UNIVERSIDAD DE GRANADA

decsai.ugr.es

Minería de Medios Sociales

DECSAI

Departamento de Ciencias de la Computación e Inteligencia Artificial

UNIVERSIDAD DE GRANADA

decsai.ugr.es

Bloque II: Minería de Texto y de la Web

DECSAI

Departamento de Ciencias de la Computación e Inteligencia Artificial

UNIVERSIDAD DE GRANADA

decsai.ugr.es

Sesión II.2: Minería Web

DECSAI

Departamento de Ciencias de la Computación e Inteligencia Artificial

Minería Web

Descubrimiento de contenido en documentos web (*web content mining*)

Descubrimiento de patrones en las relaciones entre documentos hipertexto y los links (*web structure mining*)

Descubrimiento de patrones en los accesos a servidores (*web usage mining*)

Minería Web del Contenido

- Tiene los mismos principios que la recuperación de información (multimedia) basada en contenido.

  P.ej. Dame todas las imágenes/documentos sobre coches

- Es básicamente igual al descubrimiento de asociaciones en text mining, sólo que con documentos de la web.

Minería Web de la Estructura

- Se analiza la estructura de un sitio web

- Se tienen en cuenta relaciones:
  - *intra-páginas:* estructura de la página en HTML o XML para extraer información
  - *inter-páginas:* se estudia la relación mediante links entre diversas páginas

Minería Web del Uso

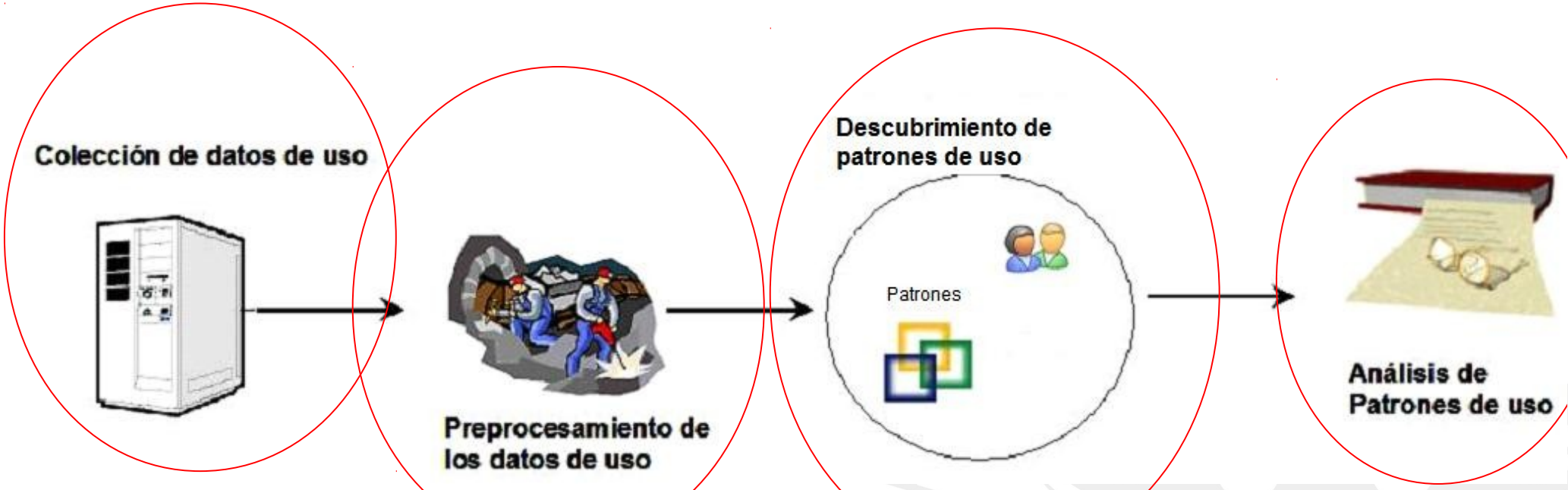
- Análisis de los archivos *logs* de datos de acceso de usuarios a un determinado servidor
- También se estudia el comportamiento en la navegación de las páginas a las que accede
- Tiene fines de marketing, generalmente
- Se extraen transacciones de los ficheros y se obtienen reglas de asociación.
- También se puede utilizar clustering

Minería Web del Uso

Tareas antes de realizar Minería

- Fusión y sincronización de los datos de varios ficheros de log
- Limpieza de datos
  - Eliminar accesos a ficheros CSS, audio y gráficos
  - Ignorar el acceso de buscadores/crawlers
- Identificación de páginas vistas.
- Identificación de sesiones
- Organizar el flujo de clicks (clickstream)
- Identificación de usuarios a través de varios accesos
  - La IP no es suficiente. Pueden usarse cookies y el campo UserAgent

Minería Web del Uso



Servidor Web

Servidor Proxy

Máquina del Usuario

Hetereogenidad y Carencia de Estructura

Forma Intermedia

Técnicas de Minería

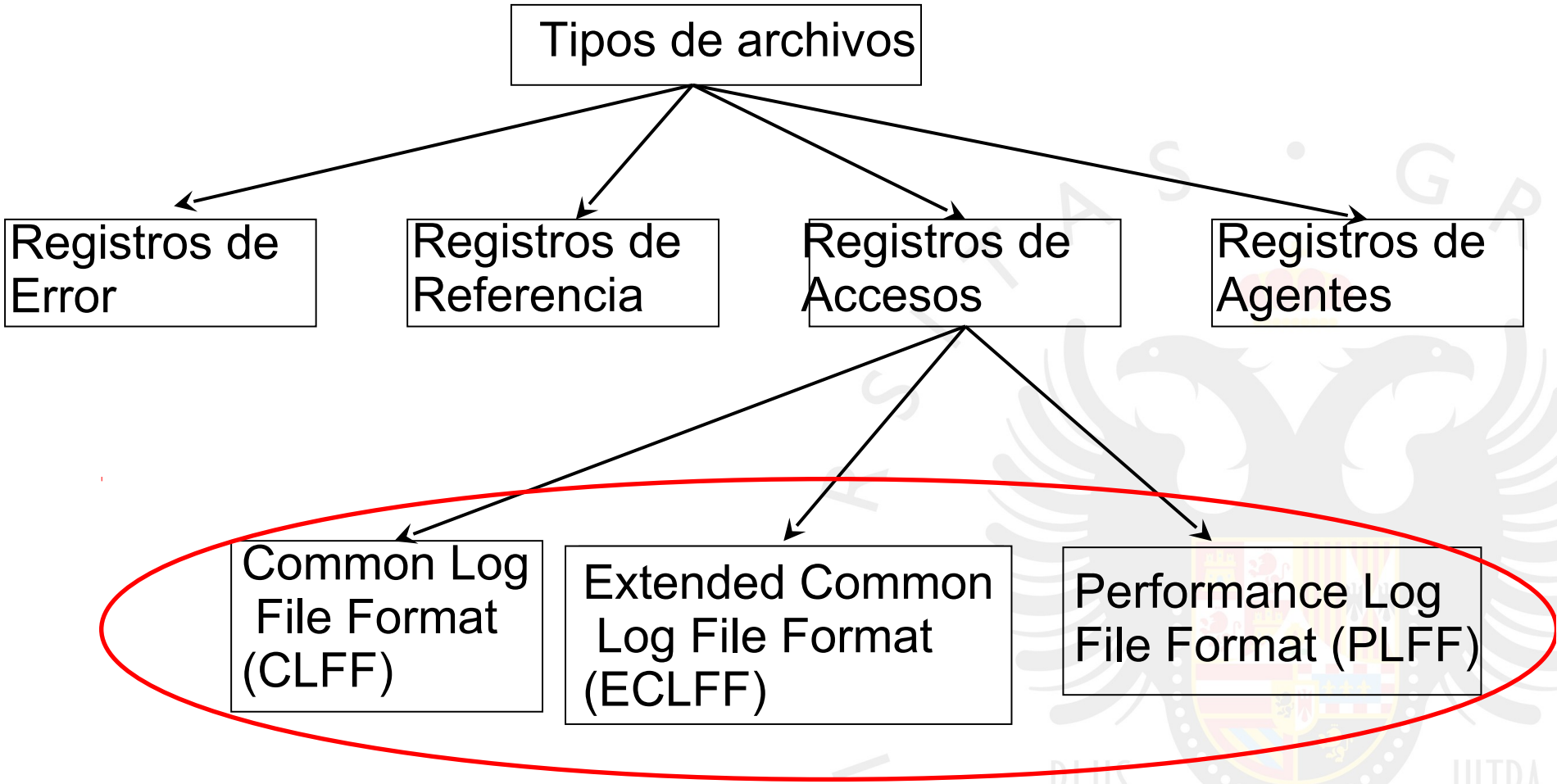
- Clustering
- Reglas de Asociación

Filtrar reglas o patrones

Construcción de perfiles

Sistemas WebMiner y WebSift

Ficheros de log



Ficheros de log

Common Log File Format (CLFF)

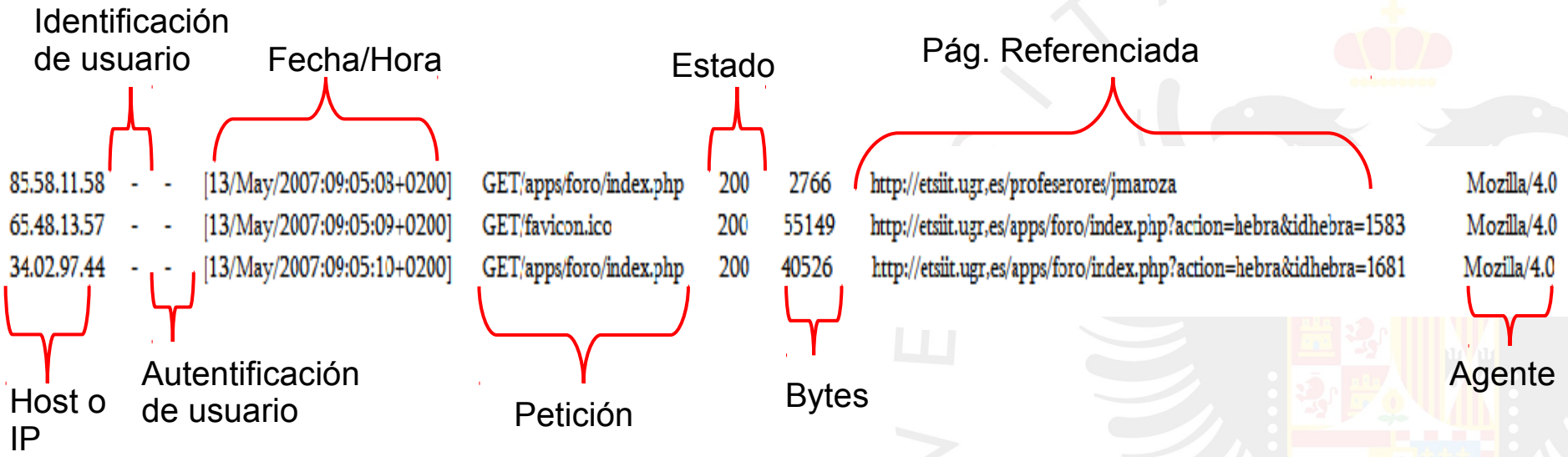
Host o IP	Identificación Usuario	Autentificación Usuario	Fecha / Hora	Petición	Estado	Byte
maquina.uji.es	-	-	[9/Feb/2016:00:56:56 +0100]	"GET /documento.html HTTP/1.0"	302	64

Extended Common Log File Format (ECLLF)

Host o IP	Id. Usuario	Aut. Usuario	Fecha / Hora	Petición	Estado	Byte	Referencia	Agente
maquina.uji.es	-	-	[9/Feb/2016:00:56:56 +0100]	"GET /documento.html HTTP/1.0"	302	64	http://www.skyweb.com	"Mozilla/46.0.1 (Win7; I)"

Ficheros de log

- Archivo Extended Common Log File Format (ECLFF)



Modelo de datos

- Una página web $p_i \in P$, $P = \{p_1, p_2, ..., p_n\}$ es un documento HTML identificable a través de la red con una URL.

- Una sesión de usuario $s_j \in S = \{s_1, s_2, ..., s_m\}$ se define como una secuencia de accesos temporales a un sitio particular de la Web por un usuario. Viene determinada por un conjunto de clicks $R = \{r_1, ..., r_q\}$

en el sitio web, denominado clickstream.

- La relación entre una página $p_i \in P$ y la sesión $s_j \in S$ en la que se visita viene dada por la matriz UP, donde cada valor representa el tiempo de la permanencia del usuario en la página p_i en la sesión s_j

$$UP = [w(s_i, p_j)], 1 \leq i \leq m, 1 \leq j \leq n$$

Modelo de datos (cont.)

- Caracterizaremos el clickstream para la identificación de sesiones de usuarios (Método timeout [Chen et al., 1996]):

  - Sea r_k el k^{th} click del clickstream R de una dirección IP en t_k segundos.
  - Sea r_{k+1} el $(k+1)^{th}$ click del clickstream de la misma dirección IP en t_{k+1} segundos después del click r_k.
  - T es el tiempo de espera calculada como la diferencia entre ambos click en un sitio web es $T = (t_{k+1} - t_k)$
  - Si $T < \beta$, siendo β el tiempo de espera máximo, entonces el click r_k y r_{k+1} son considerados partes de la sesión S_i. En otro caso, si $T > \beta$, entonces el click r_k es estimada como final de la sesión S_i, mientras el click r_{k+1} es el clickstream de la sesión S_{i+1}.

Modelo de datos (cont.)

- Identificación de sesiones de usuario:

IP o Host	Id sesión	Fecha/Hora	Tiempo
33.red-83-33-.dynamicip.rimade.net	1	[18/Jun/2006:07:41:14+0200]	0
ij2591.inktomisearch.com	2	[18/Jun/2006:07:41:20+0200]	0
70.42.51.20	3	[18/Jun/2006:07:41:35+0200]	0
33.red-83-33-8.dynamicip.rima-tde.net	1	[18/Jun/2006:07:41:39+0200]	25

clickstream	Click
/alumnos/mlii/prolog	1
/download/guia	1
/proyectos/silviaaeid/basd	1
/alumnos/oscpi/fecha	2

Patrones de navegación

- **Objetivo**

  Obtener patrones de navegación del usuario y así tener una mejor descripción de su comportamiento en la web y de esa manera saber realmente lo que sucede en el sitio web.

- **Problema**

  - Problema de navegación temporal: intervalo horario – página visitada.
  - Problema de navegación intrapágina: pagina visitada - página referenciada.

Patrones de navegación: Reglas de asociación difusas

- **Definición:** Dado I un conjunto de ítems, definiremos una transacción difusa $\tilde{\tau}$, donde $\tilde{\tau} \subseteq I$. Sea $\tilde{\tau}(i)$ el grado de pertenencia de i a $\tilde{\tau}$ y notaremos a $\tilde{\tau}(I_0)$el grado de inclusión de un ítemset en una transacción difusa definida como

$$\tilde{\tau}(I_0) = \min_{i \in I_0} \tilde{\tau}(i)$$

- **Medidas de interés:** hemos utilizado diferentes medidas para la obtención de las reglas de asociación difusas, tanto medidas objetivas como subjetivas
- **Algoritmo:** AprioriTID

Modelo asociado a la navegación temporal - Fecha-Página visitada

- Problema 1: problema de navegación temporal.

- Ítems: fecha/hora y páginas visitadas
- Transacciones: tablas transaccionales difusa para la obtención de las reglas, para los ítems fecha/hora y páginas visitadas.

Modelo asociado a la navegación temporal - Fecha-Página visitada

Hora.	Peso	Etiqueta
08:30	1.0	Mañana
12:45	0.5	Medio Día
15:25	0.4	Tarde
20:20	0.3	Noche

IP/Pag.	Madrugada	Mañana	Mediodía	Tarde	Noche	Pag_1	Pag_2	Pag_3
IP_1	**0**	**1.0**	**0**	**0**	**0**	**0.4**	**0**	**0.8**
IP_2	**0**	**0**	**0.5**	**0**	**0**	**0**	**0**	**0.4**
IP_3	**0**	**0**	**0**	**0.4**	**0**	**0.7**	**0.3**	**0**
IP_4	**0**	**0**	**0**	**0**	**0.3**	**0.2**	**0**	**0**

Modelo asociado a la navegación temporal - Fecha-Página visitada

Fecha/Hora → Página Visitada.

Mañana→
http://www.shop2.cz/ls/index.php?\&id=98\&filtr=102

Soporte =60%; confianza =1.0; FC =1.0

-Interpretación: del conjunto analizado el 60% presentaba esta regla, la cual nos indica que los usuarios se conectan por la mañana a esa página.

Modelo asociado a la navegación entre páginas - Página visitada – Página referenciada

- Problema 2: problema de navegación entre páginas.

- Ítems: páginas visitadas y páginas referenciadas.
- Transacciones: tablas transaccionales difusa para la obtención de las reglas, para los ítems páginas visitadas y páginas referenciadas:

Modelo asociado a la navegación entre páginas - Página visitada – Página referenciada

Frecuencia

IP/Pag.	Pag_1	Pag_2	Pag_3	Pag_4
IP_1	0	4	0	7
IP_2	7	0	8	0
IP_3	6	0	2	0
IP_4	0	3	0	10

Peso

IP/Pag	Pag_1	Pag_2	Pag_3	Pag_4
IP_1	0	0.4	0	0.7
IP_2	0.7	0	0.8	0
IP_3	0.6	0	0.2	0
IP_4	0	0.3	0	1

Obtención de los pesos

Modelo asociado a la navegación entre páginas - Página visitada – Página referenciada

página visitada → página referenciada

/dt/?c=11670 → http://www.shop2.cz

- Soporte =40%; confianza =1.0; FC =1.0

Interpretación: esto indica que los usuarios vistan a la página /dt/?c=11670 y luego se van a la página http://www.shop2.cz, esta regla se encuentra en un 40% dentro del conjunto analizado.

Modelo asociado a la navegación entre páginas - Página visitada – Página referenciada

Conjuntos de datos	Entrada de datos originales	Entrada de datos preprocesadas	Preprocesamiento
Conjunto 1	100900	100810	Eliminación entradas idénticas
Conjunto 2	100810	46950	Eliminación entradas sin el campo de referencia
Conjunto 3	46950	16518	Eliminación de imágenes
Conjunto 4	16518	12910	Eliminación javascript
Conjunto 5	98202	15676	Preprocesamiento completo

Modelo asociado a la navegación entre páginas - Página visitada – Página referenciada

N° Regla	Reglas Obtenidas
Regla 1	GET/apps/tablon/ → http://etsiit.ugr.es
Regla 2	GET/apps/foro/index.php → http://etsiit.ugr.es
Regla 3	GET/apps/foro/index.php?idforo=asignaturas → http://etsiit.ugr.es/apps/foro/index.php
Regla 4	GET/apps/foro/index.php? action=foro&idforo=escuela → http://etsiit.ugr.es/apps/foro/index.php
Regla 5	GET/apps/foro/index.php?idforo=general → http://etsiit.ugr.es/apps/foro/index.php

Modelos de Markov

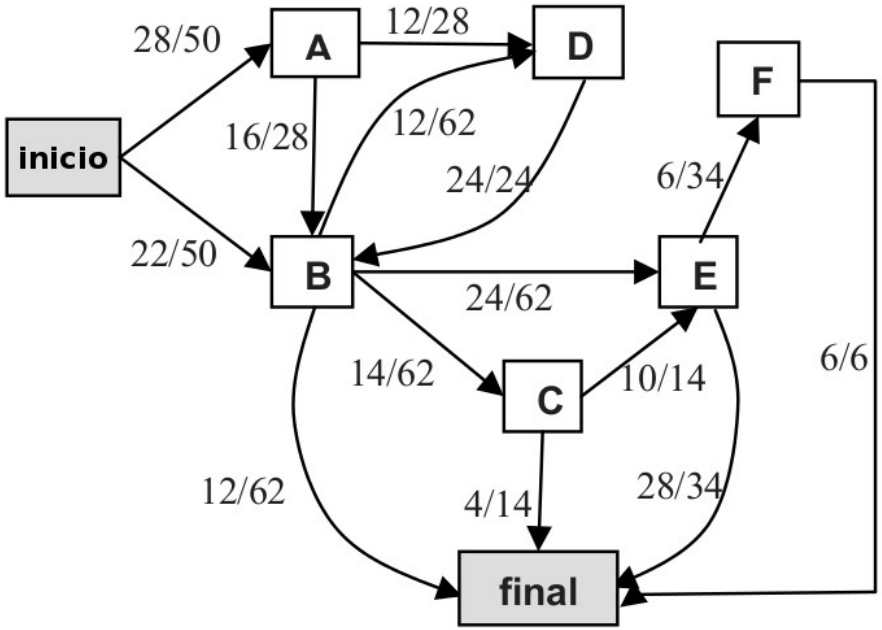
Página visitada – Página referenciada

Los modelos de Markov vienen definidos por un conjunto de estados $\{s_1, s_2, \ldots, s_n\}$ y una matriz $[Pr_{i,j}]_{nxn}$ de probabilidades de transición entre dichos estados.

Se utilizan principalmente para modelado predictivo basado en secuencias contiguas de eventos.

En este caso podemos considerar una página como un estado, y las transiciones entre páginas tendrán una probabilidad determinada.

Transacción	Frecuencia
A, B, E	10
B, D, B, C	4
B, C, E	10
A, B, E, F	6
A, D, B	12
B, D, B, E	8



UNIVERSIDAD DE GRANADA

decsai.ugr.es

Sesión II.2: Minería Web de uso: clustering y perfiles de usuario

Departamento de Ciencias de la Computación e Inteligencia Artificial

Contenidos

- Tecnología de minería web de uso en un portal web
- Clustering de sesiones de usuario
- Definición de perfiles de usuario en XML
- Obtención de perfiles de usuario a partir del clustering

Usuarios de negocios/marketing

- Los usuarios de negocios en las empresas son los que tratan con la información extraída de los procesos de minería
  - Identificación de grupos sociales
  - Relación entre los usuarios (clientes) y sus áreas de interés
- Mala comunicación entre los analistas de minería de datos y los usuarios de negocios
- Necesidad de software específico para el desarrollo de herramientas de minería para usuarios de negocios

Perfiles de usuario en la web

- Se almacena información sobre el comportamiento de los usuarios al navegar
  - Información de minería para el usuario
  - Información de minería para el usuario de negocios o de marketing

Perfiles de usuario en la web

- *Información de minería para el usuario*

  En sitios web comerciales, la empresa puede extraer las preferencias del usuario y personalizar el sitio web, ofreciendo al usuario algunos productos de acuerdo a sus preferencias.

Perfiles de usuario en la web

- *Información de minería para usuarios de negocios o de marketing*
  - Proceso de Clustering para agrupar perfiles de usuarios por áreas de interés
  - Se puede generar un conjunto de reglas para realizar inferencia de la relación entre usuarios y términos
  - Conexión de usuarios a grupos sociales basados en información extraída de los meta-datos de la página web

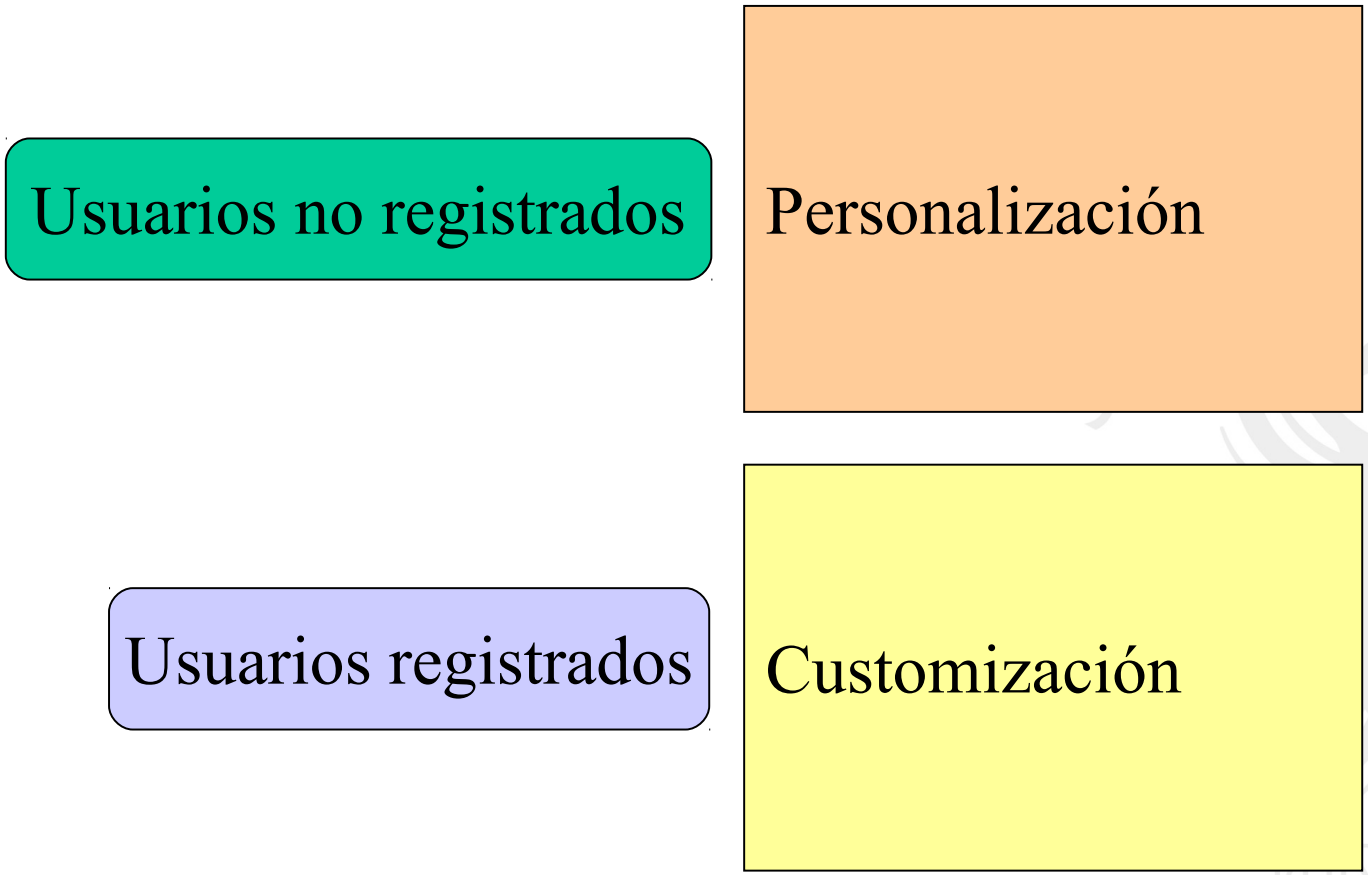
Usuarios registrados y no registrados

- ¿Identidad del usuario?

- Los usuarios con la misma dirección IP tienen diferentes perfiles

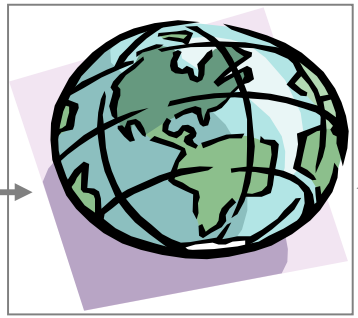
- Los usuarios pueden proporcionar nicknames

Usuarios registrados y no registrados



Procesamiento de la actividad de ratón
(clickstream)

Usuario



Web

Servidores de Web y aplicaciones

Servidor de contenidos

Identificación de Sesión

Actividad de página

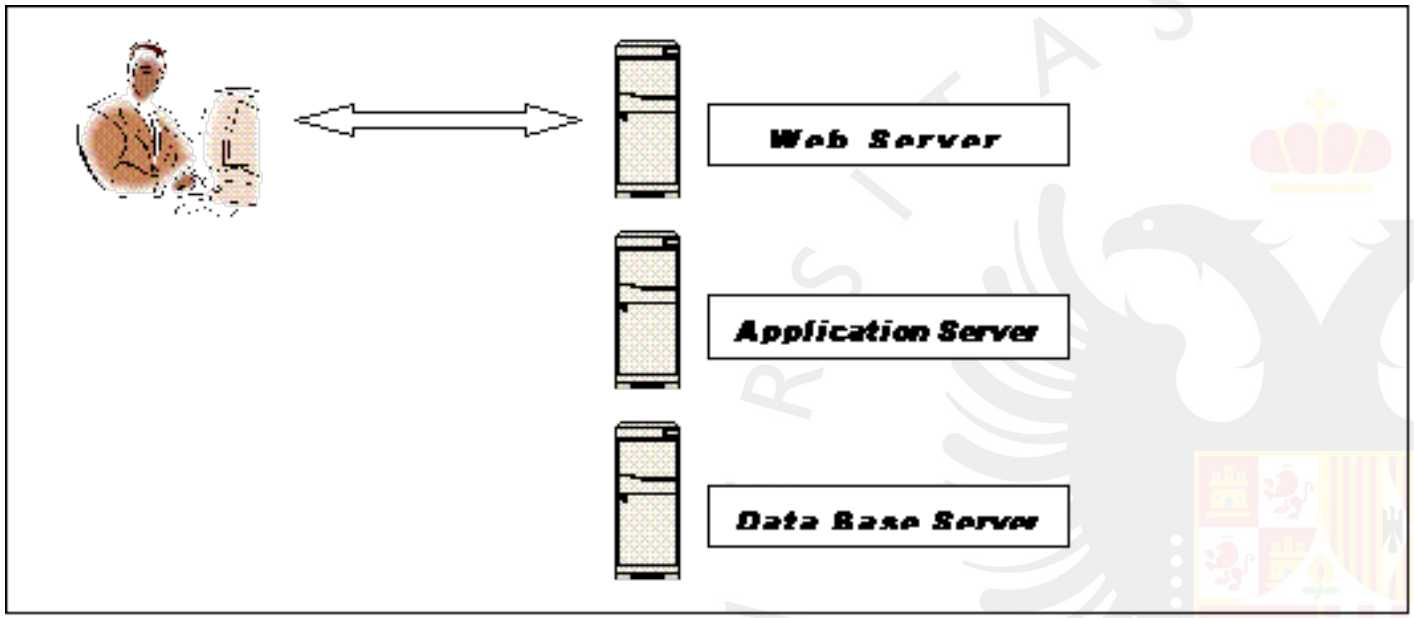
Descriptores de contenido

Identificación y clasificación del usuario (Perfil Extendido)

Base de Datos para análisis

Tecnología de minería en Web en un portal

- El área del portal



- El área de análisis de datos

Tecnología de minería en Web en un portal

- El área de análisis de datos
  - Entrada de datos:

    Explícita: Formularios, Opiniones, registros,...

    Implícita: Cookies, ficheros de log
  - Generación de datos:

    Ficheros de logs

    Ficheros de metadatos

- Objetivo.

  Obtener grupos de sesiones de usuarios que navegan por la web con características similares.

Clustering en minería web de uso

- Objetivo:

  Obtener grupos de sesiones de usuarios que navegan por la web con características similares.

- Modelo de datos asociado:
  - Conjunto de sesiones S definida como: $S = \{s_1, s_2, \ldots, s_m\}$
  - Conjunto de páginas P, definida como: $P = \{p_1, p_2, \ldots, p_n\}$
  - Matriz sesión-página m x n : $UP = [w(s_i, p_j)], 1 \leq i \leq m, 1 \leq j \leq n$

- A partir de esta matriz de peso sesión-página obtendremos una matriz de semejanza entre las sesiones definidas

$$SS = [sim(s_i, s_j)], i \leq j \leq m$$

y para obtener esta semejanza aplicaremos la medida del coseno y coseno extendido.

Medidas de semejanza

- Coseno. $$S_{1,kl} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} S_i^k \cdot S_j^l}{\sqrt{\sum_{i=1}^{N} S_i^k}\sqrt{\sum_{j=1}^{N} S_j^l}}$$

- Coseno extendido. $$S_{2,kl} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} S_i^k S_j^l Sn(i,j)}{\sqrt{\sum_{i=1}^{N} S_i^k}\sqrt{\sum_{j=1}^{N} S_j^l}}$$

$$Sn(i,j) = \min\left(\frac{|p_i \cap p_j|}{\max\left(1, \max\left(|p_i| \cdot |p_j| - 1\right)\right)}\right)$$

Clustering en minería web de uso

- Conjuntos de datos:

Conjuntos de datos	Entrada de datos originales	Entrada de datos preprocesadas	N° Sesiones
Conjunto 1	100900	12910	2024
Conjunto 2	98202	15676	2780

- Medidas: coseno y coseno extendido.
- Técnica: clustering difuso con el algoritmo c-medias difuso.
- Número de particiones iniciales: 12 (obtenidos mediante el análisis jerárquico previamente)

Clustering en minería web de uso

Nº Cluster	Sesiones (Grado de Pertenencia)	Sesión Centroide
Cluster 0	2 (0.95) 437 (0.98) 508 (0.98) 512 (0.96)	GET/apps/tablon GET/apps/foro/index.php GET/apps/foro/index.php?action=foro&idforo=escuela GET/apps/foro/index.php?action=foro&idforo=general GET/apps/foro/index.php?action=hebra&idhebra=1920 GET/apps/foro/index.php?action=foro&idforo=asignaturas GET/apps/foro/index.php?action=hebra&idhebra=1937 GET/apps/foro/index.php?action=hebra&idhebra=1920 GET/apps/foro/index.php?action=hebra&idhebra=1916H
Cluster 3	21 (1.00) 65 (1.00) 6 (1.00) 51(1.00) 136 (1.00) 13 (1.00) 68 (0.939) 569 (0.939)	 GET/js/protWindows/themes/default.css GET/apps/foro/index.php GET/apps/tablon GET/page.php?pageid=departamentos GET/apps/foro/index.php?action=hebra&idhebra=1583 GET/apps/foro/index.php?action=hebra&idhebra=1874 GET/apps/foro/index.php?action=foro&idforo=escuela GET/apps/foro/index.php?action=hebra&idhebra=1709 GET/apps/foro/index.php?action=foro&idforo=general

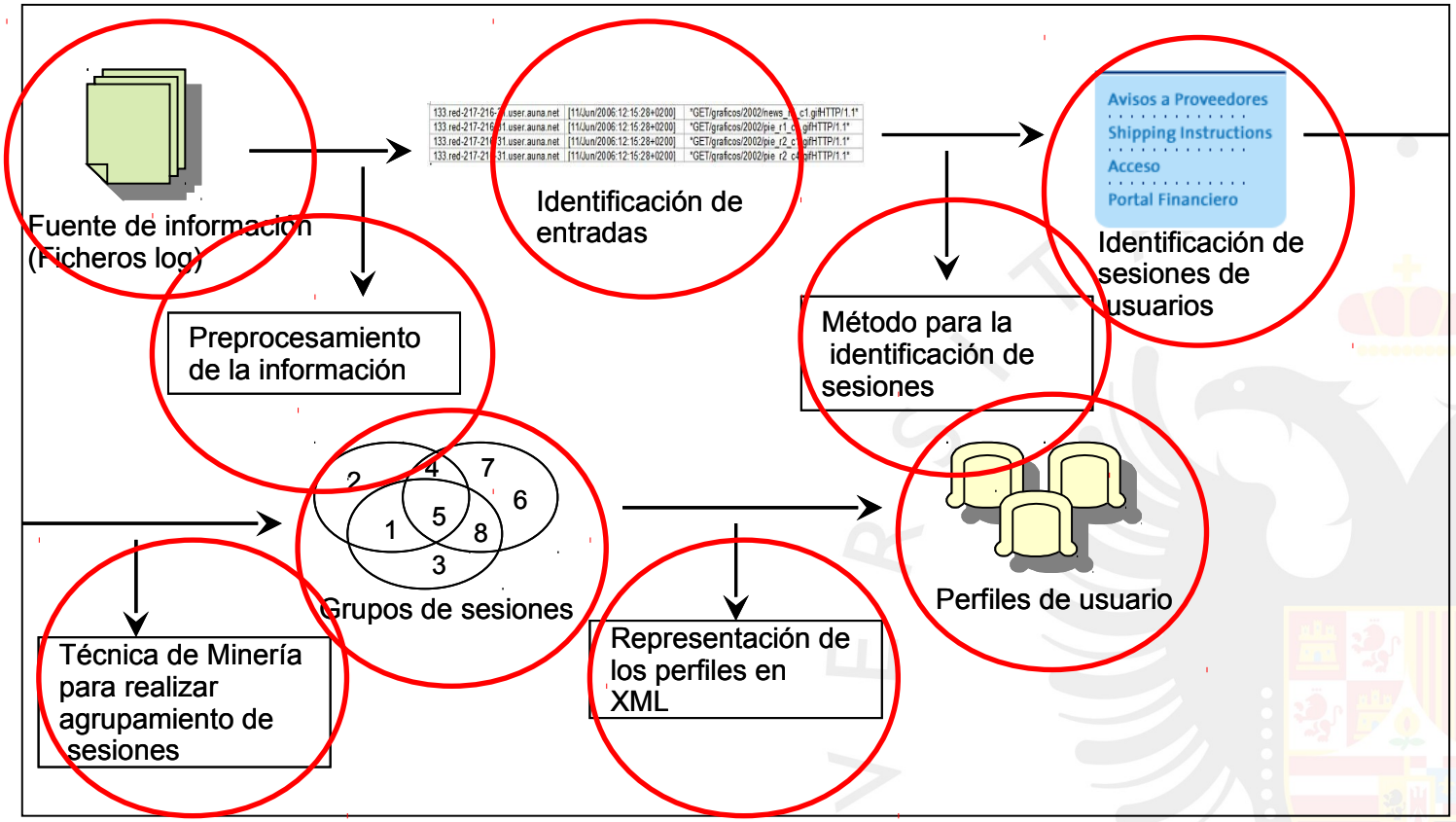
Coseno extendido

Nº Cluster	Sesiones (Grado de Pertenencia)	Sesión Centroide
Cluster 0	2 (0.75) 437 (0.90) 508 (0.90) 512 (0.85)	GET/apps/tablon GET/apps/foro/index.php GET/apps/foro/index.php?action=foro&idforo=escuela GET/apps/foro/index.php?action=foro&idforo=general GET/apps/foro/index.php?action=hebra&idhebra=1920 GET/apps/foro/index.php?action=foro&idforo=asignaturas GET/apps/foro/index.php?action=hebra&idhebra=1937 GET/apps/foro/index.php?action=hebra&idhebra=1920 GET/apps/foro/index.php?action=hebra&idhebra=1916H
Cluster 3	21 (0.97) 65 (0.97) 6 (0.97) 51(0.97) 136 (0.97) 13 (0.97) 68 (0.85) 569 (0.85)	 GET/js/protWindows/themes/default.css GET/apps/foro/index.php GET/apps/tablon GET/page.php?pageid=departamentos GET/apps/foro/index.php?action=hebra&idhebra=1583 GET/apps/foro/index.php?action=hebra&idhebra=1874 GET/apps/foro/index.php?action=foro&idforo=escuela GET/apps/foro/index.php?action=hebra&idhebra=1709 GET/apps/foro/index.php?action=foro&idforo=general

Coseno

- Discusión de los resultados: la medida del coseno extendido da mejores resultados con respecto a los centroides

Obtención de perfiles de usuario en minería web de uso



UNIVERSIDAD DE GRANADA

decsai.ugr.es

Visualización en Minería de Textos y de Web

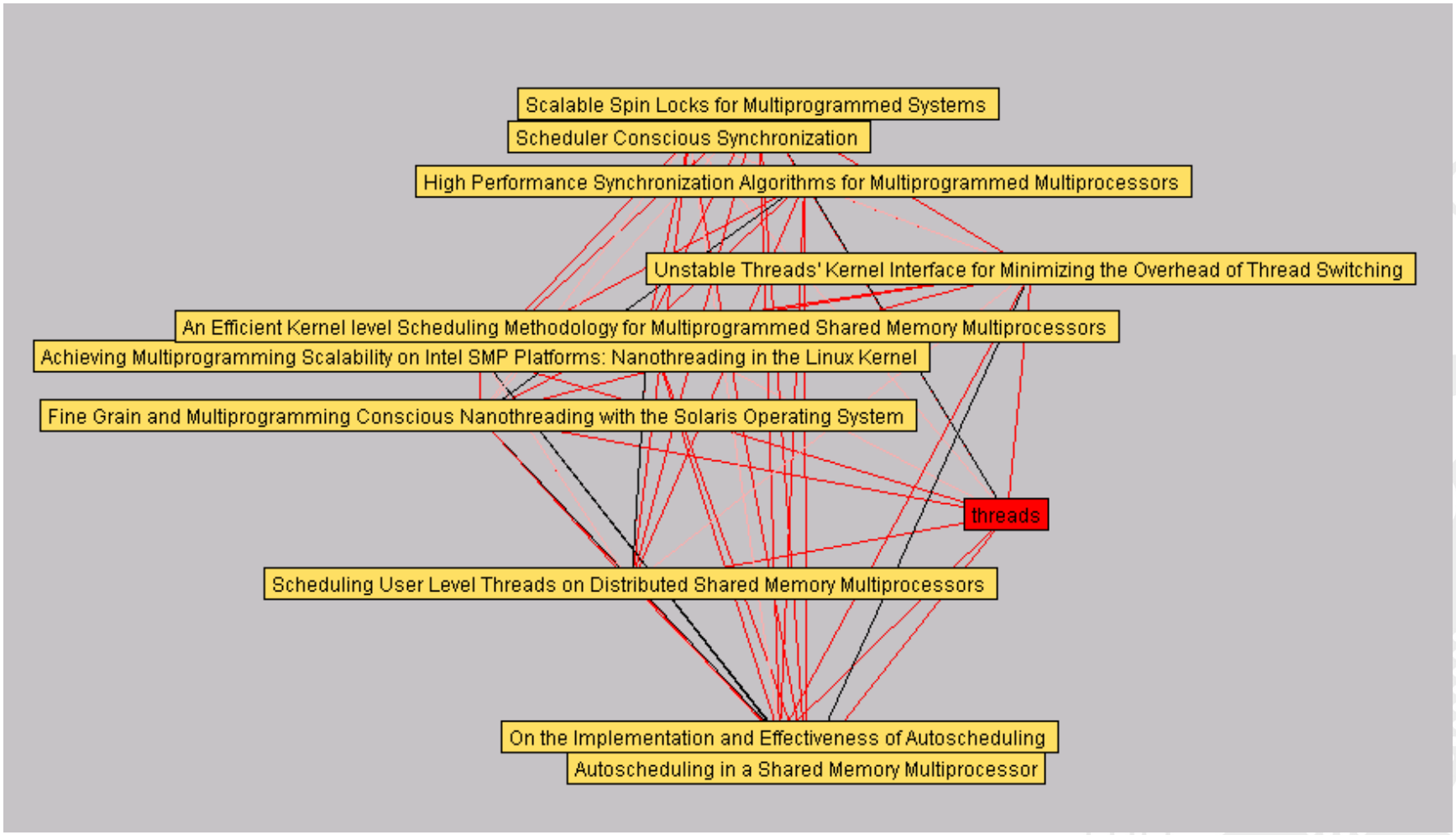
DECSAI

Departamento de Ciencias de la Computación e Inteligencia Artificial

Visualización

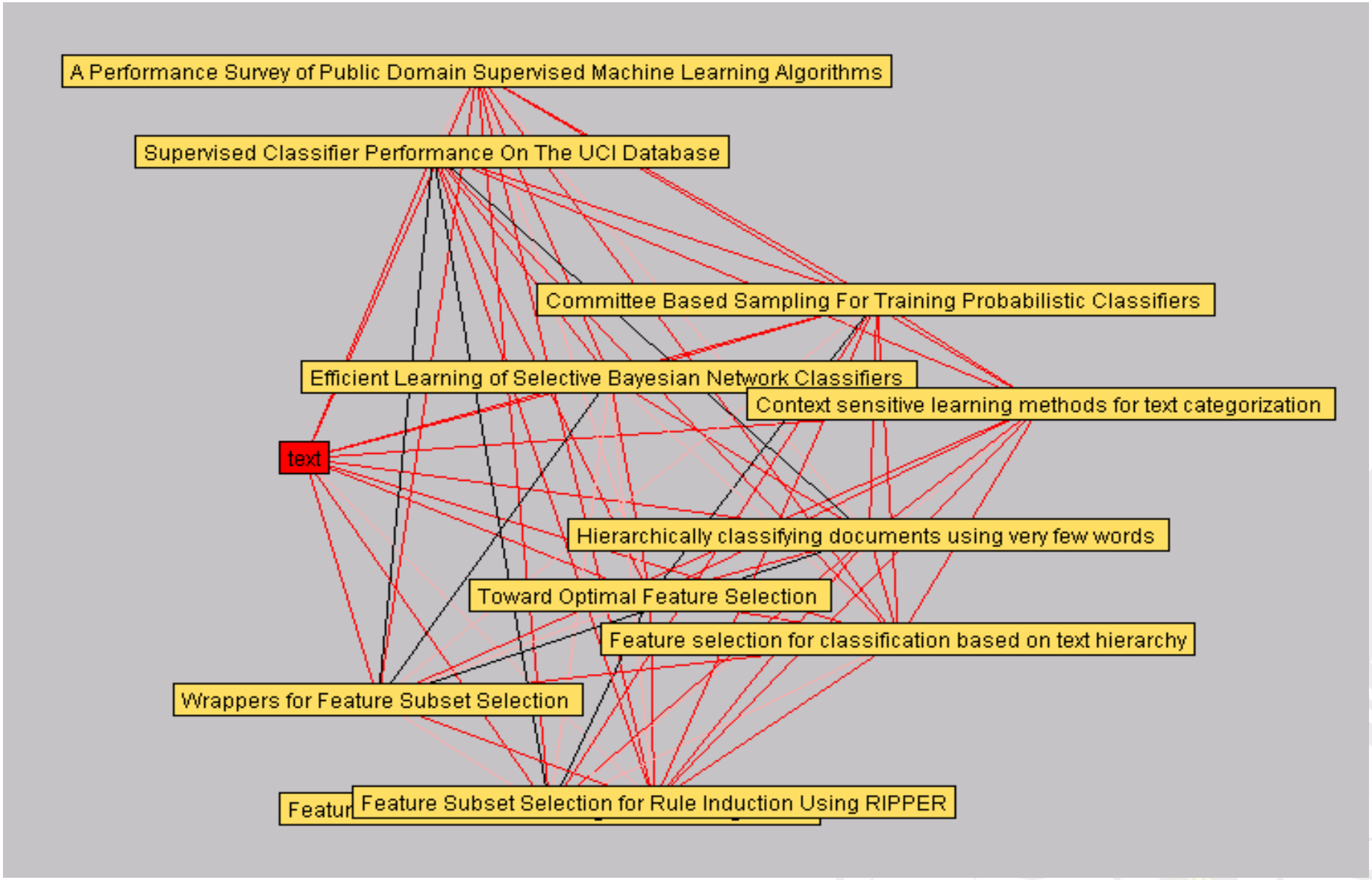
- Clustering
- Reglas de asociación
- Tag Clouds

Visualización de clústeres



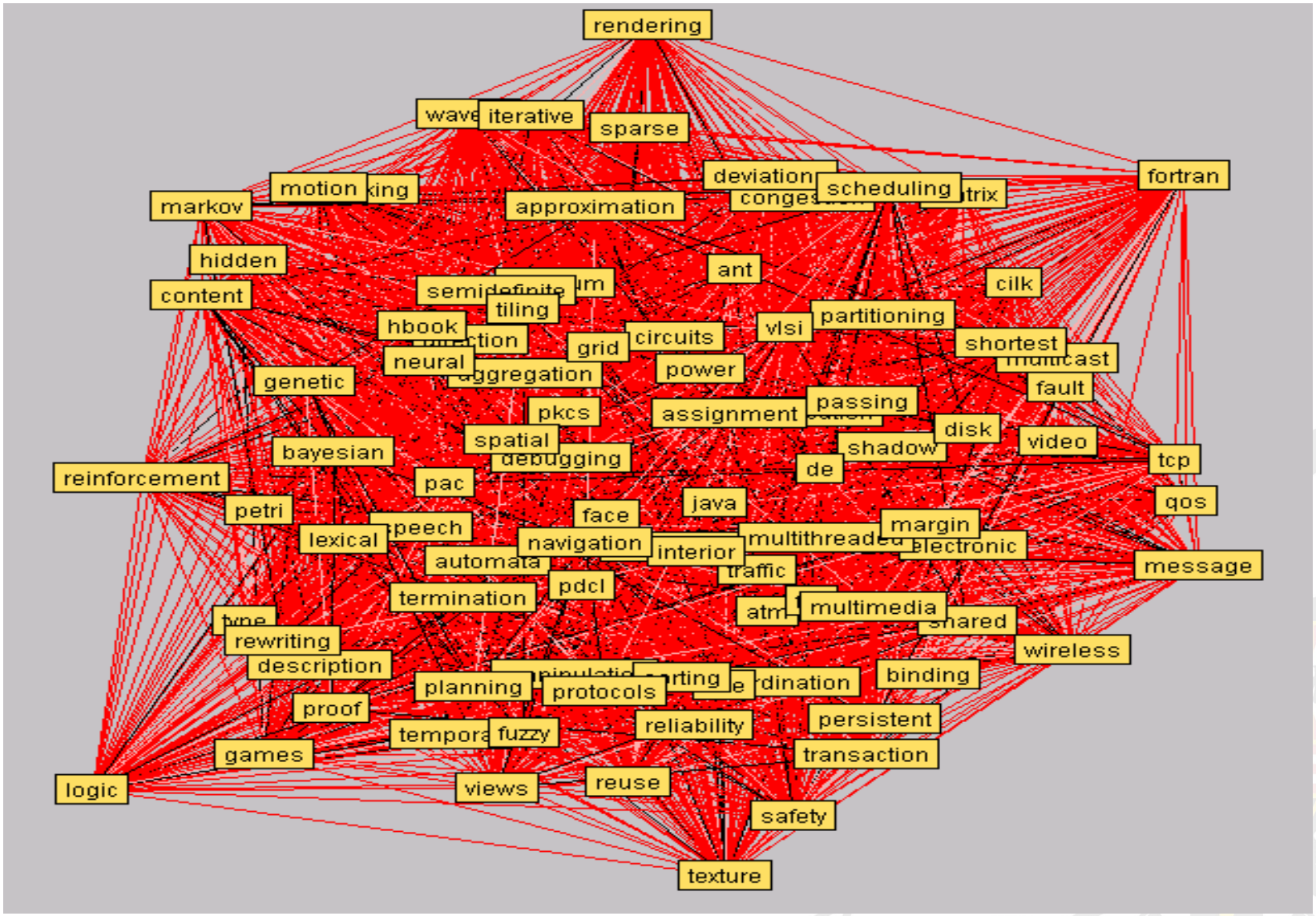
Fuente: http://nlp.stanford.edu/courses/cs224n/2003/fp/millersj/cs224nfp.pdf

Visualización de clústeres



Fuente: http://nlp.stanford.edu/courses/cs224n/2003/fp/millersj/cs224nfp.pdf

Visualización de clústeres



Fuente: http://nlp.stanford.edu/courses/cs224n/2003/fp/millersj/cs224nfp.pdf

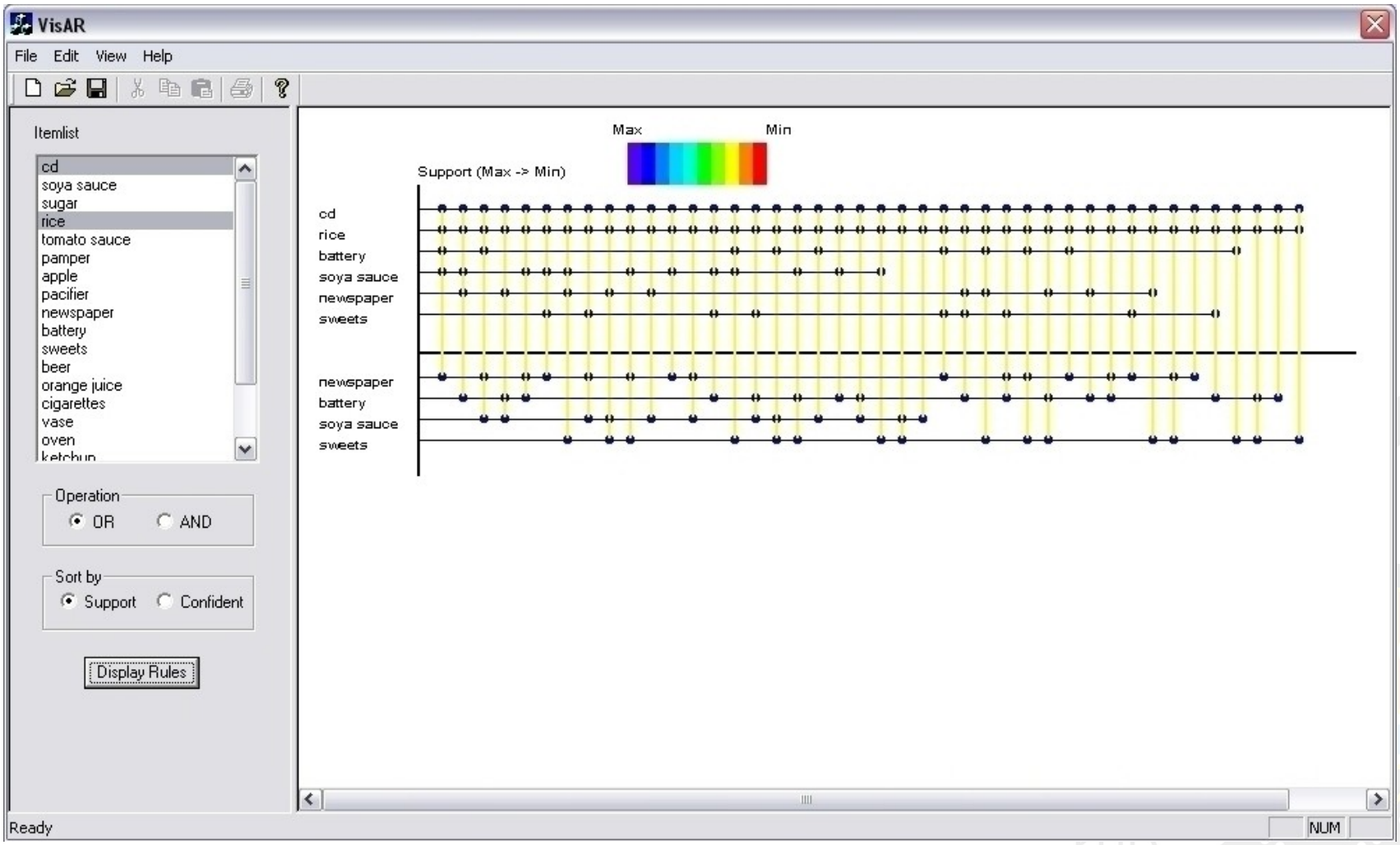
Visualización de clústeres

- Ejemplos:
  - Visualizador de Clústeres de Tweets

    http://www.csc.ncsu.edu/faculty/healey/tweet_viz/

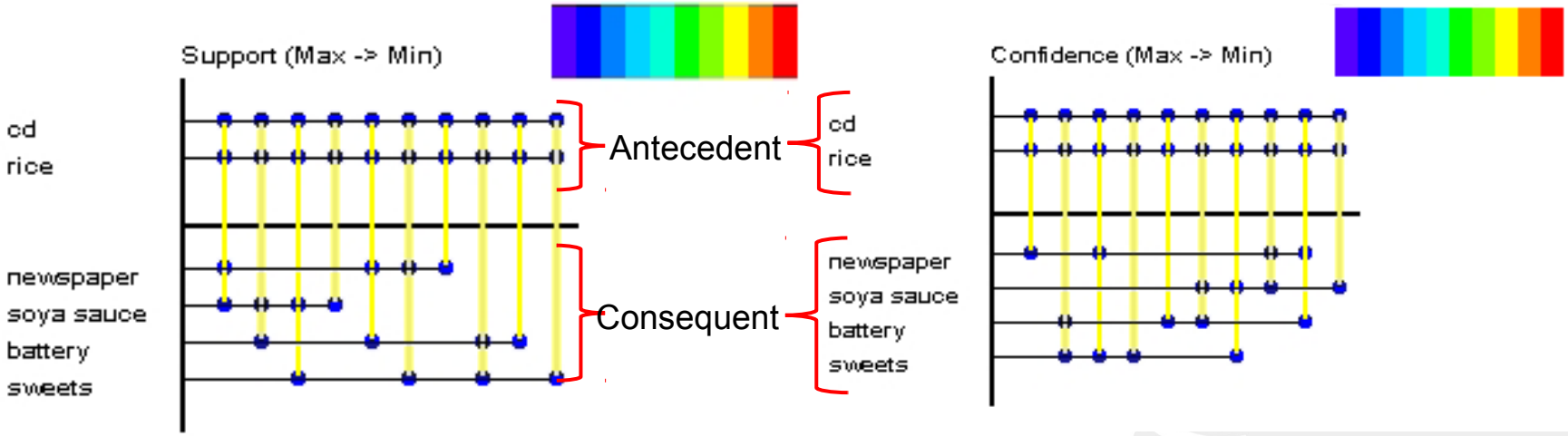
    http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

Visualización de reglas de asociación



Fuente: K. Techapichetvanich and A. Datta. VisAR : A New Technique for Visualizing Mined Association Rules. X. Li, S. Wang, and Z.Y. Dong (Eds.): ADMA , LNAI 3584, pp. 88–95, 2005.

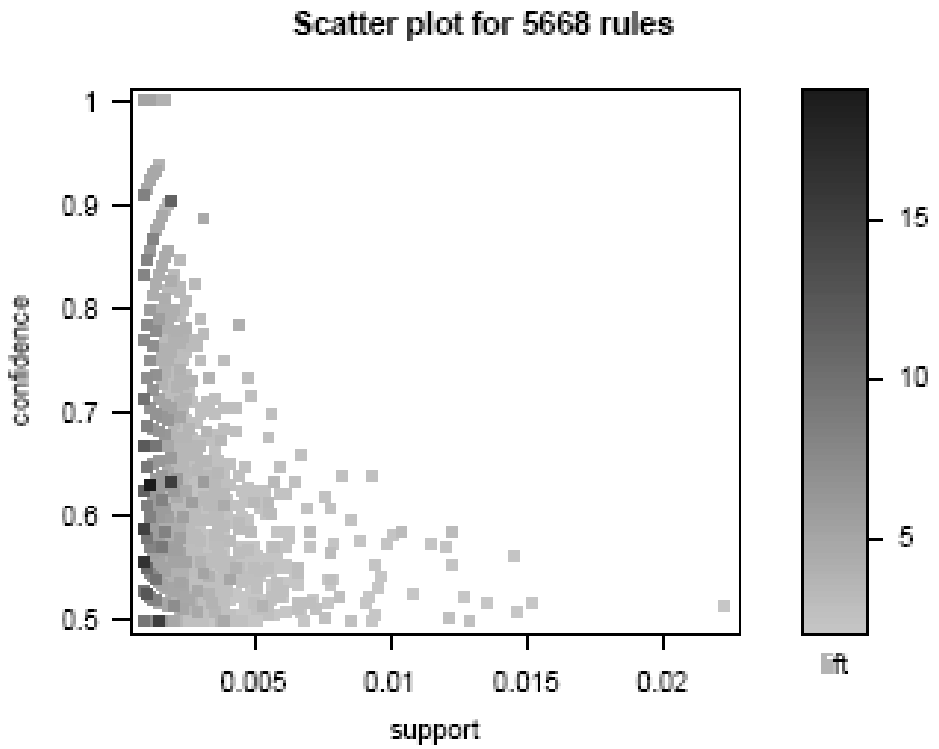
Visualización de reglas de asociación



Las reglas se ordenan de izquierda a derecha ordenadas por soporte y por confianza

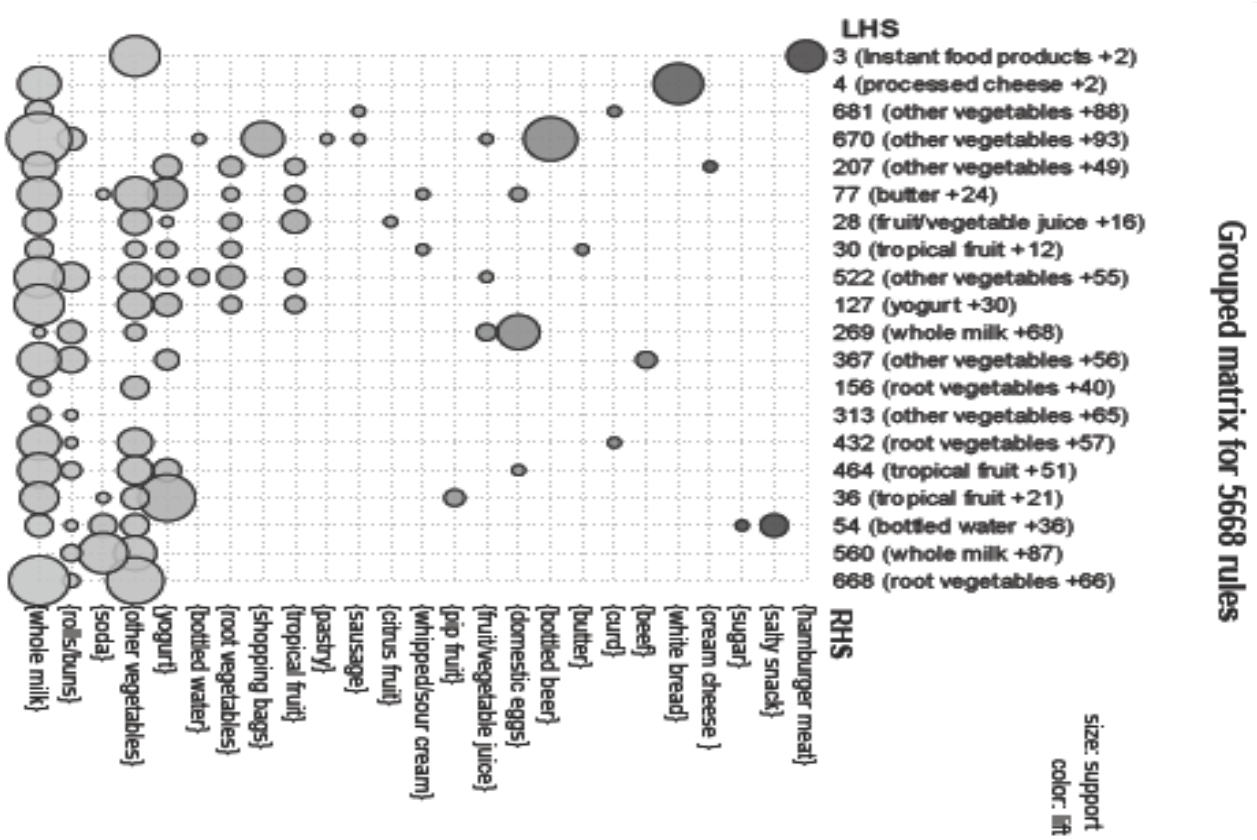
Visualización de reglas de asociación

- Ejemplo:
  - Arulesviz Package



Visualización de reglas de asociación

- Arulesviz Package -> reglas de asociación en clústeres



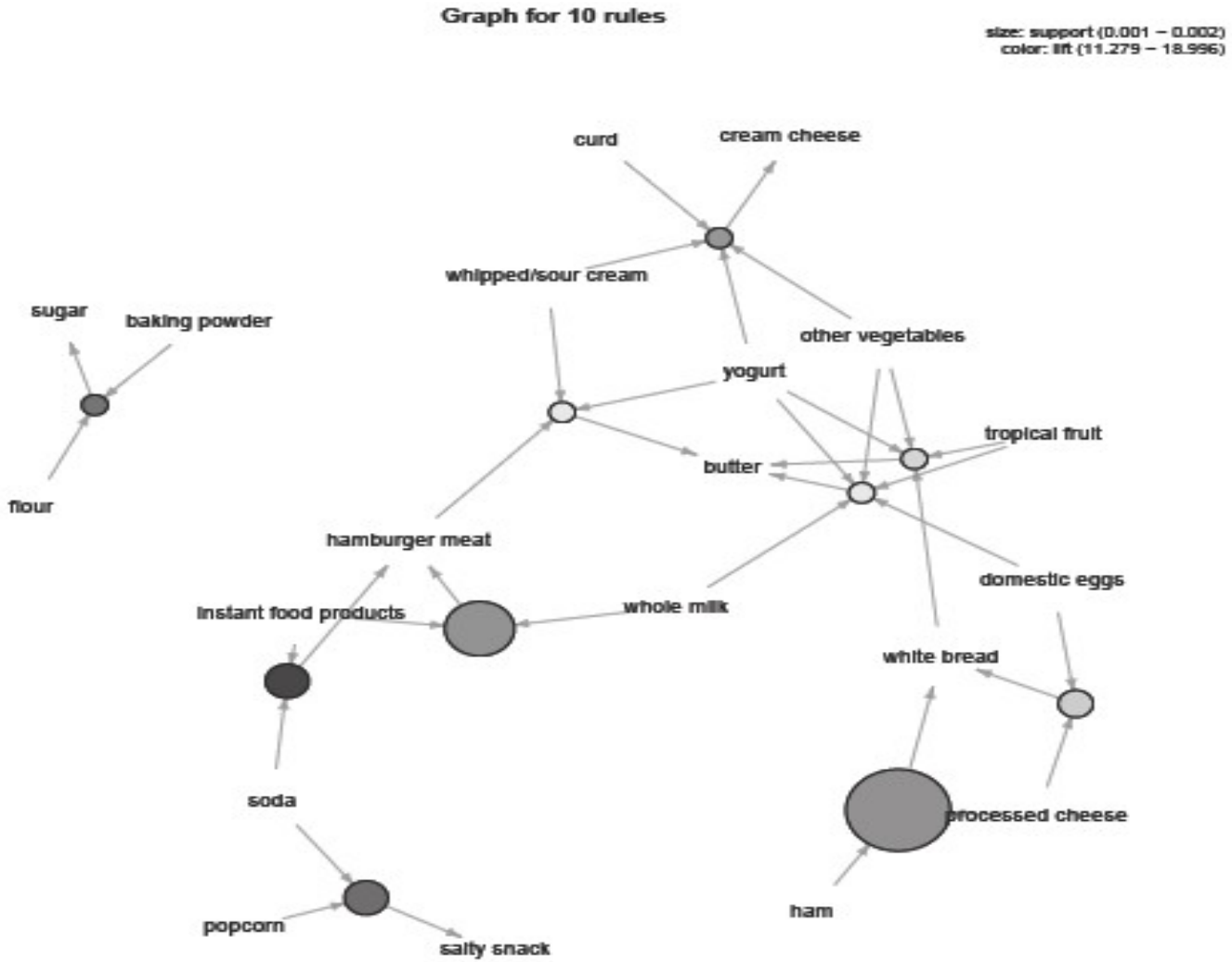
Visualización de reglas de asociación

- Arulesviz Package (CRAN) -> reglas de asociación para R.

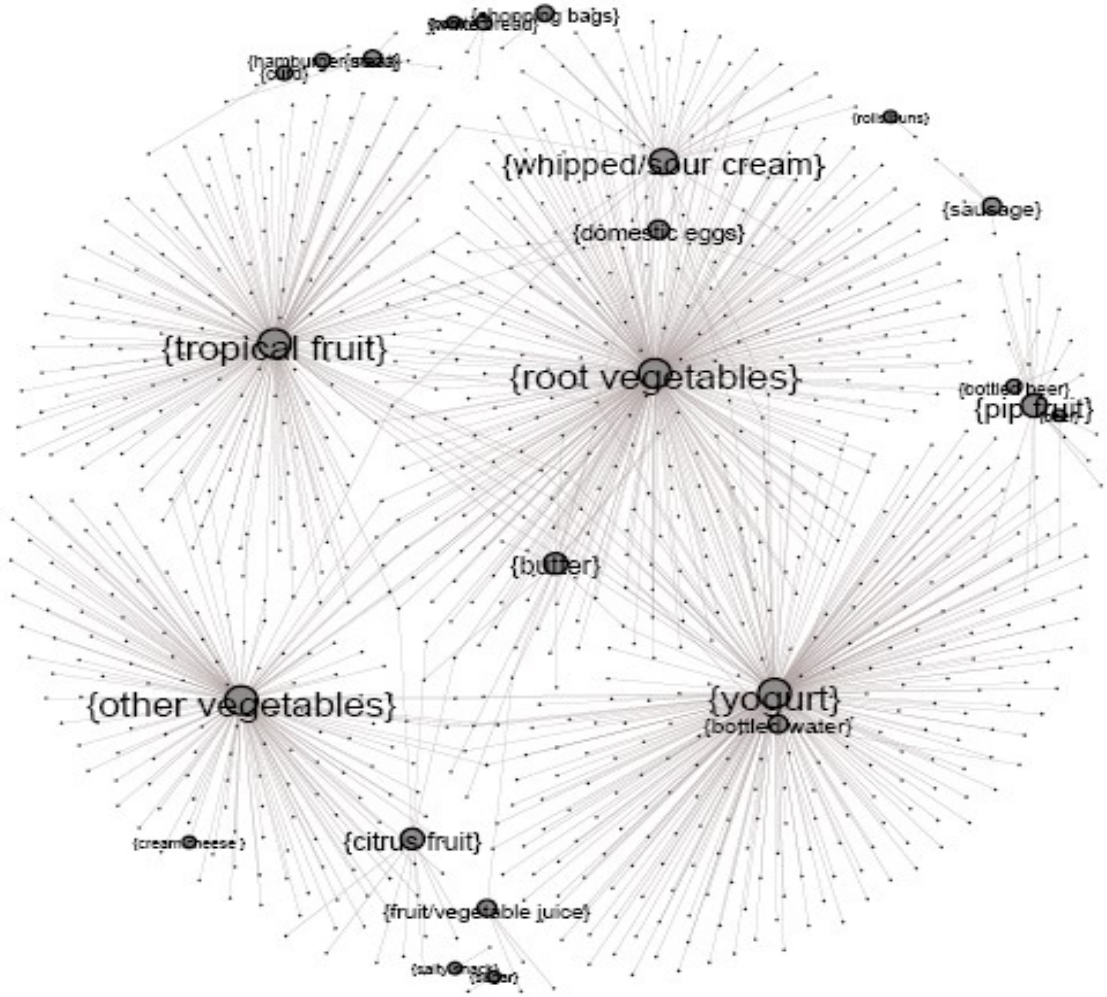
  https://cran.r-project.org/web/packages/arulesViz/index.html

- Visualización:
  - Los vértices representan items o itemsets
  - Las flechas representan relaciones en las reglas
  - El nivel de gris mide la fuerza de la asociación
  - Esta visualización no se recomienda para grupos de reglas muy grandes

Visualización de reglas de asociación



Visualización de reglas de asociación



Tag clouds

- Conceptualmente son similares a los histogramas
- Utilizan diferentes fuentes para representar el texto en base a:
  - Tamaño
  - Fuente
  - Orientación del texto (vertical u horizontal)
  - La proximidad de unas palabras a otras

Tag clouds

document
ribbon
quickly
want
performance
clicking
files
set
photo
see
design
task
new
first
easy
wufoo
list
well
include
within
even
allows
working
multiple
mobile
much
provides
tasks
simple
control
approach
fotoflexer
interaction
visual
feedback
feature
work
number
given
including
menu
software
experience
system
specific
users
less
color
large
information
computer
like
create
dialog
instance
end
two
additional
right
pane
help
user
web
hale
campaign
features
always
focus
problem
development
applications
time
order
file
sugarsync
error
access
print
interface
without
team
editing
buttons
options
usability
many
content
designed
process
elements
management
one
page
supersaas
customers
sql
testing
alerts
product
data
available
good
controls
also
different
changes
used
text
main
common
view
pages
server
company
form
use
key
functions
using
menus
possible
way
display
several
application
rather
dashboard
function
screen
although
button
since
manager
creating
wizard
get
need
cmsbox
look
level
just
table
navigation
online
important
make
xero
program
forms
familiar

Tag Clouds

- Ejemplos:
  - Generador de TagClouds On-line

    http://tagcrowd.com/

  - Visualizador de Tag Clouds de Tweets

    http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

Referencias y Bibliografía

- [Chen et al., 1996] CHEN, M. S., PARK, J. S., AND YU, P. S. 1996. Data mining for path traversal patterns in a web environment. In Proceedings of the Sixteenth International Conference on Distributed Computing Systems (May), 385–392.
- B. Liu: Web Data Mining: Exploring Hyperlinks, Contents and usage data. Springer, 2011