Capítulo Sexto

Conclusiones

6. Conclusiones

6.1. Conclusiones

1. El algoritmo ActiveRank es un método efectivo y de alta eficiencia en la clasificación automática de contenido dentro de una red de información conformada por múltiples subconjuntos o categorías siempre y cuando se cumplan las siguientes condiciones:
    a) La información por clasificar pertenezca a un conjunto antagónico al resto del universo, y sea posible determinar su pertenencia a través de operaciones binarias (pertenece o no pertenece). Dicho antagonismo puede presentarse de diferentes formas, siendo las más comunes por diferencia de idioma y temática.
    b) Cuando el vector o los vectores de referencia han sido correctamente generados y son realmente representativos del conjunto que se quiere discriminar.
    c) Cuando la dimensión de los vectores ActiveRank es óptima y la pérdida de información por nodos no significativos no altera de manera importante el comportamiento del sistema y en su caso la matriz de rankings ActiveRank.
    d) Cuando los sistemas de filtrado en el proceso de crawling funcionen correctamente retirando información inválida, inconsistente o con errores, y permitiendo el paso de aquella que resulta fundamental para el correcto perfilamiento de un documento dado.
    e) Cuando la información disponible en red pueda ser parseada correctamente y se excluyan errores de codificación, comunicación, interpretación o formato.
2. El proceso de clasificación a través de ActiveRank es una característica adicional en su utilización como motor de relación de información y administración en sistemas de análisis de redes de información, lo que permite disminuir los costos operativos en dichos sistemas al utilizar el mismo núcleo tecnológico para múltiples propósitos.
3. El algoritmo ActiveRank como método automático de clasificación de información es económicamente rentable debido a que no requiere incrementar la infraestructura necesaria para desempeñar dicha tarea, aumentando de manera linealmente su consumo de poder de cómputo con respecto el aumento de nodos en la red, lo cual a comparación del aumento cuadrático en el procesamiento de información cuando se utiliza como motor de relación es significativamente menor.

4. La eficiencia del algoritmo ActiveRank como método automático de clasificación puede disminuir dramáticamente debido a:
    a) Errores en el sistema de crawling que inserten vectores ActiveRank de dimensiones menores al resto de los elementos; lo anterior produce un error sistemático en ActiveRank debido a que dichos elementos tienen mayor probabilidad de obtener una medida de ranking de mayor cercanía por el hecho de que con menos nodos coincidentes se incrementa dicha magnitud, y que además podrían ser no representativos.
    b) El aumento del número de conjuntos que conforman el universo de información analizada, y la reducción de los intervalos entre los valores de ranking medios de cada conjunto, lo que trae como consecuencia directa una reducción en los posibles valores del ranking del umbral de clasificación, que en conjunto con la dispersión de los rankings dentro de los diferentes conjuntos reducen significativamente la capacidad de discriminar la pertenencia de un documento dentro de un conjunto determinado.
    c) Otros errores no predecibles en la operación de los sistemas y subsistemas involucrados en el proceso de indexación, procesamiento y la plataforma DARE, de esto la importancia de mantener un constante y estricto monitoreo de su funcionamiento.
5. En el análisis desarrollado para los 70,000 documentos pertenecientes o accesibles a través de vínculos dentro del dominio *unam.mx* no se encontró contenido pornográfico, utilizando vectores de referencia en inglés y en español, además de un análisis manual de la base de datos recolectada.

6.2. Contribuciones

El desarrollo de la presente tesis permitió encontrar nuevas aplicaciones de la tecnología desarrollada por la empresa Ondore, en particular del algoritmo ActiveRank, así como detectar y plantear mejoras a cada uno de los sistemas utilizados en el proceso. Lo anterior ayudará a continuar las diferentes líneas comerciales y de investigación, que a su vez impactará positivamente al generar nuevos empleos en nuestro país y seguir desarrollando oportunidades para los profesionales y la economía de México.

En términos académicos, la presente tesis abarca un gran número de temas relevantes en la investigación y desarrollo de sistemas de análisis de información, y podrá ser utilizada como referencia para futuros trabajos y proyectos al interior de la Universidad Nacional Autónoma de México.

6.3. Trabajo futuro

Durante el desarrollo de esta tesis se encontraron diversos puntos que podrían ser de interés futuro tanto en la explotación de la tecnología utilizada, como en la comprensión de algunos fenómenos poco convencionales que podrían no ser triviales y derivar en nuevas líneas de investigación del algoritmo ActiveRank; a continuación se describen en términos generales para sentar base de trabajos futuros.

- *Velocidad y probabilidad de encontrar una página pornográfica en subgráficas de la WWW*

A través de la utilización de la tecnología de análisis de información de Ondore en una implementación muy similar a la desarrollada en esta tesis, es posible detectar el momento y ubicación de una página pornográfica al iniciar una exploración de la WWW desde un punto aleatoriamente seleccionado; repitiendo este experimento un número de veces suficientemente grande sería posible encontrar la velocidad, distancia y probabilidad promedio de llegar a una página web pornográfica al iniciar desde cualquier punto de la WWW. Como primera aproximación, se deberían encontrar valores altos para la velocidad y bajos para la distancia, es decir, que partiendo de cualquier página de la WWW se puede llegar en pocos saltos a una página pornográfica. Es posible seleccionar los puntos de inicio de análisis dentro de subgráficas específicas como podrían ser páginas de universidades, de gobierno, personales, redes sociales, etc., con el fin de obtener un mejor rango de valores para los diferentes tipos de escenarios.

- *Efectos estadísticos de la normalización de un vector*

Durante las primeras etapas del desarrollo del algoritmo ActiveRank se estudió el comportamiento de cada una de sus operaciones en vectores generados de manera aleatoria. Si se genera un vector de dos dimensiones cuyas componentes han sido seleccionadas aleatoriamente del intervalo [0,1] con una distribución uniforme (todos los valores entre 0 y 1 tienen la misma probabilidad de ser obtenidos), este puede ser representado como un punto en el plano unitario del espacio coordenado de 2 dimensiones, al incrementar el número de vectores todos quedarían repartidos uniformemente en dicho plano; análogamente, si habláramos de vectores de 3 dimensiones generados con el mismo procedimiento, estos quedarían uniformemente distribuidos en el cubo unitario del espacio coordenado de 3 dimensiones.

Al normalizar los vectores la condición que se hace cumplir es que la suma de sus componentes sea igual a 1, lo que genera una dependencia o relación limitativa entre las variables, modificando la función de

distribución de cada una de ellas, que a su vez condensa los puntos en una estructura de 1 dimensión menor a la que originalmente tenían. Lo anterior es fácilmente apreciable en las figuras a continuación mostradas, donde para 2 variables, los puntos uniformemente distribuidos en el plano y cubo unitario quedan restringidos a un segmento de línea y plano respectivamente. [1]

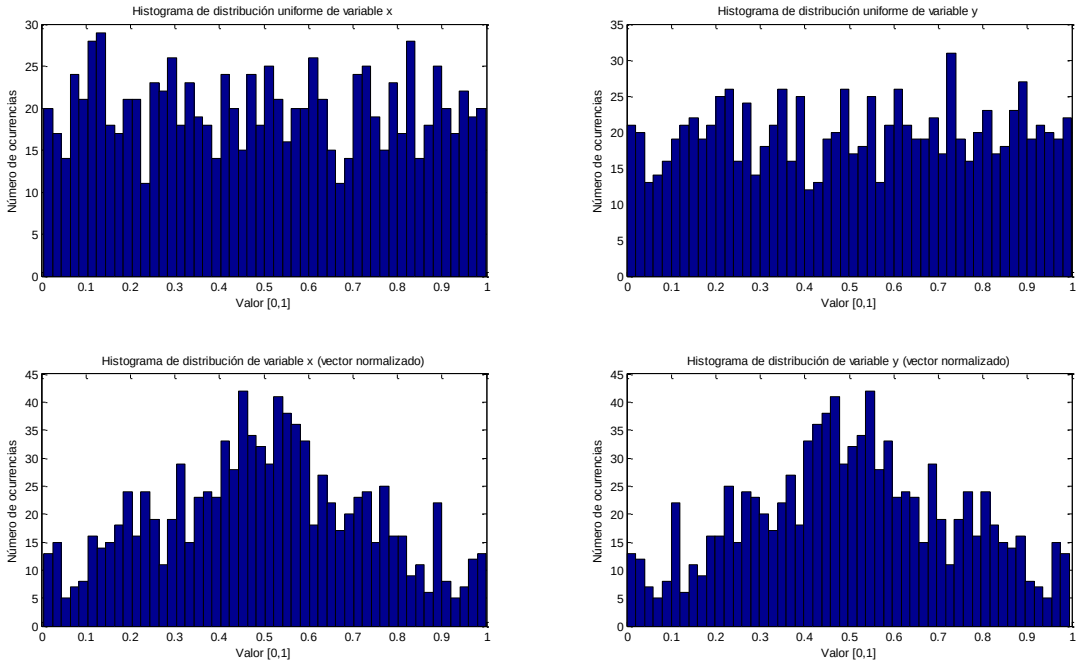

Figura 6.3.1. – Histograma de valores obtenidos por las 2 variables de 1000 vectores de 2 dimensiones generados aleatoriamente antes y después de la normalización.

5. A partir de mi apreciación y sin haber desarrollado ninguna comprobación, la función de distribución parece cambiar de uniforme a distribución chi, la cual es una generalización de la distribución de Rayleigh para *n* variables. Lo anterior se basa en que para el escenario de 2 dimensiones, la nueva distribución parece ser normal, sin embargo, al incrementar el número de dimensiones, la campana se desplaza hacia los valores inferiores, lo que resulta lógico dado que es más probable que todas las variables tengan valores chicos, a que todas tengan valores grandes debido a la relación que produce la condición de normalización.

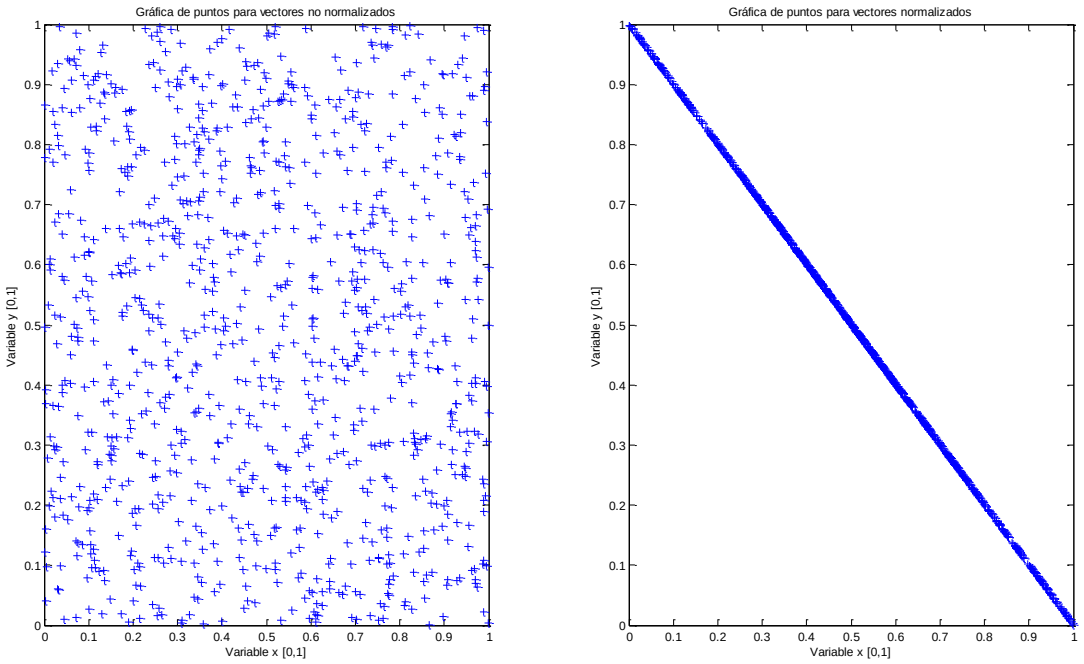

Figura 6.3.2. – Distribución de 1000 vectores de 2 dimensiones generados aleatoriamente antes y después de la normalización.

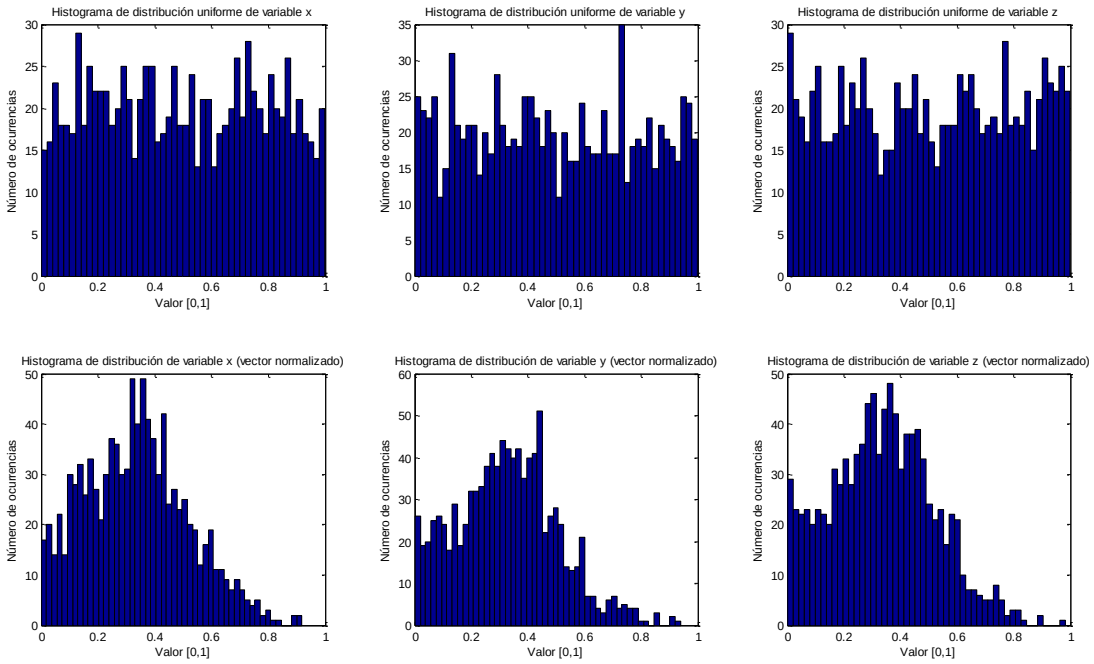

Figura 6.3.3. – Histograma de valores obtenidos por las 3 variables de 1000 vectores de 3 dimensiones generados aleatoriamente antes y después de la normalización; se puede apreciar el desplazamiento lateral de la curva en la distribución de valores después de la normalización.

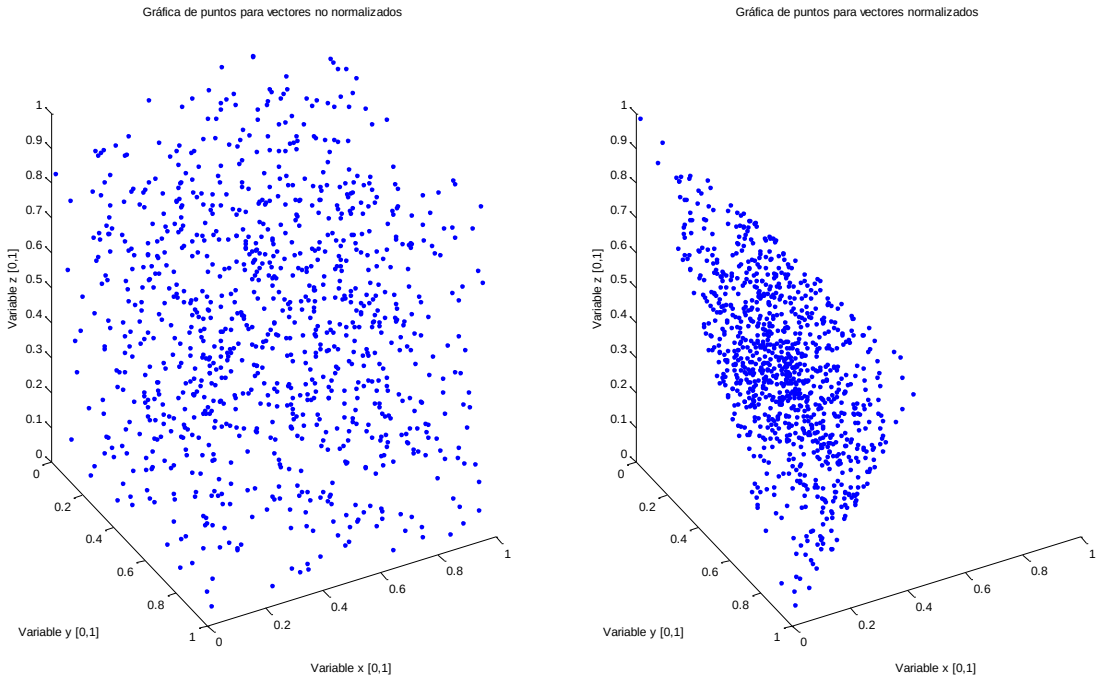

Figura 6.3.4. – Distribución de 1000 vectores de 3 dimensiones generados
aleatoriamente antes y después de la normalización; se puede apreciar como los puntos quedan limitados
a un plano triangular concentrando la densidad en la sección donde las 3 variables tienen un valor aproximado de 0.4

El tema central de investigación no es en realidad el por qué se modifica la función de probabilidad de cada una de las componentes, dado que se conoce que se trata de un efecto de correlación causado por la operación de normalización, sino el impacto que puede tener en la pérdida de información estructural que se da al colocar un punto de *n* dimensiones en una estructura de *n-1* dimensiones.

- *Patrones lineales en la matriz de rankings de ActiveRank*

Una observación importante realizada durante el desarrollo del algoritmo ActiveRank fue la presencia de patrones lineales al representar en una escala de 256 grises la matriz de rankings (ver figura 6.3.5.) de un universo de vectores generados de manera aleatoria contra sí mismo. La interpretación de los patrones lineales observados en la figura antes mencionada es que existe una relación de similitud homogénea entre un vector dado y todos los demás, lo que se opone en gran medida al hecho de que cada uno de los vectores fue generado de manera completamente independiente por lo que la relación entre ellos también debería encontrarse

uniformemente distribuida en intervalo de valores entre 0 y 1.

Si se observa con detenimiento la imagen, y su contraparte numérica, también se obtienen pocos valores extremos, es decir, la mayor concentración de calificaciones se da para los valores cercanos a 0.5, por esto es que se observa un tono homogéneo de gris en la mayor parte de la figura; la diagonal principal blanca se debe a que el ranking de un vector contra sí mismo es en todos los casos igual a 1.

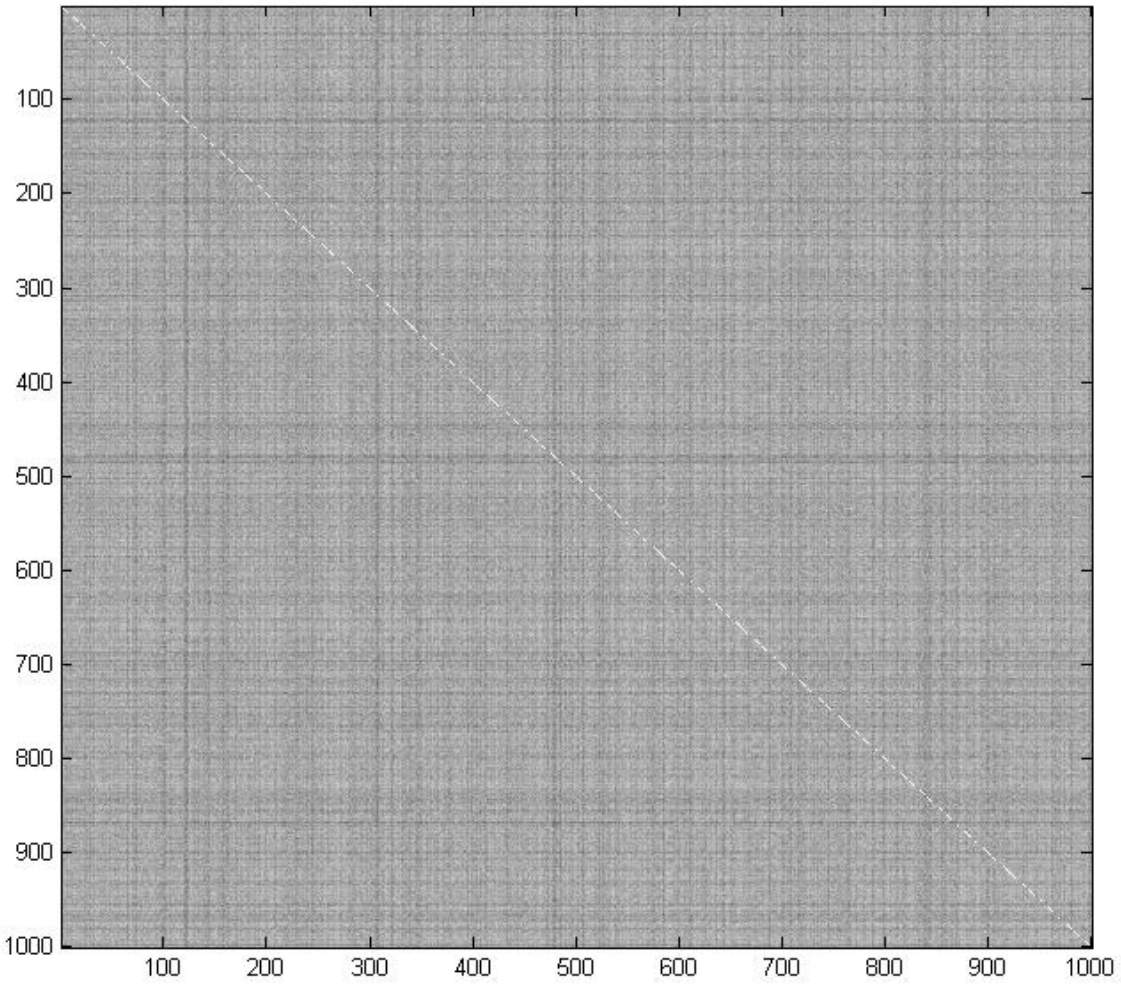

Figura 6.3.5. – Patrones lineales detectados al representar la matriz de rankings de un universo de vectores generados aleatoriamente contra sí mismo en una escala de 256 tonos de gris.

El trabajo futuro en este aspecto es, al igual que el punto anterior, analizar con detalle el efecto de cada una de las operaciones que se llevan a cabo en el algoritmo ActiveRank y determinar si la afectación en la distribución estadística de los valores que produce la normalización es causante de este tipo de correlaciones en

la matriz de rankings, o si son eventos independientes que deben ser analizados separadamente.

- *Implementación de algoritmos de aceleración y mejoras en general en los sistemas de análisis de información de Ondore*

Por último, uno de los principales puntos de desarrollo futuro encontrados tras la realización de esta tesis fue la amplia gama de mejoras en el desempeño de los sistemas de análisis de información que pueden ser implementadas, no porque actualmente sean considerados ineficientes, sino por la gran cantidad de oportunidades de mejora existentes, entre las que destacan:

- Sistemas de ruteo de peticiones web entrantes/salientes y servidores proxy que ayuden a reducir el tiempo de conexión entre la aplicación de análisis y las páginas deseadas, permitiendo la incorporación de múltiples servidores de análisis coordinando la exploración.
- Plantear un modelo de cómputo paralelo análogo al utilizado por el sistema DARE (Distributed ActiveRank Engine) en los sistemas Analyzer, de tal manera que fuera posible el mantener mayores ritmos de procesamiento a un costo técnica y económicamente accesibles.
- Trabajar en la simplificación de consultas a bases de datos, así como la revisión de la correcta utilización de índices en las tablas y consideración de algoritmos de aceleración de consultas como podría ser BWA (*Business Warehouse Accelerator*); desincorporar los servidores de base de datos de los servidores donde están instalados los sistemas Analyzer para aumentar la escalabilidad de los procesos al poder utilizar servidores NAS (*Network Attached Storage*) con plataformas de bases de datos distribuidas.
- Seguir analizando el potencial del algoritmo ActiveRank al ser utilizado en tiempo real y desarrollar métodos simplificados para su aceleración. En el ámbito del procesamiento en paralelo del sistema DARE, es posible continuar extensivamente la investigación de su viabilidad operando en infraestructura de arquitectura mixta (tanto en software como en hardware), utilizando plataformas convencionales para cómputo distribuido, entre otras.

- Expandir la funcionalidad y usabilidad de las interfases gráficas actualmente disponibles para mejorar su aprovechamiento, simplificando la navegación y la obtención de resultados, así como la evaluación de los diferentes indicadores estadísticos sobre el desempeño y eficiencia de los sistemas; la creación de *backpanels* para la administración, configuración y depuración de las plataformas de análisis de información pueden reducir significativamente el costo que tienen estas actividades al disminuir el tiempo que se invierte en estos procesos.