Descubrimiento de Conocimiento a partir de Datos

Prof. Dra. Silvia Schiaffino
ISISTAN

Contenido del Curso

- Introducción al KDD
- Etapas
- Pre-procesamiento de datos
- Data Mining
  - Reglas de Asociación
  - Redes de Bayes
  - Clasificación
  - Modelos de Markov
  - Clustering
- Web Mining
- Social Mining

Agenda

- Motivación de la tecnología
- Taxonomía
- Web Content Mining
  - Motivación
  - Enfoques
- Web Structure Mining
- Web Usage Mining
  - Motivación
  - Beneficios
  - Aplicaciones
- Text Mining
  - Motivación
  - Aplicaciones

Motivación

- Riqueza enorme de información en la Web
- La Web es una enorme colección de:
  - Documentos de todo tipo (texto, HTML, bases de datos, …)
  - Información hipervinculada
  - Información de uso y acceso
- Desafíos: no hay estándares, no estructurada, crece y cambia rápidamente

Se necesitan mejores métodos para extracción de conocimiento y acceso a los recursos, para transformar a la Web en un entorno útil

Motivación

- Hoy en día, la World Wide Web se ha convertido en un medio popular e interactivo para distribuir información
- Los usuarios de esta información interactuando con la Web, se pueden encontrar con algunos de los siguientes objetivos:
  - Encontrar información relevante
  - Extraer nuevo conocimiento a partir de la información disponible en la Web
  - Personalización de la información
  - Aprender de los consumidores o usuarios individuales
  - Analizar comportamiento social…

WWW y Web Mining

Web

Repositorio de información hipermedia/hipertextual enorme, ampliamente distribuido, heterogéneo, semi-estructurado, interconectado, en evolución.

Problemas:

- Abundancia: 99% de la información no es interesante para el usuario
- Cobertura limitada: recursos Web ocultos, la mayoría en DBMS
- Limitada personalización a usuarios individuales
- Limitada interfaz de consulta basada en búsqueda orientada a la palabra clave

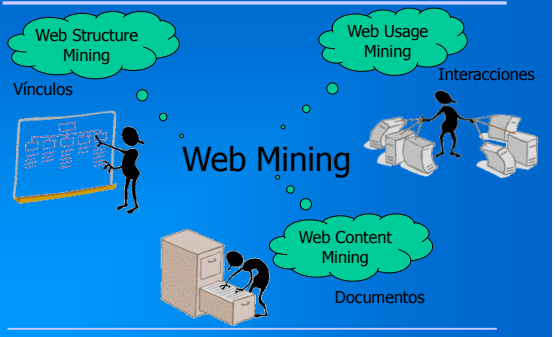
Datos en la Web

- Tipos de datos Web conceptuales según Madria et al.(1999)
  - Datos contenidos en la WWW: toda la información disponible on-line
  - Datos de logs en la Web: actividades on-line de los usuarios (Cookies)
  - Datos de estructura de la Web: información de links en la Web
- Además, Srivastava et al. (2000) and Eirinaki and Vazirgiannis (2003) definieron un cuarto tipo
  - Datos de perfil de usuario: información demográfica (nombre, edad, etc.) e información de perfil de cliente (intereses del usuario, preferencias, etc.)

Web Mining: Concepto y Categorías

- Web Mining consiste en el uso de técnicas de Data Mining y de procesamiento de información para descubrir y extraer información relevante automáticamente a partir de documentos y servicios de la Web

Categorías de Web Mining



Categorías de Web Mining

- Web Content Mining

Es el proceso automático que extrae patrones a partir de información on-line tal como archivos HTML, XML, texto, imágenes o correo electrónico, y va más allá de la extracción de palabras clave, o estadísticas de palabras o frases en documentos.

Categorías de Web Mining

- Web Structure Mining

Se centra en usar el análisis de la estructura de links de la Web y uno de sus propósitos es identificar documentos más buscados o más importantes para un tema. Se usan técnicas de data mining para reconstruir la estructura de un sitio o sitios.

Categorías de Web Mining

Web Usage Mining

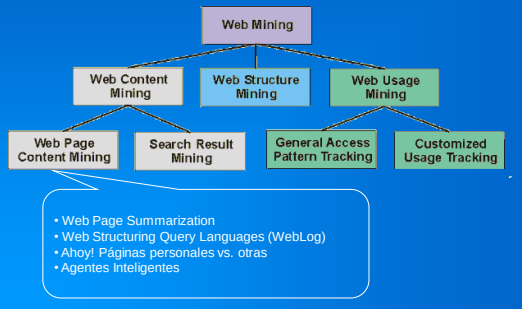
Los Web servers almacenan datos sobre las interacciones del usuario cada vez que se reciben solicitudes de recursos.

El análisis de los logs de acceso de diferentes sitios Web puede ayudar a entender el comportamiento de los usuarios y la estructura de la Web, mejorando el diseño de esta colección de recursos.

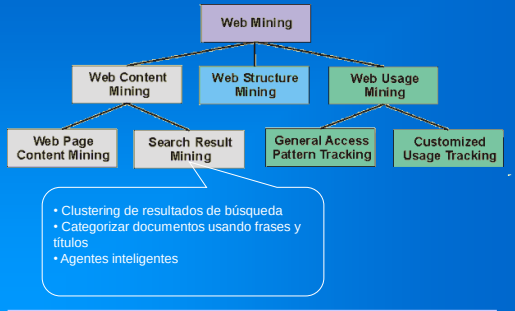
Personalization mining

- Razones para considerarlo como una cuarta categoría:
  - Datos del usuario → Perfil del usuario
  - La personalización se centra en las actividades de cada usuario (cliente de un sitio electrónico) mientras que Web Usage Mining se centra en actividades de grupos
  - Mucha investigación en el área de personalización y de sistemas adaptativos

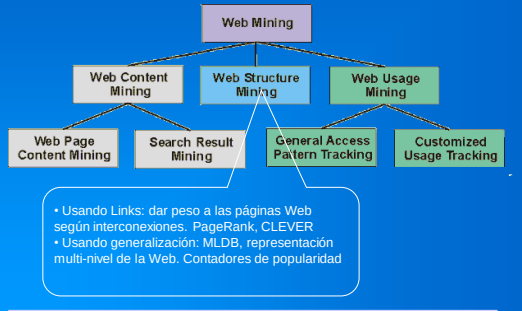
Taxonomía



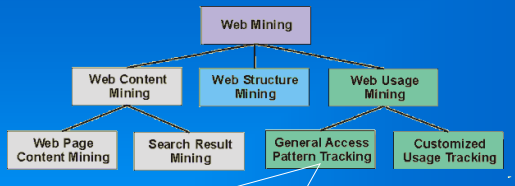
Taxonomía



Taxonomía

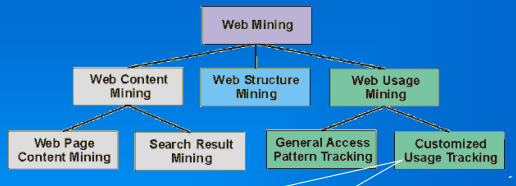


Taxonomía



• Web Log Mining: entender los patrones generales de acceso y las tendencias. Permite dar una mejor estructura al sitio y agrupar recursos

Taxonomía



• Sitios adaptativos: patrones de acceso de usuarios individuales.
• Agentes Inteligentes

Web Content Mining

- Naturaleza dinámica y gran cantidad de documentos hacen difícil el descubrimiento de conocimiento, la organización y administración de información Web automáticos.
- Una consulta Web puede devolver miles de páginas. Se necesitan métodos para presentar estos resultados de manera de ayudar al usuario a seleccionar aquellos que le interesen.

Web Content Mining

- Los motores de búsqueda y herramientas de indexado tradicionales (Lycos, AltaVista, WEbCrawler, ALIWEB, MetaCrawler) han provisto alguna ayuda a los usuarios, pero generalmente no proveen información estructural ni categorizan, filtran o interpretan documentos.
- Estos factores han llevado a los investigadores a desarrollar herramientas más inteligentes para recuperación de información, tales como agentes Web inteligentes, y a extender técnicas de bases de datos y data mining para proveer un mayor nivel de organización a los datos semi-estructurados disponibles en la Web

Índices

- Buscadores que mantienen una organización de las páginas incluidas en su base de datos por categorías
- Directorio navegable de temas
- Dentro de cada directorio o tema se pueden encontrar páginas relacionadas con ese tema
- Administradores humanos que se encargan de visitar las páginas y vigilan que todas se encuentren clasificadas en su lugar correcto.
  - Para que una página quede registrada en un índice, es preciso enviarles la dirección a los administradores humanos de ese índice
  - Ej.: Terra y Yahoo (aunque a este último se lo utiliza más habitualmente como un motor de búsqueda).

Motores de búsqueda

- Buscadores que basan su recolección de páginas en un robot (spider, araña) que recorre constantemente Internet en busca de páginas nuevas
- Introduce las páginas en su base de datos de forma automática.
- No necesariamente tienen un índice, aunque cada vez es más habitual.
- Para clasificar una página Web, los motores de búsqueda leen el contenido y encuentran aquellos datos que permitan su clasificación
  - Cuando el robot recorre una Web guarda sus datos, y luego se dirige a las distintas páginas que están enlazadas.
  - Las páginas se vuelven a recorrer cada cierto tiempo de su base de datos en busca de los cambios
  - Ej.: Altavista y Google.

Multibuscadores

- Los multibuscadores no poseen una base de datos propia
- Buscan la página en muchos motores de búsqueda e índices
- Combinan los resultados de la búsqueda.
- Ej.: MySearch (que busca simultáneamente en Google, Yahoo, Altavista y muchos otros).

Representación de contenido

- Selección de palabras clave manual. Requiere expertos en varios dominios. Evaluación imparcial de la página e imparcial selección de palabras clave. Limitado a la velocidad y cantidad de expertos
- Selección de palabras clave automática. Buscar en todo el documento o concentrarse en partes como título, primer párrafo, o usando tags HTML o XML. Meta tags.

Web Content Mining: Agentes Inteligentes

- Agentes de búsqueda inteligentes

  Sistemas que pueden actuar autónoma o semi-autónomamente en nombre de un usuario particular. Buscan información relevante usando características de un dominio dado (y un perfil de usuario) para organizar e interpretar la información descubierta

  – FAQ Finder, Harvest, ParaSite: usan información de dominio específica sobre tipos de documentos particulares, basándose en estructura dada de las fuentes de información
  – ShopBot, ILA: aprenden la estructura de fuentes de información traduciéndola a una representación conceptual propia

Web Content Mining: Agentes Inteligentes

- Filtrado y categorización de información

  Usan técnicas de recuperación de información para recuperar, filtrar y categorizar documentos (TF-IDF)

- Agentes Web personales

  Aprenden las preferencias del usuario y descubren fuentes de información que corresponden a estas preferencias y posiblemente las de otros individuos con intereses similares (filtrado colaborativo)

  – Syskill & Webert
  – Web Watcher
  – News Agent
  – Personal Searcher

Agenda

- Motivación de la tecnología
- Taxonomía
- Web Content Mining
  – Motivación
  – Enfoques
- Web Structure Mining
- Web Usage Mining
  – Motivación
  – Beneficios
  – Aplicaciones
- Text Mining
  – Motivación
  – Aplicaciones

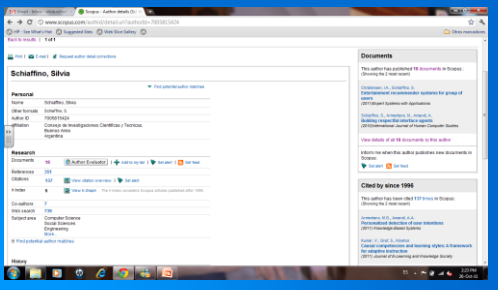
Web Structure Mining

- Los links que apuntan a un documento indican la popularidad del documento
- Los links que salen de un documento indican la riqueza o la variedad de temas cubiertos por el documento.
- Útil para generar información que relacione documentos entre diferentes sitios
- Descubrimiento de páginas "authoritative" (de calidad, que son "autoridad" en un tema dado). Mismo enfoque que referencias a artículos

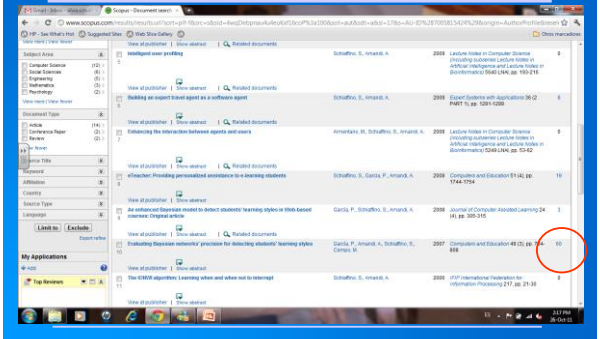
Análisis de citas (referencias)

- Fue muy estudiado en el área de recuperación de información antes de que apareciera la Web
- Factor de impacto (Garfield 72)
  - Provee una evaluación numérica de las revistas (qué tan referenciado es un journal)
- Sin embargo, no todas las citas son igualmente importantes (Pinski y Narin 96)
  - Un journal es influyente si, recursivamente, es muy citado por otros journals influyentes
  - Peso de la influencia: la influencia de un journal j es igual a la suma de la influencia de todos los journals que citen a j, con la suma ponderada entre la cantidad que cada uno cite a j

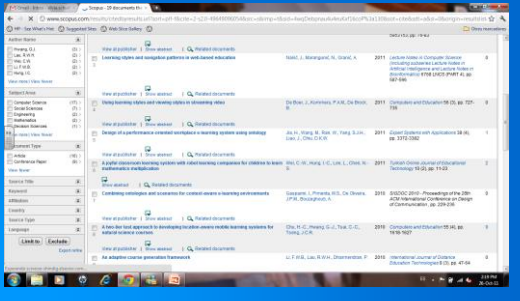
Análisis de referencias: Scopus



Análisis de referencias: Scopus



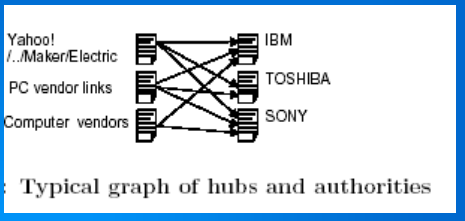
Análisis de referencias: Scopus



Descubrimiento de páginas “authoritative”

- Método Page-Rank (Brin y Page 98)
  - Usado en Google
  - Hacen un ranking de la importancia de páginas Web, tomando como “votos” los links de una página a otra (Ecuación de millones de variables...)
  - Considera pesos de las páginas que votan.
- Método hub/authority (Kleinberg98)
  - Página buena “authoritative” con respecto a una consulta si está referenciada por muchas (hub) páginas relacionadas a las consultas
  - Página es buena “hub” con respecto a una consulta si apunta a muchas páginas “authoritative” con respecto a la consulta
  - Relación de refuerzo mutuo entre ambos tipos de páginas
  - Algoritmo HITS (Hyperlink Induced Topic Search)

Hubs and authorities



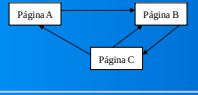
Typical graph of hubs and authorities

PageRank

- Mecanismo utilizado por Google para medir la “importancia” de una página
- Google utiliza PageRank para ajustar los resultados de las búsquedas de manera que los sitios de mayor importancia son puestos en los primeros lugares
- Pasos del sistema de ranking de Google
  1. Encontrar todas las páginas que concuerdan con las palabras clave de la búsqueda.
  2. Rankear consecuentemente utilizando factores propios de la página, como ser las keywords o palabras clave.
  3. Calcular el peso del texto de entrada.
  4. Ajustar los resultados mediante los puntajes de PageRank

¿Cómo se determina el PageRank?

- Si una página A tiene un link hacia una Página B, entonces la Página A está diciendo que la Página B es importante
- El texto real del link de A a B es irrelevante para el PageRank
- Si una página tiene muchos links importantes que entran en ella, los links hacia otras páginas también se vuelven más importantes.



Algunos puntos clave de PageRank

- El PageRank es un número
- Se calcula en base a la capacidad de votación de todos los links entrantes a una página, y cuánto la recomiendan
- Todas las páginas indexadas en Google tienen su propio PageRank (no necesariamente el de la Home Page)
- Los links internos cuentan en el pasaje de PageRank a otras páginas del sitio.
- El PageRank es independiente del texto de los links o el título de la página en cuestión

Valores de PageRank

- La única manera de saber el PageRank oficial de un sitio es a través de la barra de herramientas que ofrece Google (aunque no es exacta)
- Número de 0 a 10 que aproxima al verdadero valor
  - Limitaciones de la barra:
    - Si la página no se encuentra indexada
      - valor estimativo de acuerdo con alguna página similar
      - se muestra un 0, correspondiente a que no hay información
    - Se muestra una representación del valor verdadero (escala exponencial)

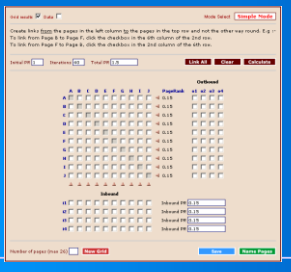
Valor Mostrado	PageRank aproximado entre
0	La página no está indexada
1	0.00000001 - 5
2	6 - 25
3	25 - 125
4	126 - 625
5	626 - 3125
6	3126 - 15625
7	15626 - 78125
8	78126 - 390625
9	390626 - 1953125
10	1953126 - Infinito

Ejemplos de la Barra de Google

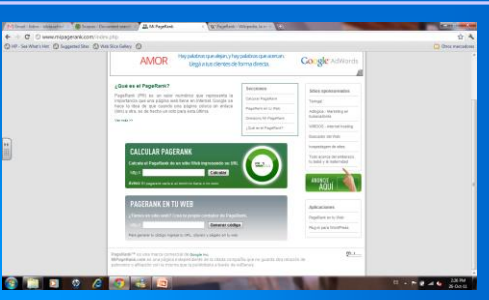
Ejemplos de la barra de Google

Calculadoras de PageRank

- Si bien no se puede determinar exactamente el PageRank, se puede aproximar con "Calculadoras de PageRank"



Mi Page Rank



Cómo se calcula el PageRank

- Única información oficial publicada originalmente por Brin y Page ("The Anatomy of a Large-Scale Hypertextual Web Search Engine")
- Fórmula:

$$PR(A) = (1-d) + d\left(\frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)}\right)$$

PR(A) es el PageRank de la Página A, es decir, la que queremos calcular.

d es el factor de debilitación, usualmente 0,85.

(1-d) asegura que cualquier página indexada, aunque no reciba enlaces, tendrá un PR mínimo de 0,15.

$PR(T_1)$ es el PageRank de un sitio que apunta a A.

$C(T_1)$ es el número de links salientes desde la página 1.

Análisis de la fórmula del PageRank

- Fórmula simple pero iterativa
- Es necesario conocer el PageRank de todas las páginas que apuntan hacia ella
- Según el paper de Google:

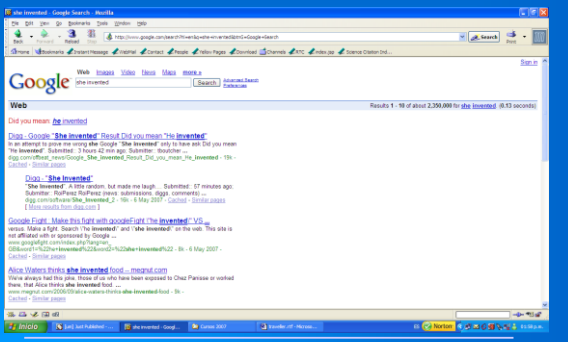
"El PageRank de una Página A puede ser calculado usando un algoritmo iterativo simple y se corresponde con el vector característico de la matriz normalizada de links de la web"

- Se comienza con todas las páginas un PageRank de 1
- Iterando sobre todas las páginas un cierto número de veces
- Los valores comienzan a converger.
- No es un proceso incremental sencillo
- Aproximaciones provisorias para mantener resultados
- Una vez al mes, Google reprocesa toda la información

Manipulación de PageRank

- No es una tarea sencilla, pero es posible
- Miles de sitios se ponen de acuerdo: Google Bombing
- En general se hace para realizar bromas

¿Humor?



Agenda

- Motivación de la tecnología
- Taxonomía
- Web Content Mining
  - Motivación
  - Enfoques
- Web Structure Mining
- Web Usage Mining
  - Motivación
  - Beneficios
  - Aplicaciones
- Text Mining
  - Motivación
  - Aplicaciones

Web Usage Mining

- Implica el descubrimiento automático de patrones de acceso de los usuarios de uno o más servidores Web
- La mayoría de la información que guardan las organizaciones día a día es generada automáticamente por servidores Web y almacenada en logs de acceso (*Web logs*). Otras fuentes: *referrer logs*, que contienen información sobre las páginas que referencian a cada página; registración de usuarios; datos recogidos mediante herramientas como scripts CGI.

Web Usage Mining

Los administradores de sitios Web están muy interesados en preguntas como:

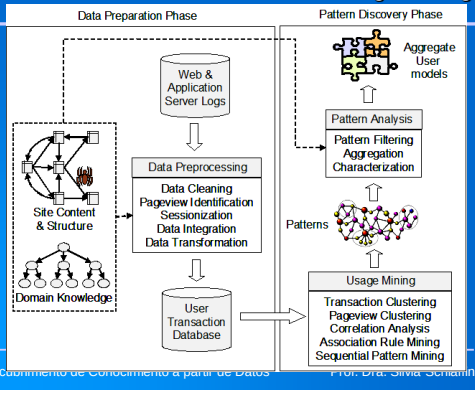
- ¿Cómo usa la gente el sitio?
- ¿Qué páginas están siendo accedidas más frecuentemente?
- ¿Quién está visitando qué documentos?
- Frecuencia de uso de cada hipervínculo
- Uso más reciente de cada hipervínculo

Web Usage Mining

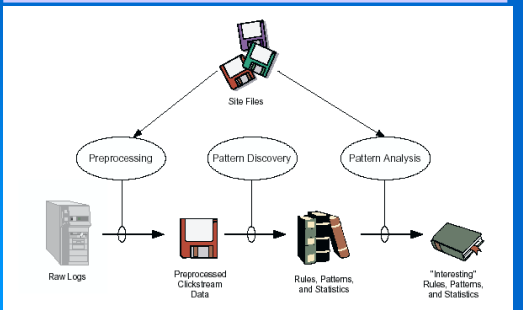
Beneficios para la organización:

- Mejorar la navegación en un sitio
- Mejorar la performance del servidor
- Campañas de publicidad efectivas
- Estructurar mejor un sitio Web para crear una presencia más efectiva de la organización
- Empresas con intranet, administración efectiva de grupos de trabajo y de la infraestructura organizacional
- Dirigir avisos y ofertas a grupos de usuario específicos

Web Usage Mining



Proceso de Web Usage Mining



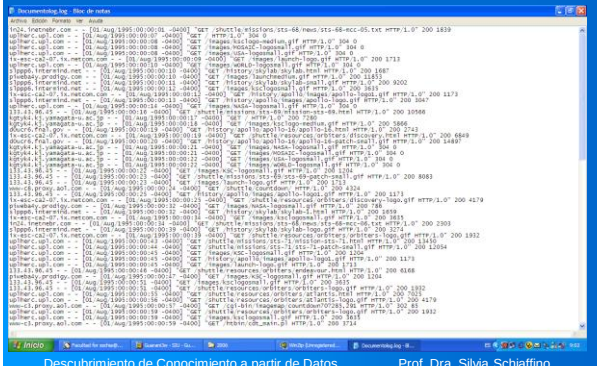
Ejemplo de log

#	IP Address	Userid	Time	Method/ URL/ Protocol	Status	Size	Referrer	Agent
1	123.456.78.9	-	[25/Apr/1998:03:04:41 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
2	123.456.78.9	-	[25/Apr/1998:03:05:34 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.04 (Win95, I)
3	123.456.78.9	-	[25/Apr/1998:03:05:39 -0500]	"GET L.html HTTP/1.0"	200	4130	-	Mozilla/3.04 (Win95, I)
4	123.456.78.9	-	[25/Apr/1998:03:06:02 -0500]	"GET F.html HTTP/1.0"	200	5096	B.html	Mozilla/3.04 (Win95, I)
5	123.456.78.9	-	[25/Apr/1998:03:06:58 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
6	123.456.78.9	-	[25/Apr/1998:03:07:42 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
7	123.456.78.9	-	[25/Apr/1998:03:07:55 -0500]	"GET R.html HTTP/1.0"	200	8140	L.html	Mozilla/3.04 (Win95, I)
8	123.456.78.9	-	[25/Apr/1998:03:09:50 -0500]	"GET C.html HTTP/1.0"	200	1820	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
9	123.456.78.9	-	[25/Apr/1998:03:10:02 -0500]	"GET O.html HTTP/1.0"	200	2270	F.html	Mozilla/3.04 (Win95, I)
10	123.456.78.9	-	[25/Apr/1998:03:10:45 -0500]	"GET J.html HTTP/1.0"	200	9430	C.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
11	123.456.78.9	-	[25/Apr/1998:03:12:23 -0500]	"GET G.html HTTP/1.0"	200	7220	B.html	Mozilla/3.04 (Win95, I)
12	209.456.78.2	-	[25/Apr/1998:05:05:22 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
13	209.456.78.3	-	[25/Apr/1998:05:06:03 -0500]	"GET D.html HTTP/1.0"	200	1680	A.html	Mozilla/3.04 (Win95, I)

Ejemplo de log

IP Address	Timestamp	Method	Requested Page	Protocol	Referrer Page	User Agent
171.64.68.115	10/Apr/2000 01:00:10	GET	http://www-users.cs.umn.edu/~kumar/	HTTP/1.0	-	Mozilla/4.61 [en] (WinNT; I)
171.64.68.115	10/Apr/2000 01:00:10	GET	http://www-users.cs.umn.edu/~kumar/krn.gif	HTTP/1.0	http://www-users.cs.umn.edu/~kumar/	Mozilla/4.61 [en] (WinNT; I)
171.64.68.115	10/Apr/2000 01:00:10	GET	http://www-users.cs.umn.edu/~kumar/book.gif	HTTP/1.0	http://www-users.cs.umn.edu/~kumar/	Mozilla/4.61 [en] (WinNT; I)
160.94.1.194	10/Apr/2000 01:43:05	GET	http://www.cs.umn.edu/robots.txt	HTTP/1.0	-	Ultraseek
160.94.1.194	10/Apr/2000 01:43:05	GET	http://www.cs.umn.edu/~heimdahl/csci5802	HTTP/1.0	-	Ultraseek
212.27.205.29	10/Apr/2000 04:13:27	GET	http://www.cs.umn.edu/robots.txt	HTTP/1.0	-	Mozilla/4.0 (compatible; MSIE 5.0; Windows 95; DigExt)

Ejemplo de log



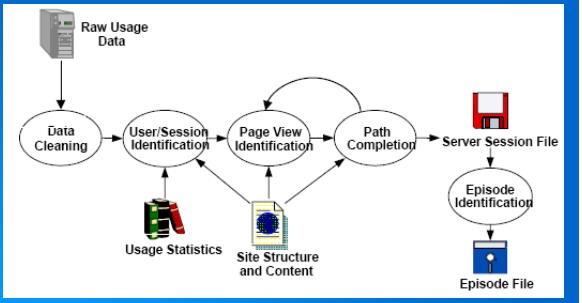
Web Log Data Mining

- Características de datos
- Comparación de clases
- Asociación
- Predicción
- Clasificación
- Análisis de series temporales
- Análisis de tráfico en la Web
  - Análisis de tendencias
  - Análisis de secuencias de eventos típicas
  - Análisis de patrones de comportamiento del usuario

Herramientas

- Descubrimiento de Patrones
  - Usan técnicas de IA, data mining, sicología y teoría de la información para descubrir conocimiento. Ejemplo WEBMINER descubre automáticamente reglas de asociación y patrones secuenciales a partir de logs de acceso a servidores
- Análisis de Patrones
  - Una vez descubiertos los patrones, los analistas necesitan las herramientas apropiadas para comprenderlos, visualizarlos e interpretarlos. Ej: WebViz system.

Pre-procesamiento



Pre-procesamiento: Limpieza de datos

- Técnicas para limpiar un log, eliminar elementos irrelevantes y eliminar "outliers". Ejemplo: entradas con extensiones gif, jpeg, jpg pueden ser eliminadas
- Determinar si hay accesos importantes que no figuran en el log de acceso. Una página que aparece solamente una vez en el log puede haber sido referenciada muchas veces por múltiples usuarios. Solucionar esto usando cookies, cache, y registración de usuarios explícita.
- Otro problema asociado con los proxy es la identificación del usuario. El uso del nombre de la máquina o IP para identificar usuarios puede resultar en que varios usuarios sean agrupados bajo uno solo.

Pre-procesamiento: Identificación de transacciones

- Las secuencias de páginas referenciadas deben ser agrupadas en unidades lógicas que representan transacciones o sesiones de usuario.
- Una sesión de usuario está compuesta por todas las páginas visitadas por un usuario durante una sola visita a un sitio.
- Una transacción difiere de una sesión de usuario en que el tamaño de una transacción puede variar de una única página a todas las páginas referenciadas en una sesión, dependiendo del criterio usado para identificar transacciones.

Tipos de transacciones

- Navegación-contenido
  - La página se usa para navegar hacia otra
  - Cada transacción consiste de una sola referencia a contenido, y todas las referencias de navegación del camino que llevan a esa referencia
- Solo contenido
  - Interesa el contenido de la página
  - Todas las referencias de contenido hechas en una sesión de usuarios. Permiten descubrir asociaciones entre páginas Web.

Identificación de transacciones

- Referencia hacia adelante maximal

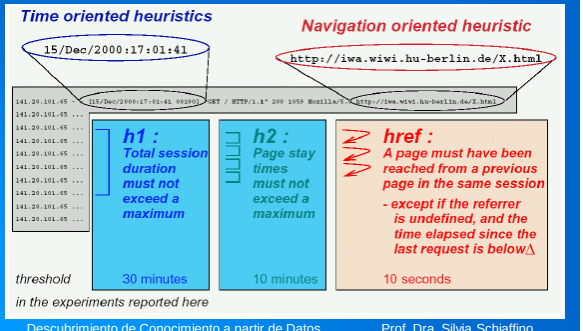
  Cada transacción es el conjunto de páginas en el camino desde la primer página en el log hasta la página anterior a una referencia hacia atrás (ya está en el conjunto).

  Comienza una nueva transacción en la próxima referencia hacia adelante (no está en la lista de páginas aún)

  Ej: A B C D C B E F E G

  3 transacciones: A B C D, A B E F, A B E G

Identificación de sesiones



Enfoque basado en tiempos

User 1

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Session 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C

Session 2

1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Fig. 12.5. Example of sessionization with a time-oriented heuristic

Identificación del usuario

Method	Description	Privacy Concerns	Advantages	Disadvantages
IP Address + Agent	Assume each unique IP address/Agent pair is a unique user	Low	Always available. No additional technology required.	Not guaranteed to be unique. Defeated by rotating IPs.
Embedded Session Ids	Use dynamically generated pages to associate ID with every hyperlink	Low to medium	Always available. Independent of IP addresses.	Cannot capture repeat visitors. Additional overhead for dynamic pages.
Registration	User explicitly logs in to the site.	Medium	Can track individuals not just browsers	Many users won't register. Not available before registration.
Cookie	Save ID on the client machine.	Medium to high	Can track repeat visits from same browser.	Can be turned off by users.
Software Agents	Program loaded into browser and sends back usage data.	High	Accurate usage data for a single site.	Likely to be rejected by users.

Ejemplo de identificación de usuario

Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:10	2.3.4.5	C	-	IE6;WinXP;SP1
0:12	2.3.4.5	B	C	IE6;WinXP;SP1
0:15	2.3.4.5	E	C	IE6;WinXP;SP1
0:19	1.2.3.4	C	A	IE5;Win2k
0:22	2.3.4.5	D	B	IE6;WinXP;SP1
0:22	1.2.3.4	A	-	IE6;WinXP;SP2
0:25	1.2.3.4	E	C	IE5;Win2k
0:25	1.2.3.4	C	A	IE6;WinXP;SP2
0:33	1.2.3.4	B	C	IE6;WinXP;SP2
0:58	1.2.3.4	D	B	IE6;WinXP;SP2
1:10	1.2.3.4	E	D	IE6;WinXP;SP2
1:15	1.2.3.4	A	-	IE5;Win2k
1:16	1.2.3.4	C	A	IE5;Win2k
1:17	1.2.3.4	F	C	IE6;WinXP;SP2
1:26	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k

User 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

User 2

0:10	2.3.4.5	C	-
0:12	2.3.4.5	B	C
0:15	2.3.4.5	E	C
0:22	2.3.4.5	D	B

User 3

0:22	1.2.3.4	A	-
0:25	1.2.3.4	C	A
0:33	1.2.3.4	B	C
0:58	1.2.3.4	D	B
1:10	1.2.3.4	E	D
1:17	1.2.3.4	F	C

Técnicas de descubrimiento de patrones

- Análisis de caminos
- Reglas de asociación
- Patrones secuenciales
- Clustering
- Clasificación

Análisis de caminos

- Hay diferentes tipos de grafos que pueden formarse para realizar análisis de caminos, dado que un grafo representa alguna relación definida sobre páginas Web pages (u otros objectos). Ej: representar el layout físico de un sitio Web con páginas Web como nodos y links de hipertexto entre páginas como arcos dirigidos.
- El análisis de caminos podría usarse para determinar los caminos más frecuentemente visitados en un sitio Web. Otros ejemplos de información que podría descubrirse son:
  - 70% de los clientes que accedieron a /company/products/file2.html lo hicieron comenzando en /company y siguiendo por /company/whatsnew, /company/products, y /company/products/file1.html;
  - 80% de los clientes que accedieron al sitio entraron por /company/products;
  - 65% de los clientes dejaron el sitio después de 4 páginas o menos visitadas.

Reglas de Asociación

- Descubrir las correlaciones entre referencias a varios archivos disponibles en el servidor hechas por un mismo cliente.
- Cada transacción comprende un conjunto de URLs accedidas por un cliente en una visita al servidor. Ejemplo, usando reglas de asociación se pueden encontrar correlaciones tales como:
  - 40% que accedieron a la página con URL /company/products/product1.html, también accedieron a /company/products/product2.html;
  - 30% de los clientes que accedieron a /company/announcements/special-offer.html, colocaron una orden online en /company/products/product1.

Reglas de asociación

Las reglas de asociación descubiertas en logs de acceso a la Web pueden dar una idea de cómo organizar mejor un sitio Web.

El 80% de los clientes que accedieron a /company/products y /company/products/file1.html también accedieron a /company/products/file2.html, pero solamente el 30% de los que accedieron a /company/products también accedieron a /company/products/file2.html.

Es probable en que file1.html haya información que lleve a file2.html, que debiera moverse a un nivel superior para aumentar el acceso a file2.html

Patrones secuenciales

- Encontrar patrones entre transacciones tales que la presencia de un conjunto de elementos es seguida por otro elemento en un conjunto de transacciones ordenadas temporalmente.
- Permite a las organizaciones basadas en la Web predecir los patrones de visita de los usuarios y los ayuda a dirigir las publicidades a grupos de usuarios basándose en estos patrones:
  - 30% de los clientes que visitaron /company/products/, hicieron una búsqueda en Yahoo, dentro de la última semana con palabra clave **w**;
  - 60% de los clientes que colocaron una orden en línea en /company/products/product1.html, también ubicaron una orden en /company1/products/product4 dentro de los 15 días.

Clasificación

- Las técnicas de clasificación permiten desarrollar perfiles de los clientes que acceden a archivos del servidor dados basándose en información demográfica disponible en estos clientes, o basándose en sus patrones de acceso.
- Por ejemplo, la clasificación de accesos Web puede llevar al descubrimiento de relaciones tales como:
  - clientes de agencias del estado o del gobierno que visitaron el sitio tienden a interesarse en la página /company/products/product1.html;
  - 50% de los clientes que colocaron una orden online en /company/products/product2, pertenecen al grupo 20-25 años y viven en la capital.

Clustering

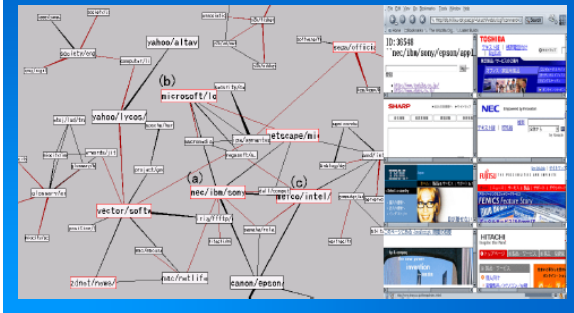
- Esta técnica permite agrupar clientes y elementos de datos que tienen características similares.
- El clustering de información de clientes o de datos en logs de transacciones puede facilitar el desarrollo y ejecución de futuras estrategias de marketing, ya sea online u off-line, tales como envío de correo electrónico a clientes dentro de un determinado cluster, o cambiar dinámicamente un sitio particular para un cliente en una visita posterior basándose en clasificaciones pasadas de ese cliente.

Identificación de comunidades Web

Una comunidad Web es un conjunto de sitios o páginas Web que tienen más links (en cualquier dirección) a miembros de la comunidad que a no miembros.

Permite a los buscadores enfocar una búsqueda de manera efectiva en un conjunto "pequeño" pero temáticamente relacionado de sitios, aumentando su precisión y recall.

Web Community Chart [Toyoda03]



Red social

- Grupos de personas vinculadas mediante lazos profesionales (red informal de colaboradores, colegas y amigos).
- "Six degree separation" la distancia entre dos individuos cualquiera en términos de relaciones personales directas es relativamente pequeña.
- Hay límites a la cantidad de información que una persona está dispuesta a publicar o hacer disponible.

Cadena de referencias

- Buscar en una red social un experto en un determinado tema a partir de la cadena de referencias de quien busca hasta el experto
- La cadena de referencia sirve al experto para acordar responder a alguien haciendo su relación explícita (colaborador mutuo)
- La cadena de referencia provee un criterio para que quien busca pueda evaluar la veracidad o validez del experto.

Referral Web

- Red social modelada como un grafo, donde un nodo representa un individuo y un arco indica una relación directa entre individuos.
- Co-ocurrencia de individuos en documentos públicamente disponibles en la Web:
  - Links en páginas personales
  - Listas de co-autores en artículos y referencias en artículos
  - Intercambios entre individuos registrados en archivos de la red
  - Organigrama de organizaciones (departamento de una universidad)

Referral Web

- Uso:
  - Cuál es mi relación con X?
  - Qué colegas míos , o colegas de colegas conocen sobre “reinforcement learning”?
  - Listar los documentos sobre el tema “reinforcement learning” de gente cercana a X

Node Ranking

- Creación de un ranking de reputación de miembros de una comunidad a través de una red social
- Cada nodo del grafo tiene un grado de autoridad que puede ser visto como una medida de importancia (reputación de un individuo dentro de la comunidad)
- El algoritmo está inspirado en los algoritmos de ranking de páginas Web en una topología Web
- Se tienen en cuenta los arcos entrantes (autoridad de nodos de estos arcos)
- La autoridad se propaga a los nodos hacia donde salen arcos

Agenda

- Motivación de la tecnología
- Taxonomía
- Web Content Mining
  - Motivación
  - Enfoques
- Web Structure Mining
- Web Usage Mining
  - Motivación
  - Beneficios
  - Aplicaciones
- Text Mining
  - Motivación
  - Aplicaciones

Text Mining

- Consiste en buscar regularidades, patrones o tendencias en texto en lenguaje natural. Generalmente implica analizar texto para propósitos específicos
- El objetivo es extraer conocimiento útil a partir de texto estructurado y semi-estructurado
  - Information Extraction (IE)
  - Natural Language Processing (NLP) and Computational Linguistics (CL)
  - Machine Learning (ML)
  - Information Retrieval (IR)
  - (Textual) Databases
  - Knowledge and Information Management
  - Information Visualization

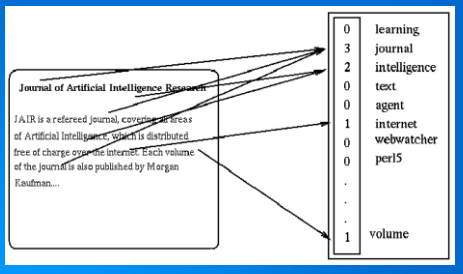
Stop words

- Son palabras que desde el punto de vista no lingüístico no contienen información
- Son dependientes del idioma
  - Inglés: a, about, above, across, after, again, against, all, almost, alone, along, ...
  - Español: de, la, que, el, en, y, a, los, del, se, las, por, un, para, con, no, una, su, al, lo, como, más, pero, sus, ...

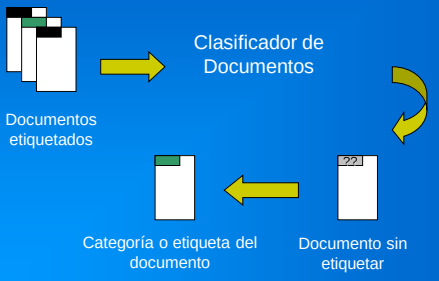
Stemming

- Diferentes variantes de una misma palabra pueden ser problemáticas al realizar análisis de texto, debido a que tienen diferente deletreo pero el mismo significado (aprender, aprende, aprendió, aprenden,...)
- Stemming es el proceso de transformar a una palabra en su raíz (stem)
- Para Inglés no es algo problemático, ya que existen varios algoritmos públicos que funcionan bien. El más conocido es el de Porter. Existen algoritmos para español, francés, portugués, italiano.

Representación de documentos



Categorización de documentos



Categorización automática de documentos

- Problema: se da un conjunto de categorías (de contenido) con los documentos que pertenecen a ellas
- Objetivo: insertar automáticamente un nuevo documento (asignar una o más categorías relevantes al documento)
- Las categorías pueden ser estructuradas (Yahoo!) o no estructuradas (Reuters)
- El problema es similar al de asignar palabras clave a documentos

Ejemplo categorización manual: Yahoo!

Aplicaciones de Text Mining

- Motores de búsqueda para grandes volúmenes de texto
- Categorización automática de documentos y asignación de palabras clave
- Clustering manual de páginas Web
- Segmentación de texto basada en contenido
- Identificación de temas y seguimiento de documentos en el tiempo
- Identificación de lenguaje natural
- Detección de autoría de documentos
- Detección de copia de documentos
- Visualización de datos textuales
- Traducción automática de texto
- Síntesis de discurso

Bibliografía

- Web Data Mining - Exploring Hyperlinks, Contents, and Usage Data - Bing Liu, Second Edition, July 2011 (First Edition, Dec 2006), Springer

http://www.cs.uic.edu/~liub/WebMiningBook.html

Preguntas

?