Text Mining*) y la Minería sobre el WEB (*WEB Mining*).

La empresa de descubrir conocimiento [8,9], consiste en la sucesión reiterativa de los pasos siguientes: selección de los datos, limpieza de los datos (eliminación de datos ruidosos, casos extremos o outliers, valores perdidos), transformación de los datos (selección de rasgos, reducción de las dimensiones), minería de datos (selección del algoritmo en correspondencia con el problema a resolver: asociación, clasificación, agrupamiento, etc), validación de los resultados (comprobación de los resultados con datos de prueba), visualización (transformación de los modelos resultantes a un formato comprensible por el hombre) y aplicación del conocimiento descubierto.

Las mayores tareas de la Minería de Datos [10] son la predicción y la descripción. Los métodos de predicción emplean algunas variables para predecir valores posibles o desconocidos de otras variables: estos incluyen la clasificación, la regresión y la detección de desviación. Los métodos descriptivos buscan patrones interpretables por el hombre que describen a los datos, estos incluyen: agrupamientos (clustering), descubrimiento de reglas de asociación y minado de sucesiones.

1.1.1 Métodos de la Minería de Datos [2], [12]

Clasificación y regresión: es el proceso de asignar nuevos objetos a clases o categorías definidas, los atributos que caracterizan a los objetos pueden tomar valor en un pequeño conjunto de valores discretos. La regresión es conceptualmente similar a la clasificación, la principal diferencia estriba en que los atributos son continuos, es decir, pueden tomar cualquier valor real o entero en un intervalo arbitrariamente grande, también se conoce como aprendizaje supervisado.

Detección de desviaciones*:* encuentra el o los objeto(s) que difieren más de los otros, es decir, encuentra todos aquellos objetos que tienen un comportamiento inusual.
Reglas de la asociación: detecta los conjuntos de atributos que frecuentemente co-ocurren y las reglas entre ellos, es decir, el 90% de la gente que compra galletas también compra leche (el 60% de todos los compradores de la tienda de comestibles compran ambos).
Minado de sucesiones: descubre las sucesiones de eventos que ocurren juntos comúnmente, es decir, en un conjunto de sucesiones de ADN, ACGTC es seguido por GTCA después de un salto de 9, con una probabilidad del 30%.

Agrupamiento *(clustering):* es el proceso de distribuir el conjunto de datos en subconjuntos o grupos tales que los elementos de un grupo comparten sistema de características común, con una alta similaridad dentro del grupo y una baja similaridad inter-grupo. También se conoce como aprendizaje no supervisado. En el agrupamiento, a diferencia del método de clasificación, las clases no están definidas, lo que hace que este proceso sea una de las tareas más complejas de la Minería de Datos y que tenga además un elevado costo computacional.

Búsqueda de la similaridad: dado un conjunto de objetos, y un objeto “solicitado", encuentran al o a los objecto(s) que están a una distancia definida por el usuario del objeto “solicitado”, o encuentra todos los pares que se encuentran a cierta distancia unos de otros.
Existen otros métodos entre los que se encuentran:
– Redes neuronales *(Neural networks).*
– Algoritmos genéticos *(Genetic algorithms)*
– Modelos de Markov ocultos *(Hidden Markov models).*
– Series temporales *(Time series).*
– Redes Bayesianas *(Bayesian networks).*
– Soft computing: conjuntos rugosos y difusos *(rough and fuzzy sets).*
– etc.

Como expresamos anteriormente, la Minería de Datos se nutre de muchos conceptos del Reconocimiento de Patrones, la Inteligencia Artificial, la Estadística, los Sistemas de Bases de Datos y la Visualización de Datos, pero muchas de las técnicas desarrolladas en estas tradicionales disciplinas no funcionan adecuadamente debido a características únicas que presentan los datos en la actualidad como son: su enorme tamaño, elevada dimensionalidad y heterogeneidad [6].

Una definición dada por D. Talia [14] que más se vincula con el procesamiento en paralelo y que complementa a las tradicionalmente conocidas, plantea de cierta forma lo siguiente: …la Minería de Datos es el análisis automatizado de grandes volúmenes de datos, buscando las relaciones o vínculos de interés y el conocimiento que está implícito en los grandes volúmenes de datos. Las investigaciones teóricas y aplicadas en el área del paralelismo aplicado a la minería de datos tienen que ver con el estudio y definición de algoritmos paralelos, métodos y herramientas, que, combinado con el uso de computadoras paralelas de gran capacidad de cómputo, permitan extraer a partir de los datos, nuevos patrones que además de ser útiles están implícitos. Cuando las herramientas de Minería de Datos se implementan sobre computadoras paralelas de alto rendimiento se puede analizar con ellas grandes volúmenes de datos en un tiempo razonable. Mayor velocidad de procesamiento significa también que los usuarios pueden experimentar con más modelos y evaluar así datos más complejos, posibilitando con ello mejorar las predicciones.

1.2 Conceptos básicos sobre procesamiento en Paralelo

La Minería de Datos es una tarea que es a la vez dato-intensiva y cálculo-intensiva. Los algoritmos de Minería de Datos cuando trabajan sobre grandes conjuntos de datos en una computadora convencional, consumen un tiempo considerable para obtener los resultados, es por ello que una aproximación para reducir este tiempo es utilizar una muestra de los datos, pero en muchos casos esto podría arrastrarnos a modelos inexactos, y en otros no es posible hacerlo porque se requiere de todos los datos [13].

Otra alternativa mayormente aceptada es la de desarrollar algoritmos escalables para diferentes técnicas de Minería de Datos, es decir, algoritmos que trabajen con la misma eficiencia ante el aumento del tamaño y la dimensionalidad de los datos [1]. Para lograr que un algoritmo de Minería de Datos sea escalable ante grandes volúmenes de datos, muchas aplicaciones y sistemas utilizan como tecnología lo que se conoce por el término de computación de alto rendimiento *(high performance computing)* o computación paralela [2]. No obstante, a pesar del uso de la computación paralela, lograr la escalabilidad de un algoritmo constituye un reto ya que es necesario conjugar diferentes factores que intervienen tanto en el diseño del algoritmo como en su implementación, entre ellos: tipo de arquitectura paralela (hardware, soporte de interconexión), modelos para la programación paralela, sistema operativo, lenguaje de programación y bibliotecas paralelas.

1.2.1 ¿Qué se entiende por paralelismo?

Una definición de paralelismo, extraída del libro "*How to build a Beowulf*" de Thomas Sterling, plantea: “El paralelismo es la habilidad de controlar muchos hilos independientes para hacer que avancen simultáneamente en la ejecución de una tarea.”

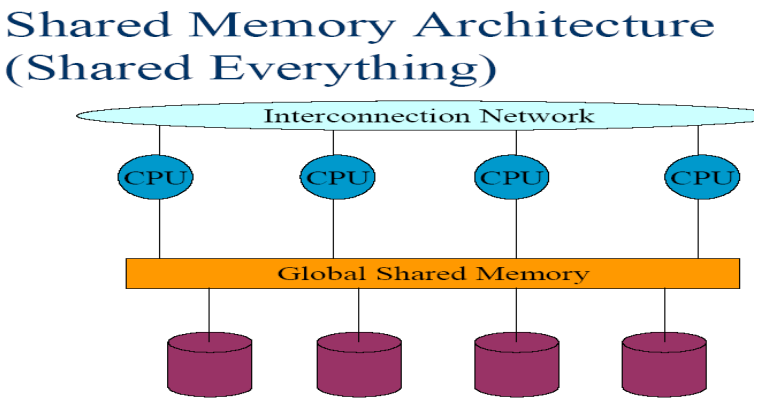
Otra definición dada por Ian Foster en su libro *Designing and Building Parallel Programs,* señala que es:

“El uso simultáneo de más de una computadora para resolver un problema.”

Ambas definiciones son válidas, porque el paralelismo puede aparecer de diferentes formas.

Las *principales arquitecturas* en las que se pueden clasificar actualmente a los sistemas paralelos son:

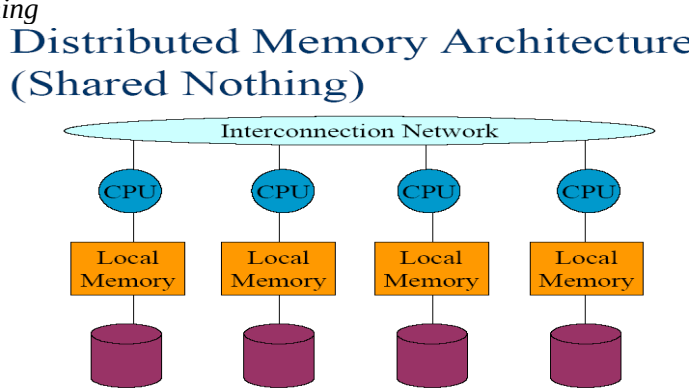
1. Arreglos de Procesadores, también conocidas como máquinas MPP (*massively parallel procesing*, mas de 1000 CPU).
2. Multiprocesadores (SMP) o máquinas de memoria compartida



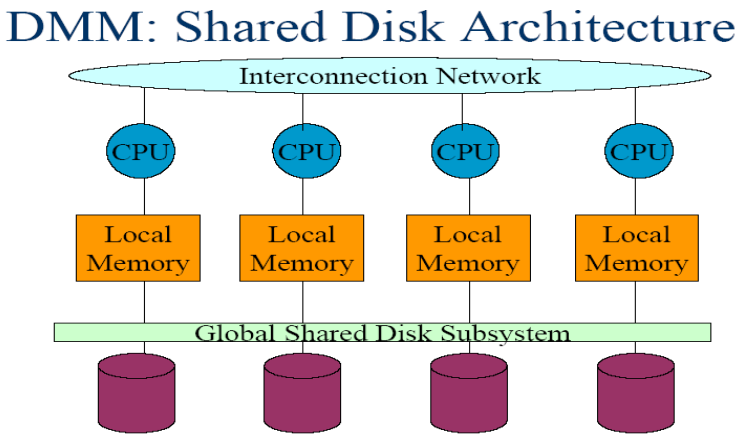
3. Multicomputadoras de memoria distribuida (cantidad de CPU menor o igual a 1000), también se les conoce como máquinas de memoria distribuida (*distributed memory machines* – DMM).

   Esta arquitectura a su vez puede ser dividida en:

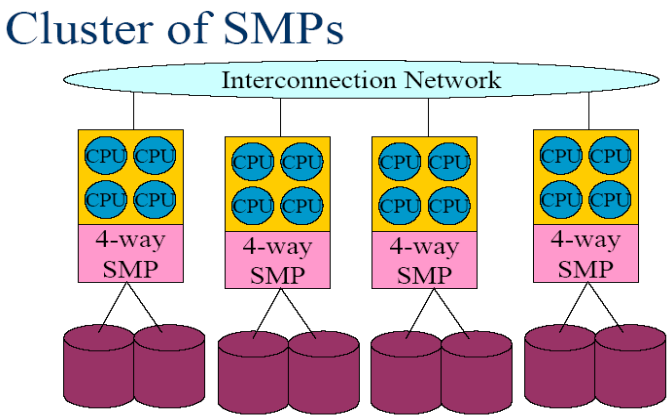
   a) Shared Nothing



   b) *Shared Disk*



4. Clusters de SMPs. A partir del año 2000 aparece el término de *Constellation*, refiriéndose a aquellos *clusters* con nodos construidos con hardware propietario y con características especiales, por ejemplo: entre 4 y 8 procesadores por nodos, tecnología de interconexión de alta velocidad, etc.



5. Máquinas de memoria compartida distribuida. Esta arquitectura es un híbrido entre la arquitectura de memoria compartida y la de memoria distribuida.

Además de las arquitecturas, otro elemento importante del paralelismo lo constituyen los *modelos para la programación paralela* que guardan estrecha relación con las arquitecturas referenciadas anteriormente, y se basan en la forma que se realiza la comunicación y el uso de la memoria. Los 3 principales modelos son:

- Hilos: arquitectura SMP, estándar Pthreads.
- Pase de mensajes: memoria distribuida, MPI-PVM.
- Memoria compartida distribuida (DSM).

Las tres formas principales de aplicar el paralelismo en los algoritmos de Minería de Datos son: ejecución de tareas en lote o paralelismo independiente [39], paralelismo de tareas o de control y paralelismo del tipo SPMD (simple program multiple data).

El **paralelismo independiente** o de ejecución de **tareas en lote** consiste en que una aplicación o varias aplicaciones corren independientemente en cada procesador y generalmente no existe la necesidad de interactuar estrechamente entre los procesadores. En este caso la aplicación normalmente realiza cálculos sobre un juego de datos y al concluir la tarea, pasa el resultado a un nodo de control (*server*), que es el responsable de recoger los resultados de todos los procesadores. La implementación de este tipo de tarea no depende de biblioteca paralela alguna, ni de un lenguaje de programación en particular, sólo requiere un sistema que sea capaz de colocar y poner a ejecutar cada tarea en cada procesador, un ejemplo de tal sistema es PBS *(Portable Batch Scheduling)*, disponible en dos versiones una *free* (*Open PBS* - www.openpbs.org/) y una comercial (PBSPro – www.pbspro.org/).

En el **paralelismo de control o de tareas**, cada procesador ejecuta diferentes operaciones sobre el conjunto de datos completo o sobre una partición del mismo, en este caso se necesita comunicación

con los otros procesadores que componen la computadora paralela, lo que conlleva una mayor actividad en la red de interconexión. El modelo de programación empleado es el de pase de mensajes, cuyos estándares más reconocidos son MPI (versiones MPI-CH y LAM-MPI) y PVM, siendo los lenguajes de programación C y Fortran los que pueden interactuar con las bibliotecas desarrolladas para dichos estándares.

En el *paralelismo del tipo SPMD,* el mismo programa se replica en todos los procesadores y se ejecuta de forma paralela sobre diferentes particiones del conjunto de datos, en algunos casos es necesaria la comunicación entre los procesadores para intercambiar resultados parciales.

Estas tres formas no son totalmente excluyentes, ellas pueden ser combinadas de forma apropiada para elevar el rendimiento y la eficiencia en la obtención de los resultados, y a partir de ello se pueden deducir diferentes estrategias para la partición de los datos:

a. particionamiento secuencial: se definen particiones separadas sin solapamiento entre ellas.

b. particionamiento con cubrimiento: una parte de los datos puede estar replicado en diferentes particiones.

c. particionamiento a solicitud basado en rango: las particiones son definidas sobre la base de una solicitud que es la que selecciona los datos en correspondencia con los valores de sus atributos.

Otros elementos a considerar en la implementación paralela de diferentes métodos de Minería de Datos son los siguientes:

- ventajas y desventajas del modelo de programación de memoria compartida contra el de memoria distribuida
- la topología de la red de interconexión de los procesadores,
- evaluar las estrategias óptimas de la comunicación,
- el balanceo de la carga de los algoritmos de Minería de Datos paralelos,
- el uso y optimización de la memoria, y
- el impacto de la E/S en el funcionamiento del algoritmo.

Los aspectos de arquitectura se relacionan fuertemente con las estrategias de paralelización, existiendo a su vez una influencia mutua entre la estrategia de extracción del conocimiento y las características arquitectónicas. Por ejemplo, incrementar el grado de paralelismo en algún caso corresponde a aumentar el sobrecosto de la comunicación entre los procesadores. Sin embargo, los costos de la comunicación pueden ser también balanceados por el conocimiento previo de cual algoritmo de Minería de Datos permite obtener mejores resultados con la paralelización.

Un sistema de minería de datos paralelos (*parallel data mining - PDM*) es un sistema fuertemente acoplado en el que se incluyen máquinas de memoria compartida (*Simetric Multi Processor - SMP*), máquinas de memoria distribuida (DMM), o un híbrido entre estas dos arquitecturas, que en sentido general se caracterizan por contar con una red de interconexión muy rápida. Por el contrario, un sistema Minería de Datos Distribuido (*Distributed Data Mining - DDM*) es un sistema ligeramente acoplado, similar a un cluster de workstations, sobre una red de área local muy lenta del tipo Ethernet; en tales sistemas se incluye también sistemas distribuidos geográficamente sobre una red de área ancha similar a Internet. El nivel de acoplamiento de un sistema lo caracteriza la infraestructura de interconexión entre sus componentes.

Las diferencias principales entre PDM y DDM son mejor comprendidas si consideramos al DDM como una transición gradual de un sistema fuertemente acoplado, que pasa por un sistema ligera-

mente acoplado de granularidad media al estilo de las redes de área local y finalmente cae en un sistema de granularidad gruesa como las redes del tipo WAN. Entiéndase que en una máquina paralela mientras la granuralidad es más refinada la latencia de las comunicaciones debe ser menor, lo que permite que la comunicación entre los procesadores sea mucho más eficiente. Visto desde otro ángulo, un DDM puede estar compuesto a su vez de otros DDM y PDM, pero lo contrario no es posible, es decir un PDM no contiene sistemas del tipo DDM.

En la sección siguiente abordaremos algunos de los resultados alcanzados en la formulación paralela de algunos de los métodos típicos de la Minería de Datos, en particular nos referiremos a: algoritmos de clasificación, algoritmos para el descubrimiento de reglas de asociación, minado de sucesiones y algoritmos de agrupamiento.

2. Principales resultados en algunos de los métodos de la Minería de Datos

Hasta la fecha, la paralelización de algoritmos de clasificación basados en árboles de decisión [22] [23][24][25][26][27][31] y la paralelización de algoritmos para el descubrimiento de reglas de asociación [15][16][17][18][19][20][21] constituyen las dos áreas temáticas más trabajadas, aunque en los últimos años ha cobrado fuerza también la paralelización de algoritmos para el minado de sucesiones.

Otra área que también ha ido cobrando fuerza, es el desarrollo de algoritmos secuenciales y paralelos para la búsqueda de los conjuntos de objetos más frecuentes (frequent itemsets), aspecto que juega un rol muy importante en muchas tareas de la Minería de Datos tales como: reglas de asociación, búsqueda de correlación, minado de sucesiones, determinación de sucesos, clasificación, agrupamientos y muchas más, cuyas reglas de asociación es uno de los problemas más populares. Una referencia actualizada de Bart Goethals sobre el desarrollo de algoritmos para la obtención de *"frequent itemsets"* puede ser obtenida de [44], mientras que un reporte técnico del 2003 de Bart Goethals y Mohammed J. Kaki sobre los resultados alcanzados en la implementación de algoritmos de éste tipo puede ser consultado en [45].

Relativamente menor ha sido el trabajo realizado en algoritmos paralelos para otras técnicas de Minería de Datos como son: algoritmos de agrupamiento, regresión, detección de desviación, algoritmos de clasificación basados en otros métodos, etc. Otras posibles áreas de investigación donde también es necesario profundizar incluyen la paralelización y el mejoramiento de muchos algoritmos secuenciales de minería de datos conocidos o nuevos, el análisis profundo y el refinamiento de algoritmos existentes en cuanto a su escalabilidad y eficiencia, el diseño de algoritmos orientado a arquitecturas de memoria compartida y distribuida con multiprocesadores simétricos e integración eficaz de algoritmos paralelos con sistemas de base de datos paralelos.

En [6] se presenta una revisión de diferentes algoritmos paralelos empleados en dos de las técnicas de Minería de Datos más comúnmente utilizadas: asociación y clasificación. Aspectos claves tales como el balanceo de carga, la localidad de los datos, alcanzar la concurrencia maximal, evitar los puntos calientes de embotellamiento y la minimización del sobrecosto de la paralelización, fueron elementos básicos analizados en dichos algoritmos, de la misma forma que se hace con los algoritmos paralelos tradicionales utilizados en diferentes aplicaciones de cálculo científico.

2.1 Algoritmos basados en reglas de asociación

Dos fuentes comparativas muy completas de diferentes formulaciones paralelas existentes para algoritmos basados en reglas de asociación se deben a M. J. Zaki [17] y al colectivo de autores compuesto por: Joshi, E.-H. Han, G. Karypis, y V. Kumar [15], resultados que le antecedieron a dichos artículos aparecen en [19, 20]. El algoritmo Apriori [21] es el algoritmo más conocido para el descubrimiento de reglas de asociación, para el que se han propuesto diferentes implementaciones paralelas. En [16] se presentan tres algoritmos paralelos diferentes llamados *Count Distribution (CD), Data Distribution* (DD) y Candidate Distribution basados en el algoritmo Apriori desarrollado por Agrawal y Shafer, posteriormente E.H. Han, G. Karypis, y V. Kumar [18] presentaron dos nuevas aproximaciones paralelas del Apriori llamadas *Intelligent Data Distribution (IDD) y Hybrid Distribution (HD).*

Morishita y Nakaya en [43] describen un nuevo algoritmo paralelo para la minería de reglas de asociación correlacionadas, para ello utilizaron el estadígrafo Chi-cuadrado, con el que utilizando un árbol de búsqueda calcularon la regla de asociación optima que maximiza el nivel de significación de la correlación entre la hipótesis y la conclusión. Para reducir el tamaño del árbol y optimizar su recorrido, desarrollaron una heurística apropiada de la técnica Ramifica y Acota, conjuntamente con la obtención de las conjunciones de una regla a partir de otras, con lo que evitaron repetir la visita a un mismo nodo del árbol y el mantenimiento de una lista de nodos visitados. Las pruebas realizadas sobre una plataforma de máquinas SMP (con hasta 128 procesadores), mostró muy buenos resultados.

2.1.1 Algoritmos para el minado de sucesiones

Una fuente muy completa, que analiza también algoritmos para el minado de sucesiones, se debe a Joshi y un colectivo de autores [15]. En dicho artículo, ellos discuten muchas de las soluciones paralelas existentes y dan una serie de retos y aspectos a considerar, para lograr formulaciones efectivas de algoritmos de descubrimiento de sucesiones patrones y conjuntos de objetos frecuentes *(frequent itemsets).*

Por otra parte en [47], Zaki presentó pSPADE, un algoritmo paralelo para la minería de sucesiones. El algoritmo pSPADE divide el espacio original de búsqueda en pequeñas clases o problemas disjuntos e independientes, cada uno de los cuales puede ser resuelto en paralelo y de manera asincrónica. Los buenos resultados alcanzados tanto en velocidad de procesamiento como en escalabilidad sobre una máquina SMP de 12 procesadores, se deben al empleo del paralelismo de tareas y a un balanceo dinámico de la carga interclase e intraclase.

2.1.2 Algoritmos de clasificación

En el caso de formulaciones paralelas para algoritmos de clasificación basados en árboles de decisión, se destacan las implementaciones paralelas propuestas en [24][25][26], los algoritmos más referenciados en la literatura sobre el tema, son el SPRINT formulado en 1996 por J. Shafer, R. Agrawal, y M. Mehta [23] y el ScalParC desarrollado en 1998 por M.V. Joshi, G. Karypis, y V. Kumar [27]. Otros resultados son la implementación paralela del algoritmo C4.5 [31] y del TDIDT (Top-Down Induction of Decision Trees) [22].

En [42], Skillicorn presenta algunas técnicas paralelas para generar predictores tanto en modelos de clasificación como de regresión. Una reciente tendencia es construir modelos de predicción

múltiple sobre diferentes muestras partiendo de un conjunto de entrenamiento y combinarlas entre si, lo que permite una inducción más rápida y razones de error más bajas. Este ambiente de trabajo es muy bueno para el paralelismo y constituye el centro de los resultados presentados por Skillicorn en su artículo.

2.1.3 Algoritmos de agrupamiento

En el caso de los algoritmos de agrupamiento, los resultados más recientes son: el *P-AutoClass* [33], la paralelización del *K-Means* [28], el algoritmo GLC [29] y más reciente la paralelización de un algoritmo de agrupamiento incremental, aplicado a la detección de sucesos en un flujo de noticias [30]. En el caso del *P-AutoClass* es una implementación paralela del tipo SPDM del algoritmo *AutoClass*. Respecto al K-Means paralelo hay que señalar que no es directamente aplicable cuando el flujo de objetos a agrupar es incremental. El algoritmo GLC busca de forma incremental las componentes conexas en grafos de semejanzas, sin embargo, este algoritmo presenta un elevado efecto de encadenamiento, lo que provoca que objetos muy poco relacionados entre sí se ubiquen en el mismo grupo. El compacto incremental paralelo presentado en [30] resuelve las problemáticas presentadas en los dos trabajos que le precedieron.

Un poco más atrás en el tiempo son los resultados de finales de la década del 80 que se presentan en [32][35]. Ya en la década del 90 se publica en [36], un análisis comparativo de la complejidad computacional de la implementación sobre diferentes arquitecturas paralelas, de un grupo de algoritmos de agrupamiento jerárquicos, artículo al que le sucedió la formulación paralela de otro algoritmo de agrupamiento al que se le denominó P-CLUSTER [34].

Uno de los métodos de finales de la década de los

90 fue el algoritmo MAFIA [41], que clasifica como un algoritmo de memoria distribuida para agrupar subespacios. A diferencia de los métodos tradicionales como el K-means y los de agrupamiento jerárquico, que determinan los agrupamientos sobre el espacio completo de los datos, es decir, utilizan todas las dimensiones para el cálculo de la distancia, el agrupamiento de subespacios se basa en encontrar grupos que se encuentran empotrados en subconjuntos de un espacio dimensional alto. En este caso MAFIA utiliza redes adaptativas (o intervalos) en cada dimensión que se combinan para encontrar agrupamientos en dimensiones más altas. El algoritmo MAFIA emplea el paralelismo de tarea, donde los datos se reparten equitativamente entre todos los procesadores y cada procesador calcula la densidad local, seguido por una reducción para obtener densidad global.

En [46], Johnson y Kargupta presentan un algoritmo de agrupamiento jerárquico colectivo para bases de datos heterogéneas que trabaja en un ambiente distribuido. En dicho algoritmo, en lugar de recoger los datos de un sitio central, ellos concibieron la generación de modelos de agrupamiento locales que son combinados de forma subsiguiente para obtener el agrupamiento global.

Aunque no era objeto de esta investigación hacer mención a implementaciones paralelas de algoritmos que se aplican en el Reconocimiento de Patrones, no queríamos terminar este epígrafe sin referenciar algunos resultados presentados por Barry Wilkinson y Michael Allen en [53]. El libro publicado por estos autores está dedicado a diferentes aplicaciones de la programación paralela, e incluye un capítulo completo (el número 11) al tema de Procesamiento de Imagen. Aunque la publicación del libro fue en el año 1999, consideramos que constituye un material didáctico de referencia obligada para el análisis e implementación paralela de algoritmos de procesamiento de imágenes y

de otras disciplinas como algoritmos de ordenación, algoritmos para el trabajo con matrices y algoritmos de búsqueda y optimización.

3. Tendencias en el desarrollo de softwares o herramientas

En la actualidad el desarrollo de softwares para la Minería de Datos sigue diferentes tendencias entre las que se pueden citar: la optimización de algoritmos ya conocidos y el desarrollo de nuevos algoritmos y la conformación de paquetes de programas o sistemas, en el que se implementan un conjunto de algoritmos y métodos que incluso puede ejecutarse sobre diferentes sistemas operativos, con relación a esta última, se puede visitar el sitio http://www.togaware.com/ datamining/catalogue.html, donde aparece un conjunto de herramientas comerciales o de libre adquisición "*free download"* que implementan diferentes técnicas, métodos o algoritmos de la Minería de Datos, es notable en este catálogo la ausencia de herramientas paralelas (sólo una). En los dos primeros casos, es decir, la optimización o el desarrollo de nuevos algoritmos se produce en un ambiente más académico, y el tipo de algoritmo sobre el que se trabaja puede ser secuencial o paralelo, sin embargo aunque se dan a conocer los algoritmos, generalmente no se pone a disposición de los interesados su implementación computacional, por lo que en muchos casos otra implementación, puede diferir de los resultados expresados por los autores. En el caso de los algoritmos paralelos hay que señalar la gran dependencia que estos poseen a la arquitectura de hardware sobre la que fueron implementados, lo que puede hacer más notable aún esta diferencia.

Para la Minería de Textos, se reconocen entre los productos comerciales más utilizados en el 2004 según evaluaciones realizadas por Gregory Piatetsky-Shapiro (ver [49]), fundador del prestigioso sitio http://www.kdnuggets.com, los softwares: *SPSS Lexiquest y SAS Text Miner*, seguidos por *Clearforest, Copernic Summarizer, dtSearch, Insightful Infact, Inxight, TEMIS y Wordstat*, ninguno de ellos tiene soporte paralelo o distribuido.

Otros softwares de gran impacto en la aplicación de diferentes métodos de la Minería de Datos son Weka y MatLab, este último también con amplias posibilidades de aplicación al Reconocimiento de Patrones. En ambos casos se trabaja en implementaciones paralelas del tipo "*free download*" como son el proyecto Weka-parallel y diferentes *plug-in* o herramientas que permiten trabajar con MatLab en ambiente paralelo o distribuido, y que pueden ser obtenidas del sitio Mathworks. Un análisis valorativo de diferentes implementaciones paralelas de MatLab aparece en un borrador escrito en el 2003 por Roy y Edelman (veáse [48]) del Massachusetts Institute Tecnology.
Otros ejemplos significativos de sistema de Minería de Datos basados en la ejecución paralela de algoritmos genéticos son GA-MINER, REGAL [38] y G-NET. En el caso de sistemas basados en redes neuronales se encuentra el *Neural Network Utility* (NNU) [37] que fue implementado sobre una máquina paralela IBM SP2.

Otras tendencias que hemos podido apreciar son el incremento del uso de la tecnología Grid, debido principalmente a que los volúmenes de datos utilizados para aplicar las técnicas de Minería de Datos, además de ser cada vez más grandes, se encuentran físicamente distribuidos, siendo precisamente una de las características principales de la tecnología *Grid*, la capacidad de unificar y potenciar las prestaciones de recursos geográficamente distribuidos, principalmente hardware y software heterogéneo, bases de datos y otros recursos; grandes compañías como IBM están enfrascados en es-

ta empresa (veáse [50]), siendo *GridMiner* otro proyecto interesante para el desarrollo de aplicaciones de *Data Mining* sobre la tecnología *Grid*.

Desde el punto de vista del hardware y aunque aun resulta una tecnología relativamente costosa, ya se dan los primeros pasos para la introducción de los FPGAs (*Field Programmable Gate Arrays*) con soporte *bus* SDRAM en el desarrollo de aplicaciones de *Data Mining* del tipo cálculo-intensiva; algunos resultados sobre este particular pueden verse en [51] y [52].

4. Personalidades más destacadas en la temática

Los investigadores que se relacionan a continuación, son los que a criterio del autor han tenido una producción científica más activa durante los últimos años, en el tema del Paralelismo aplicado a la Minería de Datos. Del listado ha sido excluido el brasileño Alex Freitas que en estos momentos no se encuentra trabajando en esta línea de investigación, aunque es meritorio reconocer su indudable contribución al tema con la publicación de su libro "*Mining Very Large Databases with Parallel Processing*", publicado en 1998, donde plasmó los resultados de su tesis doctoral asesorado por el profesor inglés Simon H. Lavington quien se encuentra retirado de la actividad profesional desde el año 2002.

VIPIN KUMAR

Ph.D en Computer Science (Universidad de Maryland., 1982).

Email: *kumar@cs.umn.edu* Web: http://www.cs.umn.edu/~kumar/

Instituciones en las que trabaja

- *Director*, *Army High Performance Computing Research Center*
- *Department of Computer Science and Engineering.*

Intereses de investigación

High performance computing (HPC) and Data Mining

Otros aspectos de interés

Ha publicado más de 150 artículos de investigación y ha sido coeditor o coautor de 8 libros, relacionados con HPC y *Data Mining*.

Ha servido como *chair/ co-chair* de muchas conferencias y talleres en el área de Computación paralela y Minería de Datos, y ha actuado como miembro del Consejo Editorial de varias publicaciones de prestigio.

MOHAMMED JAVEED ZAKI

Ph.D. Computer Science from the University of Rochester, 1998

Email: zaki.AT.cs.rpi.edu Web: http://www.cs.rpi.edu/~zaki/

Institución en la que trabaja

Associate Professor Computer Science Department Rensselaer Polytechnic Institute.

Intereses de investigación

Diseño de algoritmos paralelos escalables y eficientes para varias técnicas de Minería de Datos, está interesado también en el desarrollo de nuevas técnicas de Minería de Datos para la Bioinformática.

Otros aspectos de interés

Ha publicado más de 90 artículos sobre sus intereses de investigación y ha sido coeditor o coautor de 8 libros relacionados con HPC y *Data Mining*, una gran parte de sus artículos ha sido publicado conjuntamente con Vipin Kumar.

Ha participado como miembro del *Program committee* en más de 60 conferencias y workshops en el área de Computación Paralela y Minería de Datos, actuando como miembro del Consejo Editorial de varias publicaciones de prestigio, algunas de ellas auspiciadas por la ACM y la IEEE.

Ha recibido importantes reconocimientos entre ellos: un premio en el 2001 por su carrera profesional del *US National Science Foundation*, otro del Departmento de Energía en el 2002 por su trayectoria como Investigador Principal y un reconocimiento de la ACM en el 2003.

ROBERT GROSSMAN

Email: Web: http://www.rgrossman.com/

Instituciones en las que trabaja

- *Director of the Laboratory for Advanced Computing (LAC) and the National Center for Data Mining (NCDM) at UIC.*
- *Professor in the Department of Mathematics, Statistics, and Computer Science (MSCS) at the University of Illinois at Chicago (UIC).*

Intereses de investigación

Data mining, internet computing, high performance computing, high performance networking, distributed computing y otras areas relacionadas.

Otros aspectos de interés

Es una de las figuras más representativas en cuanto a la producción científica en las áreas de Minería de Datos y *High Performance Computing*, ha sido coeditor y coautor de varios libros relacionados

con estos temas, jugando también un papel muy activo como miembro del Comité Técnico de muchas conferencias y talleres desarrollados sobre las disciplinas antes mencionadas.

Es presidente de la compañía *Open Data Partners* que da servicios y consultorías relacionados con datos también preside *Magnify, Inc.*, una compañía que proporciona análisis a posteriori a compañías que ofertan servicios financieros.

SUDIPTO GUHA

Ph.D. in Computer Science (Maryland Univ. 1982)

Email: *sudipto@cis.upenn.edu* Web: http://www.cis.upenn.edu/~sudipto/

Institución en la que trabaja

Department of Computer and Information Science University of Pennsylvania

Intereses de investigación

Diseño y análisis de algoritmos que dependen de pocos recursos computacionales, algoritmos de aproximación de grafos para problemas NP-duros, algoritmos de una sola pasada para el análisis de flujos de datos, optimización de búsqueda y minería en bases de datos, algoritmos aleatorios y de optimización combinatoria.

GEORGE KARYPIS

Ph.D. in Computer Science.

Email: karypis@cs.umn.edu Web: http://www.cs.umn.edu/~karypis

Institución en la que trabaja

Associated Professor of Information and Computer Science of University of Minnesota.

Intereses de investigación

Parallel algorithm design; data mining; bioinformatics; information retrieval; applications of parallel processing in scientific computing and optimization; sparse matrix computations; parallel preconditioners; and parallel programming languages and libraries.

Otros aspectos de interés

Es coautor de más de noventa artículos publicados en revistas y conferencias sobre Minería de Datos y *Parallel Data Mining* y un libro titulado "*Introduction to Parallel Computing*" (Publ. Addison Wesley, 2003, segunda edición), ha servido además en los comités del programa de muchas conferencias y talleres sobre estos temas y es editor asociado de la IEEE *Transactions on Parallel and Distributed Systems.*

DAVID SKILLICORN

Ph.D. in Computer Science

Email: Web: http://www.cs.queensu.ca/home/skill/

Institución en la que trabaja

Profesor de la Escuela de Computación de la Universidad de Queen, Kingston, Ontario, Canada y Profesor adjunto del Departamento de Matemática y Computación del *Royal Military College*.

Intereses de investigación

Mining Multiprocess Data; Fraud Detection, Counterterrorism, and Intrusion Detection; Data Grid Mining and Parallel and Distributed Data Mining Algorithms.

Otros aspectos de interés

Ha actuado como miembro del Comité de Programa de varias ediciones de la IEEE *Internacional Conference on Data Mining* y de la SIAM *Conference on Data Mining*, así como de otros eventos sobre el tema de *High Performance Data Mining*. Aparece en el lugar 24 de los investigadores canadienses más citados en *Citeseer*.

DOMENICO TALIA

Ph.D. in Computer Science.

Email: talia@deis.unical.it Web: http://si.deis.unical.it/~talia/

Institución en la que trabaja

- *Professor of Computer Science at the Faculty of Engineering of University of Calabria, Italia.*
- Investigador Asociado de ICAR-CNR.
- Socio de la firma Exeura.

Intereses de investigación

Grid Computing, Parallel Data Mining, Peer-to-Peer Computing, Parallel Programming Languages, Cellular Automata, and Distributed Systems.

GAGAN AGRAWAL

Ph.D. in Computer Science, University of Maryland, College Park (Aug 1996)

Email: Web: http://www.cse.ohio-state.edu/~agrawal/

Institución en la que trabaja

Associate profesor of Computer and Information Sciences Department at the Ohio State University, trabaja también con el *Department of Bio-Medical Informatic* **.**

Intereses de investigación

Parallel and Distributed Systems; Data Mining; Online Analytical Processing (OLAP); Compiler and Middleware Systems; Grid Computing; Processing of Data Streams.

Otros aspectos de interés

Dentro de la propia Universidad de Ohio, tiene vínculos con los siguientes grupos de investigación: *High-end Computing Systems Group of Computer and Information Sciences Department and Division of Data Intensive and Grid Computing of Biomedic Informatic Department.*

HILLOL KARGUPTA

Ph.D. in Computer Science from University of Illinois at Urbana-Champaign in 1996

Email: hillol AT cs DOT umbc DOT edu Web: http://www.cs.umbc.edu/~hillol/Kargupta/

Institución en la que trabaja

Associate professor of Computer and Information Science Department in the Ohio State University

Intereses de investigación

Minería de Datos distribuida y omnipresente: desarrollo de algoritmos y sistemas experimentales; aspectos de privacidad en la minería de datos distribuida; cálculos sobre expresiones genéticas, algoritmos genéticos y sistemas evolutivos.

MAHESH JOSHI

Ph.D. of Philosophy in Computer Science at University of Minnesota, Minneapolis, 2002.
Tutores: Prof. Vipin Kumar y Dr. Ramesh Agrawal, IBM Research.

Email: Web: http://www.cs.umn.edu/~mjoshi

Institución en la que trabaja

IBM Almaden Research Center, San Jose, CA

Intereses de investigación

Data Mining Algorithms; Parallel Scientific Computing; Fuzzy Logic and Neural Networks in Control Systems.

SRINIVASAN PARTHASARATHY

Email: *srini@ cis.ohio-state.edu* Web: http://www.cis.ohio-state.edu/~srini

Institución en la que trabaja

Associate professor of Computer and Information Science Department of the Ohio State University

Intereses de investigación

Data Mining (Systems, Algorithms and Applications); Bioinformatics; Parallel and Distributed Systems.

5. Grupos de investigación más relevantes

- *Departament of Computer Sciences and Engineering* de la Universidad de Minnesota, al que pertenecen Vipin Kumar y George Karypis. En esta propia universidad aparece también el *Distributed Computing Systems Group* (DCSG), encabezado por el profesor Jon Weissman de la Universidad de Minnesota.
- *High-End Computing Systems Group*, del *Departament of Computer and Information Science* de la Universidad de Ohio, al cual pertenecen Gagan Agrawal y Srinivasan Parthasarathy
- *National Center for Data Mining*, de la Universidad de Illinois, Chicago, USA.
- *Laboratory for Advanced Computing*, dirigido por Robert Grossman, de la Universidad de Illinois, Chicago, USA.
- *San Diego Supercomputer Center*, es uno de los centros norteamericanos de mayor desarrollo en computación paralela y distribuida, con resultados importantes en el campo de la tecnología Grid.
- *High-Performance Computing Laboratory*, encabezado por el Profesor David Skillicorn es parte de la escuela de Computación de la Universidad de Queen, Kingston, Ontario, Canada.
- *Distributed Adaptive Discovery and Computation Laboratory* (DIADIC), dirigido por Hillol Kargupta.

6. Eventos científicos

Entre los eventos científicos más importantes se cuentan:

- *International Workshop on High Performance and Distributed Mining* (*HPDM*)

- *ICDM, IEEE Internacional Conference on Data Mining*.
- *EuroPar* (http://www.euro-par.org/)
- *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*
- *ACM Symposium of Applied Computing (SAC).*
- *International Conference on Very Large Databases*
- *Intelligent Data Analysis* (IDA)

7. Principales publicaciones que abordan el tema

- *IEEE Transactions on Knowledge and Data Engineering*
- *Knowledge and Information Systems: An International Journal*
- *Data Mining and Knowledge Discovery: An International Journal*
- *IEEE Transactions on Pattern Analysis and Machine Intelligence*
- *Journal on Artificial Intelligence Research*
- *ACM SIGKDD Explorations*
- *Data and Knowledge Engineering*
- *Parallel Processing for Artificial Intelligence. Elsevier Science.*
- *Lecture Notes in Artificial Intelligence. Springer Verlag.*
- *Lecture Notes in Computer Sciences. Springer Verlag*
- *Distributed and Parallel Databases. Kluwer Academic Publishers*

Conclusiones

La Computación Paralela y Distribuida, es una rama computacional que se encuentra en constante desarrollo, como consecuencia de los adelantos tecnológicos que se vienen introduciendo en el desarrollo del hardware y la evolución que van teniendo los diferentes modelos de programación paralela existentes. Todo ello repercute sin dudas en la Minería de Datos, donde ya se observa un incremento sustancial de formulaciones paralelas de algoritmos que se aplican en determinadas áreas de esta disciplina como: clasificación, reglas de asociación y clustering, así como en determinadas etapas del preprocesamiento y limpieza de los datos.

En la formulación e implementación paralela de un algoritmo de Minería de Datos u otro algoritmo aplicable a cualquier disciplina, hay que analizar varios aspectos además del modelo matemático sobre el que se sustenta, entre ellos: el modelo de programación (memoria compartida o pase de mensajes) a emplear, fuertemente vinculado a la topología de interconexión de los procesadores; eficiencia de la comunicación entre los procesos; balanceo de la carga para la ejecución de los cálculos en dependencia de la heterogeneidad del hardware; plataforma de desarrollo, dígase sistema operativo y lenguaje de programación; optimización del uso de la memoria y por último, el impacto de las operaciones de entrada y salida en la eficiencia del algoritmo. Estos y otros aspectos inciden de forma sustancial en la escalabilidad, eficiencia y explotación del paralelismo en los algoritmos de Minería de Datos y por consiguiente, constituyen líneas de investigación a nivel internacional.

Además de las anteriormente mencionadas, existen otras áreas de investigación donde es necesario la combinación del paralelismo con diferentes métodos o técnicas de la Minería de Datos, para obtener mayor capacidad de procesamiento y una elevada eficiencia, entre ellas podemos relacionar:

- El desarrollo de ambientes y herramientas para de forma interactiva aplicar la computación paralela a la Minería de Datos
- Profundizar en el uso de técnicas de Minería de Datos paralelas al “Text Mining” y a la “Multimedia Data Mining”.
- El “Web Mining” paralelo y distribuido continua siendo un área muy prometedora para la explotación de las técnicas de la computación paralela.
- La aplicación de técnicas de Minería sobre bases de datos distribuidas y Datawarehouse, constituye un aspecto vital para empresas y organizaciones públicas con grandes volúmenes de datos.
- Además de éstos, un área muy prometedora que también debemos mencionar es el desarrollo de softwares, ambientes y herramientas que permitan el uso combinado de *Clusters y Grids*, considerando que un *Grid* es capaz de aglutinar a varios clusters de computadoras que ejecutan el mismo o diferentes algoritmos de Minería de Datos y que pueden ser vistos como computadoras masivamente paralelas que minan volúmenes de datos muy grandes, distribuidos geográficamente.

Por nuestra parte, será objeto de investigación en el futuro, diferentes aspectos relacionados con:

- La eficiencia de la implementación de algoritmos de Minería de Datos y de Reconocimiento de Patrones sobre *clusters* del tipo Beowulf, así como el análisis y evaluación de diferentes herramientas de administración y desarrollo que faciliten al investigador la rápida asimilación de la programación paralela y de la plataforma de hardware sobre la que se sustenta.
- La implementación y desarrollo de algoritmos paralelos para reglas de asociación y clustering en documentos.
- El diseño e implementación de bases de datos en ambiente distribuido y la aplicación sobre ellas de técnicas paralelas de Minería de Datos.
- La evaluación periódica de nuevas tecnologías de hardware (procesador, memoria, dispositivos de interconexión, etc) que inciden en el rendimiento de aplicaciones paralelas de Minería de Datos y de Reconocimiento de Patrones.

Referencias Bilibográficas

[1] Yike Guo, Robert Grossman, "High Performance Data Mining, Scaling Algorithms, Applications and Systems", Kluwer Academic Publishers, 2002.

[2] Alex A. Freitas, Simon H. Lavington, "Mining Very Large Databases with Parallel Processing", Kluwer Academic Publishers, 1998.

[3] Mario Cannataro, "Clusters and Grids for Distributed and Parallel Knowledge Discovery" HPCN 2000, LNCS 1823, pp. 708-716, 2000.

[4] Vipin Kumar, Ananth Grama, Anshul Gupta, George Karypis "Introduction to Parallel Computing: Design and Analysis of Algorithms", Benjamin/Cummings, 1994.

[5] Ananth Grama, Anshul Gupta, George Karypis, Vipin Kumar "Introduction to Parallel Computing, Second Edition", Pearson Education Limited, 2003.

[6] Vipin Kumar, Mahesh V. Joshi, Eui-Hong (Sam) Han, Pang-Ning Tan, Michael Steinbach, "High Performance Data Mining", http://www-users.cs.umn.edu/~kumar/papers/vecpar.pdf, VECPAR-2002, Lecture Notes in Computer Science 2565 Springer 2003.

[7] Mohammed J. Zaki, "Workshop Report: Large-Scale Parallel KDD Systems". http://www.cs.rpi.edu/~zaki/WKDD99

[8] Mohammed J. Zaki, Ching-Tien Ho (Eds.), "Large-Scale Parallel Data Mining", Lecture Notes in Artificial Intelligence, Vol 1759, Springer Verlag 2000.

[9] Mohammed J. Zaki, Yi Pan, "Introduction: Recent Developments in Parallel and Distributed Data Mining", in Mohammed J. Zaki, Yi Pan (Eds.) Distributed and Parallel Databases, 11, 123–127, 2002, Kluwer Academic Publishers.

[10] M.J. Zaki. "Tutorial: Data Mining and KDD", New Directions in Bioinformatics and Biotechnology Workshop, Troy, NY, June 1999.

[11] M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from database perspective. IEEE Transactions on Knowledge and Data Eng., 8(6):866–883, December 1996.

[12] Vipin Kumar, Mohammed J. Zaki, High-Performance Data Mining, Tutorial at KDD-2000, August 20, 2000.

[13] Domenico Talia, "Technical Report - Parallel and Distributed Data Mining: From Multicomputers to Grids", University of Calabria, Italy, 23 September 2003.

[14] http://si.deis.unical.it/~talia/

[15] M.V. Joshi, E.-H. Han, G. Karypis, and V. Kumar, "Efficient parallel algorithms for mining associations", in M. J. Zaki and C.-T. Ho, editors, Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence (LNCS/LNAI), volume 1759. Springer-Verlag, 2000.

[16] R. Agrawal and J.C. Shafer. Parallel mining of association rules. IEEE Transactions on Knowledge and Data Engineering., 8(6):962–969, December 1996.

[17] M. J. Zaki. Parallel and distributed association mining: A survey. IEEE Concurrency (Special Issue on Data Mining), December 1999.

[18] E.H. Han, G. Karypis, and V. Kumar. Scalable parallel data mining for association rules. IEEE Transactions on Knowledge and Data Engineering., Vol. XX, No. Y, Month 1999.

[19] E.H. Han, G. Karypis, and V. Kumar. Scalable parallel data mining for association rules. In Proc. of 1997 ACM-SIGMOD Int. Conf. on Management of Data, Tucson, Arizona, 1997.

[20] M.V. Joshi, G. Karypis, and V. Kumar. Universal formulation of sequential patterns. Technical Report TR 99-021, Department of Computer Science, University of Minnesota, Minneapolis, 1999.

[21] R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules, Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, 1994.

[22] R. A. Pearson. A coarse grained parallel induction heuristic. In H. Kitano, V. Kumar, and C.B. Suttner, editors, Parallel Processing for Artificial Intelligence 2, pages 207–226. Elsevier Science, 1994.
[23] J. Shafer, R. Agrawal, and M. Mehta. SPRINT: A scalable parallel classifier for data mining. In Proc. of the 22nd VLDB Conference, 1996.
[24] S. Goil, S. Aluru, and S. Ranka. Concatenated parallelism: A technique for efficient parallel divide and conquer. In Proc. of the Symposium of Parallel and Distributed Computing (SPDP'96), 1996.
[25] J. Chattratichat, J. Darlington, M. Ghanem, Y. Guo, H. Huning, M. Kohler, J. Sutiwaraphun, H.W. To, and D. Yang. Large scale data mining: Challenges and responses. In Proc. of the Third Int'l Conference on Knowledge Discovery and Data Mining, 1997.
[26] R. Kufrin. Decision trees on parallel processors. In J. Geller, H. Kitano, and C. B. Suttner, editors, Parallel Processing for Artificial Intelligence 3. Elsevier Science, 1997.
[27] M.V. Joshi, G. Karypis, and V. Kumar. ScalParC: A new scalable and efficient parallel classification algorithm for mining large datasets. In Proc. of the International Parallel Processing Symposium, 1998.
[28] Dhillon, I., Modha, B. A. Data Clustering Algorithm On Distributed Memory Multiprocesor, Workshop on Largescale Parallel KDD Systems, pp. 245-260, 2000.
[29] Gil-García, R. , Badía-Contelles, J. "Algoritmo de Agrupamiento Paralelo GLC". In Pattern Recognition. Advances and Perspectives. Research on Computing Science, CIARP'2002. México, pp. 383-394, 2002.
[30] Reynaldo Gil García, José Manuel Badia Contelles, Aurora Pons Porrata., "A parallel algorithm for incremental compact clustering". LECTURE NOTES IN CONTROL AND INFORMATION SCIENCES. Num. 2790. pp. 310-317. 2003.
[31] R. Kufrin, "Generating C4.5 Production Rules in Parallel", Proc. 14th Nat. Conf. on Artificial Intelligence - AAAI-97, AAAI Press, 1997.
[32] M. Bruynooghe, Parallel Implementation of Fast Clustering Algorithms, Proc. Int. Symp. On High Performance Computing, pp. 65-78, 1989.
[33] D. Foti, D. Lipari, C. Pizzuti and D. Talia, Scalable Parallel Clustering for Data Mining on Multicomputers, Proc. of the 3rd Int. Workshop on High Performance Data Mining HPDM00-IPDPS, Cancun, LNCS 1800, pp. 390-398, Springer-Verlag, 2000.
[34] D. Judd, K. McKinley and A.K. Jain, Large-Scale Parallel Data Clustering, Proc. Int. Conf. On Pattern Recognition, Vienna, 1996.
[35] X. Li and Z. Fang, Parallel Clustering Algorithms, Parallel Computing, 11, pp. 275-290, 1989.
[36] C.F. Olson, Parallel Algorithms for Hierarchical Clustering, Parallel Computing, 21, pp. 1313-1325, 1995.
[37] J.P. Bigus, Data Mining with Neural Networks, McGraw-Hill, New York, 1996.
[38] F. Neri and A. Giordana, A Parallel Genetic Algorithm for Concept Learning, Proc. 6th Int. Conf. Genetic Algorithms, pp. 436-443, 1995.
[39] D. Talia, "Parallelism in Knowledge Discovery Techniques" in J. Fagerholm et al. (Eds.): PARA 2002, LNCS 2367, pp. 127–136, 2002.
[40] D. Talia, "High-Performance Data Mining and Knowledge Discovery", Tutorial in Euro-Par2002, Paderborn, Germany, August 2002.
[41] S. Goil, H.N., Choudhary, A.: MAFIA: Eficient and Scalable Subspace Clustering for Very Large Data Sets. Technical Report 9906-010, Center for Parallel and Distributed Computing, Northwestern University (1999).
[42] D. B. Skillicorn, "Parallel Predictor Generation" in Mohammed J. Zaki, Ching-Tien Ho (Eds.), "Large-Scale Parallel Data Mining", Lecture Notes in Artificial Intelligence, Vol 1759, p. 190-196, Springer Verlag, 2000.

[43] Shinichi Morishita, Akihiro Nakaya, “Parallel Branch and Bound Graph Search for Correlated Association Rules” in Mohammed J. Zaki, Ching-Tien Ho (Eds.), “Large-Scale Parallel Data Mining”, Lecture Notes in Artificial Intelligence, Vol 1759, p. 127-144, Springer Verlag, 2000.
[44] http://citeseer.ist.psu.edu/goethals03survey.html
[45] http://citeseer.ist.psu.edu/684802.html
[46] Erik L. Johnson, Hillol Kargupta, “Collective, Hierarchical Clustering from Distributed, Heterogeneous Data” in Mohammed J. Zaki, Ching-Tien Ho (Eds.), “Large-Scale Parallel Data Mining”, Lecture Notes in Artificial Intelligence, Vol 1759, p. 221-244, Springer Verlag, 2000.
[47] Mohamed J. Zaki, “Parallel Sequences Mining on Shared-Memory Machines” in Mohammed J. Zaki, Ching-Tien Ho (Eds.), “Large-Scale Parallel Data Mining”, Lecture Notes in Artificial Intelligence, Vol 1759, p. 161-189, Springer Verlag, 2000.
[48] Ron Choy, Alan Edelman, “Parallel MATLAB: Doing it Right” Computer Science AI Laboratory, Massachusetts Institute of Technology, 2003.
[49] http://www.kdnuggets.com/news/ 2005/n02/1i.html
[50] Viktors Berstis, “Fundamentals of Grid Computing”, IBM Red Books paper, ibm.com/redbooks 2002.
[51] Zachary K. Baker y Viktor K. Prasanna, “Efficient Parallel Data Mining with the Apriori Algorithm on FPGAs”
[52] Kelvin T. Leung1, Professor Milos Ercegovac, Professor Richard R. Muntz, “Exploiting Reconfigurable FPGA for Parallel Query Processing in Computation Intensive”
[53] Barry Wilkinson, Michael Allen, “Parallel Programming, Techniques and Applications using Networking Workstations and Parallel Computers”, ISBN: 0-13-671710-1, Prentice Hall, 1999.

RT_002, Julio 2008
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Miriela Santos Toledo
Diseño de Portada: DCG Matilde Galindo Sánchez
RNPS No. 2143
ISSN 2072-6260
Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. No. 21812 e/218 y 222, Rpto. Siboney, Playa;
Ciudad de La Habana. Cuba. C.P. 12200
Impreso en Cuba