MINERÍA WEB

MARIEL ALEJANDRA CARO
LICENCIATURA EN SISTEMAS DE INFORMACIÓN - FCEYT - UNSE

Tabla de contenido

Introducción 2

¿Qué es la Minería Web? 2

TAXONOMÍA DE LA MINERÍA WEB 3

Minería Web de Estructura 5

Análisis del artículo: "*Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval.*" 5

Objetivo: 5

Contexto de la investigación 5

Metodología aplicada: 5

Tareas 5

Métodos realizados: 5

Resultados obtenidos: 5

Métricas de evaluación: 6

Conclusiones obtenidas: 6

Minería Web de Contenido 7

Análisis de Artículo: *"Web Content Mining Techniques: A Survey."* 7

Objetivo 7

Contexto de la investigación 7

Metodología aplicada: 7

Tareas 7

Métodos realizados: 7

Métricas de evaluación: 9

Conclusiones obtenidas 10

Minería Web de Uso 11

Análisis de Artículo: *"Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data"* 11

Objetivo 11

Contexto de la investigación 11

Metodología aplicada 11

Tareas 11

Métodos realizados: 11

Métricas de evaluación: 12

Conclusiones obtenidas 12

Conclusión: 14
Bibliografía: 14

Introducción

La World Wide Web -o la Web para abreviar- ha impactado casi en todos los aspectos de nuestras vidas. Es la fuente de información más grande y conocida, de fácil y rápido acceso y búsqueda. Consiste en miles de millones de documentos interconectados -páginas web- que son creados por millones de personas alrededor de todo el mundo.

Desde su inicio, la Web ha cambiado drásticamente nuestro comportamiento de búsqueda de información, de educación, de compra, de entretenimiento y de trabajo. Antes de la Web, encontrar información significaba pedir prestado o comprar un libro para leer, recurrir a bibliotecas, y representaba una pérdida significativa de esfuerzo y tiempo.

Sin embargo, con la llegada de Internet, y posteriormente de la Web, todo está a solo unos clics de la comodidad de nuestros hogares u oficinas. La Web también proporciona un medio conveniente para que nos comuniquemos entre nosotros, expresemos nuestros puntos de vista y opiniones y discutamos con personas de cualquier parte del mundo. La web es realmente una sociedad virtual.

Este breve trabajo de investigación pretende explicar, a partir de recursos bibliográficos validados los diferentes tipos de minería web, sus aplicaciones y sus principales técnicas. Mediante un análisis de tres artículos seleccionados, se intentarán abarcar todos estos temas.

¿Qué es la Minería Web?

La minería web trata de descubrir información útil o conocimiento desde la estructura de hipervínculos de la web, el contenido de las páginas y los datos de uso. A pesar de que la minería web utiliza muchas técnicas de la minería de datos, no es una aplicación de las técnicas de minería de datos tradicional, debido a la naturaleza heterogénea, semi-estructurada y no estructurada de los datos de la web. Muchas nuevas tareas de minería y algoritmos han sido inventadas en la década pasada. En función de los tipos principales de datos utilizados en el proceso de minería, las tareas de minería web se pueden clasificar en tres tipos: **minería web de estructura, minería web de contenido y minería web de uso.** (Liu, 2011)

TAXONOMÍA DE LA MINERÍA WEB

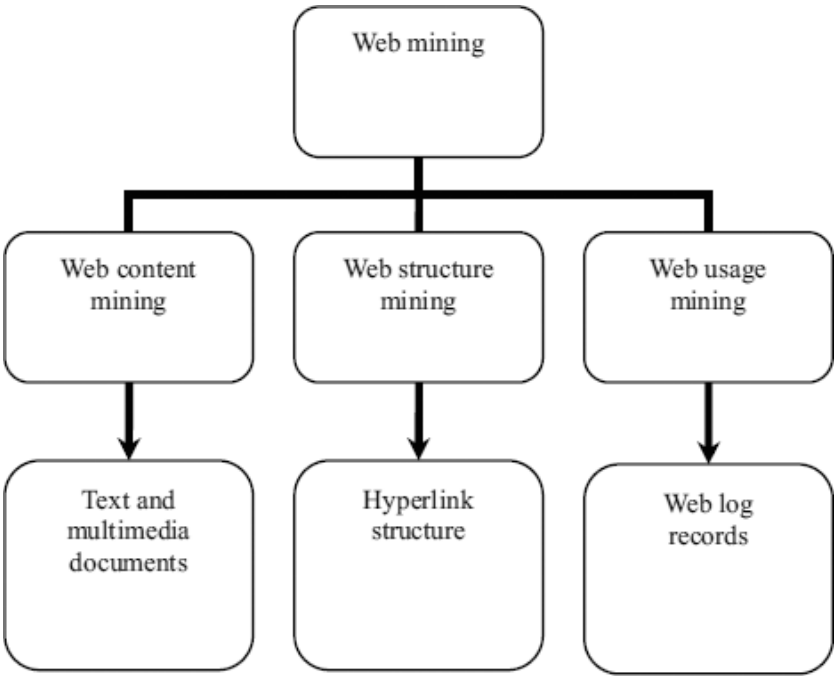

Fig 1.1- Tipos de Minería web y objetivos de uso.

- **La minería web de estructura**, descubre conocimientos útiles a partir de hipervínculos (o enlaces para abreviar), que representan la estructura de la Web. **Por ejemplo**, a partir de los enlaces, podemos descubrir páginas web importantes, que es una tecnología clave utilizada en los motores de búsqueda. También podemos descubrir comunidades de usuarios que comparten intereses comunes. La minería de datos tradicional no realiza tales tareas porque generalmente no hay una estructura de enlace en una tabla relacional.
- **La minería web de contenido,** extrae o investiga información o conocimiento útil del contenido de la página web. **Por ejemplo**, podemos clasificar y agrupar automáticamente las páginas web de acuerdo con sus temas. Estas tareas son similares a las de la minería de datos tradicional. Sin embargo, también podemos descubrir patrones en páginas web para extraer datos útiles, como descripciones de productos, publicaciones de foros, etc., para muchos propósitos. Además, podemos extraer opiniones de clientes y publicaciones en foros para descubrir las opiniones de los consumidores. Estas no son tareas tradicionales de minería de datos.

- **La minería web de uso,** se refiere al descubrimiento de patrones de acceso de usuarios a partir de registros de uso web, que registran cada clic realizado por cada usuario. La minería de uso web aplica muchos algoritmos de minería de datos. Una de las cuestiones clave en la minería de uso web es el preprocesamiento de datos de flujo de clics en los registros de uso para producir los datos correctos para la minería. **Por ejemplo,** el descubrimiento y análisis automático de patrones en flujos de clics, transacciones de usuarios y otros datos asociados recopilados o generados como resultado de las interacciones de los usuarios con los recursos web en uno o más sitios web.

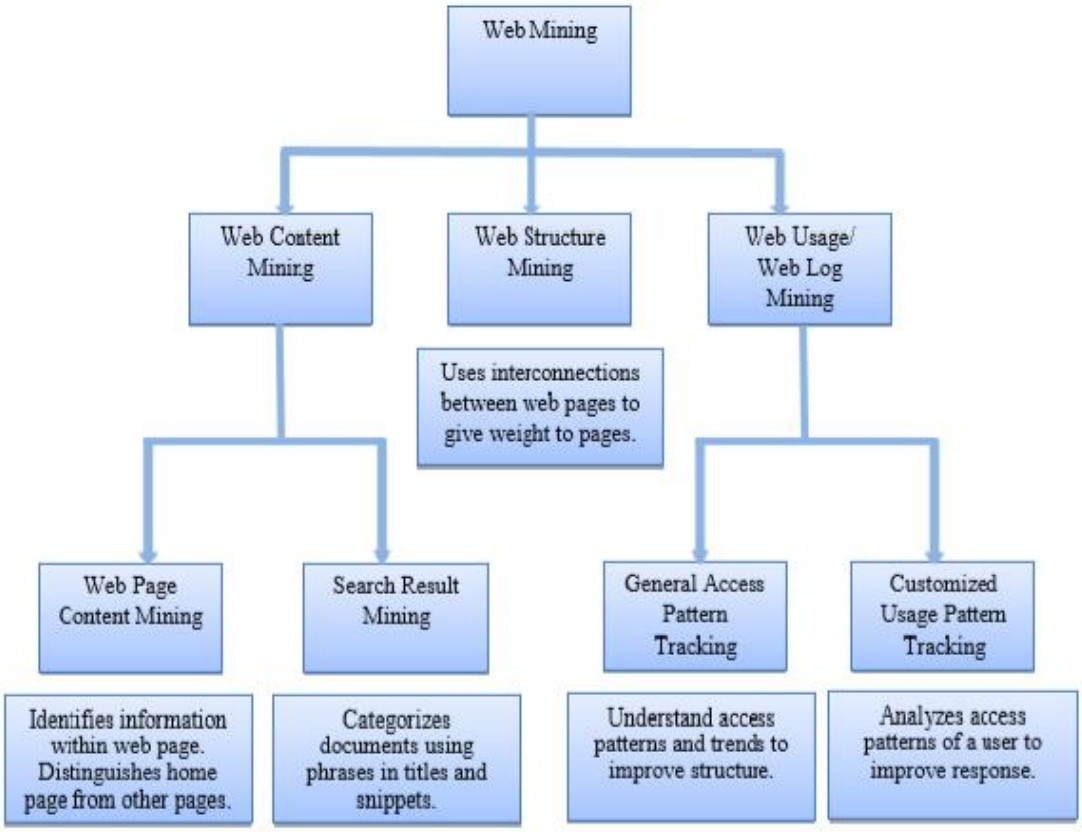

Fig 1.2- Taxonomía detallada de la minería web.

El proceso de minería web es similar al proceso de minería de datos. La diferencia suele estar en la recopilación de datos. En la minería de datos tradicional, los datos a menudo ya se recopilan y almacenan en un almacén de datos. Para la minería web, la recopilación de datos puede ser una tarea sustancial, especialmente para la estructura web y la minería de contenido, que implica rastrear (crawling) una gran cantidad de páginas web de destino. Dedicaremos todo un capítulo al rastreo.

Una vez que se recopilan los datos, pasamos por el mismo proceso de tres pasos: preprocesamiento de datos, minería de datos web y post-procesamiento. Sin embargo, las técnicas utilizadas para cada paso pueden ser bastante diferentes de las utilizadas en la minería de datos tradicional.

Minería Web de Estructura

Análisis del artículo: *"Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval."* (Ravi Kumar P;Ashutosh Kumar Singh, 2009)

Objetivo: Este artículo se centra en el análisis de hipervínculos, los algoritmos utilizados para el análisis de enlaces (PageRank y Hits), y la comparación entre ellos. Se explora el concepto de Authorities y Hubs. Los diferentes algoritmos utilizados para Link análisis como PageRank, HITS.

Contexto de la investigación: La investigación se realizó en Miri, Malasia. En el contexto en el que la WWW es hoy en día la mayor fuente de información, y cubre casi todas las necesidades de información de los usuarios. Pero esta información se presenta de manera heterogénea.

Metodología aplicada: Web Mining, metodología de recuperación de la información

Tareas: analizar la estructura de las páginas web, realizar un análisis profundo de los hipervínculos de un documento web, para descubrir un modelo subyacente en las estructuras de vínculos de las páginas web, catalogarlas y generar información.

Métodos realizados:

- **Grafo Etiquetado Web,** donde los nodos son documentos o páginas y las aristas son los hipervínculos entre ellos y sirve para analizar los vínculos entre sitios web.
- **Análisis de hipervínculos** mediante:
  - **PageRank.**
    Algoritmo basado en el análisis de citas. Es utilizado por Google. Provee maneras avanzadas para computar la importancia o relevancia de una página web que simplemente cuenta el número de páginas que están hipervinculadas a ésta (también llamadas backlinks). Puede ser calculado usando un simple algoritmo iterativo y se corresponde con un vector normalizado de una matriz.
  - **Algoritmo HITS – Hubs and Authorities.**
    Basado en identificar dos tipos de páginas web: Authorities son páginas que tienen contenido importante; y Hubs, que son páginas que actúan como listas de recursos guiando a los usuarios a los sitios Authorities. Este algoritmo directamente trata la WWW como un grafo dirigido donde los nodos o vértices son páginas y las aristas los vínculos entre ellas.

Resultados obtenidos: Se obtiene una comparación entre los dos algoritmos mencionados: PageRank y HITS.

	PageRank	HITS
Técnica usada de minería	Minería Web de estructura	Minería web de estructura y de contenido
Modo de trabajo	Calcula puntajes en el tiempo índice. Los resultados se ordenan según la importancia de las páginas.	Calcula puntajes de n páginas altamente relevantes sobre la marcha.
Parámetros	Links hacia atrás	Links hacia atrás, hacia adelante y de contenido

Complejidad	O(log N)	<O(log N)
Limitaciones	Independencia de consultas	Tópicos derivados y problema de eficiencia
Motores de búsqueda que lo utilizan	Google	Clever

Métricas de evaluación: Complejidad y limitaciones.

Conclusiones obtenidas: El artículo cubre los conceptos básicos de Web mining y la importancia de la minería web de estructura en la recuperación de la información. Se focaliza en os algoritmos para el análisis de hipervínculos.

Minería Web de Contenido

Análisis de Artículo: *"Web Content Mining Techniques: A Survey."* (Faustina Johnson;Santosh Kumar Gupta, 2012)

Objetivo: hacer un estudio de diferentes técnicas y patrones de minería web de contenido y las áreas en las que ha influido. También señala cómo utilizada la minería web de contenido en la minería web de uso.

Contexto de la investigación: La investigación se realizó para la International Journal of Computer Applications. Atendiendo al contexto en el que la complejidad de la Web aumenta debido a la enorme cantidad de datos. Entonces la extracción de datos de acuerdo con las necesidades de los usuarios se vuelve tediosa. Como resultado, la minería se convirtió en una técnica esencial para extraer información valiosa de internet. Y esta técnica fue nombrado como minería web.

Metodología aplicada: Web Mining, metodología de recuperación de la información

Tareas: descubrir información útil de los contenidos de la web tales como texto, imágenes, videos, etc. Examinar contenido de la web como los resultados de la búsqueda. Utilizar conceptos y técnicas de minería de datos y de texto para extraer el contenido útil.

Métodos realizados:

Técnicas de minería de datos no estructurados.

Se emplean para datos no estructurados como los textos, devuelve información desconocida, que se extrae desde diferentes textos, requiere la aplicación de técnicas de minería de datos y de texto. Las técnicas usadas en minería de texto son:

- **Extracción de Información:** se utiliza pattern matching. Traza la palabra clave y las frases y luego descubre la conexión de las palabras clave dentro del texto.
  La extracción de información se puede proporcionar al módulo KDD porque la extracción de información tiene que transformar el texto no estructurado en datos más estructurados. Es la base de muchas otras técnicas para datos no estructurados.
- **Topic Tracking**: Técnica que verifica los documentos vistos por el usuario y estudia los perfiles de usuarios. De acuerdo con cada usuario, predice otros documentos relacionados con los intereses de estos. Las desventajas son que cuando buscamos por tópicos, podemos obtener información no relacionada con nuestros intereses.
- **Resumen (Summarization)**: utilizada para reducir el tamaño de un documento, manteniendo los puntos principales. Ayuda al usuario a decidir cuándo debería leer ese tópico o no. El tiempo que toma es menos que aquel que emplea el usuario para leer el primer párrafo. Inclusive, otorga la libertad de que el usuario elija el porcentaje total del texto que desea leer como resumen.
- **Categorización:** es la técnica para identificar temas principales poniendo los documentos en un conjunto de grupos predefinidos. Permite contar el número de palabras de un documento.
- **Clustering:** técnica utilizada sobre un grupo de documentos similares. Los grupos no se basan en tópicos predefinidos. El agrupamiento se realiza sobre la marcha. Mismos

documentos pueden aparecer en diferentes grupos. Ayuda a los usuarios a seleccionar rápida y fácilmente los tópicos de su interés.

- **Visualización de Información:** Utiliza la extracción de características y la indexación de términos clave para construir una representación gráfica. A través de la visualización, se descubren documentos que tienen similitudes. Los materiales de texto grandes se representan como jerarquía visual o mapas donde se permite la facilidad de navegación.

Técnicas de minería de datos estructurados.

- **Web Crawler:** se divide en dos tipos: rastreador (crawler) interno o externo. Los rastreadores son programas de computadora que atraviesan la estructura de hipertexto en la web. El rastreador externo se arrastra a través de un sitio web desconocido. El rastreador interno se arrastra a través de páginas internas del sitio web que son devueltas por un rastreador externo.
- **Wrapper Generation:** Proporciona información sobre la capacidad de las fuentes. Las páginas web ya están clasificadas por motores de búsqueda tradicionales. De acuerdo con la consulta, las páginas web se recuperan utilizando el valor de rango de página. Las fuentes son qué consulta responderán y los tipos de salida. Los wrappers también proporcionarán una variedad de metainformación. P.ej. Dominios, estadísticas, índice de búsqueda sobre las fuentes.
- **Page Content Mining:** Es una técnica de extracción que trabaja con la clasificación de páginas hechas por los motores de búsqueda tradicionales. Comparando estas clasificaciones de contenido se clasifican las páginas.

Técnicas de minería de datos semi- estructurados.

- **Modelo de intercambio de objetos (Object Exchange Model -OEM):** La información relevante se extrae de datos semiestructurados embebidos en grupos de información útil y almacenados en modelos de intercambio de objetos (OEM). Ayuda al usuario a entender la estructura de la información en la web más apropiada. Es el más adecuado para entornos heterogéneos y dinámicos. Una característica principal del modelo de intercambio de objetos es la autodescripción, no es necesario describir de antemano la estructura de un objeto.
- **Extracción Top-down**: se extraen objetos complejos desde un conjunto de fuentes web ricas y los convierte en objetos menos complejos hasta obtener unidades atómicas.
- **Lenguaje de extracción de datos web:** convierte los datos web en datos estructurados y los entrega a los usuarios finales. Almacena datos en forma de tablas.

Técnicas de minería de datos multimedia:

- **SKICAT**: es un analizador de datos astronómicos exitoso y sistema de catálogo que produce catálogos digitales de los cuerpos celestes. Utiliza técnicas de machine learning para convertir estos objetos en clases usables por humanos. Integra técnicas de procesamiento de imágenes y clasificación de datos que ayuda a categorizar un gran conjunto de datos.
- **Color Histogram Matching**: consiste en la ecualización de histograma de color y suavizado. La ecualización intenta encontrar una correlación entre componentes de color.
- **Miner Multimedia:** consta de cuatro pasos principales:

1. -Un excavador de imágenes para la extracción de imágenes y videos.
2. -Un preprocesador para la extracción de características de imágenes y las almacena en una base de datos.
3. -Un núcleo de búsqueda que utiliza consultas de matching con imágenes y videos disponibles en la base de datos.
4. -Un módulo de descubrimiento que realiza rutinas de minería de datos de imágenes para rastrear los patrones en imágenes.

- **Detección de límite de disparo:** técnica que detecta automáticamente los límites entre disparos en video.

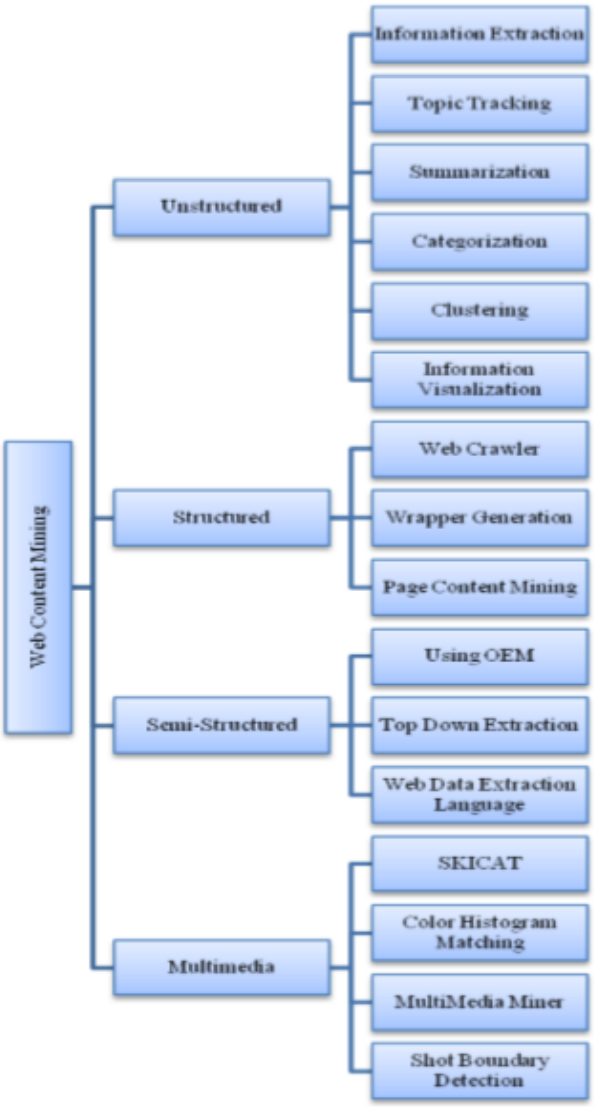

Fig 2: Web Content Mining Techniques

Métricas de evaluación: tiempo de grabado, complejidad de estructuras de datos, amigabilidad con el usuario.

Conclusiones obtenidas: El artículo discute sobre técnicas de minería web de contenido. Esta metodología es muy útil en el campo empresarial. La minería web de contenido resuelve el problema de extraer información de diferentes tipos de datos, encontrar y decidir cuál es más relevante. También, ayuda a establecer una mejor relación con el cliente al proporcionar exactamente lo que necesita.

Minería Web de Uso

Análisis de Artículo: *"Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data"* (Jaideep Srivastava; Robert Cooley; Mukund Deshpande; Pang-Ning Tan , 2000)

Objetivo: Este artículo se centra en describir las fases fe la minería web de uso: preprocesamiento, descubrimiento de patrones y análisis de patrones. Busca proporcionar una taxonomía en detalle de como trabaja el área de minería web de uso y mostrar los esfuerzos de investigación realizados en el campo.

Contexto de la investigación: La investigación se realizó en Minneapolis, Minnestoa. Es un paper que se realiza en un contexto en el que la minería tiene cada vez más interesados por los beneficios que aporta principalmente al e-commerce. La facilidad y rapidez con que se pueden realizar transacciones comerciales en la Web ha sido una fuerza impulsora clave en el rápido crecimiento del comercio electrónico. Específicamente, la actividad de comercio electrónico que involucra al usuario final está experimentando una revolución significativa. La capacidad de rastrear el comportamiento de navegación de los usuarios hasta los clics individuales del mouse ha acercado al vendedor y al cliente final más que nunca. Ahora es posible que un vendedor personalice su mensaje de producto para clientes individuales a gran escala, un fenómeno que se conoce como personalización masiva.

Metodología aplicada: Web Mining, metodología de recuperación de la información

Tareas: descubrir patrones de uso de las páginas web mediante algoritmos, así como también las direcciones IP, referencias de páginas y fecha y hora de accesos. Describir las tres fases de la minería web de uso y darle un vistazo al sistema WEBSIFT que es un prototipo de un sistema de minería web de uso.

Métodos realizados:

Se utilizan métodos similares a los de minería de datos, y el proceso de minería web de uso se divide en tres fases:

1. **Preprocesamiento:**

   1.1 Preprocesamiento de uso:

   El preprocesamiento de uso es posiblemente la tarea más difícil en el proceso de minería web de uso debido a la incompletitud de los datos disponibles. A menos que se utilice un mecanismo de seguimiento del lado del cliente, solo la dirección IP, el agente y la secuencia de clics del lado del servidor están disponibles para identificar las sesiones del servidor y de los usuarios. Si ya se pudiese identificar cada usuario, el flujo de clics debe dividirse en sesiones. Para identificar cuándo un usuario abandona un sitio web, se utiliza como parámetro treinta minutos de timeout de la conexión.

   1.2 Preprocesamiento de contenido:

   Consiste en convertir texto, imágenes, scripts y otros archivos en formas útiles para el procesamiento de minería web de uso.

A menudo, esto consiste en realizar minería de datos, como la clasificación o la agrupación. Si bien la aplicación de minería de datos sobre el contenido de los sitios web es un área de investigación interesante por derecho propio, en el contexto de la minería web de uso, el contenido de un sitio se puede utilizar para filtrar la entrada o salida de los algoritmos de descubrimiento de patrones. Por ejemplo, los resultados de un algoritmo de clasificación podrían usarse para limitar los patrones descubiertos a aquellos que contienen vistas de página sobre un determinado tema o clase de productos.

1.3 Preprocesamiento de estructura:

La estructura de un sitio se crea mediante los enlaces de hipertexto entre las vistas de página. La estructura se puede obtener y preprocesar de la misma manera que el contenido de un sitio. Nuevamente, el contenido dinámico (y, por lo tanto, los enlaces) plantean más problemas que las vistas de página estáticas. Es posible que deba construirse una estructura de sitio diferente para cada sesión de servidor.

2. Descubrimiento de patrones:

Para el descubrimiento de patrones se pueden utilizar diferentes técnicas como:

2.1 Análisis estadísticos

2.2 Reglas de asociación

2.3 Clustering (agrupamiento).

2.4 Clasificación

2.5 Patrones secuenciales

2.6 Modelado de dependencias.

3. Análisis de patrones:

Es el último paso del proceso de minería web de uso. La motivación detrás del análisis de patrones es filtrar reglas o patrones poco interesantes del conjunto que se encuentra en la fase de descubrimiento de patrones. La metodología de análisis exacta generalmente se rige por la aplicación para la que se realiza la minería web. La forma más común de análisis de patrones consiste en un mecanismo de consulta de conocimiento como SQL. Otro método es cargar datos de uso en un cubo de datos para realizar operaciones OLAP. Las técnicas de visualización, como los patrones gráficos o la asignación de colores a diferentes valores, a menudo pueden resaltar patrones o tendencias generales en los datos. La información de contenido y estructura se puede usar para filtrar patrones que contienen páginas de cierto tipo de uso, tipo de contenido o páginas que coinciden con una determinada estructura de hipervínculo.

Métricas de evaluación: Nivel de personalización obtenido, capacidad de mejora del sistema, caracterización de uso.

Conclusiones obtenidas: El artículo intenta proveer actualizaciones sobre el área de minería web de uso. Con el crecimiento de las aplicaciones basadas en la web, en especial el comercio electrónico, hay mucho interés en analizar los datos de uso de las páginas web y aplicar los

resultados para mejorar el servicio ofrecido a los usuarios. Explica el proceso de minería web de uso y lo esencial de cada etapa.

Conclusión:

En este trabajo se pudo abarcar ampliamente los tres tipos principales de minería web: de contenido, de estructura y de uso. Gran parte de estas ramas de la minería web utilizan herramientas y técnicas que vienen desde la minería de datos y la minería de texto.

A su vez, se pudo observar cómo para cada objetivo puede aplicarse los diferentes tipos de minería web o una combinación de los tres. Según lo que se desee lograr, cada variante de la minería web tiene sus herramientas y técnicas. Cada sitio web también tiene su propia estructura, su propio contenido, y en base a eso puede utilizar alguno de los métodos mencionados durante todo el trabajo para proporcionar un mejor servicio los usuarios: mayor rapidez, mayor precisión de lo que buscan, una vista personalizada y demás.

Sin duda alguna, la minería web es un conocimiento importante para la nueva era, y que se actualiza a pasos agigantados. Su aplicación en sitios web ha crecido en los últimos tiempos y es una herramienta utilizada principalmente por los grandes navegadores y buscadores web.

Bibliografía:

- Faustina Johnson;Santosh Kumar Gupta. (11 de Junio de 2012). Web Content Mining Techniques: A Survey. *International Journal of Computer Applications (0975 – 888)*, pág. 7.
- Jaideep Srivastava; Robert Cooley; Mukund Deshpande; Pang-Ning Tan . (2000). *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data.* Mineápolis, Minnesota, EEUU: SIGKDD Explorations.
- Liu, B. (2011). *Web Data Mining - Exploring Hyperlinks, Contents, and Usage Data - 2nd Edition.* Berlin, Alemania: Springer.
- M.G. da Costa, Z. G. (2005). *Web structure mining: an introduction.* Hong Kong, China, China: IEEE.
- Ravi Kumar P;Ashutosh Kumar Singh. (2009). Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval. *Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval* (pág. 5). Sarawak Campus, Miri, Malaysia: 2nd CUTSE International Conference 2009.