

Universidad Nacional de Asunción

Facultad Politécnica

Ingeniería Informática

Trabajo Final De Grado

TDRSBM: Toma de decisiones para reposición de stock basado en Business Intelligence y Machine Learning

Mayo 2017

1. **SÍNTESIS DEL PROYECTO**
   1. TÍTULO INICIAL DEL PROYECTO

|  |
| --- |
| TDRSBM: Toma de decisiones para reposición de stock basado en Business Intelligence y Machine Learning |

SINTETIZAR EL PROYECTO EN TRES PALABRAS CLAVE

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Reposición de Stock |  | Business Intelligence |  | Machine Learning |

* 1. DURACIÓN DEL PROYECTO

|  |  |
| --- | --- |
| 12 | **MESES** |

* 1. DATOS DE LOS ESTUDIANTES Y DEL ASESOR DEL PROYECTO

|  |  |
| --- | --- |
| **NOMBRE DEL ASESOR PRINCIPAL:**  Ph.D. Diego Pinto | |
| **Email:** dppintoroa@gmail.com | **Teléfono:** 0984 401 410 |
| **CARGO:**  Docente | **FIRMA:** |

|  |  |
| --- | --- |
| **NOMBRE DEL ASESOR SECUNDARIO:**  Ing. Aditardo Vázquez | |
| **Email:** vazquez.aditardo@gmail.com | **Teléfono:** 0981 358 294 |
| **CARGO:**  Docente | **FIRMA:** |

|  |  |
| --- | --- |
| **NOMBRE DEL ALUMNO:**  Alberto Gabriel Garcete González | |
| **CI:** 3.676.206 | **PLAN DE ESTUDIOS:** 2008 |
| **Email:** albertogarcetepy@gmail.com | **Teléfono:** 0971 494427 |
| **AÑO DE INGRESO: 2008** | **FIRMA:** |

|  |  |
| --- | --- |
| **NOMBRE DEL ALUMNO:**  Raúl Alberto Benítez Martínez | |
| **CI:** 4.089.446 | **PLAN DE ESTUDIOS:** 2001 |
| **Email:** raulkvd@gmail.com | **Teléfono:** 0984 196134 |
| **AÑO DE INGRESO: 2007** | **FIRMA:** |

* 1. **RESUMEN DEL PROYECTO**

Reposición de stock (métodos actuales)

En las empresas retail o de ventas minoristas uno de los principales desafíos que enfrentan es el manejo eficiente del stock, de manera a evitar tener cantidad de productos en exceso en los depósitos que incurran en sobrecostos, o en otro extremo evitar la falta de dichos productos o ruptura de stock lo cual conlleva a pérdidas de oportunidades de ventas por no disponer del producto que puede generar insatisfacción de los clientes y a su vez repercute en las utilidades de la empresa. Uno de los mayores desafíos de las empresas es la de estimar o predecir la cantidad de ventas para el próximo periodo de tiempo.

Actualmente en el proceso de gestión de compras en empresas como las retail, se utilizan algunas de las técnicas de pronósticos para determinar las cantidades de las órdenes de compra, las cuales pueden estar basadas en pronósticos cuantitativos o cualitativos. Los modelos de cantidad fija y los modelos de periodo fijo son ampliamente utilizados. Independientemente de la técnica elegida, el problema real de los pronósticos es su falta de confiabilidad, ya que por lo general no son precisos, entonces, la interrogante que siempre surge es si serán superiores o inferiores a la demanda real y en qué medida.

Business Intelligence (KPI)

Con el presente trabajo se elabora un nuevo modelo de estimación de cantidades eficientes en las órdenes de compra de productos para la reposición de stock del siguiente periodo de venta. En este nuevo modelo se integran técnicas de Business Intelligence y Machine Learning.

En la etapa de Business Intelligence el objetivo principal es calcular los Indicadores Claves de Rendimiento (KPI - Key Performance Indicators) de los productos en base a los datos históricos obtenidos de la base de datos transaccional. Luego cada serie de KPI obtenidos pasan por un proceso de etiquetado, donde el experto en compras los analiza y determina qué nivel de compra conviene para cada serie de KPI.

Machine Learning (Clasificación)

En la etapa de Machine Learning se utiliza como entrada las series de KPI obtenidos en la etapa de Business Intelligence y que constituyen las instancias que alimentan los distintos algoritmos de clasificación de Machine Learning supervisado. Luego tienen lugar los procesos propios de esta etapa que son el entrenamiento y testeo para finalmente evaluar los distintos desempeños a fin de determinar los algoritmos más adecuados que serán utilizados para estimar las cantidades de las ordenes de compra por cada producto.

Experimentos

En cuanto a las limitaciones de este nuevo modelo se puede mencionar que no toma en cuenta los costes relacionados al inventario: como costes de mantenimiento, de personal, seguros, etc. El modelo planteado se aplica a empresas retail dedicadas a la venta de productos terminados.

* 1. **OBJETIVOS DEL PROYECTO.** 
     1. OBJETIVOS GENERALES
* Modificar y extender el algoritmo TKU para que pueda ser un algoritmo paralelo distribuido y comparable con otros algoritmos, permitiendo así la optimización del tiempo de descubrimiento de high utility itemsets en volúmenes de datos muy grandes.
  + 1. OBJETIVOS ESPECÍFICOS
* Implementar una modificación del algoritmo TKU para poder realizar comparaciones con otros algoritmos que asemejan su implementación en cuanto a entrada y salida del mismo.
* Extender el algoritmo TKU [3] mediante modelos como map reduce de manera que pueda ser paralelizable.
* Probar el algoritmo desarrollado y comparar los resultados con el algoritmo TKU.

1. **FORMULACIÓN GENERAL DEL PROYECTO**
   1. **PROBLEMA E HIPÓTESIS**

2.1.1 ¿CUÁL ES EL PROBLEMA U OPORTUNIDAD?

Existen muchos algoritmos para descubrir High Utility Itemsets, con distintos enfoques, estos algoritmos han ido mejorando con el tiempo enfocándose en la generación de candidatos, pero su rendimiento se ve afectado cuando se trabajan con volúmenes de datos muy grandes debido a la gran cantidad de memoria y tiempo que se necesita.

Uno de los algoritmos que presenta un buen desempeño es llamado EFIM [6] la ventaja que presenta el algoritmo de TKU sobre éste es que no requiere la entrada de un valor denominado utilidad mínima, en vez de está requiere la entrada de cuantos resultados es deseado obtener, siendo éste el valor de K.

La ventaja de requerir un valor K frente a una la utilidad mínima es que si el valor de la utilidad mínima es muy chica la cantidad de resultados será escasa y si es muy grande los resultados no serán de gran utilidad, a diferencia de si se tiene como entrada una cantidad K el resultado será más especificó proveyendo así más utilidad de resultados.

2.1.2 ¿CUÁL ES LA SOLUCIÓN PROPUESTA POR EL PROYECTO DE INVESTIGACIÓN?

Realizar la modificación del algoritmo TKU para que su comportamiento permite la comparación con otros algoritmos, en cuanto a la entrada y salida del mismo, y extender el algoritmo para buscar un mejor rendimiento mediante la paralelización y ejecución distribuida del mismo permitiendo agilizar la búsqueda de high utility itemsets mediante el framework Apache Spark?.

2.1.3 ¿CUÁL ES LA HIPÓTESIS DEL PROYECTO?

Se espera que el algoritmo PMTKU que se pretende desarrollar optimice el tiempo de búsqueda de high utility itemsets comparado con el TKU base y otros algoritmos que sigan el mismo patrón de entrada a la hora de tratar con volúmenes de datos muy grandes

* 1. **ANTECEDENTES.**

Estado el arte.

Aborde histórico del pronóstico de la demanda.

El pronóstico de la demanda es una predicción de eventos que ocurrirán, es una técnica utilizada por las empresas para anticiparse al futuro y elaborar las estrategias apropiadas [1]. En las empresas, durante décadas han intentado optimizar la cantidad de inventario por medio de una predicción más efectiva de la demanda de sus clientes, sin embargo la naturaleza incierta de las predicciones muchas veces ha contribuido más a la incertidumbre en lugar de minimizarlo. “*Los pronósticos jamás son perfectos y serán menos confiables mientras mayor sea el lapso que se pronostique hacia el futuro*” [2]

Las técnicas empleadas para la realización de pronósticos se clasifican en:

* Técnicas Cualitativas: las técnicas cualitativas se basan principalmente en el conocimiento humano, en el juicio de un experto para elaborar las estimaciones futuras. En ocasiones son conocidas como técnicas subjetivas y son utilizadas cuando no se tienen información histórica sobre la variable que se está pronosticando.
* Técnicas Cuantitativas: las técnicas cuantitativas se utilizan cuando se dispone de información histórica de las variables a estudiar, el pronóstico se basa en modelos matemáticos principalmente los estadísticos. Los métodos cuantitativos se dividen en a) *métodos de serie de tiempo* se utiliza cuando la variable a pronosticar únicamente hace relación a la información histórica que se posee observados en periodos de tiempo iguales y b) *métodos de pronóstico causal* están basadas en el supuesto de que la variable que se quiere pronosticar tiene relación de causa y efecto con una o más variables se trata de determinar exactamente cuáles son dichas relaciones.

Aborde histórico de Business Intelligence (cómo y para qué) en empresas retatil.

La mayoría de las organizaciones contaban con un sistema de información donde se guardan las actividades diarias realizadas en la empresa, con el transcurso del tiempo estas informaciones van creando un histórico de las actividades que generalmente son almacenadas en una base de datos [3] pero que no disponían de un mecanismo que permita un rápido y fácil acceso a dicha información.

En el año 1989, Howard Dresner consultor de Gartner Group acuñó por primera vez el término de Busines Intelligence, que abarca un conjunto de conceptos, técnicas y herramientas que se utiliza para la transformación de datos transaccionales en información útil y significativa para el análisis de negocios, tiene como objetivo principal apoyar a las organizaciones a mejorar su competitividad, facilitando la información necesaria para la toma de decisiones [4]. Entre las muchas áreas donde implementan soluciones de Business Intelligence podemos citar: Ventas Minoristas en empresas retail, Inventario, Gestión de Pedidos, Contabilidad, Gestión de Relación con el Cliente, Gestión de recursos humanos, Servicios financieros, Telecomunicaciones, Educación, Medicina, Comercio Electrónico, Seguros entre otros [5].

Aborde histórico de Machine Learning para decisiones en empresas retail.

Frequent pattern mining ha sido un tema importante desde que el concepto de ‘conjunto de elementos frecuentes’ fue introducido por primera vez por Agrawal et al. [6]. Dado las transacciones de un conjunto de datos, frequent pattern mining encuentra los conjuntos de elementos cuyo soporte (es decir, el porcentaje de transacciones que contienen el conjunto de elementos) no es menor que un umbral de mínimo soporte dado. Sin embargo, ni el número de ocurrencias de un elemento en una transacción, ni la importancia de un elemento, se considera en frequent pattern mining. Los conjuntos de elementos con más ocurrencias o importancia pueden ser más interesantes para los usuarios, ya que pueden traer más beneficios.

A la luz de esto, se ha estudiado high utility itemset mining [4 - 5]. En high utility itemset mining, el término utilidad se refiere a la importancia de un conjunto de elementos; Por ejemplo, la ganancia total que el conjunto de elementos trae. Un conjunto de elementos es un conjunto de elementos de alta utilidad (HUI) si la utilidad del conjunto de elementos no es inferior a un umbral mínimo determinado.

High utility itemset mining se centra más en los valores de utilidad del conjunto de datos, que suelen estar relacionados con los beneficios del negocio. Estas utilidades son de interés para los propietarios del negocio, que podrían obtener más beneficios de ellos. Por ejemplo, los supermercados usan frequent itemset mining para encontrar productos que los clientes generalmente compran, para hacer recomendaciones a los clientes. Sin embargo, con high utility itemset mining, los supermercados podrán recomendar no sólo los productos que la gente suele comprar juntos, sino también los productos que generan más ganancias para la tienda esto sería de acuerdo a la ganancia o categoría del producto es decir si el cliente lleva una bebida de alta calidad está debería ser acompañada con algún producto comestible del mismo nivel de calidad, mientras mayor sea la calidad del producto mayor ganancia generará para el supermercado también podría ser visto desde un punto de vista educativo si bien existen materias complicadas dentro de una institución académica no todas las materias requieren el mismo esfuerzo, si se aplica la ejecución de un algoritmo que realice el cálculo de frequent itemsets se podría encontrar patrones que indiquen qué combinación de materias que requieran el mismo esfuerzo generan mayor cantidad de alumnos que deben volver a tomar la materia por segunda o inclusive tercera vez para poder avanzar en su plan curricular, teniendo estos datos la institución académica es capaz de revisar y analizar el contenido de las materias para realizar un ajuste en éste o poder estructurar una mejor planificación de las materias.

La mayoría de los algoritmos de frequent pattern mining eliminan los conjuntos de elementos en una etapa temprana, basados ​​en la popular propiedad Apriori [11]: cada sub-patrón de un patrón frecuente debe ser frecuente (también llamado propiedad de cierre descendente). Sin embargo, esta propiedad no se sostiene en high utility itemset mining, lo que hace que encontrar los conjuntos de elementos de alta utilidad sean más difíciles. Los enfoques del estado del arte alcanzan un buen rendimiento cuando el conjunto de datos es relativamente pequeño. Sin embargo, el volumen de datos puede crecer más rápido de lo esperado, que una sola máquina puede no llegar a ser suficiente para manejar grandes volúmenes de datos.

Una particularidad que posee el algoritmo de TKU o top-k high utility itemsets es que sigue un patrón de búsqueda diferente al resto de los algoritmos que realizan la búsqueda de high utility itemsets, debido a que éste realiza la búsqueda de las k mejores opciones que existen en los datos analizados sin embargo los otros algoritmos devuelven sus resultados a partir de un valor de utilidad mínima.

Una opción para resolver el problema de grandes volúmenes de datos es utilizar técnicas de computación paralela distribuida. El framework MapReduce [12] (por ejemplo, Hadoop) ha sido una solución popular recientemente, que permite el procesamiento distribuido de grandes volúmenes de datos en clusters de manera escalable y tolerante a fallos. Las aplicaciones que utilizan el framework de MapReduce tienen que conformar los protocolos de mapeo y reducción en un paradigma basado en disco, que restringe la flexibilidad así como el funcionamiento del algoritmo. Spark es también un framework de computación distribuida, basado en la memoria, que proporciona un rendimiento hasta 100 veces más rápido que Hadoop para ciertas aplicaciones [13]. Spark utiliza Resilient Distributed Dataset (RDD), que es una abstracción de memoria distribuida, para el cálculo de datos en memoria, lo que permite una reutilización eficiente de los datos.

En este trabajo, abordamos el problema de high utility itemset mining al proponer PMTKU (Parallel Modification algorithm for top-k high utility itemset). PMTKU es un algoritmo paralelo distribuido, que paraleliza a TKU y modifica la entrada del algoritmo para descubrir high utility itemset.

* 1. **METODOLOGÍAS**
* Investigación del estado del arte de:
  + High utility itemset mining
  + Parallel high utility itemset mining
* Selección del algoritmo a hacer paralelizable.
* Generación de los conjuntos de datos para las pruebas.
* Pruebas con los conjuntos de datos generados.
* Modelado del algoritmo PMTKU
* Implementación y depuración del algoritmo PMTKU
* Comparación con el algoritmo EFIM
* Redacción del documento científico
  1. **RESULTADOS ESPERADOS DEL PROYECTO**

Se espera que el algoritmo PMTKU propuesto para descubrir high utility itemsets en grandes conjuntos de datos mejore en términos de tiempo al algoritmo TKU y permita la comparación con otros algoritmos que sigan el mismo patrón de entrada.

1. **PLANEAMIENTO DEL PROYECTO**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Dic | Ene | Feb | Mar | Abr | May | Jun | Jul | Ago | Set | Oct | Nov |
| Investigación del Estado del Arte |  |  |  |  |  |  |  |  |  |  |  |  |
| Diseño del modelo propuesta. |  |  |  |  |  |  |  |  |  |  |  |  |
| Implementación. |  |  |  |  |  |  |  |  |  |  |  |  |
| Pruebas y Ajustes |  |  |  |  |  |  |  |  |  |  |  |  |
| Escritura del Libro |  |  |  |  |  |  |  |  |  |  |  |  |
| Evaluación de Resultados y elaboración de las conclusiones |  |  |  |  |  |  |  |  |  |  |  |  |
| Defensa del trabajo |  |  |  |  |  |  |  |  |  |  |  |  |

1. **TEMARIO TENTATIVO DEL REPORTE TÉCNICO FINAL**
2. Introducción
3. Descripción del problema
4. High utility itemset mining
5. Algoritmos paralelos
6. Parallel high utility itemset mining
7. Algoritmo propuesto basado en el TKU
8. Resultados Experimentales
9. Conclusiones y trabajos futuros
10. Referencias

**REFERENCIAS BIBLIOGRÁFICAS**

**[1] Anderson, David R., Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm y Kipp Martin. Métodos cuantitativos para los negocios, 11a ed.**

**[2] Narasimhan, S., McLeavey, D.W., Billington, P. Planeación de la producción y control de inventarios. 1996. Prentice-Hall Hispanoamericana**

**[3] BUSINESS INTELLIGENCE: STATE OF THE ART. Alveiro Rosado, Dewar Rico. 2010 https://dialnet.unirioja.es/descarga/articulo/4564348.pdf**

**[4] Josep Luis Cano. Business Intelligence: Competir con información. ESADE, Banesto, Banesto Pyme, 2007.**

**[5] “The datawarehouse Toolkit.” Ralph Kimball. Wiley, 1996**

[1] An introduction to frequent pattern mining. <http://data-mining.philippe-fournier-viger.com/introduction-frequent-pattern-mining/>.

[2] An introduction to High-Utility Itemset Mining. <http://data-mining.philippe-fournier-viger.com/introduction-high-utility-itemset-mining/>.

[3] Cheng Wei Wu , Bai-En Shie , Philip S. Yu , Vincent S. Tseng Mining Top-K High Utility Itemsets .

[4] Quang-Huy Duonga,, Bo Liaoa, Philippe Fournier-Viger b, Thu-Lan Dama, An efficient algorithm for mining the top-k high utility itemsets, using novel threshold raising and pruning strategies

[5] Jerry Chun-Wei Lin a, Wensheng Gan a , Philippe Fournier-Viger b , Tzung-Pei Hong c,d, Vincent S. Tseng Fast algorithms for mining high-utility itemsets with various discount strategies

[6] Souleymane Zida, Philippe Fournier-Viger, Jerry Chun-Wei Lin,Cheng-Wei Wu, Vincent S. Tseng, EFIM: A Highly Efficient Algorithm for High-Utility Itemset Mining