Hola, tesis doctoral en informtica !!!

Este es el ndice de mi ltimo informe:

Es el contenido del informe que envi a Fede el 21 de julio. Ha llovido mucho desde entonces pero hay mucho material til, al menos hasta 3.2.

ndice general

Argumentacin

Este...

SRW

Los SRW surgen de...

Catlogos

Un catlogo es...

Captulo 1

Sistemas de Recomendacin Web

En ...

1.1. Personalizacin de la Web

1.2. Minera de Uso Web

1.2.1. DATOS

1.2.2. Seleccin

1.2.3. Preproceso

1.2.4. Transformacin

1.2.5. Minera de Datos (DM)

1.2.6. Evaluacin e Integracin

1.2.7. CONOCIMIENTO

1.3. Minera de Datos

1.3.1. Mapas de Navegacin Web

1.3.2. Reglas de Asociacin

1.4. Publicaciones

1.4.1. Actas de HCII05

1.4.2. Actas de Interaccin05

1.4.3. Actas de SICO05

Captulo 2

Minera de Reglas de Asociacin (ARM)

Una breve introduccin para enlazar con el captulo anterior y paso a relatar...

2.1. Conceptos bsicos

2.1.1. Tipo de Datos

...

2.1.2. Primeros algoritmos

...

2.1.3. Formato de D

...

2.1.4. Fases de ARM

...

2.2. Minera de Itemsets Frecuentes

Frequent Itemset Minning...

2.2.1. Algoritmos y estructuras

...

2.2.2. Evaluacin de diferentes implementaciones

...

2.3. Generacin de Reglas de Asociacin

...

2.3.1. `genrules()`

...

2.3.2. Apriori2

...

2.4. El tem Raro

...

2.4.1. Estudio de tems raros

...

2.4.2. Reglas de Oportunidad

...

2.5. Publicaciones

...

2.5.1. HCII07

...

2.5.2. ESWA, vol. 35(3) 2008

...

2.5.3. Interaccin10

...

Captulo 3

ARM sobre Catlogos

Los trabajos expuestos en los captulos anteriores muestran una dificultad presente en muchas investigaciones en el mbito de la informtica, la imposibilidad de comprobar si los resultados obtenidos son correctos y aplicables con la tecnologa actual. Todas las Ciencias recogen una Teora que la sustenta, la Informtica tambin, pero las demostraciones tericas basadas en otras Ciencias no siempre son vlidas para la Informtica. Hay muchos artculos tericos sobre Minera de Datos, pero algunos de ellos no evolucionan en un artculo posterior que muestre cmo se ha realizado el experimento con la tecnologa actual usando datos reales. Es fcil calcular tericamente el nmero de reglas de asociacin presente en una coleccin concreta de datos, pero si el algoritmo propuesto no es capaz de almacenar en RAM todas las reglas de asociacin del problema no servir de nada ese algoritmo en esa situacin. `mushroom` fue el primer caso que encontr curioso en el ambicioso campo de la Minera de Datos, una coleccin de tan solo 5 644 registros de 23 valores que slo contena 100 valores diferentes no poda ser analizada a fondo por el mejor algoritmo de ARM que conozco, Apriori, cuando el nmero de transacciones distintas que se puede obtener bajo estas circunstancias es X XXX XXX y con ellas tenemos un mximo de X XXX XXX XXX XXX reglas de asociacin. La tecnologa actual de un equipo de escritorio no puede gestionar en RAM tanta informacin, pero me negaba a creer que no se pudiera extraer *toda la informacin* que contuviera un problema tan pequeo. Al profundizar en `mushroom` y los artculos que lo mencionaban y otras colecciones de datos que encontr publicadas en los mismos portales encontr un modo de indicar al algoritmo de ARM que estoy tratando con un tipo de colecciones de datos especial. No es simplemente una coleccin de transacciones como las analizadas en 2.1 si no que stas tienen unas restricciones muy fuertes en su definicin.

Este informe refleja el trabajo que he realizado para descubrir lo suficiente como para ser merecedor del ttulo de doctor en Informtica. Los anteriores captulos reflejan una gran labor de documentacin y exposicin cientfica de los resultados que poda obtener con colecciones de datos fijas, no dispona de un servidor capaz de poner a prueba nuestras aportaciones y realizar sugerencias en tiempo real a un gran nmero de usuarios. Poda comprobar que mis clculos

se podan realizar con la tecnologa actual y las colecciones de datos que yo manejaba pero no saba qu ocurrira si aumentaran mis colecciones de datos o si pudiera realizar sugerencias en tiempo real. Es un buen trabajo terico apoyado en algunas realizaciones prcticas pero del que no puedo extraer an la conclusin que es un buen trabajo de investigacin en Informtica.

3.1. Catlogo

Los catlogos son colecciones de registros preparadas para resolver informticamente un problema de clasificacin. Y muchos investigadores de esta especialidad publican sus datos para que otros investigadores puedan hacer pruebas con las mismas condiciones de partida: una coleccin de datos con ciertas caractersticas. En UCI, KEEL, LUCS... encontraremos muchos catlogos entre los datasets que publican para resolver problemas de clasificacin.

Acabo de descubrir LUCS, que discretiza las colecciones de UCI y me ofrece 97 valores distintos en adult, frente a los 27 245 que tiene el de UCI, he de analizarlo con mi cdigo y EXPLICAR MEJOR LAS CONSECUENCIAS DE APLICAR ANTES O DESPUS MI MTODO O LA AGRUPACIN DE VALORES EN ATRIBUTOS NUMRICOS ya que se obtendrn reglas y catlogos completos bastante diferentes, esto da para otro artculo y ms si tengo en cuenta que tiene datos missing por lo que puedo obtener catlogos completos usando menos atributos con ms registros o catlogos completos usando slo los atributos registrados en cada registro (a no ser que el anlisis nos diga que cierto atributo no aporta informacin....

Cuando no sabamos que esos ficheros contenan catlogos intentbamos aplicar bien conocidos algoritmos de ARM pero no podamos extraer informacin que contienen los datos porque se desbordaba la RAM del equipo en que se est aplicando el algoritmo y se abortaba el proceso tras horas de clculos que finalmente no obtenamos. Esto nos sorprenda porque el primer catlogo que intentamos analizar con Apriori slo tiene 5 644 registros de 23 datos, no son nmeros excesivos para un problema de Minera de Datos analizado con un ordenador de escritorio con cierta potencia y capacidad de RAM. Eso nos llev a descubrir cmo se cre el catlogo a travs de `UCI/mushroom`...

Los catlogos caracterizan un problema de clasificacin concreto. Si queremos plantear otro problema de clasificacin, bien etiquetando a los mismos individuos en otras clases o bien utilizando atributos diferentes no podemos utilizar directamente cualquier catlogo que tengamos sobre la misma poblacin. Si los dos problemas usaran los mismos atributos pero diferentes clases y las clases en estudio son independientes no servir de nada la informacin que tengamos sobre los catlogos completos del primer problema de clasificacin si no sabemos analizar qu informacin puede ser relevante y cul no, de hecho la informacin menos relevante en esta situacin es la distribucin de las clases en cada uno de los problemas de clasificacin por lo que debemos huir de interpretaciones errneas utilizando estos datos para estimar soportes o confianzas poblacionales.

De un catlogo se puede extraer informacin vlida para otro problema de clasificacin que utilice los mismos atributos ya que si en la muestra en que se basa el catlogo no presenta cierta relacin entre los valores de los atributos YA SABEMOS QUE NO APARECER ESA RELACIN AUNQUE CAMBIEMOS DE CLASES (siempre que el catlogo sea vlido, an tengo que hacer muchas definiciones sobre muestra, poblacin, distribucin de clases, problema de clasificacin, atributos, clases, catlogos, catlogos completos, validez de un catlogo...).

Aunque la ARM busca cualquier relacin entre cualquier par (o k-itemset) de valores de D, el objetivo del problema de clasificacin es siempre el mismo, etiquetar cada registro con una clase basndose en la informacin disponible sobre otros registros con valores idnticos en sus atributos.

3.2. Catlogo comprimido

Aprovechando las restricciones implcitas de los catlogos como `mushroom`...

3.2.1. Lectura de catlogos comprimidos

3.3. Catlogo completo

En [...] expusimos...

3.4. Publicaciones

3.4.1. Interaccin'12

En Interaccin'12 expusimos...

3.4.2. [...]

En [...]

3.4.3. [...]

Estamos terminando [...]

Captulo 4

Conclusiones y Trabajo Futuro

Las mejores ideas expuestas en esta tesis son muy simples. Desde el primer algoritmo que entra en juego, Apriori, hasta la elaboracin de catlogos completos nfimos son ideas muy simples que implementadas de forma eficiente pueden hacer lo que se le pide a la Minera de Datos: buscar una aguja en un pajar.

Los catlogos completos tienen un potencial fcil de descubrir mediante sencillas tcnicas informticas de Minera de Datos. Este trabajo presenta una teora en torno a un tipo de datos muy utilizado que posibilita la obtencin extrema de la informacin que contienen grandes colecciones de datos utilizando la tecnologa actual en tiempo real.

Los datos bien recogidos reflejan el estado actual del mundo que nos rodea, por eso es importante poder analizarlos rpidamente utilizando en algunos casos informacin histrica sobre el mismo problema o bien partiendo de un nuevo problema y analizando rpidamente las caractersticas de los datos que proporciona su estudio. Si sabemos qu puede descubrir la Minera de Datos a partir de la observacin de los datos que hemos recogido podremos crear algoritmos que descubran lo que estamos buscando en tiempo real y con un uso aceptable de recursos de un servidor dedicado.

Este trabajo presenta unos antecedentes que encaminan al investigador a descubrir, quiz por casualidad, las caractersticas especiales de un modelo matemtico de almacenamiento de informacin y el uso que se est dando a estas colecciones de datos por parte de especialistas en el problema de clasificacin. La aparicin del problema del Minado de Reglas de Clasificacin Asociativas en [...] era previsible, todas las reglas de asociacin tienen un aspecto muy simple que sugiere a cualquier investigador que puede ser utilizado en el problema de clasificacin. El hecho de que yo, especializado en el problema de asociacin, observara los mismos datos que los especialistas en clasificacin tendra que llevarnos al mismo resultado si ellos haban alcanzado el ptimo o a un mejor resultado si yo era capaz de aportar ideas sobre cmo utilizar los elementos de ARM.

El primer descubrimiento simple y til de esta tesis son los *catlogos comprimidos* expuestos en la seccin 3.2. Con ellos descubr que el modo de aplicar tcnicas de ARM en los artculos que consultaba no era del todo correcto [...]. No soy especialista todava en el Problema de Clasificacin por lo que algunas conclusiones de esos artculos y, sobre todo, las pruebas de eficiencia de los algoritmos que proponan, estaban fuera de mi alcance. Se me ocurri incorporar las restricciones iniciales del problema de clasificacin a un problema general de asociacin. Los problemas de asociacin se resuelven mediante la fuerza bruta leyendo todos los datos que tenemos y mirndolos desde distintas perspectivas, si quiero resolver un problema distinto, un problema de clasificacin, usando tcnicas de minera de reglas de asociacin debera aprovechar, al menos, la rgida estructura de los datasets usados para clasificacin (en asociacin slo hay una norma: en un registro no se cuentan los datos repetidos, lo que hace que el nmero de reglas de asociacin que se puede buscar sea tan grande que provoque desbordamiento de memoria en los programas que intentan analizar grandes colecciones de datos). Se me ocurri que si todos los registros han de tener un valor para cada uno de los atributos en estudio poda reducir el nmero de datos a procesar y las dimensiones del dataset eliminando nicamente un valor de cada atributo en todo el dataset. Al hacerlo y comprobar que la nueva coleccin de datos, compresin sin prdidas de la coleccin original, s se poda analizar utilizando el clsico Apriori y obtener todas las reglas de asociacin que contena empec a asimilar mejor las caractersticas de un catlogo.

Primero descubr caractersticas matemticas, restricciones tericas que me permitan reducir las dimensiones del problema original y, usando muchos recursos, obtener toda la informacin que contienen esas pequeas colecciones de datos en cuanto a reglas de asociacin se refiere. Pero tena que haber algo ms, las caractersticas matemticas que utilic en 3.4 me exigan usar muchos recursos y no me ofrecan informacin demasiado relevante, adems segua necesitando mucha RAM para trabajar con colecciones pequeas de datos, a pesar de que ya saba que contenan muchsima informacin. Quera encontrar mejor informacin en menos tiempo y usando menos RAM por lo que introduje la STL a mi desarrollo y comprob en la primera aplicacin que la teora de conjuntos tena mucho que aportar al anlisis de catlogos.

Tantos aos de trabajo han dado lugar a muchas ideas tericas sobre la aplicacin de tcnicas de DM por lo que quedan abiertas muchas lneas de investigacin que podran ser continuacin de este trabajo. Como *trabajar a nivel de* `bits` buscando la mxima eficiencia en el uso informtico de grandes colecciones de datos, o *profundizar en el desarrollo de Clasificadores*, de *lgica difusa para agrupacin de valores en atributos numricos o de amplios rangos* y de tantas otras cosas que han ido apareciendo en el estado del arte de esta tesis y que no he podido abarcar para centrarme en obtener algo tangible mediante el mtodo cientfico.

La investigacin mostrada en el ltimo captulo de esta tesis est avalada por su implementacin en el campo de la Minera de Datos utilizando la tecnologa actual. El preproceso de cualquier catlogo permite crear colecciones de catlogos que pueden ser utilizadas en tiempo real en grandes problemas de clasificacin que pueden ser escalados sin tener que renunciar en cada nuevo estudio a todo el

Es evidente que tengo que reescribir este prrafo. La idea es interesante pero...

conocimiento adquirido en estudios sobre las mismas clases. En este trabajo se ha demostrado que cualquier subconjunto de un catlogo completo puede ser tratado como catlogo completo considerando siempre la incertidumbre que puede contener, si quisiramos utilizar los datos de un problema de clasificacin en otro problema de clasificacin con otras clases podramos comenzar con los registros-tipo del primer catlogo, todos los que puedan ser clasificados en el segundo problema se incorporan al catlogo del segundo problema pero as no voy bien, lo que quera decir es que si empezamos con el menor de todos los catlogos nfimos y vamos catalogando en la segunda clase todos sus registros podemos llegar a no tener incertidumbre (caso ideal y poco probable si la segunda clase es independiente de la primera, dato interesante) pero al menos si tenemos incertidumbre es posible que sea poca, si hacemos lo mismo con otros catlogos nfimos podramos descubrir qu atributos aportan ms determinacin al segundo problema y plantear un catlogo inicial para el segundo problema.

Tambin queda para el futuro la agrupacin de valores en los atributos numricos. Hay ya muchas investigaciones en torno a este campo y creo que con los primeros anlisis hechos a un dataset se puede obtener informacin que pueda ayudar al investigador a hacer las agrupaciones de modo que se pueda seguir trabajando con catlogos completos ya que el agrupamiento puede generar incertidumbre. Este aspecto es muy importante pero es mucho lo que hay que investigar para llegar a conclusiones y resultados tiles, como en “Using Conjunction of Attribute Values for Classification”.

Apndice A

Notacin

La notacin usada en este informe se ha intentado ajustar a la ms utilizada en la bibliografa revisada a lo largo de estos aos de investigacin. Por este motivo no es uniforme en los tres captulos de investigacin en que se divide esta tesis.

A.1. Sistemas de Recomendacin Web

En este captulo...

A.2. Minera de Reglas de Asociacin

En este captulo...

A.3. Catlogos

En este captulo...

Apndice B

Cdigo

Se ha desarrollado mucho cdigo para poder comprobar todo lo que se afirma en esta tesis. Se ha trabajado del modo ms estndar posible para conseguir un cdigo eficiente y que pueda ser incorporado a otras investigaciones. Se publicar como cdigo abierto bajo la licencia... en...

B.1. Sistemas de Recomendacin Web

En este captulo...

```
c+cpinclude stdlib
c+cpusing std::cout;
c+cpusing std::endl;

k+ktint n+nfmainp()
p
   ncout o l+sHola, mundo o nendlp;

   kreturn l+m+mi0p;
p
```

B.2. Minera de Reglas de Asociacin

En este captulo...

B.3. Catlogos

En este captulo...

Apndice C

Datos utilizados

Para llevar a cabo las pruebas de rendimiento y aplicabilidad de nuestras propuestas se han usado datos propios y datos procedentes de diferentes repositorios pblicos como UCI, KEEL, LUCS-KDD...

C.1. Sistemas de Recomendacin Web

En este captulo usamos datos de un servidor propio con la intencin de poder utilizar las recomendaciones sugeridas por nuestra metodologa en el servidor del que se obtuvieron. Son los ficheros `TAL` y `CUAL` que no publicaremos por carecer de inters su contenido, ya que el servidor del que se obtuvieron ya no est disponible y no se podra dar ninguna interpretacin a los resultados obtenidos...

Tambin usamos `TAL`...

C.2. Minera de Reglas de Asociacin

En este captulo seguimos trabajando con los mismos datos que en el anterior e incorporamos...

C.3. Catlogos

En este captulo es en el que ms opciones hemos tenido a la hora de seleccionar datos y probar la eficiencia y posibilidades de nuestros desarrollos. Existen muchos repositorios pblicos bien documentados sobre el diseo y contenido de estos datasets y es una informacin que enriquece mucho la investigacin...

ndice de figuras

ndice de cuadros