



Universidad Autónoma de Querétaro
 Facultad de Informática
 Maestría en Sistemas de Información Gestión y Tecnología

PROPUESTA DE UN MODELO PARA LA VISUALIZACIÓN DEL PARADIGMA ESTRUCTURAL DE LA WEB

TESIS

Que como parte de los requisitos para obtener el grado de:
 Maestría en Sistemas de Información Gestión y Tecnología

Presenta:
 Sandra Patricia Arreguín Rico

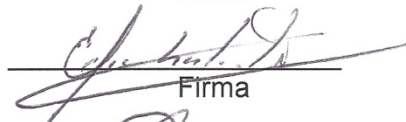
Dirigido por:
 M.S.I Elisa Morales Portillo

SINODALES


M.S.I. Elisa Morales Portillo
 Presidente


 Firma

Dr. Efrén Gorrostieta Hurtado
 Secretario


 Firma

M.I.S.D. Carlos Alberto Olmos Trejo
 Vocal



 Firma

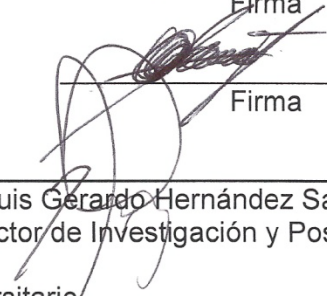
M.I.S.D. Juan Salvador Hernández Valerio
 Suplente


 Firma

M.I.S.D. Jesús Armando Rincones
 Suplente


 Firma


 M.C. Alejandro Santoyo Rodríguez
 Director de la Facultad


 Dr. Luis Gerardo Hernández Sandoval
 Director de Investigación y Posgrado

Centro Universitario
 Querétaro, Qro.
 Junio 2009
 México

RESÚMEN

La información en general, sea un libro, una página web o el contenido de una base de datos, respeta formatos de organización. Cuando se quiere operar la información de una base de datos por ejemplo o un informe financiero, resultan contenidos sencillos para el usuario ya que se sabe con qué tipo de información se cuenta desde un principio. La minería de datos se crea para extraer información que el usuario no sabe que existe. El no saber del usuario, lo que de antemano extrae la minería de datos, resulta en un proceso de comprensión un poco complicado. El propósito de la visualización de la información posterior al proceso de la minería es mostrar patrones y mejorar el entendimiento. Estas alternativas florecen con el objeto de hacer digerible la representación de grandes conjuntos de información textual. La telaraña semántica (Web) requiere convertir la información en conocimiento codificando los datos con metadatos (datos sobre los datos) legibles de forma automática. El motivo del presente es investigar los procedimientos aplicados en la minería visual de datos enfocados al paradigma estructural de la web, definida también como minería estructural de la web. Como primer paso se realizó una investigación bibliográfica acerca de las herramientas que permiten este tipo de análisis para detectar alcances u oportunidades. La aportación consiste en proponer un modelo basado en el análisis de una página web de fácil acceso, exponer la minería visual de datos en la misma por la estructura con la que se construye, representada como un árbol, es decir como estructura jerárquica, observar en ella el contenido, sea textual o multimedia y del que se pueda obtener información sobre la utilización que se hace del mismo, poder concluir que la minería de datos visual no solo es bonita sino también útil.

(Palabras clave: Minería de datos, minería web, visualización de la información, minería estructural de la web, minería visual de datos)

ABSTRACT

In general, information is presented in organized formats, whether it be found in a book, on a web page or in the content of a data base. For example, when one wishes to work with data base information or a financial report, the contents are quite simple for the user, since he/she knows beforehand what type of information he/she is dealing with. Data mining is created to obtain information which the user did not know existed. Since data mining usually involves extracting hidden information, the understanding process can become somewhat complicated. The purpose of information visualization after data mining shows patterns and increases understanding. These alternatives are available with the objective of making the representation of large amounts of textual information easier to comprehend. The semantic web requires the conversion of information to knowledge by coding the data with metadata (data about data) which are automatically readable. The goal of this research work is to investigate the procedures applied in visual data mining focusing on the structural paradigm of the web, better defined as web structure mining. The first step was to carry out bibliographic research in order to become familiar with the tools that allow for this type of analysis in detecting ranges or opportunities. The contribution of this work consists of proposing a model based on the analysis of an easily accessed web page, showing the visual data mining in the page by the way it was made – represented as a tree, in other words as a hierarchical structure – observing the contents, textual or multimedia from which one can obtain information on its utilization, and concluding that visual data mining is both interesting and useful.

(Key words: Data mining, web mining, information visualization, web structure mining, visual data mining)

DEDICATORIAS

A Dios que manifiesta su grandeza y amor para conmigo en cada día de mi vida y a través de mi esposo por decir lo más.

Lalo a ti por el tiempo robado, las ausencias aún estando en casa y porque te has encargado de que nuestros hijos sientan menos mi ausencia, por nuestra hermosa familia.

A Angel, Ana Paula, Lalito y María Karol, son mi motivo, bálsamo y cómplices.

A mis padres, que no terminan de darme amor, a mis hermanos y sus familias que llenan de orgullo.

A mis Suegros a Ale, por su inmensurable apoyo.

A mi familia, abuelos benditos, tías, tíos, primos, sobrinos, gran herencia.

Nena no hay palabras suficientes, las oportunidades y la fuerza que a mi transmites con tu ejemplo, QDTB.

AGRADECIMIENTOS

A Dios, por volver mis pasos a la Universidad Autónoma de Querétaro, que me albergó en el bachillerato y ahora me brinda la oportunidad de concluir este posgrado.

M.S.I. Elisa Morales Portillo por el tiempo y paciencia dedicada a dirigir y apoyar la presente tesis. Así mismo al M. en C. Alejandro Santoyo Rodríguez por las oportunidades que me ha otorgado. M. en C. Ruth Angélica Rico Hernández, porque con sus exigencias pedagógicas ha guiado mi espíritu de superación.

Efrén, por la compañía y la experiencia que has compartido a través de los años. CharlyO el tiempo, las ideas aportadas al trabajo, pero sobre todo por tu amistad. Valerio agradezco también tu amistad, la experiencia e invaluable apoyo para realizar la tesis. Armando por tus consejos y ecuanimidad.

Dr. Jesús Carlos Pedraza por haberme ayudado en todo lo que estuvo a su mano para realizar esta tesis. Maestra Elba, por su ánimo y contribución. Anabel Palacios, por tu compañía en esta nueva etapa y por tanto tiempo. Angélica A., fueron lindos fines de semana de trabajo.

A mis maestros y compañeros de maestría, por las experiencias compartidas a lo largo de este tiempo.

Í N D I C E

RESÚMEN	i
ABSTRACT	ii
DEDICATORIAS	iii
AGRADECIMIENTOS	iv
Í N D I C E	v
INDICE DE CUADROS	vi
INDICE DE FIGURAS	vii
1 INTRODUCCIÓN	1
1.1 Planteamiento del problema.	1
1.2 Justificación de la investigación.	3
1.3 Objetivos de la investigación	8
1.4 Organización de la tesis	9
2 MARCO TEORÍCO	10
2.1 Minería de Datos	10
2.2 Minería Web	18
2.2.1 Semántica web	25
2.3 Visualización de la información	28
2.3.1 Visualizando la web semántica	38
2.4 Minería visual de datos	42
2.4.1 Descubriendo los patrones secuenciales.	45
2.5 Árboles de decisión	50
3 CASO DE ESTUDIO	57
3.1 Método de Investigación	57
3.2 Justificación de los métodos	58
3.3 Proceso de extracción del conocimiento de un sitio web	59
3.3.1 Análisis de archivos de registro	65
3.4 Consideraciones iniciales	72
3.5 Visualizando con la herramienta WET <i>Website Exploration Tool</i>	77
4 RESULTADOS	82
4.1 Diseño de la investigación	82
4.2 Recolección de los datos	83
4.3 Conclusiones	86
BIBLIOGRAFIA	87

INDICE DE CUADROS

Cuadro		Página
1	Software comercial para Minería Web [Galeas, P. 1996]	22
2	Software público para Minería Web [Galeas, P. 1996]	26
3	Diseños con el mismo contenido, diferente visualización. [Redish, J 2001]	37

INDICE DE FIGURAS

Figura	Página
1 Mapa conceptual de la minería de datos web [C, Dürsteler 2005]	5
2 Geometría hiperbólica en 3 D [Walrus, 2005]	7
3 Etapas en el proceso de extracción del conocimiento [Fayad 96]	12
4 Proceso para extraer el conocimiento de sitios web [Galeas, P. 1996]	21
5 Transformación de información en conocimiento [C, Dürsteler. 2005]	28
6 Estudio de la lectura hecha [Redish, J 2001]	35
7 Esquema basado en el diseño de sitios web [Van Harmelen 2001]	39
8 Visualización de sitio web [Van Harmelen 2001]	40
9 Algoritmo aplicado a una base de datos educacional [Van Harmelen 2001]	41
10 Trama de diferentes tópicos de Julio a Septiembre 1990 [Wong, P.C2000]	46
11 Minería secuencial de patrones en el tiempo [Wong, P.C 2000]	47
12 Sistema de minería visual de datos [Wong,P.C 2000]	47
13 Características base de la minería visual de datos [Keim D. A. 2002]	48
14 Representaciones para estructura de llamadas [Keim D. A. 2002]	49
15 Diferentes representaciones gráficas [Keim D. A. 2002]	49
16 Estructura en tres capas del Web Map Viewer [C, Dürsteler Juan. 2005]	64
17 Estructura en tres capas del Web Map Viewer [C, Dürsteler Juan. 2005]	65
18 Análisis por visitas a un sitio	67
19 Resultados arrojados del análisis de registro de archivo	68
20 Imagen de VISVIP por John Cugini [C, Dürsteler Juan. 2005]	69
21 Línea generada por un archivo de registro (log file)	70
22 Esquema conversión de datos en sabiduría [Wurman R.S 1997]	72
23 Diagrama referente al proceso de visualización [C, Dürsteler Juan. 2005]	75
24 Diagrama construcción de visualizaciones [C, Dürsteler 2005]	76
25 Ventana de visualización utilizando la herramienta WET	79
26 Visualización de construcción para páginas web.	80
27 Errores de la página visualizada	83
28 Variables visuales asociadas a la métrica	84
29 Estadísticas de consulta	85

1 INTRODUCCIÓN

1.1 Planteamiento del problema.

En 1996 surge la minería de datos como herramienta y técnica para las tecnologías de la información, la cual ha abierto grandes posibilidades en la administración y gestión de las mismas. La minería de datos es utilizada en diversos ámbitos, en el presente trabajo, se presenta como el punto de partida para el análisis de la evolución de la web y el comportamiento de los usuarios, el cual es un trabajo relevante para que los administradores y los propietarios mejoren los sitios en términos de estructura, contenido y usabilidad.

Las estadísticas, así como la gran información que se genera de la minería de datos son de suma importancia, pero no son fáciles de interpretar y entender, la mayoría de la información debe enfocarse con la finalidad de descubrir tendencias y modalidades para lo cual se requiere de herramientas adicionales.

Una metodología basada en prototipos existentes puede permitir la representación visual que facilitaría el estudio y decisión de la forma en que se han de representar los datos disponibles [V Pascual, 2007].

Una parte importante de la minería web consiste en atacar el problema que representa la comprensión de cualquier método de minería. Por ejemplo, para mejorar la navegación del usuario es aún más difícil si uno no puede visualizar las diferencias a través de una gran colección de páginas Web o en una parte importante dentro de la estructura existente.

Las herramientas de visualización son desarrolladas con el fin de facilitar la interpretación de los resultados en la minería, combinan varias estrategias de otras herramientas que se aplican específicamente para el análisis de la web, unificando metodologías de visualización de la estructura y de la minería.

El valor de estas herramientas de visualización es representado en los siguientes ámbitos:

- A lo largo de las dimensiones de escala en:
 - Pequeñas estructuras
 - Grandes estructuras
- En la navegación dinámica puede visualizar :
 - una navegación dinámica o estáticamente.
- Y con el uso acumulativo que también puede distinguir entre:
 - Un individuo

y el uso de la web.

El análisis de la evolución de un sitio y el comportamiento de sus usuarios se ha convertido también en un trabajo crucial para los administradores y propietarios de sitios web que desean mejorar su sitio en términos de estructura y contenido. Las técnicas de minería de datos ofrecen muchas cifras y estadísticas útiles para comprender la estructura de un sitio web, y el uso que sus usuarios hacen de ella, pero que todavía no son fáciles de interpretar y de entender. La mayoría de estas cifras también pueden combinarse, con el fin de descubrir nuevas tendencias y modalidades.

Existe una diversidad de prototipos a analizar que permiten una exploración de contenidos en la web, para que la minería, cuyo objetivo principal es proporcionar un conjunto de herramientas y representaciones visuales, permitan realmente estudiar y decidir la forma de representar los datos disponibles para el usuario. Esta investigación pretende definir un modelo que ayude a evaluar la usabilidad de la aplicación de las interacciones y las metáforas visuales.

La existencia de herramientas que ayudan a evaluar la usabilidad de la aplicación de las interacciones y las metáforas visuales representan una efectiva aportación al conocimiento, con el estudio de ellas y la propuesta de una metodología para optar por la mejor representación.

En el presente trabajo, se realizó la investigación bibliográfica para justificar la creación de un espacio con una métrica de los modelos a consideración. Para visualizar el modelo de comparación el primer paso es crear un mapeo de entrada como de salida del proceso de modelado. El segundo paso es asignarle a este proceso al espacio visual de los humanos: El resultado conjunto de los pares se puede visualizar mediante una trama de puntos en un gráfico, este enfoque debe ser suficiente si se limita a la atención de una estructura bien definida.

Proponer un modelo para cubrir las oportunidades basándose en el análisis de una página web de fácil acceso, analizar la técnica y exponer la minería visual de datos aplicada a la misma. Realizar el estudio, proponer el modelo y conocer si la minería visual de datos representa una aportación significativa o es puramente estética, es la aportación del presente.

1.2 Justificación de la investigación.

La información conlleva a ejercer un control de la misma, éste se alcanza cuando se define una metodología aplicada al conocimiento mediante técnicas que permiten un resultado palpable de la transformación. Es incomparable el flujo del conocimiento hoy en día con el de los tiempos de antaño, actualmente es posible sentirse abrumado por la cantidad de información que se maneja.

En este contexto surge la minería de datos disciplina que se encarga de la gestión de los procesos involucrados en la explotación de la información que puede obtenerse mediante algoritmos para la búsqueda de la misma o bien con el registro de la cantidad de veces que se accede a ella, por ende es posible concebir modelos que lleven a acceder a juicios o conclusiones imposibles de determinar sin la minería de datos. Su desarrollo ha potenciado la convergencia de la administración estratégica y la optimización de la misma. Alcanzar las conclusiones y administración que permite la minería de datos requiere de estrategias no estandarizadas pues son tan variables como el conocimiento mismo.

En el presente trabajo el uso de la minería de datos es fundamental para el análisis de la red mundial, conocida como minería web, apoyada y también fundamentada en la minería de textos. La información accesada mediante equipo informático o vía internet es la materia prima de la minería web, por ser el nuevo vertedero de información e indicador de las tendencias más importante en el mundo.

El campo de la minería web abarca una amplia gama de oportunidades apuntando sobre todo, al proceso del conocimiento derivado de la investigación, recuperación de información, bases de datos e inteligencia artificial. Jaideep Srivastava lo define como “*La aplicación de técnicas de minería de datos para extraer el conocimiento de la información en la Web, en el que por lo menos una estructura de datos (Web log) es usada en el proceso de minería.*” Los miembros de la fundación de la ciencia en minería de datos para la venta del trabajo de la siguiente generación, [J.Srivastava 2002], [R. Kosala 2000] definen los paradigmas con los que se identifica a la minería Web:

- Minería de utilización web : Analiza los resultados de las interacciones entre el usuario y los servidores web, incluyendo los registros (*Web log*) de los accesos a la web.
- Minería del contenido web : Es la aplicación de las técnicas de minería de datos a la información publicada en internet, localización de patrones de texto en los documentos que usualmente se encuentran en versión HTML (semiestructurado), texto plano o XML (no estructurado).
- Minería de estructura web: Es el proceso de inferir conocimiento de los enlaces y estructura de las ligas en la web.

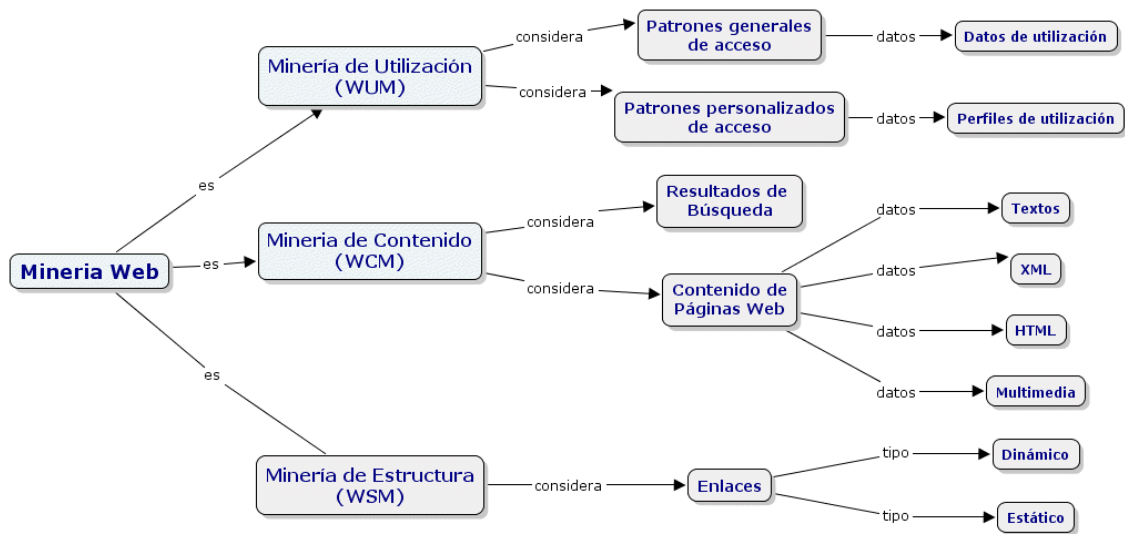


FIGURA 1 Mapa conceptual de la minería de datos web [C, Dürsteler 2005]

El mapa conceptual mostrado en la figura 1, es relevante para delimitar el paradigma de la minería de estructura web. Estos tres paradigmas que se nutren del campo del descubrimiento del conocimiento.

Una vez que se selecciona, y explora la información con las técnicas de análisis de datos es necesario moldearlas, para esto se requiere considerar la mayor parte posible de los resultados. Es muy común que las técnicas de análisis de datos produzcan extensos reportes, los cuales no siempre se interpretan de forma adecuada y pueden ser utilizados ineficientemente [Chi, E. H. 2002]. Una de las razones es porque los humanos son muy buenos identificando patrones visuales pero no textuales [Ansari, 2001].

Sin embargo diferentes técnicas de visualización son resultado de utilizar patrones estadísticos de visita a los sitios, estadísticas simples, topología de los sitios web, y el procesamiento analítico en línea OLAP web [Z. Pabarskaite 2003].

Es la visualización del contenido de la web un área compleja, no sólo por lo vasto y diverso de los contenidos sino por la complejidad semántica.

La minería web sufre de los mismos problemas que la avalancha de datos general, hacen falta herramientas de visualización que permitan digerir e interpretar los muchos resultados que proporciona. La minería de estructura web pertenece por su propia naturaleza a la estructura de hipervínculos en conjunto con la red, esto es la estructura de inter documentos en vez de la estructura intradocumentos que utiliza la minería de contenidos, es una estructura describible mediante grafos, esta visualización es importante para el análisis y la comprensión de la minería web. En la minería de estructura las gráficas representan las ligas en un sitio o entre sitios [Sankar. K. 2002].

Este es un vasto campo con posibilidades que exigen aprovecharse. Cualquier sitio web, especialmente la telaraña global se compone de páginas a las que se llamarán nodos y enlaces entre ellas a las que se denominan arcos, matemáticamente se puede considerar a la minería de estructura web como un grafo, dibujar a esté permitirá representarla.

La visualización de los grafos nos permite entender estructura de la web pues, y a su vez nos lleva a la posibilidad de tomar decisiones e implementarlas. La mayoría de las visualizaciones representan la estructura como jerarquía obviando los enlaces que se devuelven unas páginas a otras, formando circuitos, con lo que la estructura se puede representar como un árbol (un grafo conexo acíclico).

Estos circuitos están dentro de la categoría de representaciones de foco, contexto dado que permiten ver todo el contenido de la web y a la par, establecer un foco de atención al que acceder en más detalle. Walrus es una herramienta que ha ido más lejos en la representación de grandes jerarquías usando la geometría hiperbólica [Walrus, 2005].

En la figura 2 se muestra un ejemplo de geometría hiperbólica en 3D de un complejo árbol de directorios. En esta geometría cuanto más cerca de la periferia (el infinito) nos encontramos, más pequeño se hace el tamaño de lo que se desea representar y ello permite representarse manteniendo las propiedades de foco y contexto.

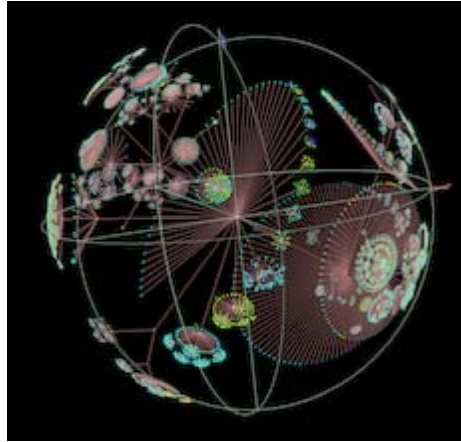


FIGURA 2 Geometría hiperbólica en 3 D [Walrus, 2005]

Los problemas que presentan las representaciones más interesantes son:

- Cada nodo hijo se convierte a su vez, en un nodo padre, cuyos hijos se representan de la misma manera, entonces el algoritmo recursivo da a todos los nodos el mismo tratamiento.
- Cuando el número de nodos es muy grande la estructura es difícil de percibir, por ejemplo cuando hay muchas páginas a esos niveles.
- Oclusión: unos nodos tapan a los otros.
- Distribución de los nodos en cada círculo, si en un círculo exterior hay muchos nodos, y pocos en los interiores, resulta difícil evitar el solapamiento resultando en que no se puede distinguir bien las líneas genealógicas.

1.3 Objetivos de la investigación

La hipótesis para sustentar el presente trabajo de investigación fue la siguiente:

“La representación visual de la información permite incrementar y facilitar la capacidad de análisis de la estructura de la web, mediante las técnicas de la minería de datos, proporcionando métricas y estadísticas para el entendimiento de la estructura de un sitio web así como del uso que se da de ella.”

Objetivo General:

Realizar un estudio comparativo entre los sistemas de visualización aplicados a la minería de estructura web, para el descubrimiento del conocimiento; consistentes en realizar una transformación del espacio de información en una representación simplificada.

Objetivos Específicos:

- Estudio de la visualización de sistemas que posean estructura representable como un árbol para disponer del contenido y obtener información sobre la utilidad de este, aplicados a la minería de estructura web.
- Evaluar la herramienta WET (Website Exploration Tool) mediante la visualización de la estructura de un sitio web de fácil acceso.
- Analizar la metáfora visual de árbol radial y mapeo de arboles.
- Establecer razonamientos y conclusiones a partir del estudio realizado.
- Demostrar la hipótesis mediante las técnicas de la minería de datos, proporcionando métricas y estadísticas para el entendimiento de la estructura de un sitio web así como del uso que se da de ella.

1.4 Organización de la tesis

En el primer capítulo se desarrolla el planteamiento del problema, así como la justificación y objetivos de la misma. El segundo capítulo define el marco teórico con los conceptos técnicos necesario para alcanzar el objetivo; estos son: Minería de datos, Minería web, Visualización de la información, Minería visual de datos, y árboles de decisión. En el capítulo 3 se hace referencia a los métodos que son parte del caso de estudio, el proceso de extracción del conocimiento en las etapas involucradas, así como el análisis mediante la visualización estructural del sitio para comparar métricas. Por último en el capítulo 4 se muestran los resultados, la evaluación de los objetivos y el análisis de la hipótesis propuesta, reflejado en las conclusiones.

2 MARCO TEORÍCO

2.1 Minería de Datos

La minería de datos está predestinada a ser uno de los desarrollos revolucionarios de la próxima década, de hecho es seleccionada como una de las diez tecnologías emergentes que han de cambiar al mundo. De acuerdo con el grupo Gartner [Larose D 2005], “La minería de datos es un proceso que se utiliza para descubrir tanto nuevas correlaciones significativas, como patrones y tendencias a través del análisis de grandes cantidades de datos almacenados en los repositorios, mediante tecnologías de reconocimiento de patrones, así como de técnicas estadísticas y matemáticas”. La minería de datos es cada vez más generalizada, ya que faculta a empresas rentables permitiéndoles descubrir patrones y tendencias de sus bases de datos.

Las empresas que no están aplicando estas técnicas corren el peligro de quedarse atrás, perdiendo cuota de mercado, porque sus competidores están aplicando esta nueva tecnología y por tanto obtienen la ventaja competitiva [Larose D 2005]. El desarrollo que promete esta disciplina está basado en las técnicas de análisis y extracción de modelos, a su vez varias son las disciplinas tradicionales que se distinguen por orientarse a extraer patrones, tendencias, regularidades y predicción de comportamientos, sacando todo el provecho posible de la vasta información computarizada existente. La minería de datos permite comprender y modelar de una forma eficiente el contexto en el que se deben actuar y tomar decisiones.

El término minería de datos es sin embargo una sola etapa importante en el proceso de extracción del conocimiento a partir de datos, este proceso consta de varias fases e incorpora diversas técnicas de los campos del aprendizaje automático, la estadística, las bases de datos, los sistemas de toma de decisión, la inteligencia artificial y otras áreas de la informática así como de la gestión de la información.

En nuestra realidad tomamos decisiones en base a la información y ésta conforme el tiempo avanza es más vasta, por lo que es necesario organizarla incluso de forma automática para lograr el aprovechamiento de la misma, esto se viene exigiendo por la omnipresencia de la información electrónica, esta se presenta en la mayoría de las ocasiones en completa desorganización y con algo de suerte podemos obtenerla de bases de datos de forma estructurada.

La minería de datos pone al alcance de las pequeñas organizaciones la tecnología para madurar las técnicas de aprendizaje automático y tratar volúmenes de información de tal forma que los usuarios puedan tomar decisiones. Las herramientas para analizar datos son aplicables en áreas como la biomedicina, ingeniería, control, administración, domótica, etc. Algunos ejemplo de estas herramientas son: Clementine de SPSS, Intelligent Miner de IBM, Mine Set de Silicon Graphics, Enterprise Miner de SAS, DM Suite (Darwin) de Oracle, WEKA (de libre distribución) Knowledge Seeker, etc., cuya característica es facilitar o poner al alcance el uso de la minería de datos a usuarios con estudios o no en medios informáticos, aunque requiriendo de un conocimiento o visión global del uso de las mismas en diferentes dominios y con diversas técnicas pues es necesario discernir entre las técnicas y metodologías de forma propia adecuándose a las limitaciones existentes [Hernández 2004].

Las fases necesarias para la extracción del conocimiento a partir de bases de datos son: *la recopilación de datos*, mediante almacenes de datos o de manera directa, *la preparación de datos*, mediante visualización, agregación, limpieza o transformación, *la minería de datos*, mediante técnicas descriptivas o predictivas, *la evaluación y mejora de modelos*, mediante validación cruzada, combinación o análisis de costos y finalmente *la difusión y uso del conocimiento extraído*, mediante estándares de intercambio de conocimiento, XML, modelos convertidos a lenguajes de programación u otras herramientas. La minería de datos surge de la necesidad de interpretar la gran cantidad de datos que se genera y almacenan en los sistemas de información, el aumento de volumen y variedad se encuentra en bases de datos digitales.

Esta información de carácter histórica representa transacciones necesarias para explicar el pasado, entender el presente y predecir la información futura. La toma de decisiones se ha de basar en estos históricos extraídos de fuentes diversas y diferentes dominios, por tanto es necesario un análisis de los mismos para que se logre la utilidad de la información. En su mayoría el método tradicional de convertir los datos en conocimiento se basa en el análisis e interpretación realizada de forma manual, forma que se presenta lenta, cara y subjetiva, de tal forma que se da imposible en situaciones donde el volumen de los datos crece de forma exponencial. Esta abundancia en los datos rebasa la capacidad humana para comprenderlos si no se cuenta con la ayuda de herramientas potentes, prácticamente muchas decisiones relevantes se toman, no sobre la cantidad de datos disponibles sino por la intuición del usuario al no disponer de las herramientas necesarias. [Hernández 2004]

Este es el contexto en el que la minería de datos surge para resolver el problema analizando los datos presentes en las bases de datos. Es importante observar en las fases del descubrimiento del conocimiento, el momento en el que se está trabajando con los datos hasta dejarlos listos para la creación de modelos por parte de la minería de datos y la forma en que transmite mediante decisiones basadas en estos patrones el conocimiento mismo, ejemplo de esta división se visualiza en la figura 3.

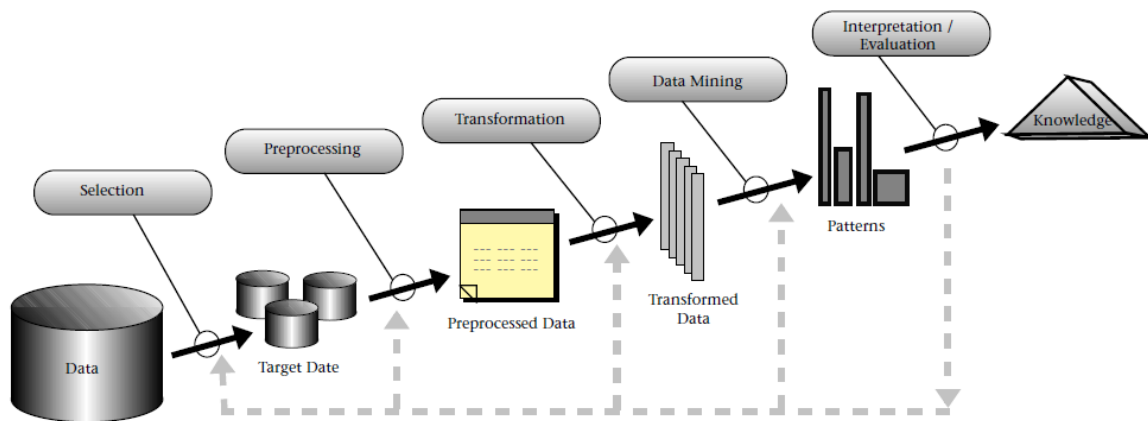


FIGURA 3 Etapas en el proceso de extracción del conocimiento [Fayad 96]

Anteriormente el análisis de los datos de una base se realizaba mediante consultas efectuadas con lenguajes generales de consulta, como lo es el SQL, y se producía sobre la base de datos operacional, junto al procesamiento de la transacción en línea (On-Line Transaction Processing, OLTP) de las aplicaciones de gestión. Sin embargo esta costumbre solo permite generar información resumida de una manera previamente establecida (generación de informes), poco flexible y sobre todo, poco escalable a grandes volúmenes de datos. La tecnología de bases de datos ha respondido a este reto con una nueva arquitectura surgida recientemente: el almacén de datos (data warehouse). Se trata de un repositorio de fuentes heterogéneas de datos, integrados y organizados bajo un esquema unificado para facilitar su análisis y dar soporte a la toma de decisiones. Esta tecnología incluye operaciones de procesamiento analítico en línea (On-Line Analytical Processing, OLAP), son técnicas de análisis como el resumen, la consolidación o agregación o el hecho de ver la información desde distintas perspectivas. Sin embargo, a pesar de que las herramientas OLAP soportan cierto análisis descriptivo y de sumarización que permite transformar los datos en otros agregados o cruzados de manera sofisticada, no generan reglas, patrones, pautas, es decir, conocimiento que se pueda aplicar a otros datos.

En muchos contextos como los negocios, la medicina o la ciencia, los datos por sí solos tienen un valor relativo. El interés está en el conocimiento, por ejemplo si conocemos estadísticamente el porcentaje de alguna enfermedad puede ser útil, pero lo que en realidad ayudaría a la población es tener un conjunto de reglas que a partir de antecedentes, hábitos y características relevantes en el paciente indicase si este tiene o no la enfermedad. Existen otras herramientas analíticas que se han empleado para analizar los datos y que tienen su origen en la estadística, algo lógico teniendo en cuenta que la materia prima de esta disciplina son precisamente los datos. Hay entonces paquetes estadísticos con la capacidad de inferir patrones a partir de los datos (utilizando modelos estadísticos paramétricos o no) el problema es que resultan especialmente encriptados para los no estadísticos, generalmente no funcionan bien para el tamaño de gigabytes de las bases de datos actuales y el tipo de algunos datos frecuentes en ellos (atributos nominales con muchos valores, datos textuales, multimedia, etc.) y no se integran bien con los sistemas de información. [Hernández 2004]

Entonces se puede presentar a la estadística como la madre de la minería de datos, y ésta ha ido ganando el prestigio y concepción de disciplina integradora. La minería de datos surge de los problemas y limitaciones de las aproximaciones clásicas, nace como una nueva generación de herramientas y técnicas para soportar la extracción de conocimiento útil desde la información disponible, distinguiéndose de las aproximaciones anteriores por no obtener información extensional (datos) sino intencional (conocimiento), y siendo el conocimiento un modelo novedoso y original extraído por la herramienta sin ser una parametrización de ningún modelo preestablecido. El resultado de la minería de datos son conjuntos de reglas, ecuaciones, árboles de decisión, redes neuronales, grafos probabilísticos, etc., los cuales pueden usarse para responder cuestiones de comportamiento diferenciado, secuenciación de tratamientos, asociaciones o calificaciones.

Retomando la definición estricta con que se inicio el tema entonces, la minería de datos definida por Witten y Frank es el proceso de extraer conocimiento útil y comprensible, que era previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos, entonces la tarea fundamental es encontrar modelos inteligibles a partir de los datos. Para que el proceso sea efectivo debiese ser automático, o semi-automático y el uso de los patrones descubiertos debería ayudar a la toma de decisiones seguras que reporten beneficio a la organización. Por tanto dos son los retos de la minería de datos: primero el trabajar con grandes volúmenes de datos, procedentes mayoritariamente de sistemas de información, con los problemas que ello conlleva (ruido, datos ausentes intratabilidad, volatilidad de los datos, etc.), y segundo usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil. En muchos casos la utilidad del conocimiento minado está íntimamente relacionada con la comprensibilidad del modelo inferido. Es importante entonces que la información descubierta sea comprensible para los humanos utilizando técnicas de visualización o representaciones gráficas para convertir los patrones a lenguaje natural de los datos. Definir el objetivo de la minería de datos es hablar de convertir datos en conocimiento. Ahora es necesario diferenciar entre datos estructurados provenientes de bases de datos relacionales, otros tipos de datos estructurados en bases de datos (especiales, temporales, textuales y multimedia) y datos no estructurados provenientes de la web o de otros tipos de repositorios de documentos.

Una base de datos relacional es una colección de relaciones que se encuentran generalmente en forma de tablas, cada tabla consta de un conjunto de atributos; una de las principales características de las bases de datos relacionales es la existencia de un esquema asociado, los datos siguen una estructura y por tanto son estructurados. Aunque las bases de datos relacionales (recogidas o no en un almacén de datos, normalizadas o estructuradas de una manera multidimensional) son la fuente de datos para la mayoría de aplicaciones de minería de datos, muchas técnicas no son capaces de trabajar con toda la base de datos, sino que sólo tratan con una sola tabla a la vez. Aunque las bases de datos relacionales son, con gran diferencia, las más utilizadas hoy en día, existen aplicaciones que requieren otros tipos de organización de la información.

Ya mencionadas las bases de datos espaciales con información relacionada con el espacio físico en un sentido amplio (una ciudad, una región montañosa, un atlas cerebral, etc.). Estas bases incluyen datos geográficos, imágenes médicas, redes de transporte o información de tráfico entre otros. La minería de datos sobre estas bases de datos permite encontrar patrones entre los datos, como por ejemplo, características de las casas en una zona montañosa o por ejemplo la planificación de nuevas líneas de metro en función de la distancia de las distintas áreas a las líneas existentes. Las bases de datos temporales almacenan datos que incluyen muchos atributos relacionados con el tiempo o en el que éste es muy relevante. Estos atributos pueden referirse a distintos instantes o intervalos temporales. En este tipo de bases de datos las técnicas de minería pueden utilizarse para encontrar las características de la evolución o las tendencias del cambio de distintas medidas o valores de la base de datos.

Las bases de datos documentales contienen descripciones para los objetos (documentos de texto) que pueden ir desde las simples palabras clave a los resúmenes. Estas bases de datos pueden contener documentos no estructurados como una biblioteca digital de libros, semi-estructurados o estructurados.

Las técnicas de minería de datos pueden utilizarse para obtener asociaciones entre los contenidos, agrupar o clasificar objetos textuales. Para ello los métodos de minería se integran con otras técnicas de recuperación de información y con la construcción o uso de jerarquías específicas para datos textuales, como los diccionarios. Las bases de datos multimedia almacenan imágenes, audio y vídeo. Soportan objetos de gran tamaño ya que los vídeos pueden necesitar varios gigabytes de capacidad para su almacenamiento. Para la minería es necesario para estos casos integrar los métodos con técnicas de búsqueda y almacenamiento. Las bases de datos objetuales y las objeto-relacionales son aproximaciones generales a la gestión de la información y, por tanto pueden utilizarse para los mismos usos que las relacionales o para algunas de las bases de datos especiales. [Hernández 2004]

La minería de datos tiene como objeto analizar los datos para extraer conocimiento, este puede ser en forma de relaciones, patrones o reglas inferidos de los datos previamente desconocidos o bien en forma descriptiva concisa, como un resumen. Estas relaciones o resúmenes constituyen el modelo de los datos analizados. Existen muchas formas diferentes de representar los modelos y cada una de ellas determina el tipo de técnica que puede usarse para inferirlos. En la práctica, los modelos pueden ser de dos tipos: predictivos y descriptivos. Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que denominamos variables objetivo o dependientes, usando otras variables o campos de la base de datos, a las que nos referiremos como variables independientes o predictivas.

Un modelo predictivo es aquel que permite estimar la demanda un producto en función de la inversión en publicidad. Los modelos descriptivos identifican patrones que explican o resumen los datos, es decir, sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos, retomando el ejemplo anterior se puede identificar el producto a vender en base a los gustos de los clientes reflejados en un histórico de ventas.

Los modelos inferidos por los árboles de decisión y las redes neuronales por ejemplo pueden inferir modelos predictivos.

Igualmente, para una misma técnica se han desarrollado diferentes algoritmos que difieren en la forma y criterios concretos con los que se construye el modelo. La minería de datos es un campo multidisciplinar que se ha desarrollado en paralelo o como prolongación de otras tecnologías.

Por tanto la investigación y los avances en la minería de datos se nutren de los que se producen en estas áreas relacionadas.

- ❑ Las bases de datos: conceptos como los almacenes de datos y el procesamiento analítico en línea, tienen una gran relación con la minería de datos, aunque en este último caso no se trata de obtener informes avanzados a base de agregar los datos de cierta manera compleja pero predefinida, sino de extraer conocimiento novedoso y comprensible. Las técnicas de indexación y de acceso eficiente a los datos son muy relevantes para el diseño de algoritmos eficientes de minería de datos.
- ❑ La recuperación de información: Consiste en obtener información desde datos textuales, por lo que su desarrollo histórico se ha basado en el uso efectivo de bibliotecas digitales y en la búsqueda por Internet. Una tarea típica es encontrar documentos a partir de palabras claves, lo cual puede verse como un proceso de clasificación de los documentos en función de estas palabras clave. Para ello se usan medidas de similitud entre los documentos y la consulta. Muchas de estas medidas e han empleado en aplicaciones más generales de minería de datos.
- ❑ La estadística: Esta disciplina ha proporcionado muchos de los conceptos, algoritmos y técnicas que se utilizan en minería de datos, como la media, la varianza, las distribuciones, el análisis univariante y multivariante, la regresión lineal y no lineal, la teoría del muestreo, la validación cruzada, etc. De hecho algunos paquetes de análisis estadístico se comercializan como herramientas de minería de datos.
- ❑ El aprendizaje automático: Área de la inteligencia artificial que se ocupa de desarrollar algoritmos capaces de aprender y junto con la estadística es el corazón del análisis inteligente de los datos.

Los principios seguidos en el aprendizaje automático y en la minería de datos son los mismos: la máquina aprende un modelo a partir de ejemplos y lo usa para resolver el problema.

- ❑ Los sistemas para la toma de decisión: Son herramientas y sistemas informatizados que asisten a los directivos en la resolución de problemas y en la toma de decisiones. El objetivo es proporcionar la información necesaria para realizar decisiones efectivas en el ámbito empresarial o en tareas de diagnóstico.
- ❑ La visualización de datos: El uso de técnicas de visualización permite al usuario descubrir, intuir o entender patrones que serían más difíciles de ver a partir de descripciones matemáticas o textuales de los resultados. [Hernández 2004]

2.2 Minería Web

El término de minería web se debe a O. Etzioni quien en 1996 la definió como la integración de información obtenida mediante los métodos tradicionales de la minería de datos con la información recogida sobre la web. Es la minería de datos aplicada a las especificidades de las páginas web, una técnica de análisis que se usa para el estudio de varios aspectos esenciales del sitio, ayuda a descubrir tendencias y relaciones en el comportamiento de los usuarios que sirven como pistas para mejorar la usabilidad de un sitio. Se asocia con la minería por la idea de excavar en busca de los datos. Generalmente se analizan grandes volúmenes de información utilizando algoritmos y luego se les representa en modelos para que puedan ser analizados, la minería web traslada este modelo al análisis de sitios, procesando los datos disponibles para su posterior examen.

Cuando un sitio es navegado por los usuarios los archivos de registro de los servidores que lo alojan van guardando información acerca de esa visita:

- »» Día y hora en la que el usuario navega por el sitio,
- »» número de visita y reincidencia,
- »» archivo que le interesa del sitio ,
- »» tiempo de visita ,
- »» país de origen del usuario,

- »» navegador del usuario,
- »» sistema operativo del usuario,
- »» ingreso directo o por enlace,
- »» ingreso por buscador, palabras solicitadas,
- »» etc.

Esta información es procesada por programas estadísticos que brindan información para la mejora del sitio, mediante información estructurada y significativa acerca de la navegación, por ejemplo:

- »» Cantidad de visitas ordenadas cronológicamente,
- »» clasificación de horarios por frecuencia,
- »» clasificación de popularidad por páginas,
- »» orígenes de uso frecuente,
- »» etc.

Las reglas de clasificación, agrupamiento, asociación y sucesos frecuentes son técnicas de minería de datos aplicadas a minería web y permiten clasificar y agrupar a los usuarios, asignando patrones de comportamiento según la reiteración de acciones que se detectan como usuario y clave, para poder ofrecerles productos o servicios acordes a sus perfiles.

Para mejorar la efectividad de la minería web, se permite descubrir patrones relacionados por áreas: la estructura, el contenido y la utilización de los sitios web. Esta herramienta es imprescindible para los desarrolladores de páginas web, sin embargo como en todas las tecnologías de información aun hay mucho por realizar, para continuar con este desarrollo la visualización de la información es una gran herramienta. La minería web se nutre de tres ámbitos dentro del campo del descubrimiento del conocimiento. La minería de la estructura de la web, revela la estructura real de un sitio web a través de recoger datos referentes a la misma y principalmente la conectividad.

De forma típica se consideran dos tipos de enlace, los estáticos y los dinámicos. La minería de contenido de la web se enfoca en recoger datos e identificar patrones relevantes a los contenidos de la web, así como a las búsquedas que se realizan sobre los mismos. Hay dos estrategias principales para la minería de contenido de la web, pueden extraerse patrones directamente de los contenidos existentes en las páginas. Los datos que se utilizan en este caso puede ser el texto libre, páginas escritas en HTML o bien en XML, elementos multimedia, o cualquier otro tipo de contenido presente en la web.

Una segunda estrategia es identificar los resultados de búsqueda, para intentar identificar patrones en los resultados de los motores de búsqueda. La minería de la utilización de la web intenta sumergirse en los registros de los servidores con el fin de encontrar patrones sobre el uso que se le da a la web. Se realiza un seguimiento de patrones generales de acceso pues el objetivo es obtener datos sobre la integración de ellos en tendencias generales para lograr una reestructuración del sitio y permitirse ofrecer un acceso ligero a los clientes.

La minería web es una disciplina con un importante potencial. Pese al creciente y enorme volumen de sitios web existentes aún es baja la proporción de sitios que emplean herramientas de minería para analizar su estructura, contenido y utilización en aras de un mejor servicio al usuario y de la mejora de sus productos. Se cuenta con diversas herramientas de minería de datos como el *WebLogMiner* y se propone un registro de técnicas de sitios web para utilizar la minería de datos y el proceso analítico en línea OLAP transformándose en archivos de acceso a la web. Por otro lado el seguimiento personalizado permite obtener datos sobre el comportamiento y la interacción con la web por parte de visitantes individuales a fin de establecer perfiles de acceso.

La figura 4 nos muestra los procesos necesarios para aplicar la minería web. Es importante fomentar y ver el potencial de de las diversas aplicaciones de análisis de los archivos de registro así como saber que el éxito de dichas aplicaciones depende de que, cuanto y con qué frecuencia se puede descubrir conocimiento confiable.

La personalización de secuencias de comandos para algunos sitios puede almacenar información adicional. Sin embargo, para una minería web eficaz puede ser importante el paso de transformación, limpieza y análisis de los datos.

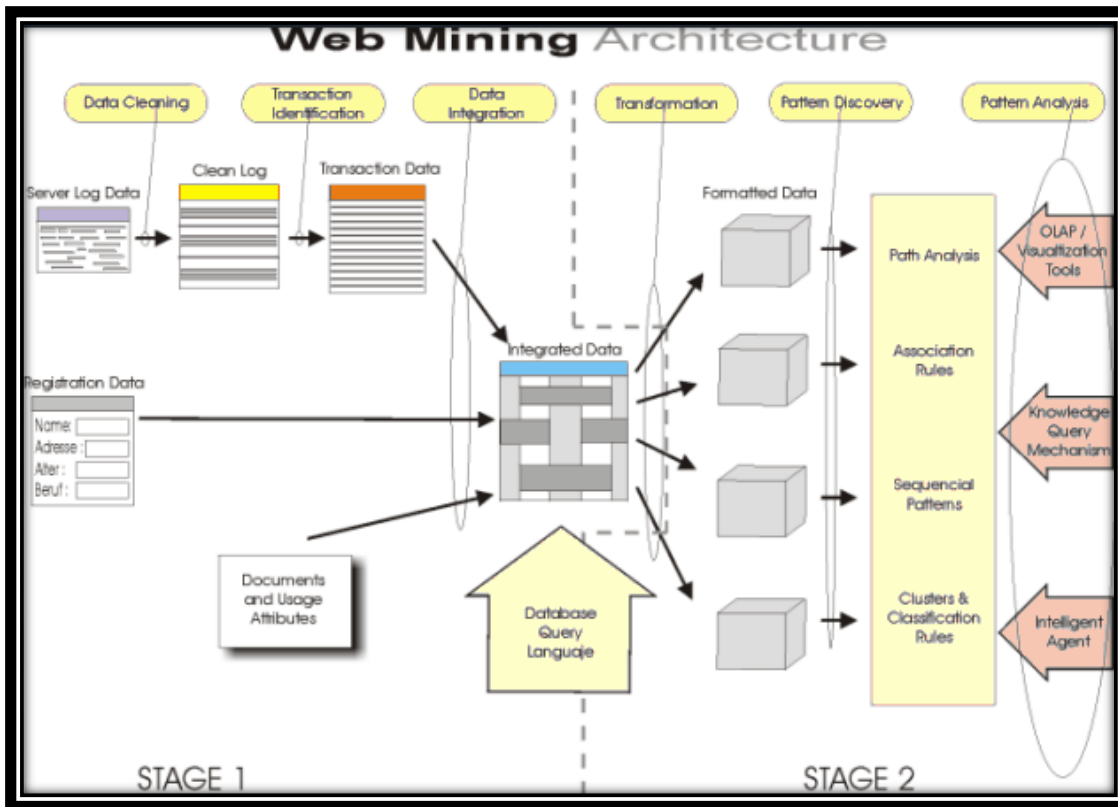


FIGURA 4 Proceso para extraer el conocimiento de sitios web [Galeas, P. 1996]

En la tabla 1[Galeas, P. 1996] se enlistan los nombres y atributos de software comercial importante para esta técnica.

La minería web es finalmente un análisis significativo de los archivos de registro guardados en los servidores cuyo proceso genera información de valor acerca del sitio y los usuarios ante cierta indexación de contenidos o estructuras de texto, preferencias e inconsistencias.

Nombre	Firma	Tipo	Comentarios
DB Miner	Simon Fraser Universidad, Canadá	Herramienta de minería de datos	Proporciona un potente herramienta al almacén de datos y bases de datos relacional, es rápida y eficaz con múltiples funciones de minería de datos. Esta versión del software utiliza a Microsoft SQL Server 7.0 Platón para construir los cubos de datos en el que realiza tareas de minería---una dramática modificación, mejora la versatilidad y la eficiencia de DBMiner.
Speed Tracer	IBM	Herramienta de minería de datos	“El SpeedTracer es una herramienta para el uso de la minería web y el análisis del usuario, con patrones de pistas de navegación, generación de informes para ayudar a afinar la estructura del sitio Web de los administradores y la navegación. La aplicación utiliza algoritmos de inferencia innovadores para reconstruir las rutas de recorrido del usuario e identificar las sesiones. La Avanzada de algoritmos de minería para descubrir el movimiento de los usuarios a través de un sitio Web. El resultado final es una colección de patrones de navegación valiosos que ayudan a a los administradores de la web a comprender mejor el comportamiento del usuario. SpeedTracer genera tres tipos de estadísticas: basados en el grupo, usuario-en y basada en la ruta de acceso. Basado en el usuario, las estadísticas señalan los recuentos de referencia por el usuario y duraciones de acceso. Las estadísticas basadas en la ruta de acceso identifican las rutas transversales frecuentes en las presentaciones de la Web. Las estadísticas basadas en el grupo de información sobre grupos frecuentes que visitan las páginas del sitio Web. “

Tabla 1 Software comercial para Minería Web [Galeas, P. 1996].

Nombre	Firma	Tipo	Comentarios
Funnel Web Pro	Active Concepts	Analizador de archivo de registro	Web del embudo 4.0 es la última liberación de nuestro análisis inteligente clásico y presentación de informes de internet en el software. Diseñado con una interfaz completamente nueva la versión 4.0 es incluso más fácil de usar y configurar que las versiones anteriores de Web.Plus Funnel. Este producto contará con una serie de impresionantes capacidades (como la administración remota totalmente basado en web) más mucho más!.Con una nueva interfaz intuitiva, atractiva y más poder que nunca, Web 4.0 es todo lo que necesita permanecer encima de su en línea Imperio.
Knowledge Studio	Angoss	herramienta de minería de datos	Knowledge STUDIO es una nueva generación de software de minería de datos. Se integra técnicas avanzadas de minería en entornos corporativos en donde las empresas pueden lograr el máximo beneficio de sus inversiones en los datos. KnowledgeSTUDIO es una herramienta de minería de datos que incluye el poder de árboles de decisión, análisis de clúster y varios modelos predictivos que permiten a los usuarios comprender sus datos de diferentes perspectivas. Incluye herramientas de visualización de la información potente apoyando y explicando los descubrimientos.
Clementine	SPSS	herramienta de minería de datos	La aplicación utiliza innovadores algoritmos de inferencia para reconstruir las rutas de recorrido del usuario e identificar las sesiones del mismo. Algoritmos de minería avanzados para descubrir el movimiento de los usuarios a través de un sitio Web. El resultado final es una colección de navegación valiosa para los patrones que ayudan a que los administradores de la web comprendan mejor el comportamiento del usuario.

Nombre	Firma	Tipo	Comentarios
Sawmill 5	Flowerfire	Analizador de archivo de registro	Sawmill es un registro jerárquico, potente herramienta de análisis para Windows 95/98/NT/2000, MacOS, UNIX, OS/2 y BeOS. Es especialmente adecuado para los registros de acceso web y la referencia del servidor, pero puede procesar casi cualquier registro. El informe que genera son jerárquicos, atractivo para la fácil navegación. Es la documentación completa integrando directamente el programa
WUM	Universidad de Humboldt Berlín	herramienta de minería de datos	WUM es una secuencia de la minería. Su principal objetivo es analizar el comportamiento de navegación de los usuarios en un sitio Web, pero es apropiado para el descubrimiento del patrón secuencial en cualquiera tipo de registro. Descubre patrones compuestos pero no necesariamente eventos adyacentes y satisface los criterios de específica del usuario. WUM es un entorno integrado para la preparación de registro, consulta y visualización. Su lenguaje de consulta de minería MINT es compatible con la especificación de criterios que describe patrones dominantes o estadísticamente raros. Su visualización es un mecanismo que muestra los nodos que comprende el modelo deseado y las diferentes rutas no-frecuentes situando medios. Esto es bastante importante al examinar el sitio web.
Commerce Trends	Web Trends	herramienta de minería de datos	CommerceTrends proporciona al comercio electrónico una poderoso inteligencia con informes disponibles, lo que permite a los clientes realizar el seguimiento, administrar y optimizar las estrategias de comercio electrónico. CommerceTrends es avanzado en funcionalidad incluyendo potentes, análisis de tráfico web de empresa-escalable, gestión de la campaña, previsión, eMarketing ROI y web, almacén de datos de ingresos de comercio electrónico así como las capacidades, que permite a los clientes aplicar principios de almacén de datos a datos de tráfico web de correlacionarse con otra información corporativa de CRM, ERP y sistemas de personalización.

Nombre	Firma	Tipo	Comentarios
Net Analysis	Net Genesis	herramienta de minería de datos	NetAnalysis, el galardonado. Da solución al análisis del comportamiento en línea de NetGenesis, proporciona la escalabilidad superior y una extensibilidad potente requerida por el administrador electrónico, cada vez es más competitiva en la línea del medio ambiente. Con mayor flexibilidad y funcionalidad, NetAnalysis puede personalizarse para satisfacer al cliente electrónico específico para las necesidades de cualquier empresa de inteligencia, al aprovechar fácilmente su arquitectura de apoyo.

Tabla 1 Software comercial [Galeas, P. 1996]

En la tabla 2 [Galeas, P. 1996] se enlistan los nombres y atributos de software público para minería web.

2.2.1 Semántica web

La telaraña o semántica web se presenta como la nueva revolución de Internet. La promesa es convertir información en conocimiento. La telaraña mundial (*www*) o web, para el caso, es un espacio de información que ha permitido nuevos niveles de comunicación humana. Por ello mismo la información que existe en la misma se ha diseñado básicamente para el consumo humano y utiliza un lenguaje que hace difícil su utilización por parte de las máquinas para el intercambio y elaboración efectiva de datos.

Las aplicaciones de comercio electrónico, por ejemplo, requieren el flujo de datos entre proveedores, distribuidores, comercios e incluso con el usuario final. Actualmente los intercambios consisten en simples transacciones de datos separados por tabuladores o aplicaciones muy específicas.

La visión que hay detrás de la idea de la Telaraña Semántica es la de que los datos que hay en la red estén definidos de tal forma que puedan ser utilizados y comprendidos por las máquinas sin necesidad de intervención humana.

Nombre	Firma	Tipo	Comentarios
STstat	ST Software	Reportes y estadísticas	Es un conjunto de scripts CGI (escrito en C), que producen informes HTML, basados en los registros de acceso que mantiene el servidor HTTP, y es adecuado para casi todos los software http de servidor (Unix y Windows), apoya ahora tres formas de inicio de sesión (común, extendida e IIS).
weblog_parse	ACME Labs software.	Procesamiento de los archivos de registro	Extrae campos específicos de un archivo de registro de la Web. Lee un archivo de registro de servidor web, en cualquier "formato de archivo de registro común" o "Combina el formato de archivo de registro". Lo analiza y sólo escribe el campo especificado por el usuario, separado por fichas para el manejo fácil
WebLog	Darryl C. Burgdorf	Herramienta de análisis de los archivos de registro	Es una herramienta de análisis completa para los registros de acceso. Permite el seguimiento de la actividad del sitio por mes, semana, día y hora, para supervisar el número de visitas totales, bytes transferidos, página de opiniones y mantener un seguimiento de las páginas más populares.
Analógico	Universidad de Cambridge estadística Laboratorio	Analizador de los archivos de registro	Analógico es un programa para analizar los archivos de registro del servidor web. Que le dice qué páginas son las más populares, que personas y de que países realizan la consulta, sitios que se ha intentado seguir, así como vínculos rotos, etc..

Tabla 2 Software público para Minería Web [Galeas, P. 1996]

La web se convertiría en un espacio auto-navegable y auto-comprensible. Pero la cosa aún va más allá, de lo que se trata es de convertir la información en conocimiento codificando los datos con metadatos, datos sobre los datos legibles de forma automática. Esta codificación viene de la mano de la definición de diferentes Ontologías. Una Ontología es, en este contexto (no confundir con el concepto filosófico), la especificación de una conceptualización, esto es de un conjunto de definiciones de conceptos. Las Ontologías se expresan mediante lenguajes de representación que, en la telaraña semántica, se construye encima de XML.

La creación de Ontologías está dando lugar al desarrollo de Editores de Metadatos o Editores Ontológicos como Protégé o Webonto y a sistemas para favorecer la interoperabilidad, la transformación entre unas ontologías y otras. También se trabaja activamente en el procesado de las mismas mediante motores de inferencia que permiten deducir nuevos conocimientos sobre conocimientos ya especificados. En principio crear software sería cuestión de encontrar los componentes apropiados en la red junto con la especificación de cómo enlazarlos. Un agente apropiado (no necesariamente humano) podría realizar esta operación. Pero para las organizaciones podría ser un verdadero salto cualitativo al permitir codificar su conocimiento interno y usarlo apropiadamente para su relación a través de la red con sus proveedores y clientes.

La realización de esta visión, y la organización semántica web, necesitará de la estandarización e incorporación de las muchas herramientas y tecnologías sobre las que se está trabajando y de la adopción de unos y otras por parte del mercado. El primer ejemplo del estándar oficial es el ISO/IEC 13250 sobre Mapas Temáticos, en esta norma se definen los requerimientos para la arquitectura de documentos . En este estándar se provee una notación para la representación en el intercambio de la información referente a la estructura usual de las fuentes de información para definir tópicos, los cuales definen las ocurrencias que existe entre las direcciones y la asociación o relación entre ellas [ISO/IEC 1999].

Las empresas de la red eléctrica en los EEUU han adoptado RDF para intercambiar modelos de generación de corriente. La diferencia entre información y conocimiento es lo que hace efectiva a una organización.

2.3 Visualización de la información

De entrada hay que decir que muchas definiciones del término lo ligan al uso del ordenador y de la función visual, ello no es así. Para entender mejor hay que comprender por un lado qué es visualización y por el otro qué es información. Seguir las definiciones es el camino. Visualización: Formación de la imagen mental de un concepto abstracto. El Collins English Dictionary sustituye “concepto abstracto” por “algo incapaz de ser visto o no visible en ese momento”. Así, pues, para visualizar no hace falta usar la visión, es más, al ser un fenómeno mental, la visión no interviene en ese momento, aunque sí lo haya hecho en la formación de los símbolos y elementos que acaban componiendo la imagen mental. Podemos observar en la figura 5 que la Visualización de Información incide en la transformación de información en conocimiento.



FIGURA 5 Transformación de información en conocimiento [C, Dürsteler. 2005]

Una imagen es, en filosofía, la conciencia de un objeto ausente o inexistente y en psicología es la representación construida al margen de los correspondientes estímulos sensoriales (Diccionario Enciclopédico Salvat).

En definitiva, la imagen y por tanto la visualización, es una construcción mental que va más allá de la percepción sensorial y que como tal construcción mental se acerca al conocimiento, que es la aprehensión intelectual de las cosas. Comprender quiere decir rodear, incluir una cosa, interiorizarla.

De la información: Para algunos diccionarios es el “conocimiento adquirido a través de la experiencia o el estudio” para otros es la “comunicación o adquisición de conocimientos”. La información consiste en la elaboración de los datos, las señales en bruto que se pueden recoger de los objetos o los fenómenos, para construir el conocimiento. Así pues la Visualización de la Información se define como el “Proceso de interiorización del conocimiento mediante la percepción de información” o, si se quiere mediante la elaboración de los datos. No se ha hecho referencia para nada a la visión o a ningún otro canal perceptivo en particular. Por tanto, para la visualización de información vale cualquier sentido, no sólo la vista; aunque hay que reconocer que es el sentido con más “ancho de banda”, es decir el que es capaz de aportar más datos a la mente por unidad de tiempo. Igualmente la elaboración de datos en forma de información, aunque enormemente facilitada por el ordenador, no requiere para nada de éste.

Los datos se han venido transformando en información durante toda la historia de la humanidad. Por ello el ordenador no figura tampoco en esta definición. El proceso que arranca en los datos, sigue con su elaboración en forma de información para constituir conocimiento y extraer de éste, finalmente la sabiduría. Así como los datos y la información se pueden traspasar, el conocimiento y, todavía más, la sabiduría requieren de la construcción de un cuerpo de experiencias y de una intuición que no son transferibles.

La visualización de la información, en tanto que la construcción de una imagen mental a partir de la información destilada de los datos y de la detección de patrones subyacentes a la información, incide plenamente en la formación del conocimiento.

En este contexto tanto la arquitectura de información como el diseño de información se pueden considerar como elementos fundamentales en el proceso de la visualización de

información, orientadas a producir la imagen mental, la chispa que enciende el fuego del conocimiento.

"El lenguaje natural es el sistema de símbolos más elaborado y universal del que disponemos" aunque esto no es siempre según Colin Ware, director del Data Visualization Research Lab, lo aprendemos desde pequeños y lo utilizamos constantemente [Ware, C. 1999].

Por ello la visualización es necesariamente un híbrido de imágenes y palabras. La decisión de cuándo y de qué forma utilizar unas u otras es una de las más importantes que ha de tomar el diseñador de visualización.

El extraordinario auge de la telefonía móvil nos permite hablar en todas partes y en todo momento. Hoy por hoy es difícil imaginar un sistema de comunicaciones más ubicuo y fácil de usar. La potencia del lenguaje reside en que con una sola palabra podemos evocar imágenes, sensaciones o experiencias completas vividas previamente. La información verbal y la información visual estimulan distintas zonas del cerebro. El lenguaje es esencialmente secuencial y dinámico.

Leer o explicar algo que no se conoce de antemano requiere un cierto tiempo. En contraste con ello "secciones relativamente grandes de imágenes estáticas y diagramas se pueden entender de un golpe"[Ware, C. 1999]:.

La descripción básica del ornitorrinco ocupa 85 palabras en una enciclopedia y, francamente, si no fuera por la foto no habría modo de aclararse. En cambio, para una persona que ha visto un ornitorrinco la mención de una palabra es suficiente para enlazar un conjunto de experiencias relacionadas con dicho animal. Es necesario utilizar ventajosamente ambos sistemas (verbal y visual) para mejorar la comprensión.

En el capítulo 9 del libro "*Information Visualization, Perception for Design*", se menciona [Ware, C. 1999]. Imágenes estáticas:

- Imágenes mejor que textos:

- »» Las imágenes son mejores para mostrar relaciones estructurales como enlaces entre entidades y grupos de entidades.
- »» Las tareas que involucran información de localización se representan mejor mediante imágenes
- »» La información visual se recuerda generalmente mejor que la información verbal siempre que no sea para imágenes abstractas.
- »» Las imágenes son mejores para suministrar detalle y apariencia. Ciertos estudios sugieren que primero asimilamos la forma y estructura general de un objeto y después los detalles.

- Texto mejor que imágenes:

- »» El texto es mejor que los gráficos para los conceptos abstractos como libertad o eficiencia.
- »» La información procedural, como los algoritmos de los programas de ordenador, se presentan mejor utilizando textos.
- »» Las imágenes estáticas por sí mismas no son efectivas para presentar instrucciones complejas no espaciales. No obstante hay excepciones como por ejemplo la planificación mediante diagramas de Gantt.
- »» La información que especifica condiciones bajo las cuales se ha de hacer o se ha de dejar de hacer algo se expresan mejor con texto que con imágenes.

Imágenes animadas:

- El trabajo se pueden representar mejor utilizando el movimiento y la animación:

- »» La animación permite representar efectivamente la causalidad. Con una animación apropiada una relación causal se puede percibir de forma directa e inequívoca.
- »» Los actos que expresan comunicación o flujo se explican mejor mediante animaciones, Por ejemplo, un acto de comunicación se puede representar mediante animaciones un símbolo (el mensaje) moviéndose de la fuente al receptor.
- »» La reorganización o reestructuración se adapta bien a la animación, siempre que la complejidad de las interacciones no sea muy alta. Una estructura se puede transformar gradualmente utilizando la animación haciendo explícitas las etapas de la reorganización.
- »» En visualización de software, una secuencia de movimientos en las estructuras de datos de un programa se puede representar mediante animación para mejorar la comprensión de su funcionamiento. Un ejemplo de ello, es la película *Sorting Out Sort* (Ordenando la Ordenación) que representa los movimientos de los datos en la memoria, para diferentes algoritmos de ordenación.
- »» Secuencias de acciones espaciales complejas se pueden representar de forma muy clara mediante la animación. Un ejemplo extraordinario esta en las instrucciones de montaje que vienen en el CD-ROM de los juguetes *Legó Mindstorms*. Permiten ver de forma animada y paso a paso como montar las piezas de los robots. Un gran trabajo.

Saber cuándo usar texto y cuando usar imágenes no es una tarea sencilla. En una visualización híbrida, además, ¿cómo mostrar el enlace entre texto e imágenes? [Ware,C.1999].

Para colocarlo en términos coloquiales, la visualización de la información puede resolver el problema de la indigesta generada por el exceso de información, a la hora de interpretar los datos, es más útil una representación abstracta que una metáfora reconocible.

Lo interesante de la tecnología destinada a la “digestión” inteligente de grandes volúmenes de información es el concepto, la idea de que hay detrás, más que la tecnología sobre la que se apoya. Los árboles hiperbólicos de Inxight,, se sustentan en la filosofía de foco y contexto, que permite tener en pantalla toda la información a la vez. Aquello que está en el foco de nuestra atención ocupa un espacio mucho mayor.

La empresa SEMIO se basa en la creación de una taxonomía (categorización por conceptos) flexible que se genera combinando diferentes técnicas de análisis lingüístico por ordenador. El resultado se visualiza con el SemioMap, que muestra el contenido de las agrupaciones de datos como una especie de Universo de estrellas y constelaciones de datos por donde puede moverse.

Cada una de estas entidades puede pulsarse y ver su contenido. En dicha web y previo envío de un e-mail se proporcionan un demo guiado del producto. En el MIT MediaLab , Flavia Sparacino trabaja, entre otros, en los proyectos Wearable City y Wearable Cinema. Wearable City es la versión de City of News, un Web Browser tridimensional que utiliza la metáfora de la ciudad para presentar la información en forma de rascacielos ordenados por distritos. Wearable Cinema es un sistema de realidad aumentada en la que la información audiovisual, por ejemplo, del contenido de un museo se superpone a la experiencia real dentro del museo. A medida que se pasa por las salas el sistema lo detecta y nos pasa las explicaciones audiovisuales convirtiendo ambas experiencias, real y virtual, en una sola.

Cartia utiliza la tecnología ThemeScape que permite la organización automática de colecciones de documentos basándose en la información que contienen, el resultado se visualiza en forma de mapa topográfico de terreno donde las elevaciones corresponden a las palabras o ítems más relevantes.

Todos estos sistemas representan sus datos basándose en ciertas metáforas. Las metáforas permiten relacionar ámbitos muy dispares que comparten estructuras o comportamientos similares. Nada más lejos de un escritorio real que el escritorio de Windows. Sin embargo la metáfora de las carpetas funciona porque pone en contacto la estructura jerárquica de los directorios y subdirectorios del sistema operativo (de comprensión y utilización no trivial) con la experiencia de las carpetas y archivadores del mundo real al que estamos acostumbrados.

Del mismo modo las metáforas que hay detrás de todas las técnicas antes mencionadas nos permiten poner en conexión complejos modelos matemáticos y lingüísticos con experiencias más apropiadas al común de los mortales. Visualizar la información no necesariamente quiere decir añadir gráficos al contenido. Especialmente en la Web algunas técnicas elementales ayudan a explicarse, principios básicos sobre cómo hacer que la información que se presenta en la web sea más visual, más adaptada al medio y por tanto más fácil de asimilar.

Existen aspectos fundamentales que hoy por hoy están bastante consensuados sobre lo que hay que tener en cuenta a la hora de visualizar información en la web[Redish, J 2001].

Hechos:

- El tiempo es un bien escaso.
- Los profesionales están demasiado ocupados para leer mucho, buscan respuestas rápidas a sus necesidades.
- Los usuarios de la web fundamentalmente ojean la información
- Incluso con los documentos en papel sólo un 15% lo leen completo y el 81% lo "sobrevuelan" de una forma u otra.
- En muchos casos se lee "para hacer" en vez de "leer para aprender", véase la figura 6.

- El contenido de muchos sitios web no encaja con lo que los usuarios buscan.

Para adaptar la creación y presentación de contenido a la web se propone las técnicas siguientes:

1. Da a los usuarios sólo lo que necesitan con las mínimas palabras posibles. Para ello destila la esencia del mensaje sin perder el significado. Morkes y Nielsen redujeron el texto de un sitio web al 54 % y los usuarios lo encontraron más completo.

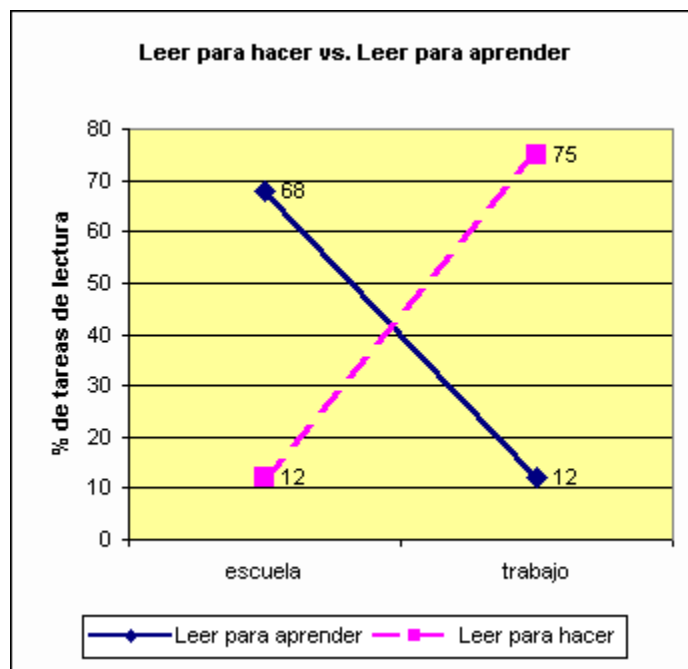


FIGURA 6 Estudio de la lectura hecha [Redish, J 2001]

2. Huye de la prosa.

En vez de ello utilizar la "escritura visual", que incluye el uso de

- » enlaces,
- » fragmentos,
- » gráficos,
- » listas y
- » tablas.

3. Si se ha de usar la prosa escribir sentencias y párrafos cortos. Una sentencia por párrafo mejor que dos.
4. Apilar la información en capas. Algunos usuarios necesitan sólo un mordisco, otros un aperitivo y otros la comida entera. Usar enlaces para separar las capas y dar a cada usuario el "alimento que necesita"
5. Utilizar listas.
 - »» Listas con viñetas para elementos individuales o posibles selecciones.
 - »» Listas numeradas para instrucciones o procedimientos.
6. Usar tablas
 - »» cuando hay que comparar números cuando haya que usar sentencias condicionales del tipo "si A entonces B" o
 - »» para relacionar resultados con operaciones, por ejemplo, "para conseguir A hacer B".
7. Pensar visualmente.
 - »» Considerar los párrafos como si fueran imágenes.
 - »» Combinar los espacios en blanco con los textos y las imágenes para conseguir que el significado del mensaje que se pretende hacer llegar quede claro de la forma más sencilla posible.
8. Usar los gráficos para ahorrar espacio si con ello se hace la información más accesible y fácil de entender. Por ejemplo los mapas de provincias o países pueden funcionar mejor que una lista con los nombres.
9. Utilizar los gráficos como ejemplos en sitios informacionales y de comercio electrónico.

Por ejemplo, en la tabla No. 3 se muestra un comparativo en el diseño de textos.

<p>SZWXY S.L. en su constante búsqueda de la rentabilidad ha facturado en el año 2000 3,27 Millones de Euros, alcanzando de esta forma una rentabilidad del 7% gracias a sus beneficios antes de impuestos de 230.000 euros. La plantilla de la empresa es de 420 empleados, altamente motivados en la consecución de los objetivos de la empresa que se centran en el diseño de sistemas de visualización para corporaciones industriales con grandes bases de datos, entidades financieras y portales de B2B.</p>	<p>SZWXY S.L. produce sistemas de visualización para:</p> <ul style="list-style-type: none">• corporaciones industriales con grandes bases de datos,• entidades financieras y• portales de B2B. <table border="1"><tr><td colspan="2">Año 2000</td></tr><tr><td>Facturación:</td><td>3.270.000 €</td></tr><tr><td>Beneficios (BAI):</td><td>230.000 €</td></tr><tr><td>Rentabilidad:</td><td>7%</td></tr><tr><td>Num. empleados:</td><td>420</td></tr></table>	Año 2000		Facturación:	3.270.000 €	Beneficios (BAI):	230.000 €	Rentabilidad:	7%	Num. empleados:	420
Año 2000											
Facturación:	3.270.000 €										
Beneficios (BAI):	230.000 €										
Rentabilidad:	7%										
Num. empleados:	420										

Tabla 3 Diseños con el mismo contenido, diferente visualización [Redish, J 2001]

2.3.1 Visualizando la web semántica

La visualización de información se beneficia de los avances de la semántica web. La semántica web, liderada por Tim Berners-Lee permite que la información residente en la red sea accesible y “comprensible” no sólo por los humanos sino también por las máquinas.

Algunas empresas están empezando a utilizar el contenido semántico (el significado de sus contenidos) que se ha empezado a codificar en las sedes web para mejorar sus procesos de búsqueda de información relevante. Ejemplos de ello son Amazon.com o Yahoo shopping.

Pero el contenido semántico de una web se puede utilizar ventajosamente para la visualización de su información, ya que disponemos entonces del conocimiento sobre su significado y las relaciones existentes entre los conceptos que se manejan en la misma.

Entre las empresas que trabajan en este sentido en Europa destaca Administrator, una empresa holandesa dedicada a la ingeniería del conocimiento que dispone de un interesante sistema de visualización basado en el contenido semántico (la ontología) de una sede web.

La aproximación que utiliza Administrator es la de considerar que en cualquier procedimiento de visualización de información se parte de una materia prima, los datos que después se organizan, filtran y clasifican hasta constituir una estructura semántica que la dota de un significado coherente. Conseguido esto la estructura obtenida sirve de base para proceder a la visualización.

En la figura número 7 se observa el proceso de visualización, partiendo de la materia prima como en cualquier proceso, los datos se tabulan, filtran, seleccionan y estructuran, hasta obtener un contenido semántico coherente que permita la visualización utilizando una metáfora visual conveniente.

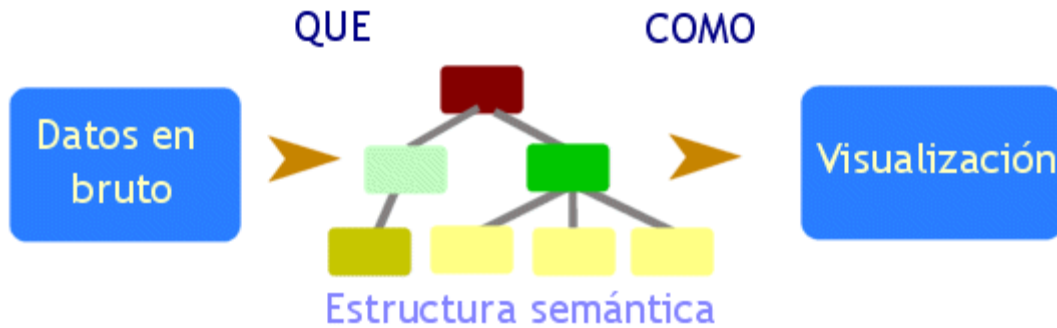


FIGURA 7 Esquema basado en el diseño de sitios web [Van Harmelen 2001]

Como ellos mismos indican en su artículo “Ontology based Information Visualisation” [Van Harmelen 2001] este es un proceso genérico que se usa “desde para preparar una clase hasta diseñar un sitio web”, primero se decide qué mostrar (la estructura) y luego cómo hacerlo (la visualización). Lo que hace el sistema de Administrator es representar la ontología codificada en una sede web (véanse las imágenes adjuntas) como un sistema de etiquetas y esferas conectadas por líneas.

- Cada etiqueta representa una clase en la ontología (una categoría de páginas por ejemplo, deportes)
- Cada esfera representa una instancia de una clase (cada página de esa categoría).
- Las líneas indican que una instancia es miembro de una clase o que una clase es una subclase de otra (por ejemplo fútbol es miembro de deportes y deportes podría ser subclase de entretenimiento).

Todos estos elementos se disponen en el espacio mediante un sistema de atracciones y repulsiones (como si hubiera unos muelles dentro de las líneas) entre los objetos que los mantiene en un equilibrio dinámico que hace que los objetos cercanos semánticamente estén juntos y los objetos lejanos semánticamente estén separados.

En este contexto, "Semánticamente cercano" quiere decir que dos clases comparten muchas instancias o bien que dos instancias pertenecen a la misma clase, podemos observar un ejemplo en la fig. 8.

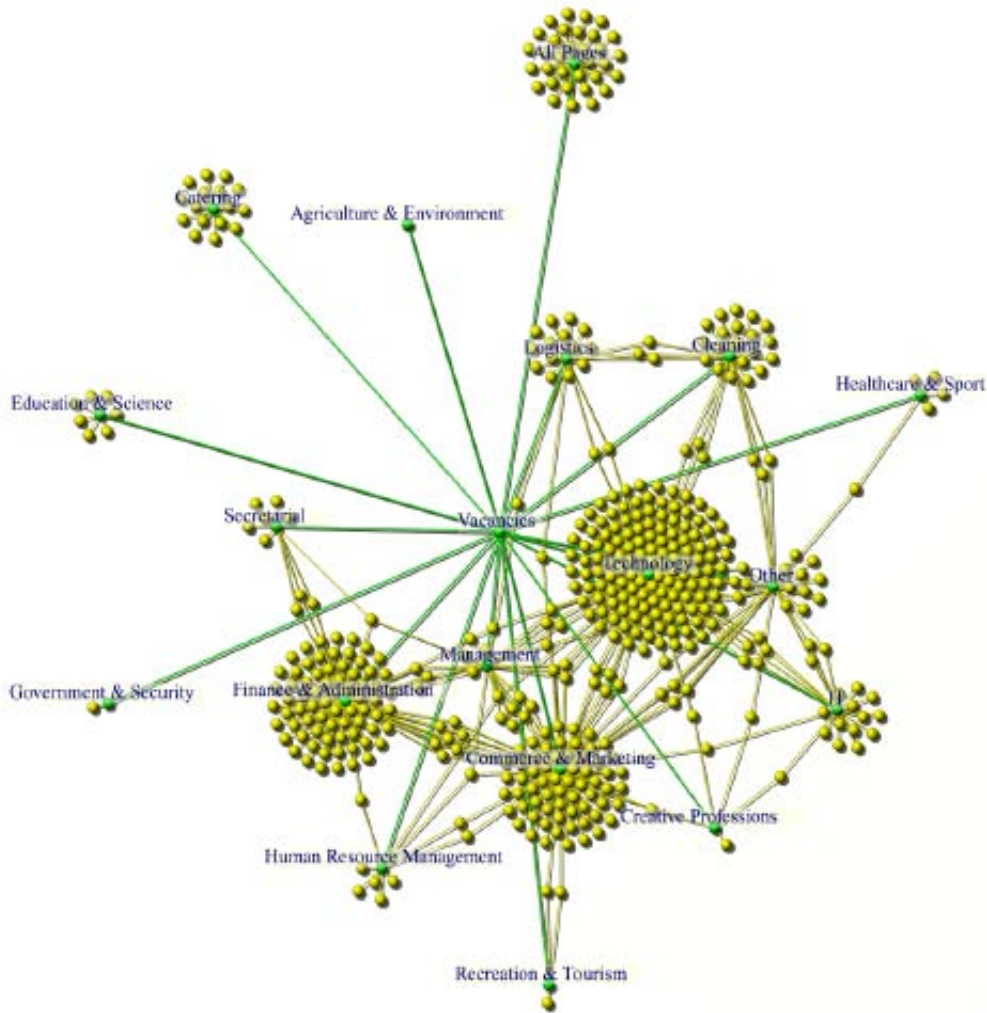


FIGURA 8 Visualización de sitio web [Van Harmelen 2001]

Por ejemplo, la figura 9 representa el contenido de una web holandesa de búsqueda de trabajo. La ontología clasifica los trabajos por sector económico: ocio, finanzas, educación, etc.

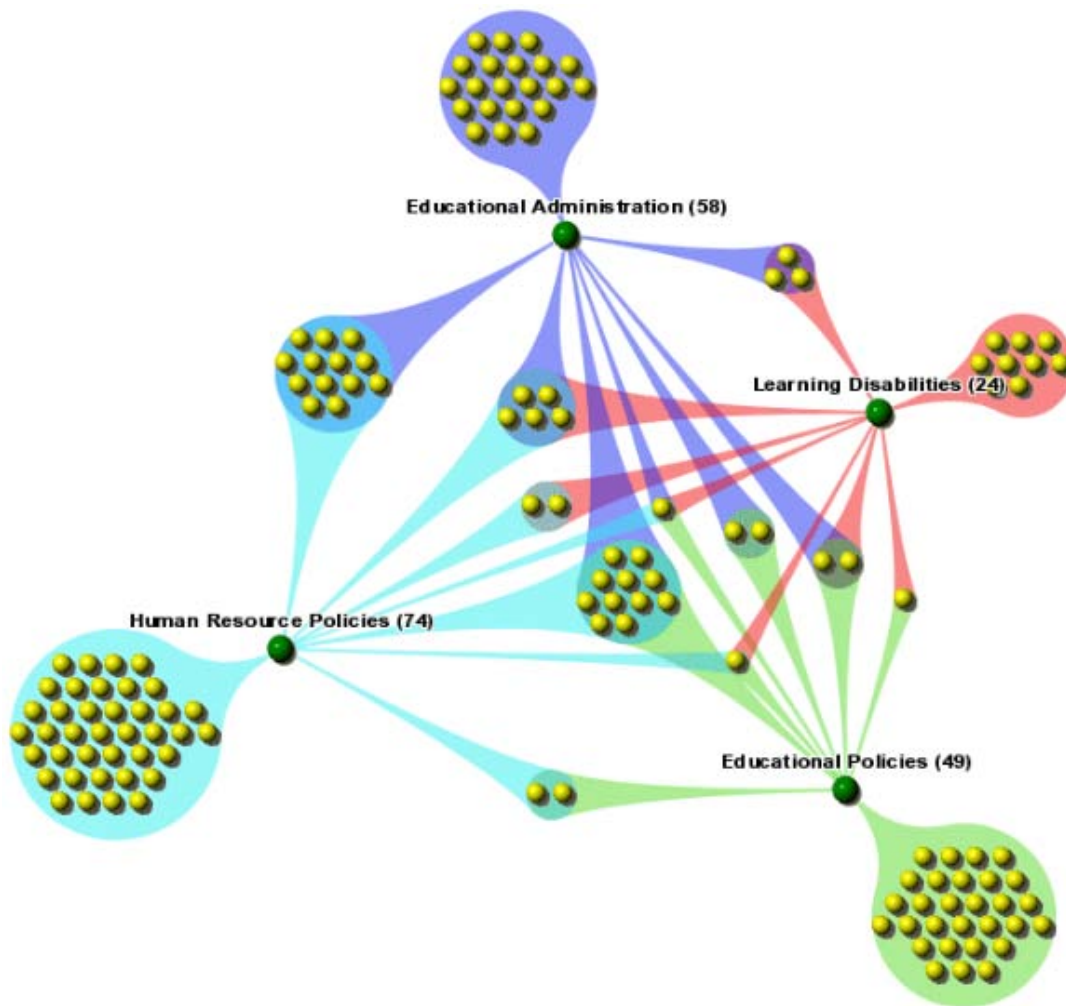


FIGURA 9 Algoritmo aplicado a una base de datos educacional [Van Harmelen 2001]

La visualización de esta ontología mediante la metáfora visual de las esferas conectadas por “muelles” permite enseguida detectar ciertas incidencias.

- El tamaño de las clases queda inmediatamente de relieve por la aglomeración de esferas.
- Se puede ver también que muchos trabajos están clasificados en más de una clase, por ejemplo uno de ellos pende de finanzas, gestión y secretariado, con mayor cercanía a este último, lo que da una idea bastante clara sobre si es de nuestro interés o no.
- Finalmente podemos ver que las clases que están en lugares opuestos del diagrama no contienen ningún miembro en común (gobierno y seguridad vs. Salud y deporte).

El objetivo es utilizar esta técnica de visualización para:

- Soportar la navegación en sitios web u otras colecciones de documentos.
- Posiblemente utilizando múltiples vistas de los mismos datos.
- Proporcionar vistas de conjunto de conjuntos de datos para tareas de análisis, posiblemente en el tiempo (ver cómo cambia un mapa en el tiempo).
- Soportar la interrogación (en motores de búsqueda o bases de datos regulares) mediante la visualización de la estructura de un conjunto de resultados en una forma más inteligente que simplemente ojeando una lista o una tabla.

La web semántica empieza su despliegue tímidamente. Al igual que para otros campos, el conocimiento semántico ayudará enormemente a visualizar la información.

2.4 Minería visual de datos

El punto de visualización de datos es permitir que el usuario entienda lo que sucede. Los datos normalmente implican la extracción de la minería oculta, por lo general hay muchas maneras de representar gráficamente un modelo, las visualizaciones que se usan deben elegirse para maximizar el valor al espectador, asumiendo que el espectador es un experto en la materia, pero no en el modelado de datos, entonces es necesario traducir el modelo en una representación más natural para él.

Para ello se sugiere la utilización de los principios de orientación como un modelo para la visualización [Fayyad 2001].

1. Mostrar las distancias pequeñas en el modelo en donde cerca de las rutas no dan lugar a fines diferentes.
2. Mostrar, la demanda, el efecto de diferentes perspectivas, (el cambio o inclusión de probabilidades) sobre el modelo de estructura.

3. Hacer los cambios dinámicos en el tramado dinámico.
4. Puntualizar relaciones conocidas (de referencia) en todo el modelo de paisaje.
5. Permitir la interacción que proporcionan los detalles de las preguntas y respuestas sobre la demanda.

Las ventajas de este enfoque múltiple, incluye la habilidad de explorar en forma óptima algunos caminos, con la capacidad de reducir los modelos independientes de coordinar un conjunto y medir la conveniencia del modelo de forma natural.

La fuerza motriz detrás de la visualización de modelos de minería de datos puede desglosarse en dos ámbitos fundamentales: la comprensión y la confianza. Entendiendo sin duda que la motivación fundamental detrás de la visualización es el modelo. Aunque la forma más sencilla de hacer frente a un modelo de minería de datos es dejar la salida en forma de un cuadro negro, el usuario no necesariamente obtiene comprensión de las conductas en las que está interesado.

Si toman el modelo de la caja negra como resultado de una base de datos, puede obtenerse una lista de clientes y ocuparla comercialmente, no hay mucho para el usuario que pueda hacer sino sentarse y ver los sobres salir para el envío a sus clientes sin reducir significativamente la tasa de respuesta.

La forma más interesante de utilizar un modelo de minería de datos es conseguir que el usuario realmente entienda lo que está pasando para que pueda tomar medidas directamente. La visualización de un modelo que permita a los usuarios debatir y explicar la lógica detrás del modelo con colegas, clientes y otros usuarios.

Obtener financiamientos con lógica o razón es parte de la construcción de la confianza para los usuarios en los resultados.

Las decisiones acerca de dónde colocar publicidad de dólares son el resultado directo de la comprensión de modelos de minería de datos en el comportamiento de los clientes [Fayyad 2001]. Todo en la cabeza del director de marketing a menos que la producción de la minería de datos del sistema pueda entenderse cualitativamente, no será de ninguna utilidad. El modelo debe ser entendido de manera que las acciones tomadas sean el resultado que pueda justificar a los demás.

Entendimiento significa algo más que comprensión, también implica el contexto, si el usuario puede entender lo que se ha descubierto en el contexto de su empresa, logrará la confianza y el éxito. Existen dos partes en el problema:

1. La visualización de la salida de la minería de datos
2. Y que el usuario permita interactuar con la simple visualización de modo que las preguntas puedan ser contestadas.

Soluciones creativas para la primera parte han sido incorporadas a una serie de productos comerciales como MineSet. Graphing impulsa responde y da indicadores financieros proporcionando al usuario un sentido de contexto en el que puede alcanzarse rápidamente los resultados en el terreno real. Después de eso simples representaciones de los resultados de minería de datos permiten a los usuarios conocer los resultados de la minería de datos.

Graficar y desplegar arboles de decisión para cambiar significativamente la forma en que la minería de datos se utiliza en un programa. La interacción es para muchos usuarios el Santo Grial de la visualización de datos en la minería, esto y la visualización dinámica de los resultados permite al usuario obtener una idea de la dinámica y comprobar si algo realmente contrai-intuitivo esta ocurriendo.

Al ver un árbol de decisiones es bueno, pero lo que realmente se quiere es arrastrar y soltar los ejores segmentos en un mapa con el fin de ver si hay datos que se descuidan [Fayyad 2001].

La interacción continuará hasta que el usuario entienda lo que pasa con sus clientes. A menudo se tiene el deseo de perforar a través de modo que se puedan ver los datos reales detrás de un modelo, aunque probablemente es más una cuestión de percepciones que una realidad útil.

Por último la integración con otras herramientas de apoyo a la decisión permite a los usuarios ver los resultados de minería de datos de manera que se utilicen para el fin de comprender el comportamiento del cliente. Mediante la incorporación de la interacción en el proceso el usuario podrá conectarse a la minería de datos con los resultados de los clientes [Fayyad 2001].

2.4.1 Descubriendo los patrones secuenciales.

Aunque diferentes aplicaciones tienen varias definiciones de patrones secuenciales, todas comparten una intención básica para encontrar los patrones repetidos en eventos discretos en una línea del tiempo.

Se presenta un ejemplo de visualización simple, para destacar las fortalezas y debilidades de utilizar la visualización para descubrir patrones secuenciales. La figura 10 muestra una trama de diferentes tópicos con respecto al tiempo. Si se reemplaza la combinación de tópicos con un tópico individual se puede encontrar la secuencia de patrones de este por comparación de iconos individuales en columnas a lo largo de la línea del tiempo [Wong, P.C 2000].

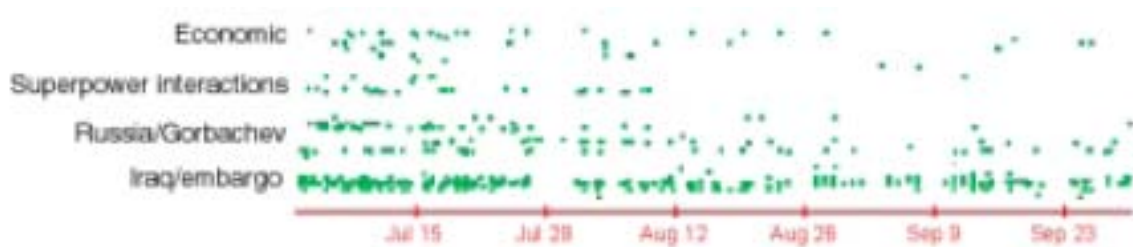


FIGURA 10 Trama de diferentes tópicos de Julio a Septiembre 1990 [Wong, P.C2000]

La ventaja de acercarse a la visualización es que se puede obtener rápidamente una impresión general de la estructura al ver el patrón de los tópicos y sus distribuciones. No tan solo se pueden ver las frecuencias de los patrones sino también la periodicidad de los datos en eventos individuales. Como una gran desventaja se presenta la precisión de los patrones. No se puede saber exactamente cuál es la conexión entre los patrones.

Otra desventaja es la falta de soporte estadístico en los patrones individuales, porque solo se puede visualizar una parte en el tiempo, esto hace imposible la predicción de la distribución y la concentración de los patrones cuando no se encuentran desplegados. Sin embargo estos problemas se resuelven con la minería de datos. El descubrimiento de patrones secuenciales juega un importante rol hoy en día, para la minería de datos en la industria. Personalizar la minería de textos para construir patrones estructurales usando arboles binarios con n-ramas, también se sabe que es un caso de estudio para los algoritmos computacionales.

Cada nodo del árbol representa un elemento o tópico en este caso para la secuencia de patrones. El patrón es válido si esta soporta por un valor largo predefinido por un umbral. El soporte de este valor es calculado como un número de precisión de patrones inicial. Los elementos de los patrones son una secuencia de entradas, una ruptura en la los datos de entrada representa una falla en las reglas. En la figura 11 se representa un ejemplo básico de minería secuencial de patrones del caso con tópicos en el dominio de A,B y C, con un periodo de 6 días.

La ventaja de la minería visual de datos es que se puede usar para compensar las debilidades en el descubrimiento de patrones secuenciales.

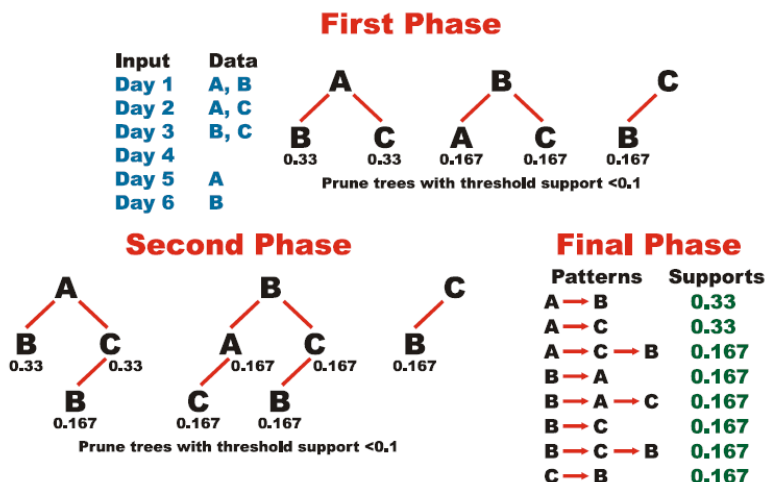


FIGURA 11 Minería secuencial de patrones en el tiempo [Wong, P.C 2000]

Un sistema de minería visual de datos es para visualizar el incremento de procesos de minería y tomar ventaja de las debilidades del descubrimiento de patrones secuenciales.

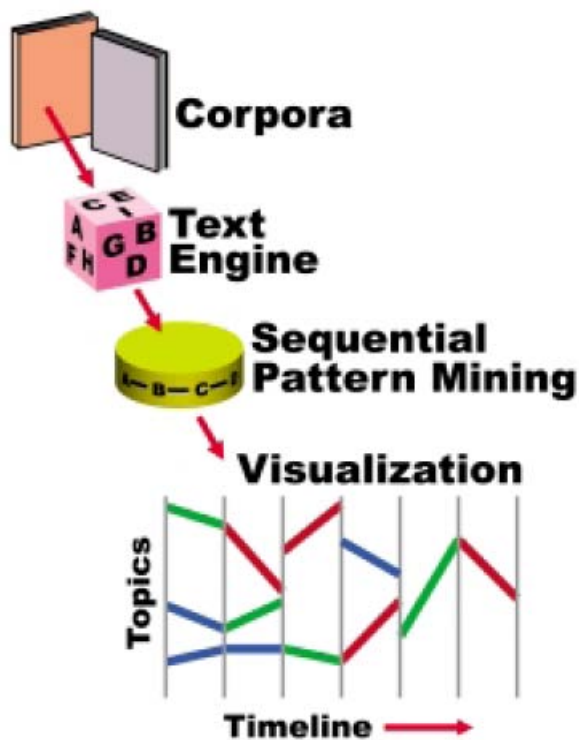


FIGURA 12 Sistema de minería visual de datos [Wong,P.C 2000]

La figura 12 muestra una introducción visual de alto nivel para un sistema de minería visual de datos en el descubrimiento de patrones secuenciales. Al aplicar técnicas de minería visual de datos se puede observar un incremento drástico en el número de patrones secuenciales. Utilizar minería de datos visual es ambicioso porque se puede aprovechar la tecnología computacional, el almacenamiento de datos, los algoritmos computacionales, utilizar la lógica para planear, diagnosticar y predecir, relacionándolo con las habilidades humanas para lograr la percepción la creatividad y la generación del conocimiento, todas estas características las observamos en la figura 13. Los métodos tradicionales de minería de datos generan resultados precisos cuando el problema involucra estadísticas.

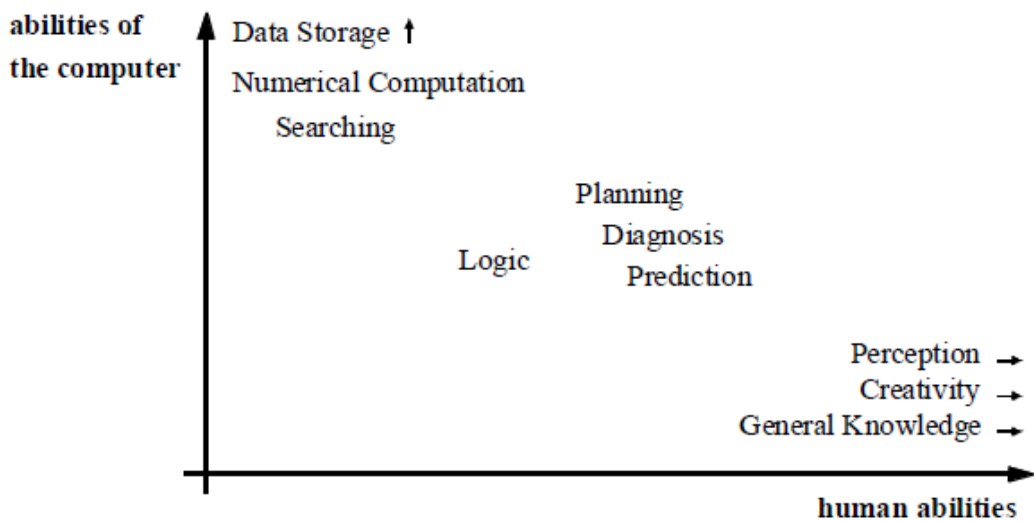


FIGURA 13 Características base de la minería visual de datos [Keim D. A. 2002]

La visualización de la información es intuitiva, el usuario se involucra de forma directa en el proceso de exploración y se aplica cuando se conocen pocos datos o las metas no están definidas.

La minería visual de datos nos aproxima y toma ventajas de ambas herramientas, el poder realizar cálculos de forma automática y las capacidades de la percepción humana para extraer las estructuras de un gráfico. En la figura 14 se muestra un ejemplo claro basado en una lista de llamadas vía teléfono celular.

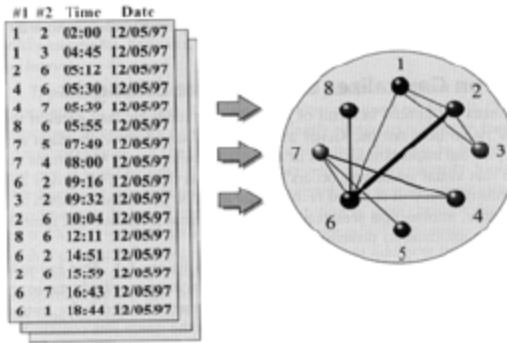


FIGURA 14 Representaciones para estructura de llamadas [Keim D. A. 2002]

Se requiere extraer datos, representar objeto de información en un proceso sistemático de la misma. Determinar, los atributos, establecer cuáles son las características que proporcionan información de los objeto. Información del espacio por atributos, cada uno de estos representa una dimensión en la información espacial. Definir relaciones para fijar consideraciones y relaciones entre los objetos. Estos conceptos podemos observarlos en la figura 15 en donde cada punto representa un objeto e información específica, los bordes determinan la estructura de la información contenidos en un espacio de información

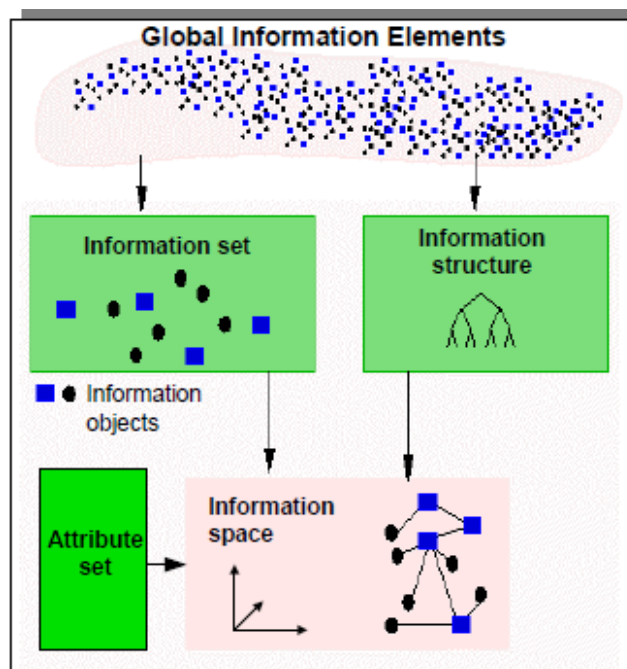


FIGURA 15 Diferentes representaciones gráficas [Keim D. A. 2002]

2.5 Árboles de decisión

Hasta hace no demasiado tiempo se utilizaba el término “procesamiento de datos” para describir la utilización de los ordenadores en distintos ámbitos. Hoy se utiliza otro término, IT [Information Technology] que se refiere a lo mismo pero implica un cambio de enfoque. Se hace énfasis no únicamente en el procesamiento de grandes cantidades de datos, sino en la extracción de información significativa de esos datos.

Los datos son información cruda, colecciones de hechos que deben ser procesados para que sean significativos. La información se obtiene asociando hechos (en un contexto determinado). El conocimiento utiliza la información obtenida en un contexto concreto y la asocia con más información obtenida en un contexto diferente. Finalmente, la sabiduría aparece cuando se obtienen principios generales de fragmentos de conocimiento.

Hasta ahora, la mayor parte del software ha sido desarrollado para procesar datos, información a lo sumo. En el futuro se trabajará con sistemas que procesen conocimiento. La clave reside en asociar elementos con información proveniente de distintas fuentes y sin conexión obvia de tal forma que la combinación nos proporcione beneficios. Este es uno de los desafíos más importantes de la actualidad: la construcción de sistemas que extraigan conocimiento de los datos de forma que sea práctico y beneficioso.

El aprendizaje en Inteligencia Artificial se entiende como un proceso por el cual un ordenador acrecienta su conocimiento y mejora su habilidad. En él se resaltan dos aspectos complementarios: el refinamiento de la habilidad y la adquisición de conocimiento. El aprendizaje denota cambios en el sistema que son adaptativos en el sentido de que permiten al sistema hacer la misma tarea o tareas a partir de la misma posición más eficiente y/o efectivamente la siguiente vez.

Muchas de las técnicas de aprendizaje usadas en IA están basadas en el aprendizaje realizado por los seres vivos. Para ellos la experiencia es muy importante, ya que les permite no volver a cometer los mismos errores una y otra vez.

Además, la capacidad de adaptarse a nuevas situaciones y resolver nuevos problemas es una característica fundamental de los seres inteligentes. Por lo tanto, podemos aducir varias razones de peso para estudiar el aprendizaje: en primer lugar, como método de comprensión del proceso de aprendizaje y, en segundo término, aunque no por ello sea menos interesante, para conseguir programas que aprendan (desde una perspectiva más propia de la Inteligencia Artificial).

Una primera clasificación de las técnicas de aprendizaje existentes se puede realizar atendiendo a la filosofía seguida en el proceso de adquisición del conocimiento:

- ❑ En el aprendizaje supervisado, los ejemplos de entrada van acompañados de una clase o salida correcta. Esta familia de técnicas engloba al aprendizaje memorístico [rote learning], a los modelos de aprendizaje por ajuste de parámetros y a los algoritmos de aprendizaje inductivo.

- ❑ En el aprendizaje no supervisado (aprendizaje por observación, sin profesor) se construyen descripciones, hipótesis o teorías a partir de un conjunto de hechos u observaciones sin que exista una clasificación a priori de los ejemplos. Este tipo de aprendizaje es el que realizan los métodos de agrupamiento o clustering.

El aprendizaje supervisado, también conocido como clasificación, es uno de los principales problemas en Inteligencia Artificial. En concreto, el objetivo de cualquier algoritmo de aprendizaje inductivo es construir un modelo de clasificación a partir de un conjunto de datos de entrada, denominado conjunto de entrenamiento, que contiene algunos ejemplos de cada una de las clases que intentamos modelar. Los casos del conjunto de entrenamiento incluyen, además de la clase a la que corresponden, una serie de atributos o características que se utilizarán para construir un modelo abstracto de clasificación: el objetivo del aprendizaje supervisado es la obtención de una descripción precisa para cada clase utilizando los atributos incluidos en el conjunto de entrenamiento.

El modelo que se obtiene durante el proceso de aprendizaje puede utilizarse para clasificar nuevos ejemplos (casos cuyas clases se desconozcan) o, simplemente, para comprender mejor la información de la que disponemos. Un modelo de clasificación puede construirse entrevistando a expertos, tal como se suele hacer para construir muchos sistemas basados en conocimiento [SBC en castellano o KBS en inglés] a pesar de la dificultad que entraña la extracción manual del conocimiento. Por desgracia, cuanto mejor es el experto peor suele describir su conocimiento (la paradoja de la Ingeniería del Conocimiento). Además, los expertos en un tema no siempre están de acuerdo (Ley de Hiram: “Si se consultan suficientes expertos, se puede confirmar cualquier opinión”).

No obstante, si se dispone de suficiente información registrada (almacenada en una base de datos, por ejemplo), el modelo de clasificación se puede construir generalizando a partir de ejemplos específicos mediante algún proceso inductivo automático.

De hecho, podemos encontrar numerosos ejemplos de algoritmos de aprendizaje inductivo en la construcción de árboles de decisión. Los casos de entrenamiento utilizados en la construcción del modelo de n suelen expresarse en términos de un conjunto finito de propiedades o atributos con valores discretos o numéricos, mientras que las categorías a las que han de asignarse los distintos casos deben establecerse de antemano (al tratarse de aprendizaje supervisado). En general, estas clases serán disjuntas (aunque pueden establecerse jerarquías) y deberán ser discretas (para predecir atributos con valores continuos se suelen definir categorías discretas utilizando términos imprecisos propios del lenguaje natural). Las técnicas inductivas de clasificación se basan en el descubrimiento de patrones en los datos de entrada.

Para que el aprendizaje inductivo sea correcto, entendiendo éste como un proceso de generalización a partir de ejemplos concretos, hemos de disponer de suficientes casos de entrenamiento (bastantes más que clases diferentes). Si las conclusiones obtenidas no están avaladas por bastantes ejemplos, entonces la aparición de errores en los datos podría conducir al aprendizaje de un modelo erróneo que no resultaría fiable.

Por tanto, cuantos más datos obtengamos más fácilmente podremos diferenciar patrones válidos de patrones debidos a irregularidades o errores. Si suponemos que todos los patrones a reconocer son elementos potenciales de J clases distintas denotadas W_j , llamaremos $\Omega = \{W_j | 1 < j < J\}$ al conjunto de las clases informacionales. En determinadas ocasiones extenderemos Ω con una clase de rechazo W_0 a la que asignaremos todos aquellos casos para los que no se tiene una certeza aceptable de ser clasificados correctamente en alguna de las clases de Ω . De este modo, denotamos $\Omega^* = \Omega \cup \{W_0\}$ al conjunto extendido de clases informacionales. Un clasificador o regla de clasificación es una función $d : P \rightarrow \Omega^*$ definida sobre el conjunto de posibles ejemplos P tal que para todo ejemplo X se verifica que $d(X) \in \Omega^*$.

Los árboles de decisión, clasificación o identificación constituyen uno de los modelos más destacados de aprendizaje supervisado. Su principal virtud radica en que son modelos de clasificación de fácil comprensión. Su dominio de aplicación no está restringido a un ámbito concreto sino que pueden utilizarse en diversas áreas, desde aplicaciones de diagnóstico médico hasta juegos como el ajedrez o sistemas de predicción meteorológica. Los algoritmos de construcción de árboles de decisión suelen construir de forma descendente los árboles de decisión, comenzando en la raíz del árbol. Por este motivo se suele hacer referencia a este tipo de algoritmos como pertenecientes a la familia TDIDT [Top-Down Induction of Decision Trees] que obtienen modelos de clasificación excelentes a la vez que pueden trabajar adecuadamente con los enormes conjuntos de datos que suelen utilizarse para resolver problemas de extracción de conocimiento en bases de datos [KDD: Knowledge Discovery in Databases] y minería de datos.

Los árboles de decisión constituyen probablemente el modelo de clasificación más popular y utilizado. El conocimiento obtenido durante el proceso de aprendizaje inductivo se representa mediante un árbol en el cual cada nodo interno contiene una pregunta acerca de un atributo particular (con un nodo hijo para cada posible respuesta) y en el que cada hoja se refiere a una decisión (etiquetada con una de las clases del problema).

Un árbol de decisión puede utilizarse para clasificar un ejemplo concreto comenzando en su raíz y siguiendo el camino determinado por las respuestas a las preguntas de los nodos internos hasta que se llega a una hoja del árbol. Su funcionamiento es análogo al de una aguja de ferrocarril: cada caso es dirigido hacia una u otra rama de acuerdo con los valores de sus atributos al igual que los trenes cambian de vía según su destino (las hojas del árbol) en función de las agujas de la red de ferrocarriles (los nodos internos). Los árboles de clasificación son útiles siempre que los ejemplos a partir de los que se desea aprender se puedan representar mediante un conjunto prefijado de atributos y valores, ya sean estos discretos o continuos. Sin embargo, no resultan demasiado adecuados cuando la estructura de los ejemplos es variable.

Tampoco están especialmente indicados para tratar con información incompleta (cuando aparecen valores desconocidos en algunos atributos de los casos de entrenamiento) y pueden resultar problemáticos cuando existen dependencias funcionales en los datos del conjunto de entrenamiento (cuando unos atributos son función de otros).

En principio, se busca la obtención de un árbol de decisión que sea compacto. Un árbol de decisión pequeño no nos permite comprender mejor el modelo de clasificación obtenido y, además, es probable que el clasificador más simple sea el correcto, de acuerdo con el principio de economía de Occam : “los entes no han de multiplicarse innecesariamente”. Este principio, si bien permite la construcción de modelos fácilmente comprensibles, no garantiza que los modelos así obtenidos sean mejores que otros aparentemente más complejos.

Por desgracia, no podemos construir todos los posibles árboles de decisión derivados de un conjunto de casos de entrenamiento para quedarnos con el más pequeño. La construcción de un árbol de decisión a partir del conjunto de datos de entrada suele realizarse de forma descendente mediante algoritmos greedy de eficiencia de orden $O(n \log n)$, siendo n el número de ejemplos incluidos en el conjunto de entrenamiento.

Los árboles de decisión se construyen recursivamente siguiendo una estrategia descendente, desde conceptos generales hasta ejemplos particulares. Esa es la razón por la cual el acrónimo TDIDT, que proviene de “Top-Down Induction on Decision Trees”, se emplea para hacer referencia a la familia de algoritmos de construcción de árboles de decisión.

Una vez que se han reunido los datos que se utilizarán como base del conjunto de entrenamiento, se descartan a priori aquellos atributos que sean irrelevantes utilizando algún método de selección de características y, finalmente, se construye el árbol de decisión recursivamente. El método de construcción de árboles de decisión mediante particionamiento recursivo del conjunto de casos de entrenamiento tiene su origen en el trabajo de Hunt a finales de los años 50.

Este algoritmo “divide y vencerás” es simple y elegante:

- ❑ Si existen uno o más casos en el conjunto de entrenamiento y todos ellos corresponden a objetos de una misma clase $c \in \text{Dom}(C)$, el árbol de decisión es una hoja etiquetada con la clase c . Hemos alcanzado un nodo puro.
- ❑ Si no encontramos ninguna forma de seguir ramificando el árbol o se cumple alguna condición de parada (regla de parada), no se sigue expandiendo el árbol por la rama actual. Se crea un nodo hoja etiquetado con la clase más común del conjunto de casos de entrenamiento que corresponden al nodo actual. Si el conjunto de casos de entrenamiento queda vacío, la clasificación adecuada ha de determinarse utilizando información adicional .
- ❑ Cuando en el conjunto de entrenamiento hay casos de distintas clases, éste se divide en subconjuntos que sean o conduzcan a agrupaciones uniformes de casos (instancias de una misma clase). Utilizando los casos de entrenamiento disponibles, hemos de seleccionar una pregunta para ramificar el árbol de decisión.

Dicha pregunta, basada en los valores que toman los atributos predictivos en el conjunto de entrenamiento, ha de tener dos o más respuestas alternativas mutuamente exclusivas R_i . Generalmente, el árbol se ramificará s evaluando el valor de algún atributo concreto. De todas las posibles alternativas, se selecciona una empleando una regla heurística a la que se denomina regla de división. El árbol de decisión resultante consiste en un nodo que idéntica la pregunta realizada del cual cuelgan tantos hijos como respuestas alternativas existan. El mismo método utilizado para el nodo se utiliza recursivamente para construir los subárboles correspondientes a cada hijo del nodo, teniendo en cuenta que al hijo H_i se le asigna el subconjunto de casos de entrenamiento correspondientes a la alternativa R_i .

En resumen, cuando se construye un nodo se considera el subconjunto de casos de entrenamiento que pertenecen a cada clase (estadísticas del nodo). Si todos los ejemplos pertenecen a una clase o se verifica alguna regla de parada, el nodo es una hoja del árbol. En caso contrario, se selecciona una pregunta basada en los atributos predictivos del conjunto de entrenamiento (usando una regla de división heurística), se divide el conjunto de entrenamiento en subconjuntos (mutuamente excluyentes siempre y cuando no existan valores desconocidos) y se aplica el mismo procedimiento a cada subconjunto.

3 CASO DE ESTUDIO

3.1 Método de Investigación

La presente se basa principalmente en investigación documental realizada en artículos de publicaciones científicas y del conocimiento de la propuesta hecha en uno de ellos con la herramienta de visualización de la web WET aplicado a la página administrada por uno de los autores de la misma herramienta.

La fase de resolución corresponde a la creación de la hipótesis en un método empírico o a la elaboración de la teoría en un método deductivo. De este modo la verificación equivale a la validación del método, mientras que la verificación de la teoría en un método deductivo equivale a la comprobación de la consistencia del método, es decir a la verificación de las tareas expuestas [E Marcos 2002]:

- ☉ Sobre la representación de la estructura es posible superponer información referente al contenido y/o a su utilización, representando una serie de métricas, como pueden ser el ranking de Google de cada página, o el número de descargas, número de enlaces que entran o salen etc., mediante su asociación con ciertas variables visuales.
- ☉ Variables visuales:
 - El tamaño que modifica el radio de cada nodo haciéndolo proporcional a alguna de las métricas o variables que queramos representar.
 - El color de relleno que puede ser proporcional o estar en relación con otra métrica.
 - El color del borde y/o
 - La forma del nodo que puede ser diferente de la circular y adoptar el hexágono, cuadrado, triángulo, etc..

Esto permite cruzar en cada nodo información multidimensional y encontrar nodos que cumplen conjuntos de requisitos más o menos complejos.

En WET cada metáfora visual corre en su propia ventana que permite hacer zoom, mover etc. Dado que sobre el mapa de arboles se centra en el árbol radial y en la representación de la estructura contenido y uso sobre el mismo. La visualización de la estructura es por si sola de gran interés. Cualquier jerarquía suficientemente grande se hace inabarcable rápidamente para quien la ha creado. La web es un ejemplo claro. Mediante la visualización de la estructura es posible localizar errores rápidamente.

3.2 Justificación de los métodos

Para los fines de la presente investigación se dividió el estudio en un análisis de teorías y prácticas utilizadas en minería visual de datos aplicados a la estructura de la web.

Posteriormente se retomaron y enfocaron estas teorías y prácticas al analizar un caso de estudio para describir el comportamiento de los sistemas existentes. El caso de estudio se utilizo para capturar los requisitos de comportamiento de las herramientas detallando un escenario hilo conducido a través de los requerimientos funcionales.

El método de resolución propuesto se acerca más a cualquiera de los métodos de desarrollo de software (refinamientos sucesivos) que a los métodos tradicionales de investigación científica, sin abandonar por ello a la misma. El motivo es que el carácter del problema que se desea resolver es cercano a los problemas que se plantean en el ámbito de la ingeniería de software. Un método de desarrollo de software da las pautas para la construcción de nuevos objetos, al igual que el método de investigación da las pautas para la construcción de metodologías y modelos [E Marcos 2002].

Por ejemplo, la filosofía de desarrollo propuesta por Beck en su metodología, cercana a los planteamientos basados en los métodos de investigación en acción [Beck, K. 1999]. Se trata de obtener la metodología de desarrollo de el caso de estudio mediante la creación directa del mismo (programación o concepción y diseño), en colaboración con los usuarios (del software o de la metodología), probándolo y adaptándolo continuamente durante el propio proceso de creación.

Se encuentran similitudes en un método de desarrollo de software con respecto al método de caso de estudio. Por ejemplo, si se desea construir una biblioteca digital, en primer lugar se analizaran otras y mediante un proceso de imaginación y creatividad se diseña la nueva biblioteca. Mediante un juego de pruebas se verificará el buen funcionamiento. La validación con el usuario puede hacerse mediante el uso de un prototipo. En el presente la etapa de verificación tiene dos tareas: la validación, comprobar que el modelo se ha construido según la hipótesis planteada, y la verificación consistente en comprobar que se ha construido correctamente, esto es, que es consistente. Igualmente, se puede realizar una comparativa con cualquier método de desarrollo de software donde la validación permite comprobar el cumplimiento de las especificaciones y la verificación para asegurar la corrección del sistema [E Marcos 2002].

3.3 Proceso de extracción del conocimiento de un sitio web

La minería de datos es un término actual como ya se ha mencionado, pero que es sólo una parte de el descubrimiento del conocimiento. La necesidad de extraer conocimiento en forma automática a partir de grandes bases de datos. Retomemos pues los conceptos que ya se han definido para analizarlos ahora en la web.

El descubrimiento del conocimiento en las bases de datos es el proceso de descubrir conocimiento útil dentro de los datos. La minería de datos es la aplicación de algoritmos específicos para extraer patrones de los datos.

Es posible enunciar una forma para explicar matemáticamente una serie de conceptos conducentes a una definición operativa del conocimiento.

- Datos: Un conjunto de hechos D .
- Patrón: Una expresión E , en algún lenguaje L que describe un subconjunto de los datos d , siempre que sea más sencilla que la simple enumeración de todos los hechos que componen d .
- Validez: La certeza de que el patrón sigue siendo válido cuando se aplica a datos nuevos. Se define como una función $C(E, D)$ que asigna una calificación (un número) al patrón.
- Novedad: Una función $N(E, D)$ que devuelve verdadero si el patrón no es simplemente una recombinación de patrones ya detectados o falso en caso contrario.
- Utilidad: La definición de utilidad es más resbaladiza y subjetiva. Un patrón es útil si nos permite decidir o realizar una acción.

De nuevo se puede representar como una función que califica la utilidad $U(E,D)$. Por ejemplo el dinero ahorrado o ganado al descubrir un patrón de compra en un supermercado.

- Comprensibilidad: Los patrones han de ser comprensibles por los seres humanos. De nuevo un concepto subjetivo y difícil de evaluar. Fayyad sugiere como medida cuantitativa la sencillez del patrón, de nuevo una función $S(E, D)$ que asigna un valor.

Todo ello conduce finalmente al importante concepto de “medida del interés” de un patrón, que se define como una combinación de Validez, Novedad, Utilidad y Comprensibilidad que nos permite valorar y clasificar los patrones [Fayyad 96].

$$i = I(E, D, N, U, S)$$

Ciertos aspectos de este concepto requieren la intervención humana, ya que no admiten una cuantificación objetiva.

La medida del interés es fundamental para la definición de Conocimiento:

- Conocimiento: Un patrón E se llama conocimiento si su medida del interés I supera un cierto umbral “u” definido por el usuario.

Puede parecer una definición muy alejada de nuestra experiencia de lo que es conocimiento. El conocimiento lo constituyen aquellos patrones que hemos aprendido a detectar y que hemos guardado por que nos permiten aplicarlos a nuevos datos y, por tanto, predecir el comportamiento de los fenómenos o las personas que nos rodean. De ahí deriva la utilidad del conocimiento. Un ejemplo es el diagnóstico médico, cada enfermedad tiene un conjunto de síntomas, un patrón, que la diferencia de otras, lo que permite diagnosticar y aplicar el tratamiento. Cuesta años hacerse con el bagaje de patrones clínicos ser un buen diagnosticador.

Los fraudes siguen patrones que se apartan del comportamiento habitual de las transacciones legales en las bases de datos financieras.

En marketing es importante descubrir los grupos de usuarios y sus patrones de comportamiento para definir productos y/o servicios específicos con resultados predecibles. Por ejemplo los usuarios que compran el artículo A y también el B probablemente compren también el C. Al final resulta que el conocimiento no es tan mágico como parece.

Disponemos de medios para aproximarnos a él y encontrar patrones interesantes para diversos campos. La Arquitectura de la Información parece indisolublemente unida al diseño de sitios web, pero en realidad es algo mayor, es una parte fundamental de la conversión de información en conocimiento.

Según el glosario del *Argus Center for Information Architecture*, la Arquitectura de Información es “el arte y la ciencia de organizar la información para ayudar a la gente a satisfacer sus necesidades de información”.

Richard Saul Wurman, el acuñador del término, la definió como “*el estudio de la organización de la información con el objetivo de permitir al usuario encontrar su vía de navegación hacia el conocimiento y la comprensión de la información*” [Wurman R.S 1997].

Hay muchas otras definiciones, otras que se ligan de forma específica al diseño de sitios web, Aaron West, define: “la práctica de diseñar la infraestructura de un sitio web, específicamente su navegación”. Cada vez hay más “arquitectos de información” y explosión demográfica de sitios web, y se muestra necesaria la especialidad. Debido a que precisamente es en un sitio web donde es imprescindible una organización de la información para evitar su fracaso, parece que arquitectura de información y páginas web sean dos facetas de lo mismo. Sin embargo la arquitectura de información es algo esencialmente independiente y tiene que ver con lo el negocio de la comprensión.

Nathan Shedroff considera el proceso que lleva al entendimiento como un continuo que arranca en los datos y finaliza en la sabiduría, pasando por la información y el conocimiento. Del libro "*Information Anxiety 2*" [Wurman R.S 1997].

Resumimos aquí los puntos principales.

- **Datos:** A pesar de su abundancia no es la fuerza que define de nuestro tiempo. Los datos sin contexto no son información y como tal son simplemente la materia en bruto del que partimos para la comprensión.
- **Información:** Proviene de la forma en que se organizan y se presentan los datos, lo que le confiere, o permite que se revele, su significado o al menos su interpretación.

El paso de datos a información representa el paso de lo puramente sensorial a lo conceptual. Es la “destilación” de los datos.

- **Conocimiento:** Lo que diferencia el conocimiento de la información es la complejidad de las experiencias que se necesitan para llegar a él. Para que un alumno llegue al conocimiento de una cierta asignatura, se ha de exponer al mismo conjunto de datos de diferentes maneras, con diferentes perspectivas y ha de elaborar su propia experiencia del mismo.
- **Sabiduría:** Es el nivel último del entendimiento. En él entendemos un abanico suficientemente amplio de patrones y meta-patrones de forma que los podemos utilizar y combinar de formas y situaciones nuevas totalmente diferentes a las que nos sirvieron para aprender. La sabiduría es, como el conocimiento, algo personal que se elabora íntimamente y que va con las personas y se pierde con ellas, a diferencia de los datos y la información. Por ello su transmisión directa es casi imposible.

La visualización de Información interviene en el paso de datos a información y en la posibilidad de la construcción del conocimiento, al revelar los patrones que subyacen a los datos.

En base a estas necesidades se han desarrollado programas que permiten acceder visualmente a Internet. Estas tecnologías permite la realización de aplicaciones que visualizan grandes cantidades de datos. Web Map Viewer permite conectar en forma de mapa enlaces a sitios web en Internet que contienen mapas de estos.

La apariencia visual de WebMap se puede apreciar en la figura 16. Esta organizado en tres capas que permiten incluir información personalizada, categoría y relevancia.

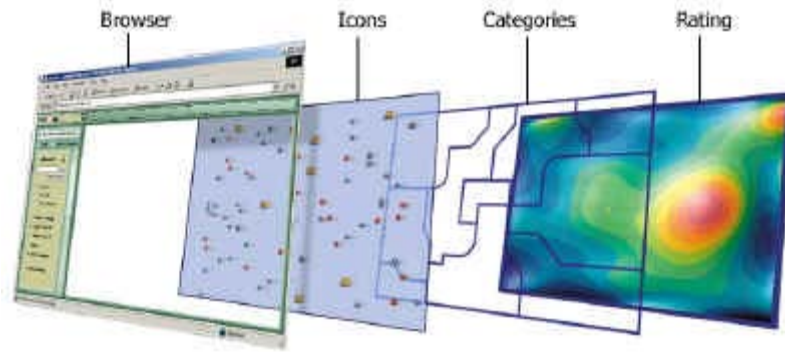


FIGURA 16 Estructura en tres capas del Web Map Viewer [C, Dürsteler Juan. 2005]

Se puede observar como la primera capa es la representación visual de los directorios de información en un espacio bidimensional. Los elementos de información se representan por píxeles definidos. La distancia entre los elementos representa distancia existente por los puntos de la pantalla. La información se encuentra de acuerdo a una serie de categorías representadas por áreas cerradas junto a los nombres correspondientes. La segunda capa representa la elevación en el eje z, y la altura es relevante a la información contenida, por último la tercera capa se compone de iconos que corresponden a la información de sitios favoritos recordado o almacenados en el histórico del navegador.

Este software proporciona algoritmos que determinan la distancia entre los elementos de información y categorías, para establecer las categorías y ubicar la información dentro de estas.

El sistema proporciona diversos algoritmos para determinar la "distancia" entre los elementos de información y entre las categorías, para establecer el tamaño y forma de los delimitadores de las categorías y para situar la información dentro de una categoría, entre otras. Permite interactuar con la tecnología y encontrar la información que se busca en forma visual.

Estas potentes tecnologías tienen que captar la atención de los usuarios y demostrar que son superiores al paradigma tradicional a la hora de facilitar la navegación a través del océano de la información.

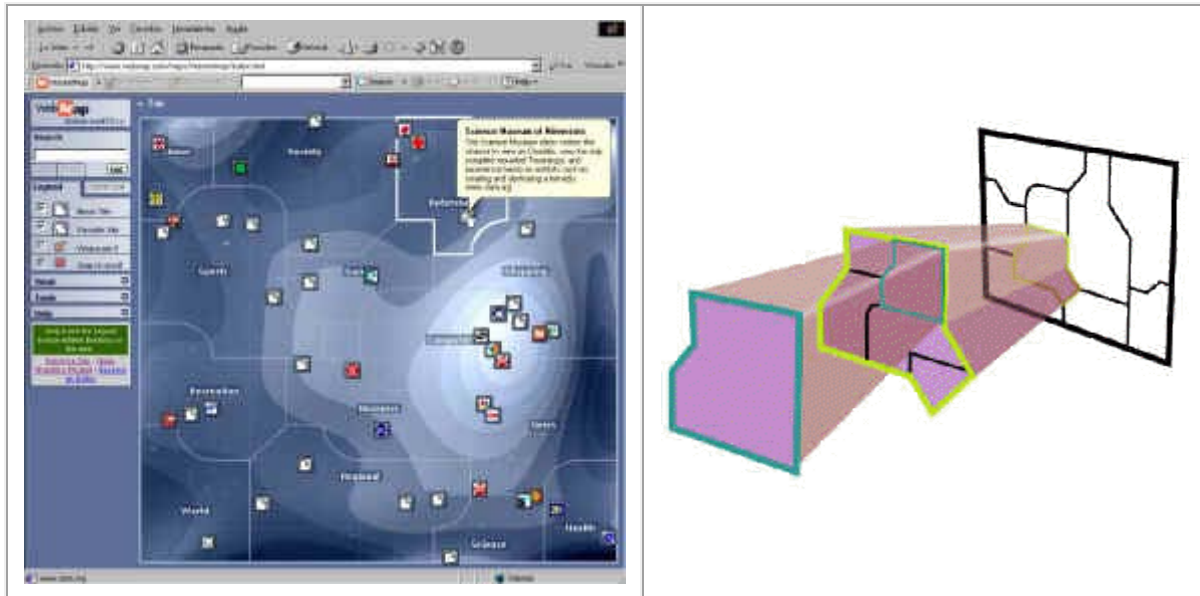


FIGURA 17 Estructura en tres capas del Web Map Viewer [C, Dürsteler Juan. 2005]

En la figura 17 se presenta otra vista de la herramienta que permite ver la información de la misma a medida que se pasa por puntos específicos. La posibilidad de lograr acercamientos es otra de las características.

3.3.1 Análisis de archivos de registro

Conocer la forma en que los usuarios de la web utilizan sus claves para el conocimiento del servicio de una página, saber si los servicios son encontrados de forma directa y medir la funcionalidad de la misma es el objetivo del análisis de los archivos de registro, éste es un método habitual a pesar de algunos problemas. Existen múltiples ejemplos de sitios web cuidadosamente diseñados en los que, sin embargo, los usuarios se pierden, no encuentran aquello que buscan, existiendo o peor aún, buscan algo que debiera estar y no está. Por otra parte los gestores del sitio web desconocen en muchos casos lo que hacen sus usuarios dentro de la web así como si encuentran lo que buscan, si buscan conceptos que no están en la web pero que podrían estar o si simplemente se pierden y se aburren abandonando el sitio. Sin conocer el impacto que tienen la campaña de marketing en un sitio web de comercio electrónico, difícilmente lo podremos hacer progresar en la dirección adecuada.

El tratamiento de estos datos es labor del software de análisis especializado, del que hay abundantes ejemplos. Uno muy popular de este tipo de software es Analog, un software gratuito que produce gran cantidad de estadísticas. Existe en más de 28 idiomas, si bien su visualización de datos es bastante elemental. Otros proveedores como Nedstat colocan una pieza de software en la sede web y son sus servidores (a diferencia de Analog) los que hacen las estadísticas.

La salida gráfica es también bastante básica. El análisis de los registros de archivo de motores de búsqueda es uno de los puntos sobre los que hay menos literatura y menos software, quizá porque pertenece al dominio de los grandes portales. (Si alguien me puede dar noticia, le quedaré muy agradecido.)

En cuanto a la interpretación, si nos basamos solamente en el análisis del archivo de registro que produce cada servidor.

Hay una serie de datos que no se pueden saber a ciencia cierta, entre ellos:

- el número de visitantes
- el tiempo que han pasado en la web
- el verdadero navegador que han estado usando.

En rigor estos datos solamente se pueden estimar. Se presenta una breve descripción de los problemas de interpretación que hace Stephen Turner, el creador de Analog. Así pues, el análisis de los archivos de registro ha sido el punto de partida para empezar a conocer que pasa dentro de la web. Pero es totalmente insuficiente. Si bien es cierto que es importante saber cuantas páginas se sirven, cuales son los países que más consultan, no dice lo más importante de cara a modificar la web: qué les gusta a los usuarios y qué no, qué páginas tienen más atractivo, como se mueven los usuarios por dentro de la web. El puro análisis de archivo de registro se está quedando corto. Es importante utilizar las propuestas de visualización para debilitar el problema.

Hay tres áreas donde la visualización de información puede ayudar respecto a la comprensión de la actividad de los visitantes:

1. Visualización de la estructura del sitio web. Sirve de elemento clarificador para los creadores del sitio y en forma de mapa del sitio, es un elemento de navegación para el usuario.
2. Revelación de los flujos y las trayectorias de los visitantes. Posibilita la creación de un sitio más efectivo, en el que no hayan páginas "muertas" o cuellos de botella.
3. Monitorización de la actividad en tiempo real del sitio web. Ayuda a los operadores del sitio a mejorar el rendimiento del negocio gracias al conocimiento de que se está comprando en cada momento, que páginas tienen éxito, etc.

Uno de los objetivos de las herramientas que se dedican a la representación de la web, es el análisis de por dónde entran los usuarios a las páginas que están visitando y ya en ella de qué forma salen, el ejemplo de lo mencionado es el motivo de la figura 18.

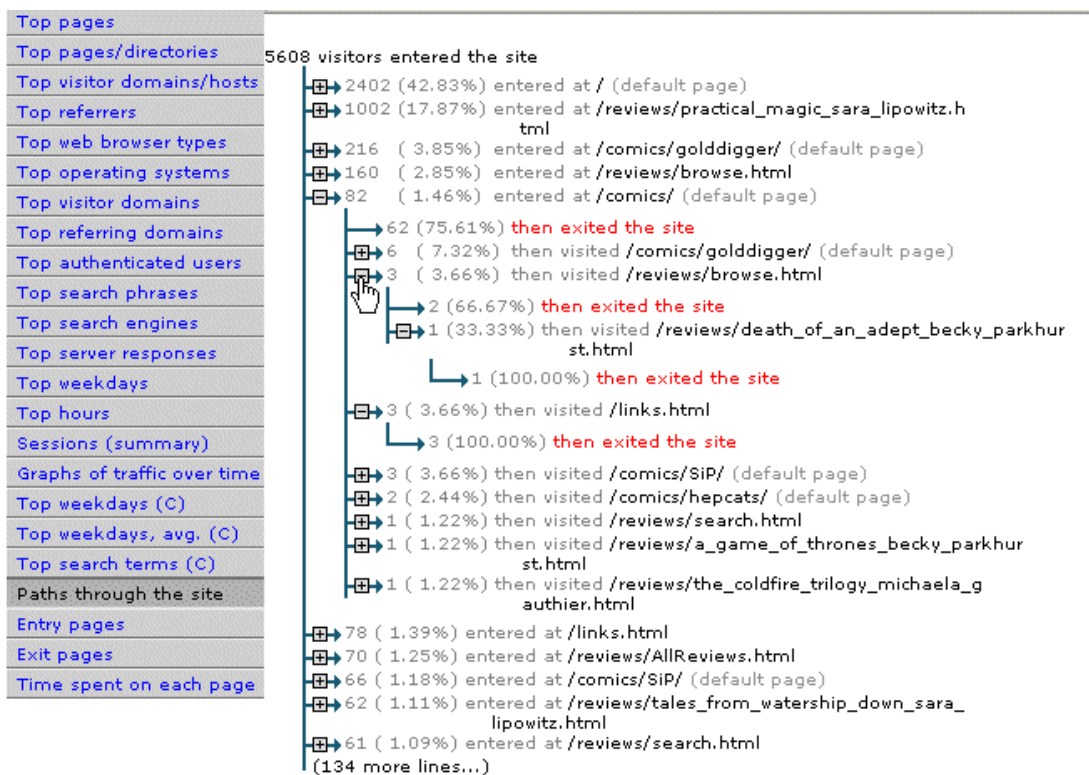


FIGURA 18 Análisis por visitas a un sitio

User info and path	Time (secs)
Visitor 62.11.121.42 came at 25-Dec-00 12:57:44 PM with agent Mozilla/4.0 (compatible; MSIE 5.01; Windows NT 5.0) from http://www.google.com/search?q=extract icon from executable&hl=it&lr= and visited pages:	
/eei.htm	0.00
Visitor 203.198.23.26 came at 26-Dec-00 2:15:20 AM with agent Mozilla/4.0 (compatible; MSIE 5.5; Windows 98; Win 9x 4.90) from - and visited pages:	
/afr/	37.00
/products.htm	71.00
/about.htm	10.00
/afr/awards.htm	15.00
/afr/scrshots.htm	56.00
/download.htm	0.00
Visitor 24.26.126.129 came at 26-Dec-00 2:17:33 AM with agent Mozilla/4.0 (compatible; MSIE 5.5; Windows 98) from http://search.dogpile.com/texis/search?q=grep&geo=no&fs=web and visited pages:	
/afr/	0.00
Visitor 142.176.115.223 came at 26-Dec-00 3:25:12 AM with agent Mozilla/4.0 (compatible; MSIE 5.5; Windows 98) from - and visited pages:	
/afr/	0.00
Visitor 202.109.94.91 came at 26-Dec-00 11:12:16 AM with agent Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 5.0; MyIE 2.0 Beta 2) from http://www.itamedia.com/cove/IntMenu1_.asp?CardID=6437 and visited pages:	
/	11.00
/afr/	4.00

FIGURA 19 Resultados arrojados del análisis de registro de archivo

El análisis de trayectorias se centra en lo que hacen los usuarios concretos y el flujo que integra los trayectos para revelar patrones de tráfico. La siguiente es una imagen de los últimos 30 visitantes o un agregado de los caminos frecuentes entre muchas otras estadísticas que pueden trabajar estas herramientas, la visualización es de forma tabular

La figura 19 muestra una visualización de los trayectos individuales en forma de árbol al estilo del explorador de Windows que resulta bastante clara y ocupa menos espacio que una representación en tercera dimensión. Otra Visualización por demás atractiva se presenta en la figura 20, representa la estructura de un sitio web como un grafo bidimensional sobre el que se superpone en tercera dimensión la trayectoria de un usuario en forma de línea suavemente curvada que salta de página en página. Las flechas indican el sentido del movimiento y unas rectas punteadas la longitud proporcional al tiempo que se invierte en cada visita.

Estas soluciones para analizar los archivos de registro permiten monitorizar el tráfico en la web y la importancia de la visualización para interpretar los resultados. Al ser desarrolladas eran pocas las aplicaciones gráficas que mostraban los resultados del análisis de los archivos de registro. La estructura de un archivo de registro es extremadamente simple.

Cada vez que alguien descarga un elemento de la web, como por ejemplo una página o una imagen, el servidor escribe una línea en el fichero histórico o *logfile*. Esta línea puede adoptar uno de entre varios formatos pero puede ser de la forma que se muestra en la tabla 4.

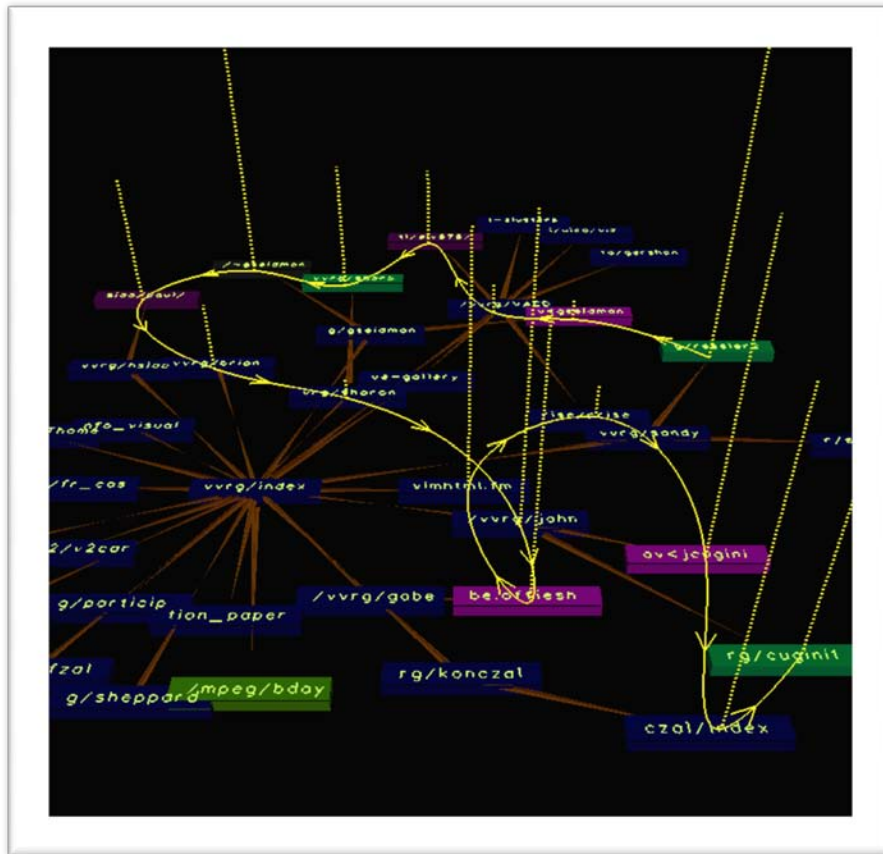


FIGURA 20 Imagen de VISVIP por John Cugini [C, Dürsteler Juan. 2005]

Lo importante es que a pesar de lo elemental que es, el estudio estadístico de la agregación de las muchas peticiones que hacen los navegadores de los usuarios al servidor permite conocer una gran cantidad de información derivada de estas simples líneas.

Entre ellas, al menos aparentemente, se puede conocer el número de páginas servidas por día, semana, mes o unidad de tiempo que se quiera, los sitios que apuntan a la web y redirigen el tráfico, las palabras que se buscan más habitualmente en la web y un largo etcétera.

IP	Identity check	ID de usuario	Fecha y hora	Método recurso y protocolo	S	T	Referrer	Agente (Browser, S.O., etc)
1 2 7. 0. 0. 1	-	f r a n k	[10/Oct/2005:13:55:36 - 0700]	"GET /apache_pb.gif HTTP/1.0"	2 0 0	2 0 3 2 6	"http://www.example.com/start.html"	"Mozilla/4.08 [en] (Win98; I;Nav)"

FIGURA 21 Línea generada por un archivo de registro (log file)

Si nos paseamos por las especificaciones de muchos analizadores de logfiles de un servidor web, como el mostrado en la figura 21, se encuentra que conoce el número de visitas únicas, de visitantes, cuánto tiempo han estado o incluso en una página parece pan comido para estos productos. Nada más lejos de la realidad. Mucha de esta información es de fiabilidad reducida debida principalmente a dos causas, entre muchas otras:

- HTML es un "stateless protocol". Cada petición resulta en una nueva conexión independiente que se abre y se cierra para la ocasión y no se puede relacionar de un modo fiel con otra hecha por la misma dirección IP. Aun más, si la IP es dinámica, es decir si la pueden usar distintos usuarios. Muchas "visitas" se crean con un "generador de sesiones" (sessionizer) que encuentra todas las entradas pertenecientes a una misma dirección IP y las considera parte de una misma

sesión si todas están alejadas entre sí menos de un cierto lapso de tiempo. Es imposible asegurar que todas pertenecen a la misma visita de la misma persona. Por lo mismo es imposible saber cuánto tiempo ha estado una página siendo vista ni cuál ha sido la secuencia real de su trayectoria dentro de nuestra web.

- Muchas páginas se reciben desde caches de servidores intermedios, sin que nuestro servidor llegue a enterarse nunca de que alguien ha visto esa página guardada en otro servidor. El uso de caches en Internet no sólo es conveniente sino la única manera de no colapsar ante un tráfico creciente, pero limita nuestro conocimiento del uso real de nuestra web.

Así pues, es imposible conocer de verdad cuantas páginas han sido vistas. En resumen. Atendiendo simplemente al análisis de un archivo de registro, no se puede conocer el número de visitantes, no se puede determinar cuántas visitas ha habido, no se puede conocer la identidad de los visitantes ni se pueden establecer fidedignamente las rutas que han seguido. Tampoco se puede saber cuánto tiempo han estado usando la web. Sin embargo ello no significa que las informaciones que se derivan del análisis de logfiles, aunque incompletas, no sean valiosas.

- Para empezar, a falta de un sitio web en el que obliguemos a nuestros usuarios a identificarse mediante un "login" y un "password", la información de los logfiles es probablemente todo lo que tenemos.
- Aunque la información sea incompleta, se puede llegar a una gran cantidad de conclusiones estudiando un logfile.
- Por ejemplo:
 - Qué conceptos buscan nuestros usuarios que no están en la web
 - Qué conceptos que sí que están no son encontrados.
 - Qué zonas de nuestra web registran más actividad.
- La aparición de patrones regulares y repetitivos en los caminos que encuentran los usuarios suele corresponder a patrones reales de comportamiento.

En definitiva, no se puede vender la idea de que los analizadores de archivo de registros encuentran visitas y visitantes "únicos", trayectorias de los mismos y hasta el sistema operativo que usan, como nos quieren hacer creer algunos fabricantes. Sin embargo si se es consciente de lo que nos dice un archivo de registro y de sus limitaciones, disponemos de herramientas poderosas para comprender un sitio web y el uso que de el hacen los usuarios, abriendo una puerta a las decisiones que pueden mejorar su rendimiento.

3.4 Consideraciones iniciales

Un diagrama resume una disciplina, se propone un diagrama sintético que englobe a la visualización de la información con las aportaciones necesarias.

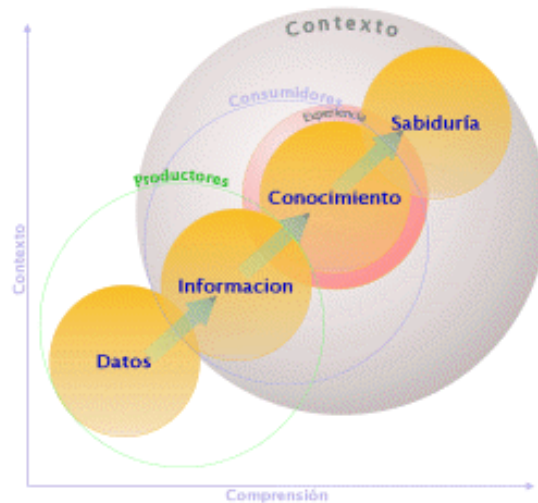


FIGURA 22 Esquema conversión de datos en sabiduría [Wurman R.S 1997]

Una de las reflexiones sobre los temas centrales de la Visualización de Información es: el esquema básico mostrado en la figura 22 mediante el cual los datos se convierten en información y ésta se transfiere a nuestro cerebro a base de estimular nuestra percepción sensorial creando en función del contexto, la cultura y la experiencia previa, una experiencia cognitiva. Estas son algunas de las aproximaciones propuestas para una extensión a los diagramas al uso. De los datos a la sabiduría.

Para ver paso a paso los elementos de dicho esquema la definición de Visualización de Información es el proceso de interiorización del conocimiento mediante la percepción de información. Nathan Shedroff considera el proceso que lleva al entendimiento como un continuo que arranca en los datos y finaliza en la sabiduría, pasando por la información y el conocimiento. La figura 23 reproduce el esquema. En el diagrama de Shedroff hay cuatro elementos conceptuales principales, que ya se mencionaron, y que se representan como una evolución en el tiempo, sobre dos ejes, el incremento de la comprensión por un lado y el aumento de la importancia que tiene el contexto entendido como la cultura, la experiencia y el conjunto de patrones adquiridos, por otro.

- **Datos** Los datos son simples hechos, carentes de contexto. Si no nos informan, no son información, o lo que es lo mismo desprovistos de contexto son simplemente la materia en bruto del que partimos para la comprensión. 07012007 es un dato, que puede tener muchos significados, una fecha, el nombre de un lote de fabricación, un cumpleaños....etc.
- **Información.** La información son los datos puestos en contexto. Es un concepto ligado al de metadato, un dato que hace referencia al significado de otro dato. Por ejemplo si en una tabla de datos una de las columnas reza número de lote, 07012007 una tira de caracteres en esa columna cobra un significado particular. La información es la destilación de los datos o los datos con su significado, pero aun no son conocimiento
- **Conocimiento.** Lo que diferencia el conocimiento de la información es la complejidad de las experiencias que se necesitan para llegar a él. Para que un conjunto de informaciones se conviertan en conocimiento hay que estar expuesto a el de diferentes maneras y hay que elaborar una experiencia propia respecto al mismo. El conocimiento se puede expresar como un patrón cuya medida de interés para el usuario supera un cierto umbral. Esto es, si no nos interesa es difícil que la información se convierta en conocimiento. El conocimiento no es transferible, se lo fabrica uno mismo experimentando la información. En este sentido Shedroff preconiza el “diseño de experiencias” como la forma de crear las experiencias que construyen conocimiento de forma más eficiente.

- Sabiduría. Es el nivel último del entendimiento. Cuando entendemos un abanico suficientemente amplio de patrones y meta-patrones de forma que los podemos utilizar y combinar de formas y situaciones nuevas totalmente diferentes a las que nos sirvieron para aprender. La sabiduría es, como el conocimiento, algo personal que se elabora íntimamente y que va con las personas y se pierde con ellas, a diferencia de los datos y la información.

En la figura 23 hay dos círculos que indican el ámbito de las personas que fabrican la información a partir de los datos (los productores) y el ámbito de los que consumen información y la procesan en conocimiento, (los consumidores). Un amplio círculo de Contexto engloba el paso de información a Conocimiento y de éste a Sabiduría. El conocimiento está englobado en un círculo de experiencia. El diagrama de Shedroff es esencialmente conceptual y no hace referencias a gráficos u otros artefactos ni a la forma en que se producen las transformaciones que dan lugar a la conversión de unas entidades en otras.

En el proceso de la visualización hay cuatro fases básicas, combinadas con un cierto número de lazos de realimentación, que se resumen en la figura 24.

- "La recolección y almacenaje de los datos mismos".
- "El pre-proceso diseñado para transformar los datos en algo que podamos entender".
- "El hardware del display y los algoritmos gráficos que producen una imagen en la pantalla".
- "El sistema humano perceptor y cognitivo".

Se pueden ver básicamente tres lazos de realimentación: la recolección, manipulación y exploración de los datos. El analista de información puede recoger una serie de datos, inyectarlos en el proceso tras un preproceso y transformación obtener una representación gráfica de los mismos que activa su sistema visual y cognitivo.

También puede manipular la forma en que el motor gráfico le muestra los datos una vez preprocesados y transformados, acaso cambiando los colores o las transformaciones geométricas que se los muestran.

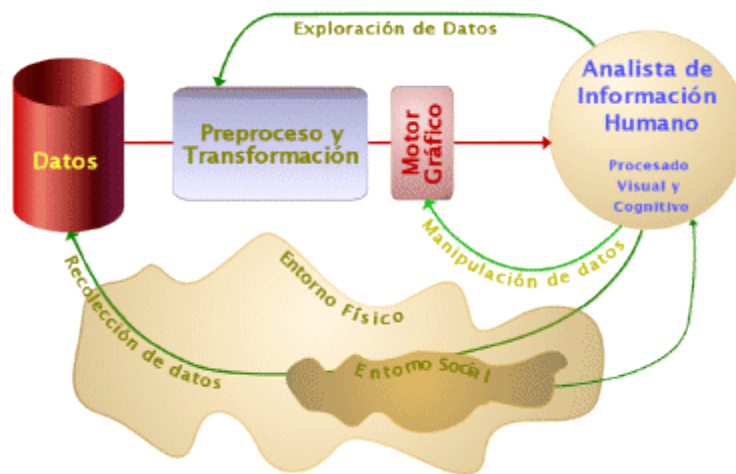


FIGURA 23 Diagrama referente al proceso de visualización [C, Dürsteler Juan. 2005]

Puede también explorarse los datos, escogiendo distintos pre-procesos, por ejemplo seleccionando distintos subconjuntos de la base de datos o transformando los datos y hallando magnitudes derivadas como, por ejemplo, diferencias entre datos o magnitudes estadísticas extraídas a partir de los mismos, y representar entonces los nuevos datos derivados. El entorno físico y el entorno social juegan, un papel en el lazo de la recogida de datos ya que el entorno físico es una fuente de datos mientras que el entorno social determina "de modos complejos y sutiles qué se recolecta y cómo se interpreta. "La construcción de Visualizaciones " propone el diagrama mostrado en la figura 24 para la visualización, entendida como "conversión ajustable de datos en forma visual para perceptores humanos." Este diagrama cubre la transformación de datos en bruto en gráficos.

Los datos en bruto en el formato que sea se transforman en tablas estructuradas de datos que tienen cierto significado representado a través de metadatos y posiblemente relaciones entre los datos. Ello se consigue utilizando "transformaciones de datos" que convierten los datos en bruto en tablas de datos mediante la adición de los metadatos apropiados.

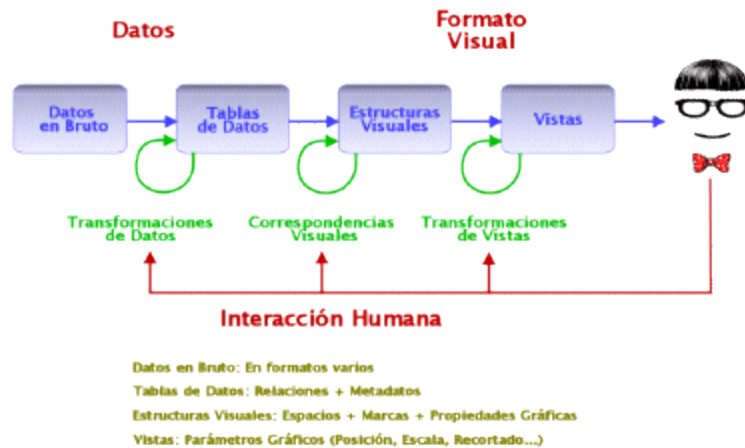


FIGURA 24 Diagrama construcción de visualizaciones [C, Dürsteler 2005]

Estas tablas de datos se traducen a su vez en estructuras visuales que producen una representación gráfica. Las transformaciones que lo hacen posible se denominan “visual mappings” o correspondencias visuales. Por ejemplo una tabla de datos tridimensional puede transformarse en un gráfico 3D usando cada una de las columnas que se asocian a una variable determinada (imaginemos que sus metadatos son consumo de gasolina, alcance y velocidad de un automóvil) o bien la misma tabla de datos puede dar lugar a una representación 2D con la tercera dimensión expresada mediante el tamaño o color de los puntos que dibujamos.

Finalmente podemos contemplar la visualización generada desde distintos puntos de vista. Esto es posible gracias a las transformaciones de pantalla que escalan, trasladan, amplifican o empequeñecen (zoom) y recortan la representación gráfica. La interacción permite al usuario proporcionar feedback al sistema cambiando los parámetros que controlan los tres tipos de transformaciones ya mencionados. Como vemos éste es un diagrama más cercano al proceso técnico de transformación de datos en bruto en representaciones gráficas.

Hasta aquí se muestra tres aproximaciones al proceso que nos interesa, la conversión de datos en entendimiento. Cada uno de los tres hace énfasis en distintos aspectos del proceso, conceptuales uno, más ligados a la percepción el otro y más cercanos a los gráficos por computador el último.

3.5 Visualizando con la herramienta WET *Website Exploration Tool*

De las visualizaciones que se han desarrollado a lo largo de la investigación, se muestran una serie de algoritmos implementados con diversas herramientas de programación que permiten sobre el papel o sobre la pantalla, la percepción de las relaciones que hay en los datos. Dichos algoritmos implementan todos o algunos de los pasos básicos necesarios para la construcción de la visualización:

- Recogida, filtrado y tratamiento de datos y compilación de estructuras de datos
- Transformación de las estructuras de datos en elementos perceptivos tales como gráficos, visualizaciones, auralizaciones u otros elementos capaces de activar nuestros sentidos.
- Presentación, posiblemente interactiva, al usuario de la visualización mencionada en el punto anterior.

Estos algoritmos se pueden ver reflejados en la herramienta que se presenta en esta investigación. La Universitat Pompeu Fabra dentro del grupo de Recuperación de Información y Minería Web, con la financiación de la empresa SPOC (ahora integrada en ALTRAN) y el soporte económico del Ayuntamiento de Zaragoza ha desarrollado una herramienta de visualización denominada WET (Website Exploration Tool).

La idea que subyace a WET es la de representar visualmente cualquier sistema que posea:

- Una estructura representable como un árbol. Es decir cualquier estructura jerárquica.
- Que disponga de un contenido, sea textual, multimedia, o de cualquier otro tipo
- Del que se pueda obtener información sobre la utilización que se hace del mismo.

Dada la ubicuidad de Internet y la facilidad de encontrar ejemplos prácticos la primera aplicación de la misma es la visualización de la estructura contenido y utilización de sitios web. Éstos obviamente cumplen las características antes mencionadas y son de fácil acceso, pero la arquitectura está pensada para cualquier sistema jerárquico digital.

La esencia de WET consiste en:

- la visualización de la estructura jerárquica en un marco que permite ver simultáneamente el foco de interés sin perder de vista el contexto, mediante el uso de varias metáforas visuales. Actualmente están implementadas dos metáforas visuales, pero la idea es tener una visión desde múltiples perspectivas enlazadas de forma que la interacción sobre una de las perspectivas se refleja simultáneamente en todas las demás (linked brushing). Dichas metáforas visuales son árbol radial y mapa de árboles, que se pueden apreciar en la figura 22 y 23, son ventanas de la aplicación de la herramienta WET, sobre la página creada por uno de los mismos autores de WET [V. Pascual 2007].

La página visualizada con el algoritmo de Radial Tree (Árbol Radial), se presenta en la figura 25. A diferencia del árbol clásico en el que desde un nodo raíz, situado en la parte alta de la visualización, aparecen las ramas hijas, enlazadas por arcos y situados por debajo del nodo raíz, en un árbol radial el nodo raíz ocupa el centro de una serie de círculos concéntricos sobre los que se sitúan las ramas. Cada círculo supone un nivel más. En el caso de una web saltar de un círculo interior al siguiente exterior supone una pulsación (click) del ratón[V. Pascual 2007]. Cada nodo del árbol se representa mediante un pequeño círculo relleno. Esta metáfora otorga un rectángulo a cada nodo del árbol de forma que el nodo raíz ocupa todo el espacio disponible y sus ramas se anidan en forma de rectángulos inscritos dentro del mismo y así sucesivamente hasta llegar a las hojas. Típicamente el área del cuadrado es proporcional a la magnitud que se quiera representar.

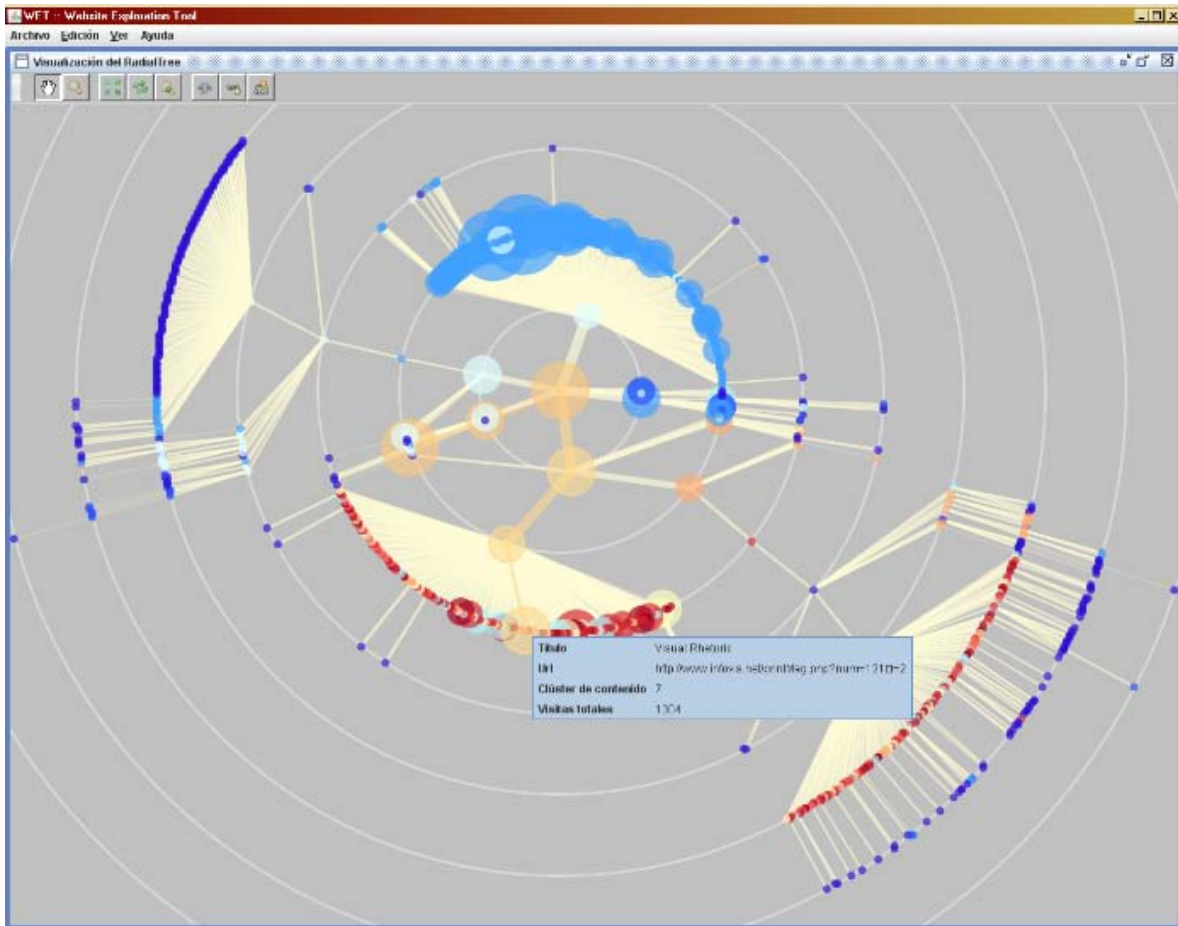


FIGURA 25 Ventana de visualización utilizando la herramienta WET

Cada nodo del árbol es representativo de una página y cada arco es un enlace que existe para ir de una página a otra. Pese a que en cada página puede haber más de un enlace, algunos de ellos cruzados, se ha implementado un algoritmo que encuentra los enlaces que con mayor probabilidad son "estructurales". Ello permite simplificar la visualización y permite conocer el camino mas corto desde la raíz hasta el nodo para así también saber cuantos son los "clics" mínimos que tiene que hacer el usuario para llegar a esa página des de la "home". El grosor de los enlaces que unen los nodos se utiliza para representar la frecuencia con que se usan dichos enlaces, y la transparencia de los nodos nos permite mostrar si la página ha sido visitada alguna vez (nodos transparentes = nunca visitados).

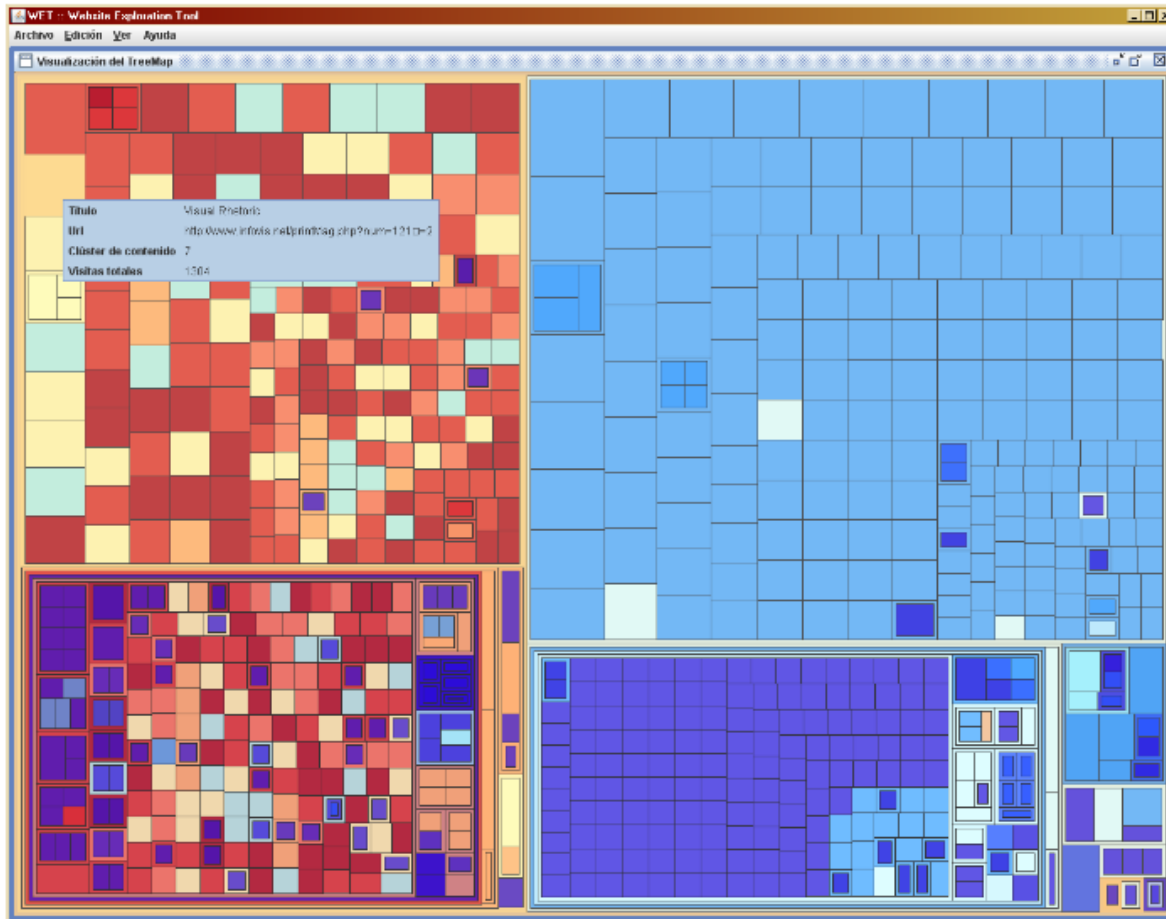


FIGURA 26 Visualización de construcción para páginas web.

Paralelamente al árbol radial se dispone de la metáfora visual del mapa de arboles. Éste representa la misma estructura jerárquica pero utilizando todo el espacio disponible, representando cada página como un bloque rectangular y anidando recursivamente en su interior sus hijos en la jerarquía. Entre ambas metáforas existe una coordinación en el resaltado. Esto permite utilizar conjuntamente todas las metáforas mostrando la misma información con las mismas características en todas las visualizaciones lo que facilita la extracción de conclusiones al combinar perspectivas diferentes y muchas veces complementarias.

Por tanto todo lo que se aplica al árbol radial se aplica simultáneamente al mapa de arboles. Cualquier asociación de métricas a variables visuales o cualquier interacción (exceptuando zoom y movimiento) que se haga con una metáfora se refleja en la otra y viceversa.

Desde el menú superior (*Ver > Navegación de los usuarios*) hay accesible la visualización de otra instancia del árbol radial que permite visualizar la estructura navegacional del sitio en vez de la estructura de diseño del mismo.

Este árbol se genera escogiendo los arcos más utilizados en vez de los más cercanos. De este modo se conoce cuales son los caminos que realmente la gente utiliza permitiendo así comparar la estructura con que se ha diseñado la jerarquía con la navegación real y observar si el diseño estructural es el más apropiado al uso que hacen los usuarios. En ocasiones los usuarios llegan a una página siguiendo caminos muy distintos de los ideados para ellos en el diseño. Al mismo tiempo disponemos de una menú lateral organizado en tres paneles diferentes. El panel superior dispone de la información sobre cada nodo en valores numéricos.

El panel intermedio contiene la herramienta de configuración que permite determinar los elementos activos de la visualización. Y el panel inferior contiene información contextual con las estadísticas de las páginas servidas durante un cierto periodo de tiempo (que por simplicidad denominamos *visitas*) y las leyendas de cada una de la métricas que se vayan visualizando durante la exploración. Si se pulsa sobre un nodo muestra las estadísticas de ese nodo, mientras que si no, se obtienen las totales del sitio.

4 RESULTADOS

4.1 Diseño de la investigación

Los componentes que se consideran para el diseño de la investigación en el presente caso de estudio son los siguientes de acuerdo a la metodología propuesta por [Yin, Robert 1994]:

Pregunta de investigación: Según la hipótesis descrita y que sustenta el presente trabajo:

“La representación visual de la información permite incrementar y facilitar la capacidad de análisis de la estructura de la web, mediante las técnicas de la minería de datos, proporcionando métricas y estadísticas para el entendimiento de la estructura de un sitio web así como del uso que se da de ella.”

Propositiones. De acuerdo con Yin los casos de estudio pueden ser explicativos, exploratorios o descriptivos. Debido a que el tipo de investigación para esta tesis es exploratorio y de acuerdo a lo propuesto por Yin, se toman en cuenta los objetivos de la investigación para la guía del estudio. Los objetivos establecidos para la presente son.

Estudio de la visualización de sistemas que posean estructura representable como un árbol para disponer del contenido y obtener información sobre la utilidad de este, aplicados a la minería de estructura web.

Evaluar la herramienta WET (Website Exploration Tool) mediante la visualización de la estructura de un sitio web de fácil acceso.

Analizar la metáfora visual de árbol radial y mapeo de arboles.

Establecer razonamientos y conclusiones a partir del estudio realizado.

□ Demostrar la hipótesis mediante las técnicas de la minería de datos, proporcionando métricas y estadísticas para el entendimiento de la estructura de un sitio web así como del uso que se da de ella.

4.2 Recolección de los datos

En el árbol radial de la web podemos constatar que hay una serie de páginas que cuelgan de los arcos más externos que, conforme al diseño original, no debieran estar ahí pero debido a errores humanos o del script utilizado para su conversión de páginas html estáticas a dinámicas han quedado descolgados. En la figura 27 se han circulado en rojo. Puede percibirse este tipo de problemas con una larga lista de nombres de página.

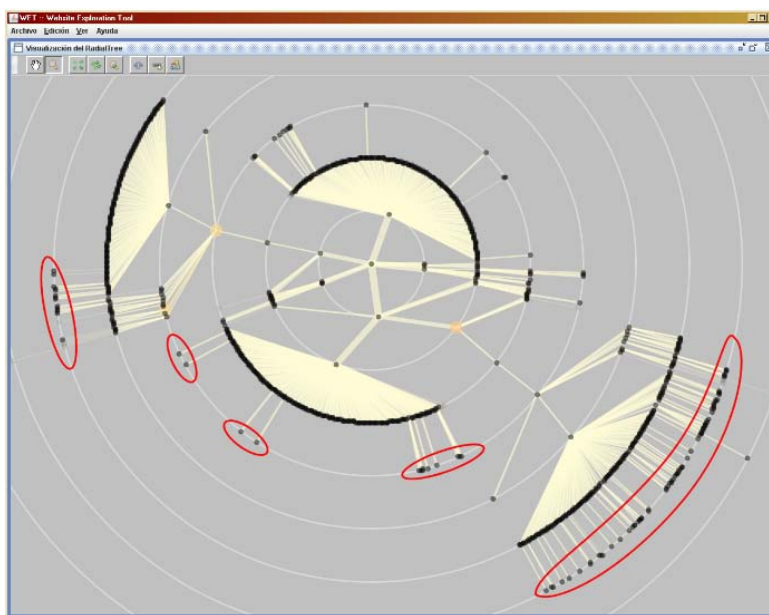


FIGURA 27 Errores de la página visualizada

Por otro lado la superposición sobre la estructura de la representación de las métricas nos permiten obtener información multidimensional cruzando distintos tipos de información, que nos pueden permitir encontrar casos particulares o bien patrones generales, según sean las necesidades.

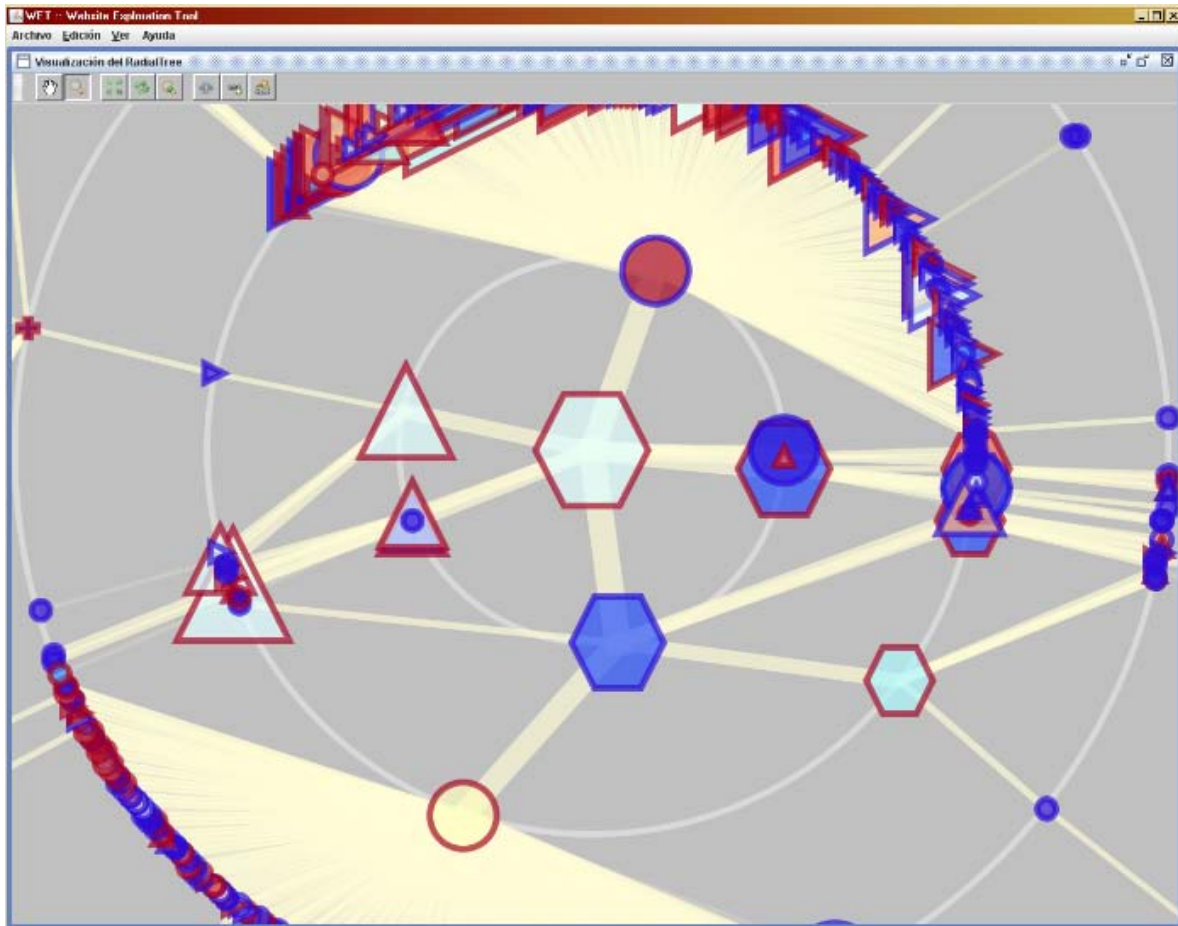


FIGURA 28 Variables visuales asociadas a la métrica

Se pueden analizar en la figura 28 las métricas como son la forma, el color con grupo de palabras clave usadas para acceder la página, el color del borde, si esta indexado en Yahoo, el tamaño, con número de páginas servidas. Es posible asignar el número de páginas servidas en un determinado periodo al tamaño del nodo y el color al clúster que agrupa las palabras clave por las que se ha llegado mayoritariamente a esa página, figura 28 y ver si las páginas más visitadas son de la misma temática (mismo color) o diferente.

O bien encontrar a golpe de vista las páginas más visitadas que tienen un ranking de Google determinado (no necesariamente las páginas con ranking más alto son las más visitadas en un periodo determinado).

Cada uno de los enlaces entre páginas tiene un grosor determinado que es proporcional al número de veces que se ha utilizado en un periodo determinado. Pasando el ratón por encima del enlace aparece la frecuencia con que se ha usado.

Si nos colocamos sobre una página y pulsamos sobre la misma podemos conocer las estadísticas de utilización de todas las métricas y la página se ilumina tanto en el árbol radial como en el mapa de arboles. Pulsando sobre el botón derecho del ratón no surge un recuadro con un pequeño menú que permite visitar directamente la página o ver los enlaces entrantes y salientes etc.

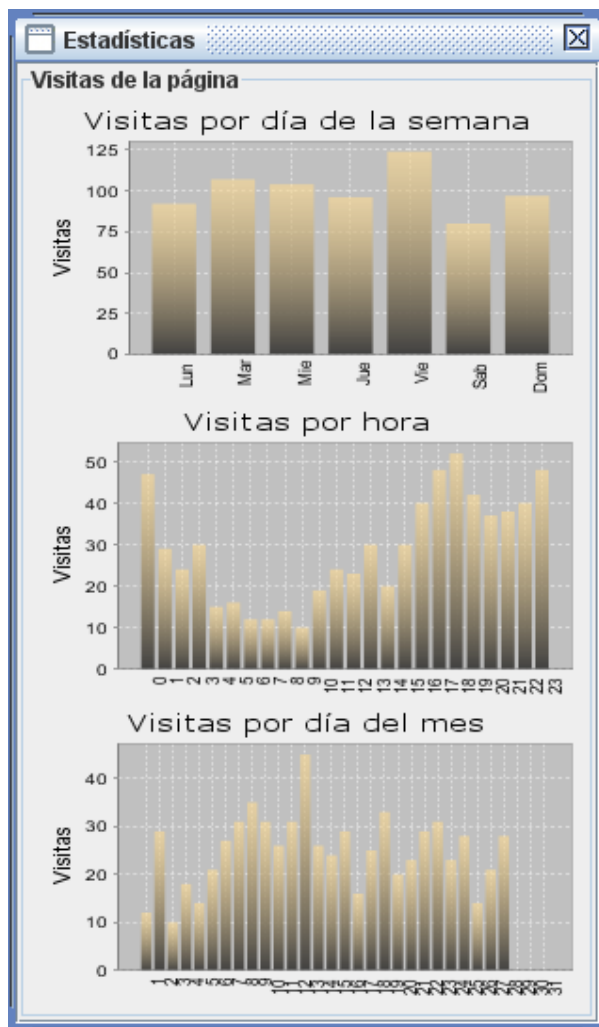


FIGURA 29 Estadísticas de consulta

WET tiene muchas posibilidades más abrir WET a la experimentación de los usuarios para pruebas y la retroalimentación al autor constituyo una gran experiencia sobre visualización de información.

4.3 Conclusiones

Mediante el uso de las herramientas de visualización y los algoritmos que puede desarrollar cada una, es posible obtener una clasificación visual del conjunto de datos o información analizada. Esto permite la administración estratégica y la toma de decisiones para futuros casos o mejoras en la estructura de un sitio web, partiendo de las condiciones generadas.

Existe una necesidad creciente o dependiente de la minería de datos para encontrar conocimiento en los volúmenes de información y lograr un aprovechamiento con la minería visual en lenguaje humano, obteniendo de esta, valores resaltados.

A través del presente trabajo se visualiza el proceso de extracción del conocimiento mediante técnicas y métodos que delimitan estrategias, los métodos de discretización de la información pueden ser supervisados y parametrizados, con el objetivo de modelar las etapas de acceso a la información basándose en atributos del uso, la forma y estructura web.

BIBLIOGRAFIA

- [Ansari, S. 2001], Kohavi, R., Mason, L., & Zheng, Z. *Integrating e-commerce and data mining: Architecture and challenges. Data mining*. San Jose, CA: IEEE Computer Society.
- [Anupam 2004] Anupam Joshi and Pranam Kolari. ,“*Web Mining: Research and Practice*”. Copublished by the IEEE Computer Society and the AIP, University of Maryland, Baltimore County. pp.49-53
- [Beck, K. 1999] *Embracing Change with eXtreme Programming*. Computer, vol.32, n° 10
- [C, Dürsteler Juan. 2005] <http://www.infovis.net> Fecha de consulta: 20/10/08
- [Chi, E. H. 2002] *Improving web usability through visualization*. IEEE Internet Computing, pp. 64-71.
- [Fayyad 96] Fayyad, U.M., *Data mining and Knowledge Discovery: Main Sense Out of Data.*, IEEE EXPERT, October 1996.
- [Fayyad 2001] Fayyad, U.M., Grinstein G., Wierse A. *Information Visualization in Data Mining and Knowledge Discovery.*, 2001, ISBN 1-55860-689-0 Morgan Kaufmann Publishers, San Francisco, CA, USA.,
- [E Marcos 2002] G KYBELE *Investigación en Ingeniería del Software– Actas de 1er Workshop en Métodos de Investigación y Fundamentos Filosóficos en IS y SI.* kybele.escet.urjc.es
- [Galeas, P. 1996] Galeas, P., 1996 *Infotrax GmbH*, fecha de consulta 29/05/2009, disponible: <http://www.galeas.de/webmining.html>
- [Hernández 2004] Hernández Orallo, J.; Ramírez Quintana, M.J.; Ferri Ramírez, C., *Introducción a la Minería de Datos.*, ISBN978-84-205-4091-7 PEARSON Prentice Hall, España.
- [ISO/IEC 1999] ISO/IEC 13250, *Topic Maps 1999* The International Organization for Standardization and The International Electrotechnical Commission.
- [J.Srivastava 2002] J.Srivastava, P. Desikan, and V. Kumar. *Web Mining Accomplishments and Future Directions*, Proc. US NaPI Science Foundation Workshop on Next-Generation Data Mining (NDGM), NAPI Science Foundation.

- [Keim D. A. 2002] Keim D. A., Wolfgang M., Heidrun S., *Information Visualization and Visual Data Mining*, January – March 2002, IEEE Transactions on Visualization and Computer Graphics, Vol. 8 No. 1
- [Larose D 2005] Larose Daniel T, 2005 *Discovering knowledge in data: an introduction to data mining.*, ISBN0-471-66657-2. John Wiley and Sons., United States of America.
- [R.Kosala 2000] R.Kosala and H. Blockeel, *Web Mining Research: A survey.*, *ACM SIGKDD Explorations.*, vol. 2, no.1 pp. 1-15
- [Redish, J 2001] Redish J, 2001 *Making Information Visual: Creating Effective Web Pages STC*, Redish & Associates, Inc. Chicago. Fecha de consulta 11/11/2008, Disponible en:
http://www.redihs.net/content/handouts/redish_shorttalk_May2001.pdf
- [Sankar. K. 2002] Sankar. K. Pal, V. Talwar, P. Mitra, 2002 *Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions.*, IEEE transactions on neural networks, vol. 13, no.5 pp. 1167
- [V. Pascual 2007] V. Pascual , Dürsteler, J.C. *WET: a prototype of an Exploratory Search System for Web Mining to assess Usability.*, Information Visualization. IV '07. 11th International Conference IEEE pp. 211
- [Van Harmelen 2001] Van Harmelen, F. Broekstra, J. Fluit, C. ter Horst, H. Kampman, A. van der Meer, J. Sabou, M. Vrije Univ., Amsterdam; *Ontology-based information visualisation* This paper appears in: Information Visualisation, 2001. Proceedings. Fifth International Conference on page(s): 546-554, ISBN: 0-7695-1195-3
- [Walrus, 2005] Walrus CAIDA: *The Cooperative Association for Internet Data Analysis.*, Fecha de consulta 21/10/2008, Disponible en
<http://www.caida.org/tools/visualization/walrus/gallery1>

- [Ware, C. 1999] Ware, C., 1999 *Information Visualization. Perception for design.*, ISBN1-55860-511-8, by Academic Press, United States of America.
- [Wong, P.C 2000] PC Wong, W Cowley, H Foote, E Jurrus, J Thomas, *Visualizing sequential patterns for text mining*, IEEE Symposium on Information Visualization ieeexplore.ieee.org
- [Wurman R.S 1997] Richard Saul Wurman et al. *Information Architects*, Watson-Guption Pubns. ISBN 1-888001-38-0
- [Yin, Robert 1994] *Case Study Research: Design and Methods*, Sage Publications, Thousand Oaks, CA, 2003, 3rd edition
- [Z. Pabarskaite 2003] A. Raudys, *A process of knowledge discovery from web log data: Systematization and critical review.*, Springer Science + Business Media, LLC.